JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

ISBN 978-88-3293-137-2

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédrick Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F
Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Tolouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I
Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingo: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Abstracts

Introduction

The *International Conference on the Statistical Analysis of Textual Data* (JADT, Journées d'Analyse Statistique des Données Textuelles) has been at its 14th edition. It was held for the third time in Rome, from 12 to 15 June 2018, organized by the DII - Department of Enterprise Engineering "Mario Lucertini" at Tor Vergata University of Rome and the DSS - Department of Statistical Sciences at Sapienza University of Rome. This biennial conference has continuously gained importance since its first occurrence in Barcelone (1992), and with the editions of Montpellier (1994), Rome (1996), Nice (1998), Lausanne (2000), Saint-Malo (2002), Louvain-la Neuve (2004), Besançon (2006), Lyon (2008), Rome (2010), Liège (2012), Paris (2014), Nice (2016). Every two years, the JADT conference presented the state of the art concerning theories, problems, methods, algorithms, software and applications in several domains, sharing a quantitative approach to the study of lexical, textual, pragmatic or discursive features of information expressed in natural language.

The proceedings of the 2018 Conference collect 113 contributions by 243 scholars from 15 countries spread all over the world. These papers include contributions open to all scholars and researchers working in the field of textual data analysis, ranging from lexicography to the analysis of political discourse, from information retrieval to marketing research, from computational linguistics to sociolinguistics, from text mining to content analysis. The invited speakers focused on the central topics of the conference, discussing open and new themes, e.g. machine learning algorithms to profiling users of social media, new multilingual approaches, textometry, and authorship. The proceedings follow an alphabetical order by the surname of the first author of the contributions.

In this edition, several innovations have been introduced with respect to the past. In a roundtable, we discussed the past, present and future of Statistical Analysis of Textual Data and Text Mining methods, by examining the point of view of Universities and enterprises. The papers, which followed a review process carried out with two and sometimes three reviewers, are maximum of 6 pages. The idea is that the papers were not yet in their final version, and the exchange with other scholars during the conference led to an improvement. For the first time, a selection of extended papers presented at

the JADT conference will be published, after another reviewing process, in a book published by Springer and in several special issues of acknowledged Journals (Advances in Data Analysis and Classification, International Review of Sociology, Italian Journal of Applied Statistics, Social Indicators Research, RPC Rivista di Psicologia Clinica). The perspective of enhancing the papers discussed during JADT conference will allow the scholar community to keep the network of active contacts and lively exchanges.

D. Fioredistella Iezzi, Livia Celardo, Michelangelo Misuraca

Acknowledgements

We express our gratitude to the 56 reviewers who offered their assistance in selecting and anonymously reviewing the papers of this volume: Massimo Aria, Barbara Baldazzi, Nadia Battisti, Valérie Beaudouin, Sergio Bolasco, Etienne Brunet, Mónica Bécue, Isabella Chiari, Livia Celardo, Michele Cortelazzo, Pasquale Del Vecchio, Francesca Della Ratta, Fiorenza Deriu, Anne Dister, Francesca Dolcetti, Annick Farina, Serge Fleury, Andrea Fronzetti, Luca Giuliano, Peter Gloor, Francesca Greco, Francesca Grippa, Serge Heiden, D. Fioredistella Iezzi, Antonio Iovanella, Sylvia Kasparian, Margareta Kastberg, Dominique Labbé, Ludovica Lanini, Alexei Lavrentev, Ludovic Lebart, Jean-Marc Leblanc, Alain Lelu, Dominique Longrée, Véronique Magri, Pascal Marchand, Damon Mayaffre, Sylvie Mellet, Silvia Micheli, Michelangelo Misuraca, Denis Monière, Gianluca Murgia, Pa-squale Pavone, Bénédicte Pincemin, Céline Poudat, Pierre Ratinaud, Piergiorgio Ricci, Maria Francesca Romano, Johanne Saint-Charles, André Salem, Massimiliano Schiraldi, Max Silberztein, Maria Spano, Arjuna Tuzzi, Mathieu Valette, Ramón Álvarez Esteban.

JADT2018 was held under the patronage of ISTAT (Istituto Nazionale di Statistica - National Institute of Statistics). We are also very grateful to the following sponsors: ISTAT, Le Sphinx, The Information Lab, Master in Data Science at Tor Vergata University, Prisma.

As regards the organisation of the conference, we would like to thank all the members of the local organising team: Francesco Alò, Silvia Castellan, Giulia Giacco, Paolo Meoli, Vittorio Palermo, Viola Talucci.

Special thanks go to Livia Celardo, Isabella Chiari, Andrea Fronzetti Colladon, Francesca Della Ratta, Fiorenza Deriu, Francesca Dolcetti, Francesca Greco, for the organisation of special tracks concerning Official Statistics, Linguistics, Applications on social and psychological domains, Social Network and Semantic Analysis.

Invited Speakers

Memorize or generalize? Searching for a compositional RNN in a haystack Adam Liška

German Kruszewski
Facebook- germank@fb.com

Abstract
Machine learning systems have made rapid progress in the past few years, as evidenced by the remarkable feats they have accomplished on fields as diverse as computer vision or reinforcement learning. Yet, as impressive as these achievements are, they rely on learning algorithms that require orders of magnitude more data than a human learner would. This disparity could be rooted in many different factors. In this talk, we will draw on the hypothesis that compositional learning — that is, the ability to recombine previously acquired skills and knowledge to solve new problems– could be one important element of fast and efficient learning (Lake et al, 2017). In this direction, we will discuss our ongoing efforts towards building systems that can learn in compositional ways. Concretely, we will present a simple benchmark based on function composition to measure the compositionality of learning systems and use it to draw insights for whether current learning systems learn or can learn in a compositional manner.

Scaling-up Sentiment Analysis through Continuous Learning

Bing Liu
University of Illinois at Chicago - liub@uic.edu

Abstract

Sentiment analysis (SA) or opinion mining is the computational study of people's opinions, sentiments, emotions, and evaluations. Due to numerous research challenges and almost unlimited applications, SA has been a very active research area in natural language processing and text mining. In this talk, I will first give a brief introduction to SA, and then move on to discuss some major difficulties with the current technologies when one wants to perform sentiment analysis in a large number of domains, e.g., all products sold in a large retail store. To tackle the scaling-up problem, I will describe our recent work on lifelong machine learning (LML) (or lifelong learning) that tries to enable the machine learn like humans, i.e., learning continuously, retaining or accumulating the knowledge learned in the past, and using the knowledge to help future learning and problem solving. This paradigm is quite suitable for SA and can help scale up SA to a large number of domains with little manual involvement.

La textométrie comme outil d'expertise : application à la négociation de crise.

Pascal Marchand
Université de Toulouse – pascal.marchand@iut-tlse3.fr

Résumé
Pour aborder la pertinence de la pratique textométrique dans des problématiques de terrains et comme outil d'expertise, on étudiera les échanges réels impliquant les négociateurs des Forces d'intervention de Police, dans des contextes de barricades, prises d'otages, terrorisme ou intention suicidaire à haut niveau de dangerosité.

Nous envisagerons donc la négociation au travers des dynamiques de choix lexical et nous chercherons à cartographier le lexique, classer des segments de textes et comparer des profils de locuteurs et de situations.
On se propose ainsi de répondre aux questions suivantes :

- Y a-t-il des thèmes récurrents dans les crises ?
- Y a-t-il une chronologie lexicale de la crise ?
- Comment se gèrent les émotions ?
- Quelles sont les spécificités des situations « radicalisées » ?

L'objectivation des échanges et la mise en évidence des séquences formelles peut alors fournir une aide au diagnostic, dans le but de tirer des éléments concrets pour des objectifs de retour d'expérience et de formalisation des pratiques des professionnels de la négociation.

Author Identification Combining Various Author Profiles. Towards a Blended Authorship Attribution Methodology

George K. Mikros
National and Kapodistrian University of Athens – gmikros@gmail.com

Abstract

The aim of this presentation is to describe a new method of attributing texts to their real authors using combined author profiles, modern computational stylistic methods based on shallow text features (n-grams) and machine learning algorithms. Until recently, authorship attribution and author profiling were considered similar methods (nearly identical feature sets and classification algorithms), but with different aims, i.e. in the former to identify the author's identity and in the latter to detect author's characteristics such as gender, age, psychological profile etc. Both of these methods have been used independently aiming at different research aims and in different real-life tasks. However, in this talk we will present a unified methodological framework where standard authorship attribution methodology and author profiling are combined so that we can approach more effectively open or semi-open authorship attribution problems, a category known as authorship verification which is particularly difficult to tackle with present computational stylistic methods. More specifically, we will present preliminary research results from the application of this blended methodology to a real semi-open authorship problem, the Ferrante's authorship case. Using a corpus of 40 modern Italian literary authors compiled by Arjuna Tuzzi and Michele Cortelazzo from the University of Padua (Tuzzi & Cortelazzo, under review), we will explore the dynamics of author profiling in gender, age and region and various methods we can combine the extracted profiles so that we can entail the identity of the real author behind Ferrante's books. Moreover, we will extend this methodology and validate its usefulness in social media texts using the English Blog Corpus (Argamon, Koppel, Pennebaker, & Schler, 2007). Using, simulated scenarios of authorship attribution cases (the real author to be included in the training data and the real author to be missing from the training corpus) we will further evaluate the usefulness of the proposed blended methodology which can lead to some exciting new possibilities for investigating author identities in both closed and open authorship attribution tasks.

From text to concepts and back: going multilingual with BabelNet in a step or two

Roberto Navigli
Sapienza University of Rome – roberto.navigli@uniroma1.it

Abstract
Multilinguality is a key feature of today's Web, and it is this feature that we leverage and exploit in our research work at the Sapienza University of Rome's Linguistic Computing Laboratory, which I am going to overview and showcase in this talk. I will describe the most recent developments of the BabelNet technology. I will introduce BabelNet live – the largest, continuously-updated multilingual encyclopedic dictionary – and then discuss a range of cutting-edge industrial use cases implemented by Babelscape, our Sapienza startup company, including: multilingual interpretation of terms; multilingual concept and entity extraction from text; cross-lingual text similarity.

Contributors

Identification automatique de l'ironie et des formes apparentées dans un corpus de controverses théâtrales

Motasem Alrahabi[1], Chiara Mainardi[2]
[1] Université Paris-Sorbonne Abu Dhabi – motasem.alrahabi@gmail.com
[2] Université Sorbonne Nouvelle – chiara.mainardi@univ-paris3.fr

Abstract
This paper presents the results of an automatic analysis on a corpus of French texts about theatre debates (16th –19th centuries). The purpose of this study is to highlight the important role of different forms of irony in the theatre controversy and to reveal the stand point of authors and established authorities towards theatre performances. Despite the difficulty of this task, our research shows encouraging results. This unprecedent comparison of these kind of texts, in which authors condemn the theatre or approve it, enables to a broader understanding of the authors' positions, arguments and rhetorical strategies relating to theatre controversies.

Résumé
Cet article présente les résultats de notre analyse automatique d'un corpus de débats sur le théâtre (16ème – 19ème siècle). L'objectif de cette étude est d'illustrer le rôle important que jouent les différentes formes de l'ironie dans la polémique autour du théâtre et de mettre en évidence la position des auteurs ou des autorités antiques citées vis-à-vis des spectacles. Les résultats obtenus sont encourageants malgré la difficulté de la tâche et ils nous permettent de comparer d'une façon inédite les textes des auteurs défenseurs avec ceux des auteurs pourfendeurs du théâtre et d'avoir une meilleure compréhension de certains arguments et stratégies d'auteurs dans le champ de la controverse.
Keywords: Ironie, théâtre, marqueurs linguistiques, annotation sémantique, système à base de règles.

1. Introduction

Nous proposons une analyse automatique d'un corpus en français qui rassemble des débats sur le théâtre depuis le milieu du 16e siècle jusque dans les années 1840. Notre objectif est d'illustrer le rôle important que jouent les expressions de l'ironie dans la polémique autour du théâtre et de mettre en évidence la position des auteurs ou des autorités antiques citées vis-à-vis des

spectacles. Nous présentons d'abord les ressources linguistiques développées, l'outil d'annotation utilisé et le corpus ; ensuite, nous commentons les résultats d'analyse automatique et, avant de conclure, nous explorons les perspectives de ce projet en cours.

2. Prémisses sur l'ironie

L'ironie est un fait de langue utilisé afin de transmettre un message directement ou indirectement opposé à ce qui est dit littéralement. Largement étudiée en philosophie, en rhétorique ou en linguistique (Berrendonner, Sperber et Wilson, Kerbrat-Orecchioni, Ducrot, Grice…), l'ironie représente un concept hétérogène extrêmement difficile à définir du fait de ses nombreuses formes et de la complexité des phénomènes qui sont en jeu. L'ironie fonctionne à l'aide d'indices laissés par le locuteur à l'interlocuteur pour lui faire comprendre ses intentions par des jeux de parallélismes, de contradictions, d'exagérations et d'hyperboles plus ou moins marqués. Ces indices – souvent pragmatiques ou extralinguistiques – sont plus ou moins évidents, d'où l'importance de la prise en compte du contexte (référentiel, locuteur, interlocuteur…), des connaissances partagées et des normes sociales et culturelles. La présente étude constitue la première étape pour une détection automatique du champ de l'ironie au sein de notre corpus. Conscients de la difficulté de la tâche et de l'absence de ressources linguistiques adaptées à notre corpus et à nos objectifs, nous nous sommes tournés vers une approche symbolique en nous basant sur un travail précédent autour de l'annotation automatique des modalités énonciatives (Riguet et Alrahabi, 2017). Employés dans les stratégies argumentatives, ces marqueurs observables aident à exprimer ou à rapporter l'ironie ou d'autres cas qui s'y apparentent (sarcasme, raillerie, satire, moquerie…). Exemple :

> *De sorte qu'on ne peut mieux définir la Comédie, qu'une « assemblée de railleurs où personne ne se connait, et où chacun rit des défauts qui les rendent tous également coupables et ridicules ». [Lelevel, 1694]*

Les marqueurs utilisés sont principalement des verbes comme *se moquer, ironiser, parodier…* Ensuite, par l'observation d'une partie du corpus, nous avons enrichi ces ressources par des substantifs, des adjectifs et des adverbes. Nous avons ensuite classé ces marqueurs dans des sous-catégories selon différentes nuances sémantiques : 1) ironie, dérision, se moquer, sarcastique, parodier… ; 2) chicaner, taquiner, narguer… ; 3) faire rire, comique, pitre, grotesque, idiot… ; 4) mordant, piquant, pinçant, aigre… ; 5) mépriser, dénigrer, sous-estimer, vilipender… ; 6) calomnier, hypocrisie, ruse, malice… ; etc. En tout, nous avons collecté autour de 70 marqueurs

linguistiques.

3. Méthodologie et choix techniques

La détection automatique de l'ironie est une tâche difficile, notamment à cause de la multitude des moyens linguistiques qui expriment, souvent de manières subtiles, l'ironie ou les autres formes apparentées. Différents travaux computationnels s'intéressent à la détection automatique de ces phénomènes linguistiques (Joshi et al., 2016): approches à base de règles, approches statistiques et approches d'apprentissage profond. Dans le présent projet, nous avons utilisé Excom2 (Alrahabi, 2010), un outil d'annotation à base de règles qui nous a permis d'avoir le contrôle sur le processus d'annotation et d'améliorer progressivement la pertinence des ressources linguistiques exploitées. Pour le système, la présence dans une phrase d'un marqueur de l'ironie déclenche les règles associées qui explorent le contexte et vérifient la présence ou l'absence de marqueurs complémentaires.

Dans la phrase qui suit, la présence de l'adverbe *moqueusement* dans le contexte d'un marqueur de parole permet à Excom2 d'attribuer à ce passage textuel l'étiquette « Ironie » :

« Il lui faut, dit-on moqueusement, cinq épithètes ! » [Corpus OBVIL]

Les règles dans Excom2 peuvent être organisées selon un ordre de priorité et utiliser en entrée les résultats d'autres règles. Avant l'étape de l'annotation, l'outil procède à la segmentation des textes afin de les découper en sections, paragraphes et phrases. Pour l'Ironie, nous avons créé 8 règles que nous avons associées aux différents marqueurs linguistiques.

4. Corpus

Cette présente étude s'appuie sur des textes à argumentation théâtrophile ou théâtrophobe, et sur des textes adoptant une stratégie « mesurée ». Cette dernière consiste à dénoncer des abus de la scène pour, ensuite, convaincre le lecteur à préserver l'utilité intrinsèque du théâtre. Ces trois types de textes possèdent une logique souvent détournée et déconcertante pour le lecteur : sous le déroulement des chapitres, on découvre parfois des connexions implicites, un usage de l'ironie très répandu et des phrases à la forme négative qui infléchissent notablement la détection des contenus. Avec ses reprises pour réitérer ou au contraire pour retourner l'argument contre l'adversaire, ce corpus de controverses théâtrales se prête bien à des analyses numériques. Le corpus rassemble 59 textes (*environ un million de mots*) écrits en langue française depuis le milieu du 16e siècle jusque dans les années

1840[1]. Ceux-ci ont été préalablement numérisés et édités dans le cadre du Labex OBVIL de Paris IV-Sorbonne et sont librement accessibles en ligne[2].

5. Evaluation

Une première phase de tests sur un échantillon du corpus a été nécessaire pour stabiliser les règles d'identification et de désambiguïsation. Afin d'évaluer la qualité des annotations obtenues, nous nous sommes focalisés dans un premier temps sur le calcul de la précision. Nous avons alors annoté avec Excom2 une autre partie du corpus (7 articles, 215675 mots) et nous avons obtenu 416 annotations. Ensuite, nous avons demandé à une personne qui connait les œuvres de cette période de juger les sorties du système selon un guide d'annotation. Pour chaque annotation, l'évaluatrice devait choisir entre : « Correct », « Incorrect » ou « Je ne sais pas ». Le critère d'évaluation était le suivant : est-ce que l'auteur du texte fait allusion à l'ironie dans la phrase en question? Nous avons obtenu 93.9 % de précision.

6. Difficultés rencontrées

Nous nous sommes heurtés à plusieurs difficultés. Au niveau du lexique, peu de changements ont été effectués sur nos marqueurs, comme par exemple le mot *satire* qui se trouve avec les deux orthographes *satire* (*88 occurrences*) et *satyre* (*68 occurrences*). En français, le dernier désigne le demi-dieu compagnon de Dionysos ou Bacchus. Cependant, dans certains textes qui n'ont pas encore été modernisés, et sont en langue française du 16e ou 17e siècle, ce mot indique plus largement la « satire ». D'un autre côté, certains marqueurs sont polysémiques et génèrent du bruit, comme *ridicule (437 occurrences, le marqueur le plus fréquent), plaisanter (176 occurrences)* et *comique (131 occurrences)*. Exemple [Rousseau, 1758] :

> *Le ridicule est l'arme favorite du vice. C'est par elle qu'en attaquant dans le fond des cœurs le respect qu'on doit à la vertu, il éteint enfin l'amour qu'on lui porte.*

Concernant la syntaxe du 17e et 18e siècle, nous avons observé une certaine complexité au niveau des phrases qui sont parfois très longues (*cinq lignes ou*

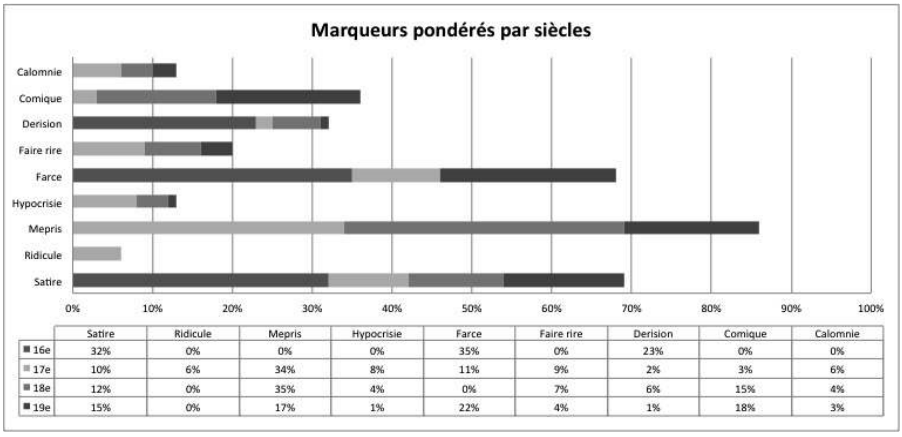
[1] Nous renvoyons à la liste de la bibliographie française qui constitue le corpus total de la Haine du Théâtre: http://obvil.paris-sorbonne.fr/corpus/haine-theatre/bibliographie_querelle-france/

[2] Il s'agit d'une partie du corpus de « La Haine du théâtre », projet dirigé, au sein du Labex OBVIL, par François Lecercle et Clotilde Thouret (Lecercle et al., 2016), http://obvil.paris-sorbonne.fr/projets/la-haine-du-theatre.

plus), et au niveau des signes de ponctuation qui ne sont pas stables. Plusieurs virgules, points virgules, *etc.* peuvent en effet se succéder dans une seule phrase. De plus, les auteurs de notre corpus utilisent des tournures complexes. Très souvent, ces phrases sibyllines sont ironiques, et cela se passe d'autant plus si elles se trouvent à la forme interrogative.

7. Interprétation des résultats

Dans l'étude des débats sur le théâtre, les expressions de l'ironie sont une voie d'entrée féconde dans le corpus. On constate d'abord que, tout au long des siècles couverts par le projet Haine du Théâtre (16e – 19e siècles), l'usage de l'ironie se situe entre les valeurs de 0,20 à 0,30 % (1265 annotations en total). Nous avons ensuite analysé les marqueurs de l'ironie en étudiant leur présence relative selon les siècles et nous avons pris en compte uniquement ceux ayant un pourcentage supérieur à 5% à l'intérieur d'un même siècle.



	Satire	Ridicule	Mepris	Hypocrisie	Farce	Faire rire	Derision	Comique	Calomnie
16e	32%	0%	0%	0%	35%	0%	23%	0%	0%
17e	10%	6%	34%	8%	11%	9%	2%	3%	6%
18e	12%	0%	35%	4%	0%	7%	6%	15%	4%
19e	15%	0%	17%	1%	22%	4%	1%	18%	3%

Figure 1 : Les marqueurs d'ironie dans le corpus HdT pondérés par siècle

Une baisse considérable a lieu au 17e siècle. S'il est prématuré d'en tirer des conclusions hâtives, nous pouvons cependant tout de suite constater que cela est probablement dû à l'affirmation de la religion, de l'ordre du classicisme ainsi qu'à l'autoritarisme étatique qui s'insinuait dans les esprits des écrivains de cette époque. En revanche, au fur et à mesure du 17e au 19e siècle, les valeurs de ces marqueurs augmentent de manière assez stable.

De manière générale, l'ironie est utilisée dans le corpus comme procédé éthique et stylistique, ce qui rend les auteurs bien efficaces dans l'élaboration de leur vision de la querelle. Qu'ils soient théâtrophobes ou théâtrophiles, ils peuvent jouer avec les nuances des marqueurs d'ironie, dissimuler un double-sens dans leurs phrases, s'exprimer figurément de manière contraire

à ce qu'ils communiquent littéralement. Par exemple, nous retrouvons une présence considérable du lemme « mépris » au 17e et 18e siècles. Il s'agit principalement d'un usage de l'ironie en tant que mécanisme de régulation de la vie sociale. Notamment, Conti et Voisin utilisent un humour inoffensif contre les excès de l'art et mettent en avant la bienséance :

> *Ceux qui vont aux Spectacles, non par hasard, mais de propos délibéré, et avec tant d'ardeur, qu'ils abandonnent l'Eglise par un mépris insupportable pour y aller, où ils passent tout le jour à regarder ces femmes infâmes, auront-ils l'impudence de dire qu'ils ne les voient pas pour les désirer [Conti 1667, Voisin 1671]*

L' « hypocrisie » commence à être utilisée au 17e et son utilisation se réduit avec le temps (jusqu'à 1% au 19e). Le lemme en question est essentiellement appliqué à des phrases où l'ironie n'est qu'un « autre nom du malheur » (Martin 2009), une manière de renforcer le point de vue de l'auteur.

> *L'hypocrisie est un vice privilégié, qui ferme la bouche à tout le monde, et qui jouit en repos d'une impunité souveraine. [Coustel 1694]*

Très répandu dans le corpus est l'usage de l'ironie comme écho satirique. Le lemme « calomnier », présent dans les textes du 17e au 19e siècle, en est l'exemple :

> *[…] cessez de calomnier vos contemporains selon l'usage immémorial de ceux qui profèrent de vaines paroles. [Senancour 1825]*

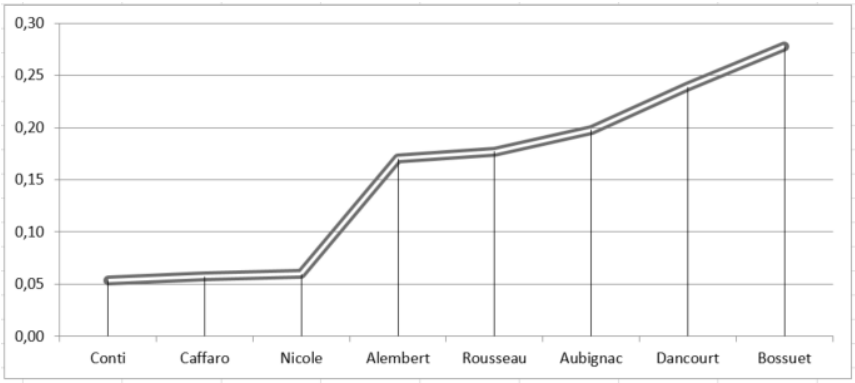


Figure 2 : Valeurs pondérées de l'annotation de l'ironie dans le corpus

Les premiers résultats nous ont ainsi permis d'effectuer des comparaisons très intéressantes entre les textes des auteurs défenseurs et les textes des auteurs pourfendeurs du théâtre. A partir du nombre d'expressions ironiques correctement identifiées comme telles, nous avons recensé leur nombre et dressé des statistiques pour chaque auteur du corpus annoté.
On constate qu'en données relatives, les auteurs qui utilisent le plus les marqueurs d'ironie appartiennent à la « querelle Rousseau » (moitié du 18e s.). Cela est à analyser en perspective mais, en l'espèce, dans cet article nous pouvons le mettre en lien avec l'usage de l'ironie au 18e siècle, comme plusieurs écrits sur Voltaire le témoignent (Loriot, 2015). Les mots de D'Alembert sont très parlants sur ce sujet et éclairent le rôle de l'ironie [Alembert, 1759] :

> *Si la satire et l'injure n'étaient pas aujourd'hui le ton favori de la critique, elle serait plus honorable à ceux qui l'exercent, et plus utile à ceux qui en sont l'objet.*

Les marqueurs linguistiques qui ont été détectés pour cette période appartiennent à la sphère sémantique du *ridicule*, de la *satire*, de la *farce* et du *comique*[3]. D'autres marqueurs verbaux, tels que *se moquer* et *plaisanter* sont présents dans cette querelle et sont communs aux écrits de la précédente controverse datant du milieu du 17e siècle. Les valeurs ironiques de cette dernière, dont les représentants théâtrophobes sont Conti et Nicole, parmi d'autres, sont cependant moins importantes (0,06 vs. 0,17). Outre ces marqueurs verbaux, nous pouvons citer les catégories de substantifs tels que le *ridicule* et le *faire rire*. A la même période, Aubignac, auteur de la stratégie offensive-défensive, part d'une critique du théâtre pour arriver à sa défense. Il s'inspire des marqueurs habituels pour la période du 17e siècle et reprend dans ses phrases les propos de ses collègues, pour ensuite les réfuter. De plus, il recourt plus spécifiquement à des marqueurs ironiques tel que *railler* et *idiot*. Contemporaine à d'Aubignac, la querelle entre Caffaro et Bossuet nous donne des résultats surprenants : si Caffaro emploie peu de marqueurs relevant de l'ironie (0,05), Bossuet est lui chef de file parmi ses contemporains (valeur de 0,27). Comme les autres auteurs, Bossuet puise dans les marqueurs du *comique* et du *ridicule*, tout comme la forme verbale *plaisanter*. Néanmoins, nous retrouvons dans ses résultats des mots appartenant à la catégorie de marqueurs piquants [Bossuet, 1694]:

[3] Signalons que le marqueur « ironie » et toutes ses variantes n'ont que 11 occurrences dans le corpus !

> *Il ne faut pas s'étonner que l'église ait improuvé en général tout ce genre de plaisirs [les spectacles…] à cause que communément, ainsi que nous l'avons remarqué, par sa bonté et par sa prudence, elle épargne la multitude dans les censures publiques : néanmoins parmi ces défenses, elle jette toujours des traits piquants contre ces sortes de spectacles, pour en détourner tous les fidèles.*

Nous comprenons ainsi que pour juger le théâtre incompatible avec la morale chrétienne, Bossuet privilégie un style vif et mordant, il appuie l'église tout en dénigrant les défenseurs du théâtre.

La recherche sur les stratégies de la querelle du théâtre, tout en se questionnant sur les modalités argumentatives et les objectifs circonstanciels de chaque auteur, nous dévoile également certaines idées récurrentes autour de la considération du théâtre. Les différents textes partagent un certain nombre de lieux communs, comme par exemple l'idée de perversion, l'inflation temporelle, ou les arguments économiques et politiques.

8. Discussion et perspectives

Dans cet article, nous avons présenté une approche à base de règles pour la détection automatique de l'ironie et des formes apparentées dans un corpus de débats sur le théâtre (16e – 19e siècle). La méthode que nous avons adoptée nous a fourni une matière abondante et des données quantitatives pour mieux cerner l'objet d'étude. Vu la particularité du phénomène langagier étudié et la simplicité de notre approche par analyse de surface, nous considérons que ces premiers résultats sont très encourageants (93.9 % de précision). A ce titre, ils méritent d'être approfondis afin d'en tirer le plus grand bénéfice en terme d'exploitation et de précision. Nous envisageons de calculer le taux de rappel dans l'annotation et d'identifier les sources des segments annotés (les locuteurs). L'un de nos objectifs consiste également à annoter les phrases négatives et à analyser leur association avec l'ironie (Mainardi et al., 2015), ce qui nous permettrait de dégager des pistes de recherche inédites dans le domaine des humanités numériques.

Références

Alrahabi, M. (2010). EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe. Thèse de doctorat, Université Paris-Sorbonne.

Joshi A., Bhattacharyya P., Carman M. J., (2016). Automatic Sarcasm Detection: A Survey ACM Comput. Surv. V, N, Article A (January 2016).

Lecercle F., Mainardi C., Thouret C. (2016). Pour une exploration numérique des polémiques sur le théâtre, RHLF, n°116 / 4 dir. Didier Alexandre, Littérature et humanités numériques, PUF.

Loriot C. (2015), Rire et sourire dans l'opéra-comique en France aux 18ème et 19ème siècles, Lyon, Symétrie.

Mainardi C., Sellami Z., Jolivet V., (2015). "A Semantic Exploration Method Based on an Ontology of 17th Century Texts on Theatre: la Haine du Théâtre", First International Workshop on Semantic Web for Cultural Heritage (SW4CH 2015), New Trends in Databases and Information Systems, 539, pp. 468-476, Communications in Computer and Information Science.

Martin L. (2009), "Le rire est une arme. L'humour et la satire dans la stratégie argumentative du Canard enchaîné", *A contrario* 2009/2 (n° 12), 26-45.

Riguet M., Alrahabi M. (2017), "Pour une analyse automatique du Jugement Critique: les citations modalisées dans le discours littéraire du XIXe siècle", in DHQ: Digital Humanities Quarterly 2017

Migrants et réfugiés : dynamique de la nomination de l'étranger

Mohammad Alsadhan, Sascha Diwersy,
Agata Jackiewicz, Giancarlo Luxardo
Praxiling UMR 5267 (Univ Paul Valéry Montpellier 3, CNRS)
muhammad.alsadhan@univ-montp3.fr, sascha.diwersy@univ-montp3.fr,
agata.jackiewicz@univ-montp3.fr, giancarlo.luxardo@univ-montp3.fr

Abstract
Intense debates arose from the migrant crisis experienced by Europe in recent years, both in the media and in the politics. We address here the issue of nomination used for the newcomers, that we propose to study based on the comparison of the two substantivations in French: *migrant* and *réfugié*. Using their combinatory profiles, we seek to highlight the contrast between the two terms and the changes in their semantics and their axiological charge. In order to do so, we rely on a large corpus of texts, established over a three-year period: the French parliamentary debates of the *Assemblée Nationale*. The comparative study of the combinatory profiles related to the two terms shows that both shared and unshared collocatives are encountered, and that their profiles overall tend to converge.

Résumé
Au cours des dernières années, la crise migratoire en Europe a suscité de vifs débats politico-médiatiques. Nous nous intéressons ici à la question de la nomination des nouveaux arrivants, que nous proposons d'étudier par la comparaison des deux substantivations migrant et réfugié. A partir de leurs profils combinatoires, nous cherchons à mettre en évidence le contraste entre ces deux termes, les changements dans leur sémantique et leur charge axiologique. Pour cela, nous nous appuyons sur un corpus, établi sur une période d'environ trois ans : les débats à l'Assemblée Nationale. L'étude comparative des profils combinatoires associés aux deux termes montre que l'on rencontre à la fois des collocatifs partagés et d'autres non partagés et que leurs profils tendent globalement à converger.
Keywords: political discourse, cooccurrences, diachronic data and hierarchical clustering, curve clustering.

1. Introduction

L'Union Européenne a connu en 2015 une arrivée massive d'étrangers extra-européens, qui a donné lieu à des formules telles que « crise migratoire » ou « crise des réfugiés ». Dans un contexte de net clivage de l'opinion publique, cette crise a entraîné des positions politiques contrastées dans chaque pays concerné et des compromis difficiles à trouver.

Les débats politico-médiatiques ont porté d'abord sur la prise en charge des victimes, le droit « d'asile » à accorder aux nouveaux venus, de même que sur la lutte contre les filières illégales, avec des positions « pro-immigration » ou « anti-immigration ». Mais ce phénomène s'expliquant en partie par les conflits en cours au Sud et à l'Est de l'Europe, la question de la désignation des intéressés a été posée. Alors que jusque-là les « migrants » étaient principalement motivés par des perspectives économiques, il a été remarqué qu'une partie de ces personnes devraient être nommés « réfugiés » ou « demandeurs d'asile ». D'autres termes, comme « clandestins », ont pu aussi être évoqués.

Nous cherchons ici à questionner la dynamique de la nomination utilisée dans les débats politiques. A partir d'un corpus de débats parlementaires nous mettons en œuvre divers procédés de classification basés sur la nature diachronique des données.

2. Les corpus de débats parlementaires

Nous faisons l'hypothèse que les discours autour de la crise migratoire font usage des deux termes *migrant* et *réfugié* en partie de façon interchangeable, en partie dans des contextes où seulement l'un des deux est possible. Cette distinction entre plusieurs emplois en discours, nous proposons de la mettre en évidence par le voisinage des deux termes et d'évaluer sa variation d'abord sur le discours politique et en fonction du temps.

Le corpus traité dans la suite est constitué à partir des transcriptions des débats en séance publique à l'Assemblée Nationale pour la période qui va de janvier 2014 à février 2017 (ce qui correspond à la fin de la XIV^{e} législature). Les données textuelles, publiées en format XML et disponibles en accès libre sur le site data.assemblee-nationale.fr, représentent environ 28,6 millions de mots occurrences. Elles ont été transformées et enrichies par des annotations linguistiques suivant une méthodologie décrite par Diwersy et al. (2018). De nombreuses métadonnées sont définies sur ce corpus, mais dans la suite nous nous concentrons sur la date (mois-année) associée à une unité structurelle de base correspondant au tour de parole (intervention d'un député).

3. Analyse chronologique

L'évolution du sémantisme des termes *migrant* et *réfugié* peut être étudiée par l'association de méthodes mettant en jeu : (i) les fréquences d'apparition de ces deux lemmes dans les corpus, (ii) leurs profils collocationnels, qui peuvent faire émerger des champs sémantiques spécifiques, (iii) la variation de la similarité de ces profils collocationnels dans le temps et la caractérisation de la contribution de chaque collocatif à l'évolution des scores de similarité obtenus.

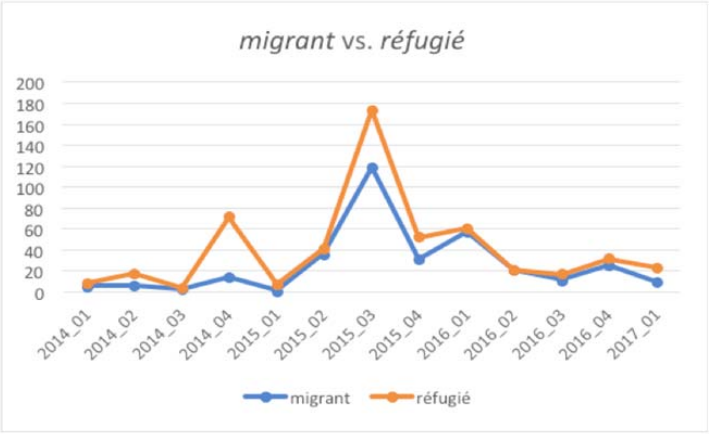


Figure 1

L'évolution des fréquences relatives des deux lemmes par trimestre dans le corpus est illustrée par le graphique en figure 1. Il met en évidence une évolution fréquentielle en parallèle avec un pic d'utilisation des deux termes autour de septembre 2015. La corrélation de rang entre les deux séries fréquentielles, mesurée par le taux de Kendall, est ici significative (environ 0,74, pour une p-valeur de 0,0005). Dans la suite, l'unité de temps choisie est le trimestre ; il en résulte des analyses sur 13 trimestres pour la période couverte. Afin de produire une périodisation plus précise, nous avons mis en œuvre une approche combinant annotations en relations de dépendance syntaxique, création de lexicogrammes représentant les profils collocationnels par trimestre des deux termes (ordonnés suivant le score d'application du test exact de Fisher) et application de Classifications Ascendantes Hiérarchiques par Contiguïtés (CAHC), cf. (Diwersy et Luxardo, 2016 ; Gries et Hilpert, 2008).

La construction d'une CAHC peut être entreprise suivant deux méthodes :

- pour chaque lemme, en calculant la similarité entre deux trimestres

successifs d'après le coefficient de Pearson (Pearson product-moment correlation coefficient),

- en calculant la variation de la similarité entre les vecteurs représentant les profils collocationnels des deux lemmes, d'après l'écart type cumulé sur deux trimestres successifs.

La première méthode révèle les variations plus importantes sur les trimestres initiaux jusqu'au pic de la crise. La deuxième méthode qui permet d'illustrer la comparaison entre les deux termes par un graphique unique est représentée par la figure 2.

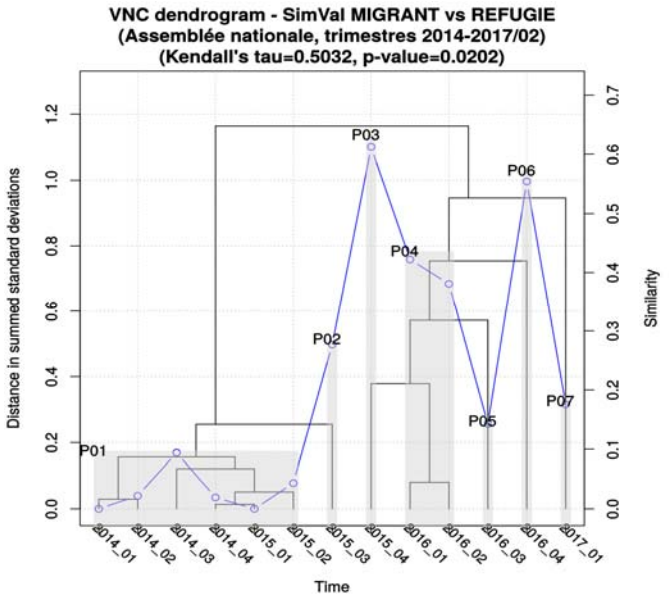


Figure 2

L'étude de cette classification hiérarchique permet de révéler sept étapes (représentées par sept zones grises). L'évolution du score de similarité est illustrée par un graphique qui se superpose au dendrogramme et qui confirme une croissance globale de 0 à 0,2 (mais avec un pic à 0,6). Le passage d'une période à l'autre est marqué par une progression jusqu'à P03 (correspondant au 3e trimestre 2015, suivant le pic de la crise) mais avec un déclin des périodes P03 à P05 et de P06 à P07.

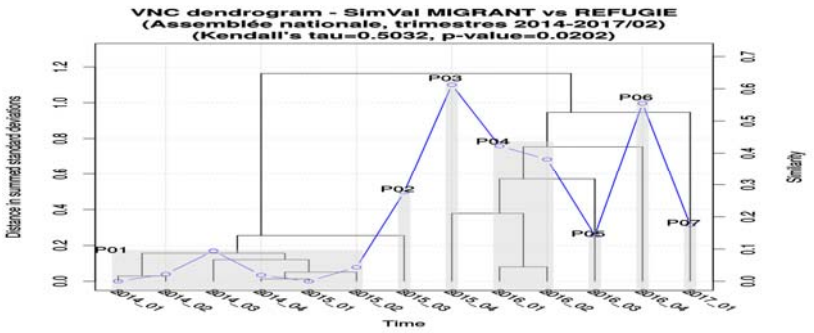


Figure 3

4. Évolution des profils combinatoires et orientations discursives

Cette section vise à expliciter les facteurs linguistiques à l'origine des tendances statistiques établies dans la partie précédente. Il s'agit, d'une part, de mettre en évidence les différences sémantiques entre *migrant* et *réfugié* telles qu'elles se manifestent à travers leurs profils différentiels et, d'autre part, de relever les points essentiels concernant leur similarité distributionnelle. Les profils différentiels sont constitués par les collocatifs exclusifs à chacun des substantifs étudiés et, de ce fait, ne contribuent à aucun moment à la similarité de leurs profils combinatoires. Le tableau 1 en donne un aperçu restreint aux collocatifs les plus saillants, situés dans le premier décile des inventaires collocationnels en termes de score d'association.

Tableau 1 - Profils différentiels constitués par le premier décile des collocatifs exclusifs à migrant et à réfugié

migrant					***réfugié***				
Dépendances en aval (régime)		Dépendances en amont (termes recteurs)		Coordination	Dépendances en aval		Dépendances en amont		Coordination
Epithètes	Compl. du nom	Compl. objet Compl. circ. Sujet	Compl. du nom		Epithètes	CDN	CO CC Sujet	CDN	
irrégulier *illégal* *clandestin* *âgé*	*Calais* *Calaisis* *situation*	*dissuader* *entasser* *refouler* *secourir*	*retour* *langue* *déferlement* *réadmission*	-	*politique* *palestinien* *afghan* *vietnamien* *irakien* *cambodgien* *persécuté* *réinstallé*	*guerre*	*affluer*	*statut* *protection* *(Haut-)Commissariat* *qualité* *relocalisation* *distinction* *concubin* *défi*	*apatride* *bénéficiaire* *déplacé* *migrant*

Parmi les collocatifs saillants du nom *réfugié*, on notera d'abord la forte présence d'une série de termes (*statut*, *qualité* ; (*Haut-*)*Commissariat* ; *protection* ; *apatride*)[1] qui renvoient au cadre des dispositions relevant du droit international qui imposent aux autorités un devoir d'assistance envers des personnes dont le départ de leur lieu de résidence habituelle est considéré comme étant contraint par une menace existentielle. Catégoriser une personne au moyen du terme *réfugié* revêt donc un enjeu juridique, administratif et politique, dont l'ampleur peut se voir régulée d'une part, par des mises en paradigme explicites avec d'autres termes dans le cadre d'une coordination (cf. les collocatifs *apatride*, *bénéficiaire*, *déplacé* et *migrant*) et, d'autre part, par des catégorisations secondaires exprimées par des expansions nominales (épithètes ou compléments du nom) caractérisant les causes du départ forcé. A travers les modifieurs du nom *réfugié* impliquant une relation causale (*politique*, *persécuté* ; (*de*) *guerre*) se construit un paradigme, et finalement une hiérarchie de causes potentiellement légitimes ou non-légitimes (et de réponses à apporter aux conséquences liées à ces causes).[2] A côté de ces modifieurs, qui dénotent directement la cause du départ forcé, on trouve toute une série d'adjectifs ethnonymiques (*palestinien*, *afghan*, *vietnamien*, *irakien*, *cambodgien*) qui la dénotent indirectement en s'appuyant sur le savoir partagé concernant l'histoire troublée de ces pays. Cet environnement discursif montre que le mot *réfugié* se présente comme la nomination d'un statut juridique et qu'il est intégré à une argumentation orientée positivement.

Les collocatifs de *migrant* révèlent un profil sémantique bien différent, en ce sens que ce terme place au centre de l'intérêt la question de la (non-)conformité à des dispositions légales imposées à des personnes dont le séjour sur un territoire différent de celui de leur lieu résidentiel d'origine est considéré comme étant le résultat d'un déplacement conditionné par des considérations utilitaires (et en premier lieu économiques). C'est bien à cette dimension sémantique que se rapporte, dans le profil différentiel de *migrant*, de façon saillante la série des collocatifs *irrégulier*, *illégal*, *clandestin* et *situation* (qui, quant à lui, s'oppose, de ce point de vue, à *statut* et *qualité*, collocatifs exclusifs à *réfugié*). Ayant hérité les traits aspectuels du participe en *–ant* dont

[1] On trouve dans les déciles inférieurs – non documentés ici – d'autres collocatifs comme *statutaire* ou *conventionnel* qui rentrent dans cette même série.

[2] On peut observer que cette sous-catégorisation va souvent de pair avec une modalisation d'appartenance catégorielle, exprimée par l'adjectif épithète *véritable* qui constitue avec *vrai* et *authentique* une série de collocatifs (appartenant à la catégorie de l'enclosure) exclusifs à *réfugié* qui sont néanmoins représentés à des rangs inférieurs de l'inventaire cooccurrentiel.

il est issu par conversion, le nom *migrant* présente le séjour momentané de la personne qualifiée en tant que telle à un endroit donné comme étant l'épisode d'une série inaccomplie de déplacements[3] – séjour et déplacements qui, à travers des collocatifs tels *dissuader*, *refouler* et *retour*, se voient caractérisés comme relevant aussi bien de la volonté des personnes en mouvement, que de la bienveillance ou du refus des autorités qui en ont le contrôle potentiel. Faut-il voir en cela la motivation inférentielle de l'évaluation négative que véhicule un terme comme *déferlement* contrairement à ses variantes axiologiquement plus neutres *afflux*, *flux* ou encore *arrivée*, qui, eux, font tous partie des collocatifs partagés des noms *migrant* et *réfugié* ?

Pour mieux cerner les collocatifs partagés qui contribuent le plus à l'évolution de la similarité distributionnelle des deux noms en question, nous avons mis en œuvre la méthode de classification proposée par Trevisani & Tuzzi (2016), en l'appliquant aux séries chronologiques des produits de scores d'association normés propres à chaque collocatif, qui entrent dans la composition des sommes donnant les produits scalaires lesquels représentent les indices de similarité retenus.

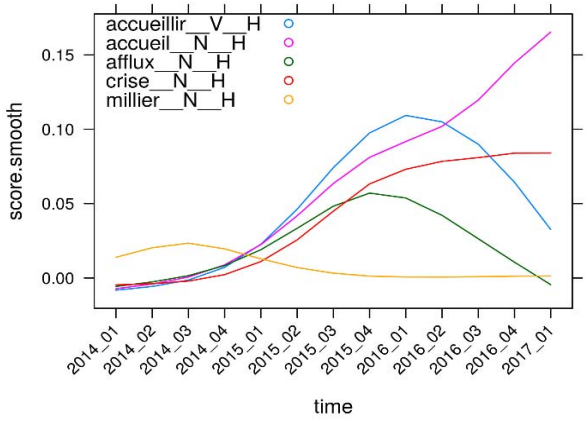


Figure 4

L'application de la méthode[4] fait ressortir, sur l'ensemble des 72 collocatifs communs à *migrant* et *réfugié*, 6 classes de profils évolutifs, dont 5 sont

[3] Contrairement à cela, *réfugié*, qui résulte de la nominalisation d'un participe passé, est associé à la représentation d'un seul épisode de déplacement accompli et envisagé en termes de son origine.

[4] Nous remercions Arjuna Tuzzi d'avoir mis à notre disposition le script R permettant de mettre en œuvre les calculs respectifs.

constituées par un seul terme, à savoir *millier*, *afflux*, *accueillir*, *crise* et *accueil* (cf. figure 3). D'un point de vue sémantique, ces 5 collocatifs, qui, à différents moments de la série chronologique analysée, occupent les premiers rangs en termes de contribution aux scores de similarités respectifs, forment tout un condensé de la trame discursive impliquant les noms *migrant* et *réfugié* au cours de la période étudiée, avec :

- *millier* et *afflux*, qui renvoient à une affluence perçue comme massive ;
- *crise*, qui caractérise ce processus comme ayant atteint un point culminant à fort potentiel de déstabilisation ;
- ainsi que *accueillir* et *accueil* qui se rapportent à la prise en charge des conséquences immédiates du processus concerné.

Facteurs distributionnels de premier ordre, ces collocatifs placent *migrant* et *réfugié* dans un rapport paradigmatique associé à plusieurs dimensions sémantiques, qui, en vue des orientations argumentatives fortement divergentes instaurées par les deux noms (cf. *supra*), fait de leur choix un véritable enjeu discursif.

5. Conclusion et perspectives

Les prolongements de cette étude exploratoire sont nombreux. En partant de Wihtol De Wenden (2016), il nous semble possible de construire un modèle d'analyse comportant cinq catégories qui sont autant de facettes du phénomène migratoire actuel : (i) origines et causes des migrations, (ii) profils des migrants, (iii) situation des migrants, (iv) gouvernance des migrations, (v) mobilité et restrictions migratoires. L'application de cette grille de lecture aux collocations impliquant les termes *réfugié* et *migrant* (ou encore leurs équivalents), peut s'avérer une piste de recherche prometteuse qui permet de donner aux résultats de l'analyse linguistique que nous venons d'effectuer une dimension transdisciplinaire, comme c'est par exemple le cas pour la différence entre facteurs « push » (poussant les individus à partir de leur pays) et « pull » (incitant les individus à venir dans un pays spécifique) établie par Wihtol de Wenden, différence qui se reflète dans la divergence fondamentale de l'orientation argumentative des programmes de sens propres aux noms étudiés, en ce que *réfugié* implique la notion de départ forcé alors que *migrant* évoque l'idée d'un déplacement volontaire.

Si la figure du réfugié ou du migrant est essentiellement une construction politique (Wihtol De Wenden, 2016, p. 50) – ce que confirme d'ailleurs le profil collocationnel du terme correspondant tel qu'il se manifeste dans le corpus de discours parlementaire analysé - les différents (et nombreux) profils des personnes en déplacement peuvent être étudiés à partir des témoignages qu'elles livrent à propos de leur expérience migratoire. C'est l'objet d'une enquête menée auprès de Syriens arrivés en France depuis 2012,

qui se situe dans le prolongement du présent article et qui comporte à ce stade un volet uniquement qualitatif, dont les résultats préliminaires (Alsadhan et Richard, 2018) montrent que, lorsque le choix se présente, c'est bien le vocable *réfugié* qui est privilégié en tant qu'auto-désignant.

Références

Alsadhan, M., Richard A. (2018, à paraître). La réception des réfugiés Syriens du discours médiatico-politique identitaire français, in Sandré M., Richard A. & Hailon F. : *Le discours politique identitaire face aux migrations*, No 8 de la revue Studii de lingvistica.

Diwersy, S., Luxardo, G. (2016). Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l'exemple des débats du Parlement européen, in Mayaffre D., Poudat C., Vanni L., Magri V. & Follette P. (éds.) : *JADT 2016 : Actes des 13es Journées internationales d'Analyse statistique des Données Textuelles*, Nice, 2016, URL : http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83638/83638.pdf.

Diwersy, S., Frontini, F., Luxardo, G. (2018, à paraître). The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse, in *Proceedings of ParlaCLARIN workshop, 11th edition of the Language Resources and Evaluation Conference (LREC2018)*.

Gries, S.T., Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora 3 (1)*, pp. 59-81.

Trevisani, M., Tuzzi, A. (2016). Analisi di dati testuali cronologici in corpora diacronici: effetti della normalizzazione sul curve clustering, in Mayaffre D., Poudat C., Vanni L., Magri V. & Follette P. (éds.) : *JADT 2016 : Actes des 13es Journées internationales d'Analyse statistique des Données Textuelles*, Nice, 2016, URL : http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/82630/82630.pdf.

Wihtol De Wenden C. (2016). *Migrations. Une nouvelle donne*, Éditions de la Maison des sciences de l'homme, Paris.

Xplortext, a R package. Multidimensional statistics for textual data science

R. Alvarez-Esteban[1], M. Bécue-Bertaut[2], B. Kostov[3], F. Husson[4], J-A Sánchez-Espigares[2]
[1]Universidad de León – ramon.alvarez@unileon.es
[2]Universitat Politècnica de Catalunya – monica.becue@upc.edu; josep.a.sanchez@upc.edu
[3]Institut d'Investigacions Biomèdiques August Pi i Sunyer – belchin3541@gmail.com
[4]Agrocampus Ouest – husson@agrocampus-ouest.fr

Abstract
We present here the package Xplortext for textual data science which provides classical and novel features for textual analysis. Starting from the corpus encoded into a lexical table, aggregate or not, several problems are dealt with: revealing both document and word structures and their mutual relationships, by applying correspondence analysis (CA); comparing several corpora structures by using multiple factor analysis for contingency tables (MFACT); uncovering complex relationships between words and contextual variables via CA for a simple or a multiple generalized aggregate lexical table (CA-GALT and MFA-GALT), clustering documents thanks to a hierarchical clustering algorithm (HCA); evaluating the evolution of the vocabulary along time thanks to a chronological constrained hierarchical clustering algorithm (CCHCA).

Resumé
Nous présentons ici le paquet Xplortext pour la science des données textuelles qui comprend des méthodes classiques et récentes d'analyse textuelle. Partant du corpus encodé sous forme tableau lexical, agrégé ou non, plusieurs problèmes sont traités: révélation des structures sur les documents et les mots ainsi comme leurs relations mutuelles, en appliquant l'analyse des correspondances (AC); comparer plusieurs structures de corpus en utilisant l'analyse factorielle multiple pour les tables de contingence (MFACT); découvrir des relations complexes entre mots et variables contextuelles via CA pour une table lexicale agrégée simple ou multiple (CA-GALT et MFA-GALT), en regroupant des documents grâce à un algorithme de clustering hiérarchique (HCA); évaluer l'évolution du vocabulaire au fil du temps grâce à un algorithme de classification hiérarchique sous contrainte chronologique (CCHCA).
Keywords: Xplortext, R package, Textual data, Contextual data, Correspondence analysis, Multiple factor analysis for contingency tables,

Generalized aggregate lexical table, Hierarchical clustering, Contiguity constrained hierarchical clustering, Labeled tree.

1. Introduction

R offers numerous tools for textual data science. However, among them, multidimensional statistics is not so well represented that it should be. Xplortext**,** a new R package, intends to fill in the gaps. Its features are based on the exploratory approach to texts, in the line of the works by Benzécri (1981) and Lebart et al. (1998). The fundamental choices behind the design of Xplortext are to offer classical and novel textual analysis methods based on multidimensional statistics in a same package. The mains issues were to consider:

- Classical multidimensional statistical methods, in which CA remains being the core method.
- Novel methods, favoring those able to jointly analyze textual and contextual data to know not only *who says what,* taking here the title of a paper by Lebart, but also *why he/she is saying that*.
- Numerous graphical outputs providing great flexibility to choose the elements to be represented.
- Specific methods to deal with chronological corpora.

2. Example

The political speech corpus used as an example consists of 11 documents of about 10,000 occurrences each one. These are the "investiture speeches" of 6 Spanish presidential candidates who have been pronounced from 1979 to 2011: Suarez (1979), Calvo-Sotelo (1981), González (1982, 1986, 1989 and 1993), Aznar (1996 and 2000), Zapatero (2004 and 2008) and Rajoy (2011).

3. Encoding the textual data and basic statistics

Xplortext takes advantage of functions of the R package tm to import the corpus. Mainly, plain text files (typically .txt) and spreadsheet-like file (.csv, .xls) are considered. By default, plain text and CSV files are assumed to use the local native system (usually latin1) on Windows and utf8 in Mac or Linux. The encoding of the file can be given in the R command read. If necessary, the corpus can be saved in a known encoding beforehand. In any format, one row corresponds to one document. The text to analyze can be filled in one or several columns; the remaining columns provide information about the documents and are automatically imported as contextual (quantitative and/or qualitative) variables. Textual and contextual data must be located in the same file. Conversion to lower/upper cases, numbers removal and punctuation are managed by Xplortext depending on the

arguments of Textdata function. Stopwords can be taken into account using the lists provided by either Xplortext (issued from tm) or the user. The importing step ends with the encoding of the corpus into a documents × words table (lexical table) and, possibly, a documents × repeated segments table (segmental table). Another option is to ask for an aggregated lexical table according to the categories of a variable. Then, elementary indicators, such as the corpus and vocabulary sizes, are computed and the words and repeated segments indices are listed and represented by a histogram, visualizing so their frequency (Fig.1). Classical summaries of the contextual variables are given.

4. Correspondence analysis as a core method

Correspondence analysis (CA) is a core method in Xplortext revealing both document and word structures and their mutual relationships.

4.1. CA and content and form of a corpus

The content and form of a corpus are both important as CA results. In fact, content is better captured when replaced into the form as, "the form is the bottom that comes back to the surface" in the words of Victor Hugo. Figure 2 shows the factor maps issued from a CA performed on the documents × words table.

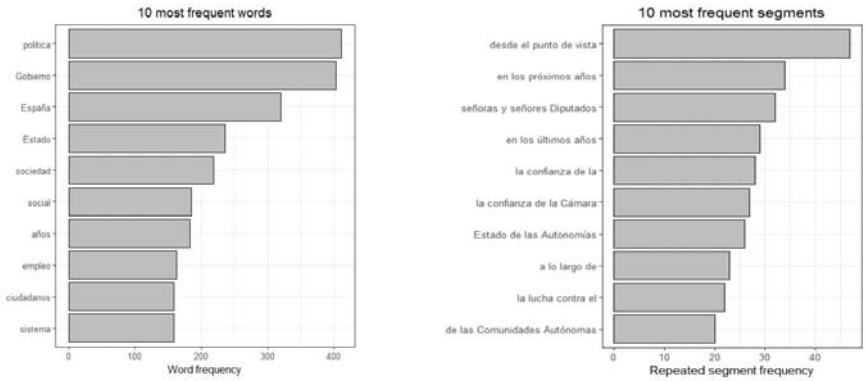


Figure 1: Most frequent words and repeated segments

The trajectory of the speeches is revealed, enhancing the existence of three temporal poles. The represented words are the most contributive and have to be read as seen along the trajectory. In this way, they clearly illustrate the three poles and allow us to capture the meaning of the evolution. Note that the confidence ellipses around the documents are very narrow.

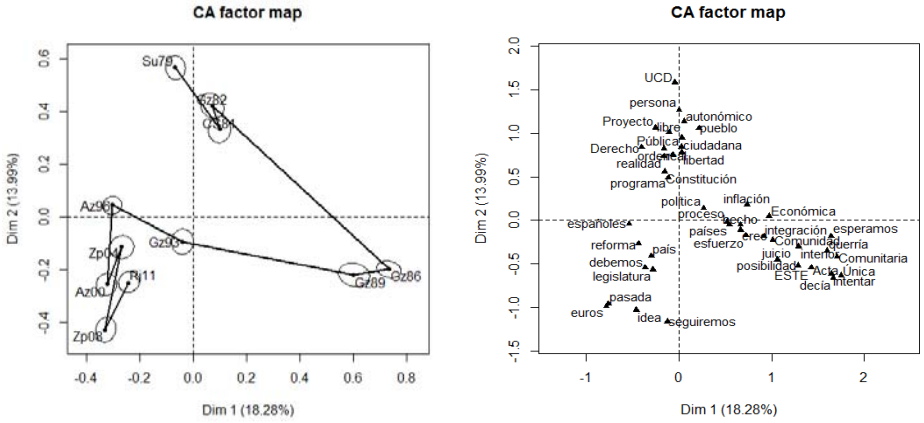


Figure 2: Documents and the most contributive words on the first CA plane

4.2. Multiple factor analysis for contingency tables

When dealing with a multiple contingency table (=juxtaposition of several contingency tables), the multiple factor analysis for contingency tables (MFACT; Bécue-Bertaut & Pagès, 2004; Bécue-Bertaut & Pagès, 2008), extension of CA, turns to be useful. Very different aims can be looked for. For example, interesting aims would be comparing the documents structures as issued either from using different thresholds on the word frequency (10, 20, 30 or 50; 4 lexical tables) or from keeping or not the tool words (2 lexical tables) or the stopwords.

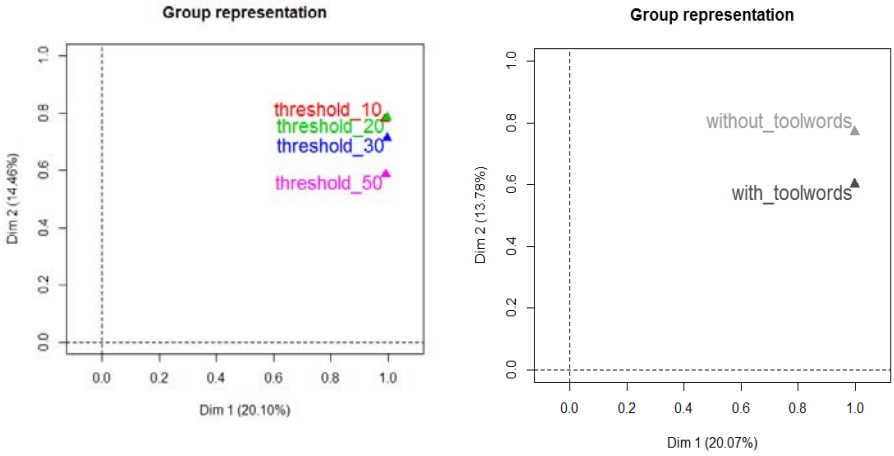


Figure 3: Synthetic representation of the groups as issued from MFATC

MFACT offers a high number of graphical and numerical results, either similar to those of any principal component methods (such as PCA or CA) or specific to the comparison of structures defined on the rows by the groups of columns. Among the latter, the representation of the groups provides a synthetic tool by representing each group with one point, revealing the global dissimilarities between the group structures (Fig. 3).

4.3. *Generalized aggregate lexical tables*

Correspondence analysis on a generalized aggregated lexical table (CA-GALT; Bécue-Bertaut & Pagès, 2015; Bécue-Bertaut, Kostov & Pagès, 2014) deals with two paired tables (frequency table, contextual variables table) observed on the same statistical units. In textual analysis, the frequency table is a lexical table and the statistical units are the documents. This method can be seen as a canonical correspondence analysis (CCA; ter Braak, 1986) approach to the texts. It enables to study the relationships between contextual variables and words but untangling the respective influences of the variables/categories on the lexical choices to avoid spurious relationships.

MFA-GALT (multiple factor analysis for analyzing a series of generalized aggregated lexical tables; Kostov, 2015) deals with several paired tables, possibly defined on several sets of statistical units while the set of variables is common to all the contextual tables. In textual analysis, MFA-GALT compares the relationships between words and variables in these several paired tables. A favored application concerns surveys answered in different languages by several samples, being common the open-ended and the closed questions.

5. Clustering algorithms

A classical hierarchical clustering algorithm (HCA) is included in Xplortext. Clustering starts from the documents coordinates on the CA dimensions. An exhaustive description of the clusters is provided, extracting their characteristic words and looking for the differentiated behavior of the variables in the clusters. The number of clusters is issued from the hierarchical tree structure. An automatic suggestion is done.

A method for chronological constrained hierarchical clustering algorithm (CCHCA) is also offered. Only chronological contiguous nodes can be grouped. Further, the tree is described by the chronological words defined as follows. The characteristic words of each node are identified but finally a word is associated to only one node, the one that it better characterizes. These words are used to label the nodes (Fig. 4). Although the tree could be used to determine clusters, its main role is to allow for capturing the evolution of the speeches and their vocabulary through a descending reading of the labels

and nodes of the tree.

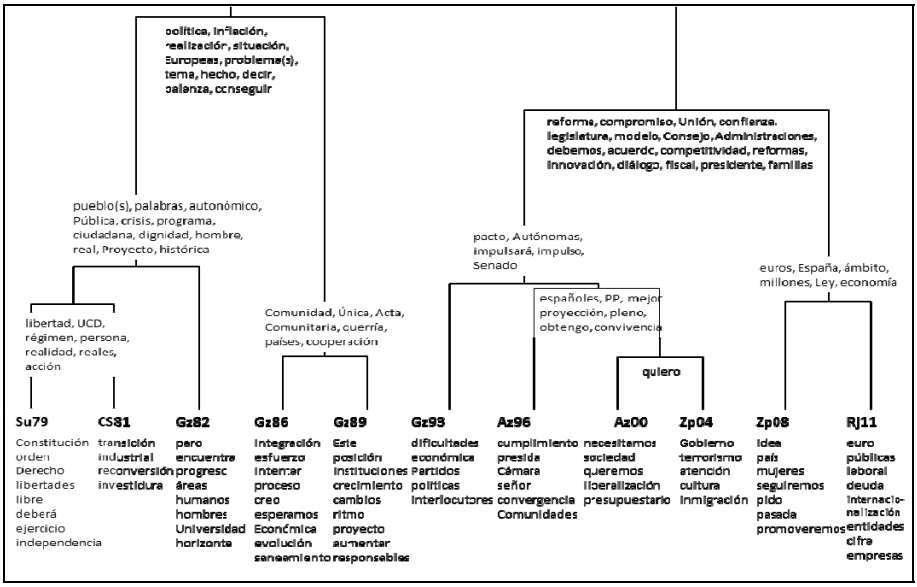


Figure 4: Labeled chronological tree

6. Works in progress

The following features will be included in a next future:

- Chronological clustering (Legendre *et al.,* 1985) has been proposed to divide a chronological series of species (=species counts operated at different moments) into homogeneous temporal parts. A same aggregation criterion as in chronological constrained clustering is used but a test is performed before aggregating two nodes to ensure their homogeneity. If homogeneity does not exist, the corresponding aggregation is not performed. As a result, the series is possibly divided into non-connected sub-series. This clustering method has been applied with benefit to the chronological series of words corresponding to a chronological corpus, allowing for dividing the corpus into non-connected homogeneous parts (Bécue-Bertaut *et al*., 2014).
- Regularized CA (Josse *et al.*) allows for recovering a low-rank structure from noisy data, such as textual data, by using regularization schemes via a simple parametric bootstrap algorithm.

7. Conclusion

Xplortext is published on R CRAN. Bécue-Bertaut, *et al*. (2018) present a

series of applications of this package through several examples whose results are interpreted in details. The corresponding bases and scripts are published on the website http://xplortext.org.

References

Bécue-Bertaut M. and coll. (2018). *Analyse textuelle avec R*. Presses Universitaires de Rennes (PUR), Rennes.

Bécue-Bertaut M., Kostov B., Morin A. and Naro G. (2014). Rhetorical strategy in forensic closing speeches. Multidimensional statistics-based methodology. *Journal of Classification, 31: 85-106.*

Bécue-Bertaut, M. and Pagès, J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, 45: 481-503.

Bécue-Bertaut M. and Pagès J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis*, 52: 3255–3268.

Bécue-Bertaut M. and Pagès J. (2015). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. *Advances in Data Analysis and Classification*, 9: 125–142.

Bécue-Bertaut M., Pagès J. and Kostov B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. *SORT – Statistics and Operations Research Transactions*, 38: 285–302.

Benzécri, J.-P. (1981). *Pratique de l'Analyse des Données. Tome III. Linguistique & Lexicologie.* Dunod, Paris.

Josse J., Sardy S. and Wager S. (2016). *denoiseR: A Package for Low Rank Matrix Estimation*. *arXiv: 1602.01206*

Kostov B. (2015). A principal component method to analyse disconnected frequency tables by means of contextual information. (Doctoral dissertation). Retrieved from http://upcommons.upc.edu/handle/2117/95759.

Lebart, L., Salem, A. and Berry, L. (1998) *Exploring textual data*. Kluwer.

Legendre, P., Dallot, S. and Legendre, L. (1985). Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton, *American Naturalist*, 125: 257–288.

ter Braak CJF. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67: 1167–1179.

L'evoluzione delle norme: analisi testuale delle politiche sull'immigrazione in Italia

Elena, Ambrosetti[1], Eleonora Mussino[2], Valentina Talucci[3]
[1] Associate Professor, Sapienza Università di Roma
[2] Associate Professor, Stockholm University
[3] Researcher, ISTAT

1. Introduzione

Nei paesi del Sud-Europa, le politiche migratorie tendono a privilegiare le questioni relative all'ingresso degli immigrati (ad esempio ingressi regolari e irregolari, sanatorie e ricongiungimento familiare) rispetto agli aspetti legati all'integrazione (Pastore 2004, Solé 2004). Questo squilibrio nell'azione politica è imputabile alla volontà dei paesi di immigrazione di poter controllare i flussi, bloccare gli ingressi non autorizzati e determinare il numero e la composizione dei migranti. Le politiche migratorie regolano in modo diretto l'esito dell'ingresso o meno nel Paese di destinazione e successivamente orientano i percorsi di inserimento nel tessuto economico-sociale e culturale degli stranieri ammessi in Italia. Attraverso lo studio delle politiche dell'immigrazione dall'Unità d'Italia a oggi possiamo analizzare come il linguaggio istituzionale nel corso degli anni e varie legislature si sia trasformato tracciando diversi aspetti legati alle migrazioni internazionali nel nostro paese. Questo argomento assume particolare importanza in quanto la scelta di un tipo di linguaggio potrebbe influenzare opinioni e atteggiamenti nei confronti degli stranieri da parte della popolazione italiana.

2. Le politiche migratorie in Italia

L'Italia, sebbene sia diventata un paese di immigrazione negli anni Settanta, soltanto nel 1986 si è dotata della prima normativa sull'immigrazione a seguito dell'adesione nel 1975 alla Convenzione alla Convenzione 143 dell'Organizzazione Internazionale del Lavoro (OIL) e dell'aumento dei flussi di immigrati nel corso degli anni Ottanta. La legge 943/1986 (Legge Foschi) riguardava in primo luogo lo status dei lavoratori, inoltre includeva il ricongiungimento familiare e l'accesso allo stato sociale di base (Colombo e Sciortino, 2004). La legge venne indirizzata ai lavoratori extra-comunitari, con l'obiettivo di equipararli ai lavoratori italiani e ai lavoratori dell'Unione europea (Nascimbene, 1988; Colombo e Sciortino, 2004). Inoltre la legge introdusse una sanatoria per i lavoratori extracomunitari che si trovavano già nel territorio senza documenti regolari. Nel febbraio 1990, la legge 39/1990 (Legge Martelli) fu approvata dal Parlamento italiano a seguito delle

pressioni dovute all'incremento degli arrivi dopo la caduta della cortina di ferro e dalla imminente ratifica del Trattato di Schengen (ratificato nel 1993 e entrato in vigore nel 1997). Al contrario della precedente legge Foschi, la legge si rivolgeva a tutte le categorie di migranti e non solo quindi ai lavoratori, per cui è considerata la prima legge organica sulle migrazioni. Nonostante ciò essa viene ricordata principalmente per la sanatoria di circa 218.000 irregolari. Ricordiamo anche alcuni altri aspetti significativi coperti dalla legge Martelli: l'introduzione dell'obbligo di visto, con conseguente inasprimento del controllo delle frontiere, che rese molto più difficile entrare in Italia, la programmazione annuale delle quote di lavoratori extracomunitari attraverso il cosiddetto Decreto Flussi, l'asilo politico, e da ultimo, l'inasprimento delle condizioni per l'ottenimento ed il rinnovo del permesso di soggiorno. Nel 1995 fu emanata la legge 489/1995 (Legge Dini): essa conteneva ulteriori misure restrittive per il controllo delle frontiere, una nuova sanatoria per i lavoratori stranieri irregolari e la regolamentazione dei flussi di lavoratori stagionali. A differenza delle misure restrittive, che non trovarono attuazione in quanto ritenute contrarie alla Costituzione, la sanatoria rappresentò il vero successo del decreto Dini, con un numero di stranieri regolarizzati pari a 248.000 persone.

Nel 1997, con l'entrata in vigore dell'accordo di Schengen è stato introdotto nell'ordinamento italiano l'adeguamento alla politica comune in materia di visti. Sempre in tema di normative comunitarie, la legge 209/1998 ha ratificato il trattato di Amsterdam, entrato in vigore in Italia in quell'anno. Nello stesso anno il governo ha approvato il Testo Unico delle disposizioni concernenti la disciplina dell'immigrazione e norme sulla condizione dello straniero, Dlgs 286/1998 (Legge Turco-Napolitano). Obiettivo della legge era quello di operare una rottura con il passato e di condurre ad una gestione del fenomeno migratorio strutturale e di lungo periodo. La legge era basata su quattro pilastri (Zincone e Caponio, 2004): 1. Prevenzione e lotta all' immigrazione irregolare: da notare in particolare l'introduzione dell'espulsione immediata di migranti irregolari e di centri di permanenza temporanea per detenere immigrati clandestini in attesa di espulsione; 2. Migrazioni da lavoro: i nuovi arrivi di lavoratori stranieri sono regolati con quote annuali di lavoratori stabilite ogni anno dal Ministero del lavoro; viene introdotto il meccanismo dello sponsor secondo il quale un cittadino italiano o uno straniero residente si fa garante dell'ingresso di uno straniero privo di contratto di lavoro; 3. Promozione dell'integrazione di migranti già residenti in Italia: creazione del Fondo Nazionale per l'integrazione dedicato al finanziamento di attività multiculturali e ad azioni antidiscriminazione; introduzione del permesso di soggiorno di lungo periodo, o carta di soggiorno per i migranti residenti da almeno 5 anni in Italia; 4. Concessione

di diritti umani fondamentali, come l'assistenza sanitaria di base, ai migranti irregolari. La legge Turco-Napolitano si fece carico della regolarizzazione di 217.000 stranieri.
Nel 2002 è stata introdotta la legge Bossi-Fini, con lo scopo di modificare in maniera restrittiva il Testo unico del 1998. Più specificamente, la legge ha modificato i primi due pilastri della legge. Con la nuova normativa sono state adottate una serie di misure volte a scoraggiare l'insediamento permanente dei migranti tra le quali: l'abolizione del sistema dello sponsor, la riduzione del periodo di validità del permesso di soggiorno e il collegamento della validità del permesso di soggiorno a un contratto di lavoro ("contratto di soggiorno"). Inoltre fu adottata una politica più repressiva nei confronti dei migranti irregolari che includeva l'applicazione del rimpatrio forzato, controlli più sistematici della polizia che includevano il pattugliamento delle coste italiane e la detenzione di coloro che rimanevano sul territorio italiano più a lungo di quanto previsto dal permesso di soggiorno (over-stayers). In linea con le precedenti leggi, la legge 189/2002 ha regolarizzato 634.728 immigrati, rappresentando la più grande sanatoria mai adottata in Europa fino a quel momento (Zincone, 2006). Dopo il 2002 sono state apportate poche modifiche alla normativa sulla migrazione, si tratta in particolare di: misure per combattere l'immigrazione clandestina, sanatorie per migranti irregolari presenti sul territorio italiano e recepimento di direttive UE che implicano modifiche alla normativa esistente.
L'acquisizione della cittadinanza per nascita (jus sanguinis) e per residenza (jus soli) era inizialmente regolata dalla legge 555/1912. Le condizioni erano molto restrittive: la cittadinanza era concessa solo al figlio di un uomo italiano e sotto condizioni specifiche al figlio di una donna italiana. La legge 123/1983 ha introdotto nella legislazione italiana l'acquisizione della cittadinanza per matrimonio e ha riformato l'acquisizione della cittadinanza per nascita, concedendo indifferentemente il diritto di cittadinanza al figlio di madre o padre italiani. L'acquisizione della cittadinanza italiana è stata ulteriormente riformata dalla legge 91/1992 riservando particolari diritti ai cittadini europei rispetto agli extra europei. La cittadinanza per matrimonio è stata riformata nel 2009 (legge 94 del 15 giugno), prolungando il periodo di residenza necessario in Italia da sei mesi a due anni dalla data del matrimonio. Negli ultimi anni si contano diversi tentativi per introdurre una nuova normativa sulla cittadinanza allo scopo di semplificare e ridurre il tempo per l'ottenimento della cittadinanza per i migranti di seconda generazione (nati in Italia). Come primo risultato, l'art. 33 del decreto 69/2013 ha semplificato la procedura di acquisizione della cittadinanza per gli stranieri nati in Italia. Nonostante ciò, fino ad oggi manca una nuova normativa in materia.

La normativa sulle migrazioni in Italia è stata costantemente caratterizzata dalla mancanza di una politica attiva degli ingressi e dal continuo tentativo di rallentare ed osteggiare il radicamento giuridico e sociale della popolazione straniera sul territorio italiano. Il ricorso continuo a strumenti ex-post come le sanatorie, l'utilizzo delle quote come sistema di emersione di lavoratori stranieri già presenti sul territorio italiano piuttosto che come norma di ingresso di nuovi lavoratori, ed il forte accento che la classe politica e i media pongono sulla lotta all'immigrazione illegale sono esempi emblematici di come il fenomeno migratorio in Italia venga affrontato in termini di contenimento e controllo e non di allargamento e integrazione. La presenza straniera, ancora oggi, è perlopiù considerata transitoria e viene percepita e gestita in termini di risposta ad eventi contestuali di emergenza.

3. Dati e Metodi

I dati testuali utilizzati per realizzare questo lavoro sono tutti i capi normativi contenuti nelle leggi approvate in Italia dal 1912 al 2014 in materia di migrazione. La metodologia di analisi proposta fa capo alla *Content Analysis* realizzata attraverso tecniche automatizzate dei dati. Si effettua applicando un insieme di *routine*, supportate da specifici software in questo caso TaLTAC2 – Trattamento automatico Lessico testuale per l'Analisi del contenuto - che consentono di automatizzarne in parte o del tutto l'esplorazione, la descrizione e il trattamento di grosse moli di dati; in questo modo vengono trasformati insiemi di testi non strutturati in insiemi di testi strutturati. Oltre alla descrizione dei contenuti del testo è possibile analizzare il corpus in base ad una o più variabili disponibili sui frammenti come l'anno e la maggioranza di governo[1]. L'estrazione dell'informazione peculiare individuata attraverso il test del p-value permetterà di avere, per ogni variabile esplicativa, una lista di parole chiave sovra o sotto rappresentate rispetto a un modello di riferimento. Inoltre tramite l'analisi delle

[1] Casa delle libertà: *centro destra*: Governo Berlusconi II, XIV Legislatura (30 maggio 2001 - 27 aprile 2006); Coalizione di centro destra: Governo Berlusconi IV, XVI Legislatura (dal 29 aprile 2008 al 23 dicembre 2012); Grande coalizione: XVII Legislatura Governi Letta e Renzi, *centro sinistra e Alternativa popolare*; Indipendente: Governo Dini - (17/01/1995 - 17/05/1996) *governo tecnico*; Indipendenti: Governo Monti (dal 16 novembre 2011 al 27 aprile 2013) *Governo tecnico*, XVI Legislatura; Liberale: Governo Giolitti (1911-1914), *UL - PR - PDC - PD - UECI – CC*, *centro destra*; L'Unione: centro sinistra, XV Legislatura (28 aprile 2006 - 6 febbraio 2008) Governo Prodi II; Pentapartito: Coalizione politica: *DC - PSI - PSDI - PRI -PLI*, IX Legislatura; Quadripartito: Coalizione politica: *DC - PSI – PSDI - PLI*, X Legislatura; Ulivo: *centro sinistra*, XIII Legislatura.

corrispondenze lessicali cerchiamo un pattern che metta in relazione in modo sistematico i lemmi e le dimensioni identificate con le caratteristiche associate ad ogni legge.

4. Risultati

Le leggi sono state analizzate come un unico corpus che soddisfa i criteri standard di dimensione minima richiesta affinché le analisi siano robuste. Ad una prima analisi lessicometrica il testo, costituito da 150.714 occorrenze e 8.113 forme grafiche, rassicura sulla sua adeguata estensione: la proporzione di parole diverse sul totale delle occorrenze (V/N*100= 5,383) si allontana notevolmente dalla soglia del 20% rispettando, quindi, la soglia minima di significatività statistica di un corpus (Bolasco, 1999). Sorprendentemente il livello di ricercatezza del linguaggio non è particolarmente elevato, come si vede dalla percentuale di hapax (V1/V*100) e dal coefficiente a di Zimpf rispettivamente 28,350% e 1,325. Guardando il vocabolario, la prima parola non vuota è comma (1529) seguita da numero (1160) e articolo (1066). Le altre parole *tema*, ovvero quei sostantivi che compaiono con maggiore frequenza nel testo, sono straniero, decreto, Stato, disposizioni, ingresso, territorio e soggiorno.

Abbiamo poi eseguito un confronto tra il nostro vocabolario e il "lessico del discorso programmatico di Governo" (Bolasco, 1999) per individuare quanto fosse peculiare il linguaggio del nostro corpus anche rispetto a un vocabolario tecnico-legislativo. Da questo confronto abbiamo ottenuto uno "scarto" che indica quanto la forma in questione sia sovra (postivo) o sotto-rappresentata (negativo) rispetto al modello di riferimento Bolasco (1999); più lo scarto è alto più le forme sono definite *peculiari* rispetto al testo analizzato, ovvero lo caratterizzano. Senza entrare nel merito delle parole chiave legate al vocabolario prettamente *giuridico* (come decreto, lettera), emerse già dalla gerarchia delle occorrenze, si possono analizzare le altre principali dimensioni del testo: oltre alla parola straniero la prima dimensione che emerge è quella di *frontiera* (ingresso, territorio, frontiera, accesso, durata) e di esercizio di diritto (regolamento, autorizzazione, disposizioni). Ma la dimensione più corposa è quella *delittuosa* (pena, delitti, reato, reati, tribunale, sentenza, condanna, violazione, esecuzione). Fa riflette invece come le parole sottorappresentate siano governo, politica, pubblico, parlamento: ovvero quelle legate alla dimensione *legislativa*.

Partendo dall'ipotesi che il linguaggio sia cambiato nel tempo abbiamo effettuato un'analisi delle specificità (vedi tavola 1). Quando una parola è sovra-rappresentata si parlerà di forma caratteristica (o specificità positiva), al contrario quando essa è sottorappresentata parleremo di specificità negativa; le forme prive di specificità in quel gruppo si definiscono banali,

mentre quelle che non sono specifiche di nessun gruppo sono considerate appartenenti al vocabolario di base del corpus (Bolasco, 1999).

Tavola 1: Specificitá positive per anno di legislatura

1912	1986	1990	1992	1995	1998	2000
cittadinanza	lavoro	entrata	cittadinanza	lavoro	lavoro	visto
legge	lavoratori	permesso	straniera	soggiorno	soggiorno	ottenimento
Stato	immigrati	materia	Stato	permesso	stranieri	professionale
italiana	extracomunitari	frontiera	italiana	entrata	permesso	consente
residenza	sociale	lavoro	figlio	penale	autonomo	seguito
cittadino	lavoratore	previdenza	estero	motivi	attuazione	transito
estero	previdenza	extracomunitari	cittadino	stagionale	motivi	presidente della Repubblica
servizio	autorizzazione	cittadini	servizio	tempo	attivita'	autonomo
Governo	collocamento	decreto	militare	sociale	sociale	tempo
straniera	extracomunitario	stranieri	figli	materia	previdenza	durata
figlio	italiani	apolidi	italiano	caso	europea	societa'
figli	occupazione	soggiorno	entrata	lire	estero	visti
militare	diritti	interno	residenza	legislativo	modalita'	sussistenza
padre	consulta	quanto	acquista	previdenza	regolarmente	requisiti
matrimonio	entrata	prima	età	pubblica	pubblica	importo
2002	**2004**	**2007**	**2008**	**2009**	**2010**	**2011**
lavoro	convalida	soggiorno	prevenzione	penale	conoscenza	seguente
decreto	successive	permesso	pubblico	codice	test	espulsione
soggiorno	legislativo	periodo	legislativo	procuratore	lingua	questore
testo	euro	sensi	ricerca	entrata	italiana	penale
legislativo	modificazioni	legislativo	penale	interno	permesso	termine
permesso	giudice	familiari	sensi	pubblico	lungo	provvedimento
penale	provvedimento	volontariato	procuratore	giudiziario	svolgimento	permesso
asilo	seguenti	unico	procedura	imputato	modalità	periodo
interno	commi	ricongiungiment	in presenza di	europea	soggiornanti	giudice
in presenza di	decorrere	familiare	funzioni	domestico	europeo	rimpatrio
stagionale	provvedimenti	motivi	sicurezza	seguenti	prefettura	respingimento
codice	parole	durata	persona	parole	sistema	lettera
caso	accompagnamen	lungo	ricercatore	comma	legislativo	legislativo
autoritÃ	composizione	nazionale	sostituite	guida	istruzione	allontanamento
procedura	decisione	rilasciato	antimafia	legislativo	rilascio	misure

Dalle specificità ottenute analizzando l'andamento del linguaggio nel tempo emerge come si sia iniziato a scrivere di migrazioni parlando di cittadinanza e residenza, introducendo progressivamente concetti connessi al lavoro e all'essere extra-comunitario arrivando da un lato a temi di integrazione e dall'altro a temi di criminalizzazione dello straniero. Il panorama lessicale nel tempo si è arricchito ma anche "estremizzato". Questa "estremizzazione" potrebbe essere il risultato delle diverse coalizioni/maggioranze e quindi non solo legato ad una dimensione temporale ma ancor di più politica, per questo motivo é bene analizzare le due dimensioni contemporaneamente.

5. Dimensioni lessicali

L'analisi delle corrispondenze lessicali[2] è stata condotta sui primi 50 lemmi estratti dal confronto tra i lemmi dei verbi del nostro vocabolario e quelli del "lessico del discorso programmatico di Governo". Attraverso l'analisi delle corrispondenze abbiamo riassunto la diversità del lessico utilizzato nelle diverse leggi rispetto all'anno e la coalizione di governo. I primi due assi fattoriali, proiettati in figura 1, rappresentano il 46% della variabilità spiegata. La prima dimensione, rappresentata dal primo fattore, è caratterizzata dalla dimensione temporale. Fatta eccezione per il 1992 e il 2007 tutte le leggi approvate dopo il 2002 si contrappongono a quelle precedenti. Il secondo asse è caratterizzato dalla contrapposizione del partito Liberale (Governo Giolitti 1911-1914) e Quadripartito (Governo Andreotti VII 1991-1992), in contrapposizione alle altre maggioranze di Governo. Le coordinate ci permettono di proiettare le classi e le forme grafiche sul piano e il posizionamento ci permette di individuare e interpretare i profili a seconda della vicinanza dei punti.

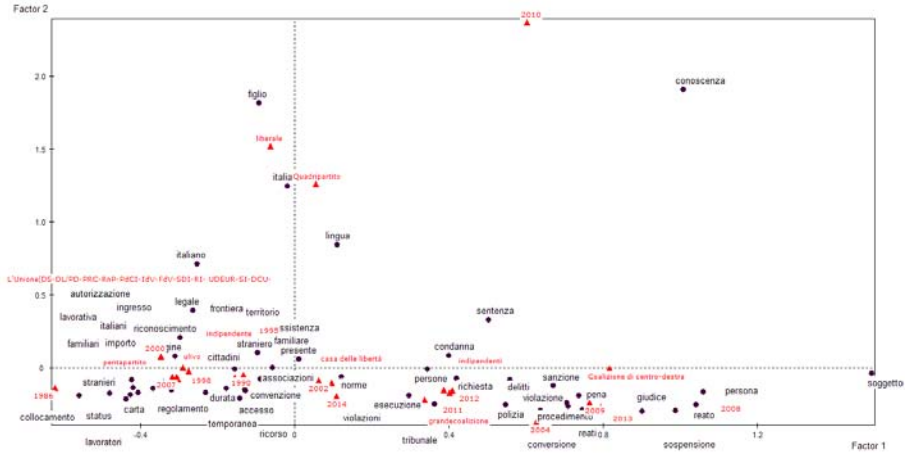


Figura 1: Dimensioni lessicali delle interviste, rappresentazione del primo piano fattoriale

Andando a vedere più in dettaglio i quadranti possiamo notare che nel primo, in cui si collocano il Quadripartito e gli anni 1992 e 2010, le forme grafiche che contraddistingono lo spazio fanno riferimento alla dimensione culturale "lingua" e "conoscenza". Le forme grafiche proiettate nel secondo piano, carattarezzato dagli anni 2002 e a seguire e dalla Casa delle libertá, la grande-coalizione e gli indipendenti del governo tecnico Monti, esprimono

[2] Con l'ausilio del programma Spad, nello specifico con il metodo CORBIT

principalmente gli aspetti legati alla delittuosità (e.g. violazioni, delitti, reato, pena) e giuridica (e.g. norme, tribunale, giudice, esecuzione). A cavallo del primo e secondo quadrante troviamo anche la coalizione di Centro-destra.
Nel terzo quadrante troviamo gli anni 1986, 1990, 1995, 1998, 2007 e il governo tecnico Dini con l'Unione e il Pentapartito. Le forme grafiche su questo piano identificano le caratteristiche del soggiorno come: carta, durata, status, temporanea. Mentre la dimensione di frontiera caratterizza il quarto quadrante: territorio, frontiera, legale, autorizzazione. A cavallo di queste due dimensioni si trovano il mondo del lavoro e quello associativo che sono parte integrante del percorso migratorio in Italia; non sorprende quindi che caratterizzano sia il terzo che il quarto quadrante.

6. Conclusioni

L'obiettivo di questo lavoro era di esplorare il panorama legislativo in riguardo alle migrazioni in un'ottica statistica, con l'obiettivo di estrarre le sue caratteristiche e le sue peculiarità. In questa prospettiva le differenze linguistiche, temporali e soprattutto dei diversi esecutivi rappresentano un'interessante bacino informativo per investigare l'evoluzione semantica delle norme. Seppur descrittivo questo lavoro assume una particolare importanza in quanto la scelta di un tipo di linguaggio potrebbe influenzare opinioni e atteggiamenti nei confronti degli stranieri da parte della popolazione italiana. I nostri risultati mostrano che il panorama lessicale della normativa italiana sull'immigrazione dal 1912 al 2014 è notevolmente mutato. In primo luogo, dal punto di vista delle specificità ottenute analizzando l'andamento del linguaggio nel tempo è emerso che inizialmente, quando l'Italia era un paese di emigrazione, la normativa sulle migrazioni era caratterizzata da temi quali la cittadinanza e la residenza. Dagli anni Ottanta del secolo scorso, con l'incremento dei flussi migratori in entrata nel nostro paese, sono stati introdotti progressivamente concetti connessi al lavoro e all'essere extra-comunitario. Alla fine degli anni Novanta del secolo scorso, a seguito del netto incremento degli arrivi di stranieri in Italia, si inizia a parlare di integrazione e di ricongiungimento familiare. Infine a partire dagli anni duemila inizia il processo di "criminalizzazione" dello straniero pertanto entrano nel vocabolario specifico temi quali sicurezza, respingimento, allontanamento. In secondo luogo, l'analisi delle corrispondenze fattoriali ha confermato che a partire dal 2002 (Legge Bossi-Fini) vi è stato un netto cambiamento del linguaggio usato nella normativa dell'immigrazione, il linguaggio è infatti caratterizzato sempre più da temi legati alla sicurezza e alla legalità. Inoltre il linguaggio usato, è stato senz'altro influenzato da altri fattori che qui non abbiamo preso in considerazione come per esempio, il recepimento delle politiche europee

sull'immigrazione, la situazione geo-politica internazionale, l'incremento degli atti terroristi di matrice islamista a partire dagli attentati negli Stati Uniti l'11 settembre 2001. Con questo lavoro abbiamo delineato un panorama lessicale che ha cambiato direzione orientandosi sempre di più verso temi di regolamentazione e contenimento (espulsione, allontanamento irregolare). Esso ha confermato un approccio negativo riguardo alle migrazioni indipendentemente dalla maggioranza di governo.

Bibliografia

Bolasco S. (1999), Analisi multidimensionale dei dati, Carocci Roma

Colombo, A., & Sciortino, G. (2004). Alcuni problemi di lungo periodo delle politiche migratorie italiane. Le Istituzioni del Federalismo, 5, 763–788.

Nascimbene, B. (1988). Lo Straniero nel diritto italiano. Milano: Giuffré Editore.

Pastore, F. (2004). A community out of balance: nationality law and migration politics in the History of post-unification Italy. Journal of Modern Italian Studies, 9(1), 27–48.

Solé, C. (2004). Immigration policies in southern Europe. Journal of Ethnic and Migration Studies, 30(6), 1209–1221.

Zincone, G., & Caponio, T. (2004). Immigrant and immigration policy-making: the case of Italy. IMISCOE Working Paper Country Report. Amsterdam: IMISCOE.

Zincone, G. (2006). The making of policies: immigration and immigrants in Italy. Journal of Ethnic and Migration Studies, 32(3), 347–375.

A bibliometric meta-review of performance measurement, appraisal, management research

Massimo Aria[1], Corrado Cuccurullo[2]
[1]University of Naples Federico II– aria@unina.it
[2]University of Campania L. Vanvitelli – corrado.cuccurullo@unicampania.it

Abstract
Performance measurement, appraisal, and management have become one of the most prominent and relevant research issues in in management studies. The emphasis on empirical contributions has resulted in voluminous and fragmented research streams. Thus, synthesizing the research literature is relevant for effectively using the existing knowledge base, advancing a line of research, and providing evidence-based insights.
In this paper, we propose a bibliometric meta-review that offers a different knowledge base for future research agenda with implications also for teaching and practice. We analyze the performance management literature through a bibliometric analysis of reviews recently published (2000 - 2017) in the scientific journals of domains, such as Management, Business and Operations. The main purpose is to map and understand the intellectual structure through co-citation analysis.
Keywords: Science Mapping; Content Analysis; Bibliometrix; Performance Measurement.

1. Introduction

Performance measurement, appraisal, and management have become one of the most prominent and relevant research issues in in management studies. They are an ongoing topic of conferences and of books and journal articles as well as of professional and popular grey literature. Researches on these topics have been conducted in different sectors and for various organizations, including public and professional ones. While the number of academic publications on these topics is increasing at a rapid pace, the emphasis on empirical contributions has resulted in voluminous and fragmented research streams that hampers the ability to accumulate knowledge and actively collect evidence through a set of previous research papers. So, literature reviews are increasingly assuming a crucial role in synthesizing past research findings to effectively use the existing knowledge base, advance a line of research, and provide evidence-based insight into the practice of exercising and sustaining professional judgment and expertise. Among the different qualitative and quantitative reviewing, bibliometrics

has the potential to introduce a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity.
In this paper, we propose a **bibliometric "review of reviews" (meta-review)** that offers a different knowledge base for future research agenda with implications also for teaching and practice. The goal of this article is to find a path and to take stock of the existing knowledge in performance measurement, appraisal, and management research.

2. Research Synthesis on performance measurement, appraisal and management

2.1 Overcoming semantic ambiguity

''Performance'' is a complex concept and can be seen from different angles. It is a multi-dimensional construct, the **measurement** of which varies depending on a variety of factors. For example, it is important to determine whether the measurement objective is to assess performance outcomes or behavior at **organizational or individual levels, in financial terms or multidimensional ones** (e.g. balanced scorecard framework), as **intermediate or final** consequence of a managerial action. In very general terms, performance is the contribution (result and how to achieve the result) that an entity (individual, group of individuals, organizational unit, organization, program, or public policy) provides through its action towards achieving the aims and objectives and also the satisfaction of the needs for which the organization was formed.
While measurement concerns performance indicators and **appraisal** is the process of evaluating the performance of individuals and teams, **performance management is a systematic process for improving organizational performance** by developing the performance of individuals and teams. It is a means of getting better results by understanding and managing performance within an agreed framework of planned goals, standards and competency requirements.

2.2 The need of a meta-review

In this work we analyze the performance management literature through a bibliometric analysis of literature reviews recently published (2000 - 2017) in the scientific journals of domains, such as Management, Business and Operations.
The main purpose is to **map and understand the intellectual structure** through **co-citation analysis** of this recent and evolving macro-topic, highlighting internal clusters. The main contribution is to understand better the state of art in terms of gaps, divergences, commonalities and tendencies

in which the field is going on. So, we provide a map to scholars in positioning their future research work and to teachers to introduce so vast topic to students.

This field of research is well suited to a bibliometric meta-review for the following reasons:

1. **there is little consensus among scholars**. For example, Franco-Santos et al. (2007) have counted 17 different definitions for business performance measurement system, while Taticchi et al. (2010) almost 25 diverse frameworks.
2. **the field is deeply multidisciplinary**. The most widely cited authors come from a variety of different disciplinary backgrounds, such as accounting, strategy, operations management and research, human resources. The scholars' background diversity brings different research questions, theoretical bases and methodological approaches. The functional silos, through which research on performance management is developing, impede to have a coherent and agreed body of knowledge. Understanding deeply the intellectual structure of the field and its evolution is a relevant challenge for researchers.
3. there is a community of dedicated scholars around the world that share the same agenda (**cohesion in dominant issues**) but use **divergent theoretical approaches and methods**.
4. **the field is still relatively immature.** As in terms of age it is relatively young, the limited professionalization is not surprising. In addition, there is not a reference journal as Strategic Management Journal for strategy scholars. In this case, our study can be contributive, showing the gaps in literature and providing some guidelines for researchers.
5. **common accepted performance management practices do not exist** (Richard et al., 2009). In many contexts performance management is dysfunctional, although this problem is known since more 50 years (Ridgway, 1956). We still miss more robust empirical and theoretical analysis of performance management frameworks and methodologies. Empirical investigations of the performance impact of frameworks, including the most diffused balanced scorecard, have failed to offer uncontroversial findings (Banker et al., 2000; Ittner et al., 2003; Neely et al., 2004). Some authors call for further and longitudinal studies for understanding the social influences and implications, but they do not show which paths follow.
6. **some publications assumed seminal roles in the evolution of the scientific field.** These articles, owing to their impact, are accelerating factors of development of the field (Berry, Parasuraman 1993). It is therefore important to identify what are the most influential performance

management articles published between 1991 and 2010, to understand better the state of art and discover the linkages among authors.

7. **there is an extended spectrum of this research field and an increased intensity of research, but most part of it also confirms the incompleteness and inconsistence of results.** There are still various open issues and unsolved problems. This depends on the fragmentation of the field of research, on different disciplinary membership of researchers and their cultural context. This diversity implies the use of different theories and methods and therefore also the emergence of different dominant themes.
8. **a profound and rapid evolution is taking place.** Not only the research has shifted from the financial performance to the multidimensional one, but a shift of scholars' attention from the organizational to the individual performance is under way. Moreover, another significant shift is ongoing. While earlier research was often normative, founded on economic rationality, more recent research is more analytical and explanatory (Cuccurullo et al., 2016).

The overwhelming volume and variety of new information, conceptual developments, and data are the milieu where bibliometrics becomes useful by providing a structured, more objective and reliable analysis to present the "big picture" of extant research.

3. Methods

Our bibliometric meta-review is a quantitative research synthesis of the reviews published on the same topic that we conducted with bibliometrix (Aria, Cuccurullo, 2017), a unique tool, developed in the R language, which follows a classic logical bibliometric workflow.

3.1 Data collection

For data retrieval, we used the Social Science Citation Index (Indexes=SCI-EXPANDED, SSCI) of Clarivate Analytics Web of Science. It is the most used database of scientific knowledge by management scholars (Zupic, Čater, 2015). Our search terms were (TS=(("performance manag*") OR ("performance measur*") OR ("performance apprais*"))). We applied our search keyword to the Timespan=2000-2017 and filtered findings for language (English) and document types (Review). Therefore, we found 783 reviews. Then, we refined our search by categories (Management or Business or Operation Research Management Science) and obtained 167 reviews.

Finally, we selected all the reviews published in the most authoritative journals as ranked as 3, 4, 4* by ABS 2015: We dropped off 31 journals for a total of 50 reviews. Our final dataset is formed by 117 reviews.

3.2 Data analysis

Our effort at delineating the intellectual structure of the discipline involves author co-citation analysis (ACA), a bibliometric technique that uses a matrix of co-citation frequencies between authors as its input. This matrix is the basis for various types of analyses.

ACA ability to reveal patterns of association between authors based on their co-citation frequencies makes it a prospective methodology for understanding the evolution of an academic discipline. Authors working in a stream of research often cite one another as well as draw on common sources of knowledge. Further, their works are likely to be frequently co-cited (i.e., cited together) by other authors working on intellectually similar themes. The citations of seminal authors provide a basis for unraveling the complex patterns of associations that exist among them as well as to trace the changes in intellectual currents taking place over time.

4. Findings

4.1 Descriptive analysis

Our dataset includes 117 reviews published in 46 journals since 2000 (table 1 and 3). They received 105 citations on average (table 2). They show fluctuating growth that reaches its peak every 5 years (table 3).

Table 1: Main Information about data

Articles	*117*		
Sources (Journals, Books, etc.)	46	Authors	297
Keyword Plus – Author's Keywords	770 – 383	Authors of single authored articles	10
Period	2000 - 2017	Co-Authors per Articles	2.65
Average citations per article	105.1	Collaboration Index	2.79

Table 2: Top manuscripts per citations

Paper	*TC*	*TCperYear*
1 BHARADWAJ AS,(2000),MIS Q.	1280	71.1
2 DIAMANTOPOULOS A;SIGUAW JA, (2006), BRIT. J. MANAGE.	588	49.0
3 MELO MT et al.(2009), EUR. J. OPER. RES.	587	65.2
4 ZHOU P et al.(2008),EUR. J. OPER. RES.	429	42.9
5 WRIGHT PM;BOSWELL WR, (2002), J. MANAGE.	379	23.7
6 WRIGHT PM et al.(2005), PERS. PSYCHOL.	347	26.7
7 ZACHARATOS A et al..(2005), J. APPL. PSYCHOL.	305	23.5

8 ADAMS R et al.(2006), INT. J. MANAG. REV.	302	25.2
9 GIBSON C;VERMEULEN F, (2003), ADM. SCI. Q.	291	19.4
10 CARDOEN B et al. (2010), EUR. J. OPER. RES.	288	36.0

Table 3: Most Relevant Sources

Sources	*Articles*
1 J. OF MANAGEMENT	11
2 INT. J. OF OPERATIONS & PRODUCTION MANAGEMENT;INT. J. OF PRODUCTION ECONOMICS	8
4 EUROPEAN J. OF OPERATIONAL RESEARCH; INT. J. OF MANAGEMENT REVIEWS	7
6 INT. J. OF HUMAN RESOURCE MANAGEMENT; INT. J. OF PRODUCTION RESEARCH	5
8 J. OF BUSINESS ETHICS; STRATEGIC MANAGEMENT J.	4
10 BRITISH J. OF MANAGEMENT; J. OF APPLIED PSYCHOLOGY; J. OF MANAGEMENT STUDIES; MANAGEMENT ACCOUNTING RESEARCH; OMEGA-INT. J. OF MANAGEMENT SCIENCE; SUPPLY CHAIN MANAGEMENT	3

4.2 Co-citation network and cluster analysis

The objective of our paper is to identify the intellectual structure of the performance measurement and management field. More specifically, our goals are to (1) delineate the subfields that constitute the intellectual structure of the field; (2) determine the relationships, if any, between the subfields; (3) identify authors who play a pivotal role in bridging two or more conceptual domains of research; and (4) graphically map the intellectual structure in a network space in order to visualize spatial distances between intellectual themes. In extreme synthesis, figure 1 shows:

1. A first cluster (red bubbles) represented by works concerning the system of multidimensional performance measurement and evaluation. At its center, we find prominent authors who contribute with specific frameworks, such as the balanced scorecard (Kaplan, Norton, 1992, 1996) and performance prism (Neely et al., 1995). Next to them, we find the contribution of Ittner et al. (2003) about one of the great problems in the multidimensional measurement: the balance between subjectivity and objectivity. Always central in the cluster, we find performance system design (Neely et al., 2000), At the upper and lower extremes of the cluster, we find other two issues of multidimensional performance systems: strategic alignment (Chenhall, 2005) and the guidelines to implement systems (Bititci et al., 1997).

2. A second cluster (blue bubbles) concerns the current prevailing perspective of studying performance measurement and management: the strategic one. In particular, figure 1 highlights the bridge contributions of two cornerstones of the resource-based view (Barney, 1991, Wernelfelt, 1984).
3. In front of this cluster, in the upper-left part of the map, we find another one (violet bubbles) that deals with theories, such as the agency theory: (Eisenhardt, 1989; Jensen, 1976) - which are the main method of investigation - (Carpenter, 2003), and psychology (Kahneman, 1979)
4. Two other neighboring clusters, located in the lower left part of the map, concerns human resources. A first cluster (green bubbles) includes almost entirely works published on Academy of Management. Their preferred theme is perceptions of organizational performance (Delaney, Huselid 1996). A second cluster concerns participation in the appraisal from psychological perspective (Cawley et al., 1998; Keeping, Levy 2000).
5. One last cluster is isolated and concerns the studies of operation research on performance measurement

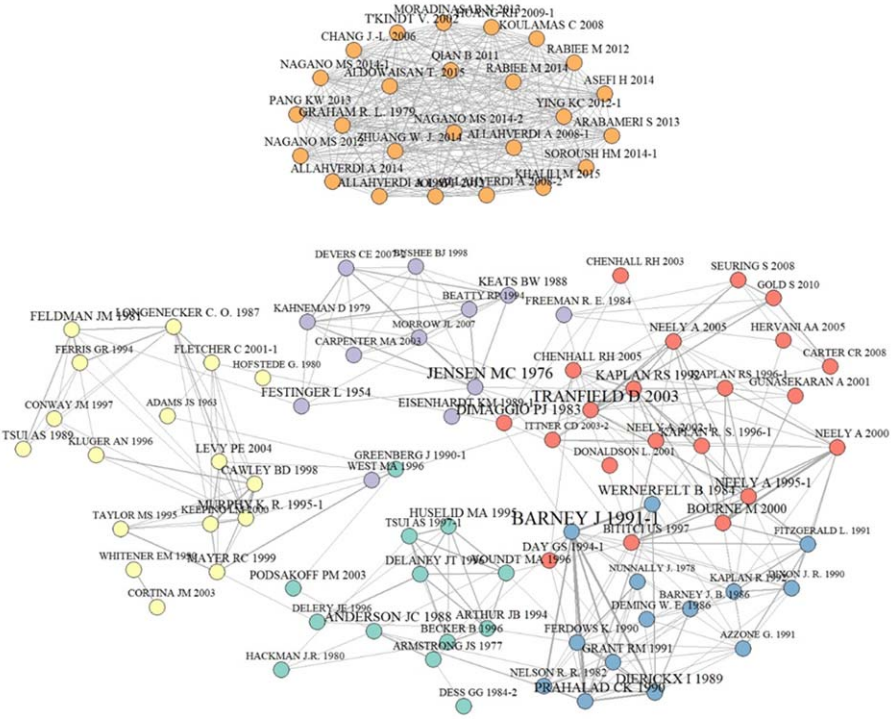


Figure 1: Co-citation network of cited references

References

Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis, Journal of Informetrics, 11(4), pp 959-975.

Barney, J. (1991). Firm resources and sustained competitive advantage. Journal of management, 17(1), 99-120.

Bititci, U. S., Carrie, A. S., & McDevitt, L. (1997). Integrated performance measurement systems: a development guide. International journal of operations & production management, 17(5), 522-534.

Cawley, B. D., Keeping, L. M., & Levy, P. E. (1998). Participation in the performance appraisal process and employee reactions: A meta-analytic review of field investigations. Journal of applied psychology, 83(4), 615.

Chenhall, R. H. (2005). Integrative strategic performance measurement systems, strategic alignment of manufacturing, learning and strategic outcomes: an exploratory study. Accounting, Organizations and Society, 30(5), 395-422.

Cuccurullo, C., Aria, M., & Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains, Scientometrics.

Delaney, J. T., & Huselid, M. A. (1996). The impact of human resource management practices on perceptions of organizational performance. Academy of Management journal, 39(4), 949-969.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. Academy of management review, 14(1), 57-74.

Ittner, C. D., Larcker, D. F., & Meyer, M. W. (2003). Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. The accounting review, 78(3), 725-758.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of financial economics, 3(4), 305-360.

Kaplan, R. S., & Norton, D. P. (1992). In Search of Excellence. Harvard manager, 14(4), 37-46.

Kaplan, R. S., & Norton, D. P. (1996). Using the balanced scorecard as a strategic management system.

Keeping, L. M., & Levy, P. E. (2000). Performance appraisal reactions: Measurement, modeling, and method bias. Journal of applied psychology, 85(5), 708.

Neely, A. (2005). The evolution of performance measurement research: developments in the last decade and a research agenda for the next. International Journal of Operations & Production Management, 25(12), 1264-1277.

Neely, A., Gregory, M., & Platts, K. (1995). Performance measurement system

design: a literature review and research agenda. International journal of operations & production management, 15(4), 80-116.

Neely, A., Mills, J., Platts, K., Richards, H., Gregory, M., Bourne, M., & Kennerley, M. (2000). Performance measurement system design: developing and testing a process-based approach. International journal of operations & production management, 20(10), 1119-1145.

Wernerfelt, B. (1984). A resource-based view of the firm. Strategic management journal, 5(2), 171-180.

Textual Analysis of Extremist Propaganda and Counter-Narrative: a quanti-quali investigation

Laura Ascone
Université de Cergy-Pontoise – laura.ascone@etu.u-cergy.fr

Abstract
This paper investigates the rhetorical strategies of jihadist propaganda and counter-narrative in English and French. Since jihadist propaganda aims at both persuading the Islamic State's sympathisers and threatening its enemies, attention was focused on the way threat and persuasion are verbalised. As far as jihadist propaganda is concerned, the study was conducted on the Islamic State's two official online magazines: *Dabiq*, published in English, and *Dar al-Islam*, published in French. As for the counter-narrative, the corpus was composed of the articles published on the main English and French governmental websites. Combining quantitative and qualitative approaches allowed to examine the general characteristics as well as the specificities of both jihadist propaganda and counter-narrative. The software *Tropes* was used to analyse the corpora from a semantic-pragmatic perspective. The results' statistical validity was then verified and synthesised with the softwares *Iramuteq* and *R*. This study revealed that the rhetorical strategies varied between both jihadist propaganda and counter-narrative, and French and English.
Keywords: jihadist propaganda, counter-narrative, discourse analysis, threat, persuasion.

1. Introduction

The recent terrorist attacks by Daesh in Western countries have led researchers and experts to examine the islamisation of radicalism (Roy, 2016). Different studies have been conducted on the psychosociological contexts that may lead someone to adhere to the jihadist ideology (Benslama, 2016; Khosrokhavar, 2014), as well as on the role played by the Internet in the radicalisation process (Von Behr, 2013). Yet, even though terrorism would not exist without communication (McLuhan, 1978), the rhetorical strategies of the jihadist propaganda have been neglected and remained unexplored. This research investigates the rhetorical strategies of both jihadist propaganda and counter-narrative published on the Internet in English and French. More precisely, this analysis focuses on the way threat and persuasion are expressed in jihadist discourse, as well as on the way French government and international institutions face and counter jihadist

propaganda. From a linguistic perspective, threat and persuasion are complex speech acts. Therefore, pragmatics and, more specifically, Searle's (1969) speech act theory, constituted the basis of this study. As far as jihadist propaganda is concerned, the analysis was conducted on the Islamic State's two official online magazines: *Dabiq*, published in English, and *Dar al-Islam*, published in French. As for the counter-narrative, the corpus was composed of the articles published on the main French and English institutional websites such as *stopdjihadism.fr* or *counterjihadreport.com*. The fact that jihadist propaganda and counter-narrative address different readerships, led us to hypothesise that differences in both content and form might be identified between the two magazines, as well as among the different governmental websites. Combining quantitative and qualitative approaches (Garric and Longhi, 2012; Rastier, 2011) (that is, lexicometry and textometry for the quantitative approach, and the interpretation of the text according to the ideology behind it for the qualitative one), allowed to examine the general characteristics as well as the specificities of both jihadist propaganda and counter-narrative. Following Marchand's (2014) work, the software *Tropes* was used to analyse the corpora from a semantic-pragmatic perspective. The results were then investigated in a qualitative way, and their statistical validity verified with the softwares *Iramuteq* and *R*. The combination of these two approaches allowed to overcome the limitations imposed by both the software's automatic analysis and the qualitative subjective interpretation. By comparing the rhetorical strategies used in both jihadist propaganda (Huyghe, 2011) and counter-narrative, the aim of this research was to identify the linguistic differences between these two discourses and these two languages, in order to determine the rhetorical strategies that might prove efficient in countering jihadist propaganda. After having presented the rhetorical pattern of jihadist propaganda, the linguistic characteristics of English and French counter-narratives will be examined. The jihadist and governmental rhetorical strategies will then be contrasted.

2. Corpus and methodology

2.1. Jihadist propaganda

The analysis of the rhetorical strategies in jihadist propaganda was conducted on Daesh's official online magazines *Dabiq*, published in English, and *Dar al-Islam*, published in French. Since these two magazines address a readership that has already adhered to the jihadist ideology, their goal is to both reinforce the reader's adhesion and incite him/her to act in the name of the jihadist ideology. The reader is then incited to adopt the behaviour a good Muslim should have, and to take revenge on who is presented by Daesh as the responsible for the Muslims' humiliation, that is the West. As

far as *Dabiq* is concerned, the corpus investigated was composed of all the articles published on the first fourteen numbers (*i.e.* 377,450 words). As for *Dar al-Islam*, the analysis was conducted on the first nine issues (229,762 words). To analyse the rhetorical strategies used in jihadist propaganda, a quanti-qualitative approach was adopted (Garric and Longhi, 2012; Rastier, 2011). More precisely, this iterative approach was composed of five stages. A first qualitative analysis of the jihadist ideology, the radicalisation process, and the linguistic characteristics of hate speech and propagandistic discourse was essential to the understanding of the jihadist discourse as well as to the advancement of our first hypotheses. The second stage corresponded to a quantitative analysis whose goal was to verify the validity of our hypotheses: the corpus was then examined with the software *Tropes*, which allows to investigate a text from a semantic perspective. More precisely, based on a pre-established lexicon, the software identifies the themes tackled in the text, and shows how these themes are linked to one another. The most frequent themes in both magazines are *religion* and *conflict*. However, in order to study the way threat and persuasion are expressed in the two corpora, a deeper qualitative analysis was conducted on the themes *sentiment* for the French corpus, and *feeling* for the English one (third stage). In other terms, the quantitative analysis constituted the basis for a qualitative study, which was then conducted only on the expressions conveying feelings. Because of their size-difference, the nine issues of the French magazine count 318 *sentiment-expressions*, whereas *Dabiq* counts 705 *feeling-expressions*. Therefore, in order to contrast the results, a normalisation was applied. Then, a quantitative analysis was conducted with the software *Iramuteq*, which is an interface of the software *R* and which performs statistical analysis of textual data based on Reinert's classification method. This way, it was possible to test the hypotheses and results issued by the qualitative study (fourth stage). Furthermore, a qualitative manual analysis of the first number of both *Dabiq* and *Dar al-Islam* allowed to identify the propositions conveying threat, persuasion, obligation, prohibition, and rewards, that had not been detected by the software *Iramuteq*. This way, it was possible to provide a lexicon specific to the corpus under investigation, that was not detected by the software because of the special features of the jihadist discourse (fifth stage). The combination and alternation of both quantitative and qualitative approaches allowed to examine Daesh's discourses in relation to the context in which it is produced (Valette and Rastier, 2006).

2.2. Counter-narrative

The analysis on the rhetorical strategies in French and English counter-narratives was conducted on the main governmental and institutional

websites. The French corpus was composed of the articles published on *www.stop-djihadisme.gouv.fr* (the platform created after the first terrorist attacks in France in 2015), *www.interieur.gouv.fr* (the website of the Minister of Interior), and *www.cpdsi.fr* (the website of the Centre de Prévention contre les Dérives Sectaires liées à l'Islam). The corpus counts 115,950 words. As far as the English corpus is concerned, it was composed of the articles published on *www.counterjihadreport.com* (a news aggregating website), *www.consilium.europa.eu* (the website of the European Council and of the European Union Council), *www.ec.europa.eu* (the website of the European Commission), and on the Radicalisation Awareness Network (this is a specific section of the website of the European Commission). The corpus counts 116,000 words. In order to conduct comparable analyses, the same quanti-qualitative approach was adopted. The qualitative analysis of the geopolitical context and of the different campaigns used to face and counter the jihadist radicalisation was essential to the understanding of both French and English counter-narratives (first stage). Then, a quantitative analysis was conducted with the software *Tropes*, which allowed to identify the most frequent themes. The themes *religion* and *droit* ("law") were the most present in the French corpus, whereas the themes *education* and *communication* were the most frequent in the English one (second stage). The third stage corresponded to the qualitative analysis that was conducted on the category *sentiment*, for the French corpus (292 propositions), and *feeling*, for the English one (370 propositions). A normalisation was then applied to compare jihadist and governmental discourse. A second quantitative analysis was then conducted with the softwares *Iramuteq* and *R* to test the results issued by the qualitative study (fourth stage). The results of the analysis on jihadist propaganda and counter-narrative were then contrasted to compare the rhetorical strategies used in jihadist propaganda and counter-narrative.

3. The rhetorical strategies used in French and English jihadist propaganda

The quantitative analysis conducted with the software *Tropes*, and the qualitative study conducted on the categories *sentiment* and *feeling*, revealed the components of the jihadist discourse. The propaganda of the Islamic State is based on five key concepts: threat, persuasion, reward, obligation, and prohibition. The assessment of interjudge agreement was necessary to determine these five concepts as well as to categorise the different propositions selected by *Tropes* as objectively as possible. Each category was examined from both quantitative (*i.e.*, its identification and distribution in the two magazines, *Dabiq* and *Dar al-Islam*, using the softwars *Tropes* and *Iramuteq*, and the corpus analysis toolkit *AntoConc*) and qualitative (*i.e.*, analysing each concept in relation to the context in which it was produced)

perspectives. Yet, these five concepts are not independent from one another. Rather, they are strongly linked to one another.

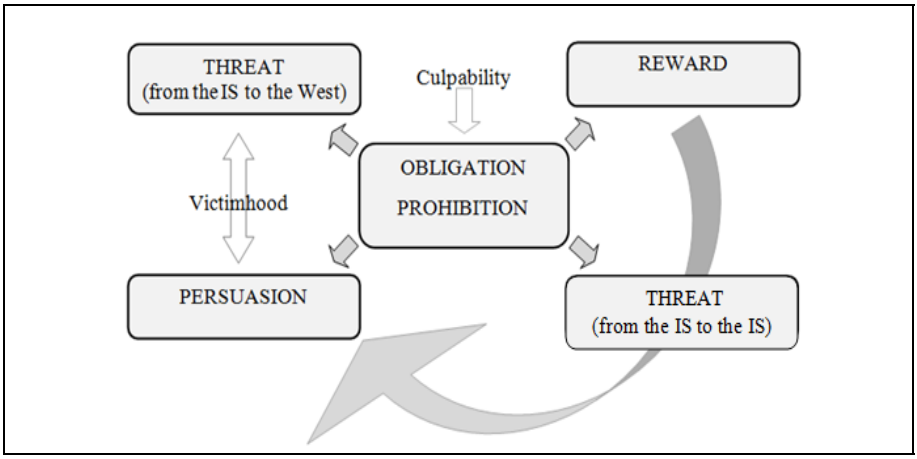


Figure 1: the rhetorical pattern of jihadist propaganda

Figure 1 shows the rhetorical pattern of jihadist discourse. Since *Dabiq* and *Dar al-Islam* aim at manipulating the reader's behaviour, jihadist propaganda is based on obligations and prohibitions. Rewards as well as guilty feelings towards the Muslims living in the Middle-East, aim at leading the reader to respect these prescriptions. Not respecting them would mean facing negative consequences. Threat may then be expressed against the members of the Islamic State themselves and, more in general, against any Muslim. Obligations are also exploited to impose the readership a hostile and violent attitude against Western countries, which is justified by the feeling of victimisation. Fighting against the Muslims' enemy is presented by jihadists as a heroic and valorising action, and therefore, a persuasive one. Furthermore, not only are attractive factors rewards for the reader's obedience. They are sometimes presented as independent from the reader's behaviour. In other terms, persuasion is presented as a positive and valorising act that, contrary to rewards, does not depend on whether the reader respects or not the prescriptions imposed The sentence "*Jihad* is necessary to obtain Allah's forgiveness", for instance, presents an obligation ("it is necessary") and a reward that will be granted if the obligation is respected ("to obtain Allah's forgiveness"). However, this sentence expresses more than an obligation and a reward. *Jihad*, which is interpreted as attractive by jihadists, tends to be associated with terrorist attacks and, consequently, it will be perceived as threatening by Western countries. Furthermore, this

sentence implies that if the obligation is not respected, the individual will not obtain Allah's forgiveness. In other terms, this sentence indirectly expresses a threat against the readership too.

4. The rhetorical strategies used in French and English counter-narratives

The large number of Daesh's sympathisers and *foreign fighters* shows that the communicative and rhetorical strategies adopted in Daesh's propaganda have an important and persuasive impact on the readership. On the contrary, the counter-narrative produced by the different governments to face and counter jihadist propaganda, has been criticised not to be as efficient as jihadist propaganda. In the French corpus, 292 propositions conveying *sentiment* ("feeling") were identified, whereas 370 propositions conveying *feelings* were identified in the English one.

The frequency of the five categories (*i.e.* of the propositions conveying threat, persuasion, reward, obligation, and prohibition) was calculated in the French and English corpora. The reward-category is the only one that was more present in the French corpus than in the English one. Contrary to the Islamic State's propaganda, the propositions conveying rewards and prohibitions are almost absent in both French and English counter-narratives. On the contrary, what these two discourses have in common is the high frequency of the propositions conveying threat (Example 1).

1. "Terrorist groups will continue to exploit the refugee crisis in their propaganda, seeking to portray Western mistreatment of Muslims, and inciting fear by alleging that their supporters are being smuggled in amongst genuine refugees." (RAN website)

As Example 1 shows, threat tends to be associated to the *other* (*i.e.*, the Islamic State), which implies that Western countries are presented as victims of the Islamic State. In the English corpus, 355 occurrences of the word *victim(s)* were identified. The corpus analysis toolkit *AntConc* showed that the most frequent collocation of this term is the word *terrorism* (57 co-occurrences). On the contrary, the French corpus, where the word *victime/s* occurs 70 times only, presents only 2 co-occurrences of the term *terrorisme*. Rather, French counter-narrative tends to talk about rescuing and helping victims (*secours/aide aux victimes*). Furthermore, differences were identified between the different websites in a same language.

Figure 2 shows the under- and overuse of the most representative terms in two French governmental websites: *stopdjihadisme* and CPDSI. More precisely, based on a *Chi2* dependence test, the graph shows the words that are significantly associated or "anti-associated" to the two websites. The figure revealed that CPDSI website focuses more on the religious dimension. The words *islam*, *jihad* and *jihadiste* ("jihadist") are significantly associated to

this sub-corpus. This implies that *jihad* and *jihadiste* are presented and interpreted as religious terms. On the contrary, the website of the *stopdjihadisme* campaign is characterised by an overuse of the words *terroriste* ("terrorist"), *terrorisme* ("terrorism"), *Syrie* ("Syria"), *radicalisation* ("radicalisation"), *Irak* ("Iraq"), *français* ("French"), and *France* ("France"). The overuse of these specific terms shows that the campaign and, consequently, its website focus more on the geopolitical dimension, where the radicalisation process is presented in relation to terrorism and not to Islam.

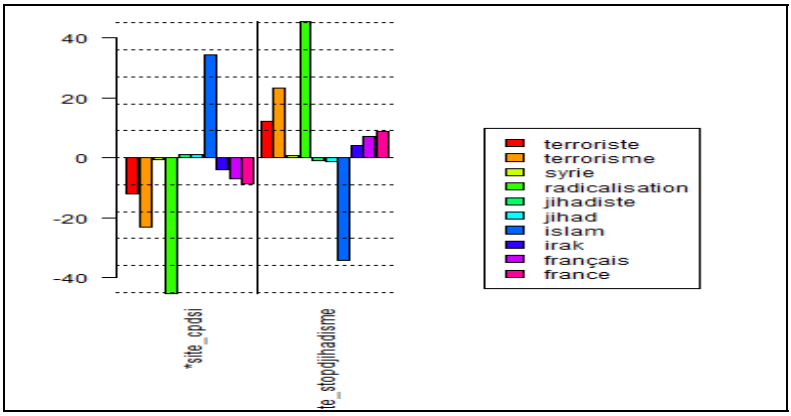


Figure 2: under- and overuse of some key-terms in French counter-narrative

5. Conclusion

This comparative analysis revealed that jihadist discourse and counter-narrative present both similarities and differences. As far as the differences are concerned, the frequency of the propositions conveying threats, persuasion, prohibitions, obligations, and rewards varied between these two discourses: they were more frequent in counter-narrative than in jihadist propaganda. The Islamic State's propaganda aims at reinforcing the reader's adhesion to the jihadist ideology, and at inciting him/her to act against its enemies in the name of the jihadist ideology. On the contrary, counter-narrative does not aim at reinforcing an ideology. Rather, it aims at countering the jihadist radicalisation. This difference was confirmed by the variation of the different category-frequencies in jihadist propaganda and counter-narrative. Despite this crucial difference, similarities between these two discourses were identified. More precisely, both discourses present the respective speakers' communities as victims of the *other* and, consequently, incite the readership to fight, whether violently or not, against the enemy. As far as the methodology is concerned, the procedures adopted allowed to

investigate the general and special features of both jihadist and governmental discourses. The results obtained in the quantitative analysis constituted the starting point for a qualitative analysis, which permitted to identify the features that had not been detected by the softwares as well as to refine *Tropes*'s pre-established lexicon.

References

Angenot, M. (2008). *Dialogue de sourds. Traité de rhétorique antilogique*. Paris : Mille et une nuits.

Benslama, F. (2016). *Un furieux désir de sacrifice : le surmusulman*. Paris : Edition du Seuil.

Garric, N., & Longhi, J. (2012). L'analyse de corpus face à l'hétérogénéité des données : d'une difficulté méthodologique à une nécessité épistémologique. *Langage*, (3) : 3-11.

Huyghe, F.-B. (2011). *Terrorismes : violence et propagande*. Paris : Gallimard.

Khosrokhavar, F. (2014). *Radicalisation*. Paris : Editions de la maison des sciences de l'homme.

Marchand, P. (2014), Analyse avec IRaMuTeQ de dialogues en situation de négociation de crise : le cas Mohammed Mehra. *Communication présentée aux 12es Journées Internationales d'Analyse statistique des Données Textuelles*, Paris, 25.

McLuhan, M. (1978). The brain and the media: The "Western" hemisphere. *Journal of Communication*, 28(4): 54-60.

Rastier, F. (2011). *La mesure et le grain : sémantique de corpus*. Champion ; diff. Slatkine.

Roy, O. (2016). *Le djihad et la mort*. Le Seuil.

Searle, J. (1969). *Speech acts: an essay in the philosophy of language*. London: Cambridge University Press.

Valette, M., & Rastier, F. (2006). Prévenir le racism et la xénophobie : propositions de linguistes. *Langues modernes*, 100(2),68.

Von Behr, I. (2013). *Radicalisation in the digital era: the use of the Internet in 15 cases of terrorism and extremism*.

Analyse de données textuelles appliquée à des problématiques de sécurité et d'enquête judiciaire

Laura Ascone[1], Lucie Gianola[1]
[1]AGORA, Université de Cergy-Pontoise – laura.ascone@etu.u-cergy.fr, lucie.gianola@u-cergy.fr

Abstract
This presentation investigates two cases of textual analysis applied to security contexts: - the analysis of the rhetorical strategies adopted in the Islamic State's official online magazines: Dabiq, published in English, and Dar al-Islam, published in French; - the use of methods for named entities' automatic extraction, and the conception of a textual exploration software for criminal analysis.

Résumé
Nous présentons deux cas d'application de l'analyse de données textuelles dans des contextes liés à la sécurité :
- l'analyse des stratégies rhétoriques de propagande djihadistes à travers l'étude des revues Dabiq et Dar-al-Islam,
- l'utilisation de méthodes d'extraction automatique d'entités nommées et la conception d'un outil d'exploration textuelle pour l'analyse criminelle.
Keywords: analyse de données textuelles, radicalisation, analyse criminelle

1. Introduction

L'essor de préoccupations sécuritaires liées aux actes de terrorisme perpétrés à travers le monde depuis le début du XXIème siècle pousse les chercheurs, acteurs publics et sociaux à rechercher de nouveaux moyens d'analyse de ce phénomène. En France, les sciences humaines et sociales se saisissent de la question comme le démontre l'organisation de plusieurs journées d'études sur la question (« Nouvelles figures de la radicalisation », Toulouse, avril 2017, « Les SHS face à la menace », Cergy, septembre 2017, « Des sciences sociales en état d'urgence : islam et crise politique », Paris, décembre 2017).
Nous souhaitons présenter dans cet article deux sujets d'étude relatifs à ces préoccupations sécuritaires : une étude de la rhétorique de Daesh du point de vue du recours aux émotions dans les revues *Dabiq* (anglais) et *Dar al-Islam* (français), ainsi qu'une collaboration entre le Pôle Judiciaire de la Gendarmerie Nationale (PJGN) et l'Université de Cergy-Pontoise visant à fournir de nouveaux outils d'analyse textuelle des procédures judiciaires aux équipes d'analystes criminels. Le phénomène de la radicalisation djihadiste a amené chercheurs et professionnels à examiner les raisons

psychosociologiques qui sont à la base de l'adhésion à l'idéologie djihadiste (Khosrokhavar, 2014) ainsi que les stratégies adoptées par le groupe extrémiste pour diffuser les messages de propagande (Lombardi, 2015). Toutefois, bien qu'elles jouent un rôle crucial au sein de la propagande djihadiste, les stratégies rhétoriques qui visent à menacer ou à persuader les différents lecteurs restent inexplorées. La première partie de cette étude vise donc à présenter une analyse quanti-qualitative du schéma rhétorique et des émotions sur lesquels se base la propagande djihadiste. Dans la continuité des travaux de Marchand (2014), les logiciels *Iramuteq* et *Tropes* ont permis d'étudier le corpus d'un point de vue quantitatif. Les résultats issus de cette analyse quantitative ont ensuite constitué le point de départ d'une analyse qualitative sur les extraits exprimant des émotions, afin d'examiner plus en détail les stratégies rhétoriques de la propagande djihadiste.

Le cas de l'analyse des procédures judiciaires nous confronte à une problématique typique d'extraction d'information passant par la reconnaissance automatique d'entités nommées : notre travail de recherche consiste notamment à concevoir les bases d'un outil de navigation textuelle *ad hoc*. Bien que les besoins des analystes criminels soient similaires à ceux d'autres domaines d'application (analyse de la voix du client, traitement automatique de la langue biomédicale, etc.), le contexte de l'enquête judiciaire pose de nouvelles contraintes de précision dans l'extraction et dans la mise à disposition des résultats à l'expert, c'est-à-dire à l'analyste criminel.

Le besoin social et institutionnel de nouvelles approches de documents d'origines variées dans les contextes judiciaires et sécuritaires nous permet de démontrer la pertinence de méthodes d'analyse de données textuelles déjà éprouvées dans ces deux cas d'étude.

2. Description de la rhétorique djihadiste : cas des revues *Dabiq* et *Dar al-Islam*

2.1. Corpus et méthodologie

Cette recherche a été menée sur les deux revues de Daech : *Dabiq*, publié en anglais, et *Dar al*-Islam, publié en français. *Dabiq* s'adresse aux sympathisants non arabophones de Daech, tandis que *Dar al-Islam*, qui n'est pas une traduction de *Dabiq*, s'adresse à un lectorat uniquement francophone. Cette distinction nous conduit à avancer l'hypothèse que les deux revues diffèrent dans leur contenu ainsi que dans la forme du message qu'elles portent. Toutefois, l'une et l'autre s'adressent à un lectorat qui a déjà adhéré à l'idéologie islamiste. Leur objectif n'est donc pas de persuader le lecteur de s'approcher de l'islamisme, mais de renforcer son adhésion et de l'amener à agir au nom de cette idéologie. Afin d'analyser les stratégies rhétoriques du discours jihadiste, une approche quanti-qualitative a été adoptée (Rastier,

2011). Plus particulièrement, cette approche itérative était constituée de quatre étapes. Une première analyse qualitative de l'idéologie djihadiste, du processus de radicalisation et des caractéristiques linguistiques du discours de haine a été essentielle à la compréhension du discours djihadiste ainsi qu'à l'avancement des premières hypothèses. La deuxième étape correspond à une analyse quantitative qui a permis de vérifier les hypothèses avancées : le corpus a donc été examiné avec le logiciel Tropes (Ghiglione *et al*, 1998), qui permet d'analyser un texte d'un point de vue sémantico-pragmatique à partir d'un lexique préétabli, et d'identifier les thèmes les plus récurrents dans le corpus ainsi que la manière dont ces thèmes sont liés l'un à l'autre. Afin d'analyser la manière dont le discours djihadiste arrive à persuader et menacer les différents lecteurs (Giro, 2014), une analyse qualitative a été menée sur les thèmes *sentiment*, pour le corpus français, et *feeling*, pour le corpus anglais (troisième étape). En d'autres termes, l'analyse quantitative a constitué le point de départ pour une étude qualitative, qui a donc été menée sur les énoncés exprimant des émotions et des sentiments (Caffi et Janney, 1994). Enfin, une dernière analyse quantitative a été menée avec le logiciel Iramuteq (Ratinaud et Marchand, 2012) qui, basé sur la méthode Reinart, permet, par exemple, de déterminer le sous- et suremploi de certains termes au sein des différents corpus (quatrième étape). La combinaison d'approches qualitatives et quantitatives a permis d'examiner de discours djihadiste en relation avec le contexte dans lequel il a été produit (Valette et Rastier, 2006).

2.2. Résultats

L'analyse des énoncés exprimant des émotions et des sentiments dans les deux revues officielles de Daesh a permis de déterminer le schéma rhétorique sur lequel se construit la propagande djihadiste. Puisque l'objectif de *Dabiq* et de *Dar al-Islam* est de manipuler le comportement du lecteur, la propagande de Daech se fonde sur l'imposition d'obligations et d'interdictions. L'accord de récompenses ainsi que le sentiment de culpabilité visent à amener le lecteur à respecter ces indications. En revanche, tout musulman qui ne respecte pas ces indications, subira des conséquences négatives : il sera jugé d'apostat et il sera donc considéré comme un ennemi. On a ici la menace exprimée par Daech contre les musulmans. Les obligations sont exploitées également pour imposer au lecteur une action violente contre l'Occident, justifiée et alimentée par le sentiment de victimisation. Combattre l'ennemi est présenté comme une action héroïque et valorisante. En participant au combat contre l'Occident, le lecteur aura l'impression de devenir un héros qui lutte au nom d'une cause juste et noble (De Bonis 2015), et de voir ses faiblesses disparaître (Rumman, Suliman *et al* 2016). En outre, en citant des versets coraniques concernant la victoire des musulmans, l'auteur assure à

son lecteur que la communauté musulmane aura la victoire sur l'ennemi ; l'extrait suivant en est un exemple : « *Allah par vos mains les châtiera, les couvrira d'ignominie, vous donnera la victoire sur eux et guérira les poitrines d'un peuple croyant* » (*Dar al-Islam*, n° 8). La victoire sur l'ennemi est perçue par les djihadistes comme persuasive. Toutefois, cet énoncé, perçu comme persuasif par les djihadistes, le sera comme menaçant par l'Occident. De même, le *djihad*, qui est interprété comme persuasif par les membres du groupe djihadiste puisqu'il permet d'accéder au Paradis, tend à être associé aux attentats terroristes et donc à être perçu comme menaçant par les occidentaux. Cette double interprétation rejoint la définition de Perelman et Olbrechts-Tyteca (1988), qui proposent d'« appeler persuasive une argumentation qui ne prétend valoir que pour un auditoire particulier » (p. 36). Bien que *Dabiq* et *Dar al-Islam* présentent le même schéma rhétorique, leur contenu varie de manière conséquente. Cette étude a révélé, par exemple, que la revue française focalise son discours sur la figure de l'*autre* (*i.e.*, de l'ennemi). En revanche, la revue anglaise est focalisée sur la figure du musulman et, plus particulièrement, sur la conduite qu'un bon musulman devrait avoir.

3. Analyse textuelle des procédures judiciaires

Au sein d'une équipe d'enquête, le travail des analystes criminels consiste à lire et synthétiser les documents de procédures (auditions de témoins, données téléphoniques et bancaires, comptes-rendus d'expertise, etc.) afin de fournir aux enquêteurs et aux magistrats une vision plus globale des informations collectées, par le biais de schémas de représentation et de synthèses (Rossy 2011). Leur intervention est requise dans des affaires complexes comme les *cold cases* ou les affaires impliquant de larges réseaux, et permet de fournir de nouvelles pistes d'investigation pour les enquêteurs. À l'heure actuelle, les analystes s'appuient sur un logiciel de reconnaissance optique de caractères, des outils de bureautique classique (traitement de texte, tableur) ainsi que sur le logiciel de représentation graphique d'IBM Analyst's Notebook. Cet outillage ne les dispense pas d'une phase de lecture précise et chronophage de la procédure visant entre autres à repérer et extraire manuellement les informations pertinentes pour l'enquête, regroupées en différents types d'entités qui une fois extraites sont agencées en représentation graphique (chronologique ou relationnelle).

3.1. Corpus de travail

Le corpus de travail mis à notre disposition par le PJGN est une procédure judiciaire complète jugée et résolue concernant un homicide. Le dossier, comme toute procédure judiciaire, rassemble une variété de documents :

rapports d'expertise, procès-verbaux d'investigations, procès-verbaux d'auditions de témoins et de mis en cause, factures téléphoniques détaillées, données bancaires, planches photographiques, etc. Nous avons choisi de concentrer notre travail sur le sous-corpus composé des auditions de témoins et de personnes gardées à vue. Ce choix s'est fait lors de notre prise de connaissance du corpus et du domaine, les auditions représentant la masse d'information la plus dense et la plus difficilement accessible d'une procédure : le nombre des auditions (dans notre cas, 370 auditions pour environ 600 000 mots) et leur manque de structure gênent leur traitement avec des outils standards, contrairement par exemple aux données téléphoniques qui peuvent être intégrées telles quelles dans Analyst's Notebook ou à d'autres données collectées en gendarmerie sous forme de formulaires structurés.

3.2. Détection automatique d'entités nommées

La notion d'entité en analyse criminelle correspond à la notion d'entité nommée (EN) en extraction d'information : une unité linguistique monoréférentielle qui a la capacité de renvoyer à un référent unique (Nouvel & al, 2015). D'une manière générale, cinq types d'entités intéressent les analystes criminels : les personnes, les lieux, les dates et heures, les véhicules et les numéros de téléphone. Nous avons entrepris d'appliquer des techniques de détection d'EN éprouvées sur les documents de procédures judiciaires, tout en variant les approches de manière à répondre au mieux aux contraintes de chaque type d'entité. Deux fonctionnalités du logiciel UNITEX (Paumier, 2016) ont été mises en œuvres : l'édition de grammaires pour la détection des dates, l'utilisation d'un lexique pour la détection des villes, et la combinaison d'un lexique de prénoms et de règles pour les noms de personnes. Les numéros de téléphone quant à eux sont détectés à l'aide d'une expression régulière.

En l'état actuel des choses, nous sommes donc en mesure de détecter :

- Les dates normées : "le 10 janvier 2017", "l'an deux mille dix-sept, le dix janvier", "le 10/01/2017"
- Les noms et prénoms de personnes : "Blanche Rivière", "Petit Noémie", "Michel E. Dupont"
- Plus de 36000 villes figurant dans un lexique[1]

Le développement d'une approche de détection des véhicules, car leurs mentions dans le corpus combinent plusieurs types d'informations :

- genre de véhicule : moto, scooter, camionnette, voiture, etc.

1 Disponible à l'adresse : http://sql.sh/736-base-donnees-villes-francaises (janvier 2018)

- marque
- mention du modèle ou d'une forme (4X4, citadine, berline, break, etc.)
- couleurs et signes distinctifs (rouille, sérigraphie, année du modèle, etc.)

La délimitation de la mention d'un véhicule ne peut se résumer à la combinaison d'une marque et d'un modèle, comme le montrent les deux exemples suivants tirés du corpus :

- *Il s'agit d'un petit modèle comme une TWINGO pour vous donner le volume. Il était de couleur orangé. Il est petit car il a un petit coffre.*
- *M. X. m'a cependant parlé d'un véhicule 4X4 conduit par un individu qui avait un fusil.*

La détection des véhicules nous amènera donc à envisager une approche de détection plus complexe que celles déjà mises en place.

3.3 Analyse de données textuelles et analyse criminelle, une même problématique ?

Si la détection automatique des entités nommées dans le contexte de l'analyse criminelle en gendarmerie constitue une tâche habituelle de TAL, on ne peut pas pour autant en circonscrire les apports potentiels à des aspects purement techniques. La méthodologie de travail de l'analyse criminelle repose sur l'interprétation humaine pour la production d'hypothèses, et en cela nous la rapprochons de l'analyse des données textuelles (ADT) telle que définie par (Ho-Dinh, 2017) : « Avec l'ADT, nous nous situons au contraire dans une perspective de construction des connaissances, par l'interprétation humaine des résultats obtenus grâce à des outils informatiques de calcul et de visualisation. La puissance informatique vient donc en assistance de l'exploration et la fouille des données. Cette différence fondamentale permet de produire des connaissances qualitatives sur les données et non seulement quantitatives. » La poursuite de nos travaux s'oriente donc non seulement vers l'amélioration des résultats de détection d'entités et l'introduction d'approches statistiques (TF-IDF, clustering de documents, etc) mais également vers le développement d'une interface d'exploration textuelle propre, prenant en compte les spécificités du genre textuel de la procédure judiciaire (tri du texte en fonction de sa nature : texte d'en-tête, informations d'état-civil), et permettant une navigation efficace entre entités détectées, mesures statistiques, et texte original. La méthodologie de l'analyse criminelle et les pratiques du métier pourraient être à revoir en conséquence, impliquant une phase de formation des analystes criminels aux méthodes textométriques.

4. Conclusion

Nous estimons avoir soulevé des perspectives théoriques et techniques pour l'analyse de données textuelles dans les domaines judiciaires et de la sécurité, relevant aussi bien de l'analyse de discours que du TAL et de la textométrie. Dans le cas de la propagande de Daesh, l'analyse et la compréhension du discours djihadiste pourraient contribuer à la formulation d'un contre-discours qui puisse faire face et contrer la propagande djihadiste. Concernant les pratiques d'analyse textuelles en analyse criminelle, nous espérons que la mise en place de techniques d'automatisation et d'un outil d'exploration textuelle permette de repenser la méthode d'accès à l'information en analyse criminelle et soit une première étape d'une réflexion plus large sur la collecte et la circulation de l'information et des documents dans le processus judiciaire. Ces deux cas d'études illustrent la pertinence d'approches de sciences humaines et sociales dans le contexte sécuritaire et judiciaire, qui a jusqu'à présent surtout eu recours à des expertises en sciences dites « dures » (médecine légale, biologie, chimie, informatique, etc.), regroupées sous l'appellation de « sciences forensiques ». Nous espérons que de telles contributions permettront de renforcer les liens et d'ouvrir la voie à d'autres projets associant institutions judiciaires et de défense et chercheurs en sciences humaines et sociales.

References

Caffi C., & Janney R. W. (1994). Toward a pragmatics of emotive communication. *Journal of pragmatics*, *22*(3), 325-373.

De Bonis M. (2015). La strategia della paura. *Limes*, *11*.

Ghiglione, R., Landré, A., Bromberg, M., & Molette, P. (1998). *L'analyse automatique des contenus*. Paris, Dunod.

Giro M. (2015). Parigi: il branco di lupi, lo Stato Islamico e quello che possiamo fare. *Limes*.

Ho Dinh O. (2017). *Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français*. Thèse de doctorat, Inalco, Paris.

Marchand P. (2014). Analyse avec Iramuteq de dialogues en situation de négociation de crise : le cas Mohammed Mehra. *Actes des 12èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Paris, pp. 457-471.

Nouvel D., Erhmann M., Rosset S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Editions

Paumier S. (2016). Unitex 3.1 user manual, http://www-igm.univ-mlv.fr/unitex

Perelman C., & Olbrechts-Tyteca L. (1988) (5e éd.). *Traité de l'argumentation.* Bruxelles : Edition de l'Université de Bruxelles.

Rastier F. (2011). *La mesure et le grain: sémantique de corpus*. Champion; diff. Slatkine.

Ratinaud P., Marchand P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IraMuTeQ. *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Liège, 13-15 juin, p. 835-844.

Rossy Q. (2011). *Méthodes de visualisation en analyse criminelle : approche générale de conception des schémas relationnels et développement d'un catalogue de patterns*. Thèse de doctorat, Université de Lausanne, Faculté de droit et des sciences criminelles.

Rumman A., Suliman M. et al. (2016). *The Secret of Attraction: ISIS Propaganda and Recruitmenet*. Traduit par Ward, W. J. et al. Amman: Friedrich-Ebert-Stiftung.

Valette M., & Rastier F. (2006). Prévenir le racisme et la xénophobie: propositions de linguistes. *Langues modernes*, *100*(2), 68.

A two-step strategy for improving categorisation of short texts

Simona Balbi[1], Michelangelo Misuraca[2], Maria Spano[1]
[1] Università di Napoli Federico II – simona.balbi@unina.it maria.spano@unina.it
[2] Università della Calabria – michelangelo.misuraca@unical.it

Abstract
Text categorisation allows organising a collection of documents with respect to their content. When we consider short texts – e.g., posts and comments shared onto social media – this task is harder to achieve because we have few significant terms. Refer to higher-level structures, representing concepts, or topics occurring in the collection, can improve the effectiveness of the procedure. In this paper, we propose a novel two-step strategy for text categorisation, in the frame of feature extraction. Concepts are identified by using network analysis tools, namely community detection algorithms. Therefore, it is possible to organise the document collection with respect to the different concepts and describe the groups of documents with respect to terms. A case study about Pope Francis on Twitter is presented for showing the effectiveness of our proposal.
Keywords: short texts, text categorisation, textual network, community detection

1. Introduction

The ever-increasing popularity of the Internet, together with the amazing progress of computer technology, has led to a tremendous growth in the availability of electronic documents. Therefore, there is a great interest in developing statistical tools for the effective and efficient extraction of information on the Web, in a so-called Text Mining perspective.
The most common reference model for representing documents, in Text Mining, is the so-called vector space model: a document is a vector in the (extremely sparse) space spanned by the terms. Documents are usually coded as bag-of-words, i.e. as an unordered set of terms, disregarding grammatical and syntactical roles. The focus is on the presence/absence of a term in a document, its characterisation and discrimination power. In the knowledge discovery process, the core of the majority of procedures is related to dimensionality reduction, both via feature selection and/or feature extraction. Statistical tools enable an effective feature extraction. One of the most interesting tasks in Text Mining is Text categorisation which allows organising a collection of documents, grouping them with respect to their

content. Here we propose a novel two-step strategy designed for the text categorisation of short documents – e.g., posts and comments shared onto social media – when the task is harder to achieve because we have few significant terms. The basic idea is that Textual data can be processed at different levels, e.g. we can consider single terms, or subsets of terms identifying different concepts, in a feature extraction frame. Concepts are identified by using network analysis tools, namely community detection algorithms. Therefore, it is possible to organise the document collection with respect to the different concepts and describe the groups of documents with respect to terms. The effectiveness of our proposal is showed by analysing a set of tweets about the Pope Francis, posted on November 2017.

2. Background and related work

The bag-of-words encoding is characterised by high dimensionality and an inherent data sparsity. According to Aggrawal and Yu (2000), the performances of text categorisation algorithms decline dramatically due to these aspects. Therefore, it is highly desirable a previous dimensionality reduction.

In pre-processing, feature selection and/or feature extraction are often used before applying any further analysis. Via feature selection, only a subset the original vocabulary is considered, according to with some criterions. Several feature selection techniques are reported in the literature, such as *term strength* (Yang, 1995), *information gain* (Yang and Pedersen, 1997), *Chi-squared statistic* (Galavotti et al., 2000), *entropy-based ranking* (Dash and Liu, 2000). Feature extraction (also known as feature reduction) is a process for extracting a set of new features from the original vocabulary by applying some functional mapping. Common feature reduction techniques include *lexical correspondence analysis* (Lebart *et al*., 1998), *latent semantic indexing* (Deerwester *et al*., 1990). These techniques obtain dimensionality reduction, by transforming the original terms in fewer linear combinations, spanning sub-dimensional spaces, that may not have a clear meaning and sometimes results are difficult to be interpreted.

To cope with this limit, here we consider a different viewpoint. Both feature selection and feature extraction are basically founded on the analysis of a *documents* x *terms* matrix, in which the generic element is the frequency of a term in a document, or another related weight representing the importance of the term. It is possible to get back part of the use context of each term by constructing a *terms* x *terms* co-occurrence matrix. In general, each element of this latter matrix is the number of times two terms co-occur in the corpus. This particular data structure can be represented as a network, where each term is a vertex and each element of the matrix different from 0 is an edge.

The problem of reducing the original dimensionality and perform a feature extraction can be seen as a community detection problem: terms used together define a concept, as in *latent semantic indexing*, or *correspondence analysis*, but without any algebraic transformation. Differently from the approaches previously described, indeed, this method preserves the original meaning of the terms and allows a better readability of the results.
A community in a network is a set of nodes where vertices are densely inter-connected and sparsely connected to other parts of the network (Wasserman and Faust, 1994). There is no universally accepted definition for a community, but it is well known that most real-world networks display community structures. When we consider networks of terms, communities of terms densely interconnected can be interpreted as topics. From a theoretical point of view, community detection is not very different from clustering. Many algorithms have been proposed. Traditional approaches are based on hierarchical or partitional clustering (e.g.: Scott, 2000; Hlaoui and Wang, 2004). The most popular algorithm is the one proposed by Girvan and Newman (2004). The method is historically important because it marked the beginning of a new era in the field of community detection, by introducing the notion of "modularity". Originally introduced to define a stopping criterion, modularity (nowadays refers as Girvan and Newman's modularity) has rapidly become an essential element of many community detection methods, as *fast-greedy* (Clauset et al., 2004), *label propagation* (Raghavan et al., 2007), *leading eigenvector* (Newman, 2006). It measures the difference between the observed fraction of edges that fall within the given communities and the expected fraction in the hypothesis of random distribution. For a most comprehensive review of the community detection literature, it is possible to refer to Fortunato (2010).

3. Problem definition and proposed method

Text categorisation allows to group documents belonging to a collection with respect to the textual content of the documents themselves. When we consider short texts, this task is more difficult to achieve because we have few significant terms for characterising the different groups. The identification of high-level structures representing the concepts/topics occurring in the collection can improve the effectiveness of the grouping procedure. In this paper, a two-step strategy for improving the automatic organisation of a collection of documents is proposed.
Let $\mathbf{T}=\{\mathbf{d}_1, \ldots, \mathbf{d}_n\} \subset \Re^p$ be a set of n document vectors in a term space of p dimension, represented by a *documents* x *terms* matrix, where each element t_{ij} is the occurrence of an i term into a j document (i=1, ..., p; j=1, ..., n). For the purpose of our analysis, we are just interested if the term i occurs in

document *j*, or not. Then we consider a binary matrix **B**, where the generic element b_{ij} is equal to1 if the term *i* occurred at least once in document *j*, 0 otherwise. From the matrix **B** we derive the *terms* x *terms* co-occurrence matrix **A** by the product $\mathbf{A} \equiv \mathbf{B}\mathbf{B}^{\mathbf{T}}$. The generic element $a_{ii'}$ is the number of documents in which the term *i* and the term *i'* co-occur ($i \neq i'$). An element a_{ii} on the principal diagonal represents the total number of documents in the collection containing the term *i*. **A** is an undirected weighted adjacency matrix that can be used to analyse the relations existing among the different terms.

As each community can be seen as a concept/topic occurring in the collection, in order to detect a group of terms defining a concept, we perform a community detection on the matrix **A**. Each community can be seen as a concept/topic occurring in the collection.

As we said above, the greedy algorithm is based on the optimisation of a quality function known as *modularity*. Suppose the vertices are divided into communities such that vertex/term *i* belongs to the community c_i. The modularity *Q* is defined as

$$Q = \frac{1}{2h} \sum_{ii'} \left[a_{ii'} - \frac{\delta_i \delta_{i'}}{2h} \right] s(c_i, c_{i'})$$

where *h* is the total number of edges in the network, δ_i is the degree of the term *i* and the s function $s(c_i, c_{i'})$ is *1* if $c_i = c_{i'}$ and *0* otherwise. In practice, a value above about 0.3 is a good indicator of an interesting community structure in a network.

The greedy algorithm falls in the general family of agglomerative hierarchical clustering methods. Starting with a state in which each term is the sole member of one of *K* concepts, the algorithm repeatedly joins concepts together in pairs choosing in each step the join that results in the greatest increase in modularity.

At the end of the detection process, we obtain a *terms* x *concepts* matrix **C**, a complete disjunctive table where the c_{ik} element ($k=1, \ldots, K$) is 0 or 1 when a term *i* belongs or not to a community. The text categorisation is performed with a clustering algorithm on the matrix *documents* x *concepts* $\mathbf{T}^* \equiv (\mathbf{T}^{\mathrm{T}}\mathbf{C})\mathbf{D}_{\mathbf{K}}^{-1}$, where $\mathbf{D}_{\mathbf{K}}^{-1}$ is the diagonal matrix of the column marginal distribution of **C**. Each cell of $\mathbf{T}^*$ contains the proportion of terms belonging to a concept.

4. A case study

Twitter is one of the most popular – and worldwide leading – social networking service. It can be seen as a blend of instant messaging, microblogging and texting, with brief content and a very broad audience. The embryonic idea was developed considering the exchange of texts like Short Message Service in a small group of users. As of the third quarter of 2017, it has 330 million monthly active users, with an amount of daily sent tweets close to 500 million (Source: *Twitter*, *Statista*). Our aim is to categorise a set of tweets, generated by the same hashtags, with respect to the different concepts expressed in the collection itself.

4.1. Data description and pre-processing

By using the Twitter Archiver add-on[1] for Google Sheet, we collected 24588 tweets about Pope Francis, published between November 10th and December 7th 2017. We use the hashtag #papafrancesco in the query, with any kind of restriction on the language of the tweets. Moreover, we do not filter the so-called retweets, so that some texts are replicated in the *corpus*. The pre-processing was performed in two steps. First, we stripped URLs, usernames, hashtags, emoticons and RT prefixes, and we normalised the tweets by removing special characters and any separators than blanks. Second, on the 23915 cleaned tweets, we performed a lemmatisation and a grammatical tagging. The terms contained in the tweets written in other languages different from Italian were considered as noise.

In the analysis, we consider only nouns because of their content-bearing role. Moreover, we delete from the vocabulary the terms occurring less than 10 times. Thus we obtain a *documents* x *terms* matrix **T** with 23915 rows and 1603 columns, and the corresponding *terms* x *terms* co-occurrence matrix **A**.

4.2. Concept identification and categorisation process

We perform the community detection procedure on **A** in order to identify the concepts. For better highlighting relations among the terms, we fixed a threshold of 30 on the value of co-occurrence, deleting isolated terms. The greedy algorithm detected 38 different concepts. The high value of the modularity measure (Q = 0.648) supports the effectiveness of our procedure results. In Table 1, we list as an example the terms belonging to some of the detected concepts.

[1] https://chrome.google.com/webstore/detail/twitter-archiver/pkanpfekacaojdncfgbjadedbggbbphi

Table 1 – Concepts detected in the collection with corresponding terms

Concept	Terms
2	scienza, sperimentazione, accanimento, responsabilità, malato, cura, eutanasia, …
7	bangladesh, religione, viaggio, cultura, myanmar, discorso, buddista, monaco, …
10	aborto, perversione, febbraio, don, pieri, colonizzazione, crimine, mafia
19	pensiero, figlio, papà, cecilia, moser, monte
23	dramática, miedo, josé, experimentan, condición, maría, marcada, incertidumbre
27	giornatamondialedeipoveri, aula, giovanni, paolo, preparazione, pranzo
…	…

It is interesting to note that the algorithm identifies the concepts not written in Italian (e.g., *#23* contains Spanish terms) and the concepts not related to Pope Francis (e.g., *#19* refers to a popular reality show). By selecting only the terms belonging to the different communities, we obtain a 19799 x 38 matrix $\mathbf{T}^*$. On this matrix, we perform a hierarchical clustering based on the Ward criterion. In Figure 1 it is shown the histogram of the *level indices* obtained by the clustering. The indices represent the loss of inter-class inertia caused by the aggregation. The maximum gap in the distribution suggests to consider a partition in 37 clusters.

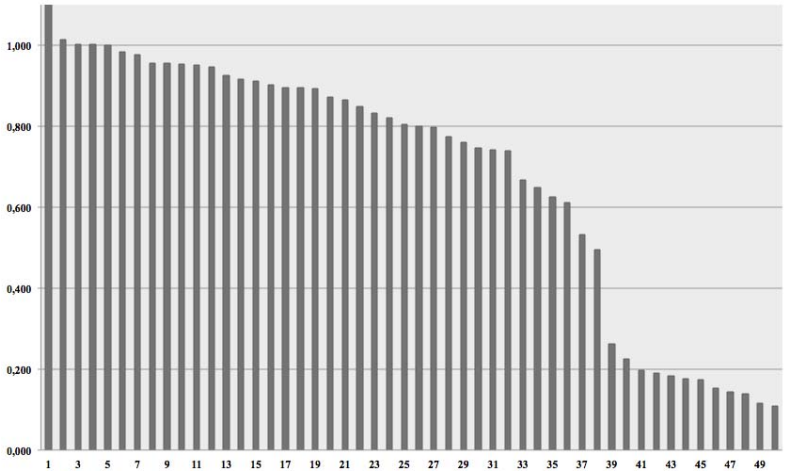


Figure 1 – Histogram of the level indices calculated on the dendrograms' nodes

Because of the unsupervised nature of the approach, the quality of the results can be investigated only by looking at the clusters' composition. Due to the limitation of 140 characters, each tweet can express one to three concepts at most. In Table 2 we can see the concepts occurring in the different clusters. The order of the concepts represents their importance in terms of statistical significance. The preliminary results seem to be very promising, but a deep investigation has to be considered in order to validate the proposal.

Table 2 – Clusters' size and composition

Cluster	Tweets	Concepts	Cluster	Tweets	Concepts	Cluster	Tweets	Concepts
1	120	*6*	**14**	8210	*4, 7*	**27**	51	*30*
2	506	*15, 6, 9*	**15**	536	*1*	**28**	150	*36*
3	95	*9, 15*	**16**	1348	*32*	**29**	163	*37*
4	62	*12*	**17**	1379	*13*	**30**	41	*21*
5	179	*29*	**18**	677	*3*	**31**	51	*28*
6	93	*14*	**19**	2699	*2*	**32**	102	*22, 4*
7	79	*16*	**20**	666	*8, 7*	**33**	71	*26, 22*
8	160	*10*	**21**	48	*24, 20, 13*	**34**	42	*17, 11*
9	445	*5*	**22**	155	*20, 4, 24*	**35**	288	*11, 34*
10	304	*19, 18*	**23**	242	*38*	**36**	125	*34, 11*
11	36	*18*	**24**	55	*25*	**37**	42	*23, 11*
12	66	*31*	**25**	71	*33*	**Total**	19799	
13	335	*27*	**26**	107	*35*			

5. Final remarks

The proposed strategy aims at categorising the documents of a collection by detecting high-level structures, i.e. concepts, as subsets of terms. The terms belonging to each concept are retained in the process and can be used for characterising the identified groups of documents. The tools are given by network analysis tools, namely community detection algorithms. The strategy is suitable when we deal with short texts. Future developments of this work are devoted to set automatically a co-occurrence threshold in the community detection step and to evaluate alternative similarity indices for measuring the relation strength among terms.

References

Aggrawal C.C. and Yu P.S. (2000). Finding generalized projected clusters in high dimensional spaces. *Proceedings of SIGMOD'00*, pp. 70-81.

Clauset A., Newman M.E. and Moore C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6), 066111.

Dash M. and Liu H. (2000). Feature selection for clustering. *Proceedings of Pacific-Asia Conference on knowledge discovery and data mining*, pp. 110-121.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshmanet R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391-407.

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, 486(3): 75-174.

Galavotti L., Sebastiani F. and Simi M. (2000). Feature selection and negative evidence in automated text categorization. *Proceedings of KDD-00*.

Hlaoui A., Wang S. (2004). A direct approach to graph clustering. *Neural Networks and Computational Intelligence*: 158-163.

Lebart L., Salem A., Berry L. (1998). *Exploring textual data*. Springer Netherlands.

Newman M.E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23): 8577-8582.

Newman M.E. and Girvan M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2): 026113.

Raghavan U.N., Albert R. and Kumara S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3): 036106.

Scott J. (2000). *Social Network Analysis: a handbook*. Sage, London.

Wasserman S. and Faust K. (1994). *Social network analysis*. Cambridge University Press.

Yang Y. (1995). Noise reduction in a statistical approach to text categorization. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 256-263.

Yang Y. and Pedersen J.O. (1997). A comparative study on feature selection in text categorization. *Proceedings of ICML-97*, pp. 412-420.

Appeler à signer une pétition en ligne : caractéristiques linguistiques des appels

Christine Barats[1], Anne Dister[2], Philippe Gambette[3],
Jean-Marc Leblanc[1], Marie Peres[1]
[1]Université Paris-Est, CEDITEC (EA 3119), Créteil, France – christine.barats@parisdescartes.fr, jean-marc.leblanc@u-pec.fr, marie.leblanc@u-pec.fr
[2]Université Saint-Louis - Bruxelles, Belgique – anne.dister@usaintlouis.be
[3]Université Paris-Est, LIGM (UMR8049), Champs-sur-Marne, France – gambette@u-pem.fr

Résumé
L'analyse des 12 522 textes d'appel d'une plateforme de pétitionnement en ligne permet d'examiner leurs caractéristiques linguistiques. Le recours à des outils textométriques met ainsi au jour certaines régularités quant aux modalités d'appel à signer. Nous nous intéressons tout particulièrement aux régularités lexicales, aux formes d'adresse ainsi qu'aux modalités d'implication des signataires.
Mots-clés : statistique textuelle, pétition en ligne, textes d'appel

Abstract
The analysis of the 12 522 petition texts of an online petition platform allows to examine their linguistic characteristics. The use of statistical textual analysis tools brings to light several regularities as for the modalities of the call to be signed. We focus on the lexical regularities, the salutations as well as the modalities of implication of the signatories.
Keywords : statistical textual analysis, online petition, petition texts

1. Introduction

Les plateformes de pétitionnement en ligne prolongent et modifient l'acte de pétitionnement (Contamin, 2001). Dans la dynamique des recherches sur l'incidence des dispositifs de participation en ligne sur les formes d'écriture numérique et d'engagement politique (Boure, Bousquet, 2011 ; Mabi, 2016 ; Badouard, 2017 ; Contamin, 2017), nous nous proposons d'interroger les caractéristiques des textes d'appel au regard d'une plateforme numérique de pétitionnement.

Le corpus que nous avons analysé est issu de l'un des principaux sites francophones de pétitions en ligne (lapetition.be). Il se compose de plus de 12 500 pétitions ayant récolté au total 3,25 millions de signatures sur la période comprise entre le 31 octobre 2006 et le 12 février 2015.

Le site propose 9 rubriques parmi lesquelles le porteur de la pétition est tenu

de classer sa pétition : Art et culture ; Droits de l'Homme ; Environnement, nature et écologie ; Humour/Insolite ; Loisirs ; Politique ; Protection animalière ; Social ; Autres. Comme nous l'avons montré ailleurs (Barats et al., 2016) et rappelé en figure 1, les différentes rubriques connaissent des variations importantes tant en termes de nombre de pétitions (figure 1) qu'en ce qui concerne la longueur des textes des appels, le nombre de signatures ou encore le nombre et le volume des commentaires laissés par les signataires. Le choix de la rubrique relève du promoteur de la pétition et témoigne d'une interprétation qui varie selon les porteurs de projet, mais débouche sur des régularités internes à chaque rubrique qui émergent de classifications automatisées du corpus.

Dans cet article, nous nous centrerons exclusivement sur les textes des appels, avec une attention particulière sur leur incipit, afin d'observer quelles sont les régularités lexicales et syntaxiques qui caractérisent les textes d'appel sur l'ensemble du corpus, mais également en contrastant les rubriques. Les 12 522 textes constituent un corpus de 2,6 millions de mots.

Humour / Insolite	397
Art et culture	652
Loisirs	795

Environnement, nature et écologie	1034
Protection animalière	1378
Droits de l'Homme	1738

Social	1806
Politique	2276
Autres	2446

Figure 1 - Distribution du nombre de pétitions par rubrique

2. Les mots les plus fréquents dans les textes d'appel

Afin d'identifier la présence ou non de formes communes aux textes d'appel, nous avons examiné les débuts des textes d'appel, indépendamment des rubriques. La répartition du premier mot des appels ne correspond pas à une loi de puissance (l'habituelle loi de Zipf) car la courbe décroit plus lentement. Les débuts des textes d'appel font donc apparaitre un vocabulaire fréquent particulier. Les 20 formes de cette liste sont en première position dans plus de la moitié des textes de pétitions : *nous, pour, bonjour, le, la, je, les, monsieur, pétition, l, il, a, depuis, non, en, cette, si, madame, contre, suite.*

Si l'on se penche maintenant sur le vocabulaire des 200 formes les plus fréquentes dans l'ensemble des textes d'appel, on constate que les premiers verbes conjugués sont *est, sont, ont, soit, peut, demandons, faut, doit, avons, sommes, demande, sera* et les premiers mots lexicaux *pétition, enfants, pays, personnes, vie, Belgique, France, temps, animaux, monsieur, monde, place, projet, jour, droit, loi, politique, mois, travail, ville, ministre, gouvernement, citoyens, cas, Bruxelles, justice, président, lieu, site, chiens, situation, rue.*

On le voit dans la figure 2, dix formes apparaissent non seulement parmi les 30 mots les plus fréquents (hors mots vides) des appels mais aussi parmi les 30 les plus fréquents en première position des textes : *nous*, *pour*, *je*, *pétition*, *non*, *contre*, *j*, *vous*, *on*, *notre*.

À l'inverse, des mots qui apparaissent avec une fréquence élevée en première position des textes d'appel ne se retrouvent pas parmi les 200 mots les plus fréquents, ou très bas dans le classement : *bonjour* (545), *monsieur* (313), *madame* (141), *chers* (111), *stop* (82), *signez* (80), *mesdames* (73), *appel* (60), *voila* (53), *marre* (45), *messieurs* (41), *cher* (40), *voici* (40), *lettre* (36), *voilà* (30), *trop* (30), *oui* (29), *sauvons* (24), *test* (23), *aidez* (22), *salut* (18).

On trouve ici des formes spécifiques de l'interpellation directe : *bonjour, salut, madame* et *mesdames*, *monsieur* et *messieurs* ou encore *chers*. La présence de *bonjour* ou *salut* rend compte de la diversité des modalités d'interpellation qui renvoient à des niveaux de langue différents et des formulations parfois inattendues. L'accessibilité en ligne du dispositif facilite le lancement d'une pétition : notre corpus se décline sur un continuum qui va des pétitions les plus sérieuses, celles qui trouvent un écho dans la presse, qui auraient sans doute existé sans le dispositif d'une plateforme en ligne, qui sont signées par plusieurs dizaines ou centaines de personnes, aux pétitions très confidentielles, « juste pour rire », dont le texte de l'appel est très réduit et qui récoltent peu de signatures. *Bonjour* apparait avec une plus grande fréquence dans la rubrique « Loisirs ». La forme *test*, quant à elle, révèle certaines difficultés liées au dispositif : il s'agit de tester si une pétition peut être mise en ligne, et le texte de l'appel comprend alors ce seul mot.

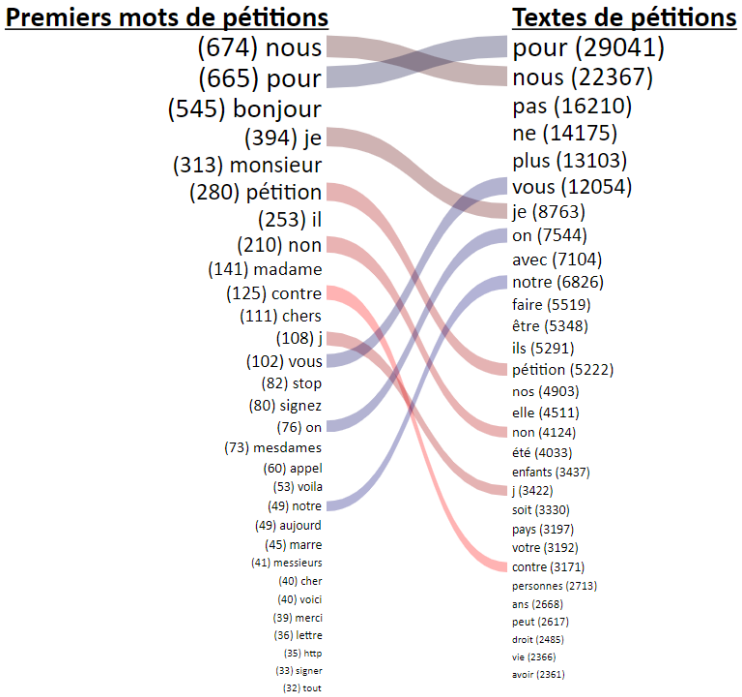


Figure 2 – Visualisation en chaines de fréquences partagées (Lechevrel & Gambette, 2016) des 30 mots les plus fréquents, hors mots vides, en première position et parmi les textes des pétitions.

Deux présentatifs (*voici* : 40 occurrences, *voila/voilà* : 83 occurrences) sont fréquemment attestés en première position des appels à pétition, en particulier dans les rubriques « Loisirs » et « Humour ». La valeur énonciative de ces deux formes est relativement différente. La forme *voilà* est dans un grand nombre d'emplois une marque de l'oralité qui introduit le propos sans en modifier fondamentalement le contenu, mais qui reste un présentatif (« *Voilà je suis une très grande fan du destin de Lisa* », « *Voilà les Tokyo Hôtel refont des tournées* »…). D'autres emplois sont le produit d'une réflexion (« *Voilà, j'ai décidé de faire une pétition* », « *Voilà, je fais cette pétition* ») ou ont valeur de conclusion : (« *Voilà pourquoi il faut avoir peur de l'avenir* »). Cette dernière configuration reste plus fréquente lorsque *voilà* se trouve dans une position autre dans la phrase (« *Voilà le problème* », « *voilà pourquoi j'ai décidé de* »…). Une deuxième catégorie d'emploi, où *voici* et *voilà* revêtent les mêmes valeurs, avec une fréquence plus importante de *voici*, concerne les marques temporelles (« *Voilà quelques années que l'on demande l'autorisation de porter des shorts* », « *Voici 22 mois que je suis papa* »). Enfin *voici* comme *voilà* (dans des

proportions bien moindres pour la seconde forme) prennent une valeur de présentatif dans un grand nombre d'emplois (« *Voilà le but de ma pétition* », « *voilà ma propre pétition* », « *voici une histoire comme tant d'autres* », « *voici une pétition à faire suivre* », « *voici le lien de ma pétition*…).
Avec les verbes à l'impératif *signez, aidez* et *sauvons*, le porteur de la pétition entre directement dans le vif du sujet : il s'agit d'inciter les signataires à agir par l'acte de pétitionnement. *Stop, marre, trop,* et *oui* participent du même mouvement : agir, mettre fin, encourager à, etc. On ajoute à cette liste *pour*, deuxième mot le plus fréquent en première position. Avec *contre*, il est très clairement une marque caractéristique de la posture pétitionnaire : on s'oppose, on soutient. Dans la majorité des rubriques, les textes qui commencent par *non* ou *contre* sont moitié moins nombreux que ceux qui commencent par *oui* ou *pour*, excepté dans la rubrique « Environnement » où ils sont plus nombreux. Nos investigations vont se poursuivre en privilégiant les fonctionnalités d'annotation du corpus offertes par TextObserver afin de davantage prendre en compte les différents contextes d'emploi de ces formes et ainsi renforcer leur désambigüisation.
Les verbes à l'impératif sont un indicateur intéressant d'implication du signataire que l'on retrouve aussi dans l'emploi des pronoms *nous*, *vous* et *je* auxquels nous allons maintenant nous intéresser.

3. L'implication des signataires et des porteurs de pétitions

Le pronom *nous* est particulièrement mobilisé dans notre corpus : mot le plus fréquent au début des appels, il est aussi le pronom le plus utilisé dans l'ensemble du corpus. Ce *nous* se veut mobilisateur : il inclut dès le texte de la pétition les futures pétitionnaires dans l'acte de pétitionnement. Une extraction des 10 mots cooccurrents les plus spécifiques du pronom *nous* placé en première position, à l'aide de l'outil TextObserver (Barats et al., 2013), permet de faire émerger par ordre décroissant de spécificité : *demandons*, *voulons*, *souhaitons*, *soussignés*, *citoyens*, *soutenons*, *réclamons*, *opposons*, *déclarons*, *appris*. Ce pronom introduit très souvent une demande ou une dénonciation, parfois des éléments de contexte (cf. *appris).*
On ne peut évidemment exclure que certains de ces *nous* ne réfèrent qu'aux porteurs de la pétition, sans l'inclusion des signataires. Néanmoins, la présence des cooccurrents *citoyens* et *soussignés* et les retours que nous avons faits aux textes montrent que la grande majorité des *nous* incluent les signataires. Une étude plus approfondie est en cours pour quantifier plus précisément les différents cas. Une interrogation par rubrique confirme l'importance quantitative de ce *nous* inclusif, en particulier dans le cas des rubriques « Environnement », « Politique » et « Social » comme le montre la figure 3(a).

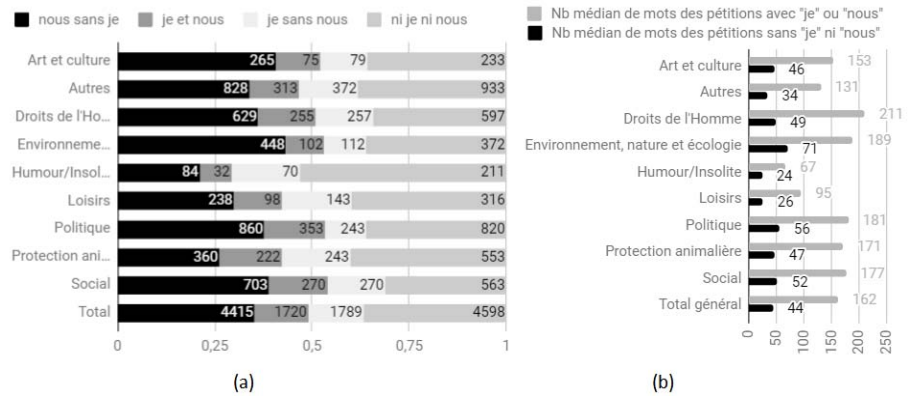


Figure 3 – Nombre de pétitions, par rubrique, dont le texte d'appel contient j', je *ou* nous *(a) et nombre médian de mots des textes de pétitions qui contiennent ou non ces pronoms (b).*

Le pronom *je* arrive quant à lui en quatrième position des mots les plus fréquents en début de texte, et il est le troisième pronom le plus mobilisé sur l'ensemble des textes après *nous* et *vous*. Il n'est pas rare que les deux pronoms *nous* et *je/j'* soient utilisés dans les textes d'appel, le porteur de la pétition passant de son expérience personnelle pour ensuite mobiliser les pétitionnaires, comme dans l'exemple de la pétition suivante intitulée « Contre la fermeture du Delhaize d'Herstal » (pet 14595) : « *Je trouve ça honteux de fermer un magasin qui est récompensé du meilleur rapport clients-Personnel! Il est temps de se serrer les coudes et de se battre jusqu'au bout! Ne nous laissons pas faire!!!!!* ».

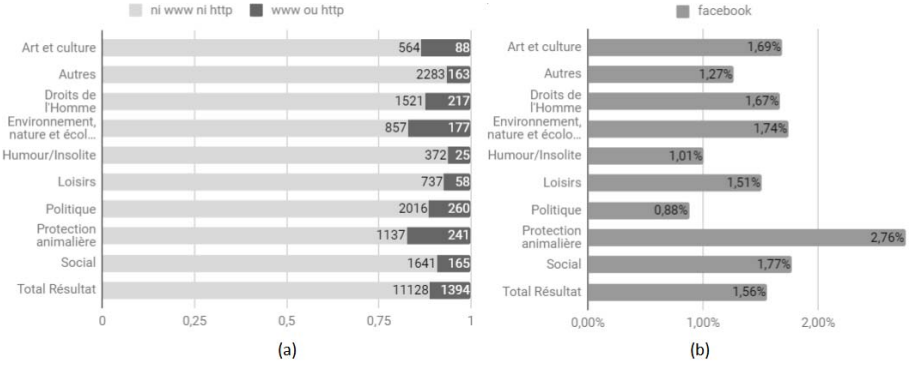


Figure 4 – Pourcentage de textes de pétitions renvoyant ou non à une URL (a) et mentionnant facebook *(b), par type de rubrique.*

Un des moyens de passer d'une implication individuelle à une mobilisation collective est de faire référence à d'autres espaces de relai d'information sur le web, ce qui se traduit par la présence d'URL, qui ciblent parfois des

réseaux sociaux. 11% des appels comprennent des URL. L'incidence des rubriques se confirme : « Protection animalière » et « Environnement » comportent le plus grand nombre d'URL (17%), comme le montre la figure 4(a). Afin d'approfondir ce résultat, nous avons prêté attention à la présence du réseau social Facebook : 1,6% des textes de pétition y renvoient, comme on le voit en figure 4(b). La rubrique « Protection animalière » est celle qui fait le plus appel à des relais via des pages Facebook, confirmant un mode de mobilisation spécifique et transmedia (Barats et al., 2016). La rubrique « Politique » est celle qui fait le moins appel au réseau social Facebook. Notons cependant que la pétition la plus signée sur l'unité de la Belgique, d'aout 2007, a proposé, à l'issue de la fermeture de la pétition, de rassembler sur un site web les photos d'une des manifestations organisées en novembre 2007. Les textes des pétitions rendent ainsi compte de l'articulation de différents dispositifs web dans la dynamique de pétitionnement, qu'une approche strictement quantitative n'indique que partiellement.

On peut s'étonner, en observant la figure 3(a), du nombre relativement important, dans chacune des rubriques, de pétitions dans lesquelles aucun de ces deux pronoms n'apparait et qui serait peut-être le signe de pétitions moins implicantes, plus impersonnelles. En effet, on constate également que moins de 15% de ces textes sans *nous* ni *je/j'* contiennent le pronom *vous*. Si l'on y regarde de plus près, on se rend compte que les textes des pétitions sans *nous* ni *je/j* sont, pour chaque rubrique, beaucoup plus courts que les textes de celles qui incluent *nous* et/ou *je/j*, comme le montre la figure 3(b).

5. Conclusions et perspectives

Notre analyse des premiers mots de textes d'appel de pétitions montre que le vocabulaire utilisé dans cette position présente davantage de régularités liées aux particularités de la pétition que la totalité des textes. Elle permet de repérer quelques caractéristiques linguistiques qui varient parfois selon les rubriques (pronoms personnels, formes d'adresse, URL, etc.).

L'approche textométrique trouve parfois ses limites, comme avec l'ambigüité du *nous* qui peut inclure ou non les promoteurs ou les signataires de la pétition, ou bien dans le cas de la polarité positive ou négative de prépositions et de verbes qui ne suffisent pas à repérer si la pétition traduit plutôt une demande ou une dénonciation.

Ce travail constitue une première étape vers une vérification systématique d'autres marqueurs qui permettent d'impliquer les signataires, comme par exemple la présence de verbes à l'impératif ou de déterminants, en vue d'une mise en relation avec le nombre de signataires et éventuellement de recommandations pour la rédaction de textes de pétitions en ligne.

Références

Badouard R. (2017). *Le désenchantement de l'internet. Désinformation, rumeur et propagande*. Paris, FYP éditions.

Barats C., Leblanc J.-M. and Fiala P. (2013). Approches textométriques du web : corpus et outils. In Barats, C., editor, *Manuel d'analyse du Web en sciences humaines et sociales*. Paris, Armand Colin.

Barats C., Dister A., Gambette Ph., Leblanc J.-M., Peres M. (2016). Analyser des pétitions en ligne : potentialités et limites d'un dispositif d'études pluridisciplinaires, JADT 2016, Nice. http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83043/83043.pdf

Boure R. and Bousquet F. (2011). La construction polyphonique des pétitions en ligne. Le cas des appels contre le débat sur l'identité nationale. *Questions de Communication*, vol. 20: 293-316.

Contamin J.-G. (2001). Contribution à une sociologie des usages pluriels des formes de mobilisation: l'exemple de la pétition en France. Thèse de doctorat, Université Paris 1.

Contamin J.-G., Léonard T. and Soubiran T. (2017). Les transformations des comportements politiques au prisme de l'e-pétitionnement. Potentialités et limites d'un dispositif d'étude pluridisciplinaire, *Réseaux*, vol. 204(4): 97-131.

Lechevrel N. and Gambette P. (2016). Une approche textométrique pour étudier la transmission des savoirs biologiques au XIX^{e} siècle. *Nouvelles perspectives en sciences sociales*, vol. 12(1): 221-253

Mabi C. (2016). Analyser les dispositifs participatifs par leur design. In Barats, C., editor, *Manuel d'analyse du Web en sciences humaines et sociales*. Paris, Armand Colin.

Newsgroup e lessicografia: dai NUNC al VoDIM*

Manuel Barbera, Carla Marello
Università degli Studi di Torino – b.manuel@inrete.it; carla.marello@unito.it

Abstract
VoDIM (Vocabolario dinamico dell'italiano moderno - Dynamic dictionary of modern Italian) represents a new development in recent Italian lexicography. In this paper we argue that NUNC corpora (www.corpora.unito.it), which contain texts from newsgroups that were downloaded at the beginning of XXI century, display aspects of "written-spoken" Italian. NUNC might offer instances of new meaning of "old" words and new collocational contexts. We discuss several examples taken from the corpora, such as the internationalism *Umwelt*, the collocation *assolutamente sì* and the abbreviation *clima* for 'climatizzatore' 'air conditioning'.

Abstract
Il VoDIM (Vocabolario dinamico dell'italiano moderno) rappresenta una grande novità nella lessicografia italiana di questi anni. Qui si argomenta che i corpora italiani della suite NUNC (www.corpora.unito.it), ricavati dai testi presenti nei newsgroup di inizio millennio, sono un buon testimone dell'italiano "scritto-parlato" e potrebbero essere utili per documentare nel VoDIM nuove accezioni e l'uso di nuove collocazioni. Si portano come esempi il caso dell' internazionalismo *Umwelt*, della collocazione di *assolutamente* con *sì* e dell'accorciamento *clima* per 'climatizzatore'.
Keywords: VoDIM – NUNC – Lessicografia – italiano

1. Introduzione

Il VoDIM (Vocabolario dinamico dell'italiano moderno), progetto capitanato dall'Accademia della Crusca[1] che coinvolge otto gruppi di ricerca di altrettante università italiane, fra cui anche il gruppo torinese, sarà un dizionario dell'italiano postunitario online, basato su corpora e su altri dizionari acquisiti in formato digitale come il Tommaseo - Bellini, la quinta Crusca ed il Battaglia, e disegnato per poter essere interrogabile anche a

* A Manuel Barbera si devono i §§ 2 e 3, a Carla Marello i §§ 4 e 5 ed il § 1 va ascritto ad entrambi; anche se ovviamente il lavoro è stato concepito insieme ed entrambi gli autori se ne sentono pienamente responsabili.

[1] Cfr. http://www.accademiadellacrusca.it/it/eventi/crusca-torna-vocabolario-lesicografia-dinamica-dellitaliano-post-unitario.

"corpus variabile", definito dall'utente.
I corpora su cui si appoggia diventano quindi essenziali. Un primo corpus di riferimento base (i cui risultati non sono ancora pubblici: http://dizionariodinamico.it/prin2012crusca/dictionary) è stato prodotto col PRIN 2012 dalla medesima Crusca (in collaborazione con le Università di Catania, Firenze, Genova, Milano, Napoli, Piemonte Orientale, Tuscia e con il CNR), ma, naturalmente, da solo è insufficiente alla bisogna.

2. I NUNC

Un corpus con cui si suggerisce di completarlo è il NUNC-IT; i NUNC (homepage: http://www.bmanuel.org/projects/ng-HOME.html), ideati da Manuel Barbera (in bmanuel.org), ed appannaggio del medesimo gruppo torinese che partecipa al VoDIM, propriamente sono una suite multilingue di corpora che vorrebbe documentare il genere testuale "newsgroup" all'inizio del terzo millennio; molte versioni ne sono state implementate (anche per tematiche specifiche), tutte reperibili dalla homepage; il risultato non è ancora del tutto soddisfacente; pure, qualche uso può già esserne fatto[2].
Un newsgroup è un forum telematico a libero accesso, gratuito, disponibile su Internet, che si manifesta nella forma di testi scritti, i post, inviati ad una "bacheca elettronica" mantenuta presso una rete di server (i newsserver che costituiscono UseNet). Gli utenti del gruppo possono scaricare, leggere e rispondere ai post, costruendo catene (thread) di botte e risposte. I newsgroup sono articolati in una tassonomia precisa, ossia in un sistema di cornici argomentative che si chiamano "gerarchie", a base geografico-nazionale e/o tematica.
I vantaggi di questa base testuale per la linguistica dei corpora sono numerosi e sono stati trattati in Barbera, 2007 e Barbera et Marello, 2009; qui ci interessa in primo luogo il fatto che presentano una *Umgangssprache* assolutamente contemporanea, reale e molto variata per registri e temi.
Per quanto riguarda il VoDIM, molte voci, neologismi, tecnicismi, prestiti, ecc., non sono attestate nel corpus base della Crusca e quindi i NUNC potrebbero risultare utile serbatoio di contesti.

3. Un case study: *Umwelt*

Si veda ad esempio un prestito tecnico, il termine *Umwelt*.
Introdotto (in tedesco) dal biologo (estone, ma di famiglia tedesca del Baltico) Jakob Johann baron von Uexküll già nel titolo della sua importante opera del 1909 (*Umwelt und Innenwelt der Tiere*), è entrato presto nella tradizione

[2] Come dimostrato da alcuni degli interventi presenti in Barbera et al. 2007; in Costantino et al. 2009, per non citare che i primi utilizzi di dieci anni fa.

filosofica (a partire da una recensione di Max Scheler del 1914): usato da Heidegger in un suo corso del 1929-30, è diventato poi moneta corrente (tra gli altri) in francese con Gilles Deleuze, Maurice Merleau-Ponty e Jacques Lacan, nonché in italiano con Giorgio Agamben. Ma è usato soprattutto in testi di biologia, naturalmente, e poi in semiotica, in cui è stato diffuso negli anni Sessanta da Thomas Albert Sebeok (born Sebők Tamás) ed è alla base della moderna biosemiotica (cfr. Kull, 2001).
Nei NUNC il termine è ripetutamente attestato.

> Per Gadamer comprendere l ' esistenza[3] - e qui c'è ancora Heidegger - significa prima di tutto pre-comprenderla , in quanto la comprendiamo con un linguaggio che non scegliamo , ma che , trascendentalmente , definisce già la realtà in cui ci muoviamo : l'Um-Welt , da un lato , e dall ' altro lato , il Mit-welt . Ma , Gadamer cerca di andare alla radice del movimento del pensiero del soggetto e tale origine sta nell ' esigenza di comprendere e farsi comprendere , cioè nel muoversi nell ' Umwelt e nel Mitwelt . Il fatto è che per Gadamer l ' Altro è visibile solo con gli " occhi nostri ", ciò con ciò che " siamo ", con la nostra " identità ", il nuovo si dà solo nel familiare . E in un certo senso è così . L ' altro è ciò che mi disturba che mi inquieta perchè non riesco a ridurlo al mio mondo : è un'eccedenza .

Quello precedente è un esempio dell'uso tecnico-filosofico del termine, che non si discosta molto da quello che si potrebbe trovare nello spogliare i testi (e le traduzioni) di quella tradizione. Più interessante è l'esempio seguente:

> Anche in Italia il consumo di televisione è vertiginosamente aumentato : […] . Oltre a due effetti di rilevanza individuale : - la caduta verticale della capacità di fissare l ' attenzione per più di un certo tempo (se a un buon insegnante occorre anche un ' ora per sviluppare un dato argomento , gli spazi televisivi obbligati in novanta secondi troncano quello stesso argomento in modo irreparabile) e - la perdita di interesse per la lettura - aspetti che coinvolgono per mimetismo inconscio (vale a dire per l ' inconscio occupazione degli spazi mentali ad opera non solo delle immagini ma dell ' intera atmosfera televisiva che foggia l ' Umwelt dell ' uomo moderno) anche persone che fruiscono della TV per tempi ben sotto la media - l ' esposizione allo "

[3] Le citazioni dal corpus sono nel prosieguo riportate *tel quel*: in particolare sono mantenute le tokenizzazioni di interpunzioni ed apostrofi, tutti gli "errori di digitazione", e le idiosincrasie ortografiche proprie del genere.

> sbarramento " delle immagni[4] televisive ha due rilevanti effetti sociali :
> - il conformismo applicato e - l ' ignoranza generalizzata . […]

Si tratta di un traslato, chiaramente fuori dai campi "tecnici" di diffusione del termine. Lessicograficamente ciò è particolarmente rilevante perché testimonia il traghettamento del prestito al di fuori del dominio originario di appartenenza, assicurandone lo sdoganamento all'uso comune, anche se colto o relativamente tale. Per questo tipo di riscontri i NUNC possono rivelarsi particolarmente utili.

4. Al di qua e al di là della parola grafica

Il VoDIM oltre che datare la comparsa di particolari lessemi o di determinate accezioni, si propone anche di attestare la comparsa di accorciamenti e combinazioni di parole: i NUNC, in effetti, presentano usi incipienti passati dal parlato a questa forma di scritto di inizio millennio.

Dal punto di vista della frequenza statistica di tali usi, i dati estratti dai corpora NUNC presentano delle criticità dovute al fenomeno del *quoting*, ma costituiscono una ricca miniera di prime attestazioni: si vedano, ad esempio, lo studio di Onesti et Squartini, 2007 sul modo di dire *tutta una serie di* o di Valle, 2006 sulla penetrazione precoce di anglismi (più o meno italianizzati).

Per quanto concerne gli accorciamenti, in particolare, in Allora et Marello, 2008 ne abbiamo dato una nutrita raccolta. Un esempio per tutti è *clima* come accorciamento di *climatizzatore*; Marello l'aveva già fatto oggetto di un breve articolo[5] e ne aveva constatato la presenza in più post del 2002 di NUNC-Motori. Si veda il brano di thread in cui compare anche un disinvolto *conce* per *concessionario*[6]:

> Qualcuno e' in grado di dirmi quanti grammi (olio/gas?) servono per la ricarica del clima per un CRD del 2002? Una spesa approssimativa? Grazie
> Ciao a tutti, scusate se mi intrometto, ma oggi dopo giorni di dubbio ho chiamato il conce per lo stesso motivo di Massimo,30 km per sentire un po' di aria fresca con il clima impostato a 5 gradi e macchina lasciata

[4] Come si diceva, le citazioni dal corpus sono riportate *tel quel*, ivi compresi gli errori presenti nella fonte. Tantopiù che la maggiore tolleranza alle cattive digitazioni, e l'aperta accettazione di alcune caratteristiche grafico-ortografiche, sono tipiche di questo genere di CMR.

[5] Apparso sul Corriere del Ticino il 23 settembre 2005

[6] Non approdato questo agli onori della registrazione nei dizionari, come invece accade per *clima* la cui data di prima attestazione è secondo il dizionario Zingarelli il 2000.

prima all'ombra

Al di là della parola grafica può, ad esempio, essere interessante documentare gli usi di *assolutamente sì*[7]: se ne trovano ben 103 nei NUNC generali. Ecco due esempi:

> Ma ti senti tanto tanto tanto depressa ??? Ci dobbiamo preoccupare ? [>]… Oggi un pò meno , però devo dire che ho passato veramente dei brutti momenti. L ' importante è riprendersi , no ? Assolutamente sì ! Riprendersi e ripartire subito !

> tu sei un troll ? […] No , perché il flame occasionale non fa di una persona un troll - werted è un troll ? Assolutamente sì , perché attua flame , insulti e provocazioni in modo sistematico e con offese che vanno oltre l ' ambito dello sfottò sportivo . In più utilizza tutte le tecniche tipiche del trollaggio , dal morphing al faking al flooding .

Stessa indagine si può fare per *anche no,* constatando che è nella stragrande maggioranza dei contesti è *ma anche no.*

5. Conclusioni

Un ulteriore fattore che rende i NUNC apprezzabili per il linguista e il lessicografo attento all'uso è la dialogicità, che si intravede soprattutto negli esempi presentati nel § 4. È un fenomeno pervasivo nei NUNC, di solito declinato nei newsgroup come *quoting* (cfr. Barbera, 2011 e Marello, 2007). Computazionalmente ciò crea, è vero, alcuni problemi (ancora non del tutto risolti), dato che il fenomeno del testo ripetuto, se incontrollato, va inevitabilmente ad intaccare l'aspetto statistico, vanificando un semplice uso quantitativo dei corpora; però testualmente è un fenomeno di grande importanza, specie se valorizzabile, come nei NUNC, con la possibilità di potere allargare i contesti fino a 2000 parole.

La capacità dei newsgroup di fissare nello scritto usi eminentemente orali, di trasferire la fluidità dell'oralità ad uno speciale tipo di scrittura, costituendo una sorta di ponte tra i due media, può rivelarsi particolarmente importante per il VoDIM, proprio perché i corpora NUNC registrano tendenze emergenti nella lingua italiana. Sulla peculiarità diamesica di questo particolare tipo di "scritto-parlato" abbiamo sostato in Barbera et Marello, 2009, ma qui non possiamo non rimarcarne l'opportunità che potrebbe presentare per il VoDIM.

I NUNC, come dicevamo, non sono ancora perfetti: i prototipi che sono stati

[7] Oggetto di un articolo sul Corriere del Ticino del 21 gennaio 2004.

messi online sono solo delle beta, ma la volontà di perfezionarli c'è: e non è da escludere che il VoDIM rappresenti l'occasione giusta per farlo.

Bibliografia

Allora A. e Marello C. (2008), "Ricarica clima". Accorciamenti nella lingua dei newsgroup, in Cresti E., editor, *Atti del IX Congresso della Società Internazionale di Linguistica e Filologia Italiana (SILFI): "Prospettive nello studio del lessico italiano" (Firenze, 14-17 giugno 2006)*. Cesati: vol. II, pp. 533-538.

Barbera M., Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it, in Barbera M., Corino E. e Onesti C., editors, *Corpora e linguistica in Rete*. Guerra Edizioni: pp. 3-20.

Barbera M., Une introduction au NUNC: histoire de la création d'un corpus, in Ferrari A. et Lala L., editors, *Variétés syntaxiques dans la variété des textes online en italien: aspects micro- et macrostructuraux*. Université de Nancy II, 2011: pp. 9-36.

Barbera M. e Marello C. (2009), Tra scritto-parlato, *Umgangssprache* e comunicazione in rete: i corpora NUNC, in Antonini A. e Stefanelli S., editors, *Per Giovanni Nencioni. Convegno internazionale di studi*. Pisa - Firenze, 4-5 Maggio 2009. Le Lettere: pp. 157-86. Poi in Barbera M., *Quanto più la relazione è bella: saggi di storia della lingua italiana 1999-2014*, Bmanuel.org - Youcanprint, 2015: pp. 157-182.

Costantino M., Marello C. e Onesti C. (2009), La cucina discussa in rete. Analisi di gruppi di discussione italiani relativi alla cucina, in Robustelli C. e Frosini G., editors, *Atti del convegno ASLI 2007 "Storia della lingua e storia della cucina. Parola e cibo: due linguaggi per la storia della società italiana". Modena, 20-22 settembre 2007*. Cesati: pp. .717-727.

Kull K. (2001), Jakob von Uexküll: An introduction. *Semiotica*, vol. 134 (1/4): pp. 1-59.

Marello C. (2007), Does Newsgroups "Quoting" Kill or Enhance Other Types of Anaphors?, in Korzen I. and Lundquist L., editors, *Comparing Anaphors between Sentences, Texts and Languages*. Samfundslitteratur Press: pp. 145-157.

Onesti C. e Squartini M. (2007), "Tutta una serie di". Lo studio di un pattern sintagmatico e del suo statuto grammaticale, in Barbera M., Corino E. e Onesti C., editors, *Corpora e linguistica in Rete*. Guerra Edizioni: pp. 271-284.

Valle L. (2006), Varietà diafasiche e forestierismi nell'italiano nei gruppi di discussione in rete, in López Díaz M. et Montes López M., editors, *Perspectives fonctionnelles: emprunts, économie et variations dans les langues. S.I.L.F. 2004. XXVIII Colloque de la Société internationale de linguistique*

fonctionnelle, tenu à Saint-Jacque-de-Compostelle et à Lugo du 20 au 26 septembre 2004. Editorial Axac: pp. 371-374.

Zingarelli N. (2017), *Lo Zingarelli 2017. Vocabolario della lingua italiana*. A cura di Mario Cannella e Beata Lazzarini. Zanichelli.

Techniques for detecting the normalized violence in the perception of refugee / asylum seekers between lexical analysis and factorial analysis

Ignazia Bartholini
Univ. of Palermo - ignazia.bartholini@unipa.it

Abstract 1
The theme of gender violence finds a peculiar declination if linked to the phenomenon of forced migrations, and intersects historical-cultural variants of neo-patriarchal nature to cultural-religious orthodoxies the newcomers often bear with them. Studying gender violence in the context of globalized migrations allows us to highlight three bias that mark the western discourse and that concern the way of conceiving its phenomenology as pre-modern (a); detaching violence interpretation from politics of intervention and contrast (b); considering gender asymmetries, sexist representations and practices in the Mediterranean hosting society as residual (c). Subsequently, the factorial structure of the questionnaire was investigated through the Principal Components Analysis (ACP) and the subsequent *Oblimin* rotation of the factorial axes, as a relation between the dimensions of the questionnaire was assumed. The reliability of the scales was verified by the *Cronbach alpha coefficient*.

Abstract 2
Il tema della violenza di genere trova una declinazione peculiare se collegato al fenomeno delle migrazioni forzate e interseca le varianti storico-culturali di natura neo-patriarcale alle ortodossie culturali-religiose che i nuovi arrivati portano spesso con loro. Studiare la violenza di genere nel contesto delle migrazioni globalizzate ci consente di evidenziare tre pregiudizi che segnano il discorso occidentale e che riguardano: il modo di concepire la sua fenomenologia come premoderna (a); la searazione fra l'interpretazione della violenza e le politiche di intervento e contrasto (b); il considerare le asimmetrie di genere, le rappresentazioni sessiste e le pratiche Mediterranee come residuali (c). Successivamente, la struttura fattoriale del questionario è stata analizzata attraverso la Principal Components Analysis (ACP) e la successiva rotazione *Oblimin* degli assi fattoriali, essendo stata ipotizzata una relazione tra le dimensioni del questionario. L'affidabilità delle scale è stata verificata dal coefficiente alfa Cronbach.
Keywords: gender violence, forced migrations, sexist representation

1. Introduction

Over the last two decades, the field of border and migration management has been characterized by the increasing interrelatedness of discourses about control practices and about humanitarian issues (Walters 2011, Fassin 2010).

Today, European policies seek to incorporate strategies to support forced migrants as key instruments for the protection of refugees (Moro 2012). Forced migration, which can also be addressed through the lens of gender (Hans 2008), is grafted onto a broader field of research, which includes welfare strategies, social representations and intercultural dynamics. According to the UNHCR, gender-based violence refers to "any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or private life" (UNHCR 2008: 201). It can take, among others, the form of "rape, forced impregnation, forced abortion, trafficking, sexual slavery, and the intentional spread of sexually transmitted infections, including HIV/AIDS" (UNHCR 2008: 7, 10).

Forms of violence happen not only inside the migratory journey by other refugees, but also by public officers, government employees, aid agencies crew (Ferris 2007; Freedman 2015).

2. The numbers of the phenomenon

According to data of the Italian Ministry of Internal Affairs, between 2015 and 2016, 154719 migrants disembarked in Italy, of which 82136 asylum seekers. From January to March 2016 9,307 migrants disembarked in Italy. Currently, migrants come mostly from Gambia, Senegal, Mali, Guinea, Ivory Coast, Morocco, Somalia, Sudan and Cameroon (Source: ANSA).

In January 2016 asylum seekers were 7,505, mostly from Pakistan (1510), Nigeria (1306), Afghanistan (665) and Gambia (625). Among these, 6739 were men, 766 women, 292 unaccompanied minors and 199 minors. 6507 requests were reviewed so far with the following outcomes: 190 people (3%) were granted the refugee status, 698 (11%) obtained a subsidiary permit, 1352 (21%) were granted with a humanitarian protection and 4266 (66%) were denied (source: Italian Ministry of Internal Affairs).

Only in the 2017, from the Hotspot Trapani-Milo, managed by "Badia Grande NGO" one of partners of the project " Provide ", have transited 21,478 refugees / asylum seekers (Source - Ministry of Interior), with 21 different nationalities. These include 16,010 men, 3177 women, 2291 children divided in 1787 males and 504 females.

Last year, two researchers from the University of Palermo submitted a questionnaire of 36 items to 465 women, temporarily hosted at the Trapani-Milo Hotspot in Sicily.

3. Objectives of research

The core question of the research concerns the identification of violence's subjective dimensions from the side of the victims and the operators, as well as the problems in building social multicultural constructions of violence.

The research wants to identify violence's subjective dimensions from the side of the victims and the operators, as well as the problems in building social multicultural constructions of violence.

For this purpose, the research investigates a specific articulation of the "migratory violence," which entails cultural specificities and contextual conditions, such as the journey and the time spent in reception facilities. In order to highlight topics and problems related to the social construction of gender violence, attention will be paid to victims' point of view concerning the 'normalized' procedural violence, even by means of operational definitions of victims' first reception treatments in the institutional arenas.

Furthermore, gender relations are biased by the whole migration experience, and this leads to various forms of direct, indirect and structural violence: forms of gender-based violence are seen not only among refugees. Finally, refugees and asylum seekers may suffer structural violence in the form of social exclusion and discrimination (Jaji 2009, Crisp, Morris & Refstie 2012), secondary victimization (Pinelli 2011, Tognetti 2016), labour exploitation (Coin 2004), forced prostitution (Naggujja et al 2014, Krause-Vilmar 2011) and sexual abuse (Crisp, Morris & Refstie 2012). Therefore, the migratory violence to which women—as well as minors and LGBT—are subjected, becomes, a particular mode of reading and interpretation of intra- and intercultural gender relations.

For the first part of the research's objective, was to assess the perception of the violence suffered of the women of sample before and during the journey to the coast of Sicily.

For the second one of the research's objective, was to individuate some effective interventions for the reduction of the migrant' exposure to different types of violence and threat, to encourage the access to physical and psychological services, to assist the violence' victims with integration, support safe and appropriate cultural instruments , to provide support for families, stable settlement in host country and to concerted actions for reducing the inequalities in access to resources.

4. Methodology

A1. Once the ethnic intersection, socioeconomic gender and status explored, an internalist perspective will be employed, based on the analysis of the narrative devices, that is the conversations' reports that migratory violence victims conduct with experts (linguistic and intercultural mediators, social assistants, psychologists and lawyers, but also doctors and police officers) or with members of the third sector.

A2. Definitions of lived or experienced violence, through interviews to refugees and operators in the first and second reception centres, that have particular acquaintance with the phenomenon;

Subsequently, the factorial structure of the questionnaire investigated through the Principal Components Analysis (ACP) and the subsequent *Oblimin* rotation of the factorial axes, as a relation between the three dimensions of the questionnaire was assumed:

a. the daily life before the trip;

b. the gender dynamics and relationships among the family members;

c. the violence normalized.

The reliability of the scales was verified by the *Cronbach alpha coefficient*.

In order to verify the hypothesis, that there are statistically significant differences to the mean scores of the different dimensions, analyzes of the variance have been carried out. Multivariate analysis techniques on variance, together with a lexical analysis, allowed us to select:

1. the keywords present in the corpus of the questionnaire using frequency indexes;

2. the meta-information contained within the text units;

3. the context units through specific data arrays for content analysis

The communication that we propose to present will describe the results of the research conducted and the methodological opportunity of the text analysis tools used by the researchers involved.

5. Some Research's results

To individuate the vulnerabilities of migrants, it was necessary to identify appropriate instruments of analysis for being able the needs of violence victims and in order to deal with them in a respectful, sensitive, professional and non-discriminatory manner. The have explained the need to receive the proper degree of assistance and a stronger support and protection. The keywords more frequently used by migrants are been: protection, fear, opportunity, work, life.

The content analysis, and the context units involved through specific data, describe the necessity to acknowledge the women/asylum seekers, who could be victims by other men after their arrive in reception center too and the opportunity to put specific procedures to prevent, identify, and respond to the different forms of proximity gender-based violence.
The content analysis, and the context units involved through specific data, describe the necessity to acknowledge the women/asylum seekers, who could be victims by other men after their arrive in reception center too and the opportunity to put specific procedures to prevent, identify, and respond to the different forms of proximity gender-based violence.

6. Conclusion

The problems that refugees face require humanitarian responses and effective interventions (Dal Lago 1999; Colombo 2012; Camarrone 2016), such as the reduction of exposure to different types of violence and threat in post-migration phase and the access to physical and psychological services (Shamir 2005; Ambrosini 2010; Bartholini 2017). From this perspective, the Mediterranean represents a peculiar field of analysis of that normalized violence – procedural and proximal – that denies refugees/asylum-seekers, minors and LGBT people to consider themselves as right holders and subjects of the same dignity and value.

Morevor, the results or content analysis shows the necesity of a stronger integration, with a support strategies of appropriate cultural s and social practices and to provide adeguate support for families in a stable settlement in our host countries (Balibar 2012). Lastly, the research highlights the need of some concerted action to reduce inequalities in access to resources (Robinson et al. 2006).

Gender violence related persecution may give rise to claims for international protection (Gilbert 2009).

Council of Europe Convention on preventing and combating violence against women (Istanbul Convention of 2011) and the Directive 2012/29/EU in establishing minimum standards on the rights, support and protection of victims, contribute to achieve the obligation to "ensure access for victims and their family members to general victim support and specialist support, in accordance with their needs".

Although member states are stepping up their work in order to streamline a gender understanding into public decision making, policy and operations, this effort is not always reflected in the asylum procedures.

References

Ambrosini M. (2010). *Richiesti e respinti. L'immigrazione in Italia. Come e perché*. Milano: il Saggiatore.

Balibar E. (2012). Strangers as enemies. Walls all over the world, and how to tear them down. *Mondi Migranti*, Vol. 6, n. 1: 7-25. DOI: 10.3280/MM2012-001001

Bartholini I (2017). *Migrations: A Global Welfare Challenge: Policies, Practices and Contemporary Vulnnerabilities*, (with F. Pattaro Amaral; A. Silvera Samiento; R. Di Rosa), Edition Corunamerica, Barranquilla (Colombia), p.1-196 (ISBN 978-9588-59812-2-5).

Camarrone D. *Hotspot di Lampedusa, la sindaca chiede al Ministero dell'interno una verifica urgente delle procedure UE*, Diritti e frontiere, 8 gennaio 2016, in http://dirittiefrontiere.blogspot.it/2016/01/la-verita-sul-sistema-hot-spot.html

Colombo A. (2012). *Fuori controllo? Miti e realtà dell'immigrazione in Italia*. Bologna: Il Mulino.

Coin F. (2004). *Gli immigrati, il lavoro, la casa*. Franco Angeli: Milano.

Convenzione di Dublino (1990), in http://www.camera.it/_bicamerali/schengen/fonti/convdubl.htm

Crisp J., Morris T. & Refstie, H. (2012). Displacement in urban areas: new challenges, new partnerships. *Disasters*, 36(1): S23-S42.

Dal Lago A. (1999). *Non Persone. L'esclusione dei migranti in una società globale*. Milano: Feltrinelli.

Fassin D. (2010). *La raison humanitaire. Une histoire morale du temps present*, Gallimard-Seuil-Hautes Études: Paris.

Gilbert L. (2009). Immigration as Local Politics: Re-Bordering Immigration and Multiculturalism through Deterrence and Incapacitation. *International Journal of Urban and Regional Research*, Vol. 33, n. 1: 26-42. DOI: 10.1111/j.1468-2427.2009.00838.x

Jaji R. (2009). *Refugee woman and the experiences of local integration in Nairobi, Kenya.* University of Bayreuth: Bayreuth.

Krause-Vilmar J. (2011). *The Living Ain't Easy, Urban Refugees in Kampala*. UN Report

Ministero dell'Interno, *Rapporto sulla protezione internazionale in Italia 2015*, in http://www.interno.gov.it/sites/default/files/t31ede-rapp_prot_int_2015_-_rapporto.pdf

Naggujja Y. et al (2014). *From The Frying Pan to the Fire: Psychosocial Challenges Faced By Vulnerable Refugee Women and Girls in Kampala*, Report of the Refugee Law Project.

Osti G. & Ventura F. a cura di (2012). *Vivere da Stranieri in Aree Fragili*. Napoli: Liguori.

Palidda S. a cura di (2011). *Il discorso ambiguo sulle migrazioni*. Messina: Mesogea.

Pinelli B. (2011). Attraversando il Mediterraneo. Il sistema campo in Italia: violenza e soggettività nelle esperienze delle donne, *Lares*, 77: 159-180.

Regolamento (CE) n. 343/2003 (Dublino II), in http://eur-lex.europa.eu/legal-content/IT/TXT/?uri=URISERV%3Al33153

Regolamento UE n. 604/2013 (Dublino III), in http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:180:0031:0059:IT:PDF

Robinson D. & Reeve K. (2006). *Neighbourhood Experiences of New Immigration. Reflections from the Evidence Base*. York: Joseph Rowntree Foundation.

Shamir R. (2005). Without borders? Notes on globalization as a mobility regime. *Sociological Theory*, Vol. 23, n. 2: 197-217. DOI: 10.1111/j.0735-2751.2005.00250.x

Tognetti M. (2016). Donne e processi migratori fra continuità e cambiamento. *ParadoXa*, X(3): 69-88.

Walters W. (2011). Foucault and Frontiers: Notes on the Birth of the Humanitarian Border. In: Bröckling U. (Ed.). *Governmentality: Current Issues and Future Challenges*. Routledge: London.

Dal corpus al dizionario: prime riflessioni lessicografiche sul Vocabolario storico della cucina italiana postunitaria (VoSCIP)

Patrizia Bertini Malgarini[1], Marco Biffi[2], Ugo Vignuzzi[3]
[1]LUMSA – p.bertini@lumsa.it
[2]Università degli Studi di Firenze – marco.biffi@unifi.it
[3]Sapienza. Università di Roma – ugo.vignuzzi@uniroma1.it

Abstract
The *Vocabolario storico della cucina italiana postunitaria* (VoSCIP) it is a historical dictionary of the language of the cooking, which has also had a considerable importance for identifying a national linguistic model after the Unity of Italy. The dictionary is based on a representative corpus (today 42 texts), but by its nature it is a work in progress, open, and it is progressively increasing. The first exemplar entries (such as *cappelletti*, *anolini*, *tagliatelle*, *bagnomaria*) had been presented in various conferences and in some articles; the entries had been based on a restricted corpus (28 texts) and they have highlighted some critical issues, so it was necessary a further methodological reflection. The aim of our paper is to propose some aspects of these investigations and this methodological reflection: a) the structure of the voice in a differentiated form ("light" and "complex"); b) the treatment of emerging positions from the statistical analysis tools of the corpus; c) the lemmatization of compound words in the face of the morphological polymorph emerging from the diachronic depth of the corpus; d) the correct balance between the examples mentioned in the voice and the possibility of a direct interrelation with the database.

Sintesi
Il *Vocabolario storico della cucina italiana postunitaria* (VoSCIP) è un dizionario storico di una lingua speciale, quella della cucina, che ha avuto una notevole importanza anche nel quadro dell'individuazione di un modello linguistico nazionale soprattutto all'indomani dell'Unità. Il dizionario si basa su un corpus rappresentativo (attualmente di 42 testi), ma che per sua natura è elastico, e aperto, e viene quindi progressivamente incrementato. Le prime voci campione (quali per esempio *cappelletti*, *anolini*, *tagliatelle*, *bagnomaria*) presentate in vari convegni e in articoli in volume e riviste, basate su un corpus ristretto a 28 testi, hanno messo in luce alcune criticità che hanno spinto a una ulteriore riflessione metodologica. Proprio alcuni aspetti di tali approfondimenti sono oggetto del contributo che proponiamo: a) la struttura

della voce in forma differenziata ("leggera" e "complessa"); b) il trattamento delle collocazioni emergenti dagli strumenti statistici di analisi del corpus; c) la lemmatizzazione di parole composte a fronte della polimorfia morfologica emergente dalla profondità diacronica del corpus; d) il corretto equilibrio tra esempi citati nella voce e possibilità di un'interrelazione diretta con la banca dati.
Keywords: lingua della cucina, lingue speciali, linguistica dei corpora, lessicografia, vocabolario, italiano, dizionario storico

1. Il VoSCIP

Il "Vocabolario storico della cucina italiana postunitaria" (VoSCIP) nasce con lo scopo di documentare il costituirsi e il fissarsi di una cultura e di una lingua unitaria della gastronomia in Italia dopo l'Unità. Si tratta di un'esigenza ben presente a tutti gli addetti ai lavori (linguisti, storici dell'alimentazione, sociologi ecc.) e che nello specifico ha preso le mosse da una precisa prospettiva di ricerca, quella di esaminare le vie e i modi dell'affermarsi di un italiano gastronomico "comune", a partire da Pellegrino Artusi e dal modello archetipico del suo fortunatissimo *La scienza in cucina e l'arte di mangiar bene*. Il progetto "L'Italiano in cucina. Per un Vocabolario storico della lingua italiana della gastronomia" è stato assunto dall'Accademia della Crusca che lo ha inserito nell'ambito degli studi che mirano alla costruzione del suo progetto strategico dedicato alla redazione di un *Vocabolario Italiano postunitario.*

Per la realizzazione del VoSCIP si è proceduto preliminarmente a fissare un corpus rappresentativo di testi, nel quale naturalmente un ruolo nodale spetta alla *Scienza in cucina*: corpus che, per motivi di fattibilità pratica, si è deciso di far arrivare alla Seconda guerra mondiale e dintorni, nell'auspicabile prospettiva di poter spostare successivamente il *terminus ad quem* alla contemporaneità (con l'inclusione, oltre che dei testi a stampa posteriori al '50, delle diverse produzioni legate al "trasmesso" nelle sue varie forme, dai ricettari presenti sul WEB, ai blog ai social media etc.). Il corpus principale di riferimento comprende al momento oltre un centinaio di volumi apparsi tra la fine del Settecento (torneremo fra poco sulle ragioni della scelta di arretrare il *terminus post quem*) e il 1950: i testi sono stati selezionati utilizzando le principali bibliografie sulla produzione gastronomica italiana del periodo considerato (preziosa in primo luogo quella di Alberto Capatti che correda l'edizione del 2010 della *Scienza artusiana* della Rizzoli). Necessariamente si è dovuto tener conto pure di fattori pratici quali in primo luogo la reperibilità delle opere e soprattutto la loro disponibilità e/o acquisibilità da parte dell'Academia Barilla, con la quale è stata a tali scopi stipulata una specifica convenzione da parte

dall'Accademia della Crusca. Al momento, i testi acquisiti informaticamente e marcati (XML/TEI) sono quaranta.
Prima di proseguire, una doverosa precisazione (già annunciata) sul *terminus post quem*: anche se il nostro obiettivo primario è, come abbiamo detto, quello di raccogliere e descrivere la lingua della tradizione gastronomica italiana postunitaria, per meglio documentare le origini di questo italiano in cucina (soprattutto per l'aspetto della fraseologia, cioè in primo luogo polirematiche e collocazioni, ma anche detti proverbiali, modi di dire, ecc.) abbiamo deciso di prendere in considerazione anche alcuni dei testi più significativi tra fine Settecento e primo Ottocento, a partire dalle due redazioni dell'*Apicio moderno* e dal *Cuoco galante* di Vincenzo Corrado. Sempre al medesimo fine, stiamo procedendo inoltre allo spoglio sistematico di tutto ciò che è pertinente all'ambito semantico del cibo nella tradizione lessicografica italiana, a partire dalle cinque impressioni del *Vocabolario degli Accademici della Crusca*, dal Tommaseo Bellini, dal Giorgini Broglio, e soprattutto dal *Dizionario moderno* (prima ed., 1905) di Alfredo Panzini. L'interesse di questo vocabolario, che offre un vero e proprio panorama della vita e della cultura italiana tra fine Ottocento e Novecento, è costituito dal nostro punto di vista proprio dallo spazio attribuito a quelle parole nuove, che già nella prima edizione lo stesso Panzini catalogava in "scientifiche, tecniche, mediche, filosofiche, [parole straniere, neologismi, parole dello sport,] della moda, del teatro, della cucina".
Imprescindibile nell'ambito lessicale del cibo (come è ben noto) è la dimensione diatopica per la quale il VoSCIP potrà utilizzare gli importanti risultati delle indagini geolinguistiche del Novecento, in primis degli atlanti linguistici: l'AIS e l'ALI, ma anche l'ASLEF, l'ALEPO, l'ALT, l'ALLI, e i preziosi materiali in corso di pubblicazione per l'ALS (tra cui si ricorderà almeno il paradigmatico volume di Ruffino 1995).
Per verificare la fattibilità del nostro progetto abbiamo realizzato alcune voci pilota: siamo partiti da *tagliatella*, cui sono seguite *agnelotto*, *cappelletto* e *anolino*; in tutt'altro ambito abbiamo recentissimamente elaborato la voce *bagnomaria*. Proprio la redazione di queste voci e in particolare dell'ultima, *bagnomaria*, ha messo in luce alcune criticità del modello di voce originariamente elaborato e reso necessario un ripensamento che sfruttasse a pieno le risorse della lessicografia *computer aided* (o della lessicografia computerizzata) e della multimedialità oggi disponibili.

2. La banca dati

I testi del corpus sono stati sottoposti a una marcatura XML/TEI leggera, mirata soprattutto a finalità lessicografiche. Attualmente sono stati acquisiti, collazionati e marcati 42 testi che coprono uniformemente l'arco cronologico

considerato. Per quanto riguarda l'*header* sono state previste le indicazioni di autore, titolo, luogo di edizione, editore, anno, tipologia testuale, indicazione diamesica, in modo che possano costituire la base per filtrare sottocorpora specifici. All'interno del testo sono state marcate le pagine di ogni volume (così che le trascrizioni possano essere di volta in volta collegate alla riproduzione in facsimile dell'originale), le eventuali figure, le parti in lingue diverse dall'italiano (perché possano essere escluse dall'interrogazione del lessicografo). Non si è ritenuto di prevedere nessuna marcatura per i forestierismi, che, al pari degli altri lessemi, devono essere analizzati opportunamente dal lessicografo in ogni loro contesto. In una seconda fase della marcatura dei primi 42 testi, in via di attuazione, è prevista anche la marcatura del testo delle singole ricette e del loro titolo. Lo scopo primario di questa marcatura è quello di ottenere una lista aperta delle ricette presenti nel corpus, che possano eventualmente essere messe a confronto tra di loro con appositi algoritmi legati alle forme presenti nel titolo. In questo modo sarà possibile individuare una linea diacronica delle singole ricette e seguire l'evoluzione della lingua in esse contenute. Per quanto concerne il trattamento informatico va tenuto conto che la banca dati è un esempio di testualità ibrida: sia in relazione all'acquisizione filologica del testo e alla sua interrogabilità, sia per quanto riguarda la possibilità di applicazione di procedure di lemmatizzazione automatica. Trattandosi di testi otto-novecenteschi la possibilità di buoni risultati nell'applicazione degli strumenti informatico-linguistici realizzati nel panorama nazionale e internazionale scema progressivamente allontanandosi dalla contemporaneità verso il 1861, ma anche per i testi ottocenteschi e primo novecenteschi si hanno garanzie sufficienti. Vista la particolare natura della banca dati, la sua cronologia e la sua finalità lessicografica, nell'equilibrio della gestione delle risorse, si è preferito quindi non investire su una lemmatizzazione controllata, che avrebbe comportato l'inserimento di correttivi legati alla lingua ottocentesca e primo-novecentesca sia sui dizionari macchina che sulle morfologie macchine attualmente in circolazione (prevalentemente di base anglofona, con tutti i limiti che questo comporta, e, anche nel migliore dei casi tarati per l'italiano scritto recente; cfr. Biffi 2016). La banca dati (attualmente in fase di *testing* nella sua versione beta) è quindi consultabile con un motore di ricerca per forme, potenziato da strumenti (caratteri jolly, ricerca *fuzzy*) che facilitino l'individuazione delle varianti formali, morfologiche e grafico-fonetiche, e da una lemmatizzazione automatica basata sulle morfologie macchina attualmente esistenti (e quindi tarate sull'italiano scritto contemporaneo, ma comunque sufficientemente funzionali per il reperimento delle forme varianti di testi otto-novecenteschi, soprattutto se a fini lessicografici). La piattaforma di interrogazione prevede

specifiche funzioni di ricerca a distanza e collocazioni, e la possibilità di accedere a dati statistici, sia in versione tabellare, sia in versione *heatmap* e *tag cloud*. Con queste caratteristiche la banca dati può peraltro essere del tutto omogenea a quelle che gravitano intorno al progetto del *Corpus di riferimento per un nuovo vocabolario dell'italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni*, finanziato su fondi PRIN 2012 e coordinato da Claudio Marazzini, offrendo così ampi margini di dialogo con gli strumenti lessicografici a essa collegati.

3. Struttura delle voci e dizionario elettronico

La struttura della voce progettata risente naturalmente delle caratteristiche dei dizionari storici. Ecco la sua architettura:

LEMMA + categoria grammaticale

0.1. Forme attestate nel corpus dei testi (con tutte le varianti)

La forma lemmatizzata per la voce principale è quella più diffusa nell'uso odierno: ci si serve del GRADIT, *Grande dizionario italiano dell'uso*, di Tullio De Mauro, con i relativi aggiornamenti.

0.2. Nota etimologica essenziale.

0.3. Prima attestazione nel corpus.

0.3.1. Indicazione numerica della frequenza (per ciascuna forma; nell'indicazione delle occorrenze, la seconda cifra, preceduta dal segno +, si riferisce alle forme presenti in eventuali indici).

0.4. Distribuzione geografica delle varianti.

Per ora si forniscono i dati relativi ai soli AIS e ALI. Aggiungiamo in nota il riscontro con le forme registrate da Touring Club Italiano 1931.

0.5. Note linguistiche/merceologiche (forestierismi; italianismi in altre lingue).

La bibliografia per ora si riferisce solo alle 'Note linguistiche', e, per quanto riguarda gli italianismi in altre lingue, al DIFIT (consultabile in versione elettronica in http://www.italianismi.org/difit-elettronico).

0.6. Riepilogo dei significati.

0.7. Locuzioni polirematiche e vere proprie (con la prima attestazione nel corpus).

0.8. Rinvii (sono previsti soprattutto 'iperlemmi', o, se si preferisce voci 'generali', di raccordo).

0.9. Corrispondenze lessicografiche (= riscontri nei dizionari e

nei corpora lessicografici in rete): si distinguono i vocabolari etimologici (compreso il LEI) da quelli descrittivi (in ordine cronologico, a partire dal Tommaseo-Bellini).

1. Prima definizione
   Contesti
   1.1. Definizione subordinata
   Contesti
   1.2. Definizione subordinata
   Contesti
[...]
2. Seconda definizione
   Contesti
[...].

La voce richiama, con gli opportuni adattamenti, quella del *TLIO Tesoro della Lingua Italiana delle origini*, dell'Istituto dell'Opera del Vocabolario Italiano del CNR di Firenze. I primi esperimenti, sui quali è basata ad esempio l'ultima voce campione relativa a *bagnomaria* (a partire da una versione iniziale del corpus, limitata a 28 testi), hanno evidenziato che la struttura rischia però di essere troppo pesante in vista di una effettiva fattibilità realizzativa del progetto.
I limiti "dimensionali" emergenti (che bene risultano evidenti in Bertini Malgarini e Vignuzzi 2017) sono legati soprattutto alla ricchezza degli esempi e all'ampiezza delle citazioni da altri strumenti lessicografici.
A entrambi questi limiti si pensa però di provvedere aumentando l'interazione con gli altri strumenti collegati e collegabili.
In primo luogo prevedendo una profonda interazione tra banca dati testuale e dizionario sia nella fase di redazione della scheda che in quella di pubblicazione. In questo modo sarà possibile limitare il numero di esempi citati per poi rimandare a un dossier completo delle occorrenze mediante il collegamento con il corpus informatizzato. Nell'ottica di creare un accesso aperto alla banca dati dei testi è opportuno porsi il problema dell'utilizzo pubblico di testi coperti da diritto d'autore. Il tema è già stato affrontato all'interno del gruppo PRIN 2008 "Il portale della TV, la TV dei portali" e in occasione del convegno conclusivo del progetto Marina Pietrangelo – ricercatrice dell'ITTIG (Istituto di Teoria e Tecniche dell'Informazione Giuridica) appositamente invitata a parlare sul tema *Per un uso legale degli audiovisivi in corpora di ricerca* – ha risposto con un sostanziale via libera previsto dalla norma nel caso di progetti con esclusiva finalità di ricerca e senza nessun risvolto economico (Pietrangelo 2017). Anche i riferimenti agli

altri dizionari vanno poi realizzati attraverso collegamenti con le versioni elettroniche in rete attualmente disponibili (ad esempio quella del Tommaseo-Bellini: *Tommaseo online*; quella delle edizioni del *Vocabolario degli Accademici della Crusca*: *Lessicografia della Crusca in rete*; e infine quella del vocabolario postunitario che si sta realizzando all'interno del progetto PRIN 2015 "*Vocabolario dinamico dell'italiano post-unitario*", coordinato da Claudio Marazzini). Sono tuttora allo studio procedure per il trattamento delle collocazioni emergenti dagli strumenti statistici di analisi del corpus, e per la lemmatizzazione di parole composte a fronte della polimorfia morfologica emergente dalla profondità diacronica del corpus. All'interno di una vera e propria stazione lessicografica tutti questi strumenti saranno integrati all'interno di un sistema di *back-office* che, tramite fasi di valutazione progressiva e di controllo, porterà alla diretta pubblicazione della voce in rete. Infine, proprio la potenziale interazione/integrazione con il citato futuro "*Vocabolario dinamico dell'italiano post-unitario*" ha suggerito al gruppo di ricerca di predisporre una scheda lessicografica variabile: alla scheda approfondita del dizionario storico si affiancheranno infatti una scheda strutturata secondo le specifiche di un dizionario sincronico per quelle voci che facciano ancora oggi parte dell'italiano dell'uso, e strumenti di calibrazione dei campi che l'utente esperto e non esperto potrà gestire in modo da avere di volta in volta una voce personalizzata.
In sede di discussione sarà presentata e discussa una voce "esemplare" del VoSCIP, anche in relazione alla selezione e all'organizzazione del materiale lessicografico e alla sua pubblicazione (in rete e in forma cartacea).

Riferimenti bibliografici

Bertini Malgarini, P. e Vignuzzi, U. (2017). *Bagnomaria nel Vocabolario storico della cucina italiana postunitaria (VoSCIP)*: < http://permariag.wixsite.com/permariagrossmann/vignuzzi>.

Biffi, M. (2016). *Progettare il corpus per il vocabolario postunitario*, in Marazzini, C. e Maconi, L. (a cura di), *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*. Accademia della Crusca, pp. 259-80.

Pietrangelo, M. (2016). *Per un uso legale degli audiovisivi in corpora di ricerca*, in Alfieri, G., Biffi, M. et alii (a cura di), *Il portale della TV. La tv dei portali*. Bonanno, pp. 171-185.

Ruffino, G. (1995). *I pani di Pasqua in Sicilia. Un saggio di geografia linguistica e etnografica*. Centro di Studi Filologici e Linguistici Siciliani.

Touring Club Italiano (1931). *Guida gastronomica d'Italia*. Touring Club Italiano [rist. anast. 2003].

Strumenti

AIS = Jaberg, K. e Jud, J. (1928-1940). *Sprach- und Sachatlas Italiens und der Südschweiz*. Ringier, 8 voll. (trad. it. 1987. *AIS. Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale*, Unicopli). Anche in rete: *NavigAIS*, <http://www3.pd.istc.cnr.it/navigais/>.

ALEPO = Telmon, T. e Canobbio, S. (1984-). *Atlante linguistico ed etnografico del Piemonte occidentale* (vedi <http://www.alepo.eu/>)

ALI = Bartoli, M. G et alii (1995-). *Atlante Linguistico Italiano*. Istituto Poligrafico e Zecca dello Stato.

ALLI = Moretti, G. et alii (1982-). *Atlante Linguistico dei Laghi Italiani* (vedi <http://www.lettere.unipg.it/ricerca/centri>

ALS = Ruffino, G. (1995-). *Atlante Linguistico della Sicilia* (vedi <http://atlantelinguisticosicilia.it/>).

ALT = Giacomelli, G. (2000). *Atlante Lessicale Toscano*. LEXIS (in CD-ROM); Ora in rete come ALT-WEB: <http://serverdbt.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm>.

ASLEF = Pellegrini, G. B. et alii (1972-). *Atlante Storico-Linguistico-Etnografico Friulano*. Istituto di glottologia e fonetica dell'Università Istituto di filologia romanza della Facoltà di lingue e letterature straniere dell'Università.

DIFIT = Stammerjohann. H. (2008). *Dizionario di italianismi in francese, inglese e tedesco*. Accademia della Crusca. Anche in rete: <http://www.italianismi.org>.

GRADIT = De Mauro, T. (2007). *Grande Dizionario Italiano dell'Uso*. UTET.

LEI = Pfister, M. e Schweickard, W. (1979-). *Lessico Etimologico Italiano*, Edito per incarico della Commissione per la Filologia romanza. Reichert.

Lessicografia della Crusca in rete = Accademia della Crusca (2004). *Lessicografia*

della Crusca in rete. <http://www.lessicografia.it>.
TLIO = Opera del Vocabolario Italiano (1997-). *Tesoro della lingua italiana delle origini.* <http://www.vocabolario.org/>.
Tommaseo-Bellini = Tommaseo, N. e Bellini V. (1861-1879). *Dizionario della lingua italiana,* Società L'Unione Tipografico-Editrice.
Tommaseo online = Accademia della Crusca (2015). *Tommaseo online.* <http://www.tommaseobellini.it>.

Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano postunitario

Marco Biffi
Università degli Studi di Firenze – marco.biffi@unifi.it

Abstract
The paper focuses on some general problems about representative corpora for the compilation of dictionaries. It starts from the concrete case of the *Vocabolario dell'italiano post-unitario*, which, due to its hybrid nature, offers a complete view of both the criticalities of synchronic lexicography and of the historical one. Therefore is introduced the concept of *Banca linguistica*, that is a platform in which different types of corpora, a search *meta*-engine of the existing databases, and tools of access to existing electronic dictionaries converge. A final paragraph is dedicated to the concept of "quantum relativity" of data of computational linguistics.

Sintesi
Il contributo mette a fuoco alcuni problemi generali relativi alla costituzione di corpora rappresentativi per la redazione di dizionari, partendo dal caso concreto del *Vocabolario dell'italiano post-unitario*, che, per la sua natura ibrida, offre un quadro completo sia delle criticità della lessicografia sincronica sia di quella storica. Si introduce pertanto il concetto di *Banca linguistica* in cui convergono diverse tipologie di corpora, un *metamotore* di ricerca per la consultazione delle banche dati esistenti e sistemi di integrazione con i dizionari elettronici esistenti. Infine ci si sofferma sul concetto di "relatività quantistica" dei dati estrapolabili dalle ricerche informatico-linguistiche.
Keywords: Linguistica dei corpora, Italiano, Dizionario sincronico, Dizionario storico, Testo elettronico, Bilanciamento, Metamotore, Banca linguistica, Relatività quantistica, Informatica linguistica, Linguistica computazionale

1. Introduzione

In questo contributo cercherò di mettere a fuoco alcuni problemi generali relativi alla costituzione di strumenti per la redazione di dizionari partendo da un caso specifico, quello del progetto di un dizionario "ibrido", insieme storico e sincronico, su cui sta lavorando un gruppo di ricerca nazionale coordinato da Claudio Marazzini. Il progetto – che ha come obiettivo finale la redazione di un vocabolario dell'italiano post-unitario che raccolga il patrimonio linguistico nazionale della lingua ufficiale dello Stato dal 1861 a

oggi – ha visto l'avvio con una prima fase finanziata sul PRIN 2012 *Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni*; e ha poi potuto continuare con un secondo finanziamento sul PRIN 2015 *Vocabolario dinamico dell'italiano post-unitario*. Ai due progetti hanno partecipato numerose università italiane: Piemonte Orientale, Milano, Genova, Firenze, Viterbo, Napoli, Catania (al progetto sul corpus ha partecipato anche l'Istituto di Teorie e Tecniche dell'Informatica Giuridica ITTIG del CNR di Firenze; al progetto sul vocabolario dinamico partecipa anche l'Università degli Studi di Torino); come partner esterno ha collaborato l'Accademia della Crusca, per la quale il dizionario post-unitario è uno dei tre progetti strategici attuali, accanto al *Vocabolario dantesco* e all'*Osservatorio degli italianismi nel Mondo* (*OIM*).
Per quanto le dinamiche di impiego di corpora per la redazione di dizionari storici siano note, soprattutto dopo l'esperienza del *TLIO Tesoro della lingua italiana delle origini* dell'Istituto dell'Opera del Vocabolario Italiano del CNR di Firenze, meno si è riflettuto sulle implicazioni pratiche della costituzione di un dizionario sincronico basato su un corpus rappresentativo, e del tutto nuovo è il caso di uno strumento ibrido come il vocabolario post-unitario, in cui le criticità della lessicografia informatica storica e sincronica si mescolano, evidenziando come si debba piuttosto muoversi nella direzione di strumenti articolati.

2. Criticità di fisionomia di un corpus rappresentativo dell'italiano post-unitario

Un primo problema da affrontare per un corpus rappresentativo per un dizionario è la sua dimensione. Se proviamo a effettuare un rapido controllo sulla situazione dei corpora di riferimento per altre lingue europee (in particolare inglese e tedesco, che hanno avuto una maggiore attenzione a questo tema), sia il *British National Corpus* (per il 10% costituito da trascrizioni dell'inglese parlato – cfr. Cresti-Panunzi 2013: 36-37) che il *DWDS-Kerncorpus* (testi del XX secolo di cinque tipologie: letteratura, 25%; giornali, 25%; prosa scientifica, 20%; guide, libri di ricette e testi analoghi, 20%; lingua parlata trascritta, 10% – cfr. Klein 2013: 18-19) hanno dimensione pari a circa 100 milioni di parole. Questa era la dimensione che nel primo decennio del secolo individuava corpora di dimensioni standard (cfr. Chiari 2007: 45; secondo la tabella ivi riportata); anzi, 100 milioni di parole era la soglia che divideva i corpora standard da quelli di grandi dimensioni. Tenendo conto dei progressi informatici e metodologici degli ultimi anni, certamente è opportuno introdurre qualche correttivo; e in effetti sia per l'inglese che per il tedesco questi correttivi esistono, perché i corpora bilanciati sono affiancati da *thesauri*. Al BNC è stata recentemente affiancata la *Bank of English* (un

monitor corpus, secondo la terminologia di Sinclair, di testi completi per un totale di 650 milioni di parole – cfr. Cresti-Panunzi 2013: 36-37); al *Kerncorpus* si sono aggiunti alcuni moderni corpora di giornali (successivi al 1995) e altre raccolte più piccole di testi, per un totale di 2,6 miliardi di parole (e, anche sul piano diacronico, si sta cercando di completare il quadro con il *Deutsche Textarchiv* in allestimento dal 2005 e ormai in via di completamento, che raccoglie 1500 libri accuratamente scelti, di solito prime edizioni e volumi di giornali, nell'arco cronologico compreso fra il 1650 e il 1900 – cfr. Klein 2013: 18-19). Per quanto riguarda la raccolta di testi si è già sottolineata l'importanza di quella che è stata definita "parabola dimensionale dei corpora" (Biffi 2016: 262).

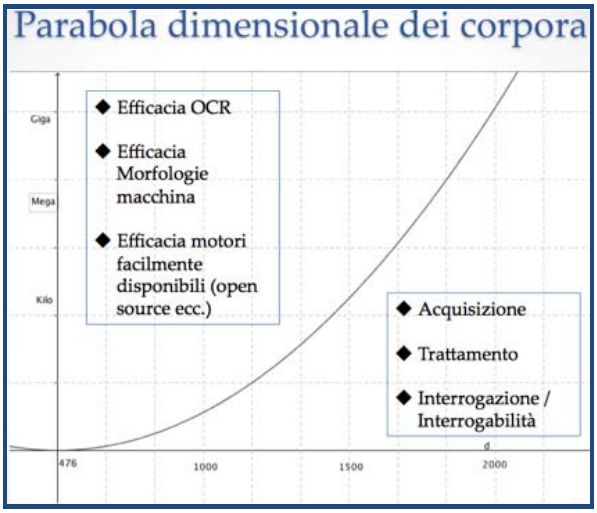


Figura 1

La rappresentazione geometrica analitica di questa parabola evidenzia il rapporto tra la lingua nei secoli (nella fattispecie l'italiano) e la possibilità di rappresentarla con un corpus dell'ordine di grandezza di 100.000 parole (*kiloparole*), di milioni di parole (*megaparole*), di miliardi di parole (*gigaparole*). La possibilità di costruire corpora di grandi dimensioni diminuisce tanto maggiormente quanto più si va indietro nel tempo, mentre aumenta vertiginosamente per la lingua dei nostri giorni, con dimensioni ormai veramente molto elevate, che non corrispondono certamente a tutto ciò che si produce in una certa lingua, perché questo è ovviamente impossibile, ma che tendenzialmente vi si avvicinano molto. La ridotta dimensione dei corpora dell'italiano del passato – questo sottolinea la curva – non è soltanto legata al fatto, oggettivo, che per il passato disponiamo di un minor numero di testi,

ma, in modo determinante, al fatto che molto più difficilmente riusciamo a riunire i testi del passato in formato elettronico per poterli interrogare con efficacia. Le difficoltà sono legate ai limiti di tutti gli strumenti informatici coinvolti nella realizzazione di corpora informatici, che paradossalmente convergono nel determinare l'andamento di questa curva: l'efficacia dell'OCR (il riconoscimento ottico, automatico, dei caratteri), l'efficacia delle morfologie macchina per la lemmatizzazione, l'efficacia dei motori di ricerca disponibili con facilità e a costo poco elevato; quindi toccano i processi che coinvolgono sia l'acquisizione dei testi, sia il loro trattamento, sia la loro interrogazione e interrogabilità (Biffi 2016: 263-267). Per il passato gli effetti della parabola rendono gestibile il problema di una reale rappresentatività del corpus di riferimento. In effetti il TLIO, che si muove in un arco cronologico che va dalle origini al 1375, può disporre come base di partenza di un corpus che riunisce una raccolta consistente di testi volgari del periodo considerato, spaziando a tutto tondo sull'asse diatopico e diafasico (e quindi garantendo una grande rappresentatività anche in diastratia). Ha fondamenta molto solide anche a fronte di dimensioni che, sulla scala di misurazione dei corpora, non sono particolarmente elevate. Le ridotte dimensioni hanno consentito infatti di abbattere gli effetti "parabolici" dell'efficacia dell'acquisizione e del trattamento del testo elettronico (i testi, ricavati dalle principali edizioni critiche hanno potuto essere sottoposti a un'attenta collazione), così come dell'efficacia delle morfologie macchina (il corpus è stato lemmatizzato di fatto manualmente con l'ausilio di procedure semiautomatiche). La possibilità di progettare e realizzare un motore per lemmi e un motore per forme personalizzato ha poi definitivamente abbattuto i problemi di interrogazione/interrogabilità. Ma è evidente che anche salendo di poco nella cronologia, proprio per l'effetto "parabolico", i problemi aumentano vertiginosamente. Per quanto riguarda le morfologie macchina, ad esempio, sarebbe opportuno ricalibrarle in base alle variazioni diacroniche delle strutture morfologiche e morfosintattiche, seguendo l'asse del tempo (ed esperimenti si stanno facendo: ad esempio per la morfologia della lingua di Leonardo in un progetto finanziato dalla Biblioteca Leonardiana di Vinci e da me curato per la parte linguistica); ma il processo è lungo e non è mai stato affrontato in modo sistematico, né metodologicamente né pragmaticamente. Questo perché, ma vale per tutti gli aspetti della linguistica computazionale e più in generale di quella che preferisco chiamare linguistica informatica, la tendenza generale è quella di lavorare per piccole monadi e non creare sistema mettendo in sinergia le competenze e gli strumenti in modo da ampliare e affinare le tecnologie disponibili rendendole sempre più potenti. Così oggi disponiamo di vari strumenti, in parte sovrapponibili, in parte complementari, ma nulla di

realmente condivisibile da migliorare con un sistema *open source*, in modo da concentrare gli sforzi su ciò che realmente manca e o è debole. Il "pezzo" delle morfologie macchina è particolarmente significativo: costruire un corpus diacronico per un dizionario storico significa infatti fornire i primi mattoni per ricalibrare le morfologie macchine esistenti tarandole sul periodo preso in considerazione; ma in nessun caso si è pensato di usare questi corpora del passato come punto di partenza per migliorare le procedure di lemmatizzazione che a loro volta potenzierebbero le possibilità lessicografiche in un circolo virtuoso destinato a raffinare gli strumenti a disposizione della comunità scientifica. Per tornare alle specificità del dizionario dell'italiano post-unitario, il suo carattere ibrido lo colloca in una posizione particolarmente delicata perché in quanto diacronico, dal 1861 al 2000, risente dei limiti informatici di cui abbiamo parlato (anche se, ad esempio, in questo segmento cronologico le procedure di riconoscimento automatico dei caratteri danno ottimi risultati). Ma diventa decisamente sincronico nel periodo 2000-2014, quando abbiamo la possibilità di creare un enorme corpus massivo (delle dimensioni delle *gigaparole*), anche con facilità, semplicemente attingendo dal web mediante programmi di *data crawling* (*web crawler*, o *spider*), come dimostra molto bene il caso di RIDIRE (www.ridire.it, diretto da Emanuela Cresti), un corpus di 1,3 miliardi di parole, realizzato con un *crowler* controllato che ha permesso un "bilanciamento" basato su domini semantici (architettura e design, arti figurative, cinema, cucina, letteratura e teatro, moda, musica, religione, sport) e domini funzionali (amministrazione e legislazione, economia e affari, informazione).

3. Dal corpus rappresentativo alla "Banca linguistica"

Da un punto di vista teorico la scelta migliore per il corpus di riferimento del dizionario dell'italiano post-unitario sarebbe quella di un corpus bilanciato nell'ordine di *megaparole* dal 1861 al 2014, da affiancare con un corpus massivo della dimensione delle *gigaparole* sul 2000-2014, un risultato, come si è visto, ormai realizzabile. Però il gruppo di ricerca è partito da una situazione pregressa di progetti già realizzati e studi già avviati con validi risultati raggiunti, per cui si è scelto di mettere a frutto al massimo le esperienze dei componenti del gruppo, recuperando tutti i materiali che ciascuno poteva portare in dote al progetto per poi ampliarli e consolidarli con competenze specifiche. La copertura quindi è "a macchia di leopardo", ed è pertanto necessario utilizzare al massimo, anche per la zona cronologica che va dal 1861 al 2000, un approccio massivo, che conduce inevitabilmente sulla strada della "banca linguistica", del *thesaurus*, dal quale poi estrarre un corpus bilanciato (o più di uno, in modo dinamico anche in relazione alle esigenze del redattore della voce assegnata).

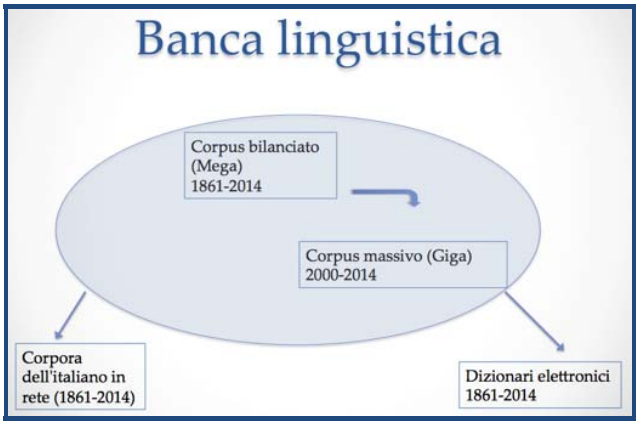


Figura 2

La "Banca linguistica" può essere una piattaforma in cui siano disponibili vari sub-corpora, in cui siano raccolti tutti i materiali con una marcatura semantica che consenta successivi bilanciamenti, con un "corpus centrale" che sarà la base primaria del lavoro del lessicografo del vocabolario postunitario, ma che andrà continuamente tarato grazie ai dati emergenti dalla consultazione del corpus massivo contemporaneo e dei sub-corpora diacronici presenti. La piattaforma dovrà anche dialogare con i dizionari elettronici di cui disponiamo dal 1861 a oggi: il *Tommaseo Online* e la versione elettronica della quinta edizione del Vocabolario degli Accademici della Crusca presente nella *Lessicografia della Crusca in rete* per la parte diacronica (nella speranza che l'accordo siglato nel settembre 2017 tra UTET e Accademia della Crusca per la digitalizzazione del *Grande Dizionario della Lingua Italiana* maturi frutti rapidi); le versioni dei dizionari sincronici presenti in rete (il *Sabatini Coletti*, il *De Mauro*, il *Treccani*, e tutto quanto sarà disponibile); tutti i corpora dell'italiano presenti sul web, inclusi quelli, preziosissimi, degli archivi elettronici delle principali testate giornalistiche nazionali (Biffi 2016: 272-273). Non va dimenticato infatti che il panorama dei corpora dell'italiano è abbastanza ampio (per un quadro generale si veda Cresti-Panunzi 2013; ma è necessario perfezionare il censimento). Però è mancata, come del resto è naturale, una politica organica di costruzione di un sistema: abbiamo quindi un'estrema eterogeneità di strumenti, piattaforme, codifiche (per fortuna in anni recenti, almeno per quest'ultimo aspetto, la forza centrifuga si sta progressivamente contenendo con il ricorso sempre più frequente, se non totale, alla codifica XML/TEI), che costringe il ricercatore a collegarsi *n* volte, su *n* piattaforme, con *n* filosofie diverse, con *n* motori diversi, per poter effettuare una ricerca a tutto campo. Diventa quindi

fondamentale un *metamotore*. Una versione beta di metamotore dei corpora dell'italiano è stata realizzata dall'unità di ricerca dell'Università degli Studi di Firenze del gruppo PRIN 2012 da me diretta (www.metaricerche.it). Come si legge nella sezione del portale intitolata "Il metamotore": «Gli strumenti individuati sono stati classificati secondo i possibili livelli di integrazione: corpora liberamente consultabili; corpora liberamente consultabili previa registrazione; corpora da scaricare. È stato poi predisposto uno studio di fattibilità per la definizione di una serie di procedure atte ad analizzare gli strumenti di partenza, determinare il livello di integrabilità (che passa anche dalla possibilità di poter interagire con lo staff tecnico della singola banca dati, a seguito di un accordo "strategico" sulla condivisione dei contenuti) e individuare delle procedure da seguire a seconda del livello. Si è passati poi a definire l'architettura del sistema, la tecnologia di riferimento e l'interfaccia di consultazione, almeno per una prima versione prototipale della piattaforma». La versione beta prevede l'integrazione di 8 banche dati, scelte come campioni delle principali tipologie di livelli di integrazione:

- Livello massimo (si è trovato accordo con lo staff tecnico che gestisce la banca dati): *LIR* (*Lessico dell'Italiano Radiofonico*), *LIS* (*Lessico dell'Italiano Scritto*) e *LIT* (*Lessico Italiano Televisivo*), Accademia della Crusca.
- Livello base (si è integrata la banca dati in una finestra, in attesa di una maggiore interoperabilità): *MIDIA* (*Morfologia dell'Italiano in DIAcronia*) Università Roma Tre; *CorDIC* (*Corpora Didattici Italiani di Confronto*) Laboratorio Linguistico Italiano Università degli Studi di Firenze.
- Livello minimo (si è integrata la banca dati in una finestra senza possibilità di maggiore interoperabilità): *Archivio* dei quotidiani «Corriere della Sera» e «La Repubblica».

Se questo strumento potrà essere potenziato fino a riunire nella lista dei risultati tutte le banche dati testuali disponibili attualmente per l'italiano, nella "Banca linguistica" del redattore del *Vocabolario post-unitario* sarà disponibile un accesso centralizzato a tutti i corpora esistenti, da integrare, modulare e bilanciare con il corpus riunito dal gruppo di ricerca PRIN, con il corpus massivo dell'italiano contemporaneo, con gli strumenti lessicografici elettronici. Rimangono da considerare alcune criticità che, se rimosse, consentirebbero un ulteriore potenziamento della "Banca linguistica", e che possiamo richiamare in questa sede solo brevemente per punti. a) La gran parte dei testi (ad esempio quelli letterari recenti) sfuggono alla possibilità di essere organizzati in corpora interrogabili per le difficoltà legate ai diritti d'autore. b) Le raccolte di corpora in diacronia, tranne rare eccezioni (come ad esempio il *CEOD*, *Corpus Epistolare Ottocentesco Digitale*) prediligono la

tradizione letteraria di registro alto. Esistono già campioni rappresentativi di italiano post-unitario, come il *DIACORIS* (25 milioni di occorrenze), ma si devono ancora integrare i vuoti legati alle lingue speciali (come è stato tentato di fare all'interno del progetto PRIN 2012). c) Resta da indagare quanto dal web si possano recuperare (in modo più o meno automatico) materiali per le sezioni in diacronia, grazie soprattutto alla presenza massiccia di testi ottocenteschi riuniti in biblioteche digitali come *Google libri* e *Archive*.

4. Informatica linguistica e relatività quantistica

Se il punto di partenza per la redazione di un dizionario non è più un corpus di riferimento omogeneo predisposto allo scopo, ma una "banca linguistica" in cui si è chiamati a gestire materiali non omogenei ed esogeni, non è inutile richiamare in questo paragrafo finale l'importanza dei risvolti "quantistici" della linguistica informatica (Biffi 2017: 545-549).

Consultando banche dati (includendo in questa categoria non solo i corpora ma anche le edizioni elettroniche dei dizionari) non è difficile imbattersi in diffrazioni nei risultati quantitativi (e quindi in quelli qualitativi, nella misura in cui possono determinarsi lacune nella ricerca di determinati contesti), che sicuramente in parte si spiegano con errori umani inseriti nelle varie fasi realizzative delle banche dati (dovuti ai moderni copisti digitali, ai programmatori, al progetto), ma anche per il concorso di fattori precisi e individuabili. Nel contributo citato (Biffi 2017) le diffrazioni riguardano i risultati relativi al numero dei lemmi nelle tre versioni elettroniche del *Vocabolario degli Accademici della Crusca* del 1612, e sono da ricondurre a diversità di tokenizzazione, diversità di approccio nella restituzione alle voci dell'intrinseca struttura di base di dati, diverse priorità nella restituzione del testo elettronico. In altre banche dati i fattori di diffrazione saranno probabilmente da ricondurre ad altro, ma si dovrà sempre tener conto delle caratteristiche e dell'architettura della banca dati così come degli strumenti di ricerca a essa applicati. Come nelle scienze esatte da Heisenberg in poi si deve tener conto dell'indeterminazione introdotta dallo strumento di misurazione, consultando le banche dati sarà opportuno ricordare che le caratteristiche dello strumento di conoscenza (in questo caso la banca dati) perturbano il risultato della ricerca costringendoci a un'inevitabile approssimazione "quantistica"; una perturbazione però dominabile, giacché si possono ricostruire le cause di diffrazione e quindi correggere il risultato finale, come avviene con la meccanica quantistica laddove è necessario sostituirla alla meccanica classica.

E allora, per poter ottenere risultati scientifici consultando una banca dati, è necessario conoscere a fondo le caratteristiche dello strumento, e tenere conto

della sua variabilità "quantistica" nel momento in cui leggiamo i dati. E, quando si leggono e gestiscono i risultati, è necessario non solo essere consapevoli di quale strumento si è usato, ma anche delle specifiche modalità di ricerca applicate; in altre parole si deve tener conto continuamente del contesto filologico della ricerca informatica, esattamente come, quando si consulta l'edizione critica di un testo, si tiene conto anche delle varianti dell'apparato.

Riferimenti bibliografici

Biffi, M. (2016). *Progettare il corpus per il vocabolario postunitario*, in Marazzini, C. e Maconi, L. (a cura di), *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*. Accademia della Crusca, pp. 259-80.

Biffi, M. (2018). *Tra fiorentino aureo e fiorentino cinquecentesco. Per uno studio della lingua dei lessicografi*, in Belloni, G. e Trovato, P. (a cura di), *La Crusca e i testi. Lessicografia, tecniche editoriali e collezionismo librario intorno al* Vocabolario *del 1612*. libreriauniversitaria.it, pp. 543-560.

Chiari, I. (2007). *Introduzione alla linguistica computazionale*, Laterza.

Cresti, E. e Panunzi, A. (2013). *Introduzione ai corpora dell'italiano*, Il Mulino.

Comparaison de corpus de langue « naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues

Annick Farina, Riccardo Billero
Università degli Studi di Firenze – annickfarina@unifi.it; riccardo.billlero@gmail.com

Abstract
The aim of this paper is to describe the work done to exploit the LBC database for the purpose of translation analysis as a resource to edit the bilingual lexical sections of our dictionaries of Cultural Heritage (in nine languages). This database, made up of nine corresponding corpora, contains texts whose subject is cultural heritage, ranging from technical texts on art history to books on art appreciation, such as tour guides, and travel books highlighting Italian art and culture. We will illustrate the different questions with the SketchEngine LBC French corpus, made up at the moment of 3,000,000 words. Our particular interest here is in research that not only orients lexical choices for translators but that also precedes the selection of bilingual quotations (from our Italian/French parallel corpus) and that we rely on for editing an optional element of the file called "translation notes." We will rely on this as much for works on "universals of translation" already described by Baker (1993) as for studies aimed at improving Translation Quality Assessment (TQA). We will show how a targeted consultation of different corpora and sub-corpora that the database allows us to distinguish ("natural language" vs "translation", "technical texts" vs "popularization texts" or "literary texts") can help us identify approximations or translation errors, so as to build quality comparative lexicographical information.
Keywords: electronic lexicography, multilingual lexical resources, corpus linguistics

Résumé
Cet article a pour but de décrire notre travail sur la base de données LBC pour ce qui concerne l'analyse de traductions comme ressources pour la rédaction de la partie bilingue de nos dictionnaires du Patrimoine (dans les neuf langues du projet). La base de données contient des corpus distincts de neuf langues composés de textes qui sont tous reliés au patrimoine italien : des textes techniques des différents domaines artistiques, des ouvrages de critique d'art ou d'histoire de l'art, des guides touristiques, des récits de

voyages, etc. Nous illustrerons différentes interrogations du corpus français (actuellement composé d'environ 3 millions de mots) dans SketchEngine. En particulier, nous nous intéresserons à des recherches qui nous guident non seulement vers la sélection de traduisants pour certains termes mais qui précèdent aussi la sélection de citations bilingues (extraites de notre futur corpus parallèle italien/français) et sur lesquelles nous nous appuyons pour la rédaction d'un élément facultatif de la fiche appelé « notes de traduction ». Nous nous appuyons pour ce faire tant sur les travaux sur les « universaux de traduction » (Baker 1993) que sur études qui visent à l'amélioration de la qualité des traduction (TQA : Translation Quality Assessment). Nous montrerons comment une consultation ciblée des différents corpus et sous-corpus que la base nous permet de distinguer (textes en « langue naturelle » vs « en traduction », « textes techniques » vs « de vulgarisation » vs « littéraires ») peut nous aider à repérer des approximations ou des erreurs de traduction, nous aidant à construire une information lexicographique comparative de qualité.
Keywords: lexicographie, ressources lexicales plurilingues, corpus linguistiques.

1. Introduction

Un des principaux buts du projet Lessico dei Beni Culturali est de constituer des dictionnaires monolingues de neuf langues différentes en fonction d'un usage précis relié à un objet particulier : la description (et traduction de descriptions) du patrimoine toscan principalement dans des textes de vulgarisation (guides touristiques, sites de musées, etc.). Pour ce faire, nous avons constitué des bases de données textuelles, que nous complétons sous la forme d'un *Work in progress*, qui nous serviront pour différentes tâches, de la création de nomenclatures à la rédaction de fiches lexicographiques/terminologiques monolingues et de fiches de traduction reliant les nomenclatures des différentes langues entre elles (pour la description de ces bases cfr. Billero *et al.* 2017). C'est l'utilisation de ces bases de données textuelles pour la rédaction de fiches bilingues de traduction que nous illustrerons ici[1], en nous basant sur l'analyse de différentes interrogations sur SketchEngine (principalement statistiques et de contexte) de notre corpus LBC français, composé actuellement d'environ trois millions

[1] Pour l'utilisation de nos bases pour la réalisation des dictionnaires monolingues, voir l'article de Nicolás et Lanini dans ce volume. Nous constituons en effet actuellement les nomenclatures des différentes langues en suivant le modèle qu'elles ont défini pour l'italien. Le lien bilingue entre ces différentes nomenclatures ne sera possible que lorsque nous aurons constitué nos bases de données parallèles.

de mots. Nous comparerons en particulier des données provenant de plusieurs sous-corpus comparables de textes « en langue naturelle » et de textes « en traduction ». Nous proposerons aussi une première comparaison de résultats provenant d'un sous-ensemble du corpus italien avec un sous-ensemble contenant les traductions françaises des mêmes textes, qui constituent un matériau fragmentaire pour le moment parce que nous travaillons encore à l'insertion des textes dans le but de créer des bases parallèles de traduction de l'italien vers toutes les langues du projet. Nous montrerons comment une consultation ciblée des différents corpus et sous-corpus que la base nous permet de distinguer (italien « langue naturelle » vs français « langue naturelle », français « en traduction » vs français « langue naturelle », français « textes spécialisés » vs français « vulgarisation » vs français « littéraire ») peut nous aider à repérer des approximations ou des erreurs de traduction, nous aidant à construire une information lexicographique comparative de qualité.

2. Comparaison entre corpus « en langue naturelle » et « en traduction » : une perspective à mi-chemin entre traductologie descriptive et prescriptive

Nous appuyant sur des analyses qui ne considèrent pas la langue de traduction comme un « troisième code » (Frawley 1984), nous estimons pour ce que des textes traduits trouvent parfaitement leur place à l'intérieur d'une base textuelle unique d'une même langue, aux côtés de textes « en langue naturelle ». Cependant, sur le modèle de propositions d'utilisation de corpus de traduction dans un but didactique, tant pour l'enseignement des langues que pour celui de la traduction, il nous semble nécessaire d'offrir la possibilité d'une consultation de la base dans des sous-corpus distincts regroupant des textes des deux types et de définir des critères d'évaluation des textes traduits à intégrer dans la base, en constituant des corpus séparés de textes traduits dans toutes les langues du projet. Ces corpus nous sont utiles comme outils de mémoire de traduction pour travailler sur la partie bilingue de nos fiches lexicographiques dans une perspective plus prescriptive que descriptive. Comme le montrera notre comparaison de résultats provenant de notre base française LBC « en langue naturelle » et « en traduction » avec un corpus de près de 100.000 mots actuellement non intégré dans la base composé de traductions d'ouvrages de « vulgarisation » traduits en français (guides touristiques de la Toscane et sites de musées surtout), certains des textes qui nous intéressent présentent des caractéristiques que l'on peut assimiler à du « translationese » et ne pourraient que fausser des interrogations de la base visant à attester des formes ou structures typiques du français tel qu'il est écrit et parlé par la majorité des locuteurs de cette langue sans interférence avec une autre langue.

2.1 Information descriptive et prescriptive dans les dictionnaires LBC : universaux et écarts

A la suite de Baker (1993), nous partons du principe qu'il existe des universaux de traduction qui nous serviront de canevas pour l'illustration des différents types d'interrogation effectués à l'intérieur de nos sous-corpus et de comparaison des résultats obtenus. C'est sur ces universaux que nous nous basons pour fournir la partie descriptive de l'information lexicographique comparative détaillée présente dans la partie bilingue de nos dictionnaires. Cette information correspond d'abord à l'observation des corpus parallèles, qui fournissent des attestations de traduction des lemmes (mots ou collocations) décrits par le dictionnaire, apparaissant dans des citations bilingues à l'intérieur de la partie bilingue de l'article. Nous analyserons en particulier :

- la simplification (principalement, pour ce qui concerne notre corpus, le choix d'hyperonymes pour traduire certains termes plus spécifiques) qui donne lieu dans nos dictionnaires à l'introduction d'une information sémantique ajoutée qui accompagne le traduisant proposé : les traits distinctifs particuliers au lemme qui ne sont pas rendus par le traduisant seront indiqués avec ou sans parenthèses après le traduisant (par ex. *tavola* traduit par *peinture (sur bois)* et *tavoletta* traduit par *(petite) peinture (sur bois)*)
- le nivellement (non-respect du registre, par exemple le choix de technicismes plutôt que de mots de la langue générale et vice versa). Toutes les entrées ont une indication de marque d'usage. Dans le cas d'une traduction qui implique un changement de registre, ce changement sera relevé dans la partie « note de traduction » ou apparaitra dans la partie réservée aux indicateurs sémantiques distinctifs dans le cas où plusieurs traductions du même lemme seraient possibles avec ou sans perte de registre. C'est le cas par exemple de *tondo* italien (non marqué) par rapport à *médaillon* (non marqué) et à l'italianisme *tondo* (technicisme utilisé principalement par les historiens de l'art). Baker analyse aussi l'explicitation qui est particulièrement fréquente dans les textes qui nous intéressent parce qu'elle est quasi systématiquement utilisée lors de l'usage d'un italianisme, en particulier pour les *realia* qui ont un traitement particulier dans nos dictionnaires (cfr. Farina 2014, 2016). Il serait possible de rechercher d'une manière systématique ce type de données dans notre corpus en extrayant toutes les occurrences de « type de » ou « sorte de » ou les éléments indiqués entre parenthèses, mais nous avons volontairement laissé de côté cette catégorie qui est trop fortement reliée à l'objet décrit par nos textes et à des choix stylistiques partagés entre les auteurs de textes « en langue naturelle » et les traducteurs dans le contexte de notre base, et ne nous permettrait donc pas d'illustrer par une comparaison des deux types de ressources des

contraintes linguistiques reliées aux opérations de traduction[2]. Nous avons laissé de côté aussi la « normalisation » ou « conservatisme » qui s'adapte peu à notre matière, peu propice à la variation ou à l'exploration sur le plan lexical et stylistique. Contrairement à Baker (1993 : 243), qui définit les universaux de traduction comme des « features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems », nous avons adopté une perspective plutôt prescriptive, ou mieux didactique, en prenant en considération les phénomènes d'interférence (influence de la langue source sur la langue cible) fréquents dans des opérations de traduction qui concernent deux langues proches comme l'italien et le français et dans des textes dont la qualité est loin d'être homogène. L'interférence est en effet selon nous à la source non seulement de nombreux cas de simplification et d'écarts de nivellement trouvés dans nos comparaisons mais d'autres manifestations assimilables à des pertes découlant de l'opération de traduction, voire à des erreurs ou inexactitudes de traduction. Le modèle du TQA (Translation Quality Assessment) et, en particulier, les différents types de mesures de qualité qui peuvent orienter le traducteur vers une amélioration de la fluidité et de la précision peuvent nous servir de référence pour ce faire (cfr. « Multidimensional Quality Metrics », Uszkoreit et al. 2013). Ces analyses nous orientent principalement vers le choix d'une position qui peut sembler aller à l'encontre d'une exploitation de corpus descriptive comme celle de Baker. De fait, elle se présente comme un accompagnement permettant à l'utilisateur de nos dictionnaires d'effectuer des choix, sur la base d'une exploitation descriptive des ressources consultées telle que nous l'avons déjà décrite et de l'indication de données statistiques résultant d'analyses de fréquence comme celles que nous les présenterons ci-dessous. Le rédacteur des fiches lexicographiques pourra de plus décider, le cas échéant et lorsque nos analyses de ces données le pousseront à repérer des erreurs ou écarts qui pourraient être réduits, de ne pas proposer une forme qui apparait dans la base comme traduisant (tout en l'indiquant dans la partie de l'article fournissant des indications statistiques sur les traduisants trouvés) ou de rédiger la partie « note de traduction », facultative dans nos articles bilingues, pour conseiller les utilisateurs dans leurs choix en expliquant pourquoi certaines formes peuvent être préférées à d'autres.

[2] L'utilisation abondante d'italianismes est une caractéristique dominante dans les guides touristiques analysés, assimilable à une volonté de leurs auteurs de donner à ces textes une « touche d'italianité» (Farina 2014 : 61)

3. Langue naturelle vs langue traduite : observation du corpus

La différence de fréquence de mots ou de collocations présents dans des corpus comparables contenant des textes français en « langue naturelle » et des textes qui proviennent d'une traduction en français peuvent nous permettre de repérer des formes choisies sous l'influence de la langue source.

3.1 Fréquence zéro dans les textes en langue naturelle

Nous avons comparé la liste des mots présents dans le sous-corpus LBC de textes de vulgarisation écrits en français contenant 270.000 mots avec un corpus non intégré à la base pour le moment de textes de même type mais en traduction 93.000 mots en réalisant une liste des mots présents exclusivement dans le sous-corpus « en traduction ».

- fautes

La majorité des formes rencontrées sont assimilables à des fautes : absence d'accent (*cloitre*), influence de l'orthographe italienne le français (*baroche*), « francisation » excessive au niveau orthographique (*Caliari*) ou par l'utilisation d'une traduction française là où l'usage préconise la forme italienne (*Sainte-Réparate* désigne en français la personne ou la cathédrale de Nice mais pas l'église Santa Reparata de Florence, la forme française n'est attestée nulle part dans la base LBC) ou l'inverse (*Giove* n'est jamais utilisé en italien dans notre corpus, où il est traduit par *Jupiter*), utilisation de mots qui n'ont rien à voir avec la description du patrimoine florentin, probablement parce qu'ils correspondent à un sens du mot-source qui s'applique à d'autres contextes (*coursive* dans une description du Dôme de Florence, ou *panonceau* pour se référer aux compartiments des portes du Paradis). Ce genre d'erreurs ne donne pas lieu à la réalisation d'une information ciblée à l'intérieur des dictionnaires sauf dans le cas d'une grande fréquence de l'erreur (par ex. pour *panonceau* présent dans plusieurs sources avec un total de 8 occurrences mais pas *coursive* qui n'a qu'une attestation).

- nivellement

On peut distinguer des formes qui correspondent à une différence « pragmatique » ou stylistique entre français et italien qui ne nous intéressent pas d'un point de vue lexicographique, comme l'utilisation de *mentionnons* dans plusieurs textes en traduction qui ne se retrouve dans aucun des textes de la base complète, ou de certaines formes du passé-simple (*décora, succéda*) qui ne sont pas utilisées dans les textes de vulgarisation en français « naturel ». Il s'agit de formes qui correspondent à des normes différentes relatives aux types de texte du corpus : une analyse plus approfondie pourrait probablement nous permettre d'observer un usage peu ou pas attesté du « nous » dans les guides touristiques, et l'usage peu fréquent de formes au passé-simple par rapport au passé-composé ou au présent, etc.

Ce qui nous intéresse beaucoup plus dans cette comparaison c'est de repérer des formes qui, tout en étant parfaitement « correctes » en français, peuvent être considérées comme hors contexte par rapport aux usages attestés dans le même type de contexte en langue naturelle. La différence dans l'usage d'un mot non attesté peut faire l'effet d'un « anachronisme » (différence dans la fréquence d'usage en synchronie). C'est le cas par exemple de l'adjectif *grand-ducal* et du participe passé *paraphé* dont les équivalents italiens sont plus fréquents dans la langue d'aujourd'hui que ne le sont leurs traductions littérales françaises. L'écart dans le registre peut aussi s'appliquer dans le cas d'une différence de « technicité ». L'adjectif *autographe* présent dans plusieurs sources de vulgarisation en traduction est absent des textes de même type de notre corpus en langue naturelle, mais on en trouve quelques occurrences dans des textes plus spécialisés du corpus général. La différence de registre donnera lieu à un marquage différencié entre lemme en langue source et sa traduction attestée.

3.2 Différence de fréquence dans les textes-source par rapport aux textes-cible

Pour illustrer les phénomènes de simplification, nous avons interrogé deux sous-corpus de notre base LBC constitués de 51 vies de l'ouvrage *Le vite de' più eccellenti pittori, scultori e architettori* de G. Vasari (1568) et de leurs traductions en français (traduction Leclanché-Weiss, 1900). Ne pouvant encore nous baser sur des statistiques provenant des bases parallèles de traduction (pour la description de ces bases cfr. Zotti 2017), nous nous sommes concentrés sur des mots français qui avaient une grande fréquence en comparant cette fréquence à celle du mot le plus proche en italien (même sens, mêmes traits distinctifs). Ceci nous a permis de relever des écarts de fréquence qui nous pousseront à une étude plus approfondie dans le but de définir des réseaux analogiques dans les deux langues qui nous donnent la possibilité de proposer des liens de traduction permettant d'éviter une perte de précision. *Tableau* a par exemple une fréquence de 2232 par million de mots dans notre sous-corpus français tandis que *quadro* a une fréquence de 793 par million de mots dans le sous-corpus italien contenant les mêmes textes en langue originale. Un grand nombre d'hyponymes de *quadro* sont en effet traduits par *tableau* en français. Si cette perte est probablement compensée par l'ajout de traits distinctifs qui accompagnent le mot, nous retenons que le traducteur ne pourrait que gagner en précision si nous lui proposions d'autres formes pour rendre le sens de ces différents hyponymes.

4. Conclusion

La comparaison de résultats qui concernent la fréquence de formes à l'intérieur du corpus LBC nous a permis d'illustrer l'utilisation de différents

sous-corpus pour orienter l'information tant descriptive que normative que nous souhaitons fournir dans la partie bilingue de nos dictionnaires LBC.

« Nous considèrerons, même si cela reste à démontrer […] qu'une sur- ou une sous-représentation d'un phénomène linguistique donné peut correspondre à une violation de la contrainte d'usage […] et qu'une bonne traduction se doit de tendre vers une homogénéisation entre la langue originale et la langue traduite. » (Loock *et al.* 2013 : sp)

L'application de méthodes visant à la vérification de la qualité des traductions et la création d'outils qui se basent sur des analyses critiques de traductions existantes, en les comparant, en particulier, à des productions qui ne passent pas par la médiation d'une autre langue devrait permettre une optimisation du caractère naturel des textes traduits et de la précision, objectif essentiel pour la diffusion d'une information de qualité.

Bibliographie

Baker M. (1993). Corpus Linguistics and Translation studies. Implications and Applications. In Baker M. and al. editors, *Text and Technology*, Amsterdam/Philadelphie, Benjamins, pp.233–250.

Billero R., Nicolas Martinez, M.C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In CHIMERA Romance Corpora and Linguistic Studies, Vol.4, No. 2, pp 203-216, ISSN: 2386-2629, 2017

Farina A. (2016). Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique, PUBLIF@RUM, vol. 24 http://publifarum.farum.it/ezine_articles.php?id=335

Farina A. (2014). Descrivere e tradurre il patrimonio gastronomico italiano: le proposte del Lessico plurilingue dei Beni Culturali. In Chessa F. and De Giovanni C., *La terminologia dell'agroalimentare*, Milan, Franco Angeli, pp. 55-66.

Frawley W. (1984). Prolegomenon to a theory of translation. In Frawley W. editor, *Translation: Literary, Linguistic and Philosophical Perspectives*, Newark, Univ. of Delaware Press : 159-175

Loock R., Mariaule M. and Oster C. (2013). Traductologie de corpus et qualité : étude de cas. Tralogy - Session 5 - Assessing Quality in MT / Mesure de la qualité en TA http://lodel.irevues.inist.fr/tralogy/index.php?id=188

Johansson S. and Hofland K. (1994). Towards an English-Norwegian parallel corpus. In Fries U. and al. editors, *Creating and using English language corpora*, Amsterdam, Rodopi pp. 25-37.

Loock R. (2016), *La Traductologie de corpus*. Villeneuve-d'Ascq. Presses Universitaires Septentrion.

Uszkoreit H., Burchardt A. and Lommel A. (2013). A New Model of Translation Quality Assessment Tralogy - Session 5 - Assessing Quality in MT / Mesure de la qualité en TA http://lodel.irevues.inist.fr/tralogy/index.php?id=319

Zotti V. (2017). L'integrazione di corpora paralleli di traduzione alla descrizione lessicografica della lingua dell'arte : l'esempio delle traduzioni francesi delle Vite di Vasari. In Zotti V., Pano A. editors, *Informatica Umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali.* Firenze University Press.

Il rapporto tra famiglie di anziani non autosufficienti e servizi territoriali: un'analisi dei dati esploratoria con l'Analisi Emozionale del Testo (AET)

Felice Bisogni[1], Stefano Pirrotta[2]
[1]Associazione GAP - SPS Scuola di Psicoterapia Psicoanalitica - felice.bisogni@gmail.com
[2]Associazione GAP - SPS Scuola di Psicoterapia Psicoanalitica - stefanopirrotta@gmail.com

Abstract
In this paper the authors present a research committed by a local authority to explore the relationship between not self-sufficient elders, their family members and the community based assistance services they uses. The exploratory data analysis, conducted with the Emotional Text Analysis (ETA) (Carli, Paniccia, 2002), was used to identify emotional and cultural factors related to the experience of assisting and being assisted at home and within the community based services. The ETA has been realized on an assembled text corpus produced transcribing 45 audio recorded interviews to not self-sufficient elders and their family members, patients of general practitioners and/or users of the community based services (home-based and half-residential). The interviews has been processed with T-Lab statistic software (Lancia, 2004) and ETA has been applied to produce a clusters analysis. Four clusters of dense words related to each others on 3 factorial axes emerged. From the factorial axes emerges a emotional representation of elderlness as a continuos allert related to the risk of dyng and as a depressive prescription to survive related to the pretension to be assisted within their own family in virtue of "blood ties". The reciprocal control and contentiousness, and the desirers to transgress the obligation of care giving and being cared are some relevant emotions emerging by the ETA. The research's results shows also a demand of a new assistance model emerges, founded on the possibility to talk, to play and to have fun with others. Finally it emerges a demand of services not only dealing with medical problems but also providing psychological support and training to the families to develop relational competences and to build reliable relationship out of the family. In the conclusions of the paper some considerations regarding the relationships between the clusters on the factorial axes and between clusters and illustrative variables are highlithed.

Abstract
In questo articolo gli autori presentano una ricerca, condotta con la metodologia dell'Analisi Emozionale del Testo (AET) (Carli, Paniccia, 2002), commissionata da un ente locale al fine di esplorare i fattori emozionali che organizzano l'esperienza di relazione tra un gruppo di anziani non autosufficienti e i loro familiari e alcuni servizi socio-sanitari territoriali. L'AET è stata realizzata su un corpus di testo assemblato trascrivendo 45 interviste audio registrate ad anziani non autosufficienti e loro familiari, che utilizzano servizi di medicina generale e/o servizi sociali territoriali (di tipo domiciliare o semiresidenziale). Le interviste sono state processate con il software statistico T-lab (Lancia, 2004) e l'AET è stata applicata per produrre una Cluster analysis. Dall'analisi sono emersi 4 cluster di "parole dense" (Carli, Paniccia, 2002) in rapporto tra loro su 3 assi fattoriali, che rappresentano il modo emozionale condiviso con cui gli intervistati parlano delle loro attese sui servizi Dall'interpretazione dei dati è emerso un rapporto tra famiglia ed anziano in crisi nel condividere desiderio e piacevolezza nello stare insieme. Emerge una rappresentazione emozionale dell'anzianità come allerta continua di fronte al rischio di morire e prescrizione depressiva a sopravvivere connessa alla pretesa di essere assistiti all'interno della propria famiglia in virtù di "rapporti di sangue". A questo si contrappone il desiderio di trasgredire l'obbligo famigliare ad assistere e farsi assistere. I risultati della ricerca rilevano una domanda di nuovi modelli di assistenza fondati sulla possibilità di parlare, giocare e divertirsi. Una domanda di servizi non rivolti esclusivamente ai problemi medici ma anche a offrire supporto psicologico e formazione alle famiglie per sviluppare competenze relazionali e relazioni affidabili all'esterno della famiglia. Nelle conclusioni vengono messe in evidenza alcune considerazioni riguardanti il rapporto tra cluster sugli assi fattoriali e tra i cluster e le variabili illustrative.

Keywords: Emotional Text Analysis (ETA), assistance, elders, family, community based services.

1. Introduzione

Sono circa 2,5 milioni gli anziani non autosufficienti presenti in Italia. Secondo le più recenti previsioni ISTAT (2017), la percentuale di individui di 65 anni e più crescerà di oltre 10 punti percentuali entro il 2050, arrivando a costituire il 34% della nostra popolazione. La presenza di un anziano non autosufficiente in famiglia diventerà sempre più un'esperienza comune per le famiglie italiane. Diversi studi hanno mostrato come l'organizzazione dell'assistenza agli anziani non autosufficienti da parte dei propri familiari comporti significativi problemi emozionali (Haley, 2003). Un recente studio

ha analizzato il testo di 26 interviste a familiari di anziani non autosufficienti con esperienza di assistenza da parte di un badante (Paniccia, Giovagnoli, Caputo, 2015). Dall'analisi del testo, condotta tramite la metodologia AET (Carli, Paniccia, 2002), è emerso come i sistemi di relazione familiari entrino in crisi contestualmente all'inattività e alla malattia dell'anziano. L'autrice afferma che la domanda delle famiglie ai servizi sia quella di non essere emarginate con il loro problemi entro il solo contesto familiare, per altro in cambiamento. "Sul piano della ricerca - afferma Paniccia - va sviluppata la differenza, proposta anche dagli intervistati, tra esplorazione dei vissuti degli anziani assistiti da un lato, degli altri membri della famiglia dall'altro". In quest'ottica, la ricerca-intervento proposta risponde a questo invito, esplorando il vissuto e le attese di un gruppo di anziani non autosufficienti e loro familiari nei confronti di alcuni servizi territoriali.

2. Il progetto di ricerca-intervento psicosociale

Il progetto di ricerca-intervento è stato realizzato dagli autori per conto dell'Associazione GAP, un'organizzazione che si occupa di ricerca e intervento psicosociale nell'ambito della disabilità. Il committente è stato un ente locale interessato a coinvolgere anziani non autosufficienti e loro familiari nella costruzione di nuovi modelli di assistenza coerenti con la domanda delle famiglie stesse. L'ente locale intendeva sviluppare un'offerta di servizi d'assistenza innovativi a fronte di cambiamenti sociali e culturali che stanno profondamente modificando l'organizzazione tradizionale della famiglia. Famiglia in passato maggiormente attrezzata al proprio interno per provvedere all'assistenza degli anziani. In tale contesto la ricerca intervento psicosociale è stato proposta come strumento di esplorazione del rapporto tra servizi d'assistenza rivolti agli anziani presenti nel territorio di competenza dell'ente committente e famiglie che a tali servizi si rivolgono. In tale contesto GAP a un gruppo di familiari e anziani non autosufficienti. Tutte le interviste sono state audio-registrate e trascritte in modo da ottenere il testo su cui è stata poi applicata l'Analisi Emozionale del Testo. In questa sede presentiamo i risultati dell'Analisi Emozionale del Testo applicata al testo prodotto trascrivendo 45 interviste a familiari e anziani non autosufficienti.

2.1. La raccolta dei dati

Le interviste sono state realizzate a 45 familiari e anziani non autosufficienti in carico ai servizi di medicina generale o ai servizi di centro diurno per anziani fragili partner del progetto. Di questi circa il 60 % usufruivano di servizi di medicina generale insieme al servizio di centro diurno per anziani fragili. Il restante 40% utilizzava esclusivamente i servizi di medicina generale. Sono state realizzate 25 interviste ad anziani e 20 interviste a loro

familiari. Le interviste sono state trattate in un unico corpus e per questo in analisi è stata inserita la variabile illustrativa "ruolo dell'intervistato", differenziando le interviste ad anziani da quelle a familiari. L'età media degli anziani intervistati è di 79 anni, mentre l'età media dei famigliari è di 60 anni. Gli intervistati sono stati scelti in ordine al criterio di coinvolgere nella ricerca chi ponesse ai servizi partner problemi complessi che i servizi stessi sentivano di avere difficoltà a prendere in carico. Questo nell'ipotesi che gli intervistati potessero poi partecipare ad un intervento psicosociale fondato sulla restituzione dei risultati della ricerca e sulla loro discussione critica al fine di contribuire alla progettazione di modelli di assistenza più in linea con i problemi sperimentati. Agli intervistati è stato proposto di partecipare a un'intervista aperta, non strutturata, con una sola domanda stimolo seguita dall'invito a dire tutto quello che veniva in mente. La domanda stimolo è stata la seguente: "nell'ambito di un progetto di ricerca-intervento siamo interessati a esplorare il rapporto tra servizi di assistenza, anziani e famiglie che a tali servizi si rivolgono. In particolare ci interessa esplorare il punto di vista dei familiari e degli anziani. Aggiungiamo che stiamo intervistano anche un gruppo di medici di base e di operatori dei servizi socio-sanitari. Siamo interessati alla sua esperienza; vorremo ascoltarla e raccogliere ciò che lei ha da dire". Gli intervistatori si sono presentanti come psicologi professionisti membri di un'associazione interessata a costruire servizi per l'invecchiamento e la non auto-sufficienza. Agli intervistati è stato detto che i risultati della ricerca sarebbero stati condivisi con tutti gli interessati per capire quali iniziative sviluppare.

3. Metodologia

L'Analisi Emozionale del Testo (Carli, Paniccia, 2002) è uno strumento proprio della ricerca-intervento psicosociale, sviluppato per esplorare i modi in cui i gruppi sociali simbolizzano emozionalmente e in modo condiviso un contesto o un tema e come queste simbolizzazioni organizzino il comportamento di quel gruppo. Tale metodologia, fondata sul principio del *conoscere per intervenire,* prevede l'attivazione di un processo di esplorazione, analisi e discussione critica della "cultura locale" condivisa entro un determinato contesto, in relazione al tema posto ad oggetto della ricerca. L'utilizzo di AET implica la destrutturazione del processo narrativo e delle connessioni che costituiscono il senso intenzionale dei discorsi entro un testo posto in analisi. Questo approccio metodologico è fondato sull'individuazione di gruppi di parole in rapporto tra loro che più di altre veicolavano significati emozionali: parole definite "parole dense". Operativamente abbiamo realizzato il processo statistico e informatico attraverso il software T-lab (Lancia, 2004) scegliendo la strategia dell'Analisi

Tematica dei Contesti elementari non supervisionata. Le interviste realizzate sono state assemblate entro un unico corpus, composto da 14053 tokens e 4121 types mentre gli hapax rilevati sono stati 230. Per quanto riguarda la sua ricchezza lessicale, il TTR (Type/Token Ratio) è 0.293. Abbiamo raggruppato le occorrenze di "parole dense" entro lessemi e in questo corpus ne sono stati individuati e messi in analisi 856. Il numero di "contesti elementari" di testo classificati 1423 (= 99.58%; del totale di 1429). Il processo di elaborazione dei dati seguito dal software comporta i seguenti passi: a) costruzione di una tabella di dati di unità contesto x unità lessicali (fino a 150,000 righe x 3,000 colonne), con valori di presenza/assenza; b) TF-IDF normalizzazione e scalaggio dei vettori riga alle unità lunghezza (norma Euclidea); c) clusterizzazione delle unità contesto (misure: coefficiente coseno; metodo: bisezione K-means); d) - limatura delle partizioni ottenute e, per ciascuna di esse: e) costruzione di una tabella di contingenza di unita lessicali x clusters; f) test del chi quadro applicato a tutte le intersezioni della tabella di contingenza; g) analisi delle corrispondenze della tabella di contingenza di unità lessicali x clusters. L'analisi statistica ha permesso di individuare diversi cluster corrispondenti a raggruppamenti di parole co-occorrenti. I cluster sono quelli che hanno una ricorsività significativa entro il testo e rappresentano le dimensioni più trasversali che caratterizzano la cultura locale esplorata.

4. Risultati

Il corpus delle interviste è stato elaborato con il software T-Lab che ha proposto come ottimale una partizione a 4 Cluster (CL) in rapporto tra loro su tre fattori (le cui percentuali di inerzia sono Fattore 1= 41,24%, Fattore 2= 32,68%, Fattore 3= 26,08%). Il cluster 3 e il cluster 2 sono in rapporto su polarità opposte del primo fattore; il cluster 1 e il cluster 4 sono in rapporto su polarità opposte del secondo fattore, mentre il cluster 1 e il cluster 3 sono in rapporto sul terzo fattore. Nella tabella (fig.1) è riporta la lista per cluster delle "parole dense" e le variabili illustrative relative al gruppo delle interviste degli anziani (_ruol_anz) e al gruppo delle interviste dei familiari di anziani (_ruol_fam).

Tabella 1: Lista parole dense per cluster con i relativi valori di chi2

	CLUSTER 1 N. of e.c..: 448		**CLUSTER 2** N. of e.c.: 371		**CLUSTER 3** N. of e.c.: 383		**CLUSTER 4** N. of e.c.: 221
χ2	soit: 31.48%	χ2	soit: 26.07%	χ2	soit: 26.91%	χ2	soit: 15.83%
171,81	problema	155,86	centro	408,52	figli	122,43	imparare
167,27	casa	116,53	persona	90,02	moglie	109,56	cura
79,08	uscire	83,4	aiutare	87,15	fratello	97,87	giocare
71,56	lasciare	68,86	trovare	52,81	sposare	61,33	parlare
57,67	vivere	63,55	malattia	46,33	mangiare	49,95	fumare
41,82	bisogno	57,09	dottore	40,44	dormire	47,67	giardino
36,62	h24	55,95	psicologia	37,92	morire	44,09	dimenticare
27,08	abbandonare	52,96	supporto	36,57	mamma	42,25	insieme
26,38	libero	36,48	municipio	35,14	telefono	36,51	somatizzare
25,59	badante	31,94	gruppo	34,77	marito	35,9	gita
23,46	pulire	26,73	amicizia	31,17	maschio	32,1	simpatia
20,33	costringere	24,09	frequentare	28,96	nonni	31,21	riflettere
19,05	persona	22,05	offrire	26,96	femmina	31,21	sigaretta
18,71	autonomo	21,62	cooperativa	26,77	cadere	25,17	ascoltare
17,74	perdere	21,28	informazione	26,68	soldi	25,17	spazio
16,72	**_ruol_fam**			**27,45**	**_ruol_anz**		

Di seguito, una lettura dei raggruppamenti di parole dense e della loro collocazione sul piano fattoriale.

4.1. Cluster 3: obbligo all'assistenza intra-famigliare e prescrizione alla sopravvivenza

Il cluster è presente in percentuale statisticamente maggiore entro il testo delle interviste agli anziani (38,4%). Gli intervistati parlano del rapporto con i propri famigliari: *figli, le mogli, i fratelli.* L'assistenza viene inscritta entro il vincolo obbligante dell'essere una *famiglia* (etimologicamente da *famulo, colui che serve, che si prende cura*): emerge l'attesa che il ruolo famigliare implichi il dovere di occuparsi di chi non riesce a vivere da solo, preoccupandosi di garantire la sopravvivenza e occupandosi di bisogni inderogabili come *mangiare* e *dormire*. Emerge una rappresentazione infantilizzante dell'anziano che sollecita l'instaurarsi di rapporti di dipendenza e accudimento. In tale contesto la quotidianità, deprivata di desideri ed obbiettivi, sembra scorrere in modo depressivo in attesa di *morire*, con il rischio di una chiusura depressiva all'interno della famiglia. L'anzianità sembra identificata con la figura del vecchio morente che non ha più nulla da dare o da chiedere alla vita. L'unico riferimento alla vitalità entro il cluster è quello connesso a parole come *nipoti* e *telefonare*: laddove si allenta l'obbligo dell'assistenza sembra farsi spazio la possibilità di un rapporto piacevole e gratificante.

4.2. Cluster 2: ricerca di servizi e domanda alla psicologia

In questo cluster è rappresentato il processo di ricerca di servizi di assistenza. Si cercano *centri*, contesti estranei alla famiglia, che *aiutino* ad occuparsi dei problemi della *persona* non autosufficiente. Da un lato si guarda alla sua soggettività, dall'altro si rappresenta una ricerca affannosa di servizi fondata sull'angoscia di *trovare* soluzioni. La non autosufficienza è rappresentata come *malattia.* Ciò comporta un vissuto di urgenza e pericolo e la fantasia di dover contrastare qualcosa che mette a rischio la sopravvivenza. Su questo si chiama in causa il *dottore*, in ipotesi il medico di base, cui viene attribuita una competenza utile. Allo stesso tempo è chiamata in causa la *psicologia* cui viene richiesto un intervento di *supporto.* Si evoca in tal modo una prospettiva di intervento alternativa alla cura. Si chiede di essere aiutati a *prepararsi* e di essere accompagnati, di parlare con qualcuno poiché ci si sente impreparati, confusi.. A questo proposito i famigliari sembrano portatori di una domanda di ascolto e consulenza fondata sul *parlare.* Agli enti locali e del privato sociale gli intervistati si propongono come clienti, viene domandata l'articolazione di un'offerta di servizi, valorizzando dispositivi d'intervento di gruppo.

4.3. Cluster 1: funzione di controllo delegata alla badante e paura del cambiamento

Il cluster è presente in percentuale statisticamente maggiore entro il testo delle interviste ai familiari (39%). Gli intervistati parlano del *problema* che vivono, situato nella *casa,* un contesto chiuso che offre riparo e che al contempo costringe. Da un lato si cercano vie di uscita e d'altro lato c'è difficoltà a lasciare, ad allontanarsi da rapporti protettivi e vincolanti. Viene rappresentato un contrasto tra queste emozioni e il *vivere*: emerge un sentimento di vita contrastata, per dirla con Canguilhem (1998). In tale contesto si è presi dalla fantasia di *abbandonare:* emerge l'emozionalità della colpa. Ciò avviene entro un contesto in cui la non autosufficienza viene trattata quale *bisogno* esclusivamente fattuale e pressante, *24 ore su 24*. L'invecchiamento è rappresentato come evento che non lascia tregua, che tormenta e angoscia. In tale contesto si chiede l'intervento della *badante* per ripristinare il controllo, fare ordine. La badante è rappresentata come una necessità motivata dal bisogno. L'assistenza all'anziano è qualcosa a cui ci si sente *costretti* o da cui liberarsi, tertium non datur. Ma in questo cluster vediamo come vivendo l'invecchiamento come bisogno continuo e prescrivendo l'assistenza si generi colpa. Colpa connessa all'impotenza per il non riuscire a rapportarsi ai cambiamenti con cui la non autosufficienza confronta.

4.4. Cluster 4: domanda di costruzione di contesti dove parlare, giocare, apprendere.

In questo cluster gli intervistati esprimono una domanda di contesti e rapporti fondati sull'*apprendimento,* il *gioco* e sulla *parola*. Emergono desideri e si riconoscono risorse che evocano la possibilità di trovare motivi per cui valga la pena vivere. Emerge una rappresentazione della vecchiaia caratterizzata da vitalità e desiderio di trasgredire. Si allenta la prescrittività dell'obbligo della sopravvivenza: la vecchiaia è anche creatività, possibilità di smarcarsi dagli obblighi rituali della vita sociale. Il riconoscimento del limite del tempo, l'avvicinarsi della fine, motiva la ricerca di esperienze piacevoli che diano senso alla vita. Si evoca il *divertimento* come obbiettivo alternativo al controllo e alla sorveglianza senza obbiettivi. Sottolineiamo come la domanda divertimento implichi il riconoscimento di una verità non scontata: che si è ancora vivi fino a cinque minuti prima di morire.

5. Conclusioni

Per concludere proponiamo alcune considerazioni sul rapporto tra i cluster sui tre assi fattoriali. Ricordiamo che il cluster 3 e il cluster 2 sono in rapporto su polarità opposte del primo fattore, il cluster 1 e il cluster 4 sono in rapporto su polarità opposte del secondo fattore, mentre il cluster 1 e il cluster 3 sono in rapporto sul terzo fattore. Sul primo fattore emerge come la dimensione motivazionale che sostiene la domanda di servizi da parte della famiglia sia il desiderio di uscire dall'obbligo familiare. È il vissuto di obbligo e l'incapacità di condividere entro i rapporti desiderio ed interessi che spinge la famiglia in un'affannosa ricerca di interlocutori e professionisti esterni. Sul secondo fattore emergono diverse modalità di rapportarsi al problema della non autosufficienza. Su di un polo del fattore (cluster 1) la fattualizzazione dell'invecchiamento come bisogno continuo di assistenza che mette in pericolo la sopravvivenza mostra come i problemi associabili alla non autosufficienza non siano esplorati. Tali problemi sembrano piuttosto presunti dal familiare in modo autoreferenziale. L'emozionalità della colpa e la fantasia irrealizzabile di ristabilire il controllo su una situazione in cambiamento vissuta come persecutoria sono corollari di tale autorefenzialità sottesa dall'incompetenza a utilizzare i rapporti familiari come contesto di confronto e scambio sui problemi e sul da farsi. D'altro lato, sull'altro polo del secondo fattore il riconoscimento di limiti, quali ad esempio il tempo limitato della vita e l'ineluttabilità della fine, sembra fare spazio al riconoscimento del desiderio degli anziani di divertirsi anche concedendosi qualche trasgressione, come alternativa a sopravvivere in modo controllante e mortifero. Infine il terzo fattore suggerisce una relazione tra la dinamica di autorefenzialità dei rapporti familiari e la domanda di servizi emergente

entro la cultura in analisi, a cui si chiede non soltanto di curare ma anche di aiutare la famiglia a sviluppare competenze e confrontarsi sui propri problemi. I risultati della ricerca suggeriscono una domanda nei confronti di servizi di accompagnamento e che sostengano la famiglia – intesa come contesto di rapporti tra la persona non autosufficiente e i suoi familiari - nel riconoscimento di desideri e obbiettivi attorno a cui organizzare l'assistenza e la convivenza nel modo più piacevole, vitale e divertente possibile.

Bibliografia

Carli R., Paniccia R.M. (2002). *L'analisi emozionale del testo*. Franco Angeli, Roma.

Haley, W. E. (2003). Family caregivers of elderly patients with cancer: understanding and minimizing the burden of care. *The journal of supportive oncology,* 1(4 Suppl 2), 25-9.

ISTAT (2017), *Demografia in cifre*, Roma, Istituto Nazionale di Statistica – www.demo.istat.it.

Lancia, F. (2004). *Strumenti per l'analisi dei testi*. Franco Angeli, Roma.

Paniccia, R. M., Giovagnoli, F., & Caputo, A. (2015). In-home elder care. The case of Italy: the badante. *Rivista di Psicologia Clinica*, (2), 60-83.

Esperienza di analisi testuale di documentazione clinica e di flussi informativi sanitari, di utilità nella ricerca epidemiologica e per indagare la qualità dell'assistenza.

Antonella Bitetto[1], Luigi Bollani[2]
[1]Azienda Socio Sanitaria Territoriale di Monza – a.bitetto@asst-monza.it
[2]Università di Torino – luigi.bollani@unito.it

Abstract
This study finds reason in the now wide availability of clinical documentation stored in electronic form to track the patient's health status during his care path or for sending information to other institutions on the activities carried out for administrative purposes. The diffusion of these methods now makes available many biomedical collections of electronic data, easily accessible at low cost that can be used for research purposes in the field of observational epidemiological studies, in analogy with what was historically already practiced in studies based on the reviewing of medical records. However, since these collections are not organized according to specific survey schemes, they sometimes do not allow the index events to be discriminated with the necessary reliability between one source and another. It has always been believed that the critical re-reading of texts can partially help these informative shortcomings with the aim of bringing back - according to possibility - the words or segments contained in the texts, to statistically analyzable categories. The recent transfer of these collections from paper to electronic forms opens the possibility of carrying out this process automatically, reducing time and costs of the process and perhaps increasing its reliability. It is proposed to address the problem, showing study criteria and an example of analysis based on an empirical experience, consistent with the needs of a biomedical context.
Keywords: textual analysis; electronic health data; medical thesaurus; analysis of lexical correspondences; emergency in psychiatry

Riassunto
Questo studio trova ragione nella ormai ampia disponibilità di documentazione clinica archiviata in forma elettronica per tracciare lo stato di salute del paziente durante il suo percorso di cura o inviare informazioni ad altri enti sulle attività svolte a scopo amministrativo. La vasta diffusione di questi metodi mette a disposizione ormai numerose raccolte di tipo

biomedico, facilmente accessibili a basso costo che possono essere utilizzate a scopo di ricerca nel settore degli studi epidemiologici osservazionali, in analogia con quanto storicamente veniva già praticato negli studi basati sulla rilettura delle cartelle cliniche. Non essendo però tali raccolte organizzate secondo schemi di rilevazione specifici a volte non permettono di discriminare con la necessaria attendibilità tra una fonte e l'altra gli eventi indice. Da sempre si ritiene che la rilettura critica dei testi possa, parzialmente soccorrere a tali carenze informative nell'obiettivo di ricondurre - secondo possibilità - le parole o i segmenti contenuti nei testi disponibili a categorie statisticamente analizzabili. Il recente passaggio di tali raccolte dalla forma cartacea a quella elettronica apre la possibilità di operare per via automatica riducendo tempi e costi del processo e forse incrementandone l'attendibilità. Ci si propone di affrontare il problema, mostrando criteri di studio ed un esempio di analisi basato su un'esperienza empirica, conforme alle esigenze di un contesto biomedico.
Parole chiave: analisi testuale; dati sanitari elettronici; thesaurus medico; analisi delle corrispondenze lessicali; psichiatria d'urgenza

1. Introduzione

Il progressivo processo di dematerializzazione della documentazione clinica (valutazioni specialistiche ambulatoriali, verbali di Pronto Soccorso, referti esami diagnostici) e l'implementazione dei flussi di dati sanitari a scopo giuridico amministrativo (per il pagamento delle prestazioni erogate o per l'aggiornamento dell'anagrafe, dell'INPS etc.) hanno reso disponibili informazioni che possono essere utilizzate anche per obiettivi diversi da quelli per cui i dati sono raccolti. I dati sanitari informatizzati (EHR - "electronic health records"), vengono generalmente distinti in: a) strutturati (ad es. registrati utilizzando terminologie cliniche controllate come la Classificazione internazionale delle malattie -10ª revisione (ICD10) o la nomenclatura sistematica della medicina - Termini clinici (SNOMED-CT), b) semistrutturati (ad es. esami di laboratorio ed informazioni sulla prescrizione) che seguono uno schema che varia a seconda delle convenzioni adottate localmente, c) non strutturati (ad es. testo clinico) e d) binari (ad esempio file di immagini come Rx e TAC). La sistematicità di queste raccolte di dati, organizzati in maggioranza per entità individuali, li rende particolarmente preziosi per diversi scopi di ricerca epidemiologica che utilizza disegni di tipo osservazionale sia nell'ambito della qualità dell'assistenza che dell'epidemiologia più classica, che studia rischi ed esiti delle malattie (Mitchell J. et al., 1994). Per contro essendo tali raccolte organizzate per scopi altri da quelli del monitoraggio della qualità o della ricerca scientifica, spesso devono essere "trattate" prima di poter essere

analizzate con metodi statistici. In passato ciò veniva fatto attraverso la rilettura delle cartelle cliniche da parte di esperti della materia. Attualmente si cerca sempre più di ricorrere a metodi di analisi automatica dei testi che garantisce una miglior standardizzazione e revisione (Denaxas S. et al., 2017). A titolo di esempio si segnala che l'analisi automatica dei testi di flussi informativi e della documentazione clinica elettronica ha permesso d'indagare ambiti terapeutici e di sicurezza fondamentali come la qualità dell'assistenza infermieristica e l'occorrenza di eventi avversi come – tra i tanti - gli incidenti domestici, le reazioni allergiche e gli effetti collaterali ai farmaci (Ehrenberg A. et Ehnfors M., 1999; Coloma P.M. et al., 2011; Migliardi A. et al., 2004). Sono stati anche prodotti numerosi studi epidemiologici classici per lo più riferiti a patologie croniche ad alta prevalenza come le malattie cardiocircolatorie, il diabete o l'asma, all'estero e in Italia (Gini R. et al., 2016; Vaona A. et al. , 2017), in alcuni casi mettendo in evidenza bisogni di cura inespressi o complicanze dovute a ritardi o trattamenti inappropriati (Persell S.D., et al., 2009; Ho M.L., et al., 2012). Alcune ricerche si sono focalizzate sui disturbi mentali, area medica scelta per l'esperienza di analisi di testo di seguito presentata. In questo ambito la documentazione clinica elettronica permette di ottenere informazioni a basso costo su ampi settori di popolazione che possono ricomprendere casistiche difficili altrimenti da reclutare: questo è il caso di soggetti in fase prodromica ad alto rischio di sviluppare psicosi (Fusar-Poli P. et al., 2017) o autolesionisti (Zanus C. et al., 2017).

2. Metodi

La classificazione dei *corpora* non ancora studiati in categorie statisticamente analizzabili rappresenta un argomento controverso ma anche una sfida che giustifica, a nostro avviso, indagini di approfondimento delle procedure metodologiche da adottare. Nel seguito si propone un metodo per il trattamento di testi medici non strutturati di psichiatria, secondo criteri già in parte utilizzati in precedenti esperienze (Bitetto A. et al., 2017).

2.1. Corpus

Le informazioni provengono dai verbali di consulenze psichiatriche svolte presso il Pronto Soccorso di un ospedale universitario lombardo di grandi dimensioni (1250 letti accreditati).

Il corpus è monolingua - in italiano - composto da brevi testi scritti dallo psichiatra di turno alla fine della consulenza in urgenza. I referti sono verificati e quindi conservati dal servizio informativo ospedaliero, certificato ISO 9001/2015, che ha fornito il corpus dei dati, in forma anonima. Si sono analizzati 1721 referti, relativi al periodo 01/01/2012 – 31/12/2012.

2.2. Pretrattamento di filtraggio linguistico

Il *corpus* è stato sottoposto ad un pretrattamento di filtraggio linguistico. Dalle 177349 parole presenti nei referti originali sono state eliminate la punteggiatura, i numeri, i pronomi, gli articoli, le proposizioni, i nomi propri - anche dei farmaci- e le parole con una ricorrenza inferiore a 10. Ne è risultato un elenco di 1679 parole distinte, che è stato rivisto manualmente da un esperto per selezionare i termini in grado di descrivere i problemi/bisogni di salute mentale secondo il modello strutturale utilizzato dalla scala HoNOS (Wing J.K. et al., 1998; Lora A. et al., 2001). Si tratta di un modello di valutazione dello stato di salute mentale impostato per problemi e non sulle diagnosi, che difficilmente sono riportate nei referti di pronto soccorso. Il modello distingue 12 "problemi" riconducibili ai seguenti concetti:
item H1- COMPORTAMENTI IPERATTIVI, AGGRESSIVI; item H2 - COMPORTAMENTI DELIBERATAMENTE AUTOLESIVI; item H3 - PROBLEMI LEGATI ALL'ASSUNZIONE DI ALCOOL O DROGHE; item H4 - PROBLEMI COGNITIVI; item H5 - PROBLEMI DI MALATTIA SOMATICA; item H6 - PROBLEMI LEGATI AD ALLUCINAZIONI E DELIRI; item H7 - PROBLEMI LEGATI ALL'UMORE DEPRESSO; item H 8 - ALTRI PROBLEMI DA ALTRI SINTOMI PSICHICI; item H9 - PROBLEMI NELLE RELAZIONI SIGNIFICATIVE; item H10 - PROBLEMI NELLO SVOLGIMENTO DI ATTIVITÀ DELLA VITA QUOTIDIANA; item 11- PROBLEMI NELLE CONDIZIONI DI VITA; item H12 - PROBLEMI NELLE ATTIVITÀ LAVORATIVE E RICREATIVE.
In questo modo è stato creato un *thesaurus* composto da 214 locuzioni brevi e 81 parole singole riconducibili a 11 categorie cliniche (esclusa la H10, data la mancanza di locuzioni in grado di ricondurre ad essa). Nel *thesaurus* si sono inoltre considerate parole e acronimi che individuano accesi legati al "rifiuto delle cure". La procedura di filtraggio dei testi, basata sul *thesaurus* (ponendo anche attenzione a non includere contesti dove la parola chiave è negata), ha permesso di riclassificare 1629 referti che rappresentano la base dell'analisi.

2.3. Analisi statistica

I diversi referti sono stati esaminati per la presenza/assenza di ciascuna parola o locuzione chiave esaminata, in modo da introdurre per ogni parola una codifica binaria rispetto al complesso dei testi considerati. Successivamente tale codifica è stata estesa agli item della classificazione HoNOS valutando – in ogni referto – la presenza di ciascun item, determinata dalla presenza di almeno una parola chiave ad esso associata (l'assenza dell'item si determina per contro in mancanza di parole chiave ad esso associate). Per rappresentare l'associazione tra i diversi item, rispetto ai referti studiati, si è quindi condotta un'analisi delle corrispondenze

(Benzécri, Jean-Paul, 1973) sulla tabella testi x item HoNOS (in aggiunta ad essi si è anche incluso concetto di rifiuto/interruzione delle cure); per poter apprezzare inoltre le relazioni tra parole e comportamenti/problemi, espressi dalla classificazione introdotta, si sono aggiunte le parole e locuzioni chiave in forma supplementare.

3. Risultati

La tabella 1 mostra la distribuzione di frequenza delle aree problematiche descritte e riclassificate secondo i criteri della scala HoNOS.

Tabella 1 – Item HoNOS e percentuale di presenza del comportamento/problema riscontrato nei referti

Item HoNOS	H1	H2	H3	H4	H5	H6	H7	H8	H9	H11	H12	RifiutoCure
% di presenza nei referti	30.82	15.22	12.22	7.18	20.32	18.35	32.72	59.55	5.10	1.23	7.31	18.97

Come atteso i referti riferiscono soprattutto le manifestazioni cliniche del disagio attraverso descrizioni dettagliate dei sintomi osservati rispetto ad altri fattori di tipo ambientale (H9, H11, H12). Tra i sintomi, quelli di più frequente riscontro sono l'umore depresso (H7) e la classe che raccoglie tutte le manifestazioni cliniche non specificate "altri sintomi psichici" (H8). Molto frequente è anche la descrizione di problemi di natura organica (sintomi fisici H5) come atteso, visto che la gestione delle urgenze psichiatriche avviene presso il pronto soccorso generale in cui la richiesta di parere su accesi legati a problematiche fisiche è più alta che presso un ambulatorio di secondo livello. Molto elevata è anche l'occorrenza di comportamenti violenti ed iperattivi (H1), una delle urgenze più tipiche dell'ambito psichiatrico.

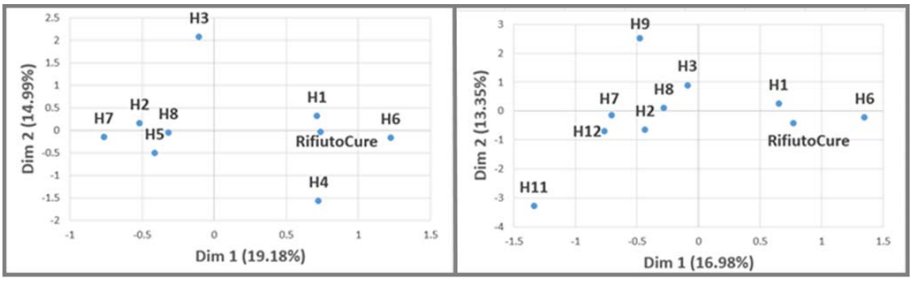


Figura 1 – A sinistra : rappresentazione congiunta dei primi 8 item HoNOS (sintomi psichici e fisici); A destra : sintomi comportamentali (H1, H2, H3), sintomi psichici (H6, H7, H8) e fattori ambientali precipitanti (H9, H11, H12)

Nella figura 1 – grafico di sinistra - sono rappresentati i risultati dell'analisi delle corrispondenze sulle categorie dei sintomi, l'area problematica di

maggior riscontro nei testi. Il primo piano fattoriale – mostrato nel grafico – spiega il 34.17 % della varianza totale. Rispetto alla dimensione 1, lungo l'asse delle ascisse, le categorie di sintomi si suddividono in due gruppi: sulla destra troviamo i problemi legati all'umore depresso (H7) vicino ad altri sintomi (H8), di cui come già detto l'ansietà rappresenta l'area più vasta, e i sintomi fisici (H5), confermando la probabile origine psicosomatica di parte di essi. Nel medesimo raggruppamento si collocano i comportamenti deliberatamente autolesivi e suicidari, che sono secondo la letteratura spesso associati a problemi di depressione. Su valori elevati di ascissa, sono invece raggruppati i sintomi psicotici (H6), i comportamenti agitati (H1), in relazione con il rifiuto delle cure, cui spesso infatti si associano. Risultano invece indipendenti dalle altre categorie di sintomi i problemi legati all'abuso di alcool e droghe (H3) e quelli dovuti alla presenza di problemi cognitivi di origine neurologica (H4), che occupano gli estremi della dimensione 2, individuate dall'asse delle ordinate. La stessa analisi è rappresentata nella figura 2, proiettando anche le parole pertinenti del *thesaurus* utilizzato.

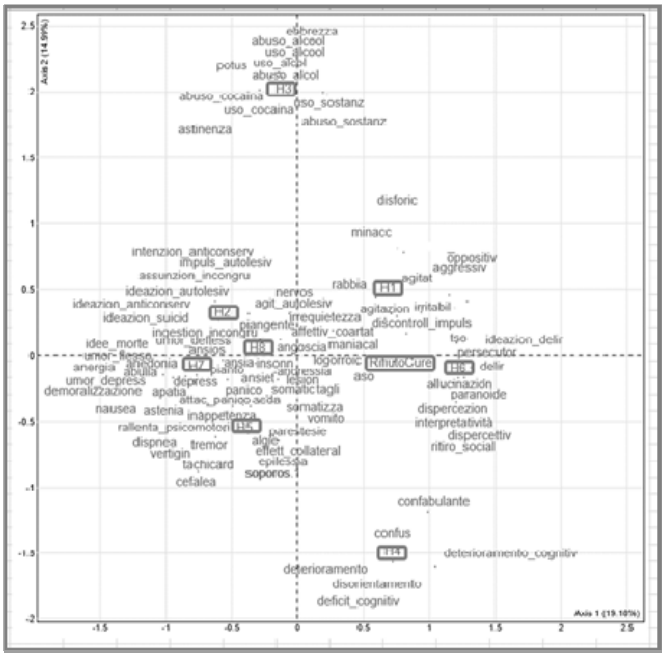


Figura 2 – Rappresentazione congiunta degli item HoNOS relativi ai primi 8 item e rappresentazione supplementare delle parole/locuzioni chiave utilizzati per individuare i diversi item

Riprendendo la figura 1 – grafico a destra – si trova una seconda analisi delle corrispondenze condotta sulle categorie di sintomi psichici e comportamentali insieme ai fattori precipitanti di tipo ambientale. In questo caso il primo piano fattoriale spiega il 30.33% della varianza totale.
La distribuzione dei sintomi psichici lungo l'asse delle ascisse conferma, come atteso, i risultati dell'analisi del primo subset di categorie. In questo caso è possibile notare la tendenza dei problemi legati all'abuso di alcool e droghe (H3) a disporsi verso il centro del grafico in prossimità della categoria altri sintomi (H8), con cui è possibile che certe manifestazioni siano in relazione. Per quanto riguarda i fattori ambientali emerge dai dati una relazione tra problemi di lavoro (H12), sintomi dello spettro depressivo (H7) e condotte deliberatamente autolesive (H2). È possibile che il Pronto Soccorso rappresenti un primo punto di accesso per un'utenza con forme reattive anche gravi, secondarie a fattori di stress occupazionale (burnout, depressioni reattive). Le altre categorie relative a problematiche ambientali (H9 e H11) si collocano agli estremi della dimensione 2, mostrando un certo grado di indipendenza rispetto all'occorrenza di sintomi comportamentali e psichici.

5. Conclusioni

L'esperienza empirica di analisi testuale automatica di referti del Pronto Soccorso conferma la sua utilità nell'indagare fenomeni complessi come le manifestazioni cliniche e i fattori di rischio dell'urgenza psichiatrica. L'analisi delle corrispondenze si dimostra un metodo semplice e utile per esplorare le relazioni tra le diverse dimensioni in esame.
Emergono per altro alcuni problemi legati alla qualità delle informazioni che, in quanto raccolte per altri scopi, presentano un eccesso di informazione rispetto ad alcune aree (manifestazioni sintomatologiche) mentre sono carenti in altre, come il grado di disabilità del soggetto non analizzabile come fattore precipitante dell'urgenza. È possibile che tali carenze possano essere superate acquisendo informazioni da altre fonti come alcuni ricercatori hanno fatto (Fusar-Poli P. et al., 2017). Resterebbe comunque aperto il problema di condividere e standardizzare i metodi di trattamento dei dati nelle diverse fasi dell'indagine, dalle modalità con cui sono raccolte le informazioni e compilati i referti, alla creazione di un *thesaurus* di parole e locuzioni chiave standard per la psichiatria sulla base di concetti teorici e criteri condivisi.

Bibliografia

Benzécri, J.P. (1973). *L'analyse des données*. *Vol. 2.* Paris: Dunod.
Bitetto A., et al. (2017). La consultazione psichiatrica in Pronto Soccorso come

fonte informativa sui bisogni inespressi di salute mentale. *Nuova rassegna studi psichiatrici* vol. 15 novembre 2017

Coloma P.M. et al. (2011). Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf.*; 20(1):1–11. 40.

Denaxas S., et al. (2017).Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *Bio Data Mining*;10:31

Ehrenberg A. et Ehnfors M. (1999). Patient problems, needs, and nursing diagnoses in Swedish nursing home records. *Nursing Diagnosis*; 10(2), 65-76.

Fusar-Poli P, et al. (2017). Diagnostic and Prognostic Significance of Brief Limited Intermittent Psychotic Symptoms (BLIPS) in Individuals at Ultra High Risk. *Schizophr Bull*; 43(1):48-56

Gini R. et al. (2016). Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian General Practitioners' electronic medical records: a validation study. *BMJ Open*; 6(12): e012413.

Ho ML, et al. (2012). The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus. *J Eval Clin Pract*. ;18(3):606–11.

Lora A. et al. (2001). The Italian version of HoNOS (Health of the Nation Outcome Scales), a scale for evaluating the outcomes and the severity in mental health services. *Epidemiology and Psychiatric Sciences;* 10.3: 198-204.

Migliardi A. et al. (2004). Descrizione degli incidenti domestici in Piemonte a partire dalle fonti informative correnti. *Epidemiologia & Prevenzione* ; 28.1: 20-26.

Mitchell J. et al., (1994). Using medicare claims for outcome research. *Medical care*; 35:589-602

Persell S.D. et al. (2009). Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. *Med Care*; 47(4):418–24.

Vaona A. et al. (2017). Data collection of patients with diabetes in family medicine: a study in north-eastern Italy. *BMC Health Serv Res.*;17(1):565

Wing J.K. et al., (1998). Health of the Nation Outcome Scales (HoNOS). Research and development. *The British Journal of Psychiatry*; 172 (1) 11-18

Zanus C. et al. (2017). Adolescent Admissions to Emergency Departments for Self-Injurious Thoughts and Behaviors. *PLoS One*.;12(1): e0170979.

Exploring the history of American philosophy in a computer-assisted framework

Guido Bonino[1], Davide Pulizzotto[2], Paolo Tripodi[3]
[1]Università di Torino – guido.bonino@unito.it
[2] LANCI, Université du Québec à Montréal – davide.pulizzotto@gmail.com
[3]Università di Torino – paolo.tripodi@unito.it

Abstract
The aim of this paper is to check to what extent some tools for computer-assisted concept analysis can be applied to philosophical texts endowed with complex and sophisticated contents, so as to yield results that are significant not only because of the technical success of the procedures leading to the results themselves, but also because the results, though highly conjectural, are a direct contribution to the history of philosophy

Sommario
Lo scopo di questo articolo è di verificare in che misura la computer-assisted concept analysis possa essere applicata a testi filosofici di contenuto complesso e sofisticato, in modo da produrre risultati significativi non solo dal punto di vista del successo tecnico delle procedure, ma anche in quanto i risultati stessi, sebbene altamente congetturali, costituiscono un contributo diretto alla storia della filosofia.
Keywords: philosophy, history of philosophy, paradigm, necessity, idealism, Digital Humanities, Text Analysis, Computer-assisted framework

1. Computer-assisted concept analysis

The development of artificial intelligence poses a methodological challenge to the humanities. Many traditional practices in disciplines such as philosophy are increasingly integrating computer support. In particular, *Concept Analysis* (CA) has always been a common practice for philosophers and other scholars in the humanities. Thanks to the development of *Text Mining* (TM) and *Natural Language Processing* (NLP), computer-assisted text reading and analysis can provide the humanities with new tools for CA (Meunier and Forest, 2005), making it possible to analyze large textual corpora, which were previously virtually unassailable. Examples of computer-assisted analyses of large corpora in philosophy are Allard et al., 1963; McKinnon, 1973; Estève et al., 2008; Danis, 2012; Sainte-Marie et al., 2010; Le et al., 2016; Meunier and Forest, 2009; Ding, 2013; Chartrand et al., 2016; Pulizzotto et al., 2016; Slingerland et al., 2017. The use of computer-

assisted text analysis is also relevant for the distant reading approach, developed by Franco Moretti in the context of literature studies (Moretti, 2005; Moretti, 2013), but which we are convinced can be usefully extended to different fields (for the application to philosophy see the Conference "Distant Reading and Data-Driven Research in the History of Philosophy" held in Turin in 2017, http://www.filosofia.unito.it/dr2/).

The main aim of this paper is to check to what extent some tools for computer-assisted CA can be applied to texts endowed with complex and sophisticated contents, so as to yield results that are significant not only because of the technical success of the procedures leading to the results themselves, but also because the results, though highly conjectural, are a direct contribution to the humanities. Philosophy, in particular the history of philosophy, seems to be a good case to be considered, because of the sophistication of its contents. Our main purpose is that of illustrating some of the different *kinds* of work that can be done in history of philosophy with the aid of computer-assisted CA.

2. Method

2.1. The corpus

To understand how TM and NLP can assist the work in history of philosophy, some standard methods have been applied to a specific corpus, which is provided by Proquest (www.proquest.com). The corpus is a collection of 20,751 PhD dissertations in philosophy discussed in the US from 1981 to 2015. It therefore contains 20,751 *documents*: each document is a text, comprising the title and the abstract of a dissertation, which are dealt with as a single unit of analysis. The corpus also contains some metadata, such as the author of the dissertation, the year of publication, the name of the supervisor, the university, the department, and so forth. In the present paper we are not going to exploit fully the wealth of information provided by these metadata, which are certainly worth being the subject of further research. However, we will use the crucial datum of the year of publication, which allows us to assume a diachronic (that is, historical) perspective on the investigated documents.

2.2. Data preprocessing

A preliminary step consists in a set of four preprocessing operations that allow us to extract the linguistic information needed for the analysis: 1) *Part of Speech* (POS) tagging; 2) *lemmatization*; 3) *vectorization*; 4) selection of the sub-corpora responding to Keyword In Context (KWIC) criteria.

The POS tagging and the lemmatization process are performed on the basis of the *TreeTagger* algorithm described by Schmid, 1994 and 1995. This

operation consists in the annotation of each word for each document according to its morphological category. Some irrelevant categories (such as determinants, prepositions and pronouns) are eliminated. Nouns, verbs, modals, adjectives, adverbs, proper nouns and foreign words are taken into account. The lemmatization process reduces a word to his lemma, according to the correspondent POS tag. At the end of this process, we can identify 17,750 different lemmas, which are called *types*.

The mathematical modeling of each document into a vector space is called vectorization. In such a model, each document is encoded by a vector, whose coordinates correspond to the TF-IDF weighting of the words occurring in that document. This weighting function calculates the normalized frequencies of the words in each document (Salton, 1971). At the end of the process, a matrix M is built, which contains 20,571 rows corresponding to each document, and 17,750 dimensions, corresponding to the types.

Finally, three sub-corpora are created on the basis of the *KWIC* criterion. These sub-corpora correspond to the set of all the text segments in which one of these three *lexical form*, each of which convey the meaning of a concept, appears: 'necessity', 'idealism', and 'paradigm'. The three concepts have been chosen because of the considerable diversity of their statuses: 'necessity' has always been a keyword of several sub-fields of philosophy; 'idealism' refers both to a philosophical current, historically determined, and to an abstract position in philosophy; 'paradigm' entered the philosophical vocabulary in relatively recent times, mainly after the publication of Kuhn, 1962, as a technical term in the philosophy of science. We obtain a set of 719 documents for 'necessity', 450 documents for 'idealism', 975 documents for 'paradigm'.

2.3. Word-sense disambiguation process

For each sub-corpus, we identify the *semantic patterns* (usually, word co-occurrence patterns) associated to each lexical form, so as to discover the most relevant *semantic structures* of that concept. This is done by using *clustering*, a common method in *Machine Learning* for pattern recognition tasks (Aggarwal and Zhai, 2012). Clustering techniques applied to texts are based on two hypotheses: a *contiguity hypothesis* and a *cluster hypothesis*. The former states that texts belonging to the same cluster form a contiguous region that is quite clearly distinct from other regions, while the latter says that texts belonging to the same cluster have similar semantic content (Manning et al., 2009, p. 289 and 350). For our purposes, clustering is an instrument for semantic disambiguation. In our experiment, we use the *K-means* algorithm (Jain, 2010, p. 50), a widely employed algorithm for *Word-Sense Disambiguation* tasks (Pal and Saha, 2015).

The main parameter that needs to be tuned in the K-means algorithm is the k

parameter, which determines the number of *centroids* to be initialized. Each execution of the K-means algorithm generates a partition P_k having a number of clusters equal to k. Since each centroid is the "center vector" of each cluster, it can also be used to identify the most "prototypical" documents in a given cluster. To complete this operation, a tool generally used to select relevant documents in *Information Retrieval* is employed, that is, the *cosine computation* among a *query* vector and a group of "document vectors" (Manning et al., 2009). In this context, each centroid of a P_k partition can be used as a query in order to identify documents with a higher cosine value.
Clustering has first been applied *synchronously* on the S_i matrices with k = {2, 3, 4, …, 50}, thus obtaining the most recurring semantic patterns; then it has been applied *diachronically*, dividing each matrix into three different periods (1981-1993, 1994-2003, 2004-2015) in order to obtain sets of documents with similar cardinality. On each sub-matrix of S_i several clusterings with k = {2, 3, 4, ..., 50} were performed, in order to identify the temporal evolution of the most important semantic patterns associated to the three concepts under study. For each generated P_k partition, we also perform the cosine computation in order to obtain a set of the most relevant PhD dissertations belonging to each cluster.

3. Analyses

In this section, we are going to present three analyses, focusing on three different concepts: paradigm, necessity and idealism. Each case illustrates a different kind of historical-philosophical result.

3.1. Necessity

After exploring both synchronically and diachronically several clusters (with different k) associated to the concept of necessity, we have focused on a clustering with k=18 in the period 1981-2015 (the clusters are not significantly different from one another in the three decades). It turns out that there are at least 16 clearly distinct and philosophically interesting meanings of 'necessity': two (maybe distinct) theological notions; physical necessity; political necessity; necessity as investigated in modal logic and possible world semantics; moral necessity; necessity as opposed to freedom in debates over determinism; the necessity of historical processes; metaphysical necessity; two notions of causal necessity (attacked by Hume); the necessity of life events; logical necessity; phenomenological necessity; necessity of the Absolute (Hegel); necessity of moral duty (Kant); ancient concept of necessity; the necessity of law. In addition to these, there is also a rather big cluster in which 'necessity' seems to occur mainly with its ordinary, not strictly philosophical meaning.

If the clustering we applied to 'necessity' were extended to a large number of philosophical words (chosen in our corpus by domain experts), that would be the first step for the construction of a bottom-up vocabulary of philosophy, and ultimately of a data-driven philosophical dictionary, in which the different (though related) meanings of philosophical terms would be determined on the basis of actual use, rather than merely on the lexicographer's discernment. This lexicographic work is also an indispensable step if one wants to overcome the "concordance approach": it seems to us that this bottom-up lexicography could be a promising starting point for the construction of semantic networks.

3.2. Idealism

Unlike 'necessity', the term 'idealism' has different distributions in the decades 1981-1993, 1994-2003 and 2004-2015. We have only considered the largest clusters (> 10 documents), since for our purpose (that of reconstructing the main historical developments of American academic philosophy), isolated cases and minor tendencies are not relevant.

The evolution of some clusters over decades suggests interesting historical reflections. First, the cluster "Kant" is persistently important. In fact, it becomes more and more important, even in wider contexts, that is, in documents that are not directly devoted to Kant. This is shown by the rising trend of the cluster "Transcendental" (a term typically, but not always directly connected with Kant). Second, the cluster "Hegel" disappears in the second decade, then it reappears: is this a real phenomenon, rather than a statistical artefact? How can it be explained? Third, the cluster "Realism" disappears in the third decade: is there a relationship between the return of "Hegel" and the disappearance of "Realism"? This is not the kind of question, which comes naturally to the mind of the historian of philosophy, on the basis on his/her knowledge of well-known developments of the history of recent American philosophy. This hypothesis can be formulated only thanks to some sort of defamiliarization (*ostranenie*) with respect to the received views in history of philosophy. Yet, it seems unlikely that philosophers in the last decade gave up speaking of realism. The received view may after all be correct, that realism is more and more central in late analytic philosophy (think, for example, of the centrality of David Lewis) (Bonino and Tripodi forthcoming). Such a view is confirmed by other data, such as the number of occurrences of 'realis-' in the abstracts of the corpus. 1981-93: 373 (5,76% of 6,471); 1994-2003: 465 (6,31% of 7,361); 2004-2015: 482 (5,6% of 8,585). Thus the focus on realism is still there, in the third decade. One is therefore led to formulate an alternative hypothesis: philosophers ceased to speak of idealism *in relation to realism*: perhaps the contrast realism-

idealism has become less important than many used to think; perhaps after Dummett, realism is contrasted with anti-realism, rather than with idealism; perhaps some sort of "interference" is here produced by the presence of a further opposition, that between realism and nominalism.
The moral of this example is that clustering applied to large and conceptually sophisticated corpora allows the historians of philosophy to concoct alternative stories to account for the historical facts. This indicates that the data-driven approach can trigger the production of conjectures one would not think about. It is usually maintained that statistical techniques are useful in that they restrict the space of possible interpretations (Mitchell, 1997), but in other cases, such as the one described in this section, at least in an early phase of the hermeneutic process, in virtue of their defamiliarizing impact they can also have the opposite effect: that of broadening that same space and discovering *nouveaux observables* (Rastier, 2011).

3.2. Paradigm

This case study deals with the term 'paradigm' in the period 1981-2015. After exploring several k in the three decades, we focus on the synchronic analysis of the set of clusters with k=16. The first result that immediately stands out is that 'paradigm' occurs rather often: 995 documents, twice as many as 'idealism' (450), and considerably more than 'necessity' (719), a concept which is widely regarded as central in the recent history of Anglo-American philosophy. Using Google Ngram Viewer, and thus taking into account a generalist, non disciplinary corpus, it turns out that such a high frequency is peculiar to the philosophical discourse (the lowest value of 'necessity' is 0.0025%, which is higher than the highest value for 'paradigm', which is 0.0016%).
Why does 'paradigm' occur so frequently? On the one hand, one could find this datum not so surprising, since 'paradigm' is a technical term in the philosophy of science, introduced by Kuhn, 1962 to refer to a set of methodological and metaphysical assumptions, examples, problems and solutions, a vocabulary, which are taken for granted, in a given period of normal science, by a scientific community. On the other hand, moving from a priori considerations to the examination of the data, a partly different landscape emerges: 'paradigm' seems to be a fashionable concept, which is used in a variety of contexts as a term that is neither technical nor simply ordinary. Only in cluster 8 has the term a straightforward technical use, derived from Kuhn's philosophy of science. Each of the other clusters (1: theology, 2: music, 3: philosophy of law, 4: education; 5: nursing; 6: philosophy of religion; 7: moral philosophy; 9: bioethics, 10: spiritualism; 11: political theory; 12: self narrative; 13: theology; 14: Kant-Leibniz; 15

aesthetics; 16: philosophy and language in Wittgenstein, Heidegger etc.) does not correspond to a different meaning of the term 'paradigm', but simply to the application of the same concept to different fields. In most cases we have to do with non-technical contexts, in which 'paradigm' has neither its original grammatical meaning nor its ordinary, non-philosophical meaning (standard, exemplar). It seems to us that its meaning and use are generic and vague, rather than precise and technical; nonetheless, they evoke Kuhn: a quasi-Kuhnian vocabulary became fashionable; it entered many philosophical discourses, often more "humanistic" than "scientific" in spirit, and much less technical than the philosophy of science.

This case study expresses an especially interesting kind of result obtainable by using TM and NLP techniques to assist research in history of philosophy: it shows how the interpretation of clusters fosters the discovery of terminological fashions as opposed to genuine conceptual developments.

References

Aggarwal C.C., and Zhai C.X. (2012). "A Survey of Text Clustering Algorithms." In *Mining Text Data*, 77–128. Springer.

Allard M. et al. (1963). *Analyse conceptuelle du Coran sur carte perforées.* Mouton.

Bonino G. and Tripodi P. (eds.), *History of Late Analytic Philosophy*, special issue of "Philosophical Inquiries", forthcoming.

Chartrand L., Meunier J.-G. and Pulizzotto D. (2016). CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis. In Mayaffre D. et al. (eds.), *Proceedings of the 13th International conference on statistical analysis of textual data*, vol. I, pp. 85-95.

Danis J. (2012). *L'analyse conceptuelle de textes assistée par ordinateur (LACTAO); une expérimentation appliquée au concept d'évolution dans l'œuvre d'Henri Bergson*. Université du Québec à Montréal (http://www.archipel.uqam.ca/4641/1/M12423.pdf).

Ding X. (2013). A text mining approach to studying Matsushita's management thought. *Proceedings of the 5th International conference on informatin, process and knowledge*, pp. 36-39.

Estève R. (2008). Une approche lexicométrique de la durée bergsonienne. *Actes des journées de la linguistique de corpus*, vol. 3: 247-258.

Jain A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, vol. 31(8): 651-666.

Kuhn T.S. (1962). *The structure of scientific revolutions.* University of Chicago Press.

Le N.T, Meunier J.-G., Chartrand L. et al. (2016). Nouvelle méthode d'analyse syntactico-sémantique profonde dans la lecture et l'analyse de textes

assistéespar ordinateur (LATAO). In Mayaffre D., et al. (eds.), *Proceedings of the 13th International conference on statistical analysis of textual data.*

Manning C.D. at al. (2009). *Introduction to Information Retrieval*. Online edition. Cambridge, UK: Cambridge University Press.

McKinnon A. (1973). The conquest of fate in Kierkegaard. *CIRPHO*, 1(1): 45-58.

Meunier J.-G. and Forest D. (2005). Classification and categorization in computer assisted reading and analysis of texts. In Cohen H. and Lefebvre C. (eds.), *Handbook of categorization in cognitive science*, pp. 955-978. Elsevier.

Meunier J.-G. and Forest D. (2009). Lecture et analyse conceptuelle assistée par ordinateur: premières expériences. In *Annotation automatique et recherche d'informations*. Hermes.

Mitchell T.M. (1997). *Machine learning*. McGraw-Hill.

Moretti F. (2005). *Graphs, maps, trees. Abstract models for a literary history*. Verso.

Moretti F. (2013). *Distant reading*. Verso.

Pal A.R. and Saha D. (2015). Word sense disambiguation: A survey. *International Journal of Control Theory and Computer Modeling*, vol. 5(3).

Pincemin B. (2007). Concordances et concordanciers: de l'art du bon KWAC. *XVIIe Colloque d'Albi. Langages et signification – Corpus en lettres et sciences sociales: des documents numériques à l'interprétation*, pp. 33-42.

Pulizzotto D. et al. (2016). Recherche de "périsegments" dans un contexte d'analyse conceptuelle assistée par ordinateur: le concept d'"esprit" chez Peirce. *JEP-TALN-RECITAL 2016*, vol. 2, pp. 522-531.

Rastier F. (2011), *La mesure et le grain. Sémantique de corpus*. Champion.

Sainte-Marie M. et al. (2010). Reading Darwin between the lines: a computer-assisted analysis of the concept of evolution in the Origin of species. *10th International conference on statystical analysis of textual data.*

Salton G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. NJ: Prentice-Hall, Upper Saddle River.

Schmid H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester; UK.

Schmid H. (1995). "Improvements In Part-of-Speech Tagging With an Application To German." In *Proceedings of the ACL SIGDAT-Workshop*, 47–50.

Slingerland E. et al. (2017). The distant reading of religious texts: A "big data" approach to mind-body concepts in early China. *Journal of the American Academy of Religion*: 1-32.

La classification hiérarchique descendante pour l'analyse des représentations sociales dans une pétition antibilinguisme au Nouveau-Brunswick, Canada

Marc-André Bouchard, Sylvia Kasparian
Université de Moncton – emb1214@umoncton.ca; sylvia.kasparian@umoncton.ca

Abstract
In this article, we apply Jean-Blaise Grize's theoretical framework and Max Reinert's descending hierarchical classification to a corpus composed of comments published as part of a petition against institutional bilingualism in New Brunswick. Using Iramuteq, we point to the lexical worlds which constitute anti-bilingualism arguments.

Résumé
Dans cet article, nous appliquons le cadre théorique développé par Jean-Blaise Grize et la classification hiérarchique descendante de Max Reinert à un corpus constitué de commentaires publiés dans le cadre d'une pétition contre le bilinguisme institutionnel au Nouveau-Brunswick. Utilisant le logiciel Iramuteq, nous dégageons les mondes lexicaux qui constituent l'argumentation anti bilinguisme.
Mots-clés: mondes lexicaux, représentations sociales, schématisation, classification hiérarchique descendante, pétition en ligne

1. Introduction

Toute analyse de discours, comme l'admet Jean-Blaise Grize dans *Logique naturelle et communications* (1998; 144-145), est confrontée au problème de la correspondance entre discours et représentations. Celui-ci serait attribuable notamment à l'importance que donne l'analyse du discours à la situation de communication, un facteur qui complique la relation de correspondance entre ce qu'on dit et ce qu'on pense « vraiment ».

Dans le cadre de cet article, nous proposons d'explorer l'intersection entre analyse de discours et étude des représentations et nous tenterons de montrer que, bien que le problème de la correspondance entre discours et représentations individuelles reste difficile à résoudre, les corpus de pétition en ligne homogénéisent le discours et jouent sur la schématisation que construit le locuteur, de façon à ce que les analyses logométriques puissent

accéder à certaines représentations sociales en jeu. À cet effet, nous aurons recours à la méthode Reinert (une classification hiérarchique descendante originalement popularisée par le logiciel ALCESTE) (1990) implantée dans le logiciel Iramuteq (Ratinaud, 2009), qui consiste à relever les mondes lexicaux d'un corpus. Plusieurs auteurs, dont Max Reinert lui-même, ont déjà établi des liens entre cette méthode et le champ d'étude des représentations sociales (1993; 13). Notre contribution à la conversation sera celle d'appliquer la méthodologie issue de la logométrie et le cadre théorique développé par Grize à un nouveau type de corpus qui gagne en popularité depuis le début du 21e siècle, celui des pétitions en ligne. L'exemple par lequel nous illustrerons notre exposé théorique sera celui de l'analyse, à l'aide d'Iramuteq, des mondes lexicaux d'une pétition en ligne lancée au Nouveau-Brunswick (Canada) en 2013, sur la plate-forme www.change.org, contre l'exigence du bilinguisme comme critère d'emploi dans la fonction publique provinciale.

2. Cadre théorique

Selon Denise Jodelet, on peut définir la représentation sociale comme « une forme de connaissance socialement élaborée et partagée, ayant une visée pratique et concourant à la construction d'une réalité commune à un ensemble social » (1997; 53). Ainsi, comme le remarque Serge Moscovici, leur étude demande des méthodes d'observation plutôt que d'expérimentation étant donné qu'elle se manifeste « comme une "modélisation" de l'objet directement lisible dans, ou inférée de, divers supports linguistiques, comportementaux ou matériels » (*idem*; 61). Bien qu'elle soit forme de connaissance, la représentation se distingue de la connaissance scientifique en ce qu'elle découle de ce que Jean-Blaise Grize nomme la logique naturelle (Grize, 1997; 171-172), donnant ainsi sur un « savoir de sens commun » (Jodelet, 1997; 53). Il faut entendre par « logique naturelle » qu'il est question d'une logique d'ordre logico-discursif, manifestée dans le discours par la schématisation, qui « prend en compte les contenus et non les seules formes de la pensée » (Grize, 1997; 171-172). Selon Grize, la schématisation compte cinq notions articulant son ensemble ainsi :

> [1] Une schématisation est la mise en discours [2] du point de vue qu'un locuteur A [3] se fait – ou a – d'une certaine réalité R. [4] Cette mise en discours est faite pour un interlocuteur, ou un groupe d'interlocuteurs, B [5] dans une situation d'interlocution donnée (*idem*).

Ainsi, Grize propose que toute communication est situation d'interlocution, dans laquelle l'orateur construit une schématisation en fonction de son préconstruit culturel, de ses représentations de l'objet en question, et de sa finalité; cette schématisation est constituée d'images de l'orateur, de l'auditeur et de l'objet dont il s'agit, et elle est ensuite reconstruite par l'auditeur en fonction de ses propres représentations, préconstruit culturel et finalité (Grize, 1993; 7). La schématisation est donc partielle et partiale : « elle est partielle dans la mesure où son auteur n'y fait figurer que ce qu'il juge utile à sa finalité, à l'effet qu'il veut produire; elle est partiale puisqu'il l'aménage de telle façon que B la reçoive » (Grize, 1997; 175). En termes de finalité, selon Patrick Charaudeau, les discours, plus particulièrement ceux de type argumentatif ont une double quête, soit le vraisemblable et l'influence, le succès de celle-ci étant fonction des « représentations socioculturelles partagées par les membres d'un groupe donné au nom de l'expérience ou de la connaissance » (1992; 784). C'est donc dire que, compte tenu de la « double quête » du mode de discours, les représentations d'objets sur lesquelles le locuteur construit sa schématisation sont choisies en raison du partage, supposé par le locuteur, de ces représentations chez le(s) destinataire(s). Dès lors, l'analyse des mondes lexicaux communs à un groupe de locuteurs dans une même situation de communication peut nous donner des indices des représentations sociales que se fait le groupe d'un objet du monde social. En effet, selon Max Reinert, dans un corpus collectif, un monde lexical serait indicateur d'un espace de référence commun à un groupe et « l'indice d'une forme de cohérence liée à l'activité spécifique du sujet-énonciateur » (Reinert, 1993; 13). La méthode de classification hiérarchique descendante (Reinert, 1990) propose une représentation de ces mondes lexicaux (ou thématiques) sous la forme de tableaux de classification obtenus par voie du croisement des unités de contexte (ou segments) et des lexèmes d'un corpus. L'hypothèse à la base de cette méthode est que « dans la mesure où une représentation collective exprime une certaine régularité de structure dans une classe de représentations singulières […] cette régularité est due aux contraintes de ce que nous appelons "un monde" » (Reinert, 1993; 29-30). La prise en compte de la fréquence et de l'environnement des formes d'un corpus permet non seulement de relever les formes lexicales les plus propices à constituer des indices de représentations sociales, mais aussi de définir ces formes lexicales en fonction de leur cotexte.

3. Corpus

Le corpus que nous analysons dans la présente recherche est issu d'une pétition en ligne. Contrairement à la pétition classique, la pétition en ligne permet à ceux qui y apposent leur nom d'y publier, s'ils le désirent, un

commentaire justifiant leur appui au titre et à la description de celle-ci. Celle dont il est question ici, *Stop the hiring discrimination against citizens who speak English only*[1], a été lancée en 2013 au www.change.org. Ses commentaires, en plus d'être signés par leurs auteurs, sont accessibles publiquement sur la page même. Cette particularité du canal de communication, que Contamin (2001) appelle « un paradoxe classique des pétitions », a une incidence sur le destinataire de la mise en discours en ce que ce dernier n'est pas seulement le gouvernement de la province, mais aussi le grand public. Ainsi, les corpus de pétitions en ligne homogénéisent les discours selon le modèle de la communication de Grize. D'abord, le groupe de locuteurs se trouve dans la même situation d'interlocution (monologues, à l'écrit, mode argumentatif) et est invité à partager son point de vue sur une même réalité (en l'occurrence, le bilinguisme institutionnel de la province du Nouveau-Brunswick). Ces mises en discours sont faites pour un public général, et la nature engagée de la pétition fait en sorte que, en théorie du moins, seuls les locuteurs partageant le point de vue énoncé dans le titre sont représentés.

Le point de vue partagé par les intervenants, dans notre corpus, est que l'exigence du bilinguisme anglais-français pour des emplois dans la fonction publique provinciale constitue une discrimination envers les Néo-Brunswickois anglophones, qui sont largement unilingues (moins de 15% de ceux-ci se considèrent bilingues, comparativement à un taux de plus de 70% dans la communauté minoritaire francophone). Ces discours s'inscrivent dans un long débat au sein de la population néo-brunswickoise sur le bilinguisme institutionnel, et historiquement le clivage se fonde sur la base linguistique : les francophones sont en faveur du bilinguisme de l'État et de l'avancement des droits linguistiques, alors que les anglophones y sont plus réticents. En tout, à son terme à la fin de l'année 2013, la pétition *Stop the hiring discrimination against citizens who speak only English* récolte 7758 signatures, pour un total de 2372 commentaires, la longueur de chacun variant d'un mot (« jobs ») à 304 mots, pour une moyenne de 37,66 unités linguistiques par commentaire. Ce corpus compte 4 425 formes différentes représentant un total de 89 338 occurrences. Le corpus nettoyé et uniformisé a été soumis à l'analyse du logiciel Iramuteq qui nous donne le dendrogramme des classes constituant les mondes lexicaux des commentaires présentés dans la section suivante.

[1]https://www.change.org/p/the-government-of-new-brunswick-stop-the-hiring-discrimination-against-citizens-who-speak-only-english

4. Analyse

Les 89 338 occurrences (4425 formes différentes) qui constituent notre corpus sont regroupées en 3492 lemmes, soit 2954 formes actives et 538 formes supplémentaires. Et l'ensemble du corpus est segmenté en un total de 2423 parties constituées d'un nombre plus ou moins égal de formes (en moyenne 36.87 formes par segment). L'analyse de la classification hiérarchique descendante avec Iramuteq produit le graphe présenté dans la Figure 1.

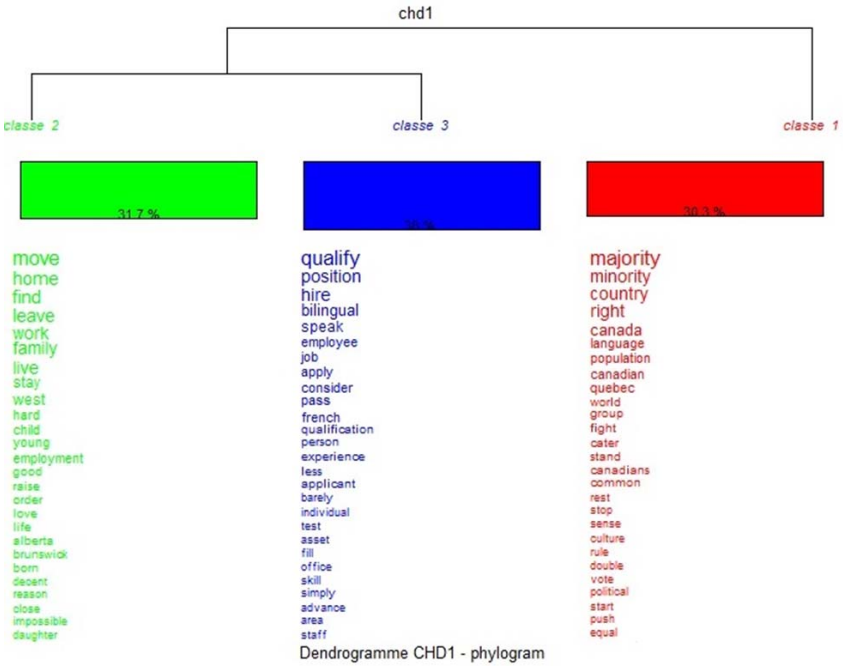


Figure 1 : Classification sur segments de textes simples

La lecture de la Figure 1 révèle que la première segmentation du corpus donne lieu à la Classe 1 (en rouge), formant une classe représentant 30.3 % des segments classés et constituée d'un lexique que nous nommons l'axe sociopolitique : on y aborde d'abord la dynamique « majority » / « minority », qui, à se fier à cette liste de formes, jouerait un rôle d'avant-plan dans les représentations du Canada et des provinces de ce pays. On remarque aussi, en plus de quelques formes relevant de la culture et de la langue, un champ lexical qui semble indiquer la présence de positionnements politiques dans le corpus (« right », « common », « sense », « rule », « vote », « political », « equal »), alors que les verbes (« fight », « cater », « stand », « stop », « start », « push »), de nature politique aussi, renforcent l'hypothèse que cette classe est

constituée de segments exprimant des représentations au sujet de la société canadienne. Une fois la Classe 1 constituée, le calcul divise le deuxième segment en deux classes : la Classe 2 (en vert), contenant 31.7 % de ceux-ci; contre 38 % dans la Classe 3 (en bleu). On observe que, collectivement, celles-ci se démarquent de la Classe 1 par leur lexique relevant de l'expérience personnelle plutôt que de l'opinion politique.

Cette caractéristique personnelle se manifeste dans la Classe 2 par des formes comme « home », « family », « child », « young », et « daughter ». Les verbes, quant à eux, précisent le contexte de cette expérience : « move », « find », « leave », « work », « live », « stay », « raise », « love », et « born »; tout comme quelques adjectifs évaluatifs et/ou axiologiques : « hard », « good », « decent », et « impossible ». On observe aussi quelques formes, en plus de « [new] brunswick », qui réfèrent à une province canadienne, soit à l'Alberta. Le contenu de la Classe 2 constitue donc l'axe biographique, rejoignant souvent le thème de l'exode vers l'Ouest canadien.

La troisième et dernière classe du corpus (en bleu) gravite autour du thème du travail, voire plus précisément de la recherche d'un emploi. C'est aussi dans cette classe qu'on trouve les seules références directes à la langue, mise à part la forme « language » dans la Classe 1 : « bilingual », « speak », et « french ». Certaines formes spécifiques à la Classe 3 laissent entendre que celle-ci est, en partie, plus impersonnelle que la Classe 2 : « employee », « person », « applicant », et « individual ».

À partir de la classification sur segments de texte, on peut parcourir, de façon automatisée, l'ensemble des segments de chaque section et leur attribuer un score selon le nombre de mots représentatifs de la classe où ils se trouvent; on tient aussi compte du degré de représentativité de ces formes.

Ainsi, les deux segments qui suivent sont caractéristiques de la Classe 1: « discrimination of the english[-]speaking white majority populace should stop with the democratic system becoming more in play with majority rules as a true reflection of the people »; « we as a province cannot afford duplicate books in 2 languages to support a minority and the need to speak french in a majority speaking english province to have a job is ridiculous »

Il apparait, dans les segments caractéristiques de la Classe 1, un renversement du rapport de pouvoir classique entre un groupe majoritaire et un groupe minoritaire : les anglophones sont ici opprimés, alors que ce sont les francophones qui sont avantagés, qui ont l'oreille attentive du gouvernement, et, ultimement, qui détiennent le marché du travail bilingue. Cette oppression serait apparente dans la difficulté pour les anglophones unilingues de se trouver un emploi, dans la fonction publique notamment, mais peut-être aussi dans le secteur privé. On remarque d'emblée une représentation de la démocratie se résumant à la règle de majorité (telle que

définie par H. B. Mayo (1957; 50) comme : « the principle that when there is a majority on a matter, then the wishes of the majority should prevail »), ce qui est explicitement communiqué au premier segment caractéristique de la Classe 1. En ce qui concerne la Classe 2, voici deux des segments les plus caractéristiques : « it is very important to me because my daughter like 1000s of other working children here in new brunswick have had to leave their home province in order to find work because they only speak their own language of english. »; et « i have been out of work for over a year. Unable to find a full time job due to bilingualism restrictions. Going to have to move west. ». Il apparait donc qu'il y a un motif récurrent dans la Classe 2 : pour trouver un bon emploi, voire un emploi tout court, il faut être bilingue, faute de quoi on s'exile, notamment dans l'Ouest canadien. On remarque que ces segments témoignent d'un sentiment d'impuissance mais aussi de réticence face à l'idée de quitter sa province natale. Certains segments caractéristiques de la Classe 2 traitent de l'expérience personnelle du commentateur, qui a dû ou qui croit avoir à déménager dans une province non bilingue, alors que d'autres racontent l'exode, accompli ou prévu, de leur(s) enfant(s). On remarque que, dans les segments de la Classe 2 qui précèdent, on attribue volontiers la pauvreté du marché de l'emploi pour les anglophones au facteur linguistique. Ensuite, les segments caractéristiques de la Classe 3 sont les suivants: « because this is a problem, i have 17 years' experience and 2 degrees and i can't even apply for the jobs i qualify for because it's mandatory bilingual positions when over 90% of the day is dealing in english, they won't even interview you unless you speak french »; et « the most qualified person for the job is not always hired because they are not bilingual ». Les différentes formes du concept de « qualification », et d'autres qui y sont liées sémantiquement, sont omniprésentes dans ces segments caractéristiques. Il apparait d'emblée qu'on exclut les compétences linguistiques de ce concept. En effet, une personne qui parle seulement l'anglais est présentée comme potentiellement aussi qualifiée, et à l'occasion plus qualifiée, qu'un candidat bilingue à un emploi qui demande le bilinguisme. Le scénario, souvent hypothétique, qui est donné à voir tend à mettre en jeu une personne unilingue qui serait plus qualifiée qu'une autre chez qui le bilinguisme est présenté comme le seul atout.

5. Conclusion

En somme, dans le cadre de cette pétition, les locuteurs ont mis en discours des représentations du bilinguisme institutionnel au Nouveau-Brunswick par l'entremise de trois mondes lexicaux, présentant ainsi trois facettes de la discrimination perçue envers les anglophones dans la fonction publique. Le premier monde lexical est sociopolitique et énonce des principes généraux

sur ce qui est juste; le deuxième est biographique et relate les effets personnels de cette discrimination; et le troisième porte sur des exemples de la façon dont se manifeste cette discrimination dans le monde du travail. Ainsi, l'échantillon des représentations sociales du bilinguisme institutionnel constituant notre corpus donne à voir un lien de causalité entre l'exigence du bilinguisme pour certains emplois et les difficultés du marché du travail de la province. Dans le but de convaincre un public général, ce point de vue est présenté sous un angle à la fois idéologique, personnel ou pratique, renvoyant ainsi à certaines images de la démocratie, de l'exode et de la compétence; images qui, bien que relativement homogènes dans notre corpus, ne seraient pas nécessairement partagées dans les représentations sociales des anglophones bilingues et des francophones.

Bibliographie

Charaudeau, Patrick (1992). *Grammaire du sens et de l'expression*. Hachette.

Contamin, J.-G. (2001). *Contribution à une sociologie des usages pluriels des forms de mobilization : l'exemple de la petition en France*. Thèse de doctorat de l'Université Paris 1.

Grize, Jean-Blaise (1998). *Logique naturelle et communications*. Presses Universitaires de France.

Jodelet, Denise (1997). Les représentations sociales. Dans Jodelet ed. *Les representations sociales* (5e ed.). Presses Universitaires de France.

Mayo, H. B. (1957). Majority Rule and the Constitution in Canada and the United States. *Political Research Quarterly*, vol. 10(1) : 49-62

Ratinaud, Pierre (2009). *Iramuteq : interface de R pour les analyses multidimensionnelles de textes et de questionnaires*. http://www.iramuteq.org.

Reinert, Max (1990). Alceste une méthodologie d'analyse des données textuelles et une application. *Bulletin de Méthodologie Sociologique*, vol. 26(1): 24-54

Reinert, Max (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, vol. 66(1) : 5-39

Reinert, Max (1997). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage et société*, no. 121/122 : 189-202

Analysing occupational safety culture through mass media monitoring

Livia Celardo[1], Rita Vallerotonda[2], Daniele De Santis[2],
Claudio Scarici[2], Antonio Leva[2]
[1] Sapienza University of Rome
[2] INAIL Research – Headquarters for Research of the Italian National Institute for Insurance against Accidents at Work

Abstract 1
In the last years, a group of researchers within the National Institute for Insurance against Accidents at Work (INAIL) has launched a pilot project about mass media monitoring in order to find out how the press deal with the culture of safety and health at work. To monitor mass media, the Institute has created a relational database of news concerning occupational injuries and diseases, that was filled with information obtained from the newspaper articles about work-related accidents and incidents, including the text itself of the articles. In keeping with that, the ultimate objective is to identify the major lines for awareness-raising actions on safety and health at work. In a first phase of this project, 1,858 news articles regarding 580 different accidents were collected; for each injury, not only the news texts but also several variables were identified. Our hypothesis is that, for different kind of accidents, a different language is used by journalists to narrate the events. To verify it, a text clustering procedure is implemented on the articles, together with a Lexical Correspondence Analysis; our purpose is to find language distinctions connected to groups of similar injuries. The identification of various ways in reporting the events, in fact, could provide new elements to describe safety knowledge, also establishing collaborations with journalists in order to enhance the communication and raise people attention toward workers' safety.

Abstract 2
Negli ultimi anni un gruppo di ricercatori all'interno dell'Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro e le malattie professionali (INAIL) ha lanciato un progetto pilota riguardante il monitoraggio dei mass media con lo scopo di analizzare come la stampa tratta la salute e la sicurezza sul lavoro. A tal fine, l'Istituto ha istituito un database relazionale delle notizie riguardanti gli infortuni e le malattie, incluso il testo stesso delle notizie. L'obiettivo finale del progetto è dunque quello di identificare le direttrici principali su cui muoversi per azioni di sensibilizzazione su salute e

sicurezza sul lavoro. Nella prima fase del progetto, 1,858 articoli di giornale riguardanti 580 infortuni sono stati raccolti; per ogni evento, non solo il testo della notizia ma anche diverse variabili sono state individuate. La nostra ipotesi è che per diversi tipi di infortunio un diverso linguaggio viene usato dai giornalisti per narrare l'accaduto. Per verificare ciò, una procedura di Text Clustering è stata implementata sugli articoli, insieme ad una Analisi delle Corrispondenze Lessicali; il nostro obiettivo è quello di individuare delle differenze nel linguaggio in relazione a diversi gruppi di infortuni. L'identificazione di diversità nel modo in cui viene riportata la notizia al lettore può fornire nuovi elementi per descrivere la cultura della sicurezza, al fine di instaurare delle collaborazioni con i giornalisti stessi per rendere migliore la comunicazione e accrescere l'attenzione del cittadino verso la sicurezza del lavoratore.
Keywords: Occupational safety; Work-related accident; Text mining; Mass media.

1. Introduction

The study described here grew out of the collaboration between the Department of Social Sciences and Economics of Sapienza University of Rome and the Headquarters for Research of INAIL (Italian National Institute for Insurance against Accidents at Work) where, since 2012 a team of researchers has developed the idea of monitoring the mass media in view of prevention against accidents at work (INAIL, 2015).

With this in mind, those researchers achieved the so-called "Repertorio Notizie SSL" (*News Repository on Occupational Safety and Health*), that is a relational database of media news related to occupational injuries and diseases. The objective of this project is to observe the culture of occupational safety and health communicated by mass media agencies in order to identify new elements for increasing prevention against accidents at work. In this study we focus on the hypothesis that there are some asymmetries in the language used to describe the injuries depending on the characteristics of the event. To test it, we performed on the repository data some Automatic Text Analysis procedures.

The article is structured as follow: in section no.2, the News Repository is presented; in section no.3, data are presented and the methodology is exposed; in section no.4, the results of the analyses are shown; in section no.5, conclusions are drawn.

2. The tool

News Repository on Occupational safety and health (NeRO) is a tool created to allow analyses of news contents and texts related to occupational diseases

and injuries. In fact, our strategic objective is to increase public awareness and safety culture through a different approach, which will be also based on the study of news articles, their composition and communication dynamics. So, the first operational purpose is to understand:

- which kind of terms are used in news articles about accidents at work or occupational diseases;
- what inspires a title;
- how the same news is treated by different sources/media;
- how the news text could be interpreted in different ways due to who communicates the news itself;
- whether or not some specific aspects of the events are considered by media.

Our study plans to analyze the cultural characteristics of mass media communication regarding occupational safety and health (OSH), observing the attitude of mass media (and journalists) towards the subject and the way users perceive the news depending on which words are used. As mentioned before, NeRO is an *ad hoc* relational database, centred on the gathering of newspaper articles regarding accidents at work, but it is also arranged to gather news on near misses, occupational diseases and incidents from all kind of sources (press, television or radio). It involves several digital interconnected *tables*, which contain structured – i.e. based on appropriate classifications – and unstructured – i.e. textual – information. Information retrieval regards events happened in Italy and it could contain both online and directly consulting newspapers, since we exploited Google Alert Service (using some suitable keywords) and a daily-newspaper subscription ("la Repubblica"). The reference unit is the event (right now, we are restricting events to accidents) and different aspects and information are linked to it: one or more articles about it, one or more workers injured, and so on. The data-entry interface consists of a series of thematic screens, starting from the opening one, which covers the list of already recorded events. These screens allow to enter the following data, step by step:

- [Screen "Event"] Text containing event description, date of the event, venue, company where accident occurred (if appropriate), economic activity;
- [Screens "News"] Texts of each article related to the event, newspaper name (or press affiliation), news title, web url, date of the article;
- [Screens "Worker" and Sub-screens "Accident" and "Harms, disorders or diseases"] Injured worker's biographical data, information about accident, type of injury, physical implication or resulting disease.

3. Methodology and data

The repository, at the end of data collection, was composed of 1,858 news, related to almost six hundreds different accidents. In order to analyse the content of the news texts in connection with the characteristics of the different events, we performed a content analysis using the Reinert's method (Reinert, 1983) for a descendant hierarchical partition. This algorithm, starting from the co-occurences matrix, generates groups of lexical units – i.e. words – that more co-occur in the texts. Then, the lexical groups were projected on the factorial axes, together with the variables modalities, using the Lexical Correspondence Analysis (Lebart, Salem and Berry, 1997); in this way, we could observe how the language is connected to the accidents features. Finally, to better understand the differences between news texts we analysed the specificities related to the modalities of the variables.

4. Main results and discussion

The cluster analysis made on news texts using the Reinert's method– choosing as segments the articles – produced three lexical groups (in order, the red, the blue and the green ones, in Figure 1):

- Cluster 1 (56.5%): in this group are included the words related to the description of the events, in terms of what happened;
- Cluster 2 (26.5%): here we have the terms connected to the road accidents;
- Cluster 3 (17%): this group is about the emotional aspects connected to the events.

We projected the lexical groups (Figure 1) and the modalities of the variables related to the events (Figure 2) on the first two factors obtained using the lexical correspondence analysis.

As shown in the figure no. 2, there are some interesting characterizations of the language used in newspapers. Some variables, like the economic activity and the accident site, present a strong lexical differentiation among the modalities; this means that who is narrating the event - i.e. the journalist - uses a specific language to describe the accident, on the basis of these characteristics. The other variables presented no particular specificities, except for the one related to the mortality of the accident. In fact, as shown in the figure no. 2, on the second factor the variable "accident mortality" is best represented because of the position and the distance of the modalities "yes" and "no" from the origin. To better understand the lexical differences, we analysed also the specificities (Bolasco and De Mauro, 2013; Lafon, 1980; Lebart, Salem and Berry, 1997) for this particular variable.

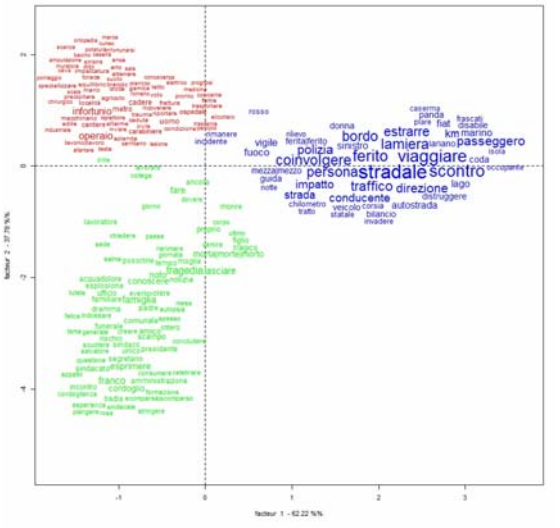


Figure 1 Lexical groups

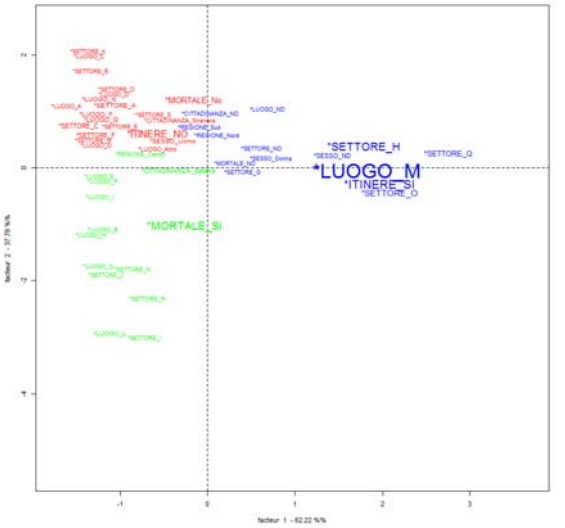


Figure 2 Lexical correspondence analysis

Starting from the results showed in table no.1, we can observe that there is a significant difference in the language utilized when the accident is fatal or not. The terms used in the case of a non-fatal event are related to the description of the injury, while in the case of a mortal accident the situation is completely different: the words utilized refer to the emotional sphere of the event, so concepts like the family or the unpredictability are very often used to describe what was happened.

Table 1 Analysis of the specificities – Variable: "accident mortality"

Fatal accident - No	*z = test-value*	Fatal accident - Yes	*z = test-value*
Hospital	59.17	Tragedy	35.68
Serious	58.84	Family	27.17
To transfer	54.90	Useless	23.62
Dangerous	28.38	To leave	19.84
Rescue	24.13	Victim	18.68
Ambulance	24.09	Tragic	17.71
Leg	23.12	Friend	14.95
Injury	22.06	Band	14.89
Trauma	20.55	Condolence	12.65
Hand	18.84	Province	12.15
Fracture	16.70	Son	11.49
Helicopter	13.70	Wife	11.48
Bus	12.23	Escape	10.63
Crossroad	10.20	Mayor	9.11

5. Conclusions

The project here presented showed how News Repository on OSH (NeRO) can contribute to analyse occupational safety and health, although in some institutions there are already databases dedicated to newspaper articles dealing with OSH. Actually, in addition to news texts, NeRO provides several systematized information, enabling to filter news according to various search criteria and, above all, to carry out a number of studies and organized analysis on textual data, too. In this paper, we showed one of the study we implemented on Repository data using Automatic Text Analysis. The results revealed that a large amount of information is contained within these data; anyway, some information asymmetries are present. For that reason, it will be essential to set up a discussion with a network of journalists and other experts, in order to improve and enhance the media communication. The challenge is to get out from the inner circle of prevention practitioners and build a bridge that could connect the Institution to a more general public, also contemplating liaison organizations (such as trade unions and employers' associations).

References

Bolasco S. and De Mauro T. (2013). *L'analisi automatica dei testi: fare ricerca con il text mining*. Carocci Editore.

Iezzi D. F. (2012). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*(16-17), 3179-3197.

INAIL. (2015). *Il monitoraggio dei mass media in materia di salute e sicurezza: Strumenti per la raccolta e l'analisi delle informazioni*.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un

corpus. *Mots*, *1*(1), 127-165.

Lebart L., Salem A. and Berry L. (1997). *Exploring textual data*(Vol. 4). Springer Science & Business Media.

Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, *8*(2), 187-198.

Is the educational culture in Italian Universities effective? A case study

Barbara Cordella, Francesca Greco, Paolo Meoli,
Vittorio Palermo, Massimo Grasso
Sapienza University of Rome – barbara.cordella@uniroma1.it; francesca.greco@uniroma1.it;
paolomeoli3@libero.it; vittorio.palermo2511@gmail.com; massimo.grasso@uniroma1.it

Abstract 1
The paper explores the professors and students' representation of professional training in Clinical Psychology in the faculty of Medicine and Psychology of the Sapienza University of Rome in order to understand whether the educational context supports students in developing their ability to enter the job market. To this aim, an Emotional Text Mining of the interviews of 30 students and 17 teachers of the Clinical Psychology Master of Science was performed. Both corpora underwent the analysis procedure performed with T-Lab, i.e. a cluster analysis with a bisecting *k*-means algorithm followed by a correspondence analysis on the keyword per cluster matrix, and the results were compared. The results show 4 clusters and 3 factors for each corpus, highlighting a relationship between student and professor representations. Both of them split the training process, distinguishing the educational process from the professional one. The emotional text mining of the interviews turned out to be an enlightening tool letting their latent dimensions emerge, setting the process and outcome of the academic training, and it proved to be very useful for educational purposes.

Abstract 2
La ricerca ha esplorato la rappresentazione della formazione in Psicologia Clinica dei professori e degli studenti della facoltà di Medicina e Psicologia della Sapienza Università di Roma al fine di comprendere se il contesto formativo supporti gli studenti nello sviluppo di competenze utili all'inserimento nel mercato del lavoro. A questo scopo è stata effettuata un'Emotinal Text Mining delle interviste di 30 studenti e di 17 professori del Corso di Laurea Magistrale in Psicologia Clinica con T-Lab (analisi dei cluster con algoritmo bisecting k-means seguita da un'analisi delle corrispondenze sulla matrice cluster per parole-chiave). I risultati mostrano 4 cluster e 3 fattori in entrambi i corpora, evidenziando una relazione tra le rappresentazioni degli studenti con quelle dei professori per quanto concerne il processo di apprendimento, distinguendo e mantenendo separati gli aspetti formativi da quelli professionali. L'Emotional Text Mining risulta essere uno

strumento utile ad evidenziare le dimensioni latenti che organizzano il processo e i risultati dell'apprendimento accademico.
Keywords: Education, Clinical Psychology, Job Market, Youth Unemployment, Emotional Text Mining.

1. Introduction

The problem of youth unemployment is relevant nowadays. In Italy, 25% of young people under 30 years of age are unemployed and this percentage grows to 40% for under 25s (Mckinsey & Company, 2014). But why is this percentage so high? According to Mckinsey's study (*ibidem*), it shows that the figure of 40% for youth unemployment does not rely on the economic cycle but on "structural causes". Among other causes, education is one of the relevant factors of youth unemployment, and is a protection factor for poverty and quality of life, as stated by ISTAT (2017). Graduates are less likely to become poor although the employability and the wages depend on the type of degree. 80% of young graduates in psychology are employed after four years (Anpal Servizi, 2017). Psychologists are more likely to become entrepreneurs than employees. Most probably, the length of time needed to get into the job market is connected to the mismatch between the educational system and enterprise (McKinsey & Company, 2014). Young people's skills are considered appropriate by 70% of Schools and Universities, but only by 42% of employers. The effectiveness of education depends in part on the representation of the professional training characterizing the University. Several studies were performed in order to investigate students' representation in the Psychology Faculty in order to improve the training process (e.g., Carli et al., 2004; Paniccia et al., 2009). Due to the change in the educational plan that took place over the past decade, this study aims to understand whether the present educational context supports students in developing their ability to enter the job market, performing an emotional text mining (Cordella et al., 2014; Greco, 2016) of the interviews of students and teachers of the Master Degree in Clinical Psychology at the Sapienza University of Rome.

2. Methodology

We know that a person's behaviour depends not only on their rationale thinking but also, and sometimes most of all, on their emotional and social way of mental functioning (Carli, 1990; Moscovici, 2005). Namely, people consciously categorize reality and, at the same time, unconsciously symbolize it emotionally (Fornari, 1976). These two thinking processes are the product of the double-logic way of the functioning of the mind (Matte Blanco, 1981) which allows people to adapt to their social environment. According to this

socio-constructivist approach, based on a psychodynamic model, the unconscious processes are social, as people generate interactively and share the same emotional meanings. The socially shared emotional symbolization sets the interactions, behaviours, attitudes, expectations and communication processes, and for this reason, the analysis of the narrations allows for the acquisition of the latent emotional meaning of the text (Salvatore & Freda, 2011). If the conscious process sets the manifest content of the narration, namely *what* is narrated, the unconscious process can be inferred through *how* it is narrated, that is to say, the words chosen to narrate and their association within the text. We consider that people emotionally symbolize an event, or an object, and socially share this symbolisation. The words they choose to talk about this event, or object, is the product of the socially-shared unconscious symbolization (Greco, 2016). According to this, it is possible to detect the associative links between the words to infer the symbolic matrix determining the coexistence of these terms in the text. To this aim, we performed a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The interpretation of the cluster analysis results allows for the identification of the elements characterizing the emotional representation of education, while the results of correspondence analysis reflect its emotional symbolization (Cordella et al., 2014; Greco, 2016). The advantage connected with this approach is to interpret the factorial space according to words polarization, thus identifying the emotional categories that generate professional training representations, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space.

3. Data collection and analysis

In order to explore the emotional representation of the education in the Master of Science in Clinical Psychology, we interviewed 30 students (13% of students) and 17 teachers (71% of teachers) of the Sapienza University of Rome accordingly to their voluntary participation. We used an open-questions interview for students and teachers. Students' interviews resulted in a medium size corpus of 57.387 tokens, and teachers' interviews resulted in a small size corpus of 28.746 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the hapax percentage ($TTR_{students}$ = 0,09; $Hapax_{students}$ = 50,3%; $TTR_{teachers}$ = 0,147; $Hapax_{teachers}$ = 53,8%). According to the size of the corpus, both lexical indicators highlight its richness and indicate the possibility to proceed with the analysis. First, data were cleaned and pre-processed by the software T-Lab (Lancia, 2017) and keywords were selected.

Due to the size of the corpus and the hapax percentage, in order to choose the keywords, we used the selection criteria proposed by Greco (Cordella et al., 2014; Greco, 2016). In particular, we used stem as keywords instead of type, filtering out the lemma of the open-questions of the interviews. Then, on the context units per keywords matrix, we performed a cluster analysis with a bisecting *k*-means algorithm (Savaresi et Boley, 2004) limited to ten partitions, excluding all the context units that did not have at least two keywords co-occurrence. The eta squared value was used to evaluate and choose the optimal solution. To finalize the analysis, a correspondence analysis on the keywords per clusters matrix was made (Lebart et Salem, 1994) in order to explore the relationship between clusters, and to identify the emotional categories setting professional training representations both for students and teachers.

4. Main results and discussion

The results of the cluster analysis show that the keywords selected allow the classification on an average of 96% for both corpuses. The eta squared values was calculated on partitions from 3 to 9, and they show that the optimal solution is four clusters for both corpora. The correspondence analysis detected three latent dimensions. In table 1 and 2, we can appreciate the emotional map of the professional training emerging from the interviews of the teachers and the students and cluster location in the factorial space.

Table 1 – Cluster coordinates on factors of the teachers' corpus (the percentage of explained inertia is reported between brackets above each factor)

Cluster (CU in Cl %)		Factor 1 1 (26,53%) Motivation	Factor 2 (19,03%) Outcome	Factor 3 (14,56%) Role
1	Training Group (22,3%)	Group -0,21	Competence 0,51	Teacher -0,50
2	Clinical Training (33,7%)	Institution 0,33	Competence 0,23	Professional 0,39
3	Institutional Obligations (20,2%)	Institution 0,65	Degree -0,66	Teacher -0,38
4	Student Orientation (23,8%)	Group -0,79	Degree -0,39	Professional 0,16

CU in Cl = context units classified in the cluster.

The teachers' corpus first factor (table 1) represents the motivation in teaching, focusing on the group of students and their specific needs or on the Institutional generic scopes; the second factor focuses on the training outcome, the degree or the professional skills; and the third factor reflects the role of the academic professor that could represent oneself as a teacher or a

professional. As regards the students corpus (table 2), the first factor represents the approach to university experience, which can be perceived as an individual experience or a social one (relational); the second factor explains how students experience vocational training, perceiving it as the fulfilment of obligations or the construction of professional skills that requires personal involvement; and the third factor reflects the outcome of the educational training that can focus on professional skills development or on the achievement of qualifications.

Table 2 – Cluster coordinates on factors of the students' corpus (the percentage of explained inertia is reported between brackets above each factor)

Cluster (CU in Cl %)		Factor 1 (23,2%) Approach	Factor 2 (15,3%) Training	Factor 3 (14,0%) Outcome
1	Idealized Product (27,6%)	Individual -0,56	Fulfilment 0,45	Skills -0,43
2	Professional Education (20,8%)	-0,04	Construction -0,63	Skills -0,24
3	Group Identity (26,3)	Relational 0,69	Fulfilment 0,22	-0,01
4	Empty Degree (25,3%)	Individual -0,32	0,01	Qualifications 0,59

CU in Cl = context units classified in the cluster.

Table 3 – Teachers' Cluster (the percentage of context units classified in the cluster is reported between brackets)

Cluster 1 (22,3%) Training Group		Cluster 2 (33,7%) Clinical Training		Cluster 3 (20,2%) Institutional Obligations		Cluster 4 (23,8%) Student Orientation	
keyword	CU	keyword	CU	keyword	CU	keyword	CU
studente	59	psicologia	94	scuola	29	domanda	42
cercare	43	lavoro	81	persona	28	idea	40
corso	43	clinico	54	laurea	19	organizzazione	33
teoria	32	insegnare	36	università	18	aggiungere	32
lezione	21	contesto	29	trovare	17	processo	30
modalità	21	problema	27	specializzazione	16	rispetto	29
organizzazione	20	intervento	27	importante	16	orientare	21
intervento	19	diverso	25	entrare	15	parlare	21
relazione	17	conoscenza	22	scegliere	14	Corso di laurea	20
				percorso	14	Attività	18
modello	16	interno	22			didattiche	

CU = context units classified in the cluster.

The four clusters of both corpuses are of different sizes (tables 1 and 2) and reflect the representations of the professional training (table 3 and 4). Regarding the teachers' corpus (table 3), the first cluster represents the group of students as a tool to teach professional skills, focusing on the group process where relational dynamics are experienced; the second cluster focuses on clinical training, teaching skills marketable in the job market; the third cluster focuses on the teachers' institutional obligations regardless of the students' training needs; and the fourth cluster represents students' orientation as a way to support students in managing their academic training regardless of professional skills. As regards the students' corpus (table 4), in the first cluster the good training involves students' adherence to lesson tasks regardless of critical thinking on the theoretical model proposed; in the second cluster, learning professional skills is strictly connected to the ability to get and respond to market demand; the third cluster reflects the relevance of belonging to a group of colleagues supporting the construction of a professional identity that, unfortunately, seems unconnected to professional skills development; and the fourth cluster represents professional training as a process in which the degree achievement is the main goal, regardless of the job market demand.

Table 4 – Students' Cluster (the percentage of context units classified in the cluster is reported between brackets)

Cluster 1 (27,6%) Idealized Product		Cluster 2 (20,8%) Professional Education		Cluster 3 (26,3) Group Identity		Cluster 4 (25,3%) Empty Degree	
keyword	CU	keyword	CU	keyword	CU	keyword	CU
esperienza	116	pensare	89	scelta	154	vivere	26
triennale	44	esame	71	studiare	153	trovare	85
percorso	43	psicologia	65	frequentare	104	tesi	20
professione	41	seguire	55	rapporto	102	sentire	91
università	37	realtà	55	piacere	98	riuscire	30
possibilità	35	vedere	55	colleghi	97	prendere	33
capire	33	iniziare	53	parlare	74	persone	105
diverso	31	triennale	53	organizzare	68	maniera	23
senso	30	lavoro	44	domanda	55	livello	35
vivere	25	interessante	44	aggiungere	36	laboratorio	18

CU = context units classified in the cluster.

Students and teachers seem to have similar representations of the training process: the academic need of building a network, highlighted by the students' cluster on *group identity*, and the teachers' cluster on *training group* and *student orientation*; the relevance of achieving a qualification, highlighted by the students' cluster on *empty degree* and the teachers' cluster on *institutional obligation*; and the development of professional skills marketable in the job market reflected by the teachers' cluster on *clinical training* and the

students' cluster on *professional education* in line with what it was found by Carli and colleagues (2004) and Paniccia and colleagues (2009) by means of a similar methodology, the emotional textual analysis (Carli et al., 2016). The awareness of the psychological demand of the labour market is an indicator of the professional training process's effectiveness. Nevertheless, students and teachers split the academic achievement from the development of professional skills. This could be a critical aspect, possibly explaining young graduates' difficulty in entering the job market, focusing more on academic context rather than on market demand. As a consequence, during the training process, students do not develop the connection between professional training (what they are learning) and professional skills (what they are going to do in the future).

5. Conclusion

Although the study results could not be generalized, due to the participants' selection criteria and the methodology we used, they highlight professional training representation characteristics, which are the elements influencing the rate of unemployment among young psychologists. Even though it is not possible to quantify the relevance of the characteristics of the representation, the emotional text mining, allowing for the identification of the words association explanatory of the education representation, allows for hypotheses definition and the identification of the resources and the issues pertaining the professional training in a specific context.

The interpretation of the text mining results lets the social unconscious process emerge, setting the education useful to defining the type of psychological intervention able to support the representation transformation toward a more effective training process. In this particular case study, the intervention would aim to develop the connection between professional qualification achievement and the professional skills development, which are currently split.

References

Anpal Servizi (2017), L*'inserimento occupazionale dei laureati in psicologia, dell'università La Sapienza di Roma, Direzione e studi analisi statistica - SAS.*

Carli R. (1990). Il processo di collusione nelle rappresentazioni sociali. *Rivista di Psicologia Clinica*, 4: 282-296.

Carli R., Dolcetti F. and Dolcetti (2004). L'Analisi Emozionale del Testo (AET): un caso di verifica nella formazione professionale. In Purnelle G., Fairon C. and Dister A., editors, *Actes JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelle*s, pp. 250-261.

Carli R., Paniccia R.M., Giovagnoli F., Carbone A. and Bucci F. (2016).

Emotional Textual Analysis. In L. A. Jason and D. S. Glenwick, editors, Handbook of methodological approaches to community-based research: Qualitative, quantitative, and mixed methods. Oxford University Press.

Cordella B., Greco F. and Raso A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In Nee E., Daube M., Valette M. and Fleury S., editors, *Actes JADT 2014 (12es Journées internationales d'Analyse Statistque des Données Textuelles, Paris, France),* pp. 173-184.

Fornari F. (1976). *Simbolo e codice: Dal processo psicoanalitico all'analisi istituzionale.* Feltrinelli.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

ISTAT (2017). *Rapporto annuale 2017*. ISTAT

Lancia F. (2017). *User's Manual : Tools for text analysis. T-Lab version Plus 2017.*

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

Matte Blanco I. (1981). *L'inconscio come insiemi infiniti: Saggio sulla bi-logica.* Einaudi

McKinsey & Company (2014). Studio ergo Lavoro, come facilitare la transizione scuola lavoro per ridurre in modo strutturale la disoccupazione giovanile in italia. Report di Ricerca "Studio ergo Lavoro", McKinsey & Company, https://www.mckinsey.it/file/2785/download?token=a3VfesjU.

Moscovici S. (2005). *Le rappresentazioni sociali.* Il Mulino.

Paniccia R.M., Giovagnoli F., Giuliano S., Terenzi V., Bonavita V., Bucci F., Dolcetti F., Scalabrella F. and Carli R. (2009). Cultura Locale e soddisfazione degli studenti di psicologia. Una indagine sul corso di laurea "intervento clinico" alla Facoltà di Psicologia 1 dell'Università di Roma "Sapienza". *Rivista di Psicologia Clinica, Supplemento n.* 1: 1-49.

Salvatore S. and Freda M. F. (2011). Affect, unconscious and sensemaking: A psychodynamic, semiotic and dialogic model. New Ideas, *Psychology*, Vol. 29, pp. 119–135.

Savaresi S. M. and Boley D. L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis* 8(4): 345-362.

Profiling Elena Ferrante: a Look Beyond Novels

Michele A. Cortelazzo[1], George K. Mikros[2], Arjuna Tuzzi[3]
[1]University of Padova – cortmic@unipd.it
[2]National and Kapodistrian University of Athens – gmikros@isll.uoa.gr
[3]University of Padova – arjuna.tuzzi@unipd.it

Abstract
Elena Ferrante represents rather a peculiar editorial and journalistic phenomenon: Today, she enjoys a wide international audience, though, on the other hand, there is surprisingly little scientific literature that discusses her works. Since Elena Ferrante is a pseudonym for an anonymous writer, some investigators have already dealt with the pursuit of her real identity and, at the moment, the main suspects that emerged are Domenico Starnone, Marcella Marmo and Anita Raja. Corpora collected in order to analyze Elena Ferrante's works and compare them with the works of other authors are usually composed of novels, however Marcella Marmo and Anita Raja are not novelists and their works are not ascribed to genres comparable with novels. One of Elena Ferrante's books, *La Frantumaglia*, is useful to collect corpora of texts of different genres (letters, essays, interviews, etc.) and they might include texts by authors that have never been taken into consideration in research studies based on novelists. Nevertheless, these texts raise specific questions that concern their exploitability in traditional authorship attribution procedures due to their limited size. This study aims at working on a corpus of texts other than novels by means of a machine learning approach, in the frame of methods for authorship attribution and profiling.

Riassunto
Elena Ferrante costituisce un fenomeno editoriale e giornalistico italiano molto particolare: attualmente gode di grande visibilità internazionale ma, allo stesso tempo, c'è sorprendentemente poca letteratura scientifica che si occupa delle sue opere. Siccome Elena Ferrante è lo pseudonimo di un/una autore/autrice ancora anonimo/anonima, alcuni si sono già confrontati con la ricerca della sua vera identità e i maggiori sospettati emersi, finora, sono Domenico Starnone, Marcella Marmo e Anita Raja. I corpora che vengono utilizzati per studiare la produzione di Elena Ferrante e confrontarla con quella di altri autori sono costituiti normalmente da romanzi ma Anita Raja e Marcella Marmo non sono scrittrici e i loro lavori non si possono ascrivere a generi confrontabili con i romanzi. Una delle opere di Elena Ferrante, *La frantumaglia*, può essere utilizzata per costituire corpora con testi di generi

diversi (lettere, saggi, interviste, ecc.) che possono includere materiali di autori non ancora considerati nelle ricerche basate su romanzieri. Tuttavia, questi testi presentano specifiche problematiche legate alla ridotta dimensione e parziale utilizzabilità con strumenti di attribuzione d'autore tradizionali. Questo lavoro ha come obiettivo studiare un corpus di testi diversi dai romanzi con un approccio machine learning nell'ambito dei metodi per l'attribuzione d'autore e il profiling.
Keywords: authorship attribution, machine learning, profiling, stylometry, support vector machine

1. Introduction

In previous works the novels signed by Elena Ferrante have already been studied in the panorama of Italian contemporary literature and they have displayed that this author has a peculiar writing style and shows relevant individual traits. Moreover, in previous investigations the Italian writer that showed the highest level of similarity with Elena Ferrante is Domenico Starnone (Galella, 2005; 2006; Gatto, 2016; Cortelazzo et Tuzzi, 2017; Tuzzi et Cortelazzo, 2018). In this study we aim at testing further hypothesis and look at texts that are not ascribed to the genre "novels". In this way we have the opportunity to consider for authorship attribution and profiling experiments new candidates, i.e. writers that are not exclusively novelists. A first reference can be made to Marcella Marmo and Anita Raja, two Italian women, that have been suspected to be the hand that hides behind the pen-name of Elena Ferrante, respectively, by Marco Santagata (2016) and Claudio Gatti (2016). The corpus collected for this new study has a specific focus on three main suspects (Marcella Marmo, Anita Raja, Domenico Starnone) and includes further suspected authors (Goffredo Fofi, Mario Martone, Valeria Parrella, Francesco Piccolo), authors that in previous analysis showed some common traits with Elena Ferrante's works (Gianrico Carofiglio, Clara Sereni), authors that provocatively claimed to be Elena Ferrante (Laura Buffoni) and members of the E/O publishing house (Sandro Ferri, Sandra Ozzola and the editorial board that is supposed to be the collective editor of the publishers' web pages).

2. Corpus

The corpus includes letters, interviews and further material written by different authors (tab. 1) that can be compared with texts included in the book *La Frantumaglia* by Elena Ferrante (2016). An innovative perspective has been adopted for analyzing texts: a Machine Learning (ML) approach based on a Support Vector Machine (SVM) method that takes into consideration 13 authors for a classical Authorship Attribution (AA) and different variables

(gender, age, geographical area) for profiling tasks.
The whole corpus adopted for this study is composed of 113 texts and includes 143,695 word tokens and 19,020 word types. In the classical ML perspective, the corpus is arranged into two groups: a "training set" and a "testing set". The training corpus (tab. 1) includes 86 texts (87,458 word tokens), 78 written by 12 authors and 8 by a collective subject (EO) that represents the editorial staff of E/O publishing house. The corpus is balanced in terms of gender and partly balanced for age and geographical area (tab. 2). Information about gender and age is not available (n.a.) for E/O, as it is presumed to be a group. The testing corpus includes 27 texts (6 essays, 7 interviews, 14 letters for a total of 56,237 word tokens in size) signed by Elena Ferrante and collected in her book *La Frantumaglia*. Five texts are chapters of the same large essay that has been written as an answer to Giuliana Olivero and Camilla Valletti's questions (Ferrante 2016).

Table 1. Authors and categories of texts included in the training corpus

Authors			Category		
	texts	tokens		texts	tokens
Laura Buffoni	3	4,477	article	53	42,124
Gianrico Carofiglio	6	4,940	essay	9	22,926
E/O	8	3,955	interview	12	15,480
Sandro Ferri	2	3,838	letter	4	1,611
Goffredo Fofi	9	7,378	web	8	5,317
Marcella Marmo	5	12,991			
Mario Martone	10	9,320			
Sandra Ozzola	4	1,879			
Valeria Parrella	7	4,676			
Francesco Piccolo	6	5,529			
Anita Raja	4	13,617			
Clara Sereni	2	2,271			
Domenico Starnone	20	12,587			
Tot	86	87,458	Tot	86	87,458

Since most stylometric measures and linguistic features are heavily influenced from text size, we decided to split our texts into equal sized text chunks. Both the training and the testing corpus were segmented into 200 words text chunks. After the chunking procedure, the training corpus inflated from 86 texts to 386 chunks of 200 words in length and the testing

corpus from 27 texts to 259 chunks of 200 word tokens in length. This enlargement had also the positive effect of making our sample space larger, giving us the opportunity to use a wider spectrum of linguistic features.

Table 2. Descriptive variables of texts included in the training corpus

Gender				Age				Naples Area			
	authors	texts	tokens		authors	texts	tokens		authors	texts	tokens
n.a.	1	8	3,955	n.a.	1	8	3,955				
f	6	25	39,911	>60old	7	46	54,561	Naples	6	52	58,720
m	6	53	43,592	≤60young	5	32	28,942	NoNaples	7	34	28,738
Tot	13	86	87,458	Tot	13	86	87,458	Tot	13	86	87,458

3. Method

In order to investigate our research aims, we developed a feature-rich document representation model comprised by the following features groups:

1) Author Multilevel N-gram Profiles (AMNP): 1,500 features, 500 features of each n-gram category (2-grams and 3-grams at the character level, and 2-grams at the word level);
2) Most Frequent Words in the corpus (MFW, 500 features).

The first feature group (AMNP) provides a robust document representation which is language independent and able to capture various aspects of stylistic textual information. It has been used effectively in authorship attribution problems (Mikros et Perifanos, 2011; 2013) and gender identification focused on bigger texts (e.g. blog posts, cfr. Mikros, 2013). AMNP consists of increasing order n-grams in both character and word level. Since character and word n-grams capture different linguistic entities and function complementary, we constructed a combined profile of 2, 3 characters n-grams and 2 words n-grams. For each n-gram we calculated its normalized frequency in the corpus and included the 500 most frequent entries resulting in a combined vector of 1,500 features. The second feature group (MFW) can be considered classic in the stylometric tradition and it is based on the idea that the MFWs belong to the functional words class and are beyond the conscious control of the author, thus revealing its stylometric finger print. In this study we used the 500 most frequent words of the corpus.
The above described features have been exploited for training a classification machine learning algorithm, Support Vector Machines (SVM, Vapnik, 1995), in both a standard authorship classification task and in three different author profiling tasks (author's gender, age, and geographical area). SVM is considered a state-of-the-art algorithm for text classification tasks. The SVM constructs hyper-planes of the feature space in order to provide a linear solution to the classification problem. For our trials we experimented with

various kernels and we ended up choosing the polynomial one as this was the most accurate in our dataset. All statistical models developed have been evaluated using 10-fold cross validation (90% training set – 10% testing set) and the accuracies reported represent the mean of the accuracies obtained in each fold. Since the feature space was sparse, we eliminated all features that showed a variance close to zero, using the two following rules: the percentage of unique values was less than 20%, and the ratio of the most frequent to the second most frequent value was greater than 20. The near-zero variance feature removal shrank the number of the employed features and led to a reduction of 47.4% (from the initial 2,000 available features we kept 1,052 features).

4. Results

4.1. Authorship Attribution Results

For the standard authorship classification task (tab. 3), first we worked with the whole corpus as training dataset and obtained an accuracy of 0.7098 on average (71%). Among the set of 13 candidates included in the corpus, a large share of testing text chunks resulted attributed to Domenico Starnone (32%), Anita Raja (21%) and Mario Martone (21%).

Table 3. Attribution of text chunks included in the testing corpora (whole and reduced corpus)

whole corpus			reduced corpus		
Authors	No. chunks	%	Authors	No. chunks	%
Starnone	84	32%	Starnone	115	44%
Raja	55	21%	Raja	73	28%
Martone	55	21%	Martone	39	15%
E/O	18	7%	E/O enlarged	32	12%
Buffoni	16	6%			
Parrella	15	6%			
Fofi	7	3%			
Carofiglio	2	1%			
Ferri	2	1%			
Marmo	2	1%			
Piccolo	3	1%			
Ozzola	0	0%			
Tot	259	100%	Tot	259	100%

Table 4. Cross-classification matrix in authorship attribution task (whole and reduced corpus)

	reduced corpus				
whole corpus	Starnone	Raja	Martone	E/O enlarged	Tot
Starnone	**77**	2	0	5	84
Raja	3	**48**	0	4	55
Martone	14	2	**30**	9	55
E/O	1	2	0	**15**	18
Buffoni	6	5	2	3	16
Parrella	8	7	0	0	15
Fofi	4	3	0	0	7
Piccolo	2	0	0	1	3
Carofiglio	0	2	0	0	2
Ferri	0	0	0	**2**	2
Marmo	0	2	0	0	2
Ozzola	0	0	0	0	0
Tot	115	73	32	39	259

We deemed useful to reduce the candidates to Starnone, Raja, Martone and rearrange the E/O collective author into a new enlarged version of the E/O group, i.e. we pool together all the members of the E/O publishing house (Sandro Ferri, Sandra Ozzola and the E/O staff). As an effect of this selection we obtained an improvement in the performance of the ML algorithm (+13%) since the accuracy rose up to 0.8408 on average (84%). With reference to this reduced version of the training corpus, that includes only four candidates, again most text chunks seem to belong to Domenico Starnone (44%) and Anita Raja (28%). From a cross comparison of the results achieved (tab. 4) with the whole and reduced versions of the training corpus we observed that the text chunks of the testing corpus that have been attributed to Domenico Starnone and Anita Raja proved more stable and consistent if compared to a more unstable and weak role of Mario Martone. The existence of an action of the publishing house was confirmed in both versions, although in some cases a confusion of the E/O editors with Starnone and Raja's hands is somewhat visible.

4.2. Profiling Results

Results achieved with profiling tasks are more schematic since the algorithm is called to work with simpler dichotomous variables (tab. 5).

With respect to gender, the ML algorithm obtained an accuracy of 0.8000 on average (80%) and the results achieved with the automatic classification of the text chunks of the testing corpus suggested that among the fragments of *La Frantumaglia* we might have different hands: at least a man (54%) and a woman (46%). If compared with the case of gender profiling, the ML

algorithm achieved a similar performance in terms of accuracy for both the classification by age (0.8027, 80%) and geographical area (0.7850, 78%) but for the most part the text chunks appeared to be written by an old author (76%) from Naples (90%).

Table 5. Profiling of text chunks included in the testing corpus

Gender			Age			Naples area		
	No. chunks	%		No. chunks	%		No. chunks	%
f	141	54%	>60 old	197	76%	Naples	233	90%
m	118	46%	≤60 young	62	24%	NoNaples	26	10%
Tot	259	100%	Tot	259	100%	Tot	259	100%

5. Discussion and conclusions

Among limitations and constrains of this method, first and foremost we have to take into account that we have different genres among the texts of this corpus (essays, interviews, newspapers articles, letters) and this feature surely affects our results. Texts show similarities when they are written by the same author or belong to the same text genre and these two effects are not easy to disentangle in our text corpus. Secondly, when the SVM prediction is called to assign testing chunks to authors and/or categories it always leads to an attribution that is the result of a formula generated by the ML algorithm (in other words it never answers "do not know"). Results depend both on quality of texts and basket of opportunities offered during the training phase. As a consequence, we have to refer to the accuracy of the model and consider the classification as the best attribution among options given by the set of reasonable candidates and available categories. Thirdly, *La Frantumaglia* represents an interesting set of texts signed by Elena Ferrante that are not ascribed to the genre "novels" and it enables new analyses to compare and contrast the author's writing style with the one of authors that are not strictly novelists. Nevertheless, we cannot be sure that all texts included in *La Frantumaglia* are written by the same hand and, moreover, we do not know whether these texts are written by the author that actually wrote also the novels signed by Elena Ferrante. From the authorship attribution viewpoint more than one hand emerged as likely and we can formulate some hypothesis. If we take into account only main suspected authors mentioned in our Introduction, Domenico Starnone and Anita Raja are confirmed; on the contrary, Marcella Marmo seems not believable. Mario Martone's role is an interesting suggestion since similarities of chunks taken from *La Frantumaglia* with his texts might be the indirect outcome of direct interactions between Martone and Ferrante (e.g. letters and interviews where they are both

speaking about the movie *L'amore molesto*). Also the E/O staff's role is engaging as it is easy to imagine the effect on the writing style of one or more editors that work as proofreaders, copyreaders and ghostwriters when Elena Ferrante has to answer many interviews and letters collected by the publishing house. From profiling experiments a composite picture of *La Frantumaglia* emerges. The procedure reveals the existence of different hands once more, suggested the involvement of at least a man and a woman, and draws the portray of an author (single or collective) from Naples that is over 60 years old.
Does the mystery about Elena Ferrante's work remain a mystery?

Acknowledgements
We thank Arianna Menin for providing us with the corpus of texts of *La Frantumaglia* collected for her first level (B.A.) 3-years degree thesis in *Communication* (University of Padova, a.y. 2016/2017, supervisor prof.ssa Arjuna Tuzzi).

References

Cortelazzo M.A. and Tuzzi A. (2017). Sulle tracce di Elena Ferrante: questioni di metodo e primi risultati. In Palumbo, G. (ed), *Testi, corpora, confronti interlinguistici: approcci qualitativi e quantitativi*, EUT – Edizioni Università di Trieste, pp. 11-25.

Ferrante, E. (2016). *La Frantumaglia*. Roma: E/O.

Galella, L. (2005). Ferrante-Starnone. Un amore molesto in via Gemito, *La Stampa*, 16 January 2005, pp. 27.

Galella, L. (2006). Ferrante è Starnone. Parola di computer. *L'Unità*, 23 November 2006.

Gatti, C. (2016). Elena Ferrante, le «tracce» dell'autrice identificata, *Il Sole 24 Ore* – Domenica, 2 October 2016, pp. 1-2.

Gatto, S. (2016). Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce, *Lo Specchio di carta. Osservatorio sul romanzo italiano contemporaneo*, 22 October 2016. www.lospecchiodicarta.it

Mikros, G.K. (2013). Authorship Attribution and Gender Identification in Greek Blogs. In Obradović, I., Kelih, E. and Köhler R. (eds.), *Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012*, Belgrade: Academic Mind, pp. 21-32.

Mikros, G.K. and Perifanos, K. (2011). Authorship identification in large email collections: Experiments using features that belong to different linguistic levels *Proceedings of PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19-*

22 September 2011, Amsterdam.

Mikros, G.K. and Perifanos, K. (2013). Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In Hovy, E., Markman, V., Martell, C. H. and Uthus D. (eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California*. Palo Alto, California: AAAI Press, pp. 17-23.

Santagata M. (2016). Elena Ferrante è …, *La lettura – Corriere della Sera*, 13 March 2016, pp. 2-5.

Tuzzi, A. and Cortelazzo, M.A. (2018), What is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer, *Digital Scholarship in the Humanities* (on line first version).

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Word Embeddings: a Powerful Tool for Innovative Statistics at Istat

Fabrizio De Fausti[1], Massimo De Cubellis[1], Diego Zardetto[1]
[1]ISTAT – Italian National Institute of Statistics
(defausti, decubell, zardetto)@istat.it

Abstract 1
In recent years, word embedding models have proven useful in many Natural Language Processing problems. These models are generated by unsupervised learning algorithms (like Word2Vec and GloVe) trained on very large text corpora. Their main purpose is to map words to vectors of a metric space in a very smart way, so that the resulting numeric representation of input texts effectively captures and preserves a wide range of semantic and syntactic relationships between words. In this paper we discuss word embedding models generated from huge corpora of raw text in Italian language, and we propose an original graph-based methodology to explore, analyze and visualize the structure of the learned embedding spaces.

Abstract 2
Il lavoro illustra le potenzialità dei modelli Word Embedding nell'analisi di grandi collezioni di dati testuali e propone un originale metodo basato sui grafi per l'esplorazione della struttura semantica catturata dai modelli.

Keywords: Word Embeddings, Word2Vec, Graphs, Text Summarization, Italian Tweets, NLP.

1. Introduction

Word embedding models represent a powerful tool that can be used as input for subsequent machine learning tasks, like text classification, topic modeling and document similarity. This work shows how we built, tested and used word embedding models (based on the Word2Vec algorithm, see Section 2.1) to achieve the following objectives:

- Istat is currently collecting streaming Twitter data on a large scale. Word embedding models helped us devise domain-specific 'filters', namely sets of keywords that we used to filter out *off-topic* tweets with respect to the intended statistical production goal. Here we will show the case of the so-

called "Europe filter", meant to measure people's mood about the European Union.

- Istat is currently exploiting textual data automatically scraped from the websites of Italian enterprises in order to *predict* whether or not they perform e-commerce. Given the huge corpus of noisy and unstructured texts derived from this web-scraping procedure, word embedding models allowed us: (i) to automatically create an "e-commerce pseudo-ontology" and to smartly summarize the input texts, (ii) to encode the summarized texts into a rich numeric representation in order to feed a Deep Learning classifier.

2. Methodology

In recent years, new successful algorithms for natural language modeling have been proposed, based on Neural Networks (e.g. Word2Vec and Glove). These algorithms, starting from very large corpora of raw text, are able to create models that map words to low-dimensional vector spaces, called *word embeddings* (Mikolov et al., 2013a). Although these algorithms do not rely on any linguistic domain-knowledge, nor on handcrafted syntactic and semantic relationships between words, they are surprisingly able to *learn* both of them from raw data. Indeed, words that are strongly related from a syntactic and/or semantic point of view are mapped to vectors that are almost parallel to each other; conversely, words that are syntactically and/or semantically loosely related are mapped to nearly perpendicular vectors. Moreover, these models perform amazingly well when it comes to solving *analogies* between words, just like a human would do. For example, if one asks a trained word embedding model «which word *X* completes the analogy: ['Paris' : 'France' = 'Madrid' : *X*]», the answer will very likely be *X* = 'Spain'. We mention here only one type of relationship (capital-nation), but word embedding models are able to capture a wide variety of relationships, such as: male-female, singular-plural, superlative-comparative, synonym-antonym, politician-party, etc.

2.1 Word2Vec

Word2Vec (Mikolov et al., 2013b) is one of the most influential word embedding algorithms. It consists of a neural network trained to solve a predictive problem according to one of the following two approaches: predicting the central word given the other words of a *context* (Cbow), or predicting the words of the *context* given the central word (Skipgram). At the end of the training the predictive ability of the network is not used; instead,

its internal structure (weights of the network) is exploited to represent the coordinates of each word of the dictionary in the embedding space.
While a large text corpus is the main input to Word2Vec, the algorithm allows also for several hyperparameters which can be tuned to improve the quality of the learned model. Some scholars (e.g. Levy et al., 2015) consider these hyperparameters as key points to understand Word2Vec's superiority as compared to previous language modeling techniques.
The main hyperparameters of Word2Vec are:

- *Embedding space dimension*: the dimension of the vector space to which the words of the corpus are mapped;
- *Window size*: the width of the sliding window used to process the corpus. It defines how large the *context* is;
- *Iteration*: how many times the weights of the neural network are updated during training;
- *Learning model*: the approach used to train the neural network, either Cbow or Skipgram.

Of course, further factors affect the performance of a Word2Vec model:

- *Size of the corpus*: bigger corpora perform better than small ones;
- *Quality of the corpus*: very noisy, fragmented and poorly curated texts generally produce lower quality embedding spaces.

At the end of the training phase, the quality of the learned word embedding model can be assessed through standard test functions. Classical examples are the *word-similarity* and the *word-analogy* functions (see e.g. Pennington et al., 2014).

2.2 Exploring and visualizing big embedding models through graphs

As sketched in Section 2, word embedding algorithms transform words into vectors of a low-dimensional metric space. The dimension of this numeric space is usually set to values in the range 100-300 (see e.g. Mikolov et al., 2013a). When input corpora are huge, taking into account inflected forms of words, the output embedding model can contain hundreds of thousands of vectors. As a consequence, the *full* structure of the embedding model is very hard to analyze. Exploration and visualization of such models requires to (i) reduce the dimensionality of the embedding space, and to (ii) focus on just a subset of vectors, namely those derived by the most relevant words for the analysis at hand. While traditional solutions exist for the first task, like PCA and t-SNE (van der Maaten, Hinton, 2008), no standard methods are available for the second one. We propose here a new technique, based on graphs (Gibbons, 1985), that simultaneously addresses both needs. It selects

just a subset of relevant words, adopting a clever filtering criterion based on their semantic proximity, and allows visualizing the resulting *sub-model* in a two-dimensional graph.

2.3 Building the graphs

Given a "node" vector/word **v** in the embedding space, let's define $G_W(\mathbf{v})$ a *base graph* of *width* $W \in \mathbb{N}$. To build $G_W(\mathbf{v})$, we connect **v** to its W nearest vectors/words in the embedding space (the cosine distance is used). The base graph $G_W(\mathbf{v})$ will thus have W + 1 nodes. Node **v** can be either the image of an *actual word* s, i.e. $\mathbf{v} = V(s)$, or the vector resulting from the *sum of multiple words*, say s_1 and s_2, i.e. $\mathbf{v} = V(s_1) + V(s_2)$. The idea is that, within the embedding space, the sum of word vectors can be exploited to *disambiguate* the meaning of *polysemous* words. An example is provided in Table 1, where the 5 closest words to the vector V('rome') are reported on the left panel, and the 5 closest words to the vector V('rome') + V('colosseum') + V('ancient') are reported in the right panel. Evidently, the addition of words 'colosseum' and 'ancient' to the polysemous word 'rome' moves the semantic area explored by the base graph G_5 from a geographical to an historical sense.

Table 1. Word disambiguation by sum of vectors: the polysemous word is 'rome'.

Closets 5 Words from V(rome)	**Cosine Similarity**	**Closest 5 Words from V(rome) + V(colosseum) + V(ancient)**	**Cosine Similarity**
turin	0.6818	roman	0.5822
palermo	0.6377	archeological	0.5318
naples	0.6212	pompei	0.5250
milan	0.6129	trastevere	0.5217
bologna	0.5857	trajan	0.5189

Our approach builds a full output graph by iteratively combining N base graphs G_W. We devised three different methods to combine base graphs according to different exploration strategies. We called these methods Geometric, Linear and Geometric-Oriented: the corresponding pseudo-codes are provided in Table 2. Besides the width parameter W and the number of iterations N, all the three methods require as input a set of *seed words* [seeds] to define the starting point for the exploration of the embedding model.

Table 2 Pseudo codes of the proposed graph generation methods. Function find_leaves() returns all the nodes with zero outdegree; function shortestPath() calculates the shortest path between two nodes.

Geometric ([seeds], N, W)	**Linear ([seeds], N, W)**	**Geometric-Oriented ([seeds], N, W)**

v = V(seed1) + V(seed2) + …
G_w(v)
for iteration in [1, …, N]:
  for leaf in find_leaves():
  G_w(V(leaf))
	v = V(seed1) + V(seed2) + …
	G_w(v)
	for i in [1, …, N]:
	  virtualNode_i = 0
	  for leaf in find_leaves():
	  addEdge(leaf, virtualNode_i)
	  virtualNode_i = virtualNode_i + V(leaf)
	  G_w(virtualNode_i)
		v = V(seed1) + V(seed2) + …
		G_w(v)
		for iteration in [1, …, N]:
		  for leaf in find_leaves()
		  virtualNode_leaf = 0
		  addEdge(leaf, virtualNode_leaf)
		  for node in shortestPath(v, leaf):
		  virtualNode_leaf = virtualNode_leaf + node
		  G_W(virtualNode_leaf)

As will be shown in Section 3, the Geometric method tends to expand the exploration range very quickly, rapidly losing the initial semantic focus provided by the seed words; the Linear method stays much more focused, but explores just a narrow sub-model; the Geometric-Oriented method provides a satisfactory compromise between the previous two methods.

3. Application

3.1 Building word embedding models on large corpora of Italian tweets

Istat is currently collecting streaming Twitter data on a large scale. Italian tweets are captured provided that they pass at least one active 'filter'. Filters are simply sets of keywords deemed to be relevant for specific statistical production goals. For instance, the 'Social Mood on Economy' filter involves 60 keywords borrowed from the questionnaire of the Italian Consumer Confidence Survey, and collects about 40,000 tweets per day.

We used a large collection of about 100 million Italian tweets to train Word2Vec with different settings of hyperparameters, therefore generating different embedding models. We subsequently analyzed the obtained models and tested their quality as discussed in Section 3.1.2. This way we managed to identify the best performing set of hyperparameters to be used for the applications described in Sections 3.2 and 3.3.

3.1.1 Process

The data processing pipeline we implemented consists of the following steps:

- Collection of Italian tweets through Twitter's streaming API as JSON files;

- Parsing of JSON files and storage of the tweets in a relational database;
- Extraction from the database of the textual content of about 100 million tweets and export to a raw text file (corpus);
- Preprocessing of the raw text (text cleaning and normalization);
- Setting of Word2Vec hyperparameters;
- Training of Word2Vec on the tweets' corpus;
- Test of the learned word embedding model.

3.1.2 Benchmark and selection of the best hyperparameters

With the aim of identifying the best hyperparameters, we customized benchmark *word-analogy* tests contributed by the Stanford University (Pennington et al., 2014), translating them in Italian and adding new word analogies involving specific terms of the Economics field. Note that our tests involved many groups of analogies, encoding a wide range of different relationships between words, of both the syntactic and the semantic kind. As a measure of model goodness, we adopted the so called "Top-1 accuracy" criterion. According to this criterion an analogy [a : b = c : x] is successfully solved by the learned model if and only if the closest (i.e. Top-1) embedding vector to V(c) - V(a) + V(b) is *exactly* V(x). We evaluated against our customized word-analogy tests many output models generated by diverse settings of hyperparameters, and eventually found the following optimal values: *embedding space dimension* = 200, *window size* = 8, *iteration* = 15, *learning model* = Cbow.

3.2 Design of the "Europe" filter

As already mentioned in Section 3.1, Istat collects only Italian tweets that match at least one active filter. So far, the keywords defining the filters have been designed by subject-matter experts. In this section, instead, we illustrate how word embedding models can be exploited to *automatically* develop new filters in a data-driven way. The idea is to leverage our graph-based exploration methodology to select the best keywords, starting from few relevant seed words. In particular, on the occasion of the 16th anniversary of the Treaties of Rome, our objective was to capture the sentiment of Italian Twitter users about European Union. In Figures 1 and Figure 2 we show the graphs resulting from the Geometric-Oriented and Geometric methods respectively. Note that both graphs were generated using the same seed words, namely: 'europa', 'ue', 'bruxelles', 'europea', 'unione', 'euro'. The Geometric-Oriented graph appears more compact and the words are indeed closely related to the semantic area of the seed words. The Geometric graph,

instead, finds many more words, which are clearly grouped in coherent clusters and represent a valuable semantic enrichment with respect to the original seeds. Given its richness, this second graph has been considered by subject-matter experts as a very good candidate to play the role of "Europe" filter.

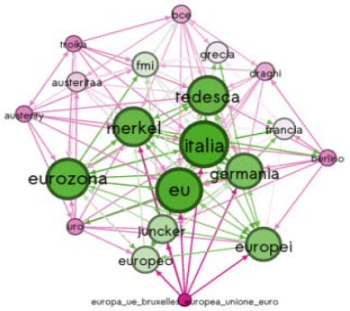


Figure 1: Geometric-Oriented
(['europa', 'ue', 'bruxelles', 'europea', 'unione', 'euro'], 8, 8)

3.3 Text Summarization and Encoding

One ongoing Istat's Big Data project aims at exploiting textual data automatically scraped from the websites of Italian enterprises in order to *predict* whether or not they perform e-commerce. To address this task, Deep Learning techniques are being used. Since input scraped texts are huge and Deep Learning algorithms are computationally intensive, a preliminary text summarization step is in order. Besides increasing efficiency, the summarization algorithm should hopefully improve accuracy by reducing the signal-to-noise ratio of input data. Word embedding models allowed us to achieve this goal with a *purely data-driven approach*.

To guide the summarization, we leveraged word embeddings trained on the whole web-scraped corpus. We used the Linear-graph illustrated in Figure 3 to select a set of *marker words* with high discriminative power for the detection of e-commerce, adopting as initial seeds the words: 'carrello', 'shopping', 'online'. (These marker words constitute what we called an "e-commerce pseudo-ontology" in the Introduction.) To summarize the texts, only input sentences containing *marker words* have been retained. This way, we obtained a 92.2% reduction of the original noisy text, along with a substantial improvement in the performance of the Deep Learning classifier (+20%, as compared to marker words defined by subject-matter experts). Lastly, we relied again on word embeddings to encode the summarized texts and feed the Deep Learning classifier. Once more, our experiments show that

word embedding models outperform more traditional text encoding approaches, like bag-of-words.

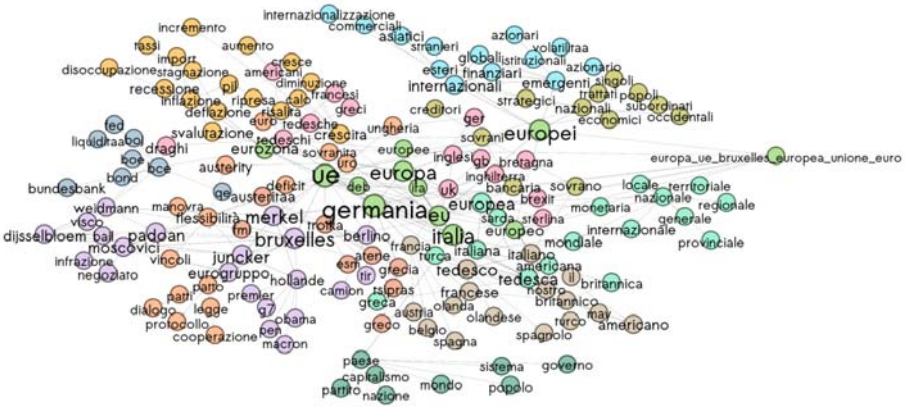


Figure 2: Geometric(['europa', 'ue', 'bruxelles', 'europea', 'unione', 'euro'], 3, 8)

Figure 3: Linear (['shopping', 'online', 'carrello'], 11, 8)

4. Conclusions

The techniques for dealing with large corpora of texts can greatly benefit from recent technology advancements. Word Embeddings are an example of this opportunity. Extensive evidence shows that Word Embedding models are indeed superior to more traditional text encoding methods like, e.g., bag-of-words. Ongoing works on textual Big Data at Istat make extensive use of these new approaches with very promising results.

References

Mikolov T., Yih W., Zweig G. (2013a). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT 2013*, pp. 746-751.

Mikolov T., Chen K., Corrado G., Dean J. (2013b). Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781.

Levy O., Goldberg Y., Dagan I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Trans. of the Association for Computational Linguistics*, vol.(3): 211-225.

Pennington J., Socher R., Manning C.D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of EMNLP 2014*, pp. 1532-1543.

van der Maaten L.J.P. and Hinton G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, vol(9): 2579-2605.

Gibbons A. (1985). *Algorithmic Graph Theory*. Cambridge University Press.

Analisi di dati d'impresa disponibili online: un esempio di data science tratto dalla realtà economica dei siti di e-commerce

Viviana De Giorgi, Chiara Gnesi
Istat – degiorgi@istat.it; gnesi@istat.it

Abstract
This work describes the process of extracting, organising and analysing detailed information on firms that trade electronic equipment on the Alibaba.com site. The first part concerns how translating unstructured information into variables organised in a statistical database by using dimensional classes, indices, indicators and classifications. A company-product matching is realised by encoding a textual variable with an international classification, and an automated analysis is applied in order to explore, describe and analyse the corpus retrieved from the Internet. In the second part a descriptive and econometric analysis shows how demographic and economic information on enterprises from Alibaba.com are very significant for competitiveness on the foreign market.
Keywords: encoding, classification, textual analysis, regression model.

Sommario
Il presente lavoro consiste nello sviluppo di un modello che consenta di trattare, organizzare ed analizzare informazioni dettagliate sulle imprese che commerciano apparecchiature elettroniche sul portale Alibaba.com.
La prima parte riguarda il processo di trasformazione dell'informazione destrutturata in variabili organizzate in un database statistico attraverso l'uso di classi dimensionali, indici, indicatori e classificazioni. Si è realizzato un abbinamento impresa-prodotto utilizzando una classificazione internazionale attraverso la codifica di una variabile testuale, su cui è applicata un'analisi automatizzata al fine di esplorare, descrivere e analizzare il corpus testuale tratto da Internet. Nella seconda parte è svolta un'analisi descrittiva ed econometrica, i cui risultati mostrano la presenza sul portale cinese di informazioni demografiche ed economiche sulle imprese altamente significative per la competitività sul mercato estero.

Parole chiave: codifica, classificazione, analisi testuale, regressione.

1. Introduzione

Questo lavoro nasce dagli spunti di riflessione e studio offerti nel corso delle lezioni di un Master universitario in Data Science[1] e si rivolge in particolare alle tecniche di trattamento, gestione ed analisi dei dati provenienti da fonti recuperabili on line[2] e fruibili in maniera gratuita. L'approccio adottato è quello della singola impresa che vuole migliorare la propria competitività nel mercato di riferimento, analizzando i dati generati dai processi aziendali nel settore in cui è presente o mira a posizionarsi. A tal fine sono preziose le informazioni dettagliate e aggiornate sui volumi prodotti, transazioni, struttura e demografia delle imprese concorrenti, presenti nei siti di commercio elettronico. Il presente lavoro è stato sviluppato utilizzando i dati estratti attraverso un'intensa attività di *web scraping* dal portale Alibaba.com, con riferimento alle imprese operanti nel settore delle apparecchiature elettroniche.

2. Dai dati destrutturati alle variabili statistiche: costruzione del database

Nel processo di trasformazione dell'informazione destrutturata acquisita online in variabili statistiche, un ruolo centrale riveste laclassificazione delle imprese a partire dal principale prodotto commercializzato. La variabile testuale – che corrisponde alla descrizione non codificata del prodotto commercializzato dalla società – è stata codificata secondo una classificazione di attività economica standardizzata a livello internazionale. Si è scelto l'elenco Prodcom con riferimento alle divisioni 26, 27 e 28, per un totale di 989 sottocategorie di prodotti[3].

L'attribuzione del codice Prodcom alla singola impresa è stata effettuata implementando un sistema di codifica *ad hoc*[4] strutturato in step successivi. La fase iniziale consiste nella normalizzazione dei testi attraverso lo sviluppo

---

[1] Master universitario in Data Science, Università Tor Vergata, Dipartimento di Ingegneria dell'impresa "Mario Lucertini", anno accademico 2015/2016. Si ringraziano Francesco Borrelli, Valentina Talucci e Domenica Fioredistella Iezzi per gli utili suggerimenti.

[2] L'acquisizione dei dati è stata effettuata nell'arco temporale che va dal 26 novembre 2016 al 7 gennaio 2017 dalla dott. Antonella Miele attraverso una attività di *web scraping*. I dati utilizzati sono relativi a 2.349 imprese presenti sul sito Alibaba.com e operanti nel settore delle apparecchiature elettroniche.

[3] http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD&StrNom=PRD_2011&StrLanguageCode=EN&StrLayoutCode=HIERARCHIC#

[4] Non avendo a disposizione software già sviluppati utilizzabili, è stato implementato un sistema di codifica *ad hoc* utilizzando il software SAS.

di un *parser* applicato alla variabile testuale e alle descrizioni della classificazione utilizzata. Successivamente si è realizzato un *matching* tra i due campi, attraverso un algoritmo che identifica l'abbinamento tra stringhe, sfruttando il dizionario al massimo livello di dettaglio possibile[5]. Infine si è realizzato l'abbinamento impresa-prodotto, assegnando a ciascuna impresa un codice Prodcom che identifica univocamente il principale prodotto commercializzato[6]. Il sistema di codifica ha permesso la classificazione del 95% delle imprese: un 30% circa vende "computer e prodotti di elettronica e ottica, apparecchi elettromedicali, apparecchi di misurazione e orologi", un quarto vende "apparecchiature elettroniche e apparecchiature per usodomestico non elettriche" e il 40% circa vende "apparecchiature elettriche diverse dalle precedenti" (Tavola 1).

Tavola 1: Imprese per divisioni Prodcom, valori assoluti e percentuali

Divisione prodcom	**n**	**%**
26 – computer e prodotti di elettronica e ottica	718	30,6
27 – apparecchiature elettroniche e apparecchiature per uso domestico non elettriche	618	26,3
28 – fabbricazione di macchinari ed apparecchiature n.c.a	893	38,0
non classificati	120	5,1
Totale complessivo	2349	100,0

L'analisi dei residui ha rivelato che la causa principale del mancato abbinamento deriva dalla presenza sul mercato di Alibaba di prodotti, elettrici e non, altamente specializzati ovvero sulla frontiera della tecnologia, non presenti nella Prodcom. Tuttavia, l'abbondanza di acronimi, abbreviazioni, slang hanno reso l'attività di standardizzazione particolarmente complessa.

In seguito alla codifica della variabile testuale, si è proceduto a una sua analisi automatizzata al fine di esplorare, descrivere e analizzare il *corpus*

[5]In questa fase si è utilizzato il dizionario al massimo livello di dettaglio possibile – 8 digit – in modo da abbracciare la descrizione del maggior numero di prodotti possibile. L'abbinamento prodotto/dizionario si è realizzato per molte sottocategorie di Prodcom; dopo aver analizzato i risultati ottenuti, si è scelto di utilizzare i 4 digit come il massimo livello di disaggregazione compatibile con una soglia di accuratezza ritenuta accettabile.

[6]L'assegnazione del codice è stata realizzata attribuendo all'impresa il codice Prodcom corrispondente alla classe in cui si è realizzato in maggior numero di *match* prodotto – dizionario, pesata per la frequenza più alta riscontrata in una determinata categoria di prodotto.

tratto da Internet. L'analisi testuale[7] consente di esplorare la struttura del testo sia come *corpus* – raccolta di frammenti testuali fra loro confrontabili – sia in relazione alla codifica ad esso attribuita. A tal fine, si è utilizzato TaLTaC2, particolarmente adattoallo studio di informazioni testuali non strutturate di grandi dimensioni e di informazioni strutturate a queste ultime collegate. Un primo approfondimento è offerto dalle misure lessicometriche, che consistono in una serie di misure e di indici statistici calcolati sul vocabolario e sulle sue classi di frequenza (Bolasco, 1999).Il *corpus*è costituito da 25.295 occorrenze, che corrispondono al numero totale di forme grafiche intese come unità di conto(Giuliano, 2004). L'ampiezza del vocabolario, pari a 4.363 forme grafiche distinte, riflettela specificità settoriale a cui attiene l'analisi. Coerentemente, l'indice di estensione lessicale percentuale, pari a 17,2, e l'indice di Guiraud normalizzato, pari a 27,4,confermano come la dimensione del vocabolario sia affetta da un *bias* determinato dalla specificità delle imprese analizzate. Tuttavia, nel settore è presente una gamma di prodotti piuttosto diversificata, come suggerito dal numero di *hapax*, pari a 50,2 (tavola 2).

Tavola 2: misure lessicometriche sul corpus

Misure lessicometriche	**Valori**
Occorrenze - N	25.395
Forme grafiche distinte - V	4.363
Type/Token (V/N)*100	17,2
% di Hapax (V1/V)*100	50,7
Frequenza media generale - N/V	5,8
G di Guiraud - V/sqrN	27,4
Coefficiente a	1,2

L'analisi lessicale, svolta a partire dall'analisi delle specificità, ha consentito di verificare, all'interno di singole classi, la rilevanza dei prodotti attraverso la sovra o sotto rappresentazione rispetto alla classificazione internazionale. L'utilizzo del dizionario della Prodcomcome risorsa statistica-linguistica esterna, ha permessoanalisi in parallelo. In effetti, l'indice *Term Frequency Inverse Document Frequency* (TFIDF) calcolato anche sul dizionario, ha consentito di evidenziare le caratteristiche peculiari dei prodotti venduti dalle imprese rispetto al panorama delle stesse che commercializzano prodotti elettronici.Inoltre, attraverso il confronto tra le forme grafiche del

[7]A tal fine, si è utilizzato TaLTaC2, un software per l'analisi automatica del testo nella duplice logica di Text Analysis e di Text Mining (TM), quindi sia come analisi del testo che come recupero e estrazione di informazione all'interno dello stesso

corpus e quelle del dizionario della Prodcom, si è potuto operare un controllo indiretto sulla qualità della codifica di cui al precedente paragrafo utilizzando lo scarto standardizzato come *proxy* di significatività[8]. Tale misura consente, infine, di caratterizzare le imprese rispetto alla peculiarità dei prodotti che le contraddistinguono all'interno del settore di riferimento (figura 1).

Figura 1: Parole chiave del corpus in base allo scarto standardizzato

Ulteriori elaborazioni sui dati reperiti dal sito hanno consentito la creazione di ulteriori variabili statistiche. Tra queste: *tenure* – una *proxy* dell'anzianità dell'impresa, costruita a partire dall'anno di iscrizione al portale; addetti e fatturato medi – a partire dal valore medio delle classi di riferimento; qualità – una variabile *dummy* che segnala la presenza di una certificazione di prodotto; propensione all'export – come quota percentuale di esportazioni sul fatturato ; ricerca e sviluppo – in termini di addetti medi impiegati nelle attività innovative; efficienza – capacità di risposta dell'impresa alle esigenze dei clienti. Il database finale è costituito da 18 variabili, che afferiscono all'Anagrafica dell'impresa, all'Attività economica, al Commercio estero, alla Dimensione economica, alla Competitività e alla Ricerca & Sviluppo.

3. Analisi descrittiva ed ecometrica dei dati

Ai dati descritti precedentemente sono state applicate le tecniche largamente adottate della ricerca statistica: un'analisi descrittiva del collettivo di riferimento, un'analisi multivariata di tipo esplorativo per la ricerca delle variabili da utilizzare in un modello econometrico e un modello di regressione che tenga conto della specificità dei dati[9]. Si riportano di seguito i principali risultati.

[8] Si è utilizzata la formula classicadellamisura di specificità in cui f_i^* è la frequenzarelativadella forma graficanell'elencoProdcom.

[9]Le informazionisulleimpresepresentisulsitovengonoaggiornate, anche se non si sa bene quando e come, e l'informazione dell'anno di riferimento è presente talvolta e solo per alcune variabili (per esempio il fatturato)

Per tutti i settori di attività, più della metà delle imprese si dichiara produttrice e venditrice, forse perché tale caratteristica tende a essere un parametro di scelta da parte di chi deve acquistare. Sono per lo più imprese medio-grandi, giovani, che in genere interagiscono con i clienti, con alte percentuali di export sul valore del fatturato, con presenza di dipendenti dedicati alla ricerca e sviluppo, disponibilità del certificato dei prodotti venduti. Cumulano un volume di esportazioni maggiore dell'80% le imprese che hanno più di 50 dipendenti, oppure sono nelle classi più elevate di fatturato, oppure rispondono almeno all'80% di richieste dal sito, o infine si dichiarano produttrici dei prodotti venduti. L'analisi condotta, e quindi il modello di regressione studiato, riguarda la dipendenza che il volume di esportazioni ha con le variabili presenti nel data set. Al fine della scelta delle variabili da utilizzare nel modello è stata effettuata un'analisi *cluster* gerarchica (SAS Institute Inc., 1999), scegliendo la variabile con minimo valore di *1-R²ratio*[10], e individuando le seguenti variabili: la produttività d'impresa, la variabile dimensionale data dal numero dei dipendenti occupati in ricerca e sviluppo e le tre variabili categoriche percentuale di risposta a richieste, attività economica e tipologia d'impresa. Le prime due risultano avere nel proprio cluster, nella suddivisione in 5 gruppi, il valore minimo di *1-R²ratio*; tra le variabili categoriche invece si evidenziano quelle aventi minore correlazione *own cluster* con le altre variabili. Il modello implementato consente di stimare i valori della variabile dipendente "volume delle esportazioni" sulla base dei valori assunti/osservati da/per alcune variabili indipendenti. Anche come conseguenza dei risultati dell'analisi *cluster* descritta precedentemente, si è scelto di includere tra queste: il fatturato per dipendente, il numero di dipendenti di ciascuna impresa, la tipologia di prodotto a 2 cifre, la percentuale di risposta alle richieste di possibili acquirenti, la quota di dipendenti d'impresa occupati in ricerca e sviluppo, la tipologia di impresa e il numero di anni di attività.
È stato stimato il seguente modello di regressione lineare (Rencher e Schaalje, 2008):

$$-\ln(export) = \alpha + \beta_1 \cdot \ln(fattxdip) + \beta_2 \cdot \ln(resp) + \beta_3 \cdot \ln\left(\frac{dip\_in\_rd}{dip}\right) + \beta_4 \cdot type + \beta_5 \cdot ateco2cifre +$$

$$\beta_6 \cdot \text{dip} + \beta_7 \cdot dip^2 + \varepsilon$$

dove: (a) $\ln(\text{export})$ è il logaritmo naturale del volume di esportazioni; (b)

[10] 1-R^2 ratio=(1-R^2 *own cluster*)/(1-R^2 *nextclosest*), dove *own cluster*=correlazione con il proprio gruppo di variabilie *nextclosest*=correlazione con il gruppo più vicino

$\ln(fattxdip)$ è il logaritmo della produttività; (c) $\ln(resp)$ è il logaritmo della percentuale di risposta; (d) $\ln\left(\frac{dip\_in\_rd}{dip}\right)$ è il logaritmo della quota di dipendenti occupati in ricerca e sviluppo; (e) $type$ è la tipologia di impresa, (f) ate è la tipologia di prodotto, (g) dip è il numero dei dipendenti.
In presenza di una variabile dipendente con distribuzione log-normale[11], l'applicazione di una trasformazione logaritmica alla variabile dipendente e alle variabili indipendenti continue ha come primo obiettivo di ottenere una distribuzione assomigliante a quella di una normale. Ciò implica, per i modelli lineari, la possibilità di estensione di tale ipotesi distributiva anche ai residui (ε) del modello e quindi consente di condurre in modo corretto i necessari test di significatività sui coefficienti stimati. Inoltre, la contemporanea trasformazione logaritmica delle variabili indipendenti (continue) consente di interpretare i valori dei coefficienti stimati direttamente in termini di elasticità. L'introduzione della variabile dip^2 è utile per verificare l'esistenza di eventuali relazioni non lineari tra *dip* e la dipendente, ovvero per capire se all'aumento del numero di dipendenti corrisponda una crescita delle esportazioni progressivamente superiore/inferiore. È stato inoltre studiato un secondo modello (modello2) introducendo l'interazione tra la quota di dipendenti occupati nella ricerca e sviluppo e la variabile categoriale relativa alla tipologia d'impresa. Tale scelta è coerente con l'idea che il livello di attività in ricerca e sviluppo possa rappresentare una fonte di valore aggiunto maggiore per le imprese che producono rispetto a quelle che vendono soltanto. I risultati ottenuti e riportati nella tavola 3 vengono di seguito descritti: (1) la relazione tra la variabile dipendente e la misura di produttività utilizzata è significativamente positiva; a una variazione dell'1% del fatturato per addetto corrisponde, mediamente, un variazione di oltre l'1% del volume delle esportazioni; (2) queste sono correlate positivamente anche con la percentuale di risposta a richieste dal sito e con il numero di anni di attività dell'impresa (coefficienti sempre significativi); (3) la stima dei due coefficienti relativi alla dimensione d'impresa evidenziano che questa accresce (come era logico aspettarsi) il volume delle esportazioni, ma con tassi progressivamente decrescenti all'aumentare del numero dei dipendenti (rendimenti decrescenti

[11] La variabile aleatoria $X = e^N$ segue la distribuzione logaritmica $logX(\mu, \sigma^2)$ solo se $N = logX$ segue la distribuzione normale $N(\mu, \sigma^2)$. La sua funzione di densità di probabilità è f(x)=e^(-[[(lnx-μ)]]^2/[[2σ]]^2)/(x√2πσ)

di scala); (4) sembrano esistere effetti differenziali tra il volume di esportazioni e le tipologie di prodotti venduti per settore di attività economica, ma non sempre i coefficienti sono significativi; (5) le *dummy* relative alla tipologia d'impresa mostrano coefficienti sempre non significativamente diversi da zero in assenza di interazione con la *proxy* di ricerca e sviluppo (modello1); (6) se fatte interagire (modello2) emerge invece come le due tipologie impresa produttrice e produttrice/venditrice abbiano un effetto positivo sulle esportazioni (rispetto alla modalità di riferimento impresa solo venditrice) e l'intensità di ricerca e sviluppo sembra accrescere significativamente le esportazioni solo per il settore delle imprese produttrici; (7) la variabile in oggetto risulta infatti correlata negativamente con la dipendente nei casi di imprese operanti esclusivamente nel settore del commercio e positivamente per quelle manifatturiere o contemporaneamente anche venditrici.

Tavola 3: Stima dei parametri del modello lineare (modello 1 e modello 2) nel data set iniziale e nel data set integrato

Variabile	**modello1**	**modello2**
ln(fattxdip)	1,024***	1,025***
Resp	0,002***	0,002***
num_anni	0,022***	0,023***
ate26 (rif,)		
at 27	-0,067*	-0,061
ate28	-0,094***	-0,085**
Others	0,067	0,063
Dip	0,012***	0,012***
dip^2	-0,001***	-0,001***
type venditrice (rif,)		
produttrice	-0,008	0,358***
produttrice/venditrice	-0,055	0,161*
ln(dip_in_rd/dip) x venditrice	-0,097***	-0,212***
ln(dip_in_rd/dip)) x produttrice		0,188***
ln(dip_in_rd/dip) x produttrice/venditrice		0,125***
Costante	2,291***	2,084***
N	1.913	1.913
r2_ajusted	**0,865**	**0,866**

**p<0,1; **p<0,05; ***p<0,01*

Le funzioni di densità della variabile dipendente osservata e stimata mostrano entrambe una forma distributiva approssimativamente normale: non emergono significative differenze tra i due modelli, ce forniscono entrambi una buona approssimazione.

Riferimenti bibliografici

Bolasco S. (1999). *L'analisi multidimensionale dei dati*, Roma, Carocci.

Giuliano L. (2004), *L'analisi automatica dei dati testuali. Software e istruzioni per l'uso*, Milano, LED.

Rencher A.C, Schaalje G.B. (2008). *LinearModels in Statistics. Second Edition*. Wiley.

SAS Institute Inc. (1999), *LogisticRegressionModeling Course Notes*, Cary, NC: SAS Institute Inc., pages 56-57.

The use of textual sources in Istat: an overview

Alessandro Capezzuoli, Francesca della Ratta, Stefania Macchia,
Manuela Murgia, Monica Scannapieco, Diego Zardetto[1]
ISTAT – Istituto Nazionale di Statistica – nome.cognome@istat.it

Abstract 1
Text Mining techniques allow a more widespread use of textual materials also in Official Statistics. We show implementations and current pilots realized in Istat, with a focus on both techniques and applications. Initially, text mining techniques were used to manage complex taxonomies or conduct open question analysis, while at the moment Big data frameworks allow to expand the different sources of data also to merge several data sources and to reduce response burden.

Abstract 2
Le tecniche di Text Mining consentono un ampio utilizzo di dati testuali anche nella Statistica Ufficiale. Sono descritte le implementazioni e le sperimentazioni realizzate in Istat in questo ambito, focalizzando sulle tecniche utilizzate e le applicazioni realizzate. Inizialmente il Text Mining veniva effettuato per gestire le tassonomie o effettuare analisi testuale delle riposte aperte, mentre più di recente il contesto dei Big data ha consentito di ampliare le fonti utilizzate e di integrarle tra loro anche in funzione del contenimento del *response burden*.

Keywords: text mining, official statistics, sentiment analysis

1. Automatic coding and semantic search of taxonomies

The first use of text mining techniques in Italian official statistics was finalized to manage complex classifications. Indeed, classifications are defined, which consist of structured lists of concepts, mutually exclusive, corresponding to codes that allow to produce a partition of the population. When the identification of the code corresponding to the concept does not present any ambiguity, it is possible to use closed questions with lists of items among which the one matching with the response is selected.

[1] This work comes from a common effort; paragraph 1.1 is written by Manuela Murgia and Stefania Macchia, par. 1.2 by Alessandro Capezzuoli; par. 2 by Francesca della Ratta, par. 3 by Monica Scannapieco and Diego Zardetto.

On the other hand, when codes belong to classifications that are complex in terms of structure, criteria and hierarchies, then the management of taxonomies is a very difficult task that implies the knowledge of the classification. Let us think, for example, of the classification of Occupation: in order to identify the code corresponding to each occupation it is necessary to consider different aspects, like the level of competences, their scope or the activities managed. In this paragraph, it is described how, with the evolution of technologies, this activity has been performed in different ways, using different software tools.

1.1. Automatic coding

Up to some years ago, statistics survey questionnaires rarely used open questions allowing textual answers because of the difficulties in processing them in order to provide *a measure* of the phenomenon. On the other hand, this could not often be avoided for some variables, like occupation, economic activity, education level that have necessarily to be coded according to official classifications for either national or cross-national data comparison.

In the past, verbal responses were manually coded, but this was very time-consuming, costly and error prone, especially for large amount of data (Macchia et Murgia, 2002). For this reason Istat decided to adopt automated coding systems that consist of two main parts: *i)* a database (*dictionary*) and *ii)* a matching algorithm. The dictionary is made of texts associated with numeric codes. Codes are those of official classifications and represent the possible values to be assigned to the verbal responses entering the coding process, while texts are the textual *labels* expressing the concepts that the classifications associate to codes. In order to improve the coding results, dictionaries are enriched with common language descriptions, resulting from answers to previous surveys. The matching algorithm is a 'weighting algorithm' that assigns a weight to each word of the verbal response to be coded. The weight indicates how much a word is informative and it depends on the word's frequency inside the dictionary: the higher its frequency the minor its weight. Then the algorithm compares the input response with all the texts inside the dictionary looking for a perfect match. If no exact match is found then it looks for a partial match with the most "similar" description, choosing the one with the highest weight.

The efficiency of the automated coding systems allowed Istat to use them not only to code responses of statistical surveys, but also to offer the coding service to a larger public such as governmental or private institutions, private citizens, who need to associate free text descriptions to official classifications codes, let's think, for instance, to businesses which have to identify their economic activity code for declarations to Chambers of Commerce. The

coding service was then made available on the Istat web site for the ATECO (the Italian version for Nace, the Economic Activity classification) variable. The software used for many years was ACTR (1998-2015) developed and distributed by Statistics Canada. In 2015 ACTR was not working anymore on the new Istat IT platform and it was substituted by CIRCE that behaves like ACTR but it is developed in house and based on R (Murgia et al., 2016). The choice of R made it possible to create a coding package freely downloadable from the website and also to offer a web service for the coding of the ATECO. The web service can be easily incorporated in any other software applications: electronic questionnaires of Istat surveys or in software systems of external organizations.

1.2 Semantic search within taxonomies

The evolution of technology allowed to explore also other software solutions suitable to represent the Statistical classifications logical structure, described within the Generic Statistical Information Model (GSIM). To this end, it was possible to exploit a very simple JSON object, to which then associate the metadata related to the classification (family, series, level, etc.). PUT and GET methods, related to the HTTP protocol, permit an easy acquisition of classification items that can then be organized through ad hoc procedures, on the basis of GSIM model, and stored into a relational database.

Being a JavaScript Object Notation, JSON is the natural environment for the construction of web applications using programming languages like e.g. Ajax/JavaScript combined with ad hoc frameworks as appropriate. Elasticsearch and Solr are the main frameworks used to search and share data. In particular, Elasticsearch provides a set of powerful and complete tools/plugins for data dissemination and the use of REST resources. Elasticsearch is well suited for the solution of some critical issues related to the use of statistical classifications in different fields (surveys, administrative registers, information systems, etc.), such as:

- acquisition, storage, management and updates of classifications;
- multilingual semantic search for coding;
- sharing and dissemination of coding tools.

Textual search is a very popular technique for users who seek information on the web. It does not require any special skill and users have already acquired through surfing the web and it is also suitable to search within statistical classifications and facilitate coding. The most common problem related to semantic searches within taxonomies concerns false-positive and false-negative results. The search is usually done through SQL queries allowing users to perform two types of operations: "exact match" and "full text". String parsing algorithms can be associated to the SQL queries.

A statistical classification can be indexed within Elasticsearch to perform complex and differentiated textual searches through DSL (Domain Specific Language) in JSON format. This solution permits to simplify the formulation of complicated SQL queries and makes the search system from any programming language usable. Elasticsearch allows users to manipulate large volumes of data thanks to an internal document management, completely independent from relational databases, and the opportunity to create distributed cluster.
Istat experience in using this methodology has been very satisfactory. The coding systems related to the main statistical classifications (ISCO, NACE, ISCED, COFOG, COICOP) were included in several Istat surveys ("Labour Force Survey", multi-purpose survey "Aspects of daily life", "Consumer prices", etc.) and Information system on occupation. Easy to use, widgets have been developed to include coding systems within web questionnaires and web applications.

2. Open questions analysis

Social research uses open questions also when category answers are not known or when researchers prefer to explore interviewees' different points of view using their own categories. This approach offers a great opportunity to realize analysis in depth, but it is difficult to be applied with the largest sample used in official statistics. So it is generally preferred using open questions only in pilot survey or small samples, to explore the possible list of answers and to obtain the closed-end list for the final survey. As an example, Istat used this approach in a survey on the female participation in parliamentary life: in 2000 an open question was introduced in a quarterly Multipurpose survey and the list of answering categories obtained with textual analysis was used in the 2005 annual Multipurpose survey.
However, in the early 2000s Text mining tools made it possible to analyse open questions also when codes does not belong to pre-defined classifications. The first example was introduced in Istat by Sergio Bolasco, who analysed the daily diaries collected in 2002-2003 Time use Survey to obtain a classification of some daily life actions (Bolasco et al., 2007). This classification was obtained using the Entity Research by Regular expression (RE) inserted in the tool Taltac2, a function that represents a very important turning point for the use of textual data in statistical surveys, because it made possible to pass from the **simple description of words contained in a corpus** (Lexical analysis) to the **classification of single records on the basis of**

words that are contained in each of them (Textual analysis[2]). The single word is no more the unit of analysis as the RE function searches or counts within the entire record a particular word or a combination of words, putting the result in a new customized variable.

This function was afterwards used in other Istat surveys. First it was used in the Survey on Occupations, developed in 2005-2006 and aimed at describing Italian labour market occupations, providing detailed information on each Occupational Unit. Researchers were interested also in tasks in which workers are daily involved, which was asked through an open question: "*What does your job consist of? Which are the activities you are involved in during your working day?*". Our aim was to provide each Occupational Units with a list of semi-standardized activities, labelling in the same way similar activities expressed in different ways by respondents. So, we used a strategy of text categorization adding in final dataset an extra column variable with a synthesis of the activities stated by interviewees: the final result was a list of over 7,000 specific activities (della Ratta, 2009).

A similar approach is currently used to check and correct the coding of economic activity carried out by interviewers in the Labour Force Survey: every quarter, 1500 records out of 24000 responses collected in the survey referred to specific Nace section are analyzed. The correctness of the codes assigned is verified from a double perspective: not only by comparing respondents' vocabulary reported in the response field of the question on economic activity with the specific dictionary of the official classification (Nace rev-2), but also considering other extra information connected with this variable collected in the same survey questionnaire. The process is completed with a thorough examination of data consistency in each session, to validate the corrections made and to assign the definitive proper code. At the end errors are transmitted to interviewers during specific training sessions in order to improve the all process of data collection, from the interview to the coding assignment (della Ratta et Tibaldi, 2014).

Other uses of Text Mining tools regarded the classification of open questions of the online survey on the dimensions of well-being (della Ratta et Tinto,

[2] The search for the textual information is run by complex queries using regular expressions with Boolean operators (AND, OR, NOT), lexeme reductions (wildcards as "*" and "?", e.g. contact* and customer?) and distances (LAGgxx) between consecutive words, that allow to identify different expressions used to convey the same concept (contact*LAG3 customer? is able to identify series such as "to contact the customer", "contacts with customers", "I contact my main customers"; the value of the new variable could be "to contact customers").

2012), or the analysis of residual answers inserted in single questions (*"Others", please specify*) that can improve the exhaustiveness of questionnaires and can be used in training activity for interviewers.
In conclusion, the availability of Text Mining tools made it possible to process open questions independently by the size of the text, being free in this way to use un-structured data in official statistics, especially in recursive analysis in which text categorization strategies can be repeated several times.

3. Dealing with Textual Big Data

Since recent years, in line with European-level strategic directives, Istat has been exploring the potential of Big Data sources for Official Statistics. Many of such sources – and notably those that seem the most promising so far – are made up of huge collections of unstructured and noisy texts.
In current Istat's projects, two types of unstructured sources were taken into account, namely: (i) textual data collected from the websites of Italian companies, obtained through automatic procedures of access and extraction performed on a large scale (hundreds of thousands of sites); (ii) messages in Italian language publicly available on Social Networks, typically collected in streaming after a preliminary selection step performed using *'filters'* (i.e. sets of keywords that a message must match to be deemed relevant).
The contexts of use of textual data from company websites include the enrichment of information in statistical business registers and the potential replacement of questions from surveys questionnaires. The possible uses of data from Social Network mainly concern the production of high-frequency (e.g. daily) sentiment indices.
At the moment the experiments with Social Networks data focused on the Twitter platform and on the development of "specific" sentiment indices: the goal is to measure the Italian mood about topics or aspects of life that might be relevant for Official Statistics (like the economic situation, the European Union, the migrants' phenomenon, the terrorist threat, and so on). The hope is that such sentiment indices can improve the quality of Istat's economic forecasting models, enrich existing statistical products (for example the BES) or create new statistical outputs in their own right.
Among the processing techniques used for these sources, a particularly promising type consists of the Word Embedding models. These models are generated by unsupervised learning algorithms (such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), both based on neural networks) trained on large collections of text documents. Their main objective is to map natural language words into vectors of a metric space, in such a way that the numerical representation of texts captures and preserves a wide range of syntactic and semantic relationship existing between words.

Istat successfully tested Word Embedding models in both the application scenarios sketched above. In the first scenario, Word Embeddings have been exploited to automatically summarize the huge text corpora scraped from company websites, and to subsequently encode the summarized texts in order to feed a Deep Learning algorithm for downstream analysis (e.g. to predict whether a given enterprise performs e-commerce). In the second scenario, Word Embedding models have been leveraged both to design the *'filters'* used to select relevant messages from Twitter and to evaluate the actual performance of the same *'filters'* after data collection.

In the following of this section a specific focus will be provided on data scraped from enterprises websites[3]. The Istat sampling survey on Information and Communication Technologies (ICT) in enterprises aims at producing information on the use of Internet and other networks by Italian enterprises for various purposes (e-commerce, e-skills, e-business, social media, e-government, etc.). In 2013, an Istat project started with the purpose of studying the possibility to estimate some indicators produced by the survey directly from the websites of the enterprises; these indicators included online sale rate, social media presence rate and job advertisement rate. The idea was to use web scraping techniques, associated, in the estimation phase, to text and data mining algorithms, with the aim of replacing traditional instruments of data collection and estimation, or to combine them in an integrated approach (Barcaroli et al., 2015). The recently achieved results are very encouraging with respect to the use of such techniques (Barcaroli et al., 2017).

The whole pipeline that has been set up for this project includes:

- A scraping activity performed by an *ad-hoc* developed software (RootJuice[4]).
- A storage step in which scraped data are stored in a NoSQL database, i.e. Apache Solr.
- A data preparation and text encoding step, performed in two different ways:
  1. tokenization, word filtering, lemmatization, generation of a term-document matrix
  2. word filtering and word embeddings.
- An analysis step, performed via machine learning methods on each of the text encodings resulting from the previous step.

[3] A more detailed focus on the processing of Twitter data is presented in the paper "Word Embeddings: a powerful tool for innovative statistics at Istat", submitted to this conference.

[4] Available on GitHub : https://github.com/SummaIstat/RootJuice/.

4. Conclusions and remarks

The techniques for dealing with large corpora of texts can greatly benefit from recent technology advancements. Word Embeddings are an example of this opportunity, giving additional possibilities to use un-structured data in official statistics for the purpose of integrating analyses or reducing response burden. Extensive evidence shows that Word Embedding models are indeed superior to more traditional text encoding methods like, e.g., bag-of-words. Ongoing works on textual Big Data at Istat make extensive use of these new approaches with very promising results.

References

Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarnò M.and Summa D. (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. *Journal of Austrian Statistics*, vol. 44, n. 2.

Barcaroli G., Scannapieco and M. Summa D. (2017). Massive Web Scraping of Enterprises Web Sites: Experiences and Solutions. *61st World Statistical Congress*, ISI.

Bolasco S., Pavone P., D'Avino E. (2007). Analisi dei diari giornalieri con strumenti di statistica testuale e text mining. In: Romano. *I tempi della vita quotidiana*, Istat, Roma, Argomenti, n. 32.

della Ratta Rinaldi F. (2009). Il trattamento dei dati, in F. Gallo, P. Scalisi, C. Scarnera. *L'indagine sulle professioni. Anno 2007, Contenuti, metodologia e organizzazione*. Collana Metodi e Norme, n. 42, Roma, Istat.

della Ratta-Rinaldi F.and Tinto A. (2012). Le opinioni dei cittadini sulle misure del benessere. Risultati della consultazione online. Roma, Istat-Cnel.

della Ratta-Rinaldi F. and Tibaldi M. (2014). Sperimentazione di un sistema di controllo e correzione per la codifica dell'attività economica. *Istat Working Paper*, n. 4, 2014.

Macchia S. and Murgia M. (2002). Coding of textual responses: various issues on automated coding and computer assisted coding. *Proc. of JADT 2002: 6es Journées Internationales d'Analyse Statistique des Données Textuelles*.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.

Murgia M. and Prigiobbe V. (2016). La nuova applicazione di codifica web dell'ATECO 2007: WITCH, un web service basato sul sistema di codifica CIRCE. *Istat Working Papers* n. *19.*

Pennington J., Socher R. and Manning C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Twitter e la statistica ufficiale: il dibattito sul mercato del lavoro

Francesca della Ratta, Gabriella Fazzi, Maria Elena Pontecorvo,
Carlo Vaccari, Antonino Virgillito[1]
Istat – Istituto Nazionale di Statistica, Rome – Italy

Abstract
The goal of the paper is to show the potential and the benefits of the integration between the big data analysis techniques and techniques used for the textual analysis, through the analysis of a corpus extracted from Twitter. The analysis is the development of a method already experimented in other works (della Ratta, Pontecorvo, Virgillito, Vaccari, 2016 and 2017), in which we started from the collection of selected tweets through a list of hashtags defined according to the theme of interest. This procedure allows to obtain in a reasonable time a selection of tweets of interest, on which to apply textual analysis techniques to describe the contents of the text and to identify its main semantic contents. The paper analyzes the role of the National Institute of Statistics in the discussion on the labor market in the periods when ISTAT spreads the monthly and quarterly press releases on employment. The analysis, already conducted at the end of 2016, has been replicated and refined in the same period of 2017, in order to show the distinctive elements of the labor market debate and to understand the changes in the perception of public opinion, also taking into account the changes in terms of the economic situation and the political scenario.

Key words: big data; text mining; twitter; Istat, labour market

1. Big data e Twitter

I dati provenienti dai Social Network sono una delle sorgenti di Big Data più utilizzate dai ricercatori: l'enorme diffusione di questi siti web, nei quali gli utenti generano grandi quantità di informazioni, li rende potenzialmente una delle fonti più interessanti anche per i dati testuali. Twitter è un Social Network nel quale gli utenti scrivono e leggono corti messaggi chiamati

[1] Questo lavoro è frutto della riflessione condivisa degli autori; il paragrafo 1 è stato redatto da Carlo Vaccari e Antonino Virgillito, il paragrafo 2.1 da Francesca della Ratta, il 2.2 da Gabriella Fazzi e Maria Elena Pontecorvo, le conclusioni da tutti gli autori.

"tweet", normalmente visibili da tutti gli utenti, che possono anche "iscriversi" ai tweet di altri utenti (diventando "follower"), inoltrare ("retweet") singoli tweet ai propri followers o aggiungere "mi piace" ad altri tweet. Twitter è oggi uno dei Social Network più diffusi, e ha superato nel 2017 i 300 milioni di utenti attivi. Secondo Alexa (2018) Twitter è oggi il tredicesimo sito più visitato al mondo, l'ottavo negli USA. Scopo di questo lavoro è applicare le tecniche dell'analisi testuale a un corpus estratto da Twitter, unendo i due mondi dei Big Data e dell'Analisi Testuale. La raccolta dei dati da Twitter è stata effettuata utilizzando una piattaforma, la "Sandbox"[2], che è il risultato finale del progetto "Big Data in Official Statistics", portato avanti nell'ambito dell'High Level Group on Modernisation of Official Statistics (HLG-MOS). La Sandbox è un ambiente web-based utilizzato per numerosi esperimenti basati su diverse sorgenti dati come le visite alle pagine di Wikipedia, i dati sul Commercio Estero del sito Comtrade dell'ONU, i siti delle imprese per ricercare annunci di lavoro e, appunto, i tweet raccolti in varie nazioni del mondo. La Sandbox è oggi ancora utilizzata per portare avanti le sperimentazioni della ESSnet on Big Data[3], un progetto europeo coordinato da Eurostat per l'utilizzo dei Big Data nella produzione di statistiche ufficiali. I tweet analizzati sono stati raccolti attraverso uno strumento online messo a disposizione gratuitamente da Twitter (Streaming API), interrogato attraverso programmi scritti in R ed eseguiti all'interno della Sandbox. Questa soluzione, per quanto semplice da utilizzare e di immediata implementazione, presenta limitazioni sia per l'ammontare dei dati che possono essere estratti, sia per la non completa aderenza dei dati ottenuti rispetto ai filtri impostati in fase di estrazione, come spiegato nella sezione successiva. I tweet acquisiti sono stati memorizzati su Elasticsearch, un database installato nella Sandbox specializzato in dati semi-strutturati, che permette di memorizzare grandi quantità di documenti ed estrarre velocemente dei sottoinsiemi attraverso query basate su parole chiave.

2. L'analisi dei post sul mercato del lavoro: l'impatto dell'Istat

2.1 Creazione del corpus

Per analizzare i dati estratti da Twitter si è replicato il metodo testato in occasione di precedenti lavori (della Ratta, Pontecorvo, Virgillito, Vaccari; 2016 e 2017). Si è deciso, in questo contesto, di focalizzare l'analisi sul ruolo

[2] I risultati del progetto Sandbox, coordinato da Virgillito nel 2014 e da Virgillito e Vaccari nel 2015, sono illustrati in Unece (2014 e 2016).

[3] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

ricoperto dall'Istat nella diffusione delle informazioni sulla tematica del lavoro, estraendo automaticamente un primo set di tweet nelle settimane in cui l'Istat diffonde i dati mensili e trimestrali sul mercato del lavoro. Tale estrazione, già effettuata a fine 2016, è stata replicata nello stesso periodo del 2017, partendo da una query piuttosto ampia[4] che ha consentito di ottenere un corpus di 58.277 tweet relativo al periodo 28 novembre-12 dicembre 2017. Da questo corpus sono stati estratti tutti gli hashtag con occorrenza maggiore di 14 (facilmente identificabili nel testo grazie alla presenza del simbolo #) tra i quali sono stati individuati quelli strettamente connessi alla discussione sul mercato del lavoro (Tabella 1). È quindi stato estratto, utilizzando il software Taltac2, un corpus di 19.398 tweet contenente almeno uno degli hashtag di interesse. Questo corpus è stato ulteriormente ripulito eliminando i tweet relativi alle offerte di lavoro (presenza degli hashtag #offertalavoro #annunciolavoro), considerati non pertinenti. Si è così arrivati a un corpus composto da 17.419 tweet, composto da 283.000 occorrenze, 18.000 forme grafiche e una ricchezza lessicale (rapporto type/token) del 6,7%.

Poco più di un terzo dei tweet sono originali, mentre il volume dei retweet costituisce il 63% del corpus complessivo, in misura maggiore rispetto al corpus del 2016. Per "misurare" l'impatto dell'Istat nel dibattito sul lavoro sono stati etichettati tutti i tweet in cui compare la forma "Istat": il 13,9% del totale, una misura quasi triplicata rispetto a quanto osservato nel 2016 (5%). Se da un lato nel 2016 l'impatto del concomitante dibattito referendario aveva ridimensionato il peso del commento del dato Istat nella discussione sul mercato del lavoro, nel 2017 i temi della ripresa occupazionale e delle sue caratteristiche sembrano aver attirato maggiormente l'attenzione degli utenti. Inoltre, la prima uscita del rapporto annuale integrato sul mercato del lavoro ha probabilmente accresciuto il peso dei commenti sui dati.

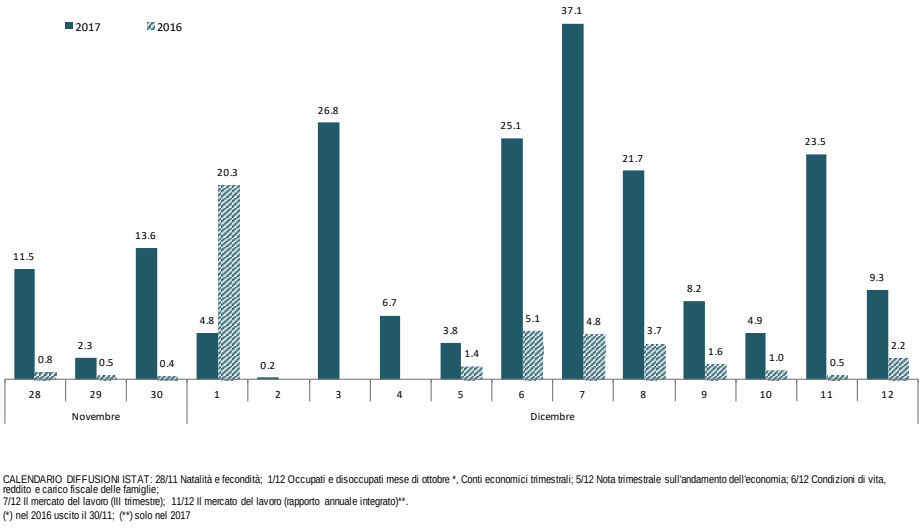
Se nel 2016 la presenza dei riferimenti a Istat si addensava in corrispondenza delle uscite ufficiali, nel 2017 è distribuita in maniera più uniforme, con un picco in corrispondenza del comunicato trimestrale del 7 dicembre (nel quale

[4] La query iniziale utilizzata è la seguente: "(istat OR inps OR #istat OR #inps OR #lavoro OR #occupati OR #disoccupati OR #disoccupato OR #jobsact OR #occupazione OR #disoccupazione OR #mercatodellavoro OR #poletti OR #cassaintegrazione)". Sul primo corpus di 58.277 tweet estratto dall'API di Twitter è stata rieseguita la stessa query in Elasticsearch, che ha consentito di effettuare una selezione ulteriore, eliminando moltissimi tweet che pur estratti attraverso la stessa query non contenevano le parole chiave, evidenziando una non completa accuratezza dell'API gratuita di Twitter nell'applicazione dei filtri di estrazione. Alla fine si è ottenuto un corpus di circa 26 mila tweet, su cui è stata effettuata la selezione successiva.

ha avuto molta eco la notizia del record assoluto di lavoratori a termine - Figura 1).

Tabella 1 – Selezione di hashtag

HASHTAG	OCC	HASHTAG	OCC	HASHTAG	OCC	HASHTAG	OCC
#lavoro	14.172	#licenziamento	190	#occupati	48	#MercatoDelLavoro	19
#jobsact	1.225	#disoccupato	173	#Occupazione	42	#precarizzazione	19
#occupazione	948	#Thyssenkrupp	164	#Cococo	42	#precarietà	19
#JobsAct	861	#contaillavoro	158	#Discoll	41	#Smartworking	18
#Jobsact	587	#lavoratori	156	#orientamento	41	#voucher	17
#disoccupazione	463	#Disoccupazione	149	#cassaintegrazione	40	#freelance	17
#povertà	414	#Melegatti	139	#mercatodellavoro	37	#Art18	15
#Poletti	278	#GaranziaGiovani	124	#JobsActSempre	32	#dipendente	15
#precari	265	#precariatodistato	110	#smartworking	31	#ScuolaLavoro	15
#LAVORO	205	#precariato	109	#thyssen	31		
#ContailLavoro	201	#pandoro	98	#RelazioniIndustrialiA	20		
#disoccupati	196	#articolo18	53	#poletti	20		



CALENDARIO DIFFUSIONI ISTAT: 28/11 Natalità e fecondità; 1/12 Occupati e disoccupati mese di ottobre *, Conti economici trimestrali; 5/12 Nota trimestrale sull'andamento dell'economia; 6/12 Condizioni di vita, reddito e carico fiscale delle famiglie;
7/12 Il mercato del lavoro (III trimestre); 11/12 Il mercato del lavoro (rapporto annuale integrato)**.
(*) nel 2016 uscito il 30/11; (**) solo nel 2017

Figura 1 – Incidenza riferimenti a Istat per giorno. Anno 2016 e 2017

Più modesto l'impatto del comunicato mensile sull'occupazione, i cui dati erano risultati sostanzialmente stabili (al contrario nel 2016, data la concomitanza con il referendum costituzionale, il mensile aveva registrato la quota massima di citazioni). Un volume consistente è stato registrato in occasione del comunicato sulle condizioni di vita e di reddito (6/12) e della

presentazione del primo rapporto integrato sul mercato del lavoro (11/12). Anomalo il picco del 3 dicembre, domenica, alimentato da un notevole tasso di retweet molto critici sulle politiche del lavoro dell'attuale governo dovuti probabilmente all'intervento del segretario del PD Matteo Renzi in una popolare trasmissione serale (Che Tempo Che Fa) centrato anche sulle politiche del mercato del lavoro degli ultimi anni. Il più citato è stato un tweet critico sul meccanismo di conteggio degli occupati, insieme ad altri più politici sull'aumento del lavoro a termine dell'ultimo periodo.

2.2 Il contenuto del corpus

Il contenuto del corpus può essere descritto utilizzando le parole chiave, calcolate rispetto all'italiano standard (Bolasco, 2013) che consentono di delimitare gli ambiti di contenuto: si incontra innanzitutto *contratti*, con riferimento all'aumento dei *contratti a termine* o *a tempo determinato.* Contribuiscono alla sovrarappresentazione del termine un numero limitato di tweet (13) che ricevono tuttavia numerosi retweet e che, riprendendo il dato Istat sulla durata dei contratti evidenziano l'aumento del *precariato*, anch'esso termine sovrautilizzato (Figura 2). Altri termini molto presenti nel testo sono *disuguaglianze* ed *esclusione*, utilizzati soprattutto in un post della Caritas che riprende il dato sulla povertà pubblicato il 6 dicembre.

Figura 2 - Tagcloud delle parole chiave

Colpisce la presenza di termini molto forti, che connotano un dibattito dai toni pesanti: *trucco*, *fraudolenta, infamia, tossico, truffa, schiavitù.* Analizzando i contesti d'uso si riscontra che ciascuno di questi termini è riferito a episodi diversi (*trucco* dei dati sulla definizione di occupazione); manodopera *fraudolente*; truffa del Governo sulle *pensioni*; accordo *tossico* con riferimento

al *CETA*; *infamia* contro il lavoro in riferimento al Jobs Act) e che proprio i tweet più forti siano quelli in grado di generare un numero elevato di retweet. Significativi anche termini utilizzati in tweet in cui si evocano storie, e in cui il dato statistico è sostituito dal caso esemplare, capace di generare empatia e, di conseguenza, retweet. Non è un caso che i termini maggiormente sovrarappresentati facciano riferimento ad un unico tweet, su un lavoratore colpito da *leucemia* che *guarisce* ma viene comunque *licenziato.* Fra gli esempi anche quello di una madre *separata,* licenziata dall'Ikea a Milano. Anche i riferimenti al *record* degli occupati e a quanti *esultano* per i dati sull'occupazione sono riportati talvolta in maniera critica; fa eccezione il riferimento al tasso di disoccupazione *giovanile*, che viene ripreso in maniera neutra dall'agenzia Ansa e retweettato numerose volte.

Prendendo in considerazione i segmenti ripetuti (ossia le sequenze di parole ripetute nel testo), si possono delimitare quattro aree semantiche principali a cui fanno riferimento i tweet (Tabella 2). In primo luogo ci sono le espressioni che rimandano alla pura diffusione delle notizie che ruotano intorno alla tematica "lavoro" e che hanno un peso rilevante anche in termini di occorrenze. In particolare emerge da un lato il riferimento ai dati diffusi dall'Istat su povertà, natalità e occupazione, dall'altro spiccano due segmenti che si riferiscono agli episodi di attualità già citati: il licenziamento da parte di Ikea di una madre separata con due figli piccoli e quello di un dipendente di una fabbrica di vernici, avvenuto dopo un lungo periodo di assenza per malattia. Accanto ai segmenti relativi alle notizie, vi sono poi i segmenti riconducibili ai commenti degli esponenti politici, ai provvedimenti legislativi e alle prime avvisaglie di campagna elettorale. A questi fanno da contraltare i tweet caratteristici del dibattito pubblico tra cui non mancano note polemiche o sarcastiche. Infine, nonostante il file sia stato in parte ripulito dagli hashtag riconducibili agli annunci di lavoro, emergono comunque alcuni segmenti inerenti la ricerca di particolari profili professionali.

Come è facilmente intuibile, peraltro, alcuni contenuti caratterizzano maggiormente i frammenti in cui si fa esplicito riferimento all'Istat. Rispetto all'analisi effettuata nello stesso periodo dello scorso anno, l'analisi delle specificità mostra la prevalenza di un linguaggio più tematico che tecnico quando si cita l'istituto (*dati, contratti, #povertà, disoccupazione*), mentre i tweet che parlano di lavoro senza citare l'Istat fanno riferimento ai fatti di cronaca e alla politica *(#jobsact, #pensioni, legge, licenziato ecc.)*, con minori riferimenti personali ai soggetti che nel 2016 erano in prima linea nella campagna referendaria. Inoltre l'analisi delle concordanze mostra che lo stesso riferimento all'Istat viene utilizzato in differenti contesti.

Tabella 2 – Segmenti ripetuti principali

Le notizie		Riferimenti politici		Dibattito pubblico e polemica		Annunci di lavoro	
Segmento	**Occ**	**Segmento**	**occ**	**Segmento**	**occ**	**Segmento**	**occ**
dati #Istat	419	Missione compiuta\ \ #JobsAct	67	Come ti trucco i dati	348	#lavoro #roma #romalavoro #lavorare	152
a rischio #povertà	392	Ministro #Poletti	60	continuano a produrre sfruttamento	119	#lavoro #adnkronos	144
guarisce e viene licenziato	253	#jobsact funziona	54	tutto da rifare	55	kijiji lavoro	53
esclusione sociale	236	Fedriga Presidente	47	essere licenziati	40	cerca socio	32
madre separata	195	campagna elettorale	43	#Bonus creano dipendenza	31	Commessa IV livello part time	27
tempo determinato	86	manovra finanziaria	28	si sono rivelate tutele inesistenti	27	#lavoro professionale	21
tempo indeterminato	80	Liberi e Uguali	18	conti non tornano	18	diventare #psicoterapeuta	13
terzo trimestre	71	presidenta #boldrini	11	politici hanno distrutto tre generazioni	3	ufficio acquisti	8
crollo della natalità	56	Politiche Attive	9	dovremmo ribellarci	2	dirigente medico	7
#Algoritmi #BigData	55	Lavori usuranti	2	giovani andati via grazie a te	2	Concorsi Pubblici Gazzetta Ufficiale	3

Oltre alla stretta diffusione delle notizie e al commento del dato sull'aumento dei contratti a termine, non manca l'uso strumentale dei dati come metro di giudizio delle politiche sul mercato del lavoro [***#Istat*** *"record di occupati a_termine: sono 2,8 milioni". ecco l' unico risultato oggettivo del #jobsact..*"; *continua a calare la #disoccupazione-i nuovi dati* ***#Istat*** *confermano le previsioni, un' altra ventata di ottimismo...*]. Rispetto al 2016 il tono sarcastico di alcuni tweet è meno rivolto esplicitamente all'Istat ma in generale alla situazione del Paese [«*record di #precari in Italia, 2,8 milioni. va tutto ben, madama la marchesa.. #lavoro* ***#Istat*** *#occupazione"*»]. Resta però un residuo polemico su alcune definizioni di occupazione e disoccupazione [«*Ricordiamo che per #istat se si lavora un'ora retribuita a settimana si è considerati occupati. #supercazzola*»; «*Come ti trucco i dati #Istat sulla disoccupazione: il 14; 6% dei contratti dura meno di 3 giorni, il 31% un_mese*»]. Infine, di interesse la valutazione del tono del testo, possibile con l'analisi degli aggettivi positivi e negativi, riconosciuti all'interno di Taltac2. Il rapporto tra aggettivi negativi e positivi è del 50,2%, un valore che denota una criticità media, pari a quella che si riscontra nel linguaggio della stampa (Bolasco, della Ratta, 2004). Il livello di criticità è variabile nelle diverse giornate: è più basso nei giorni di diffusione dei

comunicati, specie quello mensile, mentre è particolarmente elevato il 3 dicembre, a causa del "rumore" prodotto dai retweet (i retweet presentano una criticità del 63,6%), probabilmente a causa del maggiore successo dei tweet polemici. Tra gli aggettivi negativi più frequenti *precari, fraudolenta, dannoso, fallito*[5].

3. Conclusioni

L'analisi effettuata ha consentito di affinare una metodologia di trattamento dei tweet: dal punto di vista della loro estrazione, la procedura utilizzata ha consentito di ottenere in partenza un file più pulito su cui operare una selezione a partire dalla lista degli hashtag. L'analisi del testo ha poi consentito di evidenziare i diversi contesti in cui si fa riferimento al dato della statistica ufficiale. Particolarmente interessante il confronto tra i risultati dello stesso corpus a un anno di distanza. Infatti, nello stesso periodo dell'anno precedente la discussione era fortemente condizionata dal dibattito referendario che ha probabilmente "stravolto" la discussione sulle tematiche del lavoro. Nei tweet di un anno prima i livelli di criticità erano più elevati e il ruolo dell'Istat più ridimensionato (13% la presenza odierna contro il 5% di un anno prima). Il tono del testo appare in generale più neutro, con maggiori richiami all'Istat nella sua veste ufficiale di diffusore di dati e meno come oggetto di scherno e polemica. Riguardo ai contenuti, nella discussione di fine 2017 sembra avere avuto più peso la discussione sugli effetti del Jobs Act e della diffusione del lavoro precario. Il corpus odierno è inoltre caratterizzato da un più ampio ricorso al retweet.

Riferimenti

Alexa (2016). *Twitter site overview*, at http://www.alexa.com/siteinfo/twitter.com.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma, Carocci.

Bolasco S., della Ratta-Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. In *JADT 2004 - Le poids des mots, Actes des 7es Journées internationales d'Analyse Statistique des Données Textuelles*. UCL. Louvain.

della Ratta-Rinaldi F., Pontecorvo M.E., Virgillito A., Vaccari C. (2016). Big data and textual analysis: a corpus selection from twitter. Rome between the fear of terrorism and the Jubilee. In *JADT 2016 - Statistical Analysis of*

[5]Sono stati comunque eliminati i termini tecnici (riferiti a specifici aggregati statistici) che hanno una connotazione negativa, come disoccupato, scoraggiato o povero.

Textual data – Vol.2. Nice.

della Ratta-Rinaldi F., Pontecorvo M.E., Virgillito A., Vaccari C. (2017). The Role of NSIs in the Job Related Debate through Textual Analysis of Twitter Data. *NTTS 2017*. Brussels.

UNECE (2016). *Big Data in Official Statistics*.

http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

UNECE (2014). *Big Data in Official Statistics*.

http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

Vaccari C. (2014). *Big Data and Official Statistics*. PhD Thesis, School of Science and Technologies. University of Camerino.

Gauging An Author's Mood Using Hidden Markov Chains

Sami Diaf
Hildesheim Universität – sami.diaf@uni-hildesheim.de

Abstract
This paper aims to gauge the mood of an author using a text-based approach built upon a lexicon score and a hidden Markov model. The text is tokenized into sentences, each given a polarity score, yielding three evaluative factors (positive, neutral and negative) which represent the observable states. The mood of the author is considered a latent state (good, bad) and is estimated via a hidden Markov model. Tested on a psychological fiction, Franz Kafka's novel Metamorphosis, this methodology shows an interesting linkage between the author's feelings and the intent of his writing.

Keywords: Sentiment analysis, hidden Markov model, polarity.

1. Introduction (Times Bold 14 pt, left)

Sentiment analysis is defined as the general method to extract subjectivity and polarity from a text, while *semantic orientation* refers to the polarity and strength of words, phrases, or texts, meaning a measure of subjectivity and opinion in the text, capturing an evaluative factor and potency or strength of a given corpus toward a given subject (Taboada et al., 2011).

Extracting sentiment automatically usually involves two main approaches (Taboada et al., 2011): a lexicon-based approach built on computing orientation for a document from the semantic orientation of words or sentences, and a text-classification approach stemming from supervised machine learning techniques and involves building classifiers from labeled instances of texts or sentences. Lexicon-based models stress out the importance of adjectives as an indicator of a text's semantic orientation and have been preferred in the linguistic context as classifiers yielded changing results regarding their areas of application (Taboada et al., 2011).

Among many lexicon-based approaches adopted in the academic field, the one implemented by Hu and Liu (Hu and Liu, 2004) remains popular. It was built upon two hypotheses concerning the semantic orientation: independence of context (prior polarity) and being expressed as a numerical value suing an opinion lexicon.

This article uses the polarity approach of Hu and Liu to build a sequence of

evaluative factors (positive, neutral and negative), considered as the realization of an observable state x, and supposes the mood of the author could be approached via a two-state latent variable z taking two hidden states (good and bad). For this aim, hidden Markov models (Murphy, 2012) will be used to estimate the transition probabilities between hidden and observed states, to better estimate long-range correlations among the sequence of data than standard Markov models.

2. Polarity function

Polarity is defined as the measure of positive or negative intent in a writer's tone (Kwartler, 2017) and can be calculated by sophisticated or fairly straightforward methods, usually using two lists of words: one positive and one negative. Hu and Liu set up the architecture for the polarity function used to tag polarized words in the English language (Hu and Liu, 2004) and Rinkler (2017) provided a detailed description of the polarity function and its computation. A context cluster of words is pulled around a polarized word to be considered as valence shifters. Words in this context cluster are tagged as neutral, negator, amplifier or de-amplifier. Each polarized word is then weighted according to a dictionary of positive/negative words and weights, and then further weighted by the number of position of the valence shifters directly surrounding the positive or negative word. Final computation step is the sum of the context clusters divided by the square root of the word count,

which yields an unbounded polarity score.

2. Application

To illustrate this framework, we took the English version of the novella Metamorphosis written by Franz Kafka published in 1915 under the name « *Die Verwandlung* » and freely available at the *Project Gutenberg* database. This work was translated to English by David Wyllie in 2002 and belongs to the psychological fiction category.

The novella is broken down into sentences, a process called tokenization, and then we compute the polarity function for each sentence, to construct a sequence of evaluative factors (positive, neutral or negative) according to the polarity score, as shown in Figure 1.

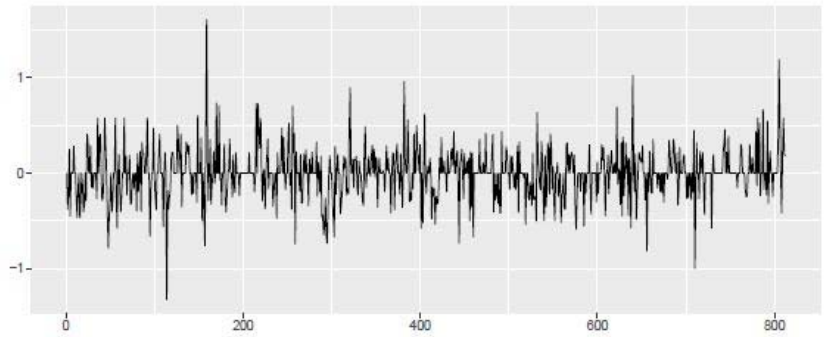


Figure 1. Sequence of data corresponding to the polarity score of each sentence.

This step generates 812 sentences where the positive and negative polarity scores represent respectively 29.1% and 28.6% of the total. The remaining sentences (42.3%) correspond to the neutral evaluative factor. Statistical tests show that the generated time series has the first two autocorrelations significantly different from zero and exhibits a slightly persistent memory as the estimated Hurst exponent is 0.587, significantly different from the value of 0.5 which corresponds to the case of a Brownian motion (Mandelbrot and Hudson, 2006). The estimated probability transition matrix of evaluative factors via the maximum likelihood shows the associated Markov chain is irreducible with no persistent states, as shown in Figure 2.

$$
\begin{array}{c} t/t+1 \\ \textit{negative} \\ \textit{neutral} \\ \textit{positive} \end{array}
\begin{array}{c}
\begin{array}{ccc} \textit{negative} & \textit{neutral} & \textit{positive} \end{array} \\
\begin{pmatrix} 0.362 & 0.375 & 0.263 \\ 0.250 & 0.474 & 0.276 \\ 0.264 & 0.396 & 0.340 \end{pmatrix}
\end{array}
$$

Figure 2. Probability transition matrix of the evaluative factors.

We assume the mood of the author could be modeled via a latent variable Z taking two states (good and bad). Hence, we can build a hidden Markov model explaining the interactions between observable states (positive, neutral and negative) and latent, unobservable states (good and bad). To estimate the hidden Markov model, the transition matrix of the latent state is set uniformly, that is all its elements equals 0.5, the same applies also for the initial latent vector. However, the emission matrix which describes the links between the latent and the observable states is set arbitrarily as in Figure 3.

	positive	*neutral*	*negative*
good	0.7	0.2	0.1
bad	0.1	0.2	0.7

Figure 3. Prior probability transition of the emission matrix.

Given these priors, the estimated hidden Markov model using the Baum-Welch algorithm (Murphy, 2012) yields a starting probability vector slightly skewed to good mood (51%) than bad mood (0.49). The estimated transition and emission matrices are reported in Figure 4 and 5 respectively.

	good	*bad*
good	0.727	0.273
bad	0.371	0.629

Figure 4. Estimated transition matrix via Baum-Welch algorithm.

	positive	*neutral*	*negative*
good	0.358	0.574	0.068
bad	0.199	0.218	0.583

Figure 5. Estimated emission matrix via Baum-Welch algorithm.

Results demonstrate significant links between writing without intent (neutral state) and being in a good mood, and between negative intent and the bad mood. The most probable states estimated via the Viterbi algorithm (Murphy, 2012) clearly show the dominance of the good state (71.4%) over the bad (28.6%) as shown in Figure 6.

These findings help clarify the nature of the story (thriller, roman, novella, …) and the author's narrative style which could be confirmed by analyzing the remaining works. Finally, it is worth noticing that this methodology could also be used to assess the accuracy of translations with respect to the original work, by comparing the similarities of the transition and the emission probabilities of the hidden Markov models.

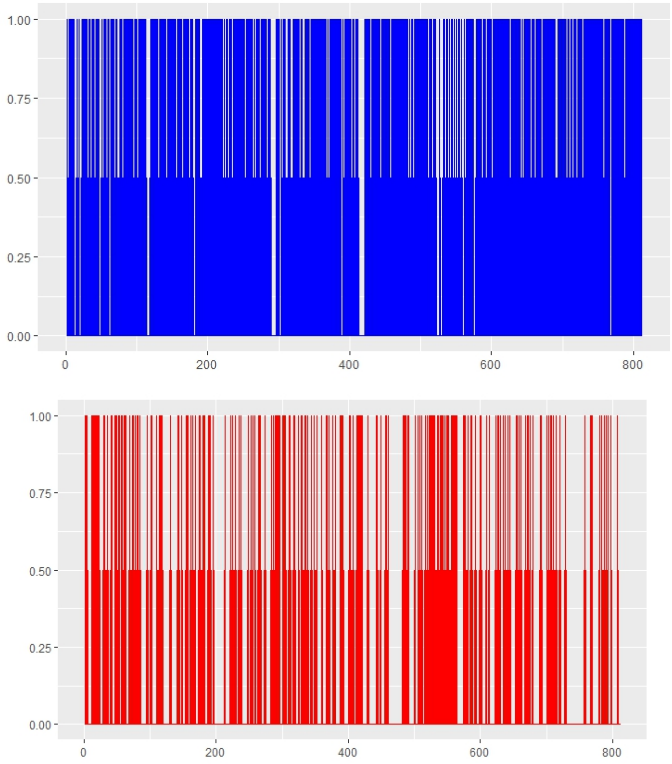


Figure 6. Most probable states estimated via Viterbi algorithm (Bad in red and Good in blue).

4. Conclusion

This works expands the application field of semantic orientation to explore a new probabilistic approach based on hidden Markov models and evaluation factors. The resulting outcomes help understanding the author's mood by examining the linkage between the evaluative factors which express the author's mindscape through his writing. The emission probabilities between the latent states and the evaluative factors helped identifying hidden structures linked to the psychological state of the author and the development of the facts. This approach could be used as a controller of translation accuracy under the condition of having a precise list of positive and negative words in the original language, to be able to compute the polarity score.

References

Hu M. and Liu B. (2004). Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD, pp. 168-177.

Kwartler T. (2017). Text Mining in Practice with R. Wiley.

Mandelbrot B. and Hudson R.L (2006). The Misbehavior of Markets: A Fractal Review of Finance Turbulance. Basic Books.

Murphy K.P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Project Gutenberg [www.gutenberg.org]

Rinkler T. (2017). Polarity score (Sentiment Analysis) [https://www.rdocumentation.org/packages/qdap/versions/2.2.9/topics/polarity]

Silge J. and Robinson D. (2017). Text mining with R: A Tidy Approach. O'reilly.

Taboada M., Brooke J., Tofiloski M., Voll K. and Stede M. (2011). Lexicon-based Methods for Sentiment Analysis. Computational Linguistics Vol. 37, Issue. 2, pp. 267-307.

Les hémistiches répétés

Marc Douguet
Université Grenoble Alpes – marc.douguet@univ-grenoble-alpes.fr

Abstract
In this paper, we propose to use the syllabic structure of classical alexandrine in order to automatically identify textual recurrences in French 17th-century theater. The two hemistichs of 6 syllables each present a syntactical unity: consequently, extracting recurrent hemistichs is a way, on the one hand, to hightlight idiomatic expressions characteristic of this period, and, on the other hand, to evaluate the influence of metric constraints on writing.

Résumé
Dans cet article, nous proposons d'utiliser les caractéristiques métriques de l'alexandrin classique afin de repérer automatiquement des récurrences textuelles dans le corpus du théâtre français du XVII^{e} siècle. Les deux hémistiches de 6 syllabes chacun qui le constituent possèdent en effet une unité syntaxique : dès lors, les réemplois fréquents des mêmes hémistiches permettent d'une part de faire émerger les éléments langages propres à ce style d'écriture, et d'autre part d'évaluer l'influence des contraintes métriques sur l'écriture.

Keywords: repeated segments, metre, verse, textual recurrences

1. Introduction

La détection des segments répétés dans un corpus est un outil particulièrement précieux pour l'analyse stylométrique : elle permet à la fois de caractériser le style propre à un auteur, un genre ou une période, et d'évaluer l'originalité d'un auteur par rapport à ses contemporains, sa capacité à s'affranchir ou non des éléments de langage de son époque (cf. notamment Salem, 1987 ; Legallois, 2009 ; Delente et Legallois, 2016). De ce point de vue, l'alexandrin classique présente une caractéristique qui nous semble n'avoir pas encore été totalement exploitée. La césure divise en effet le vers en deux hémistiches d'égale longueur (6 syllabes) qui constituent des unités à la fois rythmiques et syntaxiques. Or ces unités font l'objet de nombreuses répétitions. Par rapport à l'approche qui consiste à extraire tous les segments de n mots pour détecter les récurrences, cette approche (qui la complète) a, pour la stylistique computationnelle de la poésie, un triple avantage :

– elle permet de n'extraire que des segments qui constituent déjà des unités

syntaxiques et évite d'avoir à trier manuellement les résultats pertinents ;
– elle permet d'extraire des segments qui, quel que soit leur nombre de mots, ont le même nombre de syllabes, et sont donc, en régime poétique, d'importance strictement comparable ; – elle permet de mettre en rapport réflexion sur la répétition et analyse de la versification et d'apprécier, notamment, la contrainte que le mètre fait peser sur l'écriture.

2. Méthodologie

Nous avons travaillé sur un corpus de 200 pièces de théâtre en alexandrins publiées entre 1630 et 1680, représentatif de la diversité des genres dramatiques de cette période (tragédie, comédie, tragi-comédie[1]). Le corpus est édité en XML-TEI, avec un balisage qui décrit le découpage en actes, en scènes, en répliques et en vers[2].

Nous avons développé un syllabeur capable de césurer les vers et d'en extraire séparément chacun des hémistiches. Celui-ci est plus modeste que d'autres outils développés en analyse automatique du vers (notamment Beaudouin, 2002 ; Delente et Renault, 2015 ; Salvador, 2016), puisqu'il n'a pas pour ambition de placer avec exactitude la limite entre deux syllabes à l'intérieur d'un mot. Afin de produire un dictionnaire de diérèses et de synérèses, nous l'avons préalablement entraîné en vérifiant manuellement les résultats. Le syllabeur reconnaît automatiquement comme des vers de 12 syllabes 99,98% des 55 031 vers de Corneille dont on a préalablement vérifié qu'ils étaient des alexandrins. La marge d'erreur est uniquement due à l'ambiguïté de certains mots, dont la prononciation change en fonction de la catégorie grammaticale (par exemple « content » et « fier », selon qu'il s'agit de verbes ou d'adjectifs).

Le corpus est composé de 332 938 vers, soit en théorie 665 876 hémistiches. Nous n'en avons retenu que 624 597, après avoir exclu ceux qui était distribués sur plusieurs répliques. Le nombre d'occurrences de chaque hémistiche est calculé après avoir supprimé les ponctuations et les majuscules.

[1] La liste des pièces, les scripts utilisés ainsi que les résultats complets sont disponibles sur https://github.com/marcdouguet/dheform.

[2] Les textes sont disponibles sur https://github.com/dramacode/tcp5. Ils nous ont été fournis par le projet « Bibliothèque dramatique » (http://bibdramatique.paris-sorbonne.fr/), dirigé par Georges Forestier, et le projet « Théâtre classique » (http://theatre-classique.fr/), dirigé par Paul Fièvre. Nous les remercions tous deux d'avoir rendu accessibles leurs sources XML, sans lesquelles ce travail n'aurait pas été possible.

3. Fréquence des hémistiches répétés

Le phénomène de la reprise textuelle des hémistiches est sans commune mesure avec celui, similaire, qui concerne les vers entiers. Dans notre corpus, 499 vers sont répétés au moins une fois, soit seulement 0,1%. Pour quelqu'un qui a une connaissance approfondie du corpus, ces répétitions sont souvent repérables manuellement, et les éditions critiques en soulignent certaines (on connaît notamment le célèbre « Je suis maître, je parle, allez, obéissez » dans *La Mort de Pompée* de Corneille, repris dans *L'École des femmes* de Molière). Les enjeux de ces reprises mériteraient d'être étudiées (plagiat, parodie, citation d'un personnage par un autre, phénomène de refrain, etc.).

La répétition d'hémistiches possède des enjeux différents, à la fois en raison de la brièveté des segments répétés et du très grand nombre de répétitions : 16% des hémistiches du corpus sont répétés au moins deux fois, et un hémistiche y apparaît en moyenne 1,11 fois. L'écriture en vers utilise donc un certain nombre d'éléments de langage et d'idiomatismes préexistants, que le dramaturge combine de manière originale.

En complément des relevés quantitatifs, nous avons également développé une interface de lecture (accessible sur http://obvil.lip6.fr/dheform) : l'utilisateur peut entrer un texte, dont les hémistiches répétés seront mis en évidence à l'aide d'un code couleur.

4. Analyse des hémistiches les plus fréquents

À titre d'exemple, le tableau suivant liste les 10 hémistiches les plus fréquents du corpus, avec leur nombre d'occurrences et deux exemples en contexte :

en cette occasion	119	Que me donne l'amour en cette occasion N'offrez donc point, Seigneur, en cette occasion
en l'état où je suis	98	Que ferai-je, Philante, en l'état où je suis ? Je ne réponds de rien en l'état où je suis.
pour la dernière fois	87	Dites-lui de ma part pour la dernière fois Pour la dernière fois je me jette à vos pieds.
à votre majesté	87	Le respect que je dois à votre Majesté Je me livre, grand Prince, à votre Majesté,
que votre majesté	70	Que votre Majesté le rappelait près d'elle. Ah ! Grand Roi, se peut-il que votre Majesté
en cette extrémité	68	Mettre tout en usage en cette extrémité ; Quoi ? vous m'abandonnez en cette extrémité,
je vous l'ai déjà dit	55	Je vous l'ai déjà dit, sans vous parler de moi, Je vous l'ai déjà dit, j'estime votre flamme,
une seconde fois	51	Je renonce à choisir une seconde fois ; J'en ferais un ingrat une seconde fois.
les armes à la main	42	Les armes à la main, venez si bon vous semble, Laissez-nous lui parler les armes à la main,
de votre majesté	41	Qui vient offrir aux pieds de votre Majesté Il tira des bienfaits de votre Majesté :

Si l'on élargit l'analyse aux 470 hémistiches qui possèdent plus de 10 occurrences, on peut distinguer plusieurs catégories de récurrences. De nombreux hémistiches sont composés d'un substantif de trois syllabes ou plus, précédé de prépositions, de conjonctions et de déterminants, et placé en position de sujet, de complément de nom ou d'objet. Dans cette configuration, on repère plusieurs variations autour d'un même substantif : « à votre majesté » (87 occurrences – nous indiquerons désormais systématiquement le nombre d'occurrences d'un hémistiche entre parenthèses), « que votre majesté » (70), « de votre majesté » (41), « de générosité » (40), « la générosité » (26), « à ma confusion » (30), « cette confusion » (15), etc. Les substantifs concernés relèvent principalement d'une thématique morale ou politique, caractéristique du style d'écriture dramatique du XVII^e^ siècle.

Plus intéressants sont les compléments circonstanciels qui insistent sur le caractère exceptionnel de la situation et sur l'état émotif du locuteur et renforcent ainsi le *pathos* du discours : « en cette occasion » (119), « en l'état où je suis » (98), « en cette extrémité » (68), « en ce malheur extrême » (23), « en cette conjoncture » (22). De nombreuses expressions modalisent l'énoncé : insistance agacée (« je vous l'ai déjà dit » (55)), certitude (« il n'en faut point douter » (37), « il n'en faut plus douter » (25)), prétérition (« je ne vous dirai point » (40)). On notera également la série « pour la dernière fois » (87), « une seconde fois » (51), « pour la première fois » (29), qui relie une situation dramatique à d'autres, passées ou à venir.

Certains syntagmes figés possèdent au contraire une fonction référentielle : violence des relations (« les armes à la main » (42), « un poignard dans le sein » (27)), instinct (« la voix de la nature » (19), pouvoir (« la suprême puissance » (25), « une entière puissance » (24), « un absolu pouvoir » (22)), etc. Les expressions temporelles sont quant à elle nombreuses, et peuvent être associées à une sentence générale décrivant les mœurs du temps (« dans le siècle où nous sommes » (17)) ou à l'urgence d'une situation (« sans tarder davantage », (19)). La fréquence élevée d'« avant la fin du jour » (31) montre à quel point le dramaturges explicitent le respect de l'unité de temps dans leurs œuvres afin d'accroître la tension dramatique. Les expressions spatiales renvoient elles aussi à l'universalité (« sur la terre et sur l'onde » (16)) ou au contraire aux lieux fréquemment convoqués dans le théâtre classique (« dans son appartement » (20), « dans la chambre prochaine » (16)).

Ces expressions figées peuvent souvent être considérées comme des « chevilles », où l'on sent clairement que l'invention verbale se soumet aux contraintes de la métrique. On peut ici identifier deux cas de figure. D'une part, le sémantisme de certains hémistiches circonstanciels est parfois très faible : « en cette occasion », « en l'état où je suis » pourraient aussi bien être

supprimés sans nuire au sens du texte, ou greffés sur n'importe quel énoncé. D'autre part, même si elles sont mieux ancrées dans l'énoncé, les expressions figées que nous avons relevées (« la suprême puissance », « la voix de la nature ») doivent certainement leur succès au fait qu'elle rentrent facilement dans le moule de l'alexandrin. C'est ici l'apposition récurrente d'un adjectif (la puissance sera « entière » ou « suprême »), ou l'utilisation d'une formule imagée (« la voix de la nature », au lieu de « la nature ») qui se justifie par les contraintes de la versification. Il serait intéressant de poursuivre cette analyse en la croisant avec la théorie de la fonction poétique du langage de Jakobson, que résume en partie l'exemple suivant : « *Without its two dactylic words the combination "innocent bystander" would hardly have become a hackneyed phrase.* » (1960 : 358)

5. Vers et prose

Afin d'évaluer la spécificité de l'écriture poétique, nous avons constitué un corpus de pièces en prose de la même époque (11 tragédies de d'Aubignac[3], Baro et Puget de La Serre, et 9 comédies de Molière). Nous avons compté le nombre d'occurrences de chacune des expressions correspondant à un hémistiche récurrent, en le rapportant à la taille respective des deux corpus, calculée en nombre de mots. Certains « hémistiches » (les guillemets s'imposent ici) sont aussi fréquents en vers qu'en prose, mais il n'existe pas de corrélation nette entre les deux corpus, alors même que l'on reste dans le genre dramatique. Or les « hémistiches » que l'on trouve aussi fréquemment en prose qu'en vers, voire plus fréquemment, sont ceux qui reposent à la fois sur un substantif unique (suffisamment long pour occuper les six syllabes avec les déterminants, les prépositions et les conjonctions qui le précèdent) et qui n'ont pas une fonction de complément circonstanciel. Le fait qu'ils figurent parmi les hémistiches les plus fréquents dans le corpus en vers s'explique simplement par le fait que le substantif en question est lui-même extrêmement fréquent. En revanche, les formules figées qui reposent sur une association de plusieurs termes et qui ne font qu'apporter une modalisation sont bien surreprésentées en vers (par exemple « je vous l'ai déjà dit » : 17 occurrences pour un million de mots en vers, 0 en prose ; « il n'en faut point douter » : 12 en vers, 0 en prose ; « pour la dernière fois » : 28 en vers, 9 en prose). Ces expressions, spécifiques au théâtre en vers, semblent donc bien devoir leur suremploi à la nécessité de couler la phrase dans le moule de l'alexandrin.

[3] Nous tenons à remercier ici Bernard J. Bourque, qui nous a fourni la version numérique de son édition Abbé d'Aubignac, *Pièces en prose*, Tübingen, Gunter Narr Verlag, coll. « Biblio 17 », 2012.

6. Premiers et seconds hémistiches

Un des défauts de cette approche est de surévaluer la césure au détriment de l'unité du vers, et de la considérer comme une coupure, une pause entre deux segments indépendants. Deux écueils se profilent. D'un côté, on risque d'oublier que l'hémistiche ne constituent pas toujours, au sein d'un vers, une unité syntaxique pertinente. Les dramaturges du XVII^e^ siècle pratiquent souvent le rejet, le contre-rejet ou l'enjambement internes (par exemple : « Le temps de cet orgueil me fera la raison », dans *La Galerie du Palais* de Corneille). Cependant, notre projet est avant tout lexical, et non prosodique. Isoler les hémistiches n'est qu'une manière de faire émerger des idiomatismes, en se fondant sur le fait que, malgré des exceptions, la césure à l'hémistiche reste le plus souvent la plus forte rupture syntaxique du vers.

Il ne faudrait pas non plus oublier que l'élocution fond les deux hémistiches dans un même mouvement, et que ceux-ci ne se situent donc pas sur le même plan : un poème en alexandrins n'est pas une suite d'hémistiches. Ici, l'analyse automatique à laquelle nous nous sommes livré donne justement des arguments en faveur de l'unité du vers, car elle nous permet de faire émerger plusieurs différences entre les premiers et les seconds hémistiches, qui complètent et confirment les analyses de Beaudouin (2002 : 275-319) concernant la répartition des phonèmes et des catégories morpho-syntaxiques en fonction de la position métrique.

Ils diffèrent tout d'abord dans le taux de répétition. 13% des hémistiches placés en première position sont employés ailleurs dans notre corpus (soit en première, soit en seconde position), ce qui est moins que le pourcentage global de récurrences. Au contraire, ce pourcentage monte à 18% quand on considère les hémistiches placés en seconde position. Cette divergence s'explique facilement par le fait que le second hémistiche n'est pas seulement soumis à la contrainte du mètre, mais aussi à celle de la rime.

Si l'on considère la proportion d'hémistiches qui commencent par un son vocalique, on constate également un déséquilibre : 27% des premiers hémistiches, mais 30% des seconds. La différence est faible, mais elle nous semble permettre de quantifier la contrainte que pose la présence d'un e à la fin du premier hémistiche, qui serait fautive si le second commençait par un son consonantique. Ainsi, tandis que le premier hémistiche peut commencer par n'importe quel son, un hémistiche commençant par un son vocalique est plus facile à placer en seconde position qu'un hémistiche commençant par un son consonantique.

Enfin, les hémistiches les plus fréquents ne sont pas les mêmes selon que l'on considère ceux placés en première et en seconde position. Certains sont utilisés aussi bien à l'une ou l'autre place (par exemple, « en l'état où je suis » apparaît 40 fois en premier, 58 fois en second), mais on observe souvent une

répartition nette : les hémistiches de modalisation de l'énoncé sont plus souvent en premier (« je ne vous dirai point » : 39 pour 1, « je vous l'ai déjà dit » : 52 pour 3 ; « je vous le dis encor » : 20 pour 2), les hémistiches ayant fonction de compléments, en second (« à votre majesté » : 85 pour 2 ; « de votre majesté » : 40 pour 1 ; « à mon ressentiment » : 37 pour 0).

7. Conclusion et perspectives

La détection automatique des récurrences d'hémistiches permet donc de mettre en valeur les contraintes spécifiques qui pèsent sur l'écriture en vers. Même si les conclusions que l'on peut tirer ne font que confirmer un savoir déjà existant, cette méthode nous offre aussi un point d'entrée original dans le corpus du théâtre classique. Elle nous amène à lire autrement ces textes et rend particulièrement sensible, derrière la voix d'un auteur, la voix diffuse d'un style d'époque. À travers ces expressions et ces associations d'idées transparaît tout un imaginaire qui constitue en quelque sorte le « dictionnaire des idées reçus » du XVII^{e} siècle.

Nous n'avons fait là que jeter quelques pistes de réflexion. Un examen quantitatif et qualitatif plus précis est nécessaire pour mieux cerner les enjeux de ce phénomène, tout comme la prise en compte de textes versifiés non dramatiques. Il restera également à étendre le corpus de référence des textes en prose et à définir d'autres principes de comparaison pour évaluer l'influence de la métrique sur la diversité syntagmatique des textes. Envisager les récurrences au niveau, plus abstrait, du motif syntaxique (dans la lignée des travaux de Ganascia, 2001 ; Longrée et al., 2008 ; Mellet et Longrée, 2013 ; Legallois et Prunet, 2015), nous permettra par ailleurs de regrouper des occurrences présentant une structure syntaxique semblable (« la voix de la nature », « le flambeau de la guerre », « les fruits de la victoire ») ou centrées sur les mêmes termes (« qu'on le/la/les fasse venir »). Enfin, la fréquence relative de ces hémistiches récurrents nous paraît être un outil statistique particulièrement prometteur pour évaluer la spécificité du style d'écriture propre à un genre ou un auteur, ainsi que pour observer l'évolution de ces éléments de langage dans le temps.

Références

Beaudouin V. (2002). *Mètre et rythmes du vers classique. Corneille et Racine*. Honoré Champion.

Delente É. et Legallois D. (2016). La répétition littérale dans *Les Rougon-Macquart* : présentation d'un phénomène peu connu. *Excavatio*, vol.28.

Delente É. et Renault R. (2015). Outils et métrique : un tour d'horizon. *Langages*, vol.199 : 5-22.

Ganascia J.-G. (2001). Extraction automatique de motifs syntaxiques. Dans

Maurel D. (éd), *TALN - RECITAL 2001 : 8e conférence annuelle sur le Traitement Automatique des Langues Naturelles*.

Jakobson R. (1960). Closing statements: Linguistics and Poetics. Dans Sebeok T. A. (éd), *Style in Language*. The Technology Press of MIT/John Wiley and Sons, inc.

Legallois D. (2009). À propos de quelques n-grammes significatifs d'un corpus poétique du XIXe siècle. *L'Information grammaticale*, vol.121 : 46-52.

Legallois D. et Prunet A. (2015). Sequential patterns: a new corpus-based method to inform the teaching of language for specific purposes. *Journal of Social Science*, vol.44 : 127-140.

Longrée D., Luong X. et Mellet S. (2008), Les motifs : un outil pour la caractérisation topologique des textes. Dans Heiden S. et Pincemin B. (éds), *JADT 2008. 9es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 733-744.

Mellet S. et Longrée D. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages*, vol.189 : 65-79.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Klincksieck.

Salvador X.-L. (2016). Versification : outil d'analyse du mètre français (http://www.projetprada.fr/versification et https://gist.github.com/xavierLaurentSalvador).

«Mangiata dall'orco e tradita dalle donne». Vecchi e nuovi media raccontano la vicenda di Asia Argento, tra storytelling e Speech Hate

Francesca Dragotto[1] Sonia Melchiorre[2]
[1]Università di Roma Tor Vergata – dragotto@lettere.uniroma2.it
[2]Università della Tuscia – melchiorresmr@unitus.it

Abstract 1
Re-enacted and dissected in the National and International news, the narration of the rape denounced by Italian actress Asia Argento has triggered several coming outs revealing the violence perpetuated against other actors and actresses by prominent personalities of the Hollywood star system. Textually molded between diffused narration and the blink of a tweet, the story has hooked the public displaying, in the Italian media in particular, a morbid legitimation of Victim Blaming. Asia Argento has become the object of Hate Speech revealing, in turn, a cultural palimpsest of lies and guilty silences deriving from stereotypes represented in comments of the most crass and basest order. The present discussion starts therefore from a quantitative and qualitative analysis of texts, in English and Italian, reporting the story and aims to reveal the similarities and differences between language practices substantiating the discourse of violence. Another corpus derived from the social networks will also reveal the righteous indignant reactions of cybernauts concerning this story which will help identify the language patterns at the core of gender-based violence.

Abstract 2
Spolpata dalle cronache nazionali e internazionali, la narrazione della violenza sessuale denunciata dall'attrice italiana Asia Argento ha funto da detonatore di una esplosiva sequela di coming out rivelatori di episodi analoghi subiti, da altre attrici e, seppur in misura inferiore, attori, da parte di personaggi di spicco dello Star System hollywoodiano. Colata in tutti gli stampi testuali compresi tra la narrazione diffusa e il succinto tweet, la trama di questa vicenda ha tenuto e ad oggi ancora tiene significativo banco mediatico, alimentando un dibattito che, nel caso italiano, si è dimostrato spesso più interessato all'individuazione di ragioni utili a legittimare il Victim Blaming che a ricostruire le coordinate del contesto in primis psicologico nel quale si sarebbe consumata la violenza. Oggetto di innumerabili discorsi di odio, il racconto rappresentato dalla cronaca italiana

costituisce un oggetto utile a investigare il sentimento sociale nei confronti di storie di violenza con protagoniste persone (in special modo donne) famose, nei confronti delle quali si attivano reazioni di sdegno frammisto alla colatura dei più beceri stereotipi di genere. Muovendo dall'analisi quantitativa e qualitativa di un corpus di testi incentrati su questa vicenda, prodotti in lingua inglese e in lingua italiana, chi scrive si ripropone di far emergere luoghi di contatto e di separazione tra le diverse forme della cronaca, unitamente alle costellazioni lessicali, semantiche e pragmatiche che le hanno sostanziate. Correderà questa analisi quella di un secondo corpus, stavolta estrapolato dalla ricca produzione social riconducibile ad account ora individuali, ora di gruppi noti per l'indefessa attività di comunicazione indignata intorno a vicende dell'attualità. Scopo ultimo del lavoro, sarà l'intercettazione dell'eventuale pattern linguistico e concettuale della violenza di genere, del quale si testeranno i limiti di validità all'interno di sistemi diversi e di varietà diverse dello stesso sistema.

1. La narrazione

Umiliata e offesa. Questo il destino toccato all'attrice italiana Asia Argento, tra le prime a denunciare la violenza subita dal produttore cinematografico hollywoodiano Harvey Weinstein. La donna ha avuto il coraggio di esporre pubblicamente il suo stupratore assieme a una ottantina di altre, che come lei, hanno subito prima un oltraggio fisico e successivamente un'esposizione mediatica senza precedenti. Appare significativo da un punto di vista narrativo, che la vicenda sia stata innescata da un tweet e che sia successivamente rimbalzata nei media di tutto il mondo. Nel breve lasso di un cinguettìo Asia Argento rivela i nomi di tutte le donne che con coraggio hanno denunciato la violenza perpetrata nei loro confronti da un uomo che si credeva potente e intoccabile. Ed ecco che dal racconto delle vittime scaturisce una nuova narrazione in cui le donne diventano *survivors*, dando voce alla loro rabbia contro un sistema patriarcale, sessista e misogino, condensato in uno slogan già storico: *Me too*, "anche io", nel quale tutte le donne del mondo vittime di violenza si sono riconosciute. È accaduto poi che due parole si transustanziassero nella *Person of the Year* 2017 guadagnando la copertina del *Time*, che si incarnassero nei corpi abbigliati di nero di tutte le attrici che hanno partecipato al *Golden Globe 2018* e che, infine, si trasformassero nel *Time's Up*, "Il tempo è scaduto", refrain che si propone come impulso trasformatore della rabbia in forza (ri)costruttrice e che, probabilmente, accompagnerà l'afro-americana Ophra Winfrey nella corsa per la Casa Bianca. In Italia, nel frattempo, si fatica, e molto, ad ammettere perfino che le parole usate dai media nel caso Asia Argento dimostrino l'esistenza di un grave problema culturale. Nel nostro paese parole tossiche, nell'insieme dette *hate speech*, hanno condotto a un vergognoso *victim blaming*

nei confronti di Asia Argento: una etichetta eufemistica per le orecchie italiane che finisce però per assumere la forma testuale di un testo argomentativo dalle cui trame scaturisce violenza e accanimento mediatico – ironia sprezzante e spregiativa nei casi migliori – non già nei confronti degli aggressori, bensì delle persone vittime di violenza sessuale. Questa tendenza ben si evince dalla disamina, anche solo cursoria, di testi recuperabili dal web. In questa sede ne è stata raccolta una selezione, in lingua italiana e inglese, successivamente sottoposta ad analisi contrastiva. Dall'analisi è emersa la tendenza all'uso di una terminologia, sistematicamente sostenuta da toni aggressivi, rivelatrice di un sistema più complesso di collusione culturale con un sistema che sarebbe frettoloso liquidare come fallocentrico e misogino e percorso da una omosocialità maschile da spogliatoio. Portatrice di significato per quanto e come dice, ma anche per quanto non dice, la lingua di questi testi (e in generale di ogni testo), costituisce infatti una porta di accesso all'architettura ideologica che la sorregge e che sorregge le coordinate di chi se ne serve: una architettura che cela un mondo sclerotizzato, che nel caso in questione prevede un pendant tra atteggiamento aggressivo di chi offende e lesione della dignità di chi offeso/a, su cui è necessario gettare luce se si vogliono comprendere le dinamiche che guidano l'agire in questa porzione di tempo che vede la vita sociale e comunicativa governata dalle strutture dei social media. In attesa dei risultati dell'analisi di un corpus meglio strutturato e più tendente alla sistematicità – con tutti i limiti che la sistematicità applicata al testo inteso in senso cognitivo può avere – in questa prima fase si procederà con l'esposizione dei nuclei più significativi ottenuti per carotaggio. I frammenti proposti sono stati scelti perché rappresentativi ciascuno di un corpus dalle caratteristiche analoghe.

1.1 Victim blaming

Queste alcune delle domande proposte ad Argento da G.M. Tammaro, de *La Stampa* (15 ottobre 2017), a immediato ridosso della denuncia pubblica dell'attrice. Difficile non rintracciarvi lo schema narrativo plurisecolare dell'interrogatorio della vittima di violenza (si pensi, uno su tutte, al primo processo per stupro della storia, quello nei confronti della pittrice Artemisia Gentileschi, e non dell'aggressore Agostino Tassi, del 1612): il testo-genere si compone di domande alle quali chi ha subito violenza deve rispondere in maniera dettagliata per non essere tacciata di collusione con il predatore.[1] In grassetto gli elementi che si ritengono rilevanti per il discorso.

[1]http://www.lastampa.it/2017/10/15/italia/cronache/un-orco-mi-ha-mangiata-la-cosa-pi-sconvolgente-i-tanti-attacchi-dalle-donne-hUwq9t9TFgRHkmcjU8yhAL/pagina.html (ultimo accesso 11/01/2018).

1. Perché ha deciso di rivelare questa storia a distanza di tanti anni?
2. Non pensa che parlare prima avrebbe evitato che altre donne subissero come lei?
3. Che cosa l'ha ferita maggiormente?
4. E lei come reagisce?
5. Come ha vissuto questi anni di silenzio?
6. Si sente ancora in colpa per questo?
7. Che cosa temeva che le potesse accadere, in caso di denuncia all'epoca dei fatti?
8. Fabrizio Lombardo, ex capo di Miramax Italia, nega di averla portata da Harvey Weinstein, come lei invece sostiene.
9. Dopo il primo incontro in un hotel in Costa Azzurra, lei iniziò una relazione con Weinstein?
10. Weinstein cercò di contattarla ancora?
11. Lei accettò?
12. Qual era l'atteggiamento di Weinstein nei suoi confronti?
13. Come cambiò il suo comportamento, nei confronti di Weinstein?
14. Quindi vi incontraste altre volte?
15. Poi però ha deciso di farsi avanti in prima persona: come mai?
16. In Italia non tutti la pensano così. Non tutti le credono. Non tutti stanno dalla sua parte.
17. La accusano anche di aver firmato la petizione a favore di Roman Polanski, indagato per pedofilia.
18. Si è pentita?
19. Dopo essersi fatta avanti insieme alle altre donne e aver raccontato quello che le è successo, cosa spera che accada?

Poste una di seguito all'altra, le domande assumono la forma di una narrazione a se stante, caratterizzata da una costellazione di termini e da una semantica incentrata sulla vittima non in quanto tale ma in quanto teste che deve fornire spiegazioni per quanto accaduto, per giustificare il suo silenzio. Quelle a seguire sono invece alcune delle frasi pronunciate, a vario titolo, da Mario Adinolfi, Vittorio Feltri e Vittorio Sgarbi, rimbalzate tra numerosi siti e quotidiani del mondo, tra i quali il *New Yorker*, per primo, il *Guardian* e l'*Independent*. L'articolo di *The Guardian* riporta, per esempio, le seguenti parole: "Far from being hailed as brave, Argento's allegations were initially treated in some Italian media outlets with a mix of scepticism and scorn" dove colpisce il pendant tra il *brave*, 'coraggiosa', utilizzato dalla giornalista per definire Asia Argento, e l'atteggiamento generalizzato di 'scetticismo' e 'disprezzo, disdegno' (*scorn* rimanda anche all'idea di 'rifiuto', di non accettazione di qualcosa che viene proposto). La giornalista riporta poi le

parole di Asia Argento: "Here people don't understand. They'll say, 'oh it's just touching tits'. Well yeah, and this is a very grave thing for me. It is not normal. You can't touch me, I am not an object". Il pezzo non omette la descrizione dettagliata della violenza subita dall'attrice e il commento offensivo di Vittorio Feltri, che sminuisce l'atto sessuale poiché solo sesso orale (*licking* e non *oral sex* nella sua interpretazione). L'elemento più rilevante dell'articolo resta una delle frasi conclusive della giornalista: "For now, not a single fellow female actor who is well known has spoken out in support of her, even though the Italian film industry is rife with abuse", dove *rife with abuse* rimanda da un lato alla reiterazione di atti, dall'altro, significando *rife* 'pieno zeppo', allude anche a un atteggiamento collusivo di quanti con comportamento omertoso non denunciano. In un altro articolo, sull'*Independent*, sempre in Gran Bretagna, Lydia Smith scrive: "But she was subsequently criticised by some sections of the Italian media for not coming forward sooner about the alleged assaults, despite hesitation being common among survivors for fear of reprisals, among other reasons. […]". Riporta poi gli interventi di Renato Farina apparsi su *Libero* e i suoi commenti *Victim Blaming* volti cioè alla colpevolizzazione della vittima e tipico di chi è rimasto molto, troppo indietro rispendo a un mondo che va veloce:[2] "Conservative newspaper *Libero* published an op-ed by Renato Farina, with the headline: 'First they give it away, then they whine and pretend to repent'".[3]

1.2 Hate speech

"Se denunci uno stupro in Italia sei tu la troia". E, ancora, "Solo in Italia vengo considerata colpevole del mio stupro perché non ne parlai quando avevo 21 anni", denuncia Asia Argento dopo le critiche e le aggressioni verbali ricevute sui media italiani, anche da parte di star, che insinuano o apertamente dichiarano che "Si può sempre dire di no...". Il 13 ottobre **Asia Argento** torna sul caso Weinstein con un tweet amaro: "Ho denunciato uno stupro e per questo vengo considerata una tr...". Ma il mondo dello spettacolo affronta la questione in modo che è eufemistico definire prudente. "Conosco bene Asia Argento e la stimo", rivela Vladimir Luxuria. "Quando ho letto che raccontava di essere stata costretta a un rapporto orale, la prima reazione è stata di solidarietà. Ma quando ho letto che, dopo aver subito questa violenza, ha fatto un film con lui, è andata con lui sul red carpet a

[2] http://www.liberoquotidiano.it/news/opinioni/13264032/harvey-weinstein-renato-farina-scandalo-sessuale-hollywood.html (ultimo accesso 08/01/2018).

[3] http://www.independent.co.uk/arts-entertainment/films/news/harvey-weinstein-sexual-assault-asia-argento-flees-italy-public-condemn-speaking-out-a8012511.html (ultimo accesso 08/01/2018).

Cannes, l'ha frequentato per cinque anni, allora mi sono detta che c'era qualcosa che non andava. Purtroppo in queste vicende bisogna avere una credibilità totale, altrimenti basta una sola fake news a mettere in discussione tutto: […]". **Ottavia Piccolo**, stimatissima attrice di teatro e cinema, preferisce sorvolare: "Sono cose che sono sempre accadute, non voglio parlarne perché rischierei di dire solo banalità". Mentre **Rita Dalla Chiesa** affronta senza timore l'argomento: "Sicuramente la paura di perdere il lavoro può esserci. Se però una persona si è sentita realmente offesa e traumatizzata ma poi, invece di scappare, resta all'interno di questo cerchio negativo, prende treni e aerei e va agli appuntamenti in albergo, non parlerei più di stupro, ma di un rapporto cosciente". Cita poi le parole di Barbara Palombelli, con le quali afferma di concordare: "[…] Sei stata violentata? E perché lo dici dopo anni? Troppo comodo. Non facciamo battaglie femministe su cose che col femminismo non c'entrano niente". "Sarò una mosca bianca", rivela **invece Alba Parietti**, "ma a me non è mai capitato niente del genere. A volte basta l'atteggiamento per scoraggiare un uomo. Il punto centrale del problema è la paura: l'eterna paura delle donne nei confronti degli uomini, del loro potere, di non essere credute. Conosco potenti donne manager che quando tornano a casa si lasciano menare dal marito. Perché questo tipo di atteggiamento non riguarda solo il mondo dello spettacolo, ma tutti gli ambiti lavorativi. Con un'aggravante: nello spettacolo non insegui un posto da 1200 euro al mese, ma fama e successo". Il 26 ottobre 2017 Guia Soncini, editorialista della rivista *Gioia*, commenta sul *New York Times* il fallimento del femminismo italiano, riferendosi alla vicenda di Asia Argento:[4] "This episode is another example of my country just being male-run, sexist Italy […] This, in a country that has a total of zero national newspapers edited by women and zero female columnists in its main national papers. […] Where the reaction to Ms. Argento's account has been truly vicious has been on social media. And there, it has primarily come from women. […] What this tells us about Italian feminism isn't clear, but it's certainly ugly. […] There's something under-ripened about the state of feminism in my country". Peccato che Soncini avesse postato un tweet decisamente poco femminista ("Sogno un pezzo su Weinstein d'una sola riga. Quello sarà un vecchio porco, ma voli gliela tiravate con la fionda, finché pensavate servisse") qualche giorno prima (10 ottobre 2017), cosa che non sfugge propria ad Asia Argento. L'attacco più diretto è quello sferrato, via Facebook, da Selvaggia Lucarelli in un post

[4] https://mobile.nytimes.com/2017/10/26/opinion/italian-feminism-asia-argento-weinstein.html?partner=IFTTT &_r=0&referer=https://t.co/pj6FLcp4Fx (ultimo accesso 10/01/2018).

molto lungo:[5] "Ora. Francamente. Vai a letto con un bavoso potente per anni e non dici di no per paura che possa rovinare la tua carriera. Legittimo. Frigni 20 anni dopo su un giornale americano raccontando di tuoi rapporti da donna consenziente tra l'altro avvenuti in età più che adulta, dovendo attraversare oceani, con viaggi e spostamenti da organizzare, dipingendoli come "abusi". Meno legittimo. Ad occhio, sono abusi un po' troppo prolungati e pianificati per potersi chiamare tali. E se tu sei la prima a dire che lo facevi perché la tua carriera non venisse danneggiata, stai ammettendo di esserci andata per ragioni di opportunità. Nessuno ti giudica, Asia Argento. Però ti prego. Paladina delle vittime di molestie, abusi e stupri, anche no. Facciamo che sei finita in un gorgo putrido di squallidi *do ut des* e te ne sei pentita. Con 20 anni di ritardo però".[6]

1.3 La sindrome di Stoccolma

All'inizio di quest'anno i media hanno riportato la notizia dell'ennesimo femminicidio in Italia. Si scopre e ci si meraviglia che la donna, bruciata viva dal suo convivente, abbia più volte difeso il suo aggressore. Questo atteggiamento ha un nome: Sindrome di Stoccolma, una sindrome che sembra colpire tante donne e il cui effetto andrebbe per lo meno valutato anche per spiegare le reazioni delle tante donne che hanno reagito attribuendo la responsabilità di quanto accaduto alla Argento, chiamando del tutto fuori il suo aggressore. Natalia Aspesi, femminista e donna di cultura, ha sostenuto che "Se mi chiedi un massaggio in ufficio e io te lo concedo, poi non mi posso stupire su come va a finire". E, ancora, "Che i produttori, almeno da quando ho memoria di vicende simili, hanno sempre agito così. E le ragazze, sul famoso sofà, si accomodavano consapevoli. Avevano fretta di arrivare. E ancor più fretta di loro avevano le madri legittime che su quel divano, senza scrupoli di sorta, gettavano felici le eredi in cerca di un ruolo, di un qualsiasi ruolo".[7] "L'eccezione alla regola proposta è Sofia Loren, che sposò un produttore per proteggersi – afferma ancora Aspesi – da attenzioni indesiderate". A chi le chiede se stia giustificando Weinstein, risponde inoltre "Non giustifico niente. Il femminismo è ancora una delle missioni più importanti per le donne di tutto il mondo, forse la più importante in assoluto.

---

[5]https://www.leggo.it/gossip/news/asia_argento_stuprata_da_weinstein_selvaggia_lucarelli_frigni_dopo_20_anni_foto_video_11_ottobre_2017-3295503.html (ultimo accesso 10/01/2018).

[6]https://www.leggo.it/spettacoli/cinema/asia_argento_weinstein_sfogo_twitter_12_ottobre_2017-3297028.html (ultimo accesso 09/01/2018).

[7] https://www.vanityfair.it/news/approfondimenti/2017/10/11/weinstein-commento-natalia-aspesi (ultimo accesso 11/01/2018).

È qualcosa in cui ho creduto e credo ancora ciecamente. Ma non mi pare che con queste denunce possa fare un salto decisivo. Magari sbaglio, ma ho i miei dubbi". Il dubbio "Che sia una vendetta fratricida, per togliere di mezzo Weinstein. Era un produttore potente come pochi e sporcaccione come moltissimi altri. Che la storia, risaputa da decenni, sia venuta fuori con questa virulenza soltanto adesso, accompagnata da decine di testimonianze, non può essere casuale". A completare la rassegna, un articolo, senza firma, battuto da ADN Kronos (13/10/2017), che già col solo titolo riesce a sintetizzare lo stato della polemica *Donne che odiano le donne, gogna social per Asia Argento*: "[…] E nel marasma dei commenti social che la accusano di volta in volta di opportunismo, di prostituzione, di sensazionalismo, a colpire duro incredibilmente sono soprattutto le donne. Man mano che si scorrono i commenti agli articoli dedicati al caso in questi giorni dai principali quotidiani, non è infatti difficile incappare – anzi, è impossibile – nei tanti insulti lanciati contro l'attrice: a scriverli sono mamme, nonne, ragazze, studentesse, tutte convinte della colpevolezza di Asia Argento, rea nel migliore dei casi per chi commenta di aver aspettato troppo a parlare o, nel peggiore, di essersi prostituita in cambio di un posto al sole di Hollywood".[8]

1.4 La decisione di lasciare l'Italia

"Newspapers 'slut-shamed' Asia Argento so badly over the Weinstein saga that she's leaving Italy",[9] riporta spesso la stampa straniera nel dar conto dell'evoluzione della saga di Asia Argento, giudicata coraggiosa e ispiratrice di altre donne. Fuor di patria. "Part of the criticism from some Italian newspapers and social media users revolves around the counter-argument that these celebrities should have come forward years ago (we debunked this argument here). While these newspapers and internet users are hardly the only ones engaging in this form of victim-blaming, the violent tone used by some is alarming and astonishing […].". Cita quindi il caso di Renato Farina. La reazione sorprende ancor più la stampa straniera che ha un mezzo di facile paragone nella solidarietà riservata alle attrici americane protagoniste di analoghe denunce nei confronti di Weinstein. Giunta a Laura Boldrini la notizia dell'espatrio volontario, la Presidente della Camera indirizza il proprio appello all'attrice chiedendole di desistere dai suoi propositi: «Resta

[8] http://www.adnkronos.com/fatti/cronaca/2017/10/13/donne-che-odiano-donne-gogna-social-per-asia-argento_4KNSPMO49OoLtVvox04GWN.html

[9] http://mashable.com/2017/10/18/asia-argento-harvey-weinstein-sexual-harassment-slut-shaming/#YIIO

i.0cNaql

in Italia, non mollare».[10] Da sempre impegnata in attività contro la violenza sulle donne, da New York ha commentato al *Corriere della Sera*: "Non ho avuto modo di chiamare Asia Argento perché sono in missione a New York e in Canada. Le mando, però, questo messaggio**:** bisogna rimanere in Italia per rafforzare la solidarietà tra donne. Asia non mollare"**.** Ha poi aggiunto "Detesto il fatto che Asia Argento debba arrivare a giustificarsi […]. Questo è il mondo alla rovescia, non è importante se e quando una donna decide di denunciare un abuso. Queste sono sue scelte. Lo scandalo è che un uomo di potere, questo Weinstein, si sentiva libero di saltare addosso alle ragazze che volevano lavorare. Questo è il sistema marcio che va sradicato". La stessa presidente della Camera non è del resto estranea all'azione denigratrice del web, che ne ha spesso fatto la destinataria di valanghe di insulti e parole violente. Riporta, tra gli altri, l'intervento di Boldrini il quotidiano *Libero*, che,[11] il 19 ottobre 2017, titola *Laura Boldrini: "Cara Asia Argento resta in Italia, le donne sono con te"* un articolo parco di commento ma nel quale la lingua non rispettosa del genere e della morfologia della lingua italiana – su tutti *la presidenta* – comunica ben più di quanto avrebbero fatto molte parole: "'Per quanto riguarda le molestie e gli stupri', ha sottolinea[to n.d.r.] la presidenta, 'il problema sono gli uomini e il loro comportamento […]'".

2. Considerazioni finali

In attesa di uno scandalo a ruoli capovolti, che, da stereotipi culturali e linguistici dominanti, ad oggi lascerebbe prefigurare tutt'altro genere di commenti, ci si limiterà a una rosa di citazioni che se anche ampliata notevolmente non riuscirebbe a spostare di una virgola – chi scrive ne è convinta – lo stato di polarizzazione che si è venuto a prefigurare in Italia fin dai primi giorni di diffusione della vicenda. Una polarizzazione oppositiva che richiama quella tipica del tifo e più di recente della fede politica – che sembra rendere incapaci di acquisire, anche solo provvisoriamente, una prospettiva diversa, anche solo in parte, da quella originaria, – alla quale nessun commento sembra potersi sottrarre. Ragion per cui, per evitare che anche l'approccio descrittivo tipico dell'analisi del testo possa essere accusato di faziosità da una o dall'altra parte, occorrerebbe ampliare il corpus di riferimento di questo lavoro almeno con la disamina quantitativa e qualitativa di tutti i tweet presenti nell'account di Asia Argento con riferimento ai profili che li hanno generati; con la disamina almeno

[10] https://www.vanityfair.it/news/cronache/2017/10/19/caso-weinstein-laura-boldrini-asia-argento

[11] http://www.liberoquotidiano.it/news/politica/13266009/laura-boldrini-cara-asia-resta-in-italia-donne-sono-con-te-minigonna-uomini.html

quantitativa dei segmenti e dei contesti in cui il termine *vittima* compare esplicitamente o è richiamato in altro modo; con la disamina dei contesti e delle forme cui si ricorre per parlare di chi ha offeso, con l'attività social scaturita dalle cronache relative a momenti clou dell'anno in materia di violenza o di rivendicazione di genere, nello specifico nei confronti delle donne, quali la giornata contro la violenza sulle donne o l'8 marzo. Già attuata a campione, la raccolta e la successiva analisi di messaggi mostra una pervicace azione a ripetere impermeabilmente le proprie azioni comunicative, tanto nei contenuti tanto nella forma e nelle costellazioni di termini che accompagnano il focus di volta in volta oggetto di discussione. Segno inequivocabile della posizione che gli elementi da cui si irradia la costellazione stessa hanno nell'enciclopedia e nella coscienza e sensibilità della comunità linguistica italofona.

Il *cosa* e il *come* del processo narrativo. L'uso combinato della Text Analysis e Network Text Analysis al servizio della precarietà lavorativa

Cristiano Felaco[1], Anna Parola[2]
Università degli Studi di Napoli Federico II – cristiano.felaco@unina.it; anna.parola@unina.it

Abstract
This paper shows the analytic procedures in order to use jointly Text Analysis and Network Text Analysis. Text Analysis allows to detect the main themes subjects in the narrations and hence the processes of signification, Network Text Analysis permits to track down the relations between linguistic expressions of text, identifying therefore the path of flow of thoughts. Using jointly the two methods is possible not only to explore the content of narrations, but, starting from the words and concepts with higher semantic strength, also to identify the processes of signification. To this purpose, we will present a research aiming to understand high school students' perception of employment precariousness in Italy. The lexical corpus was built by narrations collected from 2013 to 2016 in blog of Repubblica "Microfono Aperto".

Riassunto
Il lavoro presenta le procedure analitiche per un uso congiunto delle tecniche di Text Analysis e Network Text Analysis. La prima permette di cogliere i temi principali affrontati nelle narrazioni e quindi i processi di significazione, la seconda di rintracciare le relazioni tra le espressioni linguistiche di un testo, individuando i percorsi dei flussi di pensiero. L'uso combinato delle due tecniche permette, dunque, non solo di esplorare i contenuti delle narrazioni, ma, lavorando su parole e concetti con una maggiore carica semantica, anche di ricostruire i percorsi attraverso i quali si costruisce il significato. A tale scopo sarà presentata una ricerca volta a comprendere la percezione degli studenti delle scuole secondarie superiori sulla precarietà lavorativa in Italia. Il corpus testuale è stato creato a partire dalle narrazioni raccolte dal 2013 al 2016 nel blog di Repubblica "Microfono Aperto".

Keywords: Thematic Analysis of Elementary Contexts; Network Text Analysis; Employment Precariousness; Students.

1. Introduzione

La narrazione, e più nello specifico il narrare, è un processo di costituzione di una tessitura testuale dotata di senso e veicolante significati. Analizzare i testi permette di cogliere da un lato la percezione di chi narra su un dato argomento e il processo di significazione attribuita all'esperienza narrata, ma dall'altro di comprendere i flussi di pensiero, entrando nello specifico delle parole utilizzate e della loro sequenzialità. L'uso della statistica testuale al servizio delle narrazioni permette, perciò, il riconoscimento in profondità del significato delle parole e del senso ivi presente (Bolasco, 2005). Tra le tecniche di analisi del contenuto, l'uso combinato della Text Analysis (TA) e Network Text Analysis (NTA) si presta bene a questi scopi. Se la TA permette di cogliere i temi affrontati, le parole scelte e utilizzate e le dimensioni di senso attribuite (Lebart et al., 1998), il *cosa* si narra, l'uso della TNA offre un ulteriore approfondimento sul *come* si narra. Analizzando, infatti, la posizione delle parole all'interno della rete testuale è possibile rintracciare le parole con una maggiore carica semantica, individuando in questo modo i diversi percorsi e contesti di significato (Hunter, 2014) mediante lo studio della natura delle relazioni tra i vari termini. Partendo dall'assunto che la struttura di relazioni tra le parole di un testo possa corrispondere ai modelli mentali e alle mappe cognitive messe in atto dagli autori del testo (Carley, 1997; Popping et Roberts, 1997), tale metodo permette di modellizzare il linguaggio come rete di parole e di relazioni attraverso la creazione di una mappa cognitiva (Popping, 2000). Il concetto è il nucleo (mentale) che viene rappresentato attraverso un termine o un'espressione linguistica; i termini possono essere in relazione tra loro formando un'affermazione. Le affermazioni che condividono uno stesso concetto formano una struttura interdipendente creando così una mappa concettuale o rete testuale costituita da punti (o nodi) che rappresentano le singole parole (o concetti) e da linee, cioè i legami che li collegano.

2. Metodologia

L'approccio proposto prevede dapprima che i testi prodotti siano sottoposti ad un'analisi statistica dei dati testuali servendosi del software di analisi automatica T-lab, e successivamente analizzati in una prospettiva di rete mediante il software Gephi.

2.1 Pre-trattamento dei testi

Raggruppati all'interno di un unico corpus, la prima fase di lavorazione del testo si compone di una fase di normalizzazione del corpus e di personalizzazione del dizionario. La prima ha l'obiettivo di riconoscere le parole come forme grafiche e ciò comporta una trasformazione del corpus

(eliminazione di spazi vuoti in eccesso, marcatura degli apostrofi, riduzione delle maiuscole), e la creazione di stringhe per le locuzioni polirematiche, insiemi di parole che hanno un significato unitario non desumibile da quello delle parole che lo compongono, arrivando alla creazione delle *multiwords*. La fase di personalizzazione del dizionario è effettuata con le procedure di lemmatizzazione e disambiguazione del testo che permettono di rinominare le forme grafiche in lemmi. Lo step della disambiguazione permette di selezionare le forme omografe per disambiguarle; quello di lemmatizzazione, partendo dal riconoscimento delle forme con la stessa radice lessicale (lessema) o appartenenti alla stessa categoria lessicale, di ricondurre ogni aggettivo e sostantivo al maschile singolare, ogni verbo alla forma di infinito presente, e così via. Terminata questa fase, si procede al controllo delle caratteristiche lessicali del corpus per comprenderne la trattabilità a livello statistico, verificando i valori del *type/token ratio*, adeguato per un valore inferiore a 0.2, e gli *hapax*, adeguato per una percentuale inferiore al 50% per corpus di grandi dimensioni, e per percentuali leggermente superiori in caso di corpus di medie o piccole dimensioni. Prima di procedere all'analisi, va, inoltre, presa visione della lista delle parole chiave, creata con una procedura automatica dal software, e alla loro occorrenza all'interno del corpus, e si fissa una soglia di occorrenza minima, escludendo dall'analisi tutte le parole presenti meno di n. volte. La scelta della soglia di occorrenza dipende dalle caratteristiche lessicali e dalle dimensioni del corpus in analisi. Le parole chiave possono dunque essere prese nella loro integrità, ridotte in relazione alla soglia di occorrenza, o ancora ulteriormente ridotte in base agli scopi della ricerca.

2.2. Analisi dei testi mediante Analisi Tematica dei Contesti Elementari

L'Analisi Tematica dei Contesti Elementari mediante una Cluster Analysis permette di costruire ed esplorare i contenuti del corpus in analisi (Lancia, 2004). I cluster sono costituiti da un insieme di contesti elementari definiti dagli stessi pattern di parole chiave e descritti attraverso le unità lessicali che maggiormente vanno a caratterizzare i contesti elementari. La cluster analysis è eseguita mediante un metodo gerarchico-ascendente non supervisionato (algoritmo bisecting K-means), caratterizzato dalla co-occorrenza dei tratti semantici. Nello specifico, la procedura d'analisi è costituita da: analisi delle co-occorrenze mediante la creazione di una tabella dati unità di contesto*unità lessicali con valori di presenza/assenza; pre-trattamento dei dati tramite TF-IDF e trasformazione di ogni vettore riga a lunghezza 1 (norma euclidea); uso del coseno e clusterizzazione tramite algoritmo bisecting K-means; analisi comparativa con creazione della tabella di contingenza unità lessicali*cluster; test del chi-quadrato agli incroci

cluster*unità lessicali. Rispetto al criterio di partizione che determina il numero dei cluster, viene utilizzato un algoritmo che utilizza il rapporto tra varianza intercluster e varianza totale assumendo come partizione ottimale quella in cui questo rapporto supera la soglia del 50%. L'interpretazione della posizione occupata dai cluster nello spazio fattoriale e delle parole che li caratterizzano permettono di individuare le relazioni implicite che organizzano il pensiero dei soggetti, consentendo di cogliere il punto di vista del narratore nei confronti dell'evento narrato. Quest'ultimo comprende anche una serie di elementi valutativi, riflessioni, significati, giudizi di valore, ma anche proiezioni affettive.

2.3. Analisi delle reti

Il secondo step d'analisi prevede l'inserimento del corpus all'interno del software Gephi. Tale software organizza i vari lemmi in una matrice di adiacenza (lemma*lemma) consentendo la creazione di una rete *1-mode*, uno strumento utile per visualizzare la struttura di relazioni tra i vari lemmi, rappresentati da cerchi o nodi, e collegati tramite legami rappresentati da linee direzionate. Tale tecnica permette di cogliere il modo con cui i nodi sono connessi tra loro, identificando così le zone di vicinato (*neighbourhood*), e individuando quei nodi che occupano una posizione di rilevanza in differenti set o nell'intero network. A tale scopo, vengono calcolate differenti misure basate sulla centralità e, tra queste, la *degree centrality* che indica le parole usate con maggiore frequenza in connessione ad altre parole all'interno delle narrazioni e nei vari contesti di significato. Più nel dettaglio, l'incidenza di ogni nodo può essere espressa sia come *in-degree*, numero di archi entranti in un punto, individuando in questo modo i cosiddetti "predecessori" di ogni unità lessicale, sia come *out-degree*, numero di archi uscenti dal punto, mostrando invece i "successori". Tale relazione tra predecessori e successori all'interno della rete testuale aiuta a comprendere la varietà semantica generata dai nodi. Altro indice utilizzato è la *betweennes centrality*, misura di centralità globale basata sulla vicinanza, che esprime il grado con cui un nodo sta "fra" gli altri nodi del grafo. I nodi collocati in queste zone del network eserciterebbero una funzione di controllo sui flussi informativi e di "passaggio" permettendo il collegamento tra due o più set del network (Freeman, 1979). Nell'ottica dell'analisi testuale, questi lemmi, infatti, giocano un ruolo centrale nella circolazione dei significati all'interno della rete, fungendo da punto di giunzione da cui si connettono zone diverse di testo e si snodano specifici percorsi di significato, andando a definire in questo modo la varietà semantica delle narrazioni.

3. Caso studio

Presentiamo uno studio condotto attraverso l'uso combinato delle tecniche allo scopo di comprendere la percezione degli studenti del mondo del lavoro nel contesto italiano. Gli ultimi dati disponibili mostrano che l'Italia è tra i paesi europei con il più alto tasso di disoccupazione giovanile (Eurostat, 2017). L'instabilità, la precarietà e la discontinuità delle entrate rendono i giovani vulnerabili ai cicli economici, modificando natura e tempi della transizione al mondo del lavoro e riducendo le opportunità di sviluppare soddisfacenti piani di vita (Leccardi, 2006). La sfiducia incide sui propulsori della transizione, cioè sul mantenimento di aspirazioni elevate, sulla cristallizzazione degli obiettivi di carriera e sul comportamento intensivo della ricerca di un lavoro (Vuolo et al., 2012). Per lo studio abbiamo utilizzato una fonte di dati testuali provenienti dal blog di Repubblica "Microfono Aperto" in cui studenti delle scuole superiori, nel periodo dal 2013 al 2016, hanno risposto al promt "Quattro giovani su dieci senza lavoro. E tu che pensi? Di chi sono le colpe? Cosa vorresti che venisse fatto al più presto per garantirti un dignitoso futuro?". Raccontarsi attraverso la Rete agevola il processo di riflessione su di sé, sul proprio ruolo e sul rapporto con ciò che accade nel contesto in cui il giovane è inscritto. In una situazione di malessere per la precarietà lavorativa, il web può essere un utile contenitore per la condivisione dell'esperienza di precarietà, costituendo un ambiente di condivisione e socializzazione delle proprie esperienze (Di Fraia, 2007).

3.1 Risultati

Il corpus conta 130 narrazioni (10110 occorrenze, 2484 forme grafiche, 1590 hapax), utilizzando come variabili descrittive la provenienza territoriale (nord, centro, sud) e il tipo di istituto frequentato (istituto tecnico-professionale e liceo) e soddisfa i criteri statistici di trattabilità. L'analisi tematica dei contesti elementari ha prodotto quattro cluster (Fig. 1; Tab. 1), rinominati CL1 "Guardare le opportunità" (14,6%); CL2 "E il governo?" (19,8%); CL3 "Dai sogni alla crisi" (38,5%); CL4 "La ricerca del lavoro, dove?" (27,1%). Le narrazioni del cluster "Guardare le opportunità" rimandano all'analisi di sacrifici e opportunità; emerge in modo marcato la necessità di una "attività", di una messa in pratica di azioni nel presente in vista di un futuro migliore. Per questo motivo, la crisi è al tempo stesso un'opportunità che i giovani devono cogliere per dimostrare le proprie capacità: *Ormai, per ciò che si sente, chiunque si chiede del proprio futuro. Per garantire che un giorno ci sia più lavoro, si deve agire ORA. […]. Anche chi cerca lavoro, però, deve volare basso e accontentarsi, per il momento, di poco, invece di restare a casa arreso.*

Secondo me i giovani devono avere l'opportunità di dimostrare ciò che valgono, dimostrare al mondo ciò che sanno essere e far capire a tutti che sono capaci "se si

impegnano" di fare qualsiasi lavoro, dal più semplice al più complesso. I testi del secondo cluster sono maggiormente orientati alla ricerca della "colpa" e ad una richiesta di soluzioni principalmente dallo Stato: *Penso che lo Stato dovrebbe dare più spazio ai giovani assicurando loro protezione e tutela. I parlamentari devono conservare i diritti e le possibilità di ogni giovane, siamo noi il futuro di questo stato, e come tali abbiamo bisogno di opportunità.*

Il cluster "Dai sogni alla crisi" rimanda alla dimensione più interna dell'essere immersi in una società che sta attraversando un momento di crisi economica. Gli studenti rimarcano che la mancanza di lavoro annulla i sogni: *Sono davvero preoccupata, tutti noi sogniamo cosa fare da grandi e sapere che il 38,7% dei giovani non riesce a trovare lavoro mi rende indignata. I giovani sono il futuro, il progresso, si impegnano […] Sappiamo tutti cosa dice il primo articolo della nostra splendida costituzione, eppure sembra sia ignorato. Bisogna dare più occasioni ai giovani, tenere in considerazione la nostra costituzione, per aprire le porte al futuro e rendere l'Italia migliore.* Le narrazioni dell'ultimo cluster riguardano trasversalmente tutte le difficoltà del cercare lavoro (la ricerca affannata, le aziende che non assumono a causa delle troppe tasse) e della necessità di andare all'estero: *L'Italia si ritrova in un periodo di profonda crisi e se non si riprende economicamente ridando la possibilità a noi giovani di far capire a chi di dovere che abbiamo le capacità e volontà di lavorare, l'Italia perderà tutti quei giovani ma soprattutto tutte quelle menti che andranno all'estero in cerca di condizioni di vita più favorevoli ma soprattutto di maggiori possibilità di lavoro.*

La posizione delle variabili descrittive mostra una differenza per la variabile provenienza territoriale e nessuna differenza per istituto frequentato. Se infatti il frequentare una scuola piuttosto che un'altra sembra non incidere sulla percezione del mondo del lavoro e sui vissuti di sfiducia, che sono invece comuni, l'appartenenza territoriale ha un suo peso. La modalità nord è, in termini di vicinanza, posta in prossimità dei cluster 1 e 4, il centro del 3 e il sud del cluster 2. Ciò indica come gli studenti del nord tendano maggiormente a problematizzare il fenomeno del precariato e la difficile ricerca del lavoro, mettendo anche l'accento sulle opportunità che i giovani hanno di dimostrare il proprio valore; le tematiche di quelli del sud vanno maggiormente nella colpevolizzazione del contesto, in linea con una maggiore risonanza del tema di discussione a causa di un'elevata incidenza della disoccupazione giovanile; le narrazioni degli studenti del centro, invece, maggiormente richiamano i propri vissuti interni.

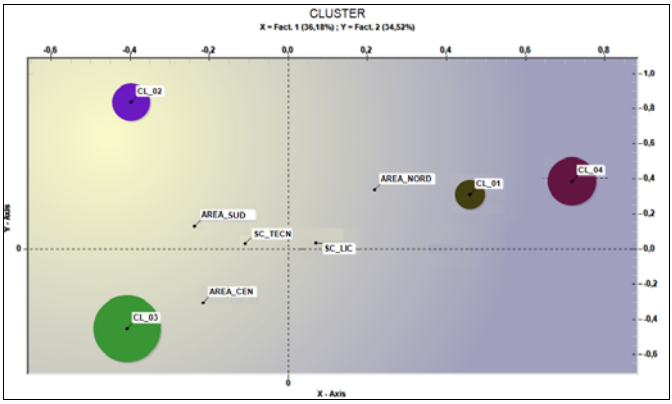
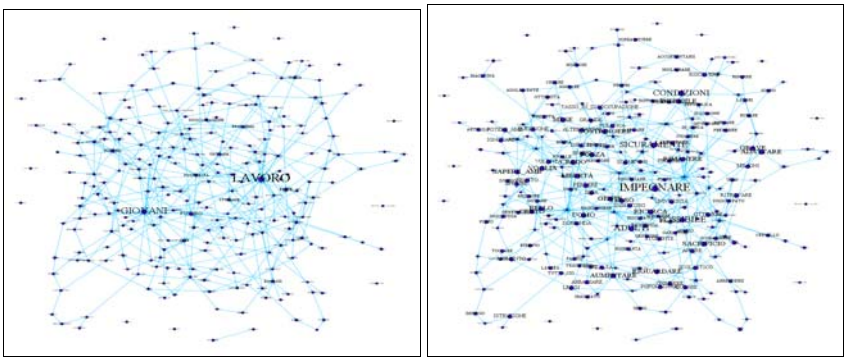


Figura 1: Cluster Analysis

La rete prodotta è composta da 259 nodi e 414 legami. Una prima approfondita forma di visualizzazione della struttura di relazioni tra i vari lemmi mostra i livelli più alti di degree centrality, in cui "lavoro", "giovani", "futuro", "problema" e "possibilità" rappresentano i nodi con maggiori connessioni. Inoltre, questi stessi nodi riportano anche i valori più alti di in-degree centrality, nodi "assorbenti" che presentano più legami in entrata che in uscita rispetto a tutti gli altri punti; gli studenti tendono a indirizzare i propri discorsi e, più in generale, il flusso di pensiero verso le tematiche relative al lavoro in termini sia di possibilità future sia analizzandone le problematiche ad esso legate. Dall'altro canto, "impegnare" (inteso come impegno messo in atto) e "condizioni" rappresentano il fulcro da cui muove la narrazione verso altre parole, nodi "sorgente" che hanno più legami in uscita che in entrata rispetto ai restanti nodi della rete. I lemmi che rimandano ai vissuti degli studenti, ai propri stati d'animo rispetto all'attuale condizione e ad una prospettiva lavorativa futura incerta sono quelli che giocano un ruolo centrale nella circolazione dei significati all'interno della rete, presentando difatti i valori più elevati di betweenness centrality. In particolare, "disoccupato", "costringere", "rimanere" e "scoraggiare" sono i nodi che fungono da principale punto di giunzione da cui si snodano specifici percorsi di significato: le diverse zone del network, e quindi diverse parti della narrazioni sono collegate tra loro da quei lemmi che ruotano intorno al tema della precarietà del presente, una situazione di costrizione e di forte scoraggiamento.



In-degree Centrality *Out-degree Centrality*

Betweenness Centrality

Figura 2

4. Conclusioni

L'uso misto della TA e NTA permette di rappresentare un quadro sintetico della struttura semantica, comprendere di cosa si parla, ma anche in che modo lo si fa: la scelta delle parole e l'ordine stesso di presentazione di un'idea o opinione rispetto al tema in oggetto. L'uso congiunto delle due tecniche fornisce: a) una sintesi delle informazioni contenute nelle narrazioni; b) l'analisi dei temi affrontati; c) un focus sulla strutturazione delle frasi in termini di relazioni tra lemmi. Permette così di mettere in relazione categorie tematiche e di contenuto in quanto struttura latente, ricostruendo a ritroso il processo discorsivo.

Bibliografia

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica*, vol. 7: 1-37.

Carley K.M. (1997). Extracting team mental models through textual analysis. *Journal of organizational behavior*, 18(1): 533-558.

Di Fraia G., a cura di, (2007). *Il fenomeno blog. Blog-grafie: identità narrative in rete*. Milano: Guerini e Associati.

Eurostat (2017). Statistics on young people neither in employment nor in education or training. Report.

Freeman L.C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*, vol. 1: 215-239.

Hunter S. (2014). A novel method of network text analysis. *Open Journal of Modern Linguistics*, vol. 4(2): 350–366.

Lancia, F. (2004). *Strumenti per l'analisi dei testi*. Milano: Franco Angeli.

Lebart L., Salem A. and Berry, L. (1998). *Exploring textual Data*. Dordrecht: Kluwer Academic Publishers.

Leccardi C. (2006). Redefining the future: Youthful biographical constructions in the 21st century. *New directions for child and adolescent development*, vol. 113: 37-48.

Popping R. (2000). *Computer-assisted Text Analysis*. London: Sage.

Popping R. and Roberts C.W. (1997). Network approaches in text analyisis. In Klar R. and Opitz O., editors, *Classification and knowledge organization*. Berlin, New York: Springer.

Vuolo M., Staff J. and Mortimer, J. T. (2012). Weathering the great recession: Psychological and behavioral trajectories in the transition from school to work. *Developmental psychology*, vol. 48(6): 1759.

Hablando de crisis: las comunicaciones del Fondo Monetario Internacional

Ana Nora Feldman
Universidad Nacional de Luján – anafeldman@gmail.com

Abstract

The annual reports of the International Monetary Fund issued annually under the name of "World Economic Outlook" from the years 2005 to 2012, are analyzed in this Paper by using the techniques of Statistical Analysis of Textual Data. The scan tool text, allows us to see the way the IMF describes in their reports the world crisis, highlighting their strengths and weaknesses in their role of the ultimate guarantor of global economic balance. Much has been discussed about the foresight of the crisis and what was the position of the IMF regarding its consequences. The denial of the crisis, only recognized in 2010, is consistent with the mission that the International Monetary Fund considers to carry out, lecturing on how governments should correct their economies (Weisbrot et al., 2009). All this ignoring that "their prescriptions failed" (Stiglitz, 2002) as their "structural adjustment policies" … "produced hunger and unrest" benefiting those who had more resources while "the poor sometimes sank more and more in misery. " In particular what is analyzed from the processing of textual corpus with Taltac2 software, developed by Prof. Sergio Bolasco from the Università di Roma "La Sapienza", are the concepts and language associated as a contribution to "a significant debate on a variety of exclusions "..." that encompass the political, economic and social fields"(Sen et Kliksberg, 2007) and considering that the World Economic Outlook reports may be useful for understanding the behavior of the IMF in the context of the financial crisis. The texts analyzed are written by technicians and bureaucrats, who possess a high level of expertise and skillful management of common codes, and are the product of a clear intention on how the global economic situation and the role of the Monetary Fund (and technicians), within this context, must be read. These reports, as will be demonstrated meet the goal of preaching the hegemonic conception on markets and policies, seeking to satisfy goals related to communication and marketing strategies in order to align public opinion, government officials and government objectives behind this concept. It is along this line that the contradictions between the more political text (the introduction and the summary) and the technical text (the body of the publication) are also shown.

Resumen
Con la ayuda de técnicas de Análisis Estadístico de Datos Textuales, se analizan los informes anuales del Fondo Monetario Internacional que se publican anualmente con el nombre de "Perspectivas de la Economía Mundial" entre los años 2005 y 2012. Se trata de evidenciar en los textos la forma en la que describe el FMI a la crisis, poniendo en evidencia sus fortalezas y debilidades en su rol de último garante del equilibrio económico mundial. Mucho se ha discutido acerca de la capacidad de previsión de la crisis y cuál fue la posición del Fondo Monetario respecto de sus consecuencias. La negación de la crisis, sólo reconocida en el año 2010, es coherente con la misión que el FMI considera que debe cumplir, aleccionando sobre la forma en que los gobiernos deben corregir sus economías (Weisbrot et al., 2009). Todo esto ignorando que "sus recetas fallaron" (Stiglitz, 2002) pues "las políticas de ajuste estructural"… "produjeron hambre y disturbios" beneficiando a quienes poseían más recursos mientras que "los pobres en ocasiones se hundían aún más en la miseria". En particular se analizan con la ayuda de Taltac2, desarrollado por el Prof. Sergio Bolasco de la Università di Roma "La Sapienza", los conceptos y el lenguaje asociado como aporte a "un debate significativo acerca de una variedad de exclusiones" … "que abarcan el campo político, económico y social" (Sen et Kliksberg, 2007) para comprender el comportamiento del FMI en el contexto de la crisis financiera.
Los textos analizados son escritos por técnicos y burócratas, que poseen un alto nivel de especialización y un manejo hábil de códigos comunes, y son producto de una clara intencionalidad acerca de cómo debe leerse la situación económica mundial y el rol del Fondo Monetario (y sus técnicos) en dicho contexto. Estos informes, como se demostrará, cumplen con el objetivo de predicación de la concepción hegemónica, sobre mercados y políticas, buscando satisfacer objetivos relacionados con estrategias comunicacionales y de marketing con el objetivo de alinear a la opinión pública, funcionarios y gobiernos detrás de esa concepción. En esa óptica es que se muestran también las contradicciones entre el texto más político (la introducción y el resumen) y el texto técnico (el cuerpo de la publicación).
Keywords: textual data analysis, content analysis, political language, economic and financial crisis.

1. Introducción

La crisis económico – financiera que comenzó en Estados Unidos en el año 2007, y que luego se extendió a Europa y otros continentes, fue reconocida de manera tardía por parte del Fondo Monetario Internacional (FMI). Considerando que la misión del Fondo es la de prever los riesgos originados en crisis económicas y brindar recomendaciones acerca de los mecanismos de

mitigación, la pregunta que se impone es ¿por qué, ante la crisis financiera de mayor envergadura después de la Gran Recesión de 1930, el Fondo ignoró la crisis, evitando declarar la emergencia de envergadura mundial? Desde el punto de vista político (y discursivo), al negar la crisis, el FMI impidió la puesta en marcha los mecanismos previstos para afrontar problemáticas de semejante envergadura. En este trabajo se analizan, con técnicas de Análisis de Datos Textuales, los informes anuales (Perspectivas de la Economía Mundial) publicados durante 8 años (2005-2012). Congruencias y contradicciones nos permiten analizar, desde un punto de vista diferente, las estrategias políticas del Fondo Monetario que ha visto muy desgastada su imagen como recurso válido e idóneo para el salvataje de economías en peligro.

2. Corpus

El criterio para la elección del período en análisis es el de relevar información en momentos diferentes de la crisis. Partiendo desde un "momento 0" (previo a su aparición), pasando por la instancia de reconocimiento del estado de situación, para finalmente considerar el cambio más importante en la política llevada adelante hasta ese momento por parte del FMI, es decir el paso del paradigma neoliberal "no intervencionista" (ninguna acción por parte del Estado para que el mercado se regule solo) a una política activa de ayuda por parte de los gobiernos (de Estados Unidos y de la Unión Europea), para "salvar las principales empresas, compañías y bancos en quiebra" (Rapoport et Brenta, 2010). Desde una óptica de análisis del contenido (Krippendorf, 1969), se realiza un análisis comparativo de dichos informes, buscando conocer cuál ha sido la forma en la que el FMI ha descrito la crisis y cuáles son las temáticas asociadas a la misma. La hipótesis, es que este lenguaje y contenido no neutral de criterios técnicos y políticos, responden al acuerdo de la que hemos llamado comunidad internacional "de peso real" (Feldman, 1995).

3. Ocho años de discursos del Fondo Monetario Internacional

Ya hemos trabajado y presentado diferentes aspectos relacionados con las comunicaciones del Fondo Monetario ante la crisis más importante tanto económica como financiera. Discursos que dependen del Director General de turno y el uso de la lexicometría como herramienta para la interpretación de los informes (Feldman, 2015 a y b).

En este trabajo analizaremos las cuestiones relacionadas con la congruencia y el uso político que se da en estas publicaciones anuales. La ambigüedad del discurso, la dificultad de previsión y reconocimiento (o negación) de la misma, sus causas y consecuencias y los reiterados anuncios del fin de la

crisis (en los años 2012, 2013 y 2014) que han sido objeto de crítica por todos los bloques de países más o menos cercanos al FMI.
El objetivo entonces es identificar las posiciones del Fondo Monetario Internacional en el tiempo. Se trata de comprender cómo habla y cómo calla el FMI sobre este crucial tema, como aporte a "un debate significativo" sobre exclusiones que "abarcan el campo político, económico y social" (Sen et Kliksberg, 2007). Subyace a esta propuesta la idea que la exploración y el análisis de textos, mediante recursos de estadística exploratoria multidimensional, permite "una concepción ecológica para el tratamiento de datos cualitativos" (Bolasco, 2007). El software utilizado es TALTAC.

3.1. El Discurso del FMI

El corpus está constituido por un total de 1.056.336 palabras (u ocurrencias).
Se trata de textos largos (más de 300 páginas incluyendo gráficos y tablas) con un promedio de 132.042 ocurrencias. Si bien la distribución entre años es aproximadamente similar, el informe del 2008 se distingue pues concentra el 16% del total de ocurrencias.

Tabla 1 – Análisis Lexicométrico

Año	Ocurrencias	%	Diferencia con $\overline{X}$
2005	133551	13	+
2006	121995	12	-
2007	147937	14	+
2008	168906	16	+
2009	114498	11	-
2010	124365	12	-
2011	120454	11	-
2012	124630	12	-
Total	**1056336**	**100**	

Así como el año 2008 se destaca por su extensión el del 2009 es el que utiliza una mayor riqueza de vocabulario. Según nuestra experiencia (Feldman, 1995), la utilización de una cantidad elevada de palabras en un informe podría estar indicando una situación de "malestar" o bien del uso de lenguaje "desvirtuado". Es decir, se deben utilizar más palabras para describir algo que aún no ha sido consensuado entre los técnicos y, por consiguiente, no ha sido conceptualizado adecuadamente.

Tabla 2 – Riqueza de Vocabulario

AÑO	Número de formas (V)	Ocurrencias (N)	Riqueza de Vocabulario (V/N)
2005	12301	133551	9,21%
2006	11203	121995	9,18%
2007	12454	147937	8,42%
2008	15283	168906	9,05%
2009	10951	114498	9,56%
2010	11332	124365	9,11%
2011	10656	120454	8,85%
2012	11112	124630	8,92%

La distribución en los años de la forma "crisis" es lo suficientemente ilustrativa acerca del uso dado, por parte del FMI, al correr de los años.

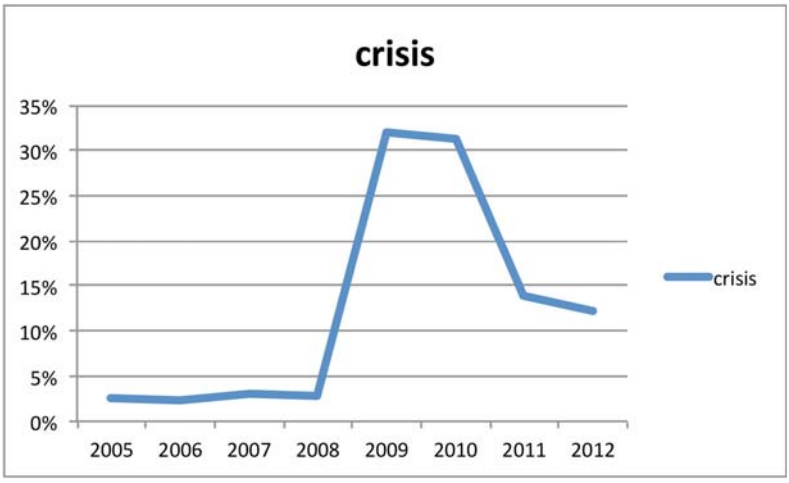


Gráfico 1 – Distribución de la forma "crisis" en el tiempo

3.2. Dos niveles de análisis: año por año los informes del Fondo

Si tomamos en cuenta sólo la Introducción y el Resumen Ejecutivo (a los que llamaremos "textos políticos"), que preceden al cuerpo del informe técnico (más de 300 páginas de textos y números) de cada Informe (a los que llamaremos "textos técnico-económicos"), éstos pueden ser considerados piezas comunicacionales que tienen un alcance público mayor, pues existe

una amplia gama de públicos que "consumen" los documentos técnicos del FMI (periodistas económicos, economistas, público en general) pero que normalmente no leen los informes completos. Muchas veces son justamente estos escritos sintéticos, los que tienen un efecto mayor en la modelación de la opinión pública internacional. ¿Existen entonces diferencias y/o inconsistencias entre los informes considerados integralmente y los resúmenes ejecutivos e introducciones? A través de la lectura de los mismos y el análisis de las principales formas estadísticamente significativas comentaremos diferencias y similitudes entre estos. Sin presagiar ninguna crisis, tanto en el año **2005** como en el **2006**, en sus textos se registra *coherencia económica* a partir de la sintonía entre los contenidos de la primera parte con aquellas formas estadísticamente significativas del documento técnico: INFLACIÓN, INVERSIÓN, AHORRO (2005), PRODUCTIVIDAD y SECTORES PRODUCTIVOS (2006). En el año **2007**, el del comienzo de la crisis el FMI comienza a hablar de un "período incierto y difícil" y las palabras estadísticamente significativas hacen referencia sobre todo a la VOLATILIDAD, contemporáneamente habla de crecimiento, registrándose *disonancia económica entre ambas partes*. El año **2008**, como ya señalado más arriba es el que concentra el 16% del total de ocurrencias del corpus. Nos encontramos aquí ante una disonancia discursiva / económica con el uso de muchos términos no habituales del FMI (VIVIENDA y CAMBIO CLIMÁTICO) para la descripción de la situación económica (disonancia y/o incongruencia en el uso de términos, cfr. Feldman, 1995). Ya estallada la crisis en el año **2009**, a partir de la presión internacional, el FMI debe comenzar a explicar aquello que no previó ni anunció (ver gráfico 1 y Tabla 2). Encontramos mayor disonancia entre texto y contexto y nuevas formas significativas (*DESPLOME*, *ALARMAS*). Intentando retomar el liderazgo político, luego de haber sufrido numerosas críticas por su falta de previsión de la crisis, el FMI, durante el año **2010**, donde – entre su parte sintética y el documento técnico – encontramos coherencia política y disonancia económica. Entre las formas significativas encontramos *CRISIS*.
A partir del año **2011** en el que encontramos más distancia entre lo que se lee en la Introducción y el Resumen Ejecutivo y el contenido del Informe completo, reaparece la política. Una vez recuperado su espacio institucional y su razón de ser, los textos del **2012** poseen coherencia tanto política como económica.

5. Conclusiones

El Fondo realiza una lectura de los indicadores económicos contradictoria, con una visión poco clara acerca de la gravedad y las consecuencias de esta crisis. El análisis del contenido de los textos (discursos e informes), con el uso

de herramientas de estadística textual, permite graficar de manera irrefutable las contradicciones y los silencios en los que incurre el FMI desde los primeros síntomas de la crisis en el año 2007. Los conceptos entonces vertidos en los Informes Perspectivas de la Economía Mundial son el producto "de una curiosa mezcla de ideología y mala economía, un dogma que en ocasiones parecía apenas velar intereses creados" recomendando "soluciones viejas, inadecuadas" con brutales efectos "sobre los pueblos de los países a los que se aconsejaba aplicarlas" (Stiglitz, 2002). Estas recetas fallaron en muchas oportunidades y produjeron situaciones sumamente graves en varios países. Un mensaje, un emisor, un objeto y una misión que falló, pues el FMI no cumplió con su rol de evitar que el mundo caiga nuevamente en una nueva Gran Depresión. Los textos analizados permiten establecer algunas pistas acerca de las motivaciones de este fracaso. En las contradicciones evidenciadas y en los intentos de negación de una realidad que no dejaba dudas acerca de la magnitud de esta crisis se afianza la idea de que existe en el Fondo Monetario y otros organismos internacionales un problema de gobernanza

Tabla 3 – Análisis de coherencia y disonancia de los Informes año por año

AÑO	COMENTARIOS
2005	**Coherencia económica**
2006	**Coherencia económica**
2007	**Disonancia económica**
2008	**Disonancia discursiva/económica** Uso de palabras poco frecuentes indica situación de entropía (Feldman, 1995). Se usan muchos términos técnicos no habituales para la descripción de la situación.
2009	**Disonancia discursiva** Mayor riqueza de vocabulario y mayor disonancia entre texto y contexto
2010	**Coherencia política** **Disonancia económica**
2011	**Coherencia económica y política**
2012	**Coherencia económica y política** Como si nada hubiera sucedido en estos años Recuperada la función política se vuelve a la coherencia económica.

Bibliografía

Bolasco S., D'Avino E. y Pavone P. (2007) Analisi dei diari giornalieri con strumenti di statistica testuale e text mining, Publicado en *I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo*. ISTAT, Roma

Feldman, A. (1995), Il concetto di sviluppo umano secondo le Nazioni Unite: analisi del contenuto in Bolasco, S., Lebart, L. e Salem, A. (eds.). JADT 1995 - *Analisi statistica dei dati testuali*, Roma, CISU, 2 voll.

Feldman, A. (2015a) Análisis del Posicionamiento del Fondo Monetario Internacional frente a la crisis del año 2007 en *Revista Latinoamericana de Opinión Pública*. Año 2016, número 6, EDUNTREF. Buenos Aires

Feldman, A. (2015b) Text Mining Strategies applied on the annual reports of the International Monetary Fund. A look at the crisis en *ISI 2015 World Statistics Congress*, Rio de Janeiro

Krippendorff, K. (1969). Theories and Anlytical Constructs en: G. Gerbner, O.R. Holsti, K. Krippendorff, W.J. Paisely y P.J. Stone (eds.) *The Analysis on Communication Content*, New York, John Wiley & Sons, p. 6 e ss.

Lebart, L y Salem, A. (2008). *Statistique Textuelle*, Dunod, Paris.

Nemiña, Pablo. (2009) Aportes para un esquema de análisis del comportamiento del FMI en crisis financieras a partir de su actuación durante la crisis argentina (2001-2002). *Documentos De Investigación Social Número 8*. ISSN 1851-8788. IDAES, UNSAM, Buenos Aires

Rapoport, M. y Brenta, N. (2010). *Las grandes crisis del capitalismo contemporáneo*. Capital Intelectual. Buenos Aires.

Sen, A. y Kliksberg, B. (2007). *Primero la Gente*. Ediciones Deusto. 9na edición Editorial Temas, Buenos Aires, Argentina.

Weisbrot, M., Cordero, J. y Sandoval, L. (2009). *Empowering the IMF: Should Reform be a Requirement for Increasing the Fund's Resources?* Center for Economic and Policy Research. Washington, D.C., Estados Unidos www.cepr.net

Brexit in the Italian and the British press: a bilingual corpus-driven analysis

Valeria Fiasco
Università Roma Tre – valeria.fiasco@gmail.com

Abstract 1 (English)
The spread of English as the Lingua Franca of international communication has given rise to meaningful language contact phenomena in the world's languages like loanwords and pseudo-loanwords, namely, words from one language (the donor language) are adopted by another language (the recipient language) sometimes becoming naturalized (Gusmani 1973). From this perspective, it is thus interesting to observe their behaviour in real language use. In particular, this study investigates Anglicisms and pseudo-Anglicisms found in the newspaper discourse of Brexit by way of a bilingual corpus collected from two Italian newspapers, i.e. *La Repubblica* and *Il Corriere della Sera* and two British newspapers, i.e. *The Independent* and *The Guardian* selected for both their authoritativeness and their extensive readership. The exit of the United Kingdom from the European Union was chosen because it is a widely covered topic both in the Italian and in the British press, thus providing abundant material for comparative analysis, as well as offering useful data in order to explore linguistic variation**.** It was useful for building an electronic corpus which was retrieved from the digital archive**s** of the newspapers' websites in order to carry out an automated text analysis.
The corpus includes articles collected during the periods that both preceded and followed the Brexit referendum. In order to carry out the analysis, corpus-driven methodology was used, namely an approach that lets hypotheses emerge from corpus observation (Tognini-Bonelli 2001). The investigation was carried out by way of the software TalTac2**,** and the automated text analysis, as a result, turned out to be invaluable in order to investigate and monitor the newspapers' vocabulary which included technical terms from the fields of politics, economics and finance as well as general language words. In order to design and sample a representative corpus, the parameters proposed by Biber (1993) were used to identify descriptive criteria so as to select and balance the population.
The aim of this study is to get an overview of the Brexit discourse as used in the two countries' newspapers' vocabulary and terminology (of the two countries) by using text mining to compare and categorize the whole corpus as a collection of texts and, then, to cluster documents on the basis of the

lexical similarity of the vocabulary to establish semantic fields or conceptual areas. Furthermore, by way of the lexical and textual analysis, this study also investigates Anglicisms and pseudo-Anglicisms in the Italian newspapers, identifying and analyzing a list of English words used in Italian. The two British newspapers serve as a reference corpus to compare to the list of Anglicisms extracted from the Italian corpus. The articles retrieved from the British newspapers serve to find out which words are typical of each corpus and to identify pseudo-anglicisms, namely new words that seem to be English form**s**, even though they do not exist in English**,** or if they do exist, they have a clearly different meaning. Lastly, the data gathered from the bilingual corpus analysis were later compared with other wider corpora included in SketchEngine and on the Brigham Young University platform in order to make generalizations about the distribution of Anglicisms and pseudo-Anglicisms in general language corpora.

Keywords: Bilingual Corpus, Textual Analysis, Anglicism, Linguistic Interference

Abstract 2 (Italian)
La diffusione e l'affermazione dell'inglese come lingua franca della comunicazione internazionale ha generato fenomeni significativi di contatto linguistico come i prestiti e i falsi prestiti, ossia parole originariamente nate in una lingua modello che entrano a far parte di un'altra lingua (lingua replica) alla quale vengono talvolta assimilate e adattate (Gusmani 1973). È quindi interessante osservarne l'uso e l'andamento in testi autentici che presentano la lingua nel suo uso corrente. Questo studio analizza gli anglicismi e i falsi anglicismi nel discorso giornalistico della Brexit, attraverso un corpus tratto dai quotidiani italiani *La Repubblica* e *Il Corriere della Sera* e dai quotidiani britannici *The Guardian* e *The Independent*, che sono stati selezionati per la loro diffusione e la loro autorevolezza. La scelta della tematica dell'uscita del Regno Unito dall'Unione Europea è stata dettata da diversi fattori, tra i quali l'ampia diffusione dell'argomento nella stampa italiana e in quella britannica, dando la possibilità di creare un corpus per realizzare un'analisi comparativa attraverso l'esplorazione della variazione linguistica. Dal momento che queste riviste offrono una versione online che mette a disposizione un archivio digitale consultabile, sono particolarmente adatte per creare un corpus che può essere esaminato attraverso l'analisi automatica del testo. Il corpus è composto da articoli raccolti durante il periodo che precede e segue il referendum della Brexit e la metodologia utilizzata per condurre l'analisi è di tipo corpus-driven, ossia un approccio esplorativo in cui, partendo dall'osservazione del corpus, si arriva alla formulazione delle ipotesi

(Tognini-Bonelli 2001). Il software TalTac2 e l'analisi automatica dei testi sono stati estremamente preziosi per esaminare e monitorare il lessico della stampa che include termini tecnici della politica, dell'economia e della finanza, insieme a parole che fanno parte del lessico comune. Per progettare il corpus, sono stati utilizzati i parametri proposti da Biber (1993) con lo scopo di identificare i criteri descrittivi per selezionare e bilanciare la popolazione all'interno del corpus. L'obiettivo di questa ricerca è offrire un'analisi del lessico e della terminologia utilizzata nel discorso sulla Brexit nei quotidiani italiani e inglesi attraverso il text mining per raffrontare i testi che compongono il corpus, categorizzarli e raggrupparli sulla base di somiglianze lessicali per individuare i campi semantici e le aree concettuali. Inoltre, l'analisi lessicale e testuale ha consentito l'identificazione degli anglicismi e dei falsi anglicismi nei quotidiani italiani, mentre il corpus dei quotidiani britannici ha svolto la funzione di corpus di riferimento per paragonare la lista degli anglicismi estratta dal corpus italiano con i dati raccolti nel corpus britannico, capire quali parole sono tipiche di ogni lingua e identificare i falsi anglicismi, vale a dire parole che presentano una forma inglese, che però non esistono nel vocabolario originario o nel caso in cui esistano, il loro significato è completamente differente. Infine, i dati raccolti dall'analisi del corpus bilingue sono stati successivamente confrontati con altri corpora più ampi, consultabili su SketchEngine e sulla piattaforma della Brigham Young University con lo scopo di fare delle generalizzazioni sulla distribuzione degli anglicismi e dei falsi anglicismi in corpora non specialistici.

Parole chiave: Corpus bilingue, analisi testuale, anglicismo, interferenza linguistica

1. Introduction

The growing influence of English on many languages in the world represents the linguistic change produced by language contact. English is used in both academic and professional settings revealing a pervasive presence of Anglicisms in European languages (Marazzini & Petralli 2015). This situation can be traced back to economic and trade developments, as well as political and social circumstances in the past decades. The Anglo-American globalization also exerts an influence on language with an increasing number of EFL (English as a Foreign Language) and ESL (English as a Second Language) learners and the English use as a Lingua Franca (ELF) for international communication giving rise to the borrowing of an increasing number of Anglicisms which have thus become the symbol of the American lifestyle, an expression of symbols, dynamism and progress. Pulcini, Furiassi,

Rodríguez Gonzàlez (2012:1) use the term *Anglicization* to stress the growing extensive research on lexical borrowing which has had a major impact on vocabulary and phraseology of English origin. Lexical borrowings adapt to their receiving language in various ways, from occasional coinages to integrated words, from more restricted circles to broad groups until reaching the totality of the speakers of the recipient language. Gusmani (1993:28) states that there are cases of complete acclimatization in which the speakers of the recipient language become so used to the foreign word that it is perceived to be part of the recipient language, i.e. *film*. One of the main sources of neologisms and borrowings is from newspapers and magazines which detect the emerging trends in contemporary language and coin new words in a creative fashion. According to Beccaria (1983:65), newspapers are one of the main forums of exchange between written and spoken language, where different varieties coexist, for example, bureaucratic, technical and literary language. Moreover, in newspapers, the interaction between the general and specialized language takes place allowing specific terms to penetrate the popular culture (Cabré 1999:17).

2. Research design

This paper stems from the assumption that the linguistic interference of English on Italian brings about significant effects giving rise to lexical borrowing phenomena like Anglicisms and false Anglicisms, especially in newspaper language. This bilingual corpus-driven analysis describes both the Italian and the British discourse of Brexit with the aim of analyzing its vocabulary and terminology as used in both the Italian and the British press. By way of text mining, patterns and trends that allow us to make connections between the two languages under investigation can be discovered. We can identify Brexit's main themes, get a picture of how corpus data are shaped and subdivided into text fragments that correspond to the newspaper article's sections (title, subtitle, summary, text). We can investigate the linguistic interference of English on Italian and the markedness between the Anglicisms/pseudo-Anglicisms retrieved in the Italian newspapers and their Italian equivalent words.

The exit of the United Kingdom from the European Union was chosen because it is a historic and momentous event which has been the focus of attention of numerous newspapers, thus, providing abundant material to collect in the corpus. The reason behind the choice of the two languages lies in the linguistic interference phenomena they are closely involved in: English performs the role of a highly productive donor language, while Italian is a recipient language which is under the influence of English.

The bilingual corpus is made up of articles retrieved from two Italian

newspapers, i.e. *La Repubblica* and *Il Corriere della Sera* and two British newspapers, i.e. *The Independent* and *The Guardian*. They were selected for their authoritativeness, their extensive readership and the possibility to access their on-line archives with a free subscription. Moreover, they all dealt with the Brexit issue thoroughly. The corpus was compiled by downloading and storing all the articles about Brexit published in the on-line versions of these newspapers from June to October 2016, that is, the period that preceded and followed the Brexit referendum. The selected articles provide a brief, but detailed overview of the Brexit, even though they are not representative of all of the Italian and the British press. The corpus is composed of two corpora, the Italian and the British one. The Italian corpus includes 42 articles from *La Repubblica* and 42 articles from *Il Corriere della Sera* for a total amount of 51,158 tokens, whereas the British corpus includes 31 articles from *The Guardian* and 31 articles from *The Independent* for a total amount of 49,995 tokens. However, a difference can be observed in the number of articles that make up the overall corpus, because the average length of the British articles was shorter than that of the Italian ones. On the whole, the corpus includes 146 articles and 101,153 tokens. The corpus was designed and sampled according to the parameters proposed by Biber (1993) in order to build up a representative corpus and to identify descriptive criteria so as to select and balance the population. The issue of whether a corpus is representative and reliable is essential, because the information included in the corpus and the way it is constructed is central in the corpus-driven approach, namely a method that lets hypotheses emerge from corpus observation (Tognini-Bonelli 2001). The automated text analysis on the corpus was carried out by way of the software TalTac2, to investigate the newspapers' vocabulary, to observe the behaviour of Anglicisms', as well as to make a detailed bilingual analysis. In order to make generalizations about the distribution of Anglicisms and pseudo-Anglicisms in general language and to retrace their routes from/into the donor and the recipient language, other general language corpora were consulted: Sketch Engine (*British National Corpus*, *itTenTen16* and *enTenTen13* corpus) and the online corpora available on the Brigham Young University website (*News on the Web – NOW*, *Global Web-Based English – GloWbE*, *TIME Magazine Corpus*). Furthermore, the software Iramuteq was used to carry out the cluster analysis of both corpora, to map them and extract the semantic associations of words according to their similarity.

3. Results

In order to identify the main themes and semantic fields of the corpus, the cluster analysis grouped its lexical content so as to maximize the similarity or

the dissimilarity of different groups of words. The analysis divided the Italian and the English corpus into 4 homogeneous clusters whose topics are economics and finance or European and British politics. The output graph was a dendrogram showing the association of all the words included in the two corpora according to their similarity. It grouped the words into two clusters: the first one concerns economics/finance and the second one is related to politics. The percentage of words included in the Italian economics cluster equals 31% compared to 23% in the English economics cluster. In both corpora, the words from the semantic field of economics are homogeneously distributed, i.e. *bank/banca, market/mercato, growth/crescita, fund/fondo, investor/investimento, rate/tasso*. As for the politics cluster, both corpora subdivide the lexical content into three clusters. In the Italian corpus, the cluster of politics generates cluster 4 (23%) grouping the words concerning the British politics and the sub-clusters 1 and 3. Sub-cluster 1 (22%) regards the European politics and the Brexit referendum, i.e. *Unione, europeo, UE, negoziati, uscire, trattativa*, while sub-cluster 3 (23%) is related to European policies linked to political integration and post-Brexit immigration policies, i.e. *difesa, migrare, integrazione, emergenza*. In the English corpus, the cluster of politics generates cluster 1 (26%) that corresponds to Italian cluster 3, i.e. *movement, immigration, person, European* and two sub-clusters (2 and 3) about the British politics. In particular, sub-cluster 3 is about the Leave campaign, i.e. *Ukip, independence, break, Farage* whereas sub-cluster 2 is about the Remain campaign of the United Kingdom in the European Union, i.e. *Cameron, conservative, labour, tory*. Moreover, the dendrogram also shows who the main actors of this event are: the European Union, David Cameron, Nigel Farage, Theresa May, Boris Johnson, and Jeremy Corbyn.

By way of its textual analysis, the software TalTac2 also identified the words occurring within specific text fragments in which the corpus has been subdivided and labelled, i.e. headline, sub-heading, lead, body. This analysis particularly focused on the headlines. On the whole, the most frequent lexical word in both corpora, *Brexit*, is mainly found in the headlines and in the body of Italian newspapers, while it can only be observed in the body of the British press. The concept of "exit, leaving the European Union" mainly appears in the body of the articles of the British press, while in Italian newspapers it is predominantly found in headlines. The brief exploration of the headlines starts with the key topics expressed by the nouns in both the Italian and the English corpus. The topics refer to the domain of politics, the governance of the UK, the debate and the negotiations between the two parties and the problems arising from the exit of the United Kingdom from the European Union (i.e. *referendum, European Union, leader, government, campaign, support/negoziato, collasso, rischio, leader, rischio, referendum*). In

particular, the most recurrent nouns in both the English and Italian headlines mirror the themes addressed in the two corpora, i.e. politics: *Brexit, EU referendum, Remain, vote/Brexit, premier, uscita;* economics and finance: *borsa, sterlina/pound*. As for verbs describing the actions, conditions or experiences linked to the Brexit, they outline a delicate and unstable situation in both corpora, i.e. *to vote, to fail, to resign, to face, to divide/uscire, crollare, affrontare, rischiare, intervenire*.

As far as the analysis of the linguistic interference is concerned, the Italian corpus includes 174 Anglicisms (types) for a total amount of 1.096 occurrences (tokens) whose percentage in the corpus is about 2.1%. As to types, their sum includes a lot of hapax legomena 91 out of 174 Anglicisms to be exact (approximately 52.3% of types). The 174 Anglicisms belong to the semantic fields of politics (22.5%), economics (27.5%), general language (45.5%), and newspaper language (4.5%). The list of Anglicisms extracted from the Italian corpus was later compared with the British one to check whether they were actually used in English and how: 81 Anglicisms out of 174 were found in the English corpus. The other 93 Anglicisms are real English words except for *neo-premier* (58.64 per million words) which can be defined as a pseudo-Anglicism. It is a loanblend or a hybrid compound (Furiassi 2010:40) formed by the English word *premier* and the Greek-derived suffix *neo-*. These two lexical elements are individually used in English, but they are not used together. The suffix *neo-* can be found in English compounds referring to political movements like *neo-socialist, neo-fascist* or regarding art and philosophy subjects, i.e. *neo-baroque, neo-Aristotelian*. The use and frequency of the compound *neo-premier* was compared with the Italian itTenTen16 corpus on SketchEngine. This online corpus displays two variants of the compound: the hyphenated word *neo-premier* (0.02 per million words) and *neopremier* (0.02 per million words). Conversely, the search of the same word in English corpora like BNC, enTenTen13, or Now corpus didn't produce any results.

The most frequent Anglicisms in the Italian corpus are *Brexit* (309 tokens, 0.6%), *referendum* (111 tokens, 0.22%), *premier* (89 tokens, 0.17%), *leader* (61 tokens, 0.12%). These four words are particularly frequent in the British corpus as well: *Brexit* (232 tokens, 0.46%), *referendum* (157 tokens, 0.31%), *leader* (71 tokens, 0.14%). In particular, the word *Brexit* is productive in both the English and the Italian corpus with numerous hyphenated compounds composed of Latin and Greek suffixs or English-derived morphemes. Some of them are common to both corpora, i.e. *post-Brexit* (English corpus 140 per million words, Italian corpus 58.6 per million words), *hard-Brexit* (English corpus 80 per million words, Italian corpus 58.6 per million words), *pro-Brexit* (English corpus 100 per million words, Italian corpus 39.1 per million words).

Other Brexit-compounds like *pre-Brexit* (39.1 per million words) and *dopo-Brexit* (19.5 per million words) are only found in the Italian corpus, while the compound *anti-Brexit* (40 per million words) is only included in the English corpus. As far as the word *premier* is concerned, in the English corpus, it only shows 1 token (20 per million words), while its synonym, *prime minister,* has a frequency of 119 tokens (2,380 per million words). The occurrence of this compound was then compared with larger English corpora like the BNC where *Prime Minister* is written both in capital letters (85.17 per million words) and in lowercase letters (8.33 per million words). On the contrary, the word *premier* is present in the BNC and occurs with a frequency of 0.23 per million words, but it mainly occurs in the semantic field of football, i.e. as a modifier of the noun *league* in the collocation *premier league*. However, it is also found in the domain of politics as a noun co-occurring with the modifiers *deputy*, *country*. Conversely, in the Italian corpus itTenTen16 in SketchEngine, *premier* always occurs in the semantic field of politics. Two different uses of the word *premier* and *Prime Minister* can thus be observed in the two languages.

4. Conclusion

The aim of this paper has been to provide an outline of the Brexit discourse as used in the vocabulary and terminology used by two Italian and two important British newspapers. By way of cluster analysis, the Brexit's main themes have been identified: economics, finance, European and British politics, and the Post-Brexit immigration policies. Another characteristic that has been explored in this paper is the distribution of the words in various newspaper article sections which was accomplished by focusing on the headlines. The analysis showed that the nouns included in newspapers' headlines refer, for the most part, to Brexit's main political issues, even though some words from the field of economics can be found as well. Whereas verbs aim at describing the difficult circumstances that both the European Union and the United Kingdom will face. As far as Anglicisms are concerned, the investigation highlighted that even though they are often used by newspapers, they represent only about 2% of the whole corpus. This percentage conforms to the most recent studies on Anglicisms in Italian by Serianni (2015), Cortellazzo (2015) and Scarpa (2015). They mirror the topic subdivision of the corpus, and in fact they mainly belong to the semantic fields of economics and politics, whereas almost half of them can be classified as general language words. In the Italian corpus, only one pseudo-Anglicism has been identified, i.e. *neo-premier*, and its status has been confirmed by numerous general English corpora. The analysis of Brexit-related Anglicisms provides a small but interesting contribution to the research on Anglicisms;

therefore, it would be interesting to keep collecting data about this historical fact so as to expand the two small corpora under investigation, to make them as comprehensible and comprehensive as possible, and to carry out an even more detailed contrastive analysis.

References

Biber D. (1993). *Representativeness in Corpus Design*. In Literary and Linguistic Computing, vol. 8 (4): 243-257.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S. (2013). *L'analisi automatica dei testi*. Carocci.

Cabré Castellví M. T. (1999). *Terminology: Theory, methods and applications*. John Benjamins

Publishing Company.

Cortellazzo M.A. (2015). Per un monitoraggio degli anglicismi incipienti. In Marazzini C., Petralli A. *La lingua italiana e le lingue romanze di fronte agli anglicismi*. Accademia della Crusca.

Furiassi C. (2010). *False Anglicisms in Italian*. Polimetrica.

Görlach M. (2001). *A dictionary of European Anglicisms*. Oxford University Press.

Gusmani R. (1973). *Analisi del prestito linguistico*. Libreria scientifica editrice.

Gusmani R. (1993). *Saggi sull'interferenza linguistica*. Le lettere.

Hunston S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Lenci A., Montemagni S. and Pirrelli V. (2007). *Testo e computer. Elementi di linguistica*

computazionale. Carocci.

Marazzini C., Petralli A. (2015). *La lingua italiana e le lingue romanze di fronte agli anglicismi*.

Accademia della Crusca.

Pulcini V., Furiassi C. and Rodríguez González F. (2012). *The Anglicization of European lexis*. John

Benjamins.

Scarpa F. (2015). *L'influsso dell'inglese sulle lingue speciali dell'italiano.* Edizioni Università Trieste.

Serianni L. (2015) Per una neologia consapevole. In Marazzini C., Petralli A. *La lingua italiana e le lingue romanze di fronte agli anglicismi*. Accademia della Crusca.

Sinclair J. (1991). *Corpus Concordance Collocation*. Oxford University Press.

Tognini-Bonelli E. (2001). *Corpus Linguistics at work.* John Benjamins Publishing Company.

Textual analysis to promote innovation within public policy evaluation

Viviana Fini[1], Giuseppe Lucio Gaeta[2], Sergio Salvatore[3]
[1] Ospedale Apuane, Massa – *vivianafini@gmail.com*
[2] Università di Napoli L'Orientale - *glgaeta@gmail.com*
[3] *Università del Salento - sergio.salvatore65@icloud.com*

Abstract
This paper illustrates the contribution by textual analysis in carrying out the research activities promoted by FORMEZ PA through the REVES (Reverse Evaluation to Enhance local Strategies) pilot project[1] that aims to innovate public policy evaluation. While evaluation usually embraces a policy/project viewpoint and adopts a sort of a top-down approach consistent with the flow of rules/resources from policy makers to citizens', REVES reverses this perspective. Indeed, it aims to assess public policies' performance in intercepting and supporting development strategies promoted by citizens/local actors. One of the three case studies carried out by the REVES project focuses on Melpignano, a small municipality in the Puglia Region of Southern Italy. Semi-structured interviews were carried out with a sample of twenty policy actors (national, regional and local policy designer and policy implementers as well as policy beneficiaries) linked with this municipality. By using the TLab software, textual analyses of responses were performed in order to identify their symbolic and latent components and to understand the actors' point of view about the world and specifically about local development. This allowed to assess how similar concepts - such as civic participation, innovation, community - are used with profoundly different cultural meanings by the actors. This contributes to understanding public policies' difficulties in enhancing local strategies.

Keywords: *Local cultures, textual analysis, innovation within evaluation.*

[1] The evaluative research was carried out within the framework of the NUVAL Project, "Actions to support the activities of the National Evaluation System and Evaluation Units" implemented by Formez PA. The case study was accomplished by Viviana Fini and Vito Belladonna, under the scientific coordination of Laura Tagle, Serafino Celano, Antonella Bonaduce, Giuseppe Lucio Gaeta. Viviana Fini carried out the cultural analysis under the supervision of Sergio Salvatore and thanks with the contribution of Giuseppe Lucio Gaeta.

Abstract
L'articolo descrive il contributo della ricerca culturale condotta attraverso lo strumento dell'analisi testuale nella realizzazione del progetto di ricerca pilota REVES (Reverse Evaluation to Enhance local Strategies) promosso da FORMEZ PA con l'intento di innovare la valutazione delle politiche pubbliche. Mentre il processo valutativo tradizionalmente segue il flusso delle risorse finanziarie e l'attuazione di norme/provvedimenti da parte dei soggetti locali, REVES propone un capovolgimento di prospettiva, intendendo valutare le performance delle politiche pubbliche nell'intercettare e valorizzare le strategie di sviluppo autonomamente elaborate dai territori. Uno dei casi studio del progetto si incentra sulla città pugliese di Melpignano. Sono state condotte interviste semi-strutturate con un campione di 20 attori di policy (policy maker e attuatori di politiche attivi sul piano nazionale, regionale e locale oltre a potenziali beneficiari delle politiche) a vario titolo connessi con la città. Con l'ausilio del software TLab sono state condotte analisi testuali aventi l'obiettivo di evidenziare le componenti latenti che orientano le visioni del mondo e dello sviluppo proprie degli attori intervistati. Ciò ha consentito di valutare come concetti simili, ad esempio partecipazione civica, innovazione, comunità – siano impiegati dagli attori con significati culturali diversi. Ciò contribuisce alla comprensione del motivo delle difficoltà delle politiche pubbliche nel valorizzare strategie localmente elaborate.

Keywords: *Culture locali, Analisi testuale, innovazione nella valutazione.*

1. Introduzione

L'articolo dà conto dell'indagine culturale - svolta attraverso analisi testuale - realizzata per supportare l'innovazione che il progetto REVES ha apportato al campo della valutazione delle politiche di sviluppo locale. Con un approccio *reverse accountability*, il progetto si è domandato se e come le politiche sovra-locali siano state in grado di cogliere e valorizzare le istanze di specifici contesti locali, indagando il caso studio "Melpignano", Comune in provincia di Lecce, noto in letteratura per aver elaborato, proposto e attuato, nel corso degli ultimi 30 anni, una visione e una strategia innovativa di intervento riguardante lo sviluppo locale (Attanasi *et al.*, 2011; Parmiggiani, 2013). Si discutono qui i risultati dell'indagine culturale e il vantaggio che l'analisi testuale ha permesso al progetto di realizzare, consentendo una lettura che è andata oltre il contenuto delle singole interviste, permettendo di cogliere come concetti simili fossero utilizzati talvolta – dagli intervistati – con significati culturalmente profondamente diversi.

2. L'indagine culturale come presupposto della ricerca valutativa

Il lavoro di ricerca realizzato mediante analisi testuale ha avuto quale fine la rilevazione delle dimensioni culturali che in modo latente hanno dato forma alle visioni e agli interventi sullo sviluppo locale. Questo tipo di indagine si inscrive in una cornice teorica psicologica ad orientamento psicodinamico e psico-culturale (Carli *et al.*, 2002; Salvatore *et al.*, 2011), che considera i comportamenti e i discorsi degli attori sociali come espressione di dinamiche culturali che solo in parte sono consce, in gran parte sono inconsce, latenti (Matte Blanco, 1975; Fornari, 1979; Carli *et al.*, 2002). Ciò che gli attori fanno, dicono, ritengono saliente - secondo tale approccio – è funzione di un campo di forze latenti, un sistema stabile di significati generalizzati, che chiamiamo cultura (Carli *et al.*, 2002; Salvatore *et al.*, 2011). L'idea di organizzare le azioni valutative sui risultati dell'indagine culturale ha risposto all'esigenza del progetto di "costruire" l'oggetto di indagine a partire da una comprensione profonda delle motivazioni alla base di certi esiti, in conseguenza della presenza/assenza di alcune iniziative. L'indagine culturale ha consentito di fare ipotesi su cosa ha avvicinato/distanziato modelli di azione appartenenti ad attori di policy diversi, consentendo di classificare i loro discorsi in relazione alla variabilità culturale che li caratterizza e che definisce lo scenario entro cui ciascuno di essi, senza la mediazione del pensiero razionale, si è mosso.

2.1 L'analisi testuale: modalità di analisi

Il metodo utilizzato per l'analisi testuale si fonda sul principio delle co-occorrenze lessicali come fonte di ricostruzione del contesto intratestuale. Tale principio è stato definito all'interno della linguistica (Reinert, 1986) e successivamente elaborato in chiave psicologica (Carli & Paniccia, 2002; Lancia, 2004). In termini generali il metodo, utilizzando il software TLab, trasforma il corpus lessicale in una matrice digitale di co-occorrenze, la quale viene a sua volta sottoposta ad una procedura di analisi multidimensionale che permette di estrapolare i cluster semantici attivi nel testo (cioè i cluster di parole co-occorrenti entro le stesse frasi, in quanto tali indicative di pattern di significato) che vengono successivamente sottoposti ad interpretazione. La procedura adottata segmenta il testo in Unità di Contesto Elementari (ECU), ossia parti di testo interrotte da punteggiatura, che possono contare da un minimo di 250 caratteri ad un massimo di 500. Attraverso una serie di operazioni il corpus testuale viene successivamente trasformato in una matrice digitale in grado di rappresentare il testo in termini di presenza/assenza dei lemmi nelle ECU che lo compongono. La matrice che si viene così a definire è sottoposta ad una procedura di analisi multidimensionale combinata, che unisce l'Analisi delle Corrispondenze

Multiple (ACM) e l'Analisi dei Cluster (AC). L'ACM permette di estrapolare le modalità nei termini delle quali i lemmi si associano all'interno delle ECU (vale a dire: le loro co-occorrenze intra - ECU). Ciascuna dimensione fattoriale individuata dalla ACM rappresenta un pattern di co-occorrenze che si ripropone attraverso il testo, o in una sua porzione sufficientemente ampia. Le dimensioni fattoriali estrapolate dalla ACM vengono quindi utilizzate come criteri classificatori dalla successiva CA. In questo modo la CA permette di raggruppare ECU (e lemmi) in base alla loro somiglianza – ossia in base alle combinazioni di parole per come si danno nelle frasi di testo. Il risultato finale della procedura è dunque l'identificazione di cluster di frasi tra loro simili in quanto caratterizzate dalla compresenza delle stesse parole; oppure, specularmente, dalla identificazione di cluster di parole simili in quanto tendenti ad essere utilizzate insieme nelle stesse frasi. Per questa loro caratteristica computazionale, i cluster individuati si prestano ad essere interpretati nei termini di nuclei tematici, tali in quanto caratterizzati dal riferimento ad un aggregato sufficientemente stabile di parole (Lancia, 2005).
L'output dell'analisi può essere considerato come una rappresentazione del campo culturale caratterizzante lo specifico contesto di policy (Carli *et al.*, 2002), dove sono visibili le dimensioni latenti che dinamizzano il campo (Fattori) e la variabilità relativa ai diversi modi di pensare dei soggetti intervistati (Cluster).

2.2 Popolazione di riferimento e campione

La popolazione di riferimento sono gli attori delle politiche. Il campione è costituito da 20 soggetti che a vario titolo hanno operato in relazione allo sviluppo locale, con i quali è stata condotta un'intervista in profondità, considerati figure chiave del contesto studiato per le seguenti variabili illustrative: *ruolo* (politici, cittadini, tecnici); *tipo di implicazione nella politica* (policy maker, policy designer, attuatori, destinatari); *livello di appartenenza* (locale, sovracomunale, regionale, nazionale). Trattandosi di uno studio pilota, al campione rappresentativo si è preferito un campione a grappolo per quote non proporzionali (Blalock jr, 1960), facendo riferimento agli attori presenti entro i contesti, distribuiti in modo tendenzialmente equivalente in relazione alle tre variabili. La scelta di un campione di questo tipo ha consentito di costruire ipotesi, più che di verificarle, enucleando lo spettro di eterogeneità culturale presente entro la popolazione di riferimento.

3. I principali risultati dell'analisi culturale

3.1 I Fattori: le principali dimensioni latenti del campo culturale

I principali fattori estratti sono tre. Di seguito, una loro interpretazione sul piano culturale.

Primo Fattore - Simbolizzazione del processo di regolazione sociale: operatività proceduralizzata vs *appartenenza valorizzata*

Invitati a parlare della propria visione dello sviluppo, del proprio ruolo in relazione ad esso, delle politiche in grado di promuoverlo, i soggetti incontrati parlano, in prima istanza, del modo in cui **regolano il processo relazionale con i propri interlocutori.** Da un lato (*operatività proceduralizzata*) lo sviluppo del territorio viene visto come esito dell'adesione, da parte degli attori locali, al frame valoriale e alle azioni proposte dalle politiche di sviluppo. Dall'altro il riferimento è al costruire un comune sentire (*appartenenza valorizzata*), governando e amministrando fatti concreti riguardanti la vita delle persone, avvalorando le valenze affettive dei legami di appartenenza. Due differenti modelli di regolazione sociale, che implicano due visioni alternative di sviluppo: tecnicalità come modello di relazione che funziona a supposto contesto dato (Carli *et al.,* 1999) - lo sviluppo qui è realizzabile per decreto - *vs* modello di regolazione sociale che funziona in modo esperienziale, - lo sviluppo è qui concepito come sviluppo endogeno del sistema (Fini *et al.,* 2015).

Secondo Fattore - Forme del desiderio: salvaguardia vs *riuscita.*

In seconda istanza, i soggetti intervistati parlano della **spinta che muove la loro azione**, ossia della forma del loro desiderio. Da un lato (*salvaguardia*) la trasformazione in mito della comunità di appartenenza sembra rispondere al desiderio di sottrarre la propria storia alla contingenza. Operazione che offre "sicurezza" in cambio di "dipendenza". Dall'altro (*riuscita*) viene messa al centro una dialettica tra identità ed estraneità, con "speranza" e "avvenire" che prendono il posto di "sicurezza". In entrambe i casi "comunità" è lemma centrale, ma mentre nella polarizzazione s*alvaguardia* le parole con cui co-occorre la fanno sembrare valore e scopo dell'azione, nel secondo caso appare più come un prodotto da costruire, dialogicamente, tra dentro e fuori, vecchio e nuovo. Due diverse modalità di entrare in rapporto con l'estraneità: nel primo caso si adatta ciò che è sconosciuto a ciò che già si sa; nel secondo caso si utilizza il noto per esplorare l'ignoto.

Terzo Fattore - Simbolizzazione della domanda di sviluppo: funzione sostitutiva vs *funzione integrativa*

I soggetti intervistati, in terza istanza parlano della **domanda di sviluppo**. Da un lato, laddove ci si propone di adeguare i destinatari alle regole della pianificazione, le regole diventano ordini invalicabili e gli operatori sentono svilito il proprio ruolo ad un mero adempimento e si sentono impotenti. Dall'altro i destinatari delle policy si propongono imprenditivamente, avendo a mente ciò che è rilevante per sé e chiedendo regole che consentano di muoversi all'interno di aspettative condivise. Emergono, polarizzate, due domande di sviluppo: la prima soggiacente ad un modello che potremmo

definire "*sostitutivo*" (Carli, Paniccia, 1999), che attribuisce alla policy un potere elevato, valutabile a prodotto finito, che mette l'impotenza al posto del desiderio. La seconda, relativa ad un modello che potremmo chiamare "*integrativo*" (Carli, Paniccia, 1999) che esprime il desiderio di contribuire al raggiungimento degli obiettivi dei destinatari, in compenetrazione di funzioni e scelte e che pensa per processi.

3.2 I principali Cluster

La *Cluster Analysis* ha individuato 4 Cluster principali.

Tab 1. Contesti Elementari

CL_1	CL_2	CL_3	CL_4
Elementary Context: 407 di 2504 (16,25%)	Elementary Context: 593 di 2504 (23,68%)	Elementary Context: 840 di 2504 (33,55%)	Elementary Context: 664 di 2504 (26,52%)

C1. Le parole con un $\chi 2$ maggiormente significativo (che riportiamo tra parentesi) per questo cluster sono: *tema* (102,4); *amministrazione* (100,6); *aspetto* (83,9); *processo* (68,5); *economico* (66,4); *contesto* (64,4); *imprenditoriale* (62,3); *azione* (50,6); *amianto* (52); *costruire* (49,6); *impresa* (48); *innovazione* (43,5). Abbiamo denominato *C1* ***"Governo imprenditivo dell'innovazione"***, per l'accento posto sull'innovazione, considerata come processo da governare proattivamente.

C2. Le parole maggiormente rappresentative sono: *io* (277,2); *tu* (154,7); *sindaco* (80,5); *parlare* (63,4); *trovare* (62,1); *sentire* (56,7); *persona* (51,7); *giorno* (45,5); *figlio* (41,6); *paese* (34,9); *riuscire* (32,6). Abbiamo denominato *C2* ***"Implicazione nella gestione della cosa pubblica"***, per l'accento posto sulla partecipazione diretta e personale, ognuno con il proprio ruolo e la propria soggettività, al governo del bene comune.

C3. Le parole maggiormente rappresentative sono: *cooperativo* (224,7); *comunità* (182); *notte* (105,3); *anno* (103,1); *Melpignano* (91,2); *fare* (87,8); *cittadino* (83,5); *acqua* (83); *bello* (78); *casa* (75,7); *pagare* (68,9); *euro* (63,7); *Taranta* (60,7). Abbiamo denominato *C3* **"Comunità come identità"** per l'accento posto su tutto ciò che ha reso possibile la costruzione di Melpignano come comunità che si riconosce nella gestione della cosa pubblica e nella valorizzazione della tradizione popolare.

C4. Le parole maggiormente rappresentative sono: *territorio* (442,4); *programmazione* (191,7); *sviluppo* (179,4); *area* (173,9); *regione* (171,1); *GAL* (118,6); *attività* (104,3); *intervento* (102,6); *livello* (90,3); *vasto* (86,9); *Puglia* (77,8); *governance* (75,2). Abbiamo denominato *C4* ***"Pianificazione come***

sviluppo" per l'identificazione del territorio con i confini amministrativi e la sovrapposizione tra sviluppo e varie forme di pianificazione, come se definire confini e pianificare azioni fosse di per sé garanzia di produzione di sviluppo.

3.3 Discussione

La Tabella 2 mostra il rapporto Cluster-Fattori.

Tab 2. Rapporto Cluster-Fattori

Cluster	Fattore1	Fattore2	Fattore3
CL_01	- 22,2374	14,7017	63,7361
CL_02	37,0788	59,5785	- 22,7426
CL_03	60,9616	- 52,9475	0
CL_04	- 81,5382	- 11,9437	- 30,8565

La proiezione dei *Cluster* sullo spazio fattoriale ha consentito di comprendere come concetti simili fossero utilizzati dagli intervistati con significati culturalmente molto diversi.

È il caso, ad esempio, di *C2* (quadrante *riuscita - appartenenza valorizzata* e quadrante *funzione sostitutiva - appartenenza valorizzata)*. I discorsi di *C2* concernono l'essere attivi nella gestione della cosa pubblica. Ma il loro differente posizionamento sullo spazio fattoriale ci ha fatto ipotizzare una differente visione e, di conseguenza, un diverso utilizzo del tema della partecipazione civica, argomento strategico per il contesto locale e per le politiche di sviluppo e strettamente connesso con l'attivazione dei cittadini. Questa ipotesi ha orientato in modo mirato le successive esplorazioni che hanno evidenziato, sotto lo stesso cappello, micro-processi socio-organizzativi molto diversi: da un lato il destinatario di policy visto come soggetto da implicare nella produzione del bene, esplorando e valorizzando il suo desiderio (in coerenza con il quadrante *riuscita-appartenenza valorizzat*a). Qui la partecipazione è considerata esito di una costruzione dialogica. Dall'altro (quadrante *funzione sostitutiva-appartenenza valorizzata*) i destinatari alternativamente visti come fruitori passivi di un bene prodotto da altri o soggetti ai quali delegare sovranità e la partecipazione trattata come strumento di rafforzamento dei sistemi di appartenenza. Questa evidenza ha consentito di superare la classica distinzione presente in letteratura tra processi top down/bottom up (Bens, 2005; Sclavi, 2002) e, in una restituzione ai soggetti locali, di discutere con loro su come lo scarto esistente stesse piuttosto nelle diverse modalità di presa in carico dell'estraneità relativa al desiderio del destinatario delle policy. Grazie al tipo di indagine è stato possibile anche cogliere come temi quali *innovazione* e *comunità*, che nelle

interviste emergevano in modo contiguo come due miti locali per certi versi sovrapponibili, evidenziassero invece posizionamenti culturali differenti: quando a prevalere è *C1-innovazione* (ad esempio: inventare una tradizione come il Festival di musica popolare La Notte della Taranta; introdurre la raccolta differenziata; promuovere presso la cittadinanza l'uso dei pannelli fotovoltaici) le pratiche raccontate sono maggiormente orientate dall'importanza attribuita al raggiungimento di obiettivi (quadrante *operatività proceduralizzata – riuscita)* e dalla necessità di capire come rendere le innovazioni appetibili per la cittadinanza (quadrante *operatività proceduralizzata – funzione integrativa*). Quando invece a prevalere è il tema *C3-comunità* (ad esempio promuovere lo sviluppo di una Cooperativa di Comunità) ciò che sembra essere motore dell'azione è l'idea di rafforzare il proprio sistema di appartenenza (quadrante *appartenenza valorizzata – salvaguardia*; e *appartenenza valorizzata-funzione sostitutiva*). Infine la proiezione di *C4* sullo spazio fattoriale nel quadrante *operatività proceduralizzata – salvaguardia* e *operatività proceduralizzata – funzione sostitutiva* ha consentito di cogliere quanto, entro questo assetto culturale, la pianificazione si muova in modo avulso dai contesti anche laddove la retorica dei programmi preveda strumenti per l'ascolto e la partecipazione dei destinatari delle policies. Da sottolineare, poi, come le variabili illustrative si siano polarizzate maggiormente sul primo fattore: *operatività proceduralizzata* vs *appartenenza valorizzata.* Tecnici da un lato e cittadini/politici dall'altro; policy designer da un lato e policy maker/destinatari dall'altro. Queste polarizzazione ci hanno fatto pensare ad una vicinanza culturale tra policy maker/politici e destinatari/cittadini, evidenziando come la politica locale, a differenza di quella centrale, sia in una posizione privilegiata per comprendere domande e interpretare esigenze, limiti, potenzialità di sviluppo dei contesti reali. Gli attuatori, invece, si posizionano in opposizione a policy maker, destinatari e policy designer. Questo ci ha interrogati sul loro difficile ruolo di cuscinetto, tra le domande dei diretti interlocutori della politica (destinatari, policy maker) e le esigenze intrinseche ai programmi.

4. Conclusioni

L'indagine culturale realizzata mediante analisi testuale ha consentito al team di ricerca di costruire l'oggetto di indagine a partire da elementi altrimenti difficilmente individuabili, dal momento che i contenuti proposti dagli intervistati si presentavano pressoché identici. Poter cogliere tali differenze sostanziali dal punto di vista culturale ci ha permesso di realizzare osservazioni, interviste, discussioni con gli attori locali in merito a quando andavamo capendo ben più mirate e interessanti, anche per i soggetti locali stessi. In ciò riposa la vera innovazione che l'indagine culturale ha consentito

al Progetto REVES di apportare nel campo della valutazione delle politiche di sviluppo locale.

Riferimenti bibliografici

Attanasi, G., Giordano, G. (2011). *Eventi, cultura e sviluppo. L'esperienza de "La Notte della Taranta"*. Milano: Egea

Bens, I., (2005). F*acilitating with ease! Core skills for facilitators, team leaders and members, managers, consultants and trainers*. San Francisco: Josey-Bass.

Blalock, Jr., H. M. (1960). *Social Statistics*. New York: McGraw-Hill Book Company.

Carli R., Paniccia, R.M (1999). *Psicologia della formazione.* Bologna: Il Mulino.

Carli, R., Paniccia, R.M. (2002). *L'Analisi Emozionale del Testo.* Milano: Franco Angeli.

Fini, V., Belladonna, V., Tagle, L., Celano, S., Bonaduce, A., & Gaeta, L.G. (2016), *Progetto Pilota di Valutazione Locale, Studio di Caso: Comune di Melpignano. Come Stato centrale, fondazioni e Regioni possono sollecitare la progettualità locale* retrieved at http://valutazioneinvestimenti.formez.it/sites/all/files/2_reves_rapporto_caso_melpignano.pdf

Fini, V., Salvatore. S. (in press). The fuel and the engine. A general semio-cultural psychological framework for social intervention. In S. Schliewe, N. Chaudhary & P. Marsico (Eds*.), Cultural Psychology of Intervention in the Globalized World*. Charlotte (NC): Information Age Publishing.

Fornari, F. (1979). *I fondamenti di una teoria psicoanalitica del linguaggio*. Torino: Boringhieri.

Lancia F. (2004). *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*. Milano: Franco Angeli.

Matte Blanco, I. (1975). *L'inconscio come insiemi infiniti. Saggio sulla bi-logica*. Torino: Einaudi.

Parmiggiani, P. (2013). Pratiche di consumo, civic engagement, creazione di comunità, in *Sociologia del lavoro*, 132, 97 – 112.

Reinert, M. (1986). Un logiciel d'analyse textuelle: ALCESTE, in *Cahiers de l'Analyse des Données*, 3.

Salvatore, S., & Zittoun, T. (2011). Outlines of a psychoanalytically informed cultural psychology. In S. Salvatore, & T. Zittoun (Eds). *Cultural Psychology and Psychoanalysis in Dialogue. Issues for Constructive Theoretical and Methodological Synergies* (pp. 3-46). Charlotte, NC: Information Age.

Sclavi, M. (2002). *Avventure Urbane. Progettare la città con gli abitanti*. Milano: Euleuthera.

A proposal for Cross-Language Analysis: violence against women and the Web

Alessia Forciniti, Simona Balbi
University of Naples Federico II - alessia.forc@libero.it

Abstract
Aim of the paper is investigating the mood on the Web with respect to one of the most relevant Human Rights violation, without any geographic distinction: the violence against women. While the literature that studies the phenomenon is rapidly growing, the action field is still fragile and the question marks are about the relationship between the public opinion and the contextual factors. In a first look at the phenomenon, we aim at mapping gender violence on the Web, in a Big Data perspective. The peculiar problem we deal with consists in analysing short documents (tweets) written in six European different languages, in the occasion of a common event: the International Day for the Elimination of Violence against Women, 25 November 2017. For our statistical analysis, we choose a multi-linguistic, cross-national perspective. The basic idea is that there are some common structures, language independent ("concepts"), which are declined in the different national natural language expressions ("terms"). Investigating those structure (e.g. factors of lexical correspondence analyses separately performed on the different collections), enables a double level analysis trying to understand and visualise national peculiarities and communalities. The statistical tool is given by Procrustes rotations.

Keywords: Big Data, Text Mining, Cross-national study, Procrustes rotations

1. Introduction

This paper proposes a statistical-linguistic analysis about the mood on Web in relation to a social issue of universal relevance: the violence against women (European Union Agency for Fundamental Rights (FRA), 2014; ONU and United Nations Population Fund, 2016, 2017). The social media, today, are becoming an important platform of the collective thought of the society and therefore, they represent an interesting container of context to study. The constant growth in unstructured information on Web makes the Text mining applications increasingly important in achieving to knowledge extraction of the phenomena. This work faces the problem of the public opinion on the phenomenon of gender-based violence, in Europe, as reply to a common

event: the International Day for the Elimination of Violence against Women (United Nations, General Assembly, 1999), 25 November 2017.The proposed method of analysis is a multi-linguistic, cross-national study of the multimedia contents extracted from Twitter through Web scraping techniques. The features of data (Wu X., Wu G-Q.,Zhu et al.,2014) propose an analysis in terms of Big Data (Zielinski et al., 2012). Considering the aspects of the comparative research (Finer, 1954; Lijphart, 1975) the choice of number of cases study does not excess the six European countries; three west countries, as United Kingdom (Uk), Italy, France and three east countries, as Bulgaria, Czech Republic and Romania.

The research takes on several methodological issues; it requires the treatment of multilingual corpora (tweets are written in six different languages) and not all the treated languages in this study are typical of the Textual Data mining application. The implications are relative to: a careful pre-processing step (corpora cleaning from URL and emoticons), it does not exists a package or software that includes a list of stop words for all investigated languages in this research and in addition the appropriate system of weights for the analysis unit in relation to the nature of data (short messages of up 140 characters). The accuracy of these choices is very important for the good result of the investigation. Therefore, this work has not only a simple cognitive function of the phenomenon but it represents an opportunity to test the scientific method. The cross-linguistic perspective is given by projection on factorial plan of the most frequent terms for couples of countries. In order to visualize the national peculiarities and communalities, the factors are projected in the two different natural languages on a common reference space, per pairwise through the Procrustes rotations.

2.Theoretical Framework

In order to visualize the relationships between document and between terms, in textual data analysis, is commonly performed a factorial approach. The starting point is a lexical table, cross-tabulating *terms and documents* (in this case *terms and tweets*).

This study in question intends propose a Procrustes analysis, such as efficient geometric technique to align lexical matrices. Our research proposes six lexical tables ($\mathbf{X}_1, ..., \mathbf{X}_6$,) as many as there are the case studies. There is an extremely wide multivariate analysis literature devoted to the problem of comparing and synthesising information contained in two or more matrices. An interesting way of approaching the problem consists in comparing geometrical configurations in some Euclidean space (Gordon, 1981). In our case, Correspondence Analysis (CA) is performed on the six tables and visualises the major themes and suggests similarities and peculiarities

between countries. In order to have a measure of this similarity for couple of countries, we can compute the sum of the square distances between corresponding points in the two configurations:

$$\Delta^2(X_{UK}, X_{IT}) \equiv \sum_{i=1}^{p}\sum_{k=1}^{n}(X_{UK} - X_{IT})^2$$

The data structure consists of two matrices, **X** *(n,p)* and **Y** *(n,p).* **X** is the lexical table having in row the *n tweets* in which the corpus is organized, and in columns some content bearing words selected among the most frequent terms in the corpus for a country. **Y** is the lexical table having in row the *n tweets*, and in columns the content bearing words selected in the natural language of the other country. Through the CA performed on each corpus, we compute the principal coordinates and create two matrices: $\mathbf{X}_1$ and $\mathbf{Y}_1$; which represent coordinate matrices of each language. The coordinates matrices have been standardized and normalized so that is not necessary "re-scaling" factor".

3. Data extraction: the Web Scraping

The Social media are a potentially infinite source of user data, and Twitter is one of the worldwide used Social network. Twitter is a micro-blogging service which messages (called tweets) of up to 140 characters. Web scraping is the process of automatically extract data from the Web by an Application Programming Interface (API) supported by software (or by packages connected to software). For our research, data extraction has been conducted with API Twitter and R, respecting specific parameters, common for each country: a keyword translated in the different 6 languages, the specification of the language, the geocode (in order to exclude urban semantic deriving from dialects or territorial slang which change the common sense of words) and finally the sample size (with technical limits; it is possible to extract until to n=3200 tweets per day). The monitoring period is a week around the International Day for the Elimination of Violence against Women, from 23 November to 30 November 2017.

4. Knowledge extraction of the phenomenon

Considering but at same time overlooking the detailed description of the methodological issues aimed at pre-processing procedures of multi-linguistic and multimedia content, the argumentation focuses on the results. The results represent one of the most interesting developments of our proposal.

However, a note deserves the attention: given the structure and the length of each document (tweet), the system of weights of elementary unit is *tf (term-frequency)* where: $w_{ij} = \frac{n_{ij}}{max\, n_j}$.
The canonical tools used for Textual Data Analysis, such as occurrence values of the most frequent terms does not represent, in this case, a useful tool to comparing relation between countries. There are other statistical tools can enable us to go deeper in understanding of the phenomenon, such as the factorial approach.

4.1. Procrustes analysis for a cross-language study

The scientific method that this research intends to test is the Procrustes Analysis by performing the overlapping of two different configurations. The configurations to comparing are two normalized CA coordinates matrices.

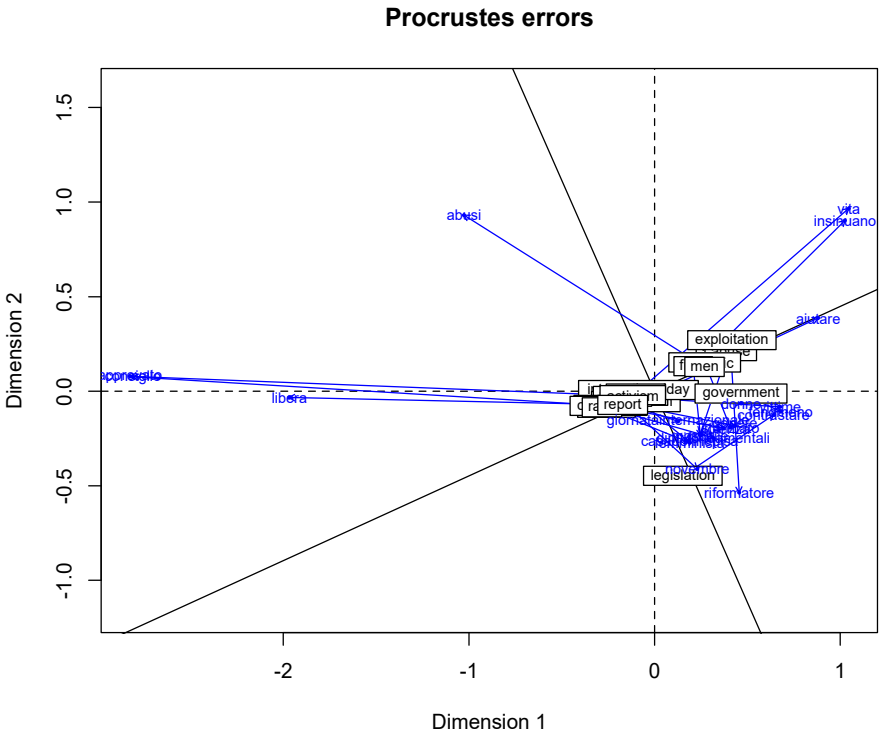


Figure 1. Procrustes errors: comparison between Italy-Uk

The graphic representation allows to observe the Procrustes errors between the two dimensions: points of Italy's normalized principal coordinates matrix and United Kingdom's points of normalized principal coordinates matrix, where Uk is the rotated matrix. Beyond the descriptive statistics about the residual scores, the graph shows how around the axes origin there is a concentration of points both $\mathbf{X}_1$ and $\mathbf{Y}_1$ and so we can affirm that there is not a wide distance between $\mathbf{X}_1$ e $\mathbf{Y}_1$.Procrustean approach confirms the similarity estimated by CA maps between Uk and Italy (*Figure 2* and *Figure 3*); where despite, third quadrant of Italy's and United Kingdom's suffer a dense overlapping of statistics entities, it is possible note similar topics, which are collocated nearly in same position on the multidimensional space.

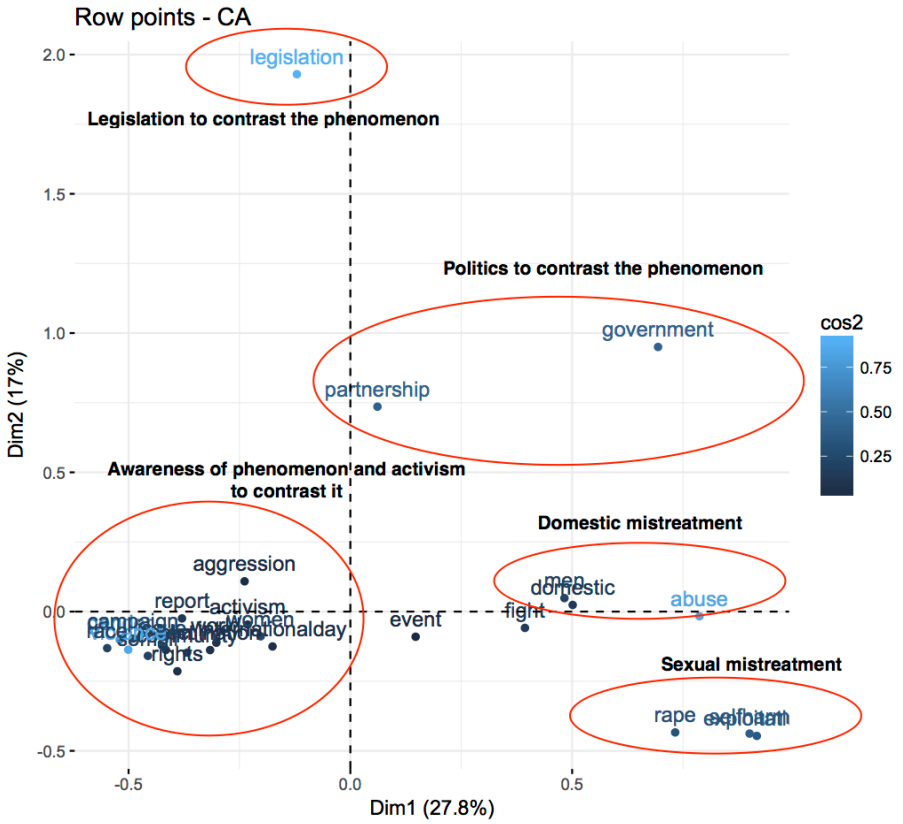


Figure 2. Correspondence Analysis Maps for Uk

Furthermore, through the CA, is possible to investigate structures, language independent ("concepts"), which is declined in the different national natural

language expressions ("terms"). In other words, even though there are terms that they are not the exact translation from a language to another and so from Italian to English o conversely, does not changes the conceptual aspect. Studying the vocabulary of the country we can consider the conceptual aspect and we can create thematic-groupings and to label the clusters. Procrustes errors and t Correspondence Analysis permits to observe the collocation of the statistic entity "abuse". In Procrustes errors plot (*Figure 1*) the "term" is distant from others statistics units; therefore it represent a Procrustes residual. Same consideration is given by observing CA maps (*Figure 2 and 3*). Despite, the word "abuse" is the relative translation of natural language from Italian to English the collocation on the multidimensional space is different. The "joint terms space" (*Figure 4*) of the comparison between Italy and Uk, allows to affirm that the terms that are the exact translation, are almost close in the projected factorial space; e.g. "women", "violence", "international day" and "rights".

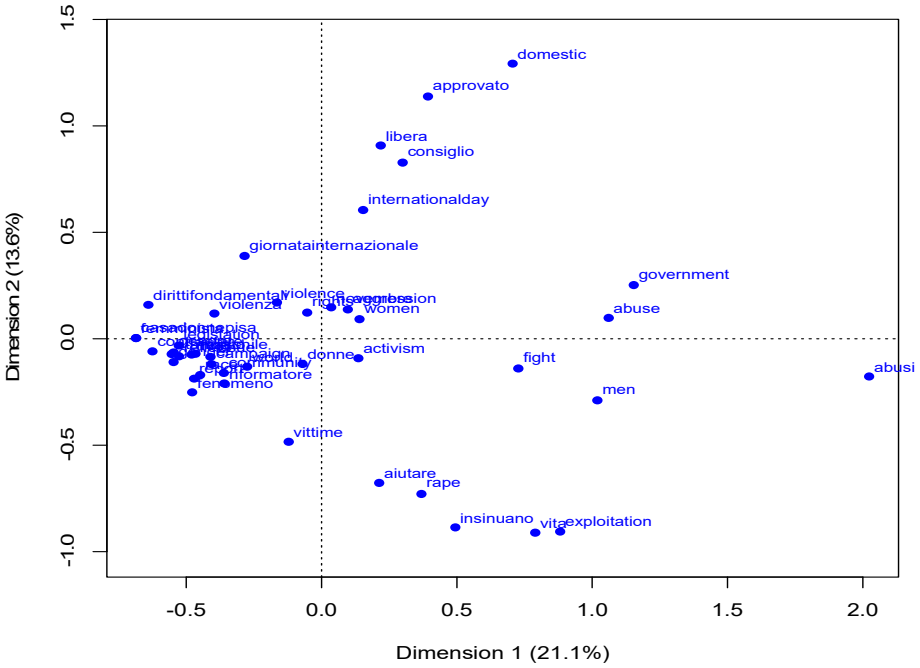


Figure 3. Correspondence Analysis Maps for Italy

Finally, by confirming the Procrustes errors plot (*Figure 1*) and the CA maps (*Figures 2 and 3*), it is possible to see the unit "abuse" (despite the exact

translation) is more distant compared to the relative translation of natural language of the other investigated context. The visualizations of Procrustes Correspondence Analysis and "Joint terms space", test the similarity between Italy and United Kingdom in a cross-linguistic perspective. The graphic intelligibility allows confirming the concordance between the two profiles in relation to public opinion on violence against women.

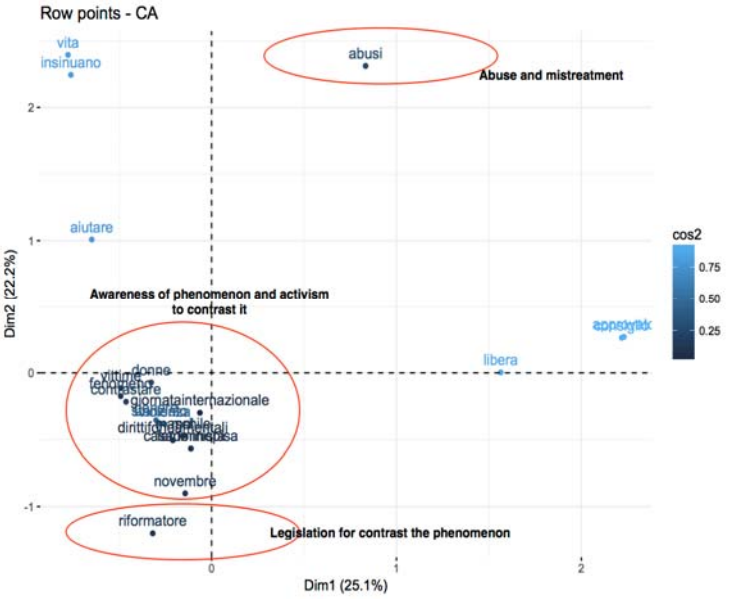


Figure 4. Joint terms space Italy-Uk

In the complex, the visualizations lead us to assert what above mentioned, while singularly they permit to investigate specific aspects of the linguistic peculiarities. The "Joint terms space" confirms the overlapping of statistics units (between countries) around the axis origin, so like the Procrustes errors graph. Therefore, it does not exist a big difference between Italy and Uk. The closeness between the "terms" of different languages collocated on the same reference space recall the thematic-groupings brought out by CA.

5. Conclusion and perspectives

In this paper we faced the problem of comparing corpora when, one is not the translation of the other. Some investigations (e.g. comparison between Uk and Italy) indicate that the Procrustes approach is a valid tool for cross-language study. However, the cross-national investigations, carried out for all case studies, bring out some limits relative to semantic of the natural language expressions of the countries. It is possible that some terms, which

are natural language expressions of a country does not coincide with the translation of the language expressions of another country. For example, in the same case Italy-Uk, we can consider that "reformer" can indicate the political aspect that Uk shows through terms such as "legislation" or "government". Different terms (in natural language expressions) could be ascribable to common conceptual labels since actually are belonging to same semantic category. The future perspective is addressed to resolve the semantic problems between countries by performing an analysis that focuses on study of thematic-axes.

References

Balbi and Misuraca (2006). Procrustes Techniques for Text Minig, in Zani et al., (Eds.*), Data Analysis, Classification and the Forward Search,* pp.227-234 Berlin, Heidelberg: Springer.

Bolasco (1999), Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione, *Carocci*, Roma.

Bolasco (2005), Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica*, Vol.7, pp. 1-37.

European Union (2017). Report on equality between women and men in the EU.

Feldman et al. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems*. Vol. 10, Issue 3, pp. 281–300.

Finer (1954). Metodo, ambito e fini dello studio comparato dei sistemi politici, in *Studi politici*, III, 1, pp. 26-43.

FRA, European Union Agency for fundamental Rights (2014). Report summary: Violence against women: an EU-wide survey. Results at a glance. *Publications Office of the European Union*.

Gower (1975). Generalised Procrustes Analysis. *Psychometrika*, vol.(40):33-51.

Lijphart (1975). The comparable-cases strategy in comparative research, in *Comparative political studies*, VIII, pp. 161-174.

Wu X., Wu G-Q., Zhu et al. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering.* Vol. 26, Issue: 1.

Zielinski et al. (2012). Multilingual Analysis of Twitter News in Support of Mass Emergency Events, *Multilingual Twitter Analysis for Crisis Management*.

La verbalisation des émotions

Béatrice Fracchiolla, Olinka Solène De Roger
University of Lorraine in Metz
beatrice.fracchiolla@univ-lorraine.fr; olinka-solene.de-roger8@etu.univ-lorraine.fr

Abstract
Our study concerns the correlation between the perception of negative emotions and discursive productions to express them. Our study is based on 26 transcribed oral interviews to be analyzed with *Lexico3* (13 men and 13 women). We study the way in which healthy volunteers react verbally to the conditioned production of negative emotions after viewing the government realized video *stop jihad*, broad casted on television after the 2015 attacks. Interviews were collected between November 2016 and February 2017 through out the COREV[1]project framework (understanding verbal violence in reception). At the same time, following an identical protocol, we showed another "neutral" video to the same people in order to have a control group. All the subjects saw both videos, but in different orders, after 11hours of intervals. According to our methodology of analysis with *Lexico3* we were able to extract the linguistic data allowing to have an over view of the emotional feelings perceived by the volunteers after viewing each neutral or violent video and to propose a synthetic card of them. The analysis was conducted with three tools for statistic alanalysis of textual data proposed by *Lexico3*:search for specificity according to the partitions using the PCLC tool (Main Lexicometric Characteristics of the Corpus), the concordances, the graphs of ventilation by partition. The over all analysis of the results shows firstly that the emotions are distributed according to the nature of the videos (neutral video: positive emotions and /or neutral - violent video: negative emotions) and that the violent video provokes a quantity of speech longer than the neutral. Then, if the intensity of perceived emotions seems to differ according to the person wehere show it also is globally correlated to the order of diffusion of the videos. We can see in the responses and the construction of the speeches a correlation of positive or negative intensity of the emotions according to the video which is seen first Like wise, the analysis

[1] The Corev project (2016-2017) which allowed us to constitute the corpus studied is an association of the CNRS, the University of Lorraine and the hospital of Pitié Salpêtrière in order to make a comparative analysis of the neurophysiological responses, emotional and discursive to exposure to (verbal) violence before / after sleep and before / after waking.

seems to show that the reception of the violence invites volunteers and urges them to express them selves more about their feelings: can we see here a correlation also between discursive productivity and negative emotions - a form of verification to the French proverb that "happy people have nothing to say " ?

Résumé
Notre étude porte sur la corrélation qui existe entre la perception d'émotions négatives et les productions discursives pour les exprimer. Elle est réalisée à partir de 26 entretiens individuels oraux retranscrits pour être analysés via *Lexico3* (13 hommes et 13 femmes). Nous étudions la manière dont des volontaires sains réagissent verbalement à la production conditionnée d'émotions négatives après avoir visionné la vidéo stop-djihad du gouvernement, diffusée à la télévision après les attentats de 2015. Les entretiens ont été recueillis entre novembre 2016 et février 2017 dans le cadre du projet COREV[2] (comprendre la violence verbale en réception). Parallèlement, suivant un protocole identique, nous avons montré une autre vidéo « neutre » aux mêmes personnes afin d'avoir un groupe contrôle. Tous les sujets ont vu les 2 vidéos, mais dans des ordres différents, à 11h d'intervalles. Suivant notre méthodologie d'analyse via *Lexico3* nous avons pu extraire les données linguistiques permettant d'avoir un aperçu des ressentis émotionnels perçus par les volontaires après le visionnage de chaque vidéo neutre ou violente et d'en proposer une carte synthétique. L'analyse par *Lexico 3* a été menée *via* trois outils d'analyse statistiques des données textuelles proposés par *Lexico3*: la recherche de particularité selon les partitions à l'aide de l'outil PCLC (Principales Caractéristiques Lexicométriques du Corpus), les concordances, les graphiques de ventilation par partition. L'analyse globale des résultats montre tout d'abord que les émotions sont réparties selon la nature des vidéos (vidéo neutre : émotion positive et ou neutre – vidéo violente : émotion négative) et que la vidéo violente suscite un temps de prises de parole plus long que la neutre. Si l'intensité des émotions perçues semble différer selon la personne nous montrons ici qu'elle est également relative à l'ordre de diffusion des vidéos. Des indices lexicaux ou discursifs nous permettent de vérifier que les sujets qui ont vu d'abord la vidéo djihad réagissent avec plus d'émotions positives

[2] Le projet Corev (2016-2017) qui nous a permis de constituer le corpus étudié est issu d'une association entre le CNRS, l'Université de Lorraine et l'hôpital de la Pitié Salpetrière dans le but de faire une analyse comparée des réponses neurophysiologiques, émotionnelles et discursives à une exposition à de la violence (verbale) avant / après sommeil et avant /après réveil.

à la vidéo « neutre »et , inversement, que celles et ceux qui ont vu la vidéo neutre en premier réagissent avec plus d'émotions négatives lors de la projection de la vidéo stop-djihad. Autrement dit : nous constatons dans les réponses et la construction des discours une corrélation d'intensité positive ou négative des émotions en fonction de la vidéo qui est vue en premier. De même, l'analyse semble montrer que la réception de la violence interpelle les volontaires et les pousse à plus s'exprimer sur leur ressenti : peut-on voir ici une corrélation également entre productivité discursive et émotions négatives – soit une forme de vérification du proverbe selon lequel « les gens heureux n'ont rien à dire ».

Keywords: verbal violence, discourse analysis, emotions, textual statistical analisis, *Lexico3*

1. Introduction

Dans cette étude, nous nous intéressons à la manière dont des sujets confrontés à des éléments violents extériorisent verbalement leurs émotions. Dans l'expérimentation que nous avons conçue pour y arriver, nous avons travaillé sur différents types de réponses émotionnelles obtenues sur 26 sujets ayant visionné une vidéo « violente » (la vidéo « stop-djihad » diffusée par le gouvernement français suite aux attentats de 2015 – désormais notée vidéo V) et une vidéo « neutre » (sur la nouvelle région Languedoc Roussillon midi Pyrénées – désormais notée N). Le protocole multimodal suivi pour récupérer nos données a été réalisé en milieu hospitalier[3]. Nous avons recueilli plusieurs entretiens individuels semi-directifs portant sur le ressenti émotionnel avant et après la vision des différentes vidéos, ainsi que de nombreuses données neurovégétatives. Cette recherche soutenue par la mission à l'interdisciplinarité du CNRS entre novembre 2016 et décembre 2017 visait plus particulièrement la compréhension et la perception de la violence verbale chez des sujets sains (Fracchiolla et al., 2013). L'expérimentation ainsi menée nous permet à la fois de mettre en évidence certains des éléments marqueurs d'extériorisation émotionnelle verbale et de comparer les types de réponses aux vidéos V et N. La présente publication porte exclusivement sur la dimension verbale de l'extériorisation des émotions, une fois le corpus des entretiens menés avec nos sujets retranscrit et étudié à l'aide du logiciel *Lexico3*. Notre approche sera ici plus

[3] Dans le service de et en collaboration avec la Professeure Isabelle Arnulf, Neurologue, directrice de l'unité des pathologies du sommeil de l'hôpital de la Pitié-Salpêtrière, professeure de neurologie à l'Université Pierre et Marie Curie (UPMC), laboratoire : ICM UMR 7225.

spécifiquement de nous demander si les mots que nous utilisons pour nous exprimer sont en adéquation avec ce que nous pensons et surtout avec les émotions ressenties. Notre corpus est ainsi constitué de 26 entretiens répartis en deux groupes comme suit : le Groupe 1 a vu les vidéos dans l'ordre : 1/ Vidéo N – 2/ Vidéo V. Le Groupe 2 : a vu les vidéos dans l'ordre inverse 1/ Vidéo V – 2/ Vidéo N[4].

2. Manifestations d'un discours « émotionné »

2.1. Analyse des PCLP

La répartition du corpus selon la partition « vidéo » avec l'outil PCLC (Principales caractéristiques lexicométriques du corpus), montre les spécificités de cette première partition par vidéo et par groupe. Les interventions des enquêtrices n'y sont pas inclues.

Tableau 1 : Principales caractéristiques de la partition « vidéo »

Partie	Occurrences	Formes	Hapax	Fréquence Max	Forme
V1 N1	8295	1227	689	300	de
V1 N2	33359	2926	1538	1049	de
V1 Neutre	***41654***	***4153***	***2227***	***1349***	***de***
V2 Dj1	7872	1224	685	260	de
V2 Dj2	40191	3325	1679	1225	de
V2 Djihad	***48063***	***4549***	***2364***	***1485***	***de***
Groupe 1	***89717***	***8702***	***4591***	***2834***	*de*
V1 Dj1	12794	1677	906	368	Et
V1 Dj2	35405	2966	1492	1096	Je
V1 Djihad	***48199***	***4643***	***2398***	***1464***	***Je***
V2 N1	5790	961	517	168	La
V2 N2	36002	3013	1561	1205	Je
V2 Neutre	***41792***	***3974***	***2078***	***1373***	***Je***
Groupe 2	***89991***	***8617***	***4476***	***2837***	***Je***

Pour le groupe 1 (N en 1 et V en 2) la forme la plus fréquente est « de » alors que pour le groupe 2, c'est « je ». Les caractéristiques sont à peu près équivalentes quelle que soit la vidéo projetée en 1. Quelle que soit la vidéo projetée, et quel que soit l'ordre, pour les deux groupes on remarque que la première exposition à la vidéo provoque moins de réactions (paroles= nombre de formes) que la seconde, ce qui est *a priori* dû au fait que les entretiens 2 (soir) et 3 (lendemain matin) contiennent un entretien de

[4] L'un des principaux critères de recherche était de voir si les émotions étaient plus ou moins mieux intégrées à 11h d'intervalle de jour ou de nuit. Tous les sujets ont donc vu les 2 vidéos deux fois, à 11h d'intervalle entre chaque projection. 13 sujets dans l'ordre vidéo V matin et soir et N soir et matin, 13 sujets au contraire dans l'ordre vidéo N matin et soir et V soir et matin.

mémoire de la vidéo, avant la seconde projection sont plus longs. Cependant, quel que soit l'ordre de passage, l'ensemble des sujets, tout groupes confondus, parlent plus (environ 7000 occurrences de plus), à propos de la vidéo V (stop djihad), qu'à propos de la N. Une tendance se dessine ainsi selon laquelle la confrontation à la violence provoquerait une prise de parole en « je » et un besoin de parler plus important.

2.2. Analyse du lexique « émotionné »

Reconnues comme des « moments » spécifiques instantanés, les émotions sont définies comme « une réaction physique et/ou psychologique due à une situation. », dont l'effet peut parfois se prolonger plus ou moins dans le temps en fonction de leur intensité (Coletta & Tcherkassof, 2016; voir aussi Bourbon, 2009 ; Feldman et al,. 2016 ou Fiehler, 2002). Pour étudier le lexique des émotions, nous avons regroupé sous formes de listes des mots identifiés dans le corpus et en fonction des concordances comme se rapportant à l'expression de 4 des 6 émotions de base selon Ekman (1972) à savoir : la joie, la colère, la tristesse et la peur (ici nommée inquiétude). Ce choix de 4 émotions et du terme « inquiétude » au lieu de « peur » a été fait en adéquation avec les tests BMIS (échelles d'auto-évaluation de l'état émotionnel par les sujets) demandés aux volontaires avant et après chaque projection de vidéo. Les termes du lexique « émotionné » sont rassemblés ci-dessous par « groupes de formes ». Ainsi par exemple agréable+ contient agréable(s)(ment) :

Bonheur/ Joie : Adoucit ; agréable+ ; allégresse ; ambiance+ ; amusé+ ; apaisant+ ; bon+ ; calme+ ; content+ : désir+ ; emballer+ ; émerveillé ; émouvoir+ ; excitant+ ; fière ; gai+ ; heureux+ ; jaloux* ; joie,+ ; marrant+ ; paisible ; ravi ; serein+ ; surpris+

Colère : aberrant+ ; agacée+ ; agressé+ ; blasé+ ; chiffonne ; choc/choquer+ ; colère ; énerver+ ; fâcher ; frappant+ ; furieux ; haine ; hard ; heurté+ ; horreur+ ; horripile+ ; hostile+ ; irriter+ ; révolter+ ; saoulé

Inquiétude/ Peur :agitation+ ; angoissant+ ; anxiété+ ; apeuré+ ; crainte ; effraiement*, effrayant+ ; flippant+ ; gêne+ ; incompréhensible+ ; nerveux+ ; perdre+ ; peur+ ; stressant+ ; terreur

Tristesse : affecter+ ; affreux+ ; attristé+ ; bouleversé+ ; déception/déçu+ ; dégoût+ ; déprimant+ ; dérange+ ;désolant+ ; impuissance ; malheureusement, malheureux ; mélancolique; navrée ; peine+ ; triste+

Nous avons ici fusionné les émotions positives et neutres dans un même groupe, ce qui explique que sous « joie » soient listés les termes « apaisante, calme, serein » qui ne signifient pas éprouver de la joie, mais dont l'axiologie est évaluée comme positive car exprimant une certaine neutralité émotionnelle (Kerbrat-Orrechioni, 1980). De même, le terme « jaloux » dans la colonne « joie » prête à interrogation : la jalousie est normalement associée

à l'expression d'un désir négatif, de l'ordre de l'inquiétude et de la colère ; mais elle traduit ici du désir, comme le montre le contexte : «[…] ça faisait, ça faisait très envie et ça rendait un peu jaloux». Ici, « jaloux », comme « envie », exprime un désir positif, qui va dans le sens d'un bien-être, contrairement à son axiologie sémantique intrinsèque. De même, le terme « chiffonne » (préoccuper, contrarier) est également une émotion négative qui devrait trouver sa place plutôt dans la colonne de l'inquiétude. Mais en contexte, il correspond ici à de la colère (« énerve » serait ici un synonyme) : « […] ça me, ça me chiffonne un peu de voir ce genre de, de, de vidéo à chaque fois ». Enfin, le néologisme « éffraiement* », substantif masculin construit sur le verbe effrayer, est ici associé à la peur, nous permettant de le classer dans la colonne inquiétude : « un petit peu de peur et, et d'effraiement[5] ». D'une manière générale, pour une étude fine, tous les termes ici listés nécessiterait une analyse développée, en contexte ; ce qui est l'objet d'une autre publication.

3. Evaluation des émotions en contexte

L'analyse en concordance du lexique émotionné relevé ci-dessus révèle des éléments significatifs avec le tri « avant », synthétisés dans le tableau ci-dessous. Ces résultats ont été doublés par des graphiques de ventilation :

Tableau 2 : synthèse des locutions adverbiales ou adverbes accompagnant les expressions des émotions

	Joie	Colère	Inquiétude	Tristesse
un (petit) peu	10	37	37	36
un (peu) plus	8	0	4	0
(encore/beaucoup) plus	20	27	8	9
aussi	0	2	2	0
assez	5	9	2	0
plutôt	8	8	1	2
moins	7	5	0	0
pas très	8	0	0	0
pas	12	0	0	7
très	13	0	1	0
vraiment	0	3	4	0
autant	0	0	3	0
surtout	0	0	0	4

[5]On peut ici interroger à un niveau plus large le principe même de la création néologique en rapport avec le contexte de l'émotion, qui peut se traduire au niveau de la production verbale comme au niveau du corps, par différentes perturbations (bégaiement, intonation, respiration changée, ne plus trouver ses mots…) (voir Plantin, 2016) ; perturbations dont la création de néologismes serait l'une des manifestations sur le plan lexical.

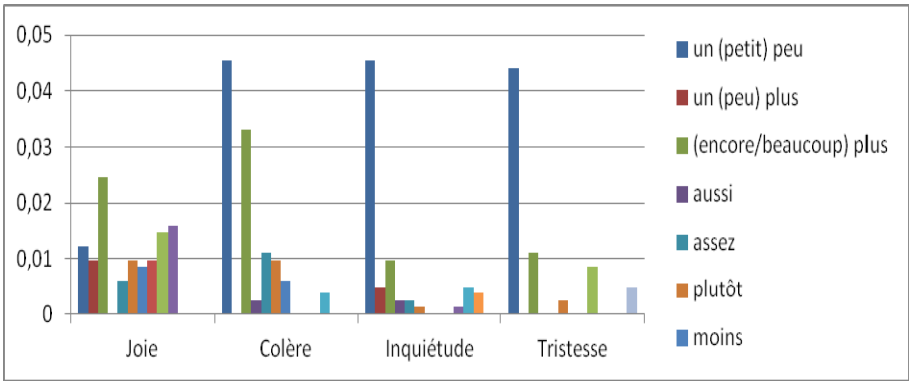


Figure 1 : Histogramme représentant les locutions adverbiales présentes à proximité des expressions d'émotion (fréquences relatives)

Le contexte interactionnel de l'étude où l'on demande aux interviewés d'évaluer les émotions ressenties, génère comme on le voit des réponses presque systématiquement accompagnées d'adverbes ou locutions adverbiales exprimant une intensité positive, équivalente, ou négative. De manière significative, on relève ensuite une accentuation de l'intensité positive lorsqu'il s'agit d'exprimer la joie (« encore/beaucoup/plus » 20 fois, « très » 13 fois) alors que « un (petit) peu » est hyper présent pour atténuer significativement les émotions négatives ressenties (colère, inquiétude, tristesse). La seconde projection graphique permet de voir que, lorsque la joie est exprimée, elle l'est de manière plus diverse, comparativement aux émotions négatives. Ces résultats indiquent que pour le corpus étudié, qui s'intéresse à la réception d'un discours violent, l'expression de l'intensité correspond à celle d'une atténuation. On peut voir par exemple que l'inquiétude et la tristesse sont les émotions qui attirent le plus la locution d'intensité « un peu » qui tend à restreindre l'intensité de l'émotion perçue par le locuteur (Coupin, 1995). Il est possible également que cela soit dû au fait que ce sont des émotions plus diffuses et plus difficiles à caractériser de manière tranchée que la joie et la colère, que l'on identifie assez facilement lorsqu'on les ressent. Cela est confirmé par le fait que les émotions positives sont accompagnées de locutions adverbiales marquant une forte intensité *(encore/beaucoup ; plus* et *très)* : les locuteur.trice.s expriment leur joie avec certitude et n'ont pas peur de la dire. De manière significative, c'est également le cas pour l'expression de la colère, qui semble être l'émotion la plus caractérisée adverbialement, à la fois par des éléments atténuateurs et par des éléments intensificateurs (*«un (petit) peu»* 37 occ. et *« encore/beaucoup/plus »* 27 occ.), ce que l'on peut interpréter comme l'expression du fait que les volontaires ne sont pas particulièrement

heureux.ses de se trouver exposé.e.s deux fois à la vidéo V et le manifestent de cette manière. Le contexte apparaît ici fondamental : la colère est liée d'une manière ou d'une autre ici à une forme d'impuissance face à la fois aux attentats terroristes, aux images montrées qui sont en lien plus ou moins direct selon les sujets, avec les attentats et l'état d'urgence et avec la situation des civils syriens.

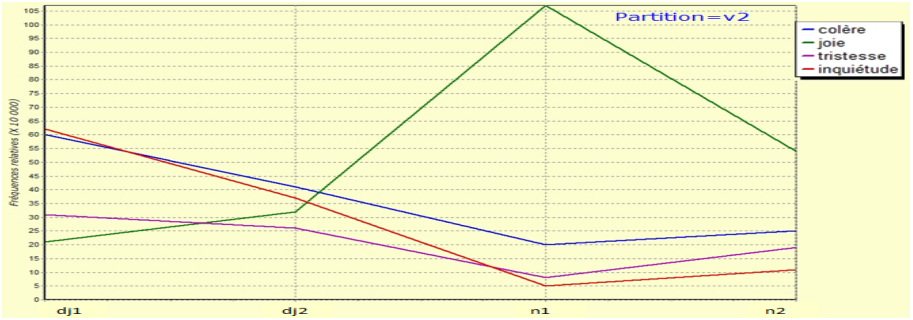


Figure 2 : Graphiques de ventilation par partition : V en N

Les graphiques de ventilation par partition vidéo V et N montrent les émotions exprimées par les volontaires selon les vidéos visualisées. Les émotions négatives (colère, inquiétude, tristesse) sont élevées en V ; à l'inverse la joie est assez élevée en N. On remarque une variation des émotions entre le premier et le second visionnage des vidéos : en effet, la verbalisation des émotions négatives tend à baisser lors du second visionnage (V1 à V2) alors que les émotions positives augmentent de V1 à V2. Le même phénomène s'observe à l'inverse :les émotions positives baissent de N1 à N2, et les négatives augmentent de N1 à N2, ce que montre le tableau ci-dessous :

Tableau 3: tableau récapitulatif des graphiques de partition v1 et v2

	Groupe 1			Groupe 2		
	V1=N	V2=DJ	V1 – V2	V1=DJ	V2=N	V1 – V2
Joie	159	154	5	245	259	14
Colère	153	215	62	167	105	62
Inquiétude	145	202	57	100	43	57
Tristesse	84	134	50	124	74	50

Conclusion

Les réactions des sujets montrent de manière attendue, que la vidéo V génère des émotions négatives et N, des émotions positives. En revanche, l'intensité des émotions exprimées tend à être influencée par l'ordre dans lequel sont vues les vidéos :dans le groupe 1, 1'expression de la joie est exprimée 159 fois ; elle est exprimée 259 fois en N dans le groupe 2. Lorsque les volontaires voient d'abord la vidéo V, il semble que leurs réactions émotionnelles tendent statistiquement à l'inverse de ce à quoi elles tendent dans l'ordre contraire : ainsi l'expression verbale d'une émotion de bonheur tend à être supérieure lorsqu'ils voient la vidéo N après la V, et l'expression de la colère, l'inquiétude et la tristesse sont nettement inférieures. L'étude du lexique émotionné tend à montrer que les sujets ressentent plus de bien être lorsqu'ils voient la vidéo N après la V, comme un soulagement, un apaisement qui arrive après une scène violente. Lorsque la vidéo N est vue en premier, néanmoins, un certain facteur de stress émotionnel demeure, dû probablement au fait que les sujets découvrent l'expérimentation et ne savent pas ce qu'ils vont voir.

References

Bourbon B., (2009). *L'expression des émotions & des tendances dans le langage*, University of Michigan Library.

Colletta J.-M. et Tcherkassof A. (2003). *Les émotions. Cognition, langage et développement.* (P. Mardaga, Éd.) Belgique:Mardaga.

Coupin C. (1995). *La quantification de faible degré : le couple peu/un peu et la classe des petits opérateurs*, thèse de doctorat, dir. Oswald Ducrot, EHESS.

Feldman B. L., Lewis M., Haviland-J. et Jeanette M. (2016). *Handbook of Emotions*, Fourth Edition, The Guildford Press.

Fiehler R. (2002). « How to Do Emotions withWords : Emotionality in Conversations », in Fussel, Susan (ed.) *The Verbal Communication of Emotions*, London, Lawrence Erlbaum,pp.87-107.

Fracchiolla B., Moïse C., Romain C. et Auger N. (2013). *Violences verbales Analyses, enjeux et perspectives.* Rennes: Presses Universitaires de Rennes.

Kerbrat-Orecchioni C. (1980) *L'énonciation. La subjectivité dans le langage*, Paris, A. Colin.

Perrin L. (2016). « La subjectivité de l'esprit dans le langage », in Rabatel A. et al. (éds) Sciences du langage et neurosciences (Acte du colloque de l'ASL 2015), Lambert-Lucas, 189-209.

Plantin Ch. (2011). *Les bonnes raisons des émotions. Principes et méthode pour l'étude du discours émotionné***.** Berne, Peter Lang.

Improving Collection Process for Social Media Intelligence: A Case Study

Luisa Franchina[1], Francesca Greco[2], Andrea Lucariello[3], Angelo Socal[4], Laura Teodonno[5]
[1]AIIC (Associazione Italiana esperti in Infrastrutture Critiche) President – blustarcacina@gmail.com
[2]Sapienza University of Rome – francesca.greco@uniroma1.it
[3]Hermes Bay Srl – a.lucariello@hermesbay.com
[4]Hermes Bay Srl – a.socal@hermesbay.com
[5]Hermes Bay Srl – l.teodonno@hermesbay.com

Abstract
Social Media Intelligence (SOCMINT) is a specific section of Open Source Intelligence. Open Source Intelligence (OSINT) consists in the collection and analysis of information that is gathered from public, or open sources. Social Media Intelligence allows to collect data gathering from Social Media web sites (such as Facebook, Twitter, YouTube etc…). Both OSINT and SOCMINT are based on the Intelligence Cycle. This Paper aims to illustrate advantages gained by applying text mining to collection phase of the intelligence cycle, in order to perform threat analysis. The first step for detecting information related to a specific target is to define a consistent set of keywords. Web sources are various and characterized by different writing styles. Repeating this process manually for each source could be very inefficient and time consuming. Text mining specific software have been used in order to automatize the process and to reach more reliable results. A partially automatized procedure has been developed in order to gather information on specific topic using the Social Media Twitter. The procedure consists in searching manually a set of few keywords to be used for a specific threat analysis. Then TwitteR of R Statistics was used to gather tweets that were collected in a corpus and processed with T-Lab software in order to identify a new list of keywords according to their occurrence and association. Finally, an analysis of advantages and drawbacks of the developed method.

Abstract
La Social Media Intelligence (SOCMINT) è una sezione specifica di Open Source Intelligence. L'Open Source Intelligence (OSINT) consiste nella raccolta e analisi di informazioni da fonti pubbliche o aperte. La Social Media Intelligence consente di raccogliere dati da siti Web di social media (come Facebook, Twitter, YouTube ecc.). Sia l'OSINT che la SOCMINT sono basate

sul ciclo di Intelligence. Il presente documento intende illustrare i vantaggi ottenuti applicando tecniche di text mining alla fase di raccolta del ciclo di intelligence, al fine di eseguire analisi delle minacce. Il primo passo per individuare le informazioni relative ad un obiettivo specifico è definire un insieme coerente di parole chiave. Le fonti Web sono varie e caratterizzate da diversi stili di scrittura. La ripetizione manuale di questo processo per ciascuna fonte potrebbe essere molto inefficiente e dispendiosa in termini di tempo. Sono stati utilizzati software specifici di text mining per automatizzare il processo e ottenere risultati più affidabili. È stata sviluppata una procedura parzialmente automatizzata al fine di raccogliere informazioni su argomenti specifici utilizzando il Social Media Twitter. La procedura consiste nella ricerca manuale di un gruppo di poche parole chiave da utilizzare per un'analisi specifica delle minacce. Quindi il pacchetto TwitteR di R Statistics è stato utilizzato per raccogliere i tweet che sono stati raccolti in un corpus ed elaborati con il software T-Lab al fine di identificare un nuovo elenco di parole chiave in base al loro verificarsi e associazione. Infine viene fornita un'analisi dei vantaggi e degli svantaggi della procedura sviluppata.

Keywords: Social Media Intelligence, Twitter, text mining, data collection

1. Introduction

"Open Source Intelligence [OSINT] is the discipline that pertains to intelligence produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement" (Headquarters Department of the Army, 2010, p. 11-1). OSINT is mainly used in the framework of national security, by law enforcement to conduct investigations, and in business field to gather important information. Social Media Intelligence (SOCMINT) is a specific section of OSINT which focuses on Social Media.

In recent years, with the spread of Internet, and the high amount of readily accessible data, which give a picture of the actual state of things, the importance of OSINT and SOCMINT has grown, becoming a key enabler of decision and policy making. To bring the best out of such flow of data, the intelligence process must take place as a systematic approach structured around clear steps: planning and direction; collection; processing; analysis and production; dissemination. These stages, each of which is vital, create the Intelligence Cycle (CIA - Central Intelligence Agency, 2013). In order to automatically collect data from both the web and the Social Media, OSINT dashboards are being developed (Brignoli et Franchina, 2017).

This paper describes the contribution provided by automated support tools in the collection phase of the Intelligence Cycle from a Social Media (Twitter) on the phenomenon of interest. To capture the real essence of text available and turn data publicly collected into valuable and reliable knowledge, text mining techniques were implemented. To this aim, text mining plays a relevant role as it enables the detection of meaningful patterns to explore knowledge from textual data. As stated by Feldman and Sanger: "Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns" (Feldman et Sanger, 2007, p. 1).

2. The use of Twitter

Twitter is a common Social Media, a microblog mainly for real time information and communication. With Social Media becoming the main tool for informational exchange, in October 2017, Twitter reached about 330 million users (Statista, 2018).

Twitter's specific characteristics makes such a social particularly suitable for SOCMINT purposes. Contents can be accessed by anyone, with no need to create an account. Its users interact with short messages called "tweet", whose length is limited to 280 characters and can be embedded, replied to, liked and unliked. Tweet quick nature, which can then be easily compared to SMS (Short Messaging Service) messaging, fosters the use of acronyms and slang, providing a real-time feel as they bring the first reaction to an event. Phrasing can be simple in structure or imply a large amount of hapax.

With Twitter becoming one of the most important web application, it provides a big amount of data and therefore it constitutes a vital source for Social Media Intelligence. Thanks to its characteristics (potential reach, one-on-one conversation, promotional impact), Tweeter gained importance over years in different social fields, from policy, to media communication and terrorism. As a result, it is commonly considered a valuable source to monitor social phenomena and their changing pattern.

3. Case Study

This paragraph illustrates how text mining tools can be integrated into the SOCMINT data collection phase. The aim of the procedure is to select a suitable and limited list of keywords allowing for an effective and efficient information retrieval in order to support the analyst work.

In this case study the analyst was interested in collecting tweets on the criminal and antagonist threat macro thematic that is related to many specific

topics as, for example, critical infrastructures or telecommunications. The collection process has to identify a list of keyword able to collect the messages concerning, for example, "the criminal and antagonist threat in relation to critical infrastructures". The process can be illustrated by a cycle of four different steps: selection of keywords related with the specific tropic performed by the analyst; tweets collection; text mining; and verification and list of keywords definition (figure 1).

Figure 1: illustration of automatic process for Twitter's data collection four steps cycle

3.2. Keywords selection

The first step is performed by the analyst and consists in defining a suitable list of words which could be used in order to collect tweets related to a specific thematic, which in our example could be *Critical Infrastructures*. To each X topic there is a set of keywords defining it (X_1, X_2, … X_n), e.g., *railway*, *station*, *airport*. The same topic is made by all possible sets, given by the formula:

$$\forall X \exists \{X_1, X_2, \ldots X_n\} : X = \bigcup X_i$$

3.1. Tweets collection

Once the keywords are selected, the second step consists collect data from Twitter repository, e.g. using the twitteR package of R statistics (Gentry, 2016), in order to identify the keywords allowing for the collection of a certain amount of tweets, that in our example was more than one hundred in a day. That is, a word could perfectly represent the topic but could be rarely used in the messages, resulting in a collection of a small sample of tweets. The aim of this step is to find these words that allows for an effective data collection ($n \geq 100$), eliminating those words that are rarely used in the

messages (n < 100). That makes information retrieval more effective as the number of keywords that can be used is limited.

3.3. Text Mining

After the keywords' data collection efficacy was checked, a ten day messages collection was performed including the retweets (49,3%), which is the data retrieval maximum limit of the Twitter repository. The large size corpus (token = 284.253) of 19.491 tweets was cleaned and pre-processed by the software T-Lab (Lancia, 2017) in order to build a vocabulary (type = 19.765; hapax = 8.947) and a list of content words (nouns, verbs, adverbs, adjectives) (table 1). Then the list of content words was checked in order to identify the new keywords and to implement the list.

Table 1: List of the first 20 lemmas of the list

Word	n	Word	n	Word	n	Word	n	Word	n
stazione	6066	elettrico	2226	treno	1198	via	825	ferrovia	659
aeroporto	4734	nuovo	1581	regione	1025	Milano	731	repubblica	632
impianti	3605	rifiuti	1536	Zingaretti	1022	autorizzare	720	giorni	627
Roma	3337	comune	1317	aiutare	896	Italia	679	centrale	605

In order to perform a content analysis, keywords were selected. In particular, we used lemmas as keywords filtering out the lemmas below ten occurrences. Then, on the tweets per keywords matrix, we performed a cluster analysis with a bisecting k-means algorithm (Savaresi et Boley, 2004) limited to twenty partitions, excluding all the tweets that did not have at least two keywords co-occurrence. The eta squared value was used to evaluate and choose the optimal solution.

The results of the cluster analysis show that the keywords selection criteria allow the classification of 98.53% of the tweets. The eta squared value was calculated on partitions from 3 to 19, and it shows that the optimal solution is 13 clusters ($\eta 2 = 0{,}19$) (figure 2). Then, the analyst controlled for the lexical profile of each cluster in order to detect the words useful to focus data collection by means of the Boolean operators.

This procedure allows for the identification of a short list of most used words (about 20) with regard to both the macro thematic and the related topic. The list of keyword was then further reduced and it was reached a set off five meaningful words for each intersection of the macro thematic with a specific topic. Such a reduction stems from the fact that the use of a bigger amount of words led to an exponential increase of false - positive production rate.

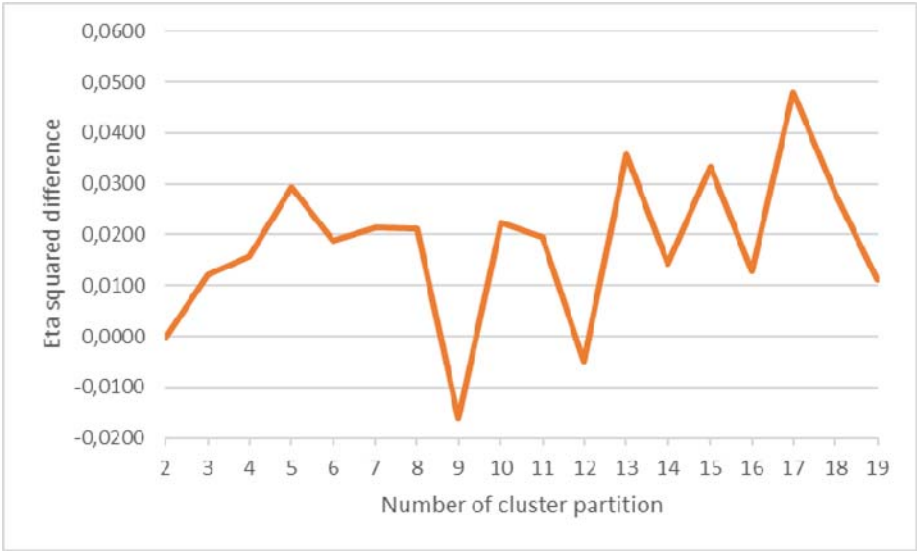


Figure 2: Eta squared difference per partition

As abovementioned, though such a work methodology effectively enables to extract more often used words, with regard to Twitter it is still necessary to test keywords to delete "noise" they produce, which however will not be eliminated entirely. In other words, this methodology affects keywords' amount on the basis of redundancies used by users. However, keywords' quality should be tested in Twitter search engine in order to reach a level of acceptance which includes both false and positive negative. Such words made up the vocabulary to be used to identify intersection between the macro thematic and a specific topic, i.e in the first case "criminal and antagonist's threat with regard to critical infrastructure", in the second case "criminal and antagonist's threat with regard to telecommunication" etc. Between words identified there is an OR relationship. Example: terrorism OR attack OR attack at station OR airport OR railway. Intersection between cluster "criminal and antagonist's threat" and "critical infrastructure is synthetized by the following formula:

$$C = A \cap B = \left| \; \right|(A_i \cap B_i) \neq 0$$

Where A is the cluster "criminal and antagonist's threat", B is "critical infrastructure" and C is the intersection, which is "criminal and antagonist's threat with regard to "critical infrastructures". The following image shows an example.

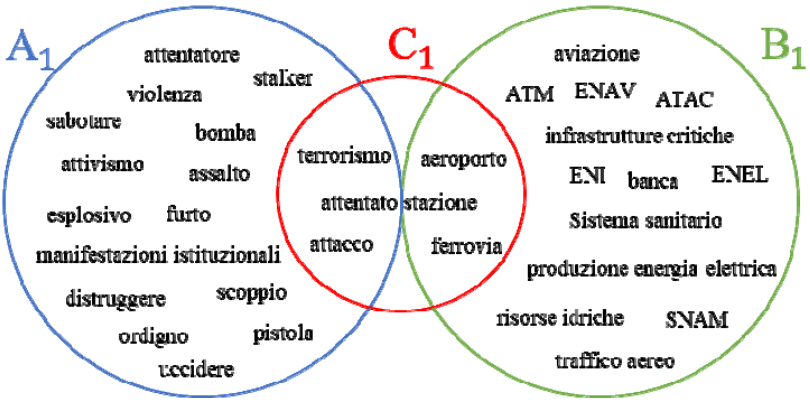


Figure 3: an example of a possible set of words defining the intersection of the cluster "criminal and antagonist's threat", with the topic "critical infrastructure"

3.4. Verification test

Finally, the list of keywords was tested on the Open Source Intelligence dashboard. Collected Tweets were analyzed in order to identify the level of its reliability to monitor the desired phenomena.

4. Conclusion

The developed process reflects the reliability of text mining software in supporting information gathering process for Social Media Intelligence purposes. The vocabulary identified for four different clusters, each of one covering a specific topic, is being tested at this very moment on an advanced dashboard in order to evaluate reliability. However, the role of the analyst is still fundamental. The relationship between OSINT dashboard and analysts must be complementary: dashboard plays a key role in gathering a big amount of tweet, but it is still necessary the analyst support in choosing the suitable keywords to be upload in the database, in order to render information collection more effective. Indeed, OSINT dashboard can't understand Twitter users' use of metaphors and similarities: keywords choice must be made in accordance with monitoring targets. It should be recalled that Italian language is really complex and it might occur that users' language don't refer to chosen target. Let's see a practical example: some keywords which usually refer to criminal threats (bomba - bomb or furto - theft) can be used in Italian language also to refer to synthetic concepts with regard to football or business offers ("bomba" might be used to mean a goal scored through a powerful strike; "furto" might be used to mean that a particular business offer is uneconomical). Another very important issue, which can't be solved without analysts, regard ironic tweets: dashboard

collects all information uploaded into database but it can't subdivide tweets into ironic and non-ironic by means of interpretation. To conclude, as dashboards don't understand textual meaning of words, analysts are required to support dashboards' capabilities, being the only ones to interpret the specific meaning of words.

References

Brignoli M. A., and Franchina L. (2017). Progetto di Piattaforma di Intelligence con strumenti OSINT e tecnologie Open Source. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, pp. 232-241.

CIA, Central Intelligence Agency (2013). Kids' Zone. CIA, https://www.cia.gov/kids-page/6-12th-grade/who-we-are-what-we-do/the-intelligence-cycle.html

Feldman R. and Sanger J. (2006), *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Gentry J. (2016). *R Based Twitter Client*. R package version 1.1.9.

Headquarters Department of the Army (2010). FM 2-0 Intelligence: Field Manual. USA Army, https://fas.org/irp/doddir/army/atp2-22-9.pdf

Lancia F. (2017). *User's Manual: Tools for text analysis.* T-Lab version Plus 2017.

Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis, 8(4):* 345-362.

Statista (2018). Twitter: number of monthly active users 2010-2017. Statista, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

The impact of language homophily and similarity of social position on employees' digital communication

Andrea Fronzetti Colladon, Johanne Saint-Charles, Pierre Mongeau

1. Introduction

Knowledge creation and organizational communication are fundamental assets to obtain strategic competitive advantage (Tucker, Meyer, & Westerman, 1996) and in modern organization most of these happen through digital communication. We know that the way employees use digital communication can predict their engagement level (Gloor, Fronzetti Colladon, Giacomelli, Saran, & Grippa, 2017) as well as future business performance (Fronzetti Colladon & Scettri, 2017). Hence there is a need to better understand what is affecting employees' participation in internal communication in order to foster the efficacy of internal communication and to deliver effective messages and campaigns in the most strategic way. Based on the idea of homophily, this paper examines if employees' participation in their organization intranet is linked with their similarity in discourse and in network positions. Communication, digital or not, encompasses both the language people are using to communicate and the interactions and relationships they have (Tietze, Cohen, & Musson, 2003; White, 2011). In the last two decades scholars have explored how people's discourse[1] and relationships are intertwined notably through the lenses of social network analysis. Among others, those studies have shown that social relationships or interactions between people are linked to the similarity of the words and expressions they use (Basov & Brennecke, 2018; Nerghes, Lee, Groenewegen, & Hellsten, 2015; Roth & Cointet, 2010; Saint-Charles & Mongeau, 2018).

Also, Gloor and colleagues have proposed a framework to study online social dynamics in which language plays an important role, especially with regards to the dimensions of sentiment, emotionality and complexity (Gloor et al., 2017). Such results align with the notion of homophily that corresponds to the tendency to relate to others on the basis of similarities (Lazarsfeld & Merton, 1954). A tendency now acknowledged as an important factors for the constitution of social networks (Mcpherson, Smith-Lovin, & Cook, 2001). It is assumed that this similarity leads to the development of relationships since similarity is linked to attraction towards the other (Montoya & Horton, 2013).

[1] Discourse is define here as "a general term that applies to either written or spoken language that is used for some communicative purpose" (Ellis, 1999, p. 81).

Considering digital communication Brown, Broderick, & Lee (2007) and Yuan & Gay (2006) showed that ties strength and computer-mediated interaction increases with homophily. Most of the studies have explored similarities with regards to sociodemographic variables but several authors have expanded this to a wide range of variables including attitudes, psychological traits, values, etc. as latent homophily factors (Lawrence & Shah, 2007; Shalizi & Thomas, 2011). Hence, given that interaction in digital communication happens through written text, we assume that discourse similarity of employees' messages is a key homophilic determinant for employees' interactions in the network of internal digital communication.

Similarity can also be observed with regard to network position. Indeed, occupying an equivalent position in a network was shown to lead to similar outcomes (attitudes, points of view, roles, etc.) (Borgatti & Foster, 2003; Burt, 1987). In the study of large on-line networks, actors' similarity in centrality has proven useful for identifying role-similarity of actors in the network (Roy, Schmid, & Tredan, 2014). According to Gloor et al. (2017), it is also important to investigate the dynamic evolution of social positions. Rotating leaders, for example, proved to play a very important role in online communities, supporting their growth and participation (Antonacci, Fronzetti Colladon, Stefanini, & Gloor, 2017). In sum, the "homophily phenomenon" has been largely demonstrated through the study of various types of similarities. This paper seeks to explore this phenomenon in the context of the use of internal digital communication system in an organization and we propose to use discourse and network position similarity measures to this avail, our overall hypothesis being that the two are correlated and that they are correlated with interactions.

2. Research Design and Methodology

We analyzed the digital communications of about 1,600 employees working for a large multinational company, mainly operating in Italy. This company has a largely popular intranet social network, structured as an online forum, where only employees can interact, exchanging opinions and ideas through the sharing of news and comments. We could extract and analyze more than 23,000 posts (news and comments), written in Italian over a period of one and a half year. Users were mostly males (68%) and a small part of them also played the role of content managers (7%).

The first step in our analysis was to build the social network which represents the forum interactions. This network is made of N nodes, one for each forum user, and M edges. In general, there is an edge between two nodes if the corresponding employees had at least one interaction – for example, they exchanged knowledge or opinion through subsequent

comments, or one answered a question of the other. We then proceeded to calculate the similarity measures for both discourse and network position.

Based on what was presented above, we looked at five aspects of discourse similarity: words use, sentiment, emotionality, complexity and length. Additionally, we studied employees' connectivity and interactivity, as suggested by Gloor and colleagues (2017). We further explored employees' use of language by looking at the sentiment, emotionality, complexity and length of their forum posts. Length is simply calculated as the average number of characters used in forum posts by an employee – after having removed stop-words and punctuation, via a script written using the Python programming language and the package NLTK (Perkins, 2014). Sentiment expresses the positivity or negativity of forum posts and is calculated thanks to the machine learning algorithm included in the social network and semantic analysis software Condor (Gloor, 2017). Sentiment varies between 0 and 1, where 0 represents a totally negative post and 1 a totally positive one. Emotionality expresses the variation from neutral sentiment and is computed by Condor using the formula presented by Brönnimann (2014). Posts that convey less neutral expressions, either positive or negative, are considered more emotional. Lastly, complexity represents the deviation from common language and is calculated as the probability of each word of a dictionary to appear in the forum posts (Brönnimann, 2014); when rare terms appear in forum posts more often, complexity is higher. Even this last measure was obtained from Condor. Concerning the study of employees' positions in the social structure, we referred to network centrality measures (Freeman, 1979). To measure centrality, we used the two well-known metrics of degree and betweenness centrality. Degree centrality measures the number of direct links of a node, i.e. the number of people an employee interacted with, in the online forum. Betweenness centrality, on the other hand, takes into account the indirect links of a node and counts how many times a social actor lies in-between the paths that interconnect his/her peers. Betweenness centrality is calculated by considering the shortest network paths that interconnect every possible pair of nodes and counting how many times these paths include a specific employee (i.e. the node for which the betweenness centrality is calculated). Employees' interactivity was operationalized by calculating rotating leadership. This variable counts the oscillations in betweenness centrality of a social actor, i.e. the number of times betweenness centrality changed reaching local maxima or minima. If an employee maintains a static position, his/her rotating leadership is zero. On the other hand, we have rotating leaders when people oscillate more between central and peripheral positions, activating or taking the lead of some conversations and then leaving space to other people in the network. As control variables, we could

access to employees' gender and forum role (content manager or not). Even if gender homophily is not always supported by social networks studies, it is very often used as a control variable, as it has been shown that gender can influence online social communication and behavior (Thelwall, 2008, 2009). Similarly, we control for content manager role, as we expect different behaviors when employees have the assignment of informally moderating the intranet social network. All the variables presented above were first calculated at the node level and subsequently transformed into similarity matrices. Like a network adjacency matrix, a similarity matrix is made of N row and columns, where each row and column represents a specific employee. For categorical attributes (gender and being a content manager or not) we have a value of 1 in a cell of the matrix if the two corresponding employees share the same attribute (for example they are both females), and 0 otherwise. For continuous variables, we populated the matrices with the absolute value of the differences in individual actor scores.

3. Results

In general, we notice a prevalence of male employees, even if more forum content managers are females (most of them working in the internal communication department, which is mostly populated by females). Being a content manager is also associated with more central and dynamic network positions: content managers have on average higher scores of degree and betweenness centrality and they rotate more. To put it in other words, they have interactions with more people, often act as brokers of information and in general do not keep a static dominant position after having fostered a conversation.

As described in the previous section, we measured similarity with respect to several characteristics of employees: their gender, content manager role, use of language, centrality and interactivity. Text similarity shows the strongest association with digital communication ($\varrho = 0.48$). Employees who more frequently use the same vocabulary communicate more between themselves. Apart from gender and sentiment, homophily effects seem to be significant for all the other variables included in our study. Employees that are more similar with respect to their use of language, degree of interactivity and network position tend to interact more between themselves.

As per agreed privacy arrangements, we are prohibited from revealing the company name or other details that could help in its identification. It might be useful to replicate our research to see if our findings are confirmed in different business contexts. Future studies could include more control variables, particularly those which are supposed to produce homophily effects – such as employees' age (Kossinets & Watts, 2009). Having more

accurate timestamps could also help in the assessment of average response time, to see if more reactive users tend to cluster. As our was mainly an association study, we advocate further research to carry out a longitudinal analysis which could tell us which actor similarity effects can be considered as significant antecedents of digital communication.

Our findings have practical implications both for company managers and administrators of online communities. For example, if a company wants to attract the attention of employees on a strategic topic, in the light of our results, it appears vital to choose a language close to that of the target people. Employees' participation in conversations can be fostered by online messages aligned with the general use of language and by choosing social ambassadors who have network positions similar to the target.

References

Antonacci, G., Fronzetti Colladon, A., Stefanini, A., & Gloor, P. A. (2017). It is Rotating Leaders Who Build the Swarm: Social Network Determinants of Growth for Healthcare Virtual Communities of Practice. *Journal of Knowledge Management*, *21*(5), 1218–1239. https://doi.org/10.1108/JKM-11-2016-0504

Basov, N., & Brennecke, J. (2018). Duality beyond Dyads: Multiplex patterning of social ties and cultural meanings. *Research in the Sociology of Organizations*, in press.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*. https://doi.org/10.1016/S0149-2063(03)00087-4

Brönnimann, L. (2014). *Analyse der Verbreitung von Innovationen in sozialen Netzwerken*. University of Applied Sciences Northwestern Switzerland. Retrieved from http://www.twitterpolitiker.ch/documents/Master_Thesis_Lucas_Broennimann.pdf

Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, *21*(3), 2–20. https://doi.org/10.1002/dir.20082

Burt, R. S. (1987). Social Contagion and Innovation: Cohesion versus Structural Equivalence. *American Journal of Sociology*, *92*(6), 1287–1335. https://doi.org/10.1086/228667

Ellis, D. G. (1999). *From Language To Communication*. New York, NY: Routledge.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*, 215–239.

Fronzetti Colladon, A., & Scettri, G. (2017). Look Inside. Predicting Stock

Prices by Analysing an Enterprise Intranet Social Network and Using Word Co-Occurrence Networks. *International Journal of Entrepreneurship and Small Business*, in press. https://doi.org/10.1504/IJESB.2019.10007839

Gloor, P. A. (2017). *Sociometrics and Human Relationships: Analyzing Social Networks to Manage Brands, Predict Trends, and Improve Organizational Performance*. London, UK: Emerald Publishing Limited.

Gloor, P. A., Fronzetti Colladon, A., Giacomelli, G., Saran, T., & Grippa, F. (2017). The Impact of Virtual Mirroring on Customer Satisfaction. *Journal of Business Research*, *75*, 67–76. https://doi.org/10.1016/j.jbusres.2017.02.010

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)* (pp. 49–56). Christchurch, New Zealand.

Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications*, *2*(6), 1930–1938. https://doi.org/10.1.1.642.7100

Kossinets, G., & Watts, D. J. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, *115*(2), 405–450. https://doi.org/10.1086/599247

Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks*, *10*(4), 359–381.

Lawrence, T. B., & Shah, N. (2007). Homophily: Meaning and Measures. In *Paper presented at the International Network for Social Network Analysis (INSNA)*. Corfu, Greece.

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a Social Process: A Substantive and Methodological analysis. *Freedom and Control in Modern Society*, *18*, 18–66. https://doi.org/10.1111/j.1467-8705.2012.02056_3.x

Mcpherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships*, *30*(1), 64–94. https://doi.org/10.1177/0265407512452989

Nerghes, A., Lee, J.-S., Groenewegen, P., & Hellsten, I. (2015). Mapping discursive dynamics of the financial crisis: a structural perspective of concept roles in semantic networks. *Computational Social Networks*, *2*(16), 1–29. https://doi.org/10.1186/s40649-015-0021-8

Perkins, J. (2014). *Python 3 Text Processing With NLTK 3 Cookbook*. *Python 3 Text Processing With NLTK 3 Cookbook*. Birmingham, UK: Packt Publishing.

Roth, C., & Cointet, J. P. (2010). Social and semantic coevolution in knowledge networks. *Social Networks*, *32*(1), 16–29. https://doi.org/10.1016/j.socnet.2009.04.005

Roy, M., Schmid, S., & Tredan, G. (2014). Modeling and measuring graph similarity: The case for centrality distance. In *Proceedings of the 10th ACM international workshop on Foundations of mobile computing, FOMC 2014* (pp. 47–52). New York, NY: ACM. https://doi.org/10.1145/2634274.2634277

Saint-Charles, J., & Mongeau, P. (2018). Social influence and discourse similarity networks in workgroups. *Social Networks*, *52*, 228–237. https://doi.org/10.1016/j.socnet.2017.09.001

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM SIGMOD Record*, *36*(2), 7–12. https://doi.org/10.1145/1328854.1328855

Thelwall, M. (2008). Social networks, gender, and friending: An analysis of mySpace member profiles. *Journal of the American Society for Information Science and Technology*, *59*(8), 1321–1330. https://doi.org/10.1002/asi.20835

Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, *60*(2), 219–231. https://doi.org/10.1002/asi.20978

Tietze, S., Cohen, L., & Musson, G. (2003). *Understanding organizations through language*. *Understanding Organizations Through Language*. https://doi.org/10.4135/9781446219997

Tucker, M. L., Meyer, G. D., & Westerman, J. W. (1996). Organizational communication: Development of internal strategic competitive advantage. *Journal of Business Communication*, *33*(1), 51–69. https://doi.org/10.1177/002194369603300106

White, H. C. (2011). *Identité et contrôle. Une théorie de l'émergence des formations sociales*. Paris: Éditions de l'École des hautes études en sciences sociales.

Yuan, Y. C., & Gay, G. (2006). Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams. *Journal of Computer-Mediated Communication*, *11*(4), 1062–1084. https://doi.org/10.1111/j.1083-6101.2006.00308.x

Looking Through the Lens of Social Sciences: The European Union in the EU-Funded Research Projects Reporting

Matteo Gerli
University for Foreigners of Perugia – matteogerli81@gmail.com

Abstract
In the last decades, European integration and scientific production have come to be deeply intertwined as a result of the Europeanization of many research activities. On one side, European institutions promote the realization of research projects aiming at developing a type of knowledge "close" to the end users' interests; on the other side, the resulting knowledge contributes to conditioning the practices that take place in the European and national institutions, according to a circular process that brings the innovations to feed back into the system that expresses them. The purpose of this paper is to explore this relationship by examining two peculiar scientific products realized by researchers operating within the broad domain of the Socio-economic Sciences and Humanities (SSH), as a part of the research projects financed by the Seventh Framework Programme (2007-2013) of the European Union: *final reports* and *policy briefs*. In other words, it aims to analyse all reports as a whole using some automatic text analysis tools, while incorporating some supplementary variables which help to define the broader context of scientific production.
Keywords: European Union, International Research Projects, Socio-economic Sciences and Humanities, Textual Data Exploration, Quantitative Discourse Analysis, IRaMuTeQ.

1. Introduction

The European Research Policy plays a strategic role for thousand of researchers and research institutions which operate within the EU borders. Thanks to the concomitant decrease in national public funds for scientific activities (see for instance, Vincent-Lacrin, 2006; 2009), the European research agenda has dramatically increased its appeal among scholars and consequently its ability to have an impact on the directions and processes of scientific knowledge production. Indeed, starting from the 90s, the European Commission has equipped itself with new means to combine and manage, on the basis of medium to long-term planning cycles, the whole set of scientific and technological initiatives financed by the European budget: the *framework*

programme (Ippolito, 1989; Ruberti and André, 1995; Guzzetti, 1995; Menéndez and Borras, 2000; Borras, 2000; Banchoff, 2002; Cerroni and Giuffredi, 2015). In short, the underlying logic is that of the programmatic intersection between research activities and other European policies, so that the promotion of scientific excellence complements the need to foster the creation of cross-border and interdisciplinary collaborations intended for producing a type of knowledge "close" to the end users' interests.

As it was observed in previous studies (Adler-Nissen and Kropp, 2015), European integration and scientific production have come to be deeply intertwined: on one side, the progress of integration process influenced (and still influences) research activities through the promotion of particular forms of knowledge and research questions (as far as we are concerned, mainly through the realization of cross-national and cross-disciplinary research projects); on the other side, the resulting knowledge contributes to conditioning the practices that take place in the European and national institutions, according to a circular process that brings the innovations to feed back into the system that expresses them. Social Sciences and Humanities, which are less directly involved in the production of knowledge with a clear practical usability, are by no means unconcerned with this kind of phenomenon. At this regard, the *Journal of European Integration* has recently published a special issue on the relationship between social sciences and European integration, hosting some important articles that have highlighted the existence of several "crossroads" between the European Union and the scientific community's "itineraries"[1]: Rosamond (2015), for instance, observed how certain theories on the political and economic integration (in particular that of the Hungarian Béla Balassa, from the economics side, and the neofunctionalism, from the political science side) had been informing the "strategic narrative" adopted by the European Commission during the 60s and 70s to legitimize its newly-formed institutional role and its economic policy position, according to a quite peculiar two-ways traffic of influences process, being the economic integration theorized while it was happening; Deem (2015) pointed out the existence of a relationship between the birth of a new field on higher education studies, the simultaneous evolution of national university systems and the launch of the so-called *Bologna process* at European level; Vauchez analysed, through a sociogenetic approach, the historical process through which the *acquis communautaire* «has been formulated, stretched, criticized, revised and finally naturalized as the most rigorous and objective measure of Europe against other possible methods» (2015: 196) thanks to the work of those who have been defined

[1] *Journal of European Integration*, 37 (2015).

"methodological entrepreneurs", that is European officials who have politically invested and succeeded in establishing Europe's cognitive and technical equipment.

Looking beyond such individual cases, what is really relevant to our purpose is the underlying idea about the possibility of studying science production from a sociological point of view, basically by rejecting what was traditionally regarded as an *internal/external division* (Adler-Nissen and Kropp, 2015: 161-163), and thus admitting that even scientific and academic concepts can be formulated in conjunction with political-economic ambitions and practical problems (see Bohme *et al.*, 1983; Funtowicz and Ravez 1993; Slaughter and Leslie 1997; Gibbons *et al.*, 1994; Ziman, 2000; Albert and Mcguire, 2014), such as those above mentioned. This does not mean that science is equal to politics or economics (Breslau, 1998); what it does mean is that, in order to understand science production, one needs to recognize that "non-academic" resources (such as, for instance, financial or material resources, ideas and beliefs, symbolic resources, political or normative resources, people, etc.) may overstep scientific boundaries and be used for the production of new knowledge. Bourdieu (1975, 1984, 1990, 1992, 1994, 1995, 2001) described this phenomenon through the concept of "fields interrelations". In few words, the social word is composed of multiple semi-autonomous *fields*, basically microcosms characterized by different *stakes*, *rules of the game* and particular resources which one needs to possess to get access to the game itself and its specific advantages. He conceptualized these sphere as partially independent, by which he means that, even though each field develops its own institutions, hierarchies, problems, tacit or explicit rules, they necessarily interact and affect each other. This is particularly true for cultural fields (art, cinema, religion, science, journalism, etc.), since they are structurally dependent and subordinated to political and economic fields. Going straight to the point, this is to say that, if one is dealing with a sociological analysis of a cultural product (e.g. a text), thus one neither can just consider its formal characteristics, nor be limited to its context of production. Instead, one should use a "relational approach", taking into account both the *internal* features of the product and its *external* determinants.

In engaging with this broad issue, this paper will try to further contribute to the understanding of the topic by examining two peculiar scientific products realized by researchers operating within the broad domain of the Socio-economic Sciences and Humanities (SSH), as a part of the research projects financed by the Seventh Framework Programme (2007-2013) of the European Union: *final reports* and *policy briefs*. By using some automatic text analysis tools, it will thus statistically explore the contents of such documents not *per se*, but in connection with some variables, which help to define the broader

context of production. In its exploratory character, this study does not have strong hypothesis to be tested. Nevertheless, following Bourdieu's approach, it aims to give an original perspective through which observing the relationship between the field of social sciences and the public policy field of the European Union (Gerli, 2017).

2. The corpus and methodology

Unlike the studies discussed earlier, which are mainly based on micro-sociological observation, our investigation covers a macro-sociological analysis of a quit large *corpus* made of 46.513 graphic forms, equal to 3.025.960 occurrences. It is an *ad-hoc* constructed *corpus*: it contains 360 texts, of which 205 belonging to *final reports* and 155 to *policy briefs*, which were collected from the digital database CORDIS[2], the main institutional source of information related to the research projects financed by the European Union.

The choice to focus on these documents is not accidental, but depends on their strict relevance to our research objectives. In fact, both include a summary of the project results and conclusions, with a description of their potential socio-economic impact (EC 2010), even though *policy brief* is strictly designed for policy makers (both European and national ones), while *final report* is addressed to a wider audience, which may include (at least potentially) lay people as well. In this perspective, they represent an effective "shortcut" through which empirically observe the way in which the research groups awarded a grant "actualized" the inputs they received from the Commission. This is, to resume the previous discussion, to analyse how European institutions and social scientists contribute together to the definition and resolution of some EU-related issues.

With regard to the methodology, both simple and multivariate analyses were performed with the IRaMuTeQ software (Lebart *et al.*, 1998; Bolasco, 2013). In particular, the lexicographical analysis was used for a first exploration of the *corpus*, that is to identify and format texts units, turn texts into text segments (TS) and classify words by their frequency. The multivariate analysis, instead, was performed to detect the associations between textual data and the following supplementary variables related to what in the 7FP was defined as *macro-activity* (MA) and *financing scheme* (FS)[3]. Going into more details, the 7FP included eight macro-activities: *Growth, employment and competitiveness in a knowledge society* (MA1); *Combining economic, social and environmental goals in Europe: towards sustainable development* (MA2); *Major*

[2] http://cordis.europa.eu/projects/home_it.html.

[3] For more details: Decision No 1982/2006/EC of the European Parliament and of the Council of 18 December 2006.

trends in society and their implications (MA3); *Europe in the world* (MA4); *The citizen in the European Union* (MA5); *Socio-economic and scientific indicators* (MA6); *Foresight studies* (MA7); *Strategic activities* (MA8). As for the financing schemes, the 7FP included five main different types, which differed from each other by the research team size and the type of purposes to be achieved (the first three mainly focused on the development of new knowledge, while the last two were mainly thought for the coordination and support of research activities and policies): *Small or medium-scale focused project* (FS1); *Small or medium-scale focused research project aimed at international cooperation* (FS2); *Large-scale integrating project* (FS3); *Coordination action* (FS4); *Support action* (FS5). Additionally, we also took into account the *starting year* of the project and the *geographic area* in which the coordinating institution was located.

As a whole, our sample (of non-probabilistic type) involves 223 research projects out of 251 realized in 2007-2013 (equal to 88.8%) and broadly covers all macro-activities and financing schemes above mentioned. In Tab. 1, a description of the *corpus* and its main subsets is provided.

Tab. 1: Description of the corpus

Type	*Number of texts*	*Graphic forms*	*Occurrences*
Final report	205	42.047	2.441.168
Policy Brief	155	19.795	584.792
Corpus	360	46.513	3.025.960

3. The main findings

At first glance, the most frequent "full" words used in the SSH research reports do not provide particularly relevant insights. The first ten (*social*, *policy*, *research*, *European*, *project*, *EU*, *countries*, *public*, *national*, *Europe*) concerns the "general context of meaning" where discourses on Europe and related issues took shape. Ten words that, without having a clear disciplinary connotation, define some "semantic coordinates" common to all research projects carried out. Interesting enough, it is the wide use of the words *country/es* (freq.=10.531) and *national* (freq.=5.527) which, compared with the words *European* (freq.=9.190), *EU* (freq.= 8.563) and *Europe* (freq.=5.408), prove the great importance of the "national" level of analysis, mainly in a comparative way. Scrolling down the list, we can also recognise some typical words of the socio-economic lexicon (*economic*, *market*, *growth*, *employment*, *financial*), the socio-political lexicon (*people*, *education*, *State*, *young*, *groups*, *cultural*, *society*, *governance*), and the methodological one, namely related to the operative context of the research activities (*date*, *case*, *results*, *impact*, *analysis*, *study*). Yet these are terms that, at this early stage of the analysis, do not provide any clear "message".

At a closer look, however, we can identify some specific words which are, in a broad sense, linked to the political macro-orientations defined by the Lisbon Strategy (European Council 2000), demonstrating the "osmosis" existing between European institutions and social sciences. Here some examples: *innovation* (freq.=5.793), cornerstone of industrial competitiveness and economic growth (EC 2003, 2006); *development* (freq.=5.176), to be understood, among the various meaning, mainly as sustainable development (EC 2005, 2009); *education* (freq.= 3.490) and *knowledge* (freq.=3.221), which, together with the already mentioned "innovation", represent the "three sides" of the so-called "knowledge triangle", from the European Commission's perspective, the ground for a greater economic and social dynamism.

For the aim of this study, what is of particular interest is also the geographical scope of the research activities. Indeed, the most frequent toponyms refer to EU based countries. Among these, the five main sponsors and recipients of the framework programs (*Germany*, *UK*, *France*, *Italy* and *Spain*) are placed at the top of the ranking. As for the extra-European countries, several of them are placed in Asia (e.g. *China*, *Japan*, *India*, *Vietnam* and *Thailand*), North Africa (*Morocco*, *Tunisia*, *Egypt* and *Libya*) and South America (*Brazil*, *Argentina*, *Colombia*, *Peru* and *Chile*). This is indicative of a globalization process, which is affecting both European institutions and researchers by expanding their interests ("political", with regard to the first ones, and "scientific", for the second ones) beyond the European borders. What matters is that they are moving together insofar we can suppose the existence of a clear synergy between the emergence of a new multipolar area of political, commercial and cultural influence, in which the European Union is now required to act, and the production of knowledge on topics with a potential "global" added value.

3.1 The main semantic groups and their connections with the "context"

To go deeper in the analysis, and to explore the relationship between the selected texts and some variables related to their context of production, we performed a Descending Hierarchical Analysis (DHA). Indeed, this method allowed us, first, to identify clusters with similar vocabulary within text segments and, then, to visualize them in conjunction with the supplementary variables (Camargo and Justo 2013; Curbelo, 2017). In Fig. 1, the output of the DHA is summarised.

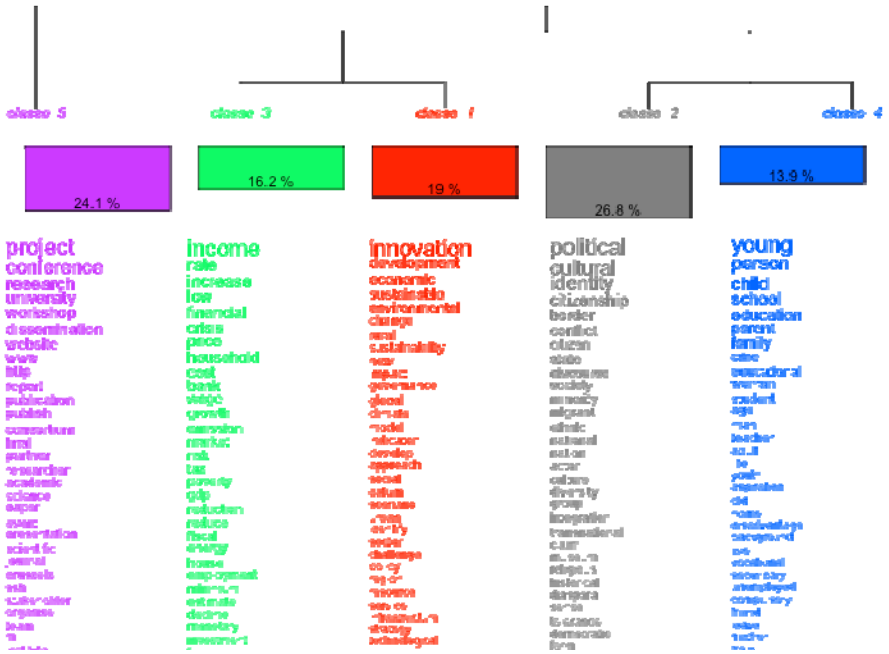


Fig. 1: Dendrogram of top-down hierarchical classification (Reinert method) of the corpus

As it can be easily seen in Fig. 1, the DHA algorithm allowed the identification of five clusters, each with its own specific semantic content. Following Reinert (1987), they can be interpreted as "lexical words", namely specific semantic structures which, in our case, refer to different and even competing scientific representations of the European Union and related issues. The second cluster has the greater representation (26,8% of the SSH discourses) and identifies a semantic sphere characterized by a language mainly oriented towards political and social issues. Indeed, the most central word in this cluster is *political*, followed by *cultural*, *identity*, *citizenship*, *border*, *conflict*, *citizen*, *State* and so on. Immigration (*migrant*) and related issues appear to be particularly relevant as well. The fifth cluster (24,1%) delineates a quite peculiar semantic sphere based on a set of words (such as *project*, *conference*, *research*, *university*, *workshop*, *dissemination*, *website*, etc.) strictly linked with the management and realization of European research projects and, more in general, with scientific research and related activities. The first cluster, third in terms of representativeness (19%), refers to the relationship between economic development and environmental protection, being the most central word *innovation*, followed by *development*, *economic*, *sustainable*, *environmental*, *change*, *rural* and so on. This interpretation seems to be supported by the presence of several words that refer to the need for a change with respect to a situation that is perceived as not desirable (*change*,

impact, *strategy*, *challenge*, *need*, *solution*, *improve*, *step*, etc.). The third cluster (16,2%), instead, covers a semantic area mainly related to the economy and the market. It is a language that involves two main branches, the one of the real economy (*income*, *price*, *household*, *wage*, *firm*, *energy*, *poverty*, etc.), and the one of the finance (*financial*, *bank*, *risk*, *monetary*, *credit*), but above all it is characterized by the large presence of technical terms and acronyms (*gdp*, *estimate*, *asset*, *inflation*, *emu*, *Eurozone*, *insurance*, *macroeconomic*, etc.). Finally, the fourth linguistic cluster (13,9%) includes words essentially associated to the relationship between education, training and employment, as shown by the presence of terms such as *young*, *person*, *child*, *school*, *education*, *aspiration*, *background*, *vocational* and *compulsory*. It is a cluster that differs from the others due to the greater concreteness of the language, as proved by the recurring use of words referring to "concrete" social actors (*child*, *parent*, *student*, *teacher*, *mother*, *friend*, *volunteer*, etc.).

Fig. 2, resulting from a Lexical Correspondences Analysis (LCA), shows the relationship between clusters (left side) and between clusters and the supplementary variables (right side). The main aim here was to verify whether or not SSH discourse exhibits clear evidence of "adaptability" with regard to the macro-activities and the financing schemes, as defined by the European Commission.

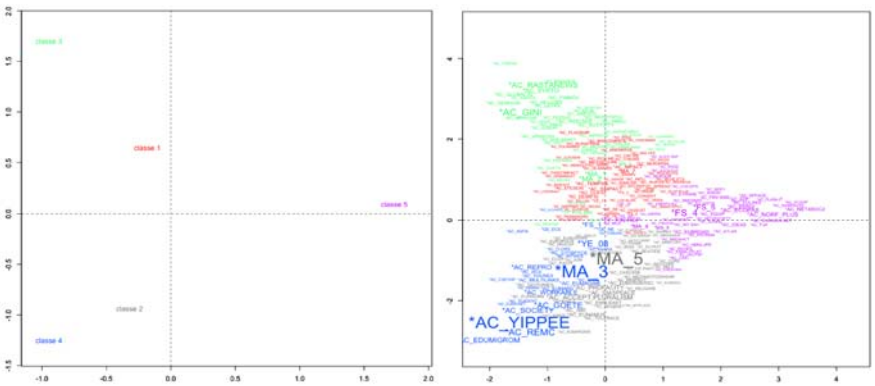


Fig. 2: Association between clusters and supplementary variables

The first two factors summarize together 67,5% of the total inertia: the first one (39,97%) marks a clear opposition between cluster 5 (positive half-plane) and the other four clusters (negative half-plane); the second factor (27,47%), instead, highlights a significant opposition between clusters 1 and 3 (positive half-plane) and clusters 2 and 4 (negative half-plane). As a whole, we can distinguish three different (partially autonomous) semantic contexts, arising from the association between the "cultural" and "socio-political" discourses

(third quarter), the "economic" discourse and that on "innovation" and "sustainable development" (forth quarter), and finally the discourse on "research activities" (in-between the first and the second quarters).
As far as the relationship between discourses (clusters) and supplementary variables, Figg. 3 and 4 show the most significant categories (those with a larger chi-square and a lower p-value), referring to the "macro-activity" and "financing scheme" variables. As shown in the first figure, MA1 and MA2 categories are only significant in the definition of clusters 1 (innovation) and 3 (economics); MA5 is the most relevant for cluster 2 (politics); similarly, MA3 category is the only significant for cluster 4 (culture); and finally, MA4 and MA8 categories predominate on cluster 5 (research activities). In short, these results strongly support the thesis of adaptability, insofar the different scientific representations of the European Union emerged from the analysis resulted strongly associated with the macro-activities defined by the European Commission.

Cluster	Category	Chi2	%	p-value
1	MA2	1226.7	25,7	<0.0001
	MA7	762.9	36.5	<0.0001
2	MA5	5220.0	54.8	<0.0001
3	MA1	1282.4	28.9	<0.0001
	MA2	1414.2	27.0	<0.0001
4	MA3	5238.5	33.0	<0.0001
5	MA4	839.9	33.6	<0.0001
	MA8	534.9	43.7	<0.0001

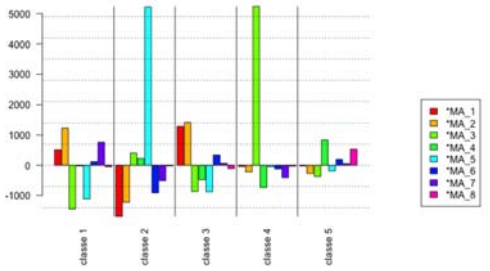


Fig. 3: Chi2 significance of variable "macro-activity" by cluster

On the other hand, the role of the "financing scheme" variable resulted much less significant in discriminating the five clusters, except for categories FS4 and FS5, which are the most significant for cluster 5, and category FS1, which instead clearly prevail on cluster 4. Nothing relevant emerged in relation to the variables "geographic area" and "starting year".

Cluster	Category	Chi2	%	p-value
1	FS2	186.3	25,7	<0.0001
	FS3	145.1	24.7	<0.0001
2	FS1	487.6	29.0	<0.0001
3	FS1	286.5	17.6	<0.0001
4	FS1	1245.0	16.7	<0.0001
5	FS4	2195.0	51.5	<0.0001
	FS5	1583.2	58.5	<0.0001

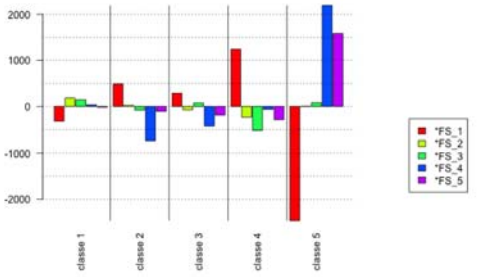


Fig. 4

4. Conclusions

The findings presented herein indicate a close relationship between the programmatic framework, defined by the Commission, and the contents of the *final reports* and *policy briefs*, supporting the thesis of a co-construction of the European integration (Adler-Nissen, Kropp 2015). The scientific discourse has come to be structured around few semantic macro-aggregates arisen from DHA, which in turn resulted associated with the variables performed in LCA. Furthermore, the SSH linguistic space shows a clear cleavage between the economic discourse and the cultural discourse, which points out the existence of a lack of interaction between these two spheres. From a more "general" point of view, all this means that, in connecting the social sciences field with the policy field, the European research projects produced a scientific discourse that, on the whole, is structurally *homologous* with the "space of possibilities" inherent to the 7PQ.

References

Adler-Nissen R., Kropp K. (2015). A Sociology of Knowledge Approach to European Integration: Four Analytical Principles. *Journal of European Integration*, 37(2): 155-173.

Albert M., Mcguire W. L. (2014). Understanding Changes in Academic Knowledge Production in a Neoliberal Era. *Political Power and Social Theory*, 27: 33-57.

Banchoff T. (2002). The Politics of the European Research Area. *ACES Working Paper 3*, Paul H. Nitze School for Advanced International Studies.

Böheme G., Van den Daele W., Hohlfeld R., Krohn W., Shafër W. (1983). *Finalization in Science. The Social Orientation of Scientific Progress*. Dordrecht: Riedel.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Borras S. (2000). Science, Technology and Innovation in European Politics. *Research Paper n. 5*, Roskilde University.

Bourdieu P. (1975). The Specificity of Scientific Field and the Social Condition of the Progress of Reason. *Social Sciences Informations*, 6: 19-47.

Bourdieu P. (1984). *Homo academicus*, trad. it. (2013) *Homo academicus*. Bari: Edizioni Dedalo.

Bourdieu P. (1992). *Les règles de l'art*, trad. it. (2013) *Le regole dell'arte*. Milano: Il Saggiatore.

Bourdieu P. (1994). *Raisons pratiques. Sur la théorie de l'action*, trad. it. (2009) *Ragioni pratiche*. Bologna: Il Mulino.

Bourdieu P. (1995). *Champ politique, champ des sciences sociales, champ*

journalistique, trad. it. (2010) *Campo politico, campo delle scienze sociali, campo giornalistico*. In Cerulo M. (a cura di). *Sul concetto di campo in sociologia*. Roma: Armando.

Bourdieu P. (2001). *Science de la science et réflexivité*, trad. it. (2003) *Il mestiere di scienziato*. Milano: Mondolibri.

Breslau D. (1998). *In Search of the Unequivocal: The Political Economy of Measurement in U.S. Labor Market Policy*. London: Praeger.

Camargo B. V., Justo A. M. (2013). R Interface for Multidimentional Analysis of Texts and Questionnaires, *IraMuTeQ tutorial*, available on: http://www.iramuteq.org.

Cerroni A., Giuffredi R. (2015). L'orizzonte di Horizon 2020: il futuro europeo nelle politiche della ricerca. *Futuri*, 6: 29-39.

Curbelo A. A. (2017). Analysing the (Ab)use of Language in Politics: the Case of Donald Trump. *Working Paper n. 2*. University of Bristol: SPAIS.

Deem R. (2015). What is the Nature of the Relationship between Changes in European Higher Education and Social Science Research on Higher Education and (Why) Does It Matter?. *Journal of European Integration*. 37(2): 263-279.

European Commission (2010). *Communicating research for evidence-based policymaking*. Bruxelles: Directorate-General for Research.

European Commission (2003). *Politica dell'innovazione: aggiornare l'approccio dell'Unione Europea nel contesto della Strategia di Lisbona*. COM(2003) 112 definitivo, 11.03.2003.

European Commission (2005). *Comunicazione della Commissione al Consiglio e al Parlamento europeo sul riesame della strategia per lo sviluppo sostenibile. Una piattaforma d'azione*. COM(2005) 658 definitivo, 13.12.2005.

European Commission (2006). *Mettere in pratica la conoscenza: un'ampia strategia per l'innovazione per l'UE*. COM(2006) 502 definitivo, 10.05.2006.

European Commission (2009). *Integrare lo sviluppo sostenibile nelle politiche dell'UE: riesame 2009 della strategia dell'Unione Europea per lo sviluppo sostenibile*. COM(2009) 400 definitivo, 24.07.2009.

Funtowicz S., Ravez J. (1993). Science for the Post-Normal Age. *Future*, 25: 735-755.

Gerli M. (2017). Il campo sociale dei progetti di ricerca europei. Il caso delle SSH. *Studi Culturali*, 1: 127-150.

Gibbons M., Limoges C., Nowotny H., Schwartzman S., Scott P. e Trow M. (1994). *The New Production of Knowledge*. London: Sage.

Guzzetti L. (1995). *A Brief History of European Union Research Po*licy. Luxembourg: Publications Office of the European Communities.

Ippolito F. (1989). *Un progetto incompiuto. La ricerca comune europea: 1958-88*. Bari: Edizioni Dedalo.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. New York: Kluwer Academic.

Menéndez L. S., Borrás S. (2000). *Explainig Changes and Continuity in EU Technology Policy: The Politics of Ideas*. In Dresner S. e Gilbert N. (eds), *Changing European Research System*. Aldershot: Ashgate.

Reinert M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud. *Bulletin de Méthodologie* Sociologique, 13: 53-90.

Rosamond B. (2015). Performing Theory/Theorizing Performance in Emergent Supranational Governance: The Live Knowledge Archive of European Integration and the Early European Commission. *Journal of European Integration*, 37(2): 175-191.

Ruberti A., André G. (1995). *Uno spazio europeo della scienza. Riflessioni sulla politica europea della ricerca*. Firenze: Giunti.

Slaughter S., Leslie L.L. (1997). *Academic Capitalism: Politics, Policies and the Entrepreneurial University*. Baltimore: The John Hopkins University Press.

Vauchez A. (2015). Methodological Europeanism at the Cradle: Eur-lex, the Acquis and the Making of Europe's Cognitive Equipement. *Journal of European Integration*, 37(2): 193-210.

Vincent-Lacrin S. (2006). What is Changing in Academic Research? Trends and Futures Scenarios. *European Journal of Education*, 41(2): 169-202.

Vincent-Lacrin S. (2009). Finance and Provision in Higher Education: A Shift from Public to Private?. *Higher Education to 2030 (vol. 2)*, Centre for Education Research and Innovation: OECD.

Ziman J. (2000). *Real Science: What It Is, and What It Means*. Cambridge-New York: Cambridge University Press.

Spécialisation générique et discursive d'une unité lexical L'exemple de *joggeuse* dans la presse quotidienne régionale

Lucie Gianola[1], Mathieu Valette[2]
[1]Université de Cergy-Pontoise – lucie.gianola@u-cergy.fr
[2]Institut National des Langues et Civilisations Orientales– mvalette@inalco.fr

Abstract
In this paper, we study the distribution of lexical items designating outdoor sport practitioners (*joggeur/joggeuse, randonneur/randonneuse, unneur/runneuse, promeneur/promeneuse*), in order to identify links between gender, semantic themes and genre across press discourse in French. The corpus is sampled from newspaper articles from regional newspapers. In press discourse, we observe a convergence between gender and genre through the actualized semantic classes.

Résumé
Nous étudions dans cet article la distribution d'unités lexicales désignant les pratiquant·e·s de sport de plein air (*joggeur/joggeuse*, *randonneur/randonneuse*, *runneur/runneuse*, *promeneur/promeneuse*) afin d'identifier les corrélations entre genres sexuels, thèmes sémantiques et genres textuels dans le discours journalistique en français. Le corpus est constitué à partir d'un échantillonnage d'articles de la presse quotidienne régionale. Il apparaît que dans le discours journalistique, on observe une convergence entre genres sexuels et genres textuels par le biais des classes sémantiques instanciées.
Keywords: Press discourse, textometrics, semantic class, genre, gender

1. Introduction

Nous proposons une étude de lexicologie textuelle sur la distribution d'unités lexicales choisies dans un corpus de textes de presse. L'étude n'a pas été réalisée dans une perspective *corpus-driven*, comme c'est souvent le cas en textométrie, mais avec une approche *corpus-based* (Biber, 2009) où les observables ont été prédéfinis. Notre objectif est en effet de nous focaliser sur les désignations des pratiquant·e·s de sport de plein air suivant une opposition en genres sexuels : *joggeur vs joggeuse*, *randonneur vs randonneuse*, *runneur vs runneuse*, *promeneur vs promeneuse*. Il s'agit d'identifier les corrélations entre genres sexuels, isotopies et genres textuels dans le discours journalistique de la presse quotidienne régionale française.

2. Problématique

2.1. Sommation des isotopies de genres et de discours en signifiés

La lexicologie textuelle consiste en l'analyse du lexique à partir des conditions textuelles de sa production. Elle repose sur l'hypothèse selon laquelle les unités lexicales subissent un ensemble de contraintes intertextuelles et infratextuelles de la même nature que les formes sémantiques diffuses et non lexicalisées et qui en conditionnent les régimes de production et d'interprétation. Dans de précédents travaux, ont été proposées les conditions théoriques d'une analyse textuelle du lexique, principalement focalisées sur l'étude de la néologie sémantique – ou *néosémie* (Rastier et Valette, 2009) et des formes sémantiques diffuses en voie de lexicalisation synthétique ou *protosémie* (Valette, 2010ab). Il s'agit ici d'étudier l'utilisation systématique d'une unité lexicale donnée dans un genre textuel précis et l'incidence de cette utilisation sur son sémantisme. En effet, tout mot placé dans un texte en reçoit des déterminations sémantiques, qui sont susceptibles de modifier son signifié (afférence de sèmes). Posant l'hypothèse selon laquelle le signifié est une forme sémantique lexicalisée (Valette 2010b), on considérera que les sèmes des isotopies du texte peuvent se propager au signifié d'une unité lexicale par le processus de *sommation* décrit par (Rastier, 2006). L'observation a pu être faite concernant les isotopies de domaine (redomanialisation d'une unité lexicale dans le cas de la néosémie par exemple) mais les isotopies génériques (relatives au genre textuel) ou discursives (relatives au discours) peuvent-elles transformer le signifié d'un mot de la même façon que les isotopies domaniales ? C'est à cette question que nous allons tâcher de répondre ici.

2.2. Présentation du corpus

Le corpus est donc constitué suivant deux axes, lexical et discursif : nous avons utilisé 8 formes considérées comme des mots-clés pour collecter des textes exclusivement issus du discours journalistique et, plus précisément, de la presse quotidienne régionale, sans considération de genre textuel. Le corpus a été collecté de manière semi-automatique à l'aide d'un script d'aspiration de pages web puis nettoyé et dédoublonné manuellement, afin d'écarter des articles constitués de reprises de dépêches AFP qui se retrouvent d'un titre à un autre. Le script, basé sur la commande Linux *cURL*, est alimenté pas une liste d'URL collectées sur les sites des titres de presse à l'aide de requêtes effectuées sur le moteur de recherche Google (*site:nomdusite forme*, modulée par un inhibiteur *-blade* dans le cas de « *runner* » afin d'écarter les articles à propos du film *Blade Runner*). Entre 100 et 130 URL ont été collectées pour chaque forme. La phase de nettoyage a permis de supprimer les en-têtes, sommaires, liens annexes, légendes

d'images, etc., pour ne conserver que le titre et le corps de l'article. Le corpus est organisé en huit sous-corpus correspondant aux 8 formes étudiées : *Joggeur*, *Joggeuse*, *Promeneur*, *Promeneuse*, *Randonneur*, *Randonneuse*, *Runner*, *Runneuse*, dont les statistiques sont présentées dans le tableau suivant.

Table 1 : Analyse factorielle des correspondances sur les parties du discours

Sous-corpus	**Nombre de mots**
Joggeur	40 671
Joggeuse	48 285
Randonneur	35 162
Randonneuse	31 931
Promeneur	44 497
Promeneuse	31 009
Runner	22 212
Runneuse	31 367
Total	**285 134**

Les articles sont issus principalement de titres de la presse quotidienne régionale comme *Nice Matin*, *Ouest-France*, *L'Est Républicain*, *La Dépêche du Midi*, *La Montagne*, *Corse-Matin*, *La Provence*. La collecte n'a pas été orientée sur une rubrique en particulier mais sur l'ensemble des titres, et nous n'avons pas défini de limite temporelle.

3. Analyses[1]

3.1. Observations générales

Une analyse factorielle préliminaire (figure 1) portant sur les seules parties du discours montre une opposition marquée sur l'axe 1 entre les sous-corpus *Runner* et *Runneuse* et les autres sous-corpus. Cet écart s'explique par les genres textuels des sous-corpus considérés. En effet, comme l'ont montré les travaux pionniers de (Biber, 1988) et, à leur suite, ceux de (Malrieu et Rastier, 2001), les variables locales que constituent les parties du discours sont des marqueurs de genre particulièrement stables. Ici, il apparaît que *Runner* et *Runneuse* relèvent du genre du compte rendu d'événements sportifs tandis que les 6 autres sous-corpus sont composés en grande majorité de faits divers. Autrement dit, la plupart des unités lexicales choisies pour nos requêtes, qui correspondent à des pratiques sportives de plein air,

[1] Le corpus a été analysé au moyen du logiciel de textométrie TXM (http://textometrie.ens-lyon.fr/) (Heiden *et al.* 2010).

n'appartiendraient pas – ou alors à la marge – au vocabulaire des genres sportifs du discours journalistique.
L'analyse factorielle des correspondances sur les formes, dont la fréquence est au moins égale à 10 occurrences, offre à voir une distribution très différente. *Runner* et *Runneuse* sont toujours très proches mais il en est désormais de même de *Randonneur* et *Randonneuse* (désormais *Randonneur·se*) (figure 2). Les sous-corpus *Joggeur*, *Promeneur* et *Promeneuse* se situent à la croisée des axes et seront étudiés individuellement, mais *Joggeuse* se singularise.

3.2. Analyses des classes sémantiques constituantes
L'analyse des spécificités (formes) des regroupements ainsi constitués nous indique les contextes d'instanciation des différentes formes.

Le regroupement a priori très homogène *Randonneur·se* offre à voir un vocabulaire associé aux *accidents de montagne*. Le corpus est structuré en 3 classes sémantiques principales,

- celle des accidents : « chute », « mortelle », « mètre », « avalanche », « fracture », « cheville », « hôpital », « blessée », « trauma », « glisser » etc.
- celle des disparitions : « disparu », « alerte », « retrouvé », « emporté », « inquiet », etc.
- celle des secours : « PGHM » (pour Peloton de gendarmerie de haute montagne), « hélicoptère », « Dragon » (un modèle d'hélicoptère) « évacué·e », « pompiers », « CRS », « secouriste », « secteur », « équipe », « sauveteur », « secourir », etc.

Le sous-corpus *Promeneur* et le sous-corpus *Promeneuse* relatent essentiellement 3 types d'événements :

- la promenade : « sentier », « phare », « littoral », « patrimoine », « chemin », etc.
- les accidents : accident de chasse essentiellement : « chasseurs », « chasse », etc.
- les découvertes : « macabres », « corps », « cadavre », « tronc », « jambe », « squelette », « ossement », « obus », « pépite », etc.

Le sous-corpus *Joggeur* ne comporte quant à lui qu'une classe sémantique principale, celle des accidents n'incluant pas de tiers humain : « arrêt, malaise, crise cardiaque », « algues vertes », attaques d'animaux (« rapace », « aigle », « buse »), sulfure d'hydrogène, H_2S, intoxication, toxique, gaz. Il est à noter que cette classe ne s'actualise pas dans le sous-corpus *Joggeuse.*

Les deux sous-corpus restants, le regroupement très homogène *Runner* et *Runneuse* (désormais *Runneur·se*) et *Joggeuse* méritent toute notre attention. D'un point vue ontologique, le jogging comme le running sont des formes similaires de course à pied relevant du domaine du sport. Mais leur usage dans le discours journalistique diffère très sensiblement. Dans le regroupement *Runneur·se*, qui comporte, comme nous l'avons vu, essentiellement des articles relatant des événements sportifs, le vocabulaire est structuré autour des classes sémantiques suivantes :

- définitoire : hyperonyme « sport », synonyme « coureur », etc. Ainsi, le sous-corpus *Runneur·se* est le seul dont le sens correspond à la signification.
- classe de la compétition : « course » « marathon », « semi-marathon », « trail », «triathlon », « championnat », « inscription », « départ », « épreuve », « km », « victoire », « podium », « médaille », « sponsors », etc.

- classe des blessures : « blessure », « foulure », « ampoule », « contracture », etc.

Il comporte également deux classes sémantiques liées aux techniques associées à la pratique :

- classe des équipements : « équipement », « baskets », « chaussures », « brassière », « connectés », « GPS » ou « montre GPS », etc.
- classe des entrainements : « entrainement », « préparation », « fractionné », « cardio », « conseils », « performances », « yoga » (comme activité complémentaire destinée à éviter les blessures), etc.

Il est à noter que le sous-corpus *Runneuse* se singularise par la mention d'événements sportifs caritatifs liés à la lutte contre le cancer du sein : « octobre rose », « prévention ».

A l'inverse, la joggeuse dans le sous-corpus éponyme n'est nullement une sportive, mais sa caractérisation textuelle est remarquablement précise : elle est une femme agressée pendant son jogging et les classes sémantiques actualisées dans ce sous-corps relèvent du crime, du droit et de l'enquête judiciaire :

- classe des agressions : « meurtre », « tentative », « agressée », « agression sexuelle », « viol », « enlèvement », « tuée »,
- classe des agresseurs : « homme », « suspect », « meurtrier », « présumé », « portrait-robot », « violeur », « exhibitionniste »
- classe des procédures judiciaires : « enquêteurs », « avocats »,

« cour », « procureur », « réquisition », « réclusion », « prison », « accusé », « interpellé », « agresseur », « condamné », « procédure », « instruction », « ADN », etc.

3.3. Synthèse

A l'issue de cette analyse, on choisit de se concentrer sur la définition en miroir de la *joggeuse* et de la *runneuse*, laissant de côté les autres unités lexicales détaillées ci-dessus. Les isotopies génériques et discursives qui constituent la trame sémantique des articles dans lesquels occurrent ces deux formes donnent lieu à la construction de deux signifiés antagonistes, par sommation :

La joggeuse apparaît :

1. /isolée/ (elle court seule),
2. /vulnérable/ (elle est sans défense face à un agresseur) et, quoi qu'il arrive, puisque le genre du fait divers l'exige,
3. /victime/ (elle est agressée, violée, tuée).

A l'inverse la runneuse est :

1. /entourée/ (elle court dans le cadre d'événement sportifs collectifs),
2. /sécurisée/ (par la technologie, notamment les montres GPS qui permettent de gérer l'effort et d'optimiser ses performances, par l'entraînement suivi. Les blessures subies apparaissent par ailleurs bénignes par rapport aux risques encourus par la joggeuse),
3. /compétitrice/ (elle participe à des compétitions).

4. Conclusion

Dans cet article, nous avons tenté de montrer comment les fonds sémantiques issus des genres et des discours pouvaient modifier, par sommation, les signifiés des unités lexicales qui sont utilisées. Pour deux unités lexicales partageant a priori un référent identique, celui d'une femme pratiquant la course à pied, l'actualisation en corpus journalistique fait émerger des contenus sémantiques très différents. Il ne s'agit pas de considérer que les joggeuses sont nécessairement des femmes en danger mais la régularité avec laquelle le mot joggeuse est actualisé dans la presse comme une /victime/, /vulnérable/ et /isolée/ pourrait avoir, à terme, une incidence sur la perception d'une pratique dont la réalité médiatique est exclusivement macabre. En d'autres termes, dans le discours de presse, pour les femmes, le jogging est une pratique dangereuse, la joggeuse une victime d'agression, alors que la runneuse une sportive impliquée dans des événements sociaux et le running une pratique sûre et valorisante.

Références

Biber, D. (1988). *Variation across Speech and Writing* . Cambridge, Cambridge University Press.

Biber, D. (2009). Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In B. Heine and H. Narrog (editors) *The Oxford Handbook of Linguistic Analysis*, 159–191. Oxford.

Heiden S., Magué J.-P., et Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement, S. Bolasco. editors., *Journées internationales d'Analyses statistiques des Données Textuelles*, vol(2), 1021-1032.

Malrieu, D. et Rastier, F. (2001). Genres et variations morphosyntaxiques, In *Traitements automatiques du langage*, 42, 2, 547-577.

Rastier, F. (2006). Passages. In *Corpus*, 6, 125-152.

Rastier, F., Valette, M. (2009). De la polysémie à la néosémie. In *Le français moderne*, vol. (77), 97-116.

Valette, M. (2010a). Propositions pour une lexicologie textuelle. In *Zeitschrift für Französische Sprache und Literatur*, vol. (37): 171-188.

Valette, M. (2010b). Méthodes pour la veille lexicale, In L. Messaoudi, *et al.* editors *Sur les dictionnaires*, Publication du laboratoire Langage et société, Université Ibn Tofail, Kénitra: 251-272.

The Transparency Engine – A Better Way to Deal with Fake News

Peter A. Gloor[1], Joao Marcos de Oliveira[2], Detlef Schoder[3]
[1]MIT Center for Collective Intelligence, Cambridge MA – pgloor@mit.edu
[2]Galaxyadvisors, Aarau Switzerland – jmarcos@galaxyadvisors.com
[3]University of Cologne, Germany – schoder@wim.uni-koeln.de

Abstract
We introduce the "Transparency Engine", a social network search engine to separate fact from fiction by exposing (1) the hidden "influencers" and (2) their "tribes". Our goals are to quantify the influence and relevancy of persons, concepts, or companies on institutions, issues or industries by tracking the dynamics and changes in the observed environment. In particular we visualize the networks of influence for a given social or economical ecosystem, thus providing a tool to both the scientific and general public (including journalists, or anyone interested to check news) to track the diffusion of new ideas, both good and bad. In particular, the Transparency Engine exposes the hidden influencers behind fake news. We propose a unique solution, which combines three subsystems we have been developing over the last five years: (I) Powergraph, (II) Tribefinder, and (III) Swarmpulse, The powergraph displays the degree and power of the spreader's position by re-constructing her/his (social) network via Web sites and social position in the Twitter-universe. The tribefinder exposes the tribal echo chambers on Twitter nurturing fake news items through social media mining, thus allowing the news consumer to develop an informed opinion for identifying the motivation of the spreaders of fake news. This is done through mining Twitter word usage of tribe members with neural networks using tensorflow. The swarmpulse system finds the most relevant fake and non-fake news on Wikipedia and Twitter by combining their emergent patterns.

Keywords: Fake News, Transparency Engine, News, Truth, Belief System, Machine Learning, Big Data

1. Introduction

According to independent investigations, Russian misinformation and fake news by Western conspiracy theorists on social media may have contributed

to the outcome of the Brexit vote[1] and the election of Donald Trump[2]. Misinforming news has become a significant threat to societal discourse and opinion formation. Mechanisms to deal with this type of fake news by making them transparent are urgently needed. The goal of this project is to understand the concept of "fake news" in the context of forming collective awareness through social media. The concept of truth is dependent on a personal belief system. On the other hand, conspiracy theories and satire is nothing new, and people who WANT to believe these have always embraced them. Categorizing news as "Fake news" happens when they are against one's innermost and most passionate beliefs. The more somebody is embedded into a predefined belief system, the more likely they are to believe fake news. For instance, people who use Facebook as their major news source, are more likely to believe fake news (Silverman & Singer-Vine, 2016).
What mental processes are happening when we embrace fake news? When embedded in a particular belief system, individuals recognize fake news immediately when they read them, because they do not want to believe them, similarly they also immediately categorize news as true news when they read them, because they perfectly fit into their belief system. For instance, Trump followers label mainstream news as "fake news", while mainstream news labels news from Trump followers as "fake news".

2. Related Work

There are many approaches to creating more transparency in societal discourses. In fact, this may be seen as the core task of quality journalism. Most if not all of these approaches, however, are not well supported by IT tools, do not scale well, and many do not reveal the applied algorithms. Fact checking Websites such as Wikitribune, Snopes.com, PolitFact, and FactCheck.org, and corporate/proprietary initiatives like Facebooks's fake news detection tools mostly rely on human volunteers and/or paid staff to do fact checking, which has major disadvantages:

- human bias: fact checkers might have a "leftist" or "right-wing" bias
- non-scalable: the human pool of fact checkers is by definition restricted
- deferred access: the machine can check any news item immediately, 24/7, and it does not take the expensive detective work of the human fact checker
- non-replicable: as the fact checking is done by different users, the reader will not be able to understand why a certain fact has been categorized in a particular way

[1] Londongrad - Russian Twitter trolls meddled in the Brexit vote. Did they swing it?. Economist, Nov. 23rd 2017

[2] https://en.wikipedia.org/wiki/Russian_interference_in_the_2016_United_States_elections

Among the automated approaches, Kloutscore (www.klout.com) gives a metric for the social media influence of a person. However the kloutscore has to be requested manually by a user who wants a kloutscore, so it is heavily skewed towards self-promoters. Another solution for finding the social media profiles of users is to leverage the Google Knowledge Graph (https://en.wikipedia.org/wiki/Knowledge_Graph), which has been employed in theoretical work by Ciampaglia et al. (2015) for fact checking by measuring the shortest path distance between related concept nodes. Another approach consists of using machine learning to identify fake news, for instance it has been shown by Ott et al. (2011) that machine learning based on word usage beats humans by wide margins to identify fake reviews in tripadvisor by computing feature vectors from the text of the reviews. More generally, (Youyou et al. 2015) have shown that to identify (tribal) attributes of people, having the computer look at their Facebook likes through machine learning will be more reliable than human judgment. A similar research question is addressed when identifying Twitter bots based on their networking pattern and word usage. For instance, Botcheck (botcheck.me) and Botometer (https://botometer.iuni.iu.edu/#!/) (Varol et al. 2017) check the likelihood of any Twitter id to be a bot, based on number of followers and friends, tweeting dynamics, and content of tweets.

3. Motivation – How Influencers Spread Fake News

Today's online social media consumers are exposed to a cacophony of fact and fiction as never before. "It is true, I read it on the Internet" is unfortunately a prominent way for information to spread. For example, immediately after the 2016 US Presidential elections, in early November 2016, Hillary Clinton was accused of running a pedophile ring out of a pizza restaurant in Washington. Called "pizzagate", this news item became a favorite call to arms among right-wing extremists and Donald Trump supporters, leading one incensed fanatic to drive a few hundred miles from Salisbury, North Carolina to Washington DC, and firing his automatic gun into the pizza restaurant. The origin of this fake news story has been well documented, starting from a white supremacist Twitter account, then picked up by the conspiracy News Web site of Sean Adl-Tabatabai, where it fell on the willing ears of the American right. Just like Google has revolutionized the way we access information, our proposed Transparency Engine intends to change the way how we look at such information, by exposing the hidden influencers like "Sean Adl-Tabatabai" who inject new information into the public discourse.

3.1 The concept of tribes and how they perceive information

Besides knowing the sources of rumors, it is essential to also know the (political) orientation of these influencers. Quantum physics suggests that there are many different universes, with our current world being embedded into just one out of infinitely many other universes. Looking at radically different interpretations of the same news item, it seems we are indeed living in different quantum universes. These different universes can be grouped into "tribes" (Sloterdijk 2011). Each of these tribes has its own reality, defining fact or fiction for the members of the tribe. Previous research (De Oliveira et al. 2017) has exemplified this idea. What is fact for one tribe is fiction for another tribe. It all depends on the tribe, and what the members of the tribe WANT to believe. Examples are the denial of human-influenced global warming, the explanation of evolution through "intelligent design", or the causal relationship between vaccination and autism where some tribes perceives related issues as "fact" and "truth" whereas other tribes perceive the objectively same issues as "fiction", "lie" or "fake news", thus creating an "alternate reality". In contrast to the power of states and corporations, the growing power and dynamics of networks is mostly invisible. Unlike hierarchical structures, the central influencers in networks are hard to identify by the "naked eye". What matters to spread any news – fact or fake – is the influence of the spreader. The main way to quantify the influence of the spreader is her/his position in a given network and with it the power to "multiply" the word to larger audiences. More specifically, the degree and power of the spreaders' position can be measured by re-constructing their (social) network via their Web sites and their social position for example in the Twitter-universe (and other social networking platforms) thus measuring the influence of Web sites and the influence of Twitter (accounts) on a specific topic.

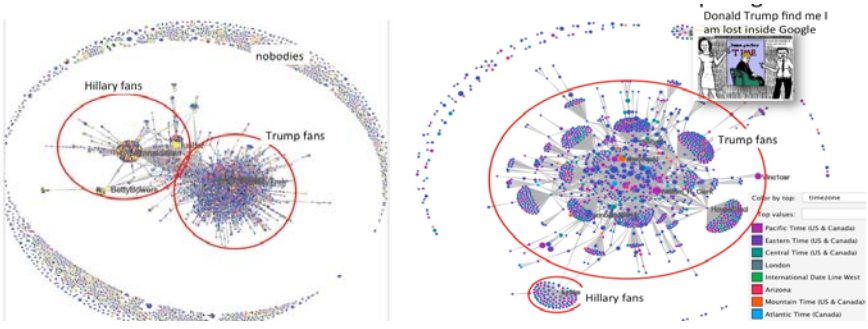


Figure 1 Twitter retweet network "pizzagate" (left), and Twitter influence network (right)

Pizzagate only spread because a moderately influential spreader, Sean Adl-Tabatabai, discovered the original tweet and posted it on his conspiracy News Web site. Figure 1 illustrates how social media analysis can increase trust and transparency by visualizing the echo chambers of fake news about pizzagate using our social media analysis system Condor (Gloor 2017). The picture at left shows the Twitter network about pizzagate, each node is a person tweeting, a link between two people means either that one person is retweeting a tweet sent by the other person, or is mentioning the other person in a tweet. There is a large cluster in the center of the network, made up of believers in the fake news. They are reinforcing each other, and increasing the traffic in their echo chamber. The few supporters of Hillary Clinton, trying to debunk the fake news, are pushed aside; their tweets are ignored by the large echo chamber of conspiracy theory believers. The people in the periphery (the "asteroid belt") are tweeting into the void, as their tweets are ignored by friends and foes alike.

Using an influencer algorithm (Gloor 2017) shows that the discourse about pizzagate on Twitter is dominated by Trump followers (the picture at right above). Our algorithm makes somebody an influencer, if the words she or he is using, are picked up by others and spread quickly through the network. As the picture at right in figure 1 shows, there is just one voice of reason left, while the proponents of pizzagate reinforce each other much more, with a cluster of influential spreaders of wild ideas in the center, and other conspiratorialists in the periphery of the cluster, being retweeted by hundreds of likeminded others (shown as "parachutes" in the graph).

4. Our Solution – Transparency Engine

We introduce the "Transparency Engine", a social network search engine to separate fact from fiction by exposing the hidden influencers and their "tribes" behind fake news. Just like Google has revolutionized the way we access information, Transparency Engine changes the way we look at such information, by exposing the hidden influencers. Our goals are fourfold: (1) Quantify the influence and relevancy of persons, concepts, companies on institutions, issues or industries. (2) Qualify the dynamics and changes in the observed environment. (3) Visualize the networks of influence for a given social or economical ecosystem. (4) Provide a tool to track the diffusion of new ideas, both good and bad.

4.1. Powergraph

Our solution combines three subsystems we have been developing over the last five years (Fuehres et al. 2012, de Oliveira et al. 2016, de Oliveira et al 2017): Power graph, tribe finder, and swarmpulse. Power graph measures the

importance of "notable" people as defined by Wikipedia through calculating the number of other Wikipedia people pages than can be reached within two degrees of separation from a particular people page on Wikipedia. This is a proxy for social capital, as it basically measures the influence of the people a person is connected to. The system also identifies those people with Twitter accounts by matching them with sources of information like Wikidata and Google knowledge graph.

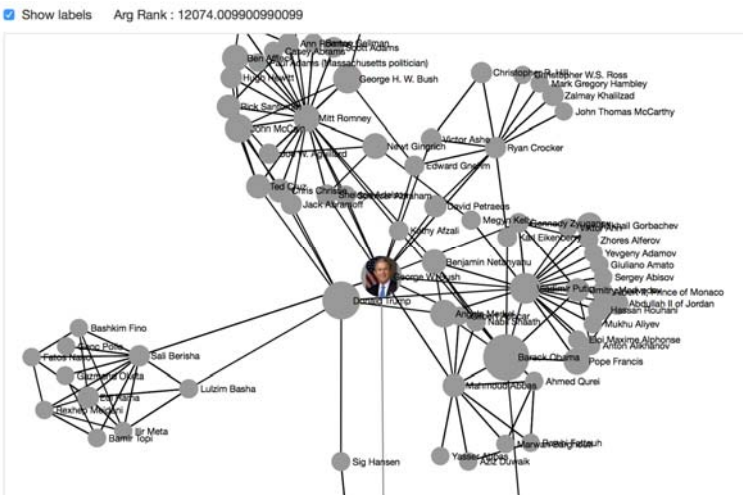


Figure 2. Sample Powergraph for "global warming"

Figure 2 illustrates our prototype version of the Powergraph, showing the social network of the most influential people about "global warming", based on their Wikipedia and Twitter presence. We find, not surprisingly, that Donald Trump and the former US presidents are most influential. We measure the importance of people through calculating the number of other Wikipedia people pages and Twitter friendship networks than can be reached within two degrees of separation from a particular people page. This is a proxy for social capital, as it basically measures the influence of the people a person is connected to (Fuehres et al. 2012).

4.2 Tribefinder

The second component of our system, tribefinder (de Oliveira et al. 2017), identifies the tribal affiliations of the opinion leaders about any news item. To assign a tribe to an influencer, our system analyzes their word usage, using deep learning. An integral component of the tribefinder system is "TribeCreator", this subsystem automatically helps the user to find people that belong to a newly defined tribe by looking at profile self-descriptions,

the content of tweets, and at followers, and Twitter friends. For example, if users wants to create a tribe for Treehuggers (people who like nature), they can search for people with profile descriptions that match the idea of this tribe: "nature lover", "I love nature", "nature", etc., for people who follow pages about nature, or tweet about nature. In the second step we calculate the vocabulary that these influentials are using in their tweets. This vocabulary is then used to match the vocabulary against the vocabulary of any Twitter user, calculating their tribal affiliates. Knowing the tribal affiliations of the thoughtleaders for a news item allows readers to correctly position the news item, deciding for themselves if they want to trust the news coming from a particular influencer.

4.3 Swarmpulse

The third component of our system is Swarmpulse (de Oliveira et al 2016). Swarmpulse finds the most recently edited Wikipedia pages and uses Twitter to see which people are talking about those subjects. This system helps users to serendipitously spot most recent news items they were not aware of, and then check their influencer network on the power graph and calculate their tribal affiliations with tribefinder.

5. Conclusion

The best approach for fact-checking is a critical, well-informed mind. Our world needs more powerful ways and tools to support the critical mind. Transparency is a key enabler for this. The Transparency Engine thus provides the foundation for informing the critical mind: The global Powergraph will display the power network of the one million globally most influential people on Wikipedia people pages and the most popular Twitter users. It will allow all other Twitter users to position themselves within the context of the Powergraph. The Tribefinder will show the "truth of tribes" by creating tribes through their use of language on social media and assigning each influencer to one or more tribes and showing the tribal affiliations in the Powergraph. Swarmpulse will build an index of most recent significant news by combining new edits on Wikipedia with the most popular tweets from influential twitterers and show the actors involved through Powergraph. The landscape of transparency generating approaches calls for a scientific, open approach such as the Transparency Engine proposes. Our aim is to substantially contribute to popularizing and democratizing fact checking for the whole world. Everyone should be enabled to do this easily and simply by themselves!

References

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. PloS one, 10(6), e0128193.

de Oliveira, J. Gloor, P. (2016) The Citizen IS the Journalist - Automatically Extracting News from the Swarm. Rome, Italy June 9-11, 2016, Designing Networks for Innovation and Improvisation: Proceedings of the 6th International COINs Conference (Springer Proceedings in Complexity)

de Oliveira, J. Gloor, P. (2017) GalaxyScope – Finding the "Truth of Tribes" on Social Media. Detroit September 11-14, 2017. Proceedings of the 7th International COINs Conference (Springer Proceedings in Complexity)

Fuehres, H. Gloor, P. Henninger, M. Kleeb, R. Nemoto, K. (2012) Galaxysearch: Discovering the Knowledge of Many by Using Wikipedia as a Meta-Search Index. Proceedings Collective Intelligence 2012, April 18-20, Cambridge, MA

Gloor, P. (2017) Sociometrics and Human Relationships: Analyzing Social Networks to Manage Brands, Predict Trends, and Improve Organizational Performance , Emerald Publishing, London 2017

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 309-319).

Silverman, C. Singer-Vine, J. (2016) "Most Americans who see fake news believe it, new survey says." BuzzFeed News; https://www.buzzfeed.com/craigsilverman/fake-news-survey

Sloterdijk, P. (2011). Bubbles: microspherology. MIT Press

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. arXiv preprint arXiv:1703.03107.

Youyou, W. Kosinski, M. Stillwell, D. (2015) Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences (PNAS)

Brexit and Twitter: The voice of people

Francesca Greco, Leonardo Alaimo, Livia Celardo
Sapienza University of Rome – francesca.greco@uniroma1.it;
leonardo.alaimo@uniroma1.it; livia.celardo@uniroma1.it

Abstract 1
There is an increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European states. Regarding the European Union membership, public opinion is divided in two. British referendum in 2016, where citizens chose to "exit" shaking the public opinion, and the following general election in June 2017, where the British Europeanist parties won the election according to the 1975 British referendum where 72% of citizens chose to "Remain", are clear examples of this fracture. There are still few studies concerning the investigation of Brexit discourses within the social media and most of them focus on the 2016 British referendum. Due to that, this exploratory research aims to identify how Brexit and the EU are nowadays discussed on Twitter, through a text mining approach. We collected all the tweets containing the terms "Brexit" and "EU", for a period of 10 days. Data collection has been performed with TwitteR package, resulting in a large corpus to which we applied multivariate techniques in order to identify the contents and the sentiments behind the shared comments.

Abstract 2
Negli ultimi anni c'è stato un aumento dell'euroscetticismo tra i cittadini dell'UE, come testimoniato dallo sviluppo di partiti ultra nazionalisti in diversi stati europei. Sul tema "Europa", l'opinione pubblica è divisa fra europeisti e euroscettici. Un chiaro esempio di questa divisione è dato dalle recenti vicende britanniche: infatti, nel referendum del 2016 i cittadini britannici hanno scelto di "uscire" dall'UE scuotendo l'opinione pubblica, mentre le successive elezioni politiche di giugno 2017 hanno visto l'affermazione dei principali partiti filo-europeisti. Vi sono ancora pochi studi in letteratura che indagano come nei social media venga affrontato il tema della Brexit in relazione all'UE, dato che la maggior parte di essi si focalizza su cause e potenziali effetti del voto di giugno 2016. In tal senso, questa ricerca esplorativa ha lo scopo di identificare in che modo Brexit e l'Unione Europea vengano discusse su Twitter in questo momento storico attraverso l'analisi automatica del testo. A questo scopo sono stati raccolti tutti i messaggi contenenti i termini "Brexit" e "EU" per 10 giorni attraverso

l'utilizzo del pacchetto TwitteR, ottenendo un corpus di grandi dimensioni a cui sono state applicate delle tecniche multivariate, al fine di individuare i contenuti e i sentimenti relativi al tema in esame.

Keywords: Brexit, Twitter, Emotional text mining.

1. Introduction

There is a growing increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European states. Regarding to European Union membership, public opinion is divided between Eurosceptics and pro-Europeans, as shown by the 2016 British referendum ("Brexit"), where 52% of citizens chose to "Leave". For further evidence of this division, the following general election of June 2017 saw the affirmation of the main Europeanist parties (especially the Labour Party) and the results led to a *hung Parliament*. Brexit has shaken the European public opinion as it revealed the relevance of the anti-Europeanist trend. During the 60th Anniversary of the Treaties of Rome in 2017, millions of citizens expressed their support to the EU participating to Europeanist demonstrations in many European cities.

One useful starting point for explaining the results of Brexit is to focus on the electoral issue: the relationship between the UK and Europe. This has always been a central and rather controversial issue in the British public debate. The media, public opinion and the political class have always been deeply critical and sceptical about the European integration. This position influences citizens' attitudes towards the Union, which is not only considered distant and inadequate to resolve everyday issues (immigration, unemployment, and so on), but it is often perceived as their major cause, by limiting the political and economic power of United Kingdom. The electoral outcome created disbelief all over the world. *Britain is the home of the term Euroscepticism* (Spiering 2004, p.127). But, while it is clear that a large proportion of UK residents are sceptical about Europe, it is not clear enough that this position coincides with the wish to leave the EU. However, Euroscepticism should not be confused with this wish. Szczerbiak and Taggart (2008) have distinguished two different types of Euroscepticism: the *Hard Euroscepticism* that is *a principled opposition to the EU and European integration* and *Soft Euroscepticism* that *concerns on one (or a number) of policy areas lead to the expression of qualified opposition to the EU*.

Although there are several studies exploring British Euroscepticism, only few of them investigate the Brexit discourses within the social media. Due to that, we decided to perform a quantitative study, where the online discourses regarding Brexit and EU are analysed using two different approaches,

Content Analysis and Emotional Text Mining. The aim is to explore not only the contents but also the sentiments shared by users on Twitter. For this paper, we used one of the most important and known blog tools, Twitter. It is an online platform for sharing real-time, character limited communication with people partaking of similar interests that, in 2017, counted over than 300 million users and an average of about 500 million of tweets sent per day.

2. Data collection and analysis

In order to explore the sentiments and the contents on Brexit and EU in twitter communications during ten days, we scraped all the messengers in English language produced from September 22nd to October 2nd, 2017, containing together the words *Brexit* and *EU*. The data extraction was carried out with the TwitteR package of R Statistics (Gentry, 2016). We started collecting 221,069 messengers, including 83% of retweets, from which two samples of tweets were extracted. The first we used for the sentiment analysis is composed of 99,812 messengers, where the retweets were limited to the threshold of 31, resulting in a large corpus of 1,601,985 of tokens; the second one we used for content analysis, where we excluded all the retweets, resulted in a large corpus of 37,318 tweets and 618,255 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the hapax percentage ($TTR_{corpus\ 1}$ = 0.02; $Hapax_{corpus\ 1}$ = 39.8%; $TTR_{corpus\ 2}$ = 0.04; $Hapax_{corpus\ 2}$ = 52.31%). According to the large size of the corpus, both lexical indicators highlighted its richness and indicated the possibility to proceed with the analysis.

2.1. Emotional text mining

We know that people sentiments depend not only on their rational thinking but also, and sometimes most of all, on the emotional and social way of functioning of people's mind. If the conscious process set the manifest content of the narration, that is what is narrated, the unconscious process can be inferred through how it is narrated, that is, the words chosen to narrate and their association within the text. According to this, it is possible to detect the associative links between the words to infer the symbolic matrix determining the coexistence of these terms in the text (Greco, 2016). To this aim we perform a multivariate analysis based on a bisecting *k*-means algorithm to classify the text (Savaresi et Boley, 2004), and a correspondence analysis to detect the latent dimensions setting the cluster per keywords matrix (Lebart et Salem, 1994) by means of T-Lab software. The interpretation of the cluster analysis results allows to identify the elements characterizing the emotional representation of Brexit, while the results of correspondence

analysis reflect its emotional symbolization. By the clusters interpretation, we classify the emotional representations in positive, neutral and negative sentiments, determining the percentage of messages for each sentiment modality. To this aim, first corpus was cleaned and pre-processed with the software T-Lab (T-Lab Plus version, 2017) and keywords selected. In particular, we used lemmas as keywords instead of types, filtering out the lemma *Brexit* and *EU* and those of the low rank of frequency (Greco, 2016). Then, on the tweets per keywords matrix, we performed a cluster analysis with a bisecting *k*-means algorithm limited to twenty partitions, excluding all the tweets that do not have at least two keywords co-occurrence. The percentage of explained variance (η) was used to evaluate and choose the optimal partition. To finalize the analysis, a correspondence analysis on the keywords per clusters matrix was made in order to explore the relationship between clusters and to identify the emotional categories setting Brexit representations.

2.2. Content analysis

Content analysis is a technique used to investigate the content of a text; in text mining, many methods exist to analyse it automatically. One of these is Text Clustering, where the corpus is splits in different subgroups based on words/documents similarities (Iezzi, 2012). In this paper, a text co-clustering approach (Celardo et al., 2016) is used. The objective is to simultaneously classify rows and columns, in order to identify groups of texts characterized by specific contents. To do that, data were pre-processed with Iramuteq software lemmatizing the texts, removing stop words and terms with frequency lower than 10. The weighted term-document matrix was then co-clustered through the double *k*-means algorithm (Vichi, 2001); the number of clusters for both rows and columns was fixed using the Calinski-Harabasz index.

3. Emotional text mining main results and discussion

The results of the cluster analysis for ETM show that the 655 keywords selected allow the classification of 88,6% of the tweets. The percentage of explained variance was calculated on partitions from 3 to 19, and it shows that the optimal solution is six clusters (η= 0.057). The correspondence analysis detected six latent dimensions. In table 1, we can appreciate the emotional map of Brexit and the EU emerging from the English tweets. It shows how the clusters are placed in the factorial space produced by five factors. The first factor represents the political and economic domain where Brexit seems to have its main impact; the second factor reproduces the possible solutions of Brexit: a separation or a new agreement; the third factor

represents the national or European level of reaction to Brexit; the fourth factor is the blame, distinguishing the blame of politicians from the one of the willingness to be independent; and the fifth factor is the political leadership, differing old and new policies.

Table 1 – Correspondence analysis results (the percentage of explained inertia is reported between brackets beside each factor).

Factor 1 (27.5%)		Factor 2 (24.3%)		Factor 3 (19.8%)		Factor 4 (15.6%)		Factor 5 (12.9%)	
NP	PP	NP	PP	NP	PP	NP	PP	NP	PP
future	negotiation	bill	try	Macron	Florence	blame	referendum	leader	people
war	Briton	Barnier	pro	European	Delay	march	Johnson	remain	Tory
support	chance	Brussel	deliver	good	withdrawal	stay	Verhofstadt	walk	hard
remainer	zero	divorce	brexiteers	miracle	blast	speech	independent	urge	voter
concern	better off	progress	help	market	states	conservative	destroy	May T.	party
save	laureate	negotiator	debate	union	finger	anti	migrant	hope	happen
proposal	leaving	pay	allow	single	finish	Blair	vow	call	Catalonia
fight	economist	chief	event	Merkel	row	reverse	adopt	time	die
0.07-0.02 ac	6.49-4.40 ac	4.72-1.50 ac	0.35-0.12 ac	1.5-1.61 ac	0.35-0.05 ac	0.55-0.29 ac	5.22-0.94 ac	3.65-1.24 ac	10.28-1.49 ac

NP =negative pole; PP = positive pole; ac = absolute contribution (10^{-3})

The six clusters are of different sizes and reflect the representations of Brexit (table 2), that correspond to three different sentiments: positive, negative for domestic reasons, and negative for foreign ones (table 1). The first cluster represents the choice to leave EU as a good option, underlining the need to proceed; the second cluster focuses on the EU political reaction fixing divorce conditions, perceiving EU political representatives as unfavourable and therefore threatening; the third cluster represents Britons' hope to improve their economic condition leaving EU as naive; the fourth cluster represents the old British political leadership as incompetent, being unable to protect and adequately inform Britons in order to support them in remaining in the EU; the fifth cluster reflects the negotiation of the divorce conditions, perceiving the negotiation as unfair and the costs of leaving EU as a punishment; and the sixth cluster represents Brexit as a Britons informed choice, highlighting that its consequences belong to the policy domain who should respect the citizens' choice.

Table 2 – Clusters (the percentage of context units classified in the cluster is reported between brackets).

Cluster 1 (10.0% CU)	Cluster 2 (14.9% CU)	Cluster 3 (20.9% CU)	Cluster 4 (13.4% CU)	Cluster 5 (19.2% CU)	Cluster 6 (21.7% CU)
Good Choice	*EU Reaction*	*Uncertain Future*	*British Leadership*	*Divorce Conditions*	*Informed Choice*
leader	Macron	negotiation	people	bill	referendum
remain	European	leaving	Tory	Barnier	Corbyn
time	good	Briton	hard	brussel	Johnson
Theresa May	market	chance	voter	progress	think
urge	warn	zero	party	divorce	independent
call	single	better off	happen	negotiator	Boris
walk	business	Nobel	Florence	pay	Verhofstadt
UKIP	minister	economist	stay	chief	Florence
government	Europe	laureate	Catalonia	demand	destroy
hope	move	tell	believe	national	try
look	Merkel	rating	Spain	Davis	policy
mean	miracle	law	Rees Mogg	offer	issue
From 1611 to 620 CU	From 2004 to 951	From 1844 to 668 CU	From 2506 to 461 CU	From 2705 to 843 CU	From 2098 to 512 CU

CU = context units classified in the cluster.

By the clusters interpretation, we detected six different representations of Brexit that correspond to three different sentiments (table 1). We have considered as positive (21,7%) the representation of Brexit as a Good Choice or an Informed Choice, and negatives all the other representations (78,3%). Among the negative clusters, we distinguished negativity according to the origin of the problem: Uncertain Future and British Leadership are negative for domestic reasons (34,2%), that is, the lack of UK political leadership's competences; and EU Reaction and Divorce Condition are negatives due to foreign factors (34,1%) as the EU after Brexit seems to be perceived as vindictive and, therefore, threatening.

4. Content analysis main results and discussion

The pre-processing phase, implemented on the second corpus, allowed us to identify a set of 1.957 keywords, representing the 97% of the tweets; so, on the term-document matrix of dimension (1.957 × 36.383) we calculated the Calinski-Harabasz Index in order to define the number of clusters for rows and columns. After calculating the index values for partitions from 2 to 10 for each dimension, the Calinski-Harabasz Index suggested to classify the words in three groups and the tweets in five groups. In table 3, the centroids of the clusters are exposed.

Table 3 – Centroids matrix (Terms × Documents).

	Cluster 1 (55%)	*Cluster 2* (20%)	*Cluster 3* (12%)	*Cluster 4* (11%)	*Cluster 5* (2%)
Cluster 1	**0,005**	0,003	0,004	0,000	0,000
Cluster 2	0,002	**0,063**	0,003	**0,149**	0,012
Cluster 3	-0,002	0,000	**0,090**	-0,003	**0,309**

Table 4 – Words groups (first 10 words listed below by frequency of occurrence).

Cluster 1 *Negotiation*	*Cluster 2* *Economic Transformation*	*Cluster 3* *British Identity*
stay	leave	home
Junker	move	sound
ambassador	transition	cake
cry	late	plan
track	deal	datum
surge	trade	live
peer	retain	finish
shape	post	Id
turmoil	Macron	idea
survive	urge	national

As shown in the table 3, the algorithm has identified five blocks of specificities; in fact; the first cluster of words is connected to the first group of tweets; the second is specific of the second and the fourth cluster of tweets and the third is related to the third and the fifth group of tweets. In table 4, the groups of words are presented.

The first group of words is related to the need of defining new rules and settlements within the negotiation and it represents more than half of the tweets; it has no strong specificities related to the texts, but in comparison to all the documents clusters, it seems to be more connected to those words. On the other hand, for the other two groups of words, there are more effective specificities; the second cluster of words is about the definition of new economic agreements, and it is connected to the 31% of the tweets, while the third one, related to the requirement in specifying a new identity after Brexit, is representative of the 14% of the corpus documents.

5. Conclusions

The results of the two analyses showed a strong relationship between the terms "Brexit" and "EU", not only in terms of sentiment, but also in terms of

contents. According to the literature, the sentiment analysis revealed the presence of both positive and negative opinions in respect to the exit of United Kingdom from the EU. On the other hand, starting from the analysis of the contents we found that the Twitter communications on Brexit focuses primarily on the concept of *negotiation*. The remaining part of the messages take into account both the Brexit economic features and the need of the national identity redefinition. To conclude, the results of the two analyses revealed that Brexit is a theme with a strong emotional charge, mostly negative. British people seem to focus their attention basically toward three issues: the new asset, the economic consequences, and the national identity. These subjects are treated positively and negatively from the users, probably because of the lack of cohesion within the country.

References

Celardo L., Iezzi D.F. and Vichi M. (2016). Multi-mode partitioning for text clustering to reduce dimensionality and noises. In *Proceedings of the 13th International Conference on Statistical Analysis of Textual Data*.

Gentry J. (2016). R Based Twitter Client. R package version 1.1.9.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Hobolt S. (2016). The Brexit vote: a divided nation, a divided continent. *Journal of European public policy*, 23(9): 1259–1277.

Iezzi D. F. (2012). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*(16-17), 3179-3197.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

Savaresi S. M. and Boley D. L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.

Spiering M. (2004). British Euroscepticism. In Harmsen R. and Spiering M., editors, *Euroscepticism: Party Politics, National Identity and European Integration*. Editions Rodopi B.V.

Szczerbiak A. and Taggart P. (2008). *Opposing Europe? The Comparative Party Politics of Euroscepticism. Volume 1: Case Studies and Country Surveys*. Oxford University Press.

Vichi M. (2001). Double k-means clustering for simultaneous classification of objects and variables. *Advances in classification and data analysis*, 43-52.

A text mining on clinical transcripts of good and poor outcome psychotherapies

Francesca Greco[1], Giulio de Felice[2], Omar Gelo[3]
[1]Sapienza University of Rome & Prisma S.r.l. – francesca.greco@uniroma1.it
[2] Sapienza University of Rome & NCU University – giulio.defelice@uniroma1.it
[3] University of Salento & Sigmund Freud University – omar.gelo@unisalento.it

Abstract
The text mining of clinical transcripts is broadly used in psychotherapy research, but is limited to top-down approaches, with a-priori vocabularies that code the transcripts according to a theoretical predetermined framework. Nevertheless, the semantic level that a word or clinical intervention can assume depends on the relational field in which the discourse is produced. Thus, bottom-up approaches seem to be particularly meaningful in addressing such a relevant issue. With the aim of investigating possible similarities and differences between good outcome and poor outcome psychotherapies, we applied a multivariate analysis on the transcripts of eight single cases of brief experiential psychotherapy (four good outcome vs four poor outcome cases), in order to identify the general core themes, and their difference according to therapy outcome. The results showed a significant difference in the number of context units classified in two of the six core themes (clusters) between good and poor outcome cases ($\chi 2$, df=5, $p<0,01$). These findings show how the bottom-up technique of text analysis on clinical transcripts turned out to be an enlightening tool to let their latent dimensions emerge, setting the clinical process and outcome, and therefore, providing a very useful tool for clinical purposes.

Abstract
L'analisi delle trascrizioni cliniche è stata ampiamente utilizzata nella ricerca in psicoterapia, sebbene prevalentemente si basi sull'utilizzo di un dizionario che consente la codifica del testo in funzione di criteri predeterminati. Tuttavia, la polisemia che una parola, o un intervento clinico," può assumere dipende dal campo relazionale in cui il discorso è prodotto. Pertanto, gli approcci bottom-up sembrano essere particolarmente utili nell'affrontare tale questione. Allo scopo di indagare gli elementi caratterizzati le trascrizioni cliniche con esito positivo e negativo, è stata effettuata un'analisi multivariata di un corpus composto da otto trascrizioni di psicoterapia breve (quattro con esito positivo e quattro con esito negativo) al fine di identificare i temi centrali generali e la distribuzione delle unità di contesto nei diversi temi in

funzione dell'esito della terapia. I risultati hanno evidenziato una differenza significativa tra i casi con esito positivo e quelli con esito sfavorevole (χ^2, df = 5, p <0,01), mettendo in evidenza come l'analisi automatica del testo delle trascrizioni dei colloqui clinici possa essere uno strumento utile a far emergere le dimensioni latenti organizzatrici del processo e del risultato, configurandosi così come un utile strumento ai fini clinici.

Keywords: Emotional Text Mining, clinical transcripts, psychotherapy outcome.

1. Introduction

The text mining of clinical transcripts is very broadly used in psychotherapy research, but is limited to top-down approaches where *a-priori* vocabularies code them according to a theoretical predetermined framework. Nevertheless, the semantic level that a word, or clinical intervention, can assume, depends on the relational field in which the discourse is produced. Thus, bottom-up approaches seem to be particularly meaningful in addressing such relevant issue. Psychotherapy can be considered a dynamic communicative exchange between the client and the therapist (e.g., Gelo et Salvatore, 2016). Within such an exchange, the content (i.e., the semantic) of what is said plays a primary role. Thus, the textual analysis of therapy transcripts may represent a very useful tool for psychotherapy process researchers as well as for clinicians (Gelo et al., 2013; Salvatore et al. 2017). In the field of psychotherapy research, some methods of text mining have been developed and applied, such as the Therapeutic Cycle Model (Mergenthaler, 2008) and Referential Activity (Bucci et al., 1992). Following a *top-down* approach, these methods use predefined content categories to semantically classify units of text. Each of these categories corresponds to a thematic dictionary containing all the words indicative of the content represented by that category. Even though these top-down methods of text mining allow for a reliable and valid investigation of the therapeutic process, they present a major limitation, disregarding the contextual nature of the linguistic meaning (Carli et al., 2004; Salvatore et al., 2012). In fact, the meaning of a word is polysemic and depends on the way it combines with other words in the communicative interaction, i.e., it depends on its association with other words. Grounded on these considerations, there has recently been a development of text mining approaches which, by means of their bottom-up logic, allow for a context-sensitive textual analysis (e.g., Salvatore et al., 2012; 2017; Cordella et al., 2014; Greco, 2016). The aim of this study is to investigate possible similarities and differences between good outcome and poor outcome psychotherapy cases by applying the Emotional Text Mining (Cordella et al., 2014; Greco, 2016). Our assumption is that it is possible to

detect the associative links between the words in order to infer the symbolic matrix determining the coexistence of the terms in the text. To this aim, we perform a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The interpretation of the cluster analysis allows for the identification of the elements characterizing the core themes of the treatment, while the results of the correspondence analysis reflect the emotional symbolisation characterising the therapeutic exchange. The advantage of such an approach is to interpret the factorial space according to word polarization, thus identifying the emotional categories that generate the core themes, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space (Greco et al., 2017).

2. Data collection and analysis

2.1. Data collection

The sample of the present study was drawn from the York Depression Study I, a randomized clinical trial to assess the efficacy of brief experiential therapy for depression (Greenberg et Watson, 1998; Watson et al., 1998).[1] From the original sample, we initially selected the six best outcome cases and the six cases worst outcome cases based on the Reliable Change Index of the Beck Depression Inventory (BDI; Beck et al., 1988). We then excluded four cases due to missing session transcripts. Our final sample was thus comprised of a total of eight cases, with four good outcomes and four poor outcomes. The treatment length was between 15 and 20 sessions (M = 17.62; SD = 1.38), for a total of 141 sessions. Patients (one man and seven women; M=37.1 years old) met the criteria for major depressive disorder assessed by means of the Structured Clinical Interview for DSM-III-R (SCID; Spitzer et al., 1989). Therapists (seven women and one man; M= 5.5 years of therapeutic experience) had six months of training in experiential psychotherapy (Greenberg et al., 1993). The transcripts were collected in a large size corpus of 1090234 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the percentage of hapax (TTR = 0.01; hapax = 35.3%). They highlighted the richness of the corpus indicating the possibility of proceeding with the analysis.

[1] We are grateful to Dr. Les Greenberg for having us provided with files of the transcripts for these cases.

2.2. Data analysis

First, data were cleaned and pre-processed with the software T-Lab and keywords selected. In particular, we used lemmas as keywords instead of type. We selected all the lemmas in the medium rank of frequency (upper frequency threshold = 933), and those of the low rank of frequency until the threshold of 17 occurrences, that is, equal to the average number of sessions made on average by the patients (Greco, 2016). Then, in order to identify the core themes common to all the psychotherapies, we performed a cluster analysis on the keywords per context units (CU) matrix, by means of a bisecting *k*-means algorithm (Savaresi et Boley, 2004), limited to ten partitions, excluding all the CU that did not have at least two keywords co-occurrences. The eta squared value was used to evaluate and choose the optimal solution. To finalize the text mining, we performed a correspondence analysis on the keywords per clusters matrix (Lebart et Salem, 1994) in order to explore the relationship between clusters, and to identify the emotional categories setting the psychotherapeutic process. The interpretation of the factorial space was performed according to the procedure proposed by Cordella and colleagues (2014) in which each keyword is considered only in the factor with the greatest absolute value. To finalise the analysis, we performed a chi squared test on the contingency table cluster per therapy outcome, calculating the standard residual in order to identify the differences between good outcome and poor outcome clinical transcripts in terms of core themes.

3. Main results and discussion

The results of the cluster analysis show that the 1351 keywords selected allow for the classification of 56.6% of context units. The high proportion of unclassified context units is due to the transcripts richness in terms of para-verbal interactions (i.e. *mhm*, *yeah*, etc). The eta squared value was calculated on partitions from 3 to 9, and it showed six clusters as the optimal solution ($\eta^2 = 0.034$). In table 1, we can appreciate the emotional map emerging from the clinical transcripts representing the clusters location in the factorial space produced by the interpretation of the five factors. The first factor reflects patient *positioning*, which can be passive or active; the second factor refers to the *relationship* that could be familiar or unfamiliar, i.e., a person facing something new and unpredictable; the third factor represents the *communication content* that can be emotional or concrete; the fourth factor reflects the *outcome* of the therapeutic work, that is, the patient's empowerment or making sense of the patient's experiences; and the fifth factor distinguishes the *issues* within the daily ones, concerning everyday life,

from the relational ones, concerning the loved ones.[2]

Table 1 – Factorial space representation (the percentage of explained inertia is reported between brackets under each factor).

Cluster	Label (CU%)	Factor 1 (26.7%) Positioning	Factor 2 (25.8%) Relationship	Factor 3 (21.5%) Content	Factor 4 (14.5%) Outcome	Factor 5 (11.5%) Issues
1	Family Structure	Passive	Familiar	Emotional		Daily
	(11.6%)	0.20	-0.56	-0.16	-0.01	-0.32
2	Transformative Process	Active	Unfamiliar		To empower	Daily
	(12.1%)	-0.46	0.29	0.06	-0.35	-0.16
3	Concrete thinking	Passive	Unfamiliar	Concrete	To empower	
	(16.1%)	0.84	0.34	0.42	-0.19	0.05
4	Therapeutic Relationship	Active	Familiar	Concrete	To understand	Relational
	(22.4%)	-0.25	-0.18	0.41	0.28	0.16
5	Relational Issues		Familiar	Emotional	To empower	Relational
	(14.6%)	0.04	-0.14	-0.47	-0.18	0.45
6	Feelings		Unfamiliar	Emotional	To understand	Daily
	(23.1%)	0.06	0.58	-0.43	0.49	-0.14

CU = context units classified in the cluster.

Table 2 – Psychotherapy core themes.

Cluster 1 Family Structure		Cluster 2 Transformative Process		Cluster 3 Concrete Thinking		Cluster 4 Therapeutic Relationship		Cluster 5 Relational Issues		Cluster 6 Feelings	
keyword	CU	keyword	CU	keyword	CU	keyword	CU	keyword	CU	keyword	CU
home	525	start	507	hear	455	week	699	mother	399	understand	416
kid	371	able to	504	money	326	sense	675	life	335	hurt	300
house	290	change	438	dollar	267	day	438	problem	333	important	298
father	241	different	396	accept	205	bad	432	hard	292	person	231
husband	213	situation	288	pay	196	angry	381	care	268	hard	213
child	205	point	237	listen	175	call	253	deal	252	support	185
parent	194	go on	216	believe	135	night	189	family	237	inside	170
stay	190	mind	213	matter	130	morning	169	relationship	233	strong	168
live	179	trying	183	sell	126	set	162	Father	195	pain	153

CU = context units classified in the cluster.

The six clusters are of different sizes (table 1) and reflect the core themes of the brief psychotherapy (table 2). The first cluster describes the *family structure* with its role and places; the second cluster reflects the *transformative*

[2] In the negative pole of the fifth factor (Daily Issues) we find the following words: *house, stay, TV, rule, street, teacher, move out, neighbour, pounds,* and in the positive pole we find words as *mother, life, problem, sister, relationship.*

process characterising a psychotherapy; the third cluster highlights the *concrete thinking* process, a way to think that could be defined as concrete thinking, which is often rational and frequently concerning economic issues; the fourth cluster represents the *therapeutic relationship* that is made of concrete limits, and the process of making sense of personal experiences; the fifth cluster reflects the *relational issues* of the patient's private life; and the sixth cluster refers to the process of detecting, recognizing, and understanding *feelings,* characterizing internal emotional experiences.

There is a significant difference in the number of content units classified in each cluster among the good and poor outcome therapies (χ^2, df = 5, $p < 0.01$). In particular, the differences lay on the relevance of two of the six core themes: the *concrete thinking* and the *feelings*. While the good outcome brief psychotherapies are characterized by a high number of context units classified in the cluster *feelings* (SE = 6.8) and a low number of context units classified in the cluster *concrete thinking* (SE = -5.8); the poor outcomes psychotherapies are characterized by a high number of context units classified in the cluster *concrete thinking* (SE = 6.8) and a low number of context units classified in the cluster *feelings* (SE = -7.0). Namely, it would seem that patients tend to dwell upon their emotional experiences in the good outcome psychotherapy, while they tend to dwell upon facts in the poor outcome psychotherapy, probably not connecting them to their emotional experiences. Given that we classified the interactions among the patients and the therapists in this analysis, the therapy outcome could derive both from the patient's ability in dealing with feelings or the therapist's ability to support the patient in doing so.

The above-mentioned differences between good and poor outcome cases are coherent with findings obtained on the same sample by means of a principal component analysis made on the transcripts coded according to three dictionaries: abstract language, emotional positive language, and emotional negative language (de Felice et al., 2018). In this study, differences in the correlation matrices between good outcome and poor outcome cases were evident. The most obvious one concerned the dynamic in which the patient made use of abstract/concrete language, interpreted very positively in poor outcome cases and very negatively in good outcome cases. In the latter, it was probably and correctly considered as a patient's defense mechanism to address. This was confirmed by the use of positive and negative emotional language, inversely proportional to abstraction, only in poor outcome cases.

4. Conclusion

Talking about concrete events without any sort of emotional involvement in the clinical literature is a defence mechanism that goes under the name of

rationalisation, and it represents a way to protect the mind from painful feelings using an abstract, intellectual and often concrete attitude in dealing with them. While the good outcome psychotherapeutic relationships seem to be capable of addressing the emotional content laying under the surface of the psychotherapeutic field (i.e. use of the therapist's negative emotional language), the poor outcome dynamics seem to be completely wrapped up in a process of avoiding it. Both the PCA (de Felice et al 2018) and text analysis on clinical transcripts confirmed the difficulty in poor outcome psychotherapies to work on the patient's emotional aspects. This bottom-up technique of text analysis on clinical transcripts turned out to be an enlightening tool to let their latent dimensions emerge, arranging the clinical process and outcome, therefore, providing a very useful tool for clinical purposes.

References

Beck A.T., Steer R.A. and Garbin M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8: 77 100.

Bucci W., Kabasakalian-McKay R. and RA Research Group (1992). Scoring referential activity. Ulm, Germany: Ulmer Textbank.

Carli R., Dolcetti F. and Dolcetti (2004). L'Analisi Emozionale del Testo (AET): un caso di verifica nella formazione professionale. In Purnelle G., Fairon C. and Dister A., editors, *Actes JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelle*s, pp. 250-261.

Cordella B., Greco F. and Raso A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In Nee E., Daube M., Valette M. and Fleury S., editors, *Actes JADT 2014 (12es Journées internationales d'Analyse Statistque des Données Textuelles, Paris, France, Juin 3-6, 2014),* pp. 173-184.

de Felice G., Orsucci F., Mergenthaler E., Gelo O., Paoloni G., Scozzari A., Serafini G., Andreassi S., Vegni N. and Giuliani A. (2018). What differentiates good and poor outcome psychotherapies? A statistical mechanics approach to psychotherapy research. *Nonlinear Dynamics, Psychology and Life Sciences*. Submitted.

Gelo O.C.G. and Salvatore S. (2016). A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes. *Journal of Counseling Psychology, 63*(4): 379-395.

Gelo O.C.G., Salcuni S. and Colli A. (2013). Text analysis within quantitative and qualitative psychotherapy process research: introduction to special issue. *Res. Psychother. Psychopathol. Process Outcome* 15: 45–53.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *Rivista Italiana di Economia Demografia e Statistica*, 71(2): 125:36.

Greenberg L., Rice L. and Elliott R. (1993). *Facilitating emotional change. The moment by moment process*. Guilford Press.

Greenberg LS, Watson JC (1998). Experiential therapy of depression: differential effects of client-centered relationship conditions and process experiential interventions. *Psychotherapy-Research* 8: 210–224.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

Mergenthaler E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, 18(2): 109-126.

Salvatore S., Gelo O., Gennaro A., Metrangolo R., Terrone G., Pace V., Venuleo C., Venezia A. and Ciavolino E. (2017). An automated method of content analysis for psychotherapy research: A further validation. *Psychotherapy Research*, 27(*1*): 38-50.

Salvatore S., Gennaro A., Auletta A.F., Tonti M. and Nitti M. (2012). Automated method of content analysis: A device for psychotherapy process research. *Psychotherapy Research*, 22(*3*): 256-273.

Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.

Spitzer R., Williams J., Gibbons M. and Firs M. (1989). *Structured Clinical Interview for DSM-III-R*. American Psychiatric Association

Watson J.C., Greenberg L. S. and Lietaer G. (1998). The experiential paradigm unfolding: Relationship & experiencing in therapy. In Greenberg L.S., Watson J.C. and Lietaer G., editors, *Handbook of experiential psychotherapy*, Guilford Press.

DOMINIO: A Modular and Scalable Tool for the Open Source Intelligence

Francesca Greco[1], Dario Maschietti[2], Alessandro Polli[3]
[1] La Sapienza University of Rome, Prisma S.r.l. – francesca.greco@uniroma1.it
[2] Prisma S.r.l – d.maschietti@prismaprogetti.it
[3] La Sapienza University of Roma – alessandro.polli@uniroma1.it

Abstract
Prisma has developed an innovative technology for the Open Source Intelligence (OSINT) which aims to provide a solution for those processes of knowledge management, which require the intervention of a human operator, unaided by information technology (IT) support, in one or more stages of the procedure. Such intervention involves a considerable waste of time and resources that could be reduced through the use of an IT tool, partially or totally automating entire stages of the procedure. DOMINIO is a platform that implements tools for automatic online information aggregation, its analysis, the possible alignment with traditional databases and the representation through infographic and georeferencing tools, in order to generate a report. This paper describes the platform architecture, the main algorithms used in the analysis stage of the contents and possible directions of development.

Abstract
Prisma ha sviluppato una tecnologia innovativa finalizzata all'Open Source Intelligence (OSINT) che intende fornire risposta alle necessità di knowledge management, che richiedono l'intervento di un operatore umano, non assistito da supporti di information technology (IT), in una o più fasi della procedura. Tale intervento comporta un notevole dispendio di tempo e risorse che potrebbe essere ridotto attraverso l'utilizzo di uno strumento di IT, automatizzando parzialmente o totalmente intere fasi della procedura. DOMINIO è una piattaforma che implementa strumenti per l'aggregazione automatica di informazioni on line, la loro analisi, l'eventuale allineamento con banche dati di tipo tradizionale, la rappresentazione attraverso tool di infografica e georeferenziazione, allo scopo di generare una reportistica. Il presente contributo descrive l'architettura della piattaforma, i principali algoritmi adottati nella fase di analisi dei contenuti e le possibili direzioni di sviluppo.
Keywords: knowledge management, Open Source Intelligence tool, Information Technology,

1. Introduction

There is a close link between data management and knowledge on the one hand, and knowledge and innovation on the other. The growing mass of unstructured information from disparate channels (search engines, RSS feeds, social networks) and traditional databases entails the need to drastically simplify the preparation, analysis and reporting stages required to structure the information. In fact, only a structured information translates into knowledge. Knowledge, in turn, is a major driver of innovation and, properly managed, it translates into a competitive advantage. The idea at the basis of the tool OSINT (Open Source Intelligence) stems from the needs expressed by analysts – mainly involved in the field of sentiment analysis and opinion mining industry. However, this idea is enough comprehensive to encompass all those activities of knowledge management, similar to the former, which require intervention by a human operator, unaided by IT support (Information Technology), in one or more stages of the procedure, the intervention of which involves a great deal of time and resources. Although in high-end solutions machine learning systems are starting to spread, the available technology is still characterized by significant limitations, especially in the presence of unstructured information. In particular, with regard to supervised machine learning systems, intervention is required by an operator in the initial stages of the procedure and, in general, with reference to any automated system applied to the analysis of a text, it is still impossible to identify complex cognitive functions (for example, irony). Of course, these problems are immanent in many fields of OSINT, and they also affect the stage of reporting, which requires a direct involvement of the analyst, unaided by IT. So, the availability of an IT tool that minimizes human operator intervention − partially or totally automates entire stages of the procedure − would result in substantial advantages, like time savings, increased productivity and the resulting increased efficiency in the allocation of human and financial resources.

Prisma has developed an innovative technology of OSINT, which aims to fix the problems briefly described above. The platform implements tools for automatic aggregation of the online information, their analysis, the alignment with traditional databases, the representation through infographic and georeferencing tools, aimed to automate also the phase of elaboration of the final report.

This paper will describe the architecture of the platform, the main analysis modules and the possible directions of development.

2. Platform Architecture

DOMINIO is an OSINT (Open Source Intelligence) platform that automatically aggregates information from online and traditional databases, analyses it and generates reports on a user-defined subject. The platform collects information by querying several channels: search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr), traditional databases. The goal of DOMINIO is to build a structured set of contents, as broad as possible, and to carry out a wide range of qualitative and quantitative analysis. DOMINIO stores these contents within a non-relational database (DB) (MongoDB, 2018; Morphia, 2018), classifying the various documents by channel of origin (Twitter, Facebook, RSS, etc.) to ensure the homogeneity of the collections.

Among the options, the DOMINIO user can make queries on-demand or in a continuous mode. The on-demand option carries out an asynchronous search, while the continuous mode option enables to aggregate periodically data and to track a subject over an extended time span. The DOMINIO's architecture allows the user to switch from one mode to another; the availability of two searching modes allows overcoming the trade-off between accuracy of analysis and speed of processing.

With regard to one or more subjects selected by the operator, DOMINIO performs synchronous or asynchronous research on a set of Internet channels, such as search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr). The user can also extend the search to the Deep Web, through specific search engines, such as Torch or Grams.

Moreover, to meet specific information needs, DOMINION can match these search results with the information achievable from the traditional databases to support many types of analysis (brand reputation, country risk assessment, opinion polls, cyber security, etc.), considerably increasing the operability and flexibility of the tool.

Among the traditional databases already available, DOMINIO includes:

- IHS Jane's (2018), which provides updates on military and political situation, terrorist acts, civil wars, transportation system, for most of the countries in the world;
- Bureau Van Dijk (2018), which collects firms data on ratings, shareholdings, equity investments and M&A;
- MIG (a geographic information database drawn up by one of the authors).

In addition, for specific information purposes, DOMINIO is open for interfacing with Enterprise Resource Planning databases (like SAP, Oracle, etc.) through market tools (Business Object, Quick View).

The search results are recalled by the analyst, who operates from a CMS (Content Management System) application to manage the structured set of content and conduct a wide range of qualitative and quantitative analyses (from simple summary statistics to sophisticated multivariate analyses and text and opinion mining techniques).
The statistical methods implemented on DOMINIO are chosen by the Prisma research team according to a set of criteria that privileges the suitability of one algorithm to automate entire stages of the procedure, in accordance with the original design idea. Moreover, the modular architecture of DOMINIO, described briefly below, allows a quick integration of the latest analysis tools and innovative methodologies produced in the academic field.
Once the stage of content analysis is completed, the CMS application generates a micro-site containing the results (geo-referenced maps, summary statistics, multivariate analysis results, textual and semantic analysis of sentiment analysis, etc.). After selecting a graphic layout for the final report, the analyst has only to write notes and final remarks.
The possibility of including features generating automatic and/or auto-completion comments, customizable by the user, is also being studied. Once the last stage is completed, the report is ready for online publication or traditional diffusion in pdf format, or linked to external services.
From an architectural point of view, DOMINIO is designed following the most modern criteria of modular software design, with the parallel development of the platform's modules. In short, in order to ensure a greater fault tolerance and high safety standards, the system is divided into three independent logical units (cfr. Figure 1):

- DOMINIO Engine Unit (MEU), which implements the features of 1) scraping information from the sources mentioned above (web, social networks, RSS feeds, traditional databases); 2) storage of results on MEDB database; 3) qualitative and quantitative analysis;
- DOMINIO RESTurl Unit (MRESTU), which receives requests from the MCMS unit, verifies the consistency and forwards the request to the unit ME. Upon receiving the response, it implements the request by adding additional fields (username, token, etc.) and returns them to the MCMS client. The MRESTU unit contains the database (MRESTDB) for user profiling;
- DOMINIO Content Management System Unit (MCMSU), which manages the stage concerning the reporting and archiving of reports according to pre-logical criteria (organization by topic, chronologically, for templates, etc.).

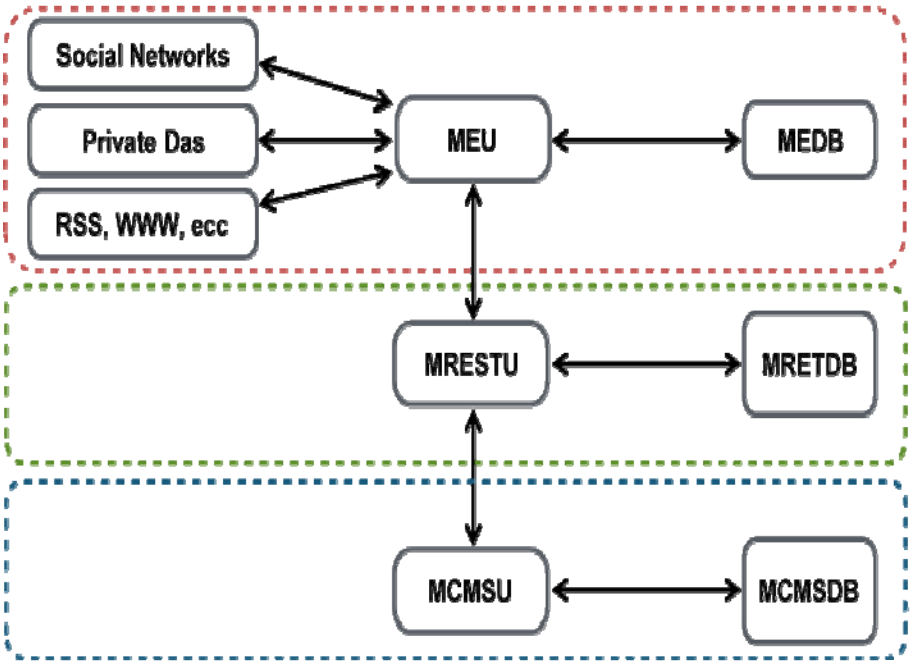


Figure 1 - DOMINIO General Overview

3. Main analysis modules

3.1. Country Threat Assessment

The Country Threat Assessment module supports the Company Intelligence and Security analyst in the country's risk assessment process. Through a responsive type interface, it aggregates information from major global industry databases (eg, IHS Jane's) giving an assessment of external and internal risk and that due to political and socio-economic factors and potential outbreaks or revolutionary movements for 192 different countries. Country Threat Assessment is integrated with intelligence information updated weekly on each country. Through an automatic report, data is aggregated into a single file by optimizing timing of risk assessment and providing a solid foundation for any further detailed analysis. DOMINIO offers the possibility of making a full or partial information download, and the generation of an automatic report, thus optimizing any drafting processes.

3.2. Due Diligence

The Due Diligence module supports the Economic Intelligence analyst in the process of business valuation in relation to suppliers, partners and customers. Among the sectors analysed in the module are included

assessments of profitability and financial performance as well as creditworthiness. Through a simple and intuitive interface, the module aggregates information from leading industry databases and returns an economic, financial and credit risk profile on hundreds of millions of businesses around the world. The Due Diligence Module also allows an assessment of individuals, through the analysis of individuals exposed politically, returning an automatic report that integrates the main aspects of each business and its economic risk analysis.

3.3. Open Source Intelligence

On completion of the aggregation of large amounts of data from major social networks (Facebook, Twitter, Youtube) and the main Italian newspapers based on predetermined keywords analyst, a statistical representation of the main trending topic is returned and an output of structured data for subsequent multivariate analysis is generated. Furthermore, the module allows the geo-referencing of content, highlighting even at geographic levels useful signs for the analyst. As for each of DOMINIO's modules, it is possible to generate automatic reporting.

3.4. Geographic Information Module

This is a module that analyses the information inferable from a dataset of basic statistical information and related indicators, with reference to a multitude of subjects, 9 of which are in a current stage of development. The basic statistical information, refers to the division of the Italian territory into provinces, covering a time period between 1995 and the latest available year, which for some subject areas is ongoing or, more frequently, the previous year to the current one. The dataset will be supportive to a wide range of applications - from forecasting and scenario analysis, counterfactual analysis to spatial analysis.

3.5. Text Mining Module

On completion of the automatic analysis of textual data using statistical methods (Lebart et Salem, 1994; Feldman et Sanger, 2006; Bolasco, 2013), in order to extract structured information, the main statistical methods of analysis of textual data implemented in DOMINIO are: factor analysis (correspondence analysis, multiple correspondence analysis); cluster analysis (k-mean, bisetting k-mean, fuzzy clustering, etc.); network analysis; Markov analysis; pattern recognition.

For example, during the French presidential campaign of 2017 we analysed the sentiment about migration, that was one of the most debated theme. We performed an Emotional Text Mining (Greco et al., 2017) in order to explore

the emotional content of the Twitter messages concerning migration written in French in the last two weeks before the first round of the presidential election in 2017. The aim was to analyse the opinions, feelings and shared comments, classifying the contents and the sentiments. We retrieved the messanges from the Twitter repository collecting a sample of over une hundred thousand tweets The large size corpus of 2.154.194 tokens (TTR = 0,01; Hapax percentage = 40,4) underwent a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The advantage connected with this approach is to interpret the factorial space according to words polarization, thus identifying the emotional categories that generate migration representations, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space (Greco, 2016).
The results interpretation allowed for the detection of seven representations of migrants that corresponded to three different sentiments: positive (42%), negative for the community (45%), and negative for migrants (13%). We considered as negative the representation of migrants as squatters, invaders, terrorists, trafficking slaves and migration victims, and positive the sport heroes and the EU solidarity target. Among the negative clusters, we distinguished negativity according to the direction of the action: squatters, terrorists and invaders are negative for the community and trafficking slaves and migration victims are negatives for migrants themselves (see Greco et al., 2017). Moreover, It was possible to highlight the connection between the real life events and the tweets production. While the terrorist attack three days before the first round of voting in the centre of Paris had slightly modified the production of messages, the candidates' interviews had a higher impact. This suggests that the medialization was more important than the terrorist attack in the production of messages (see Greco et al., 2017).

4. Conclusion

The innovative aspect that characterizes DOMINIO is the ability to aggregate data of different types and from different channels of information, automatically, simply and transparently. Moreover, its structure allows for the integration of the latest analytical tools and innovative methodologies produced in academia. By means of an automated reporting system, the analyst is supported in the assessment of risk and the collection of information in the geopolitical and economic field and from open sources. The set of modules allows the analyst to generate knowledge from an ever-growing amount of data by optimizing the processes of assessment and risk reduction.

References

Bolasco S. (2013). *L'analisi automatica dei testi: Fare ricerca con il text mining*. Carocci.

Bureau von Dijk (2018). *A Moody's Analytics Company*. Bureau von Dijk, https://www.bvdinfo.com/it-it/home

Feldman R. and Sanger J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *RIEDS*, 71(2): 125:36.

IHS Jane's (2018). *Jane's Information Group*. IHS Jane's, http://www.janes.com

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

MongoDB (2018). *MongoDB for GIANT ideas.* MongoDB, https://www.mongodb.com

Morphia (2018). *The Java Object Document Mapper for MongoDB*. MongoDB, https://mongodb.github.io/morphia/

Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.

Is training worth the trouble? A PoS tagging experiment with Dutch clinical records

Leonie Grön, Ann Bertels, Kris Heylen
KU Leuven – leonie.gron@kuleuven.be; ann.bertels@kuleuven.be; kris.heylen@kuleuven.be

Abstract
Part-of-speech (PoS) tagging is a core task of Natural Language Processing (NLP), which crucially influences the output of advanced applications. For the tagging of specialized language, such as that used in Electronic Health Records (EHRs), the domain adaptation of taggers is generally considered necessary, since the linguistic properties of such sublanguages may differ considerably from those of general language. Previous research suggests, though, that the net benefit of domain adaptation varies across languages. Therefore, in this paper, we present a case study to evaluate the effect of training with in-domain data on the tagging of Dutch EHRs.

Keywords: Electronic Health Records; Part-of-Speech tagging; medical sublanguage; Dutch

1. Background

EHRs are valuable resources for data-driven knowledge-making. To unlock the relevant information from free text, domain-specific NLP systems are required. Such systems must deal with a text genre that can be characterized by a high density of specialized terms, including non-canonical variants, and non-standard syntactic constructions. These properties affect all further steps in a processing pipeline, starting from core tasks such as PoS tagging. Since PoS values are important features for further processing, the output of many systems, such as tools for term extraction and term-to-concept mapping (e.g. Doing-Harris et al., 2015; Scheurwegs et al., 2017), crucially depends on the accuracy of the PoS tags assigned in the first place. Processing suites such as cTAKES (e.g. Savova et al., 2010), which have been developed specifically for the medical domain, are known to boost tagging performance. As most tools are only available for English, though, systems dealing with other languages, such as Dutch, are required to start the domain adaptation from scratch. Typically, this process involves the re-training of an existing tool on hand-coded data, which is time- and labor-intensive. Besides, evidence from German challenges the wide-held belief that domain training is a prerequisite to achieve good tagging performance (Wermter et Hahn, 2004).

Given these considerations, we conduct a pilot study to investigate the potential benefit of domain adaptation for the PoS tagging of Dutch EHRs. Firstly, we assess the impact of training with a hand-coded clinical dataset on the accuracy of an off-the-shelf tagger. Secondly, we evaluate how the difference in accuracy affects the output of a term extraction method based on PoS patterns.

2. Related Work

For the PoS tagging of clinical writing, the main challenges reside in the particular linguistic properties of the genre, both at the lexical and the syntactic level: On the one hand, EHRs contain a high proportion of specialized terminology and idiosyncracies, including misspellings and non-canonical abbreviations; a tagger developed for general language will thus encounter a high number of out-of-vocabulary words (Knoll et al., 2016). To complicate this matter, the PoS distributions in clinical corpora differ from those found in general language, which may be detrimental to the statistical classification of unknown or ambiguous tokens (Pakhomov et al., 2006). On the other hand, EHRs are typically composed in a telegraphic style, which can be characterized by the omission of functional syntactic elements; the lack of linguistically informative context may prevent the accurate prediction of PoS transitions within n-grams (Coden et al., 2005). At the same time, the average sentence length in EHRs is relatively short; the high number of inter-sentential transitions may pose additional pitfalls for an out-of-domain tagger (Pakhomov et al., 2006). Most previous research thus agrees that the use of off-the-shelf taggers on clinical writing is highly prone to errors, which are likely to be propagated through the different levels of an application (Ferraro et al., 2013).Therefore, many state-of-the-art systems use an annotated set of EHRs for training. The creation of training materials comes at a cost, though, and entails a range of methodological challenges in itself, such as the creation of suitable guidelines and tagsets (Albright et al., 2013). To circumvent these issues, alternative ways of domain adaptation have been explored, including the integration of a domain-specific vocabulary, and the exploitation of morphological features to classify unknown words (Knoll et al., 2016). However, other languages than English may present a different case: In an early study, Wermter & Hahn (2005) come to the conclusion that in German, taggers trained on newswire perform very well on EHRs. This surprising finding can be partly attributed to the rich inflectional system of the language, which lends itself to the prediction of PoS categories. On the other hand, the low complexity of the medical sublanguage may be a factor: In their study, the general training data subsumed all PoS transitions found in the clinical test data, so that the tagger was sufficiently equipped to handle the latter.

3. Methods

3.1. Corpus and manual tagging

Our study is based on the analysis of a mixed sample of EHRs, containing a total of 375 documents. As detailed in Table 1, the subsets of this sample differ with regard to their medical subdomain, institutional origin and document structure: The EN and RD sets cover only one medical specialty, whereas the DL, SP and GP sets are less homogeneous; the DL, EN and RD sets were composed at a single institution, while the documents in the GP and SP sets are drawn from a multi-source database, Integrated Primary Care Information (ICPI), which contains EHRs from medical practices all across the Netherlands. Finally, the EHRs in four subsets (DL, GP, RD, SP) had been split into shorter fragments to comply with privacy standards; therefore, these documents are much shorter than those in the EN set, which count 204.2 tokens on average. All EHRs are tokenized with the NLTK tokenizer[1] and manually labelled by the authors, using the Universal Tagset (Petrov et al., 2012). Finally, for each subset, the EHRs are split into a training and test set, containing 67% vs. 33% of the files respectively.

Table 1: Overview of the subsets of our file sample. The first three columns specify the name of the subset, the document types, the origin and the number of institutions involved in their creation. The remaining columns give the number of documents, the absolute length in tokens, and the average document length in tokens.

Subset	*Document types*	*Origin*	*Nr. of sources*	*Nr. of documents*	*Subset length*	*Average document length*
DL	Clinical discharge letters	EMC Rotterdam	One	88	3597	40.88
EN	EHRs from endocrinology	UZ Leuven	One	80	16337	204.2
GP	EHRs from general practitioners	IPCI (Vlug et al., 1999)	Multiple	60	1431	23.85
RD	EHRs from radiology	EMC Rotterdam	One	60	1441	24.02
SP	Specialist letters from various fields (e.g. cardiology)	IPCI (Vlug et al., 1999)	Multiple	87	4784	54.99
			Σ	375	27590	73.57

[1] http://www.nltk.org/_modules/nltk/tokenize.html

3.2. Evaluation

3.2.1. Effect of domain training on tagging performance

Firstly, we assess the impact of using in-domain data for training on tagging accuracy. For evaluation, we use the state-of-the-art Perceptron Tagger.[2] This tagger uses context tokens as well as suffix features for classification. As Knoll et al. (2016) show, this configuration outperforms a primarily sequential tagger, as used by Wermter & Hahn (2005), on clinical data. The pre-compiled model for Dutch is trained on the Alpino Treebank (van Noord, 2006). In addition, we build a domain-specific model based on the manually labelled training set. Then, we feed both models into the tagger to classify the test set. To measure the accuracy of each model, we calculate the precision, i.e. the proportion of tags that match those in the manually labelled gold standard.[3] To compare the effect across the different subsets, we calculate the gain in precision achieved with the domain model relative to the precision achieved with the Alpino baseline.

3.2.2. Effect of tagging performance on term recognition and extraction

Secondly, we quantify the effect of tagging performance on pattern-based term recognition. For the identification of term candidates, we use a set of PoS sequences that are characteristic for termhood in the domain. Similar to Scheurwegs et al. (2017), we focus on complex nominals, i.e. nouns surrounded by one or more modifiers; Table 2 provides some examples of such patterns.

Table 2: Examples of PoS patterns used for term retrieval. The left column lists the target tag sequence, the middle and right column provide Dutch examples and English translations of term candidates.

PoS pattern	***Dutch example***	***English translation***
adjective noun	'diabetische retinopathie'	*diabetic retinopathy*
noun adposition noun	'syndroom van Apert'	*syndrom of Apert*
noun noun	'zwelling enkel'	*swelling ankle*

Using a sliding-window approach, we iterate through the three tagged versions of the test set, i.e. the manually tagged gold standard, the version

[2] http://www.nltk.org/_modules/nltk/tag/perceptron.html

[3] The Alpino model uses a more fine-grained tagset than the Universal Tagset used for the manual tagging. To enable the comparison across models, the redundant labels from Alpino are mapped to the respective categories of the Universal Tagset (e.g. adj <adjective>, comparative <comparative> → ADJ <adjective>).

tagged with the Alpino model and the version tagged with the domain model. We identify all PoS sequences that match the pre-specified patterns, and extract the respective tokens for manual validation. For each version, we calculate the precision as the proportion of true positives, i.e. domain-specific phrases, relative to the total list of matches.[4] To assess the individual effect size, we also calculate the relative gain in precision for each subset.

3.3. Results

3.3.1. Effect on tagging performance

For PoS tagging, training on domain data has a sizeable effect on precision: The domain model reaches 85.8% accuracy on the test set of held-out EHRs, compared to 66.9% with the Alpino baseline. Regardless of the model, the best results are achieved for DL, followed by RD and EN; for SP and GP, precision stays at the lowest levels. To evaluate the improvement across the different subsets, we compare the increase in precision relative to the value achieved with the baseline. The comparison of these values reveals considerable differences of the individual effect sizes: In SP, the training effect is most striking, followed by GP and EN; in RD and DL, the improvement is less evident.

3.3.2. Effect on term recognition and extraction

The increase in accuracy has a strong effect on the term retrieval task: When using the tags assigned by the Alpino model, only 3.42 of the retrieved candidates are correct; with the domain model, precision jumps to 9.3%. Again, the results vary substantially across the different datasets: Overall, the best results are obtained for EN, followed by RD and DL. In SP and GP, precision remains at the lowest levels. Judging from the relative gain in precision, though, we find the strongest increase in GP, followed by DL. In RD, EN and SP, we only find weaker effects. Table 3 provides the full results for both tasks.

For error analysis, we label all false positives with the nature of misclassification, whereby we distinguish between three types of errors: Firstly, errors based on erroneous PoS tags (e.g. 'merkt hypoglycemie' *notices hypoglycemia*, whereby the verb is tagged as an adjective); secondly, segmentation errors, whereby one token is associated with an unrelated one (e.g. 'oedeem Lipitor' *edema Lipitor*, whereby two unrelated nouns are

[4] To qualify as domain-specific, a phrase must contain at least one noun that has a concept entry in the clinical terminology SNOMED-CT (International Release July 2017; http://browser.ihtsdotools.org/). For instance, 'echografie rechterschouder' *echography right shoulder*, which refers to a clinical procedure, would count as a true positive; the general expression 'pak koekjes' *bag of biscuits* would not.

mistaken for a compound); thirdly, term candidates that match a target PoS pattern, but are not domain-specific (e.g. 'kleine boterhammen' *small sandwiches*). Then, we calculate the proportion of error types among the false positives provided by both models. With the Alpino model, the vast majority of errors (74.4%) is based on false PoS tags. About 18.2% of the proposed term candidates are out-of-domain, while only a small portion (7.3%) of errors is caused by mistakes in segmentation. Conversely, with the domain model, most false positives (49.7%) are out-of-domain terms; errors in tagging and segmentation account for 30.1% and 20.2% respectively.

Table 3 : Precision of PoS tagging and term extraction across subsets. The first column specifies the subset. The second and third column provide the percentage of correct tags assigned by the domain model and the Alpino model respectively; the fourth column contains the relative increase in precision. The remaining three columns provide the corresponding values for the extraction task.

PoS tagging				***Term extraction***		
subset	***% Prec domain model***	***% Prec Alpino***	***% increase***	***% Prec domain model***	***% Prec Alpino***	***% increase***
DL	89.62	76.61	16.99	7.33	2.64	177.87
EN	86.82	67.5	28.62	21.48	8.04	167.1
GP	79.81	61.76	29.23	3.28	0.84	291.31
RD	88.98	74.1	20.08	8.89	3.31	168.52
SP	83.68	54.5	53.53	5.52	2.26	144.09
Σ	85.78	66.9	29.69	9.3	3.42	189.78

4. Discussion

Overall, the positive effect of domain adaptation is evident: Using clinical data for training improved the accuracy of PoS assignments and, as a consequence, the output of the term extraction method. Based on our results, we do not see a clear relation between the amount of training data and the global level of precision: For PoS tagging, DL and RD, which are among the smaller subsets, score highest; on the other hand, for the term extraction task, EN, which is the largest subset, produces the best results by far. This indicates that the benefit of training hinges on linguistic and semantic *qualities*, rather than the mere *quantity* of the data.

In particular, tagging performance correlates with the *homogeneity* and *well-formedness* of the data. The *homogeneity* depends, on the one hand, on the medical field: A dataset such as RD, which is confined to one clinical

specialty, only makes reference to a fairly limited number of medical concepts; by contrast, a more heterogeneous set, such as SP, covers a wider range. Besides, the number of institutions involved in data creation plays a role: In an EHR sample provided by a single hospital, such as EN, it is likely that preferred terms and phrases are perpetuated throughout the dataset. By contrast, in a set drawn from a multi-source database, such as GP, the potential for variation is higher. Both these factors affect the overall size of the vocabulary, which, in turn, determines the complexity of the tagging task. The *well-formedness*, on the other hand, depends mainly on the EHR type. The GP set, for instance, contains mostly notes intended for internal documentation; these notes are written in an informal style, whereby function words and suffixes may be left out or truncated. As these features usually serve as predictors for PoS classification, their omission may cause a drop in tagging performance. While the global level of precision is thus lowest in conceptually and lexically EHR samples, such as GP and SP, the relative benefit of domain adaptation is the greatest here.

5. Conclusion

We conclude that the training with in-domain data benefits the output of PoS taggers for clinical Dutch. Especially if the file sample covers different subdomains, or if the language used deviates strongly from the standard, the potential gain in performance is great. At the same time, considerable training efforts are required to achieve only marginal improvements. Depending on the scope of the project and the composition of the sample, it may thus be preferable to implement a cheaper alternative, for instance by integrating a domain dictionary into the tagger.

Acknowledgements

This work was supported by Internal Funds KU Leuven.

References

Albright D., Lanfranchi A., Fredriksen A., Styler W.F., Warner C., Hwang J.D., Choi J.D. et al. (2013). Towards Comprehensive Syntactic and Semantic Annotations of the Clinical Narrative. *J Am Med Inform Assoc* vol. 20: 922–30.

Coden A.R., Pakhomov S.V., Ando R.K., Duffy P.H. and Chute C.G. (2005). Domain-Specific Language Models and Lexicons for Tagging. *J Biomed Inform* vol. 38: 422–30.

Doing-Harris K., Livnat Y. and Meystre S. (2015). Automated Concept and Relationship Extraction for the Semi-Automated Ontology Management (SEAM) System. *J Biomed Semantics* vol. 6 (15): 1–15.

Fan J.-W., Prasad R., Yabut R.M., Loomis R.M., Zisook D.S., Mattison J.E. and Huang Y. (2011). Part-of-Speech Tagging for Clinical Text: Wall or Bridge between Institutions? In *AMIA Annu Symp Proc*, pp. 382–91.

Ferraro J.P., Daumé H.I., DuVall S.L., Chapman W.W., Harkema H. and Haug P.J. (2013). Improving Performance of Natural Language Processing Part-of-Speech Tagging on Clinical Narratives through Domain Adaptation. *J Am Med Inform Assoc* vol. 20: 931–39.

Knoll B.C., Melton G.B., Liu H., Xu H. and Pakhomov S.V.S. (2016). Using Synthetic Clinical Data to Train an HMM-Based POS Tagger. In *2016 IEEE-EMBS (International Conference on Biomedical and Health Informatics)*, pp. 252–55.

van Noord, G. (2006). At Last Parsing Is Now Operational. In *Proceedings of TALN 2006*, pp.20–42.

Pakhomov S.V., Coden A. and Chute C.G. (2006). Developing a Corpus of Clinical Notes Manually Annotated for Part-of-Speech. *Int J Med Inform* vol. 75: 418–29.

Petrov S., Das D. and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In Piperidis N.C., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2089–96.

Savova G.K., Masanz J.J., Ogren P.V., Zheng J., Sohn S., Kipper-Schuler K.C. and Chute C.G. (2010). Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. *J Am Med Inform*

Les outils de la statistique textuelle pour analyser les corpus de données d'enquêtes de la statistique publique

France Guérin-Pace, Elodie Baril
Institut national d'études démographiques

Abstract
For more than 20 years, textual statistic methods have been allowing us to explore and analyze data from official statistics survey and the different corpora it contains: answers to an open question, associated words, significant life events. Based on three corpora of data: Population-Lived Spaces-Environments survey (Ined, 1992), EuroBroadMap survey on representations of Europe in the world (2009), and more recently the Information and Daily Life survey on adult reading skills (INSEE, 2011), we have demonstrated the diverse use cases of these methods and the richness that helps identify the corpus content in relation to the individual characteristics of respondents as well as to the survey questions. In recent years, we have mobilized these methods to post-codify the events collected in the IVQ survey. Today we will present to you the results of this work: the benefits and limitations of textual statistic method.

Résumé
Réponses à une question ouverte, mots associés, évènements marquants de la biographie, constituent autant de corpus issus de données d'enquêtes de la statistique publique que nous avons explorés et analysés avec les méthodes de la statistique textuelle, depuis plus de 20 ans. A partir de trois corpus de données : enquête Populations-Espaces de vie-Environnements (Ined, 1992), enquête EuroBroadMap sur les représentations de l'Europe dans le monde (2009), et plus récemment l'enquête Information et Vie quotidienne sur les compétences en lecture des adultes (Insee, 2011), nous montrons la diversité d'applications de ces méthodes, leur richesse pour cerner le contenu des corpus en lien avec les caractéristiques individuelles des répondants mais aussi d'autres questions d'enquête. Plus récemment nous avons mobilisés ces méthodes pour post-codifier les évènements recueillis dans l'enquête IVQ. Nous présenterons les apports et les limites de cette démarche.
Keywords: textual statistics, open-ended questions, associated words corpus, post-coding.

1. Des corpus de nature variée

Introduire un questionnement ouvert dans une enquête en population générale est toujours un défi pour les concepteurs même si les méthodes de la statistique textuelle ont prouvé depuis longtemps leur intérêt et leur efficacité pour leur traitement. Cerner les contours et l'acception d'un mot valise était l'objectif de l'introduction de la question ouverte «Si je vous dis environnement, qu'est-ce que cela évoque pour vous? » dans l'enquête « Populations-Espace de vie-Environnements » réalisée en 1992 (INED) auprès d'un échantillon de 6 000 personnes, représentatif de la population française. Un des objectifs consistait à examiner quelles représentations les populations construisent sur la notion même d'environnement.

Une technique un peu différente de recueil est celle adoptée, par exemple, dans l'enquête EuroBroadMap conduite en 2009 dans 18 pays. Enquêter près de 10 000 étudiants à travers le monde sur leurs représentations de l'Europe est l'un des objectifs de ce projet européen. Une pièce centrale de ce dispositif est de recueillir les mots associés à l'Europe par les étudiants[1] après leur avoir demandé de délimiter, selon leur perception, ses contours sur une carte du monde. A la différence du corpus précédent, les mots ne sont pas proposés sous forme de liste et ce sont les représentations spontanées qui sont recueillies. Cette technique des mots associés a pour intérêt de contraindre davantage le format des réponses et d'obtenir un corpus plus homogène. Une des principales difficultés de ce corpus est celle de la langue de recueil des mots associés. Pour résoudre en partie ce problème, nous avons choisi de traduire les réponses en anglais pour chacun des pays au moment de la saisie, selon des consignes précises[2].

Une autre forme de matériau qualitatif intéressant à recueillir dans les enquêtes concerne les événements de vie. Pour les démographes, le recueil d'éléments des parcours individuels possède une dimension explicative très pertinente, qu'ils s'agissent de points d'inflexion, de ruptures au sein des parcours biographiques ou d'éléments ponctuels sans conséquence à long terme (Laborde et al., 2007). C'est ce que nous avons mis en place dans l'enquête Information et Vie quotidienne (Guérin-Pace, 2009). Les évènements marquants peuvent être recueillis de manière ouverte ou fermée. L'intérêt de les recueillir, sous forme de question fermée, est de pouvoir

[1] La question posée était « Quels sont les mots que vous associez le plus à l' « Europe » ? Choisissez 5 mots au maximum. »

[2] Pour des raisons de coût et de délai, l'instruction donnée aux partenaires était de traduire, eux-mêmes, en anglais les mots associés au moment de la saisie des questionnaires. Les premiers traitements textuels ont permis de repérer des incohérences et nécessité un retour vers les questionnaires dans leur langue d'origine.

effectuer des comparaisons systématiques dans la mesure où tous les enquêtés répondent à une même question. Nous avons introduit dans l'enquête sous forme de question fermée les évènements les plus fréquemment cités (divorce ou séparation des parents, décès d'un proche, problème de santé, etc.). Les événements recueillis de manière « fermée » ne permettent pas d'aborder tous les thèmes notamment ceux portant sur des sujets sensibles (cas de violence par exemple). Le recueil sous forme d'énumération devient en effet vite intrusif, parfois déplacé, si les personnes ne sont pas concernées. Par ailleurs, par cette démarche, on fait l'hypothèse de la nature a priori traumatisante d'un événement sans savoir si Ego l'a vécu comme tel durant son enfance (Laborde et al., 2007). Nous avons ainsi fait le choix de compléter ce questionnement par la question ouverte suivante « Avez-vous connu un autre événement marquant durant votre enfance ? Si oui, lequel ? ». Près d'un quart des répondants déclarent un « autre événement marquant » de leur enfance en réponse à cette question. Parmi eux, un sur deux évoque un décès, un sur dix un événement lié à un problème de santé, et dans les mêmes proportions une situation de violence vécue durant l'enfance (Baril, Guérin-Pace, 2016).

Tableau 1 : Description des corpus analysés

Enquêtes	Corpus	Nombre de réponses	Nombre d'occurrences	Nombre de mots distincts
Populations-Espaces de Vie-Environnement (1992)	Environnement	4596	28716	2130
EuroBroadMap (2009)	Mots associés à l'Europe	9343	40800	5111
Information et Vie Quotidienne (2011)	Evènements marquants de l'enfance	3167	15993	2161

2. Une étape sous-estimée : lecture des mots du corpus et les statistiques lexicales

Une première étape essentielle d'analyse est la lecture du lexique des mots les plus fréquents associé à un corpus d'enquête. Ce lexique donne à lui seul un aperçu de la tonalité du vocabulaire (positive ou négative) et des registres abordés. Par exemple, dans le corpus de mots associés à l'Europe, le premier mot à connotation péjorative n'apparait qu'en 26ème position (*colonialism*). La lecture des événements les plus fréquents indique quant à elle le caractère individuel ou collectif, le plus souvent historique, des événements perçus.

Pour les enquêtes internationales ou à passage répété, le recours aux

statistiques lexicales permet de comparer la richesse du vocabulaire de manière pertinente. Ainsi, dans le corpus « Europe », la comparaison des proportions de mots distincts (Figure 1) apporte des informations intéressantes. Il apparaît ainsi que les étudiants interrogés dans des pays les plus éloignés de l'Union européenne (Cameroun, Chine, Russie, Brésil, Inde) ont une vision plus consensuelle ou partagée de l'Europe que ceux des pays qui en sont membres, ou à la marge.

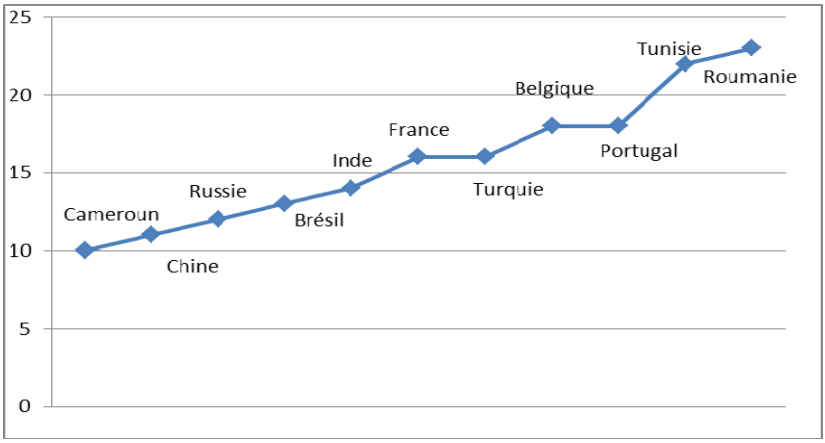


Figure 1 : Diversité des mots associés à l'Europe selon les pays d'enquête
Source : Enquête EuroBroadMap (2009)

3. Faire émerger le contenu d'une question ouverte à partir du TLE

Une autre application des méthodes d'analyse textuelle à un corpus de réponses à une question ouverte consiste à extraire les mondes lexicaux selon la méthodologie Alceste. Une CDH effectuée sur le tableau croisant les réponses à la question ouverte avec le lexique associé au mot « environnement » met en évidence deux approches fondamentalement différentes de la notion d'environnement (Figure 2). La première aborde l'environnement selon une approche cognitive concernant un espace physique et social (qualité de vie, univers local, etc.), tandis que la deuxième approche est plus symbolique ou imaginaire (iconographie de la nature, sensation de bien-être.).

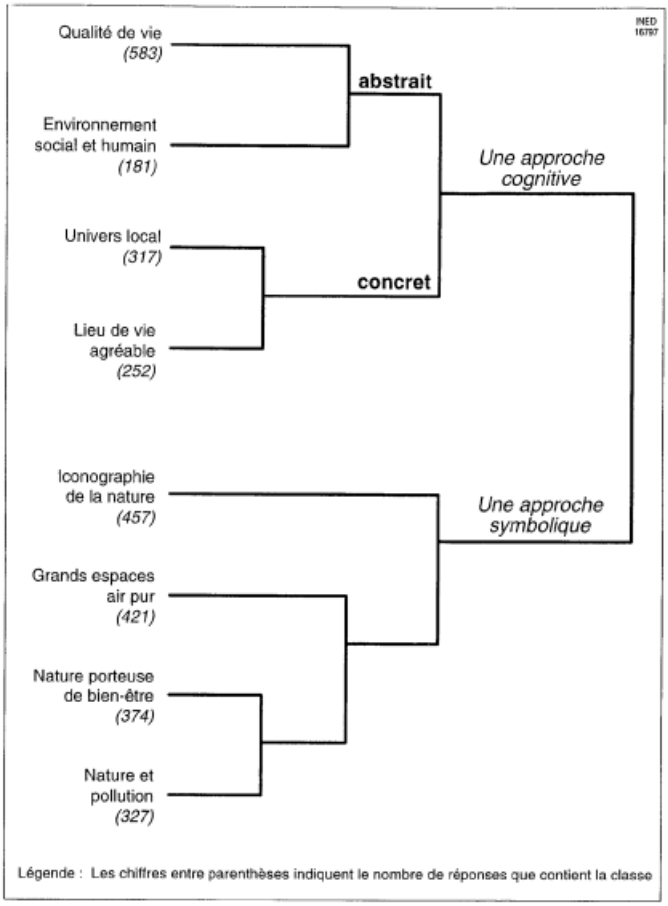


Figure 2 : Les mondes lexicaux du corpus « environnement » (Alceste)
In Guérin-Pace F., 1997

4. Croiser les réponses spontanées avec un questionnement fermé

Les limites d'interprétation d'une question ouverte résident dans l'impossibilité d'interpréter ce qui n'a pas été évoqué par les répondants. Compléter ce dispositif par un questionnement fermé permet d'y remédier. Nous avons ainsi, à la suite de la question ouverte, introduit deux questions fermées qui proposaient une liste de mots et d'adjectifs pouvant être associés ou non, par le répondant, au mot « environnement »[3]. L'observation conjointe

[3] Les questions étaient libellées de la manière suivante : « Voici une liste de noms (adjectifs). Lesquels vous semblent liés à la notion d'environnement ? (Pour chacun, précisez oui ou non).

des réponses à ces deux modes de questionnement par une ACM sur le TLA permet d'enrichir l'analyse du contenu « spontané » au regard des représentations fermées.
On observe ainsi (Figure 3) que l'opposition entre un environnement fait de « relations » et un environnement fait de « nature » (axe horizontal) s'accompagne, par exemple, du choix ou du refus de mots et d'adjectifs qui décrivent les nuisances urbaines. Sur l'axe vertical, à l'opposition entre un environnement conçu comme une proximité immédiate et un environnement basé sur les relations entre « l'homme et son milieu » correspond un vocabulaire associé qui renforce cette perception. Proche de la première perception, on relève les mots « maison-oui », « amical-oui », « sécurité-oui » et « planète-Non ».

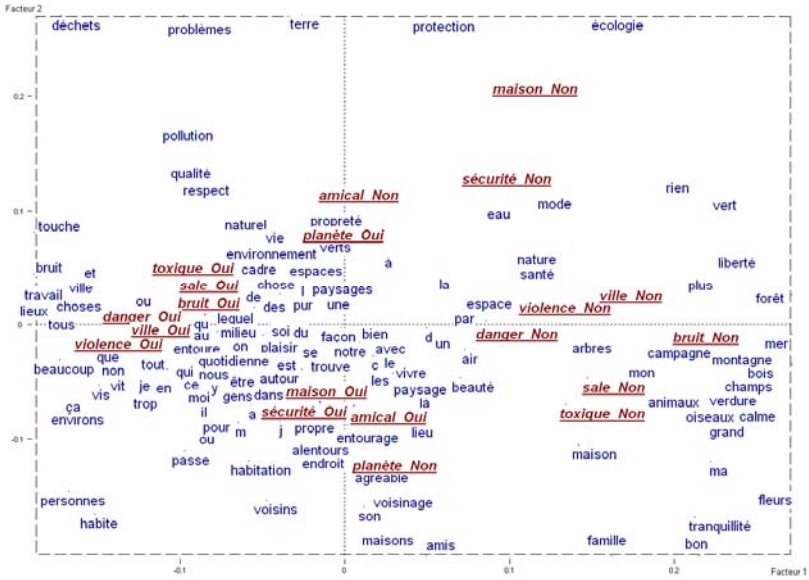


Figure 3 : Proximité entre formes du corpus « environnement » et associations proposées Guérin-Pace F., Garnier B, 1995 Lecture : à proximité des mots « santé » ou « liberté » cités en réponse à la question ouverte, on relève les réponses « non » à l'association du mot environnement aux mots « ville » ou « violence ».

5. Post-coder les événements marquants de l'enfance par la statistique textuelle

Une autre application plus récente de ces méthodes pour post-codifier des réponses à une question ouverte peut sembler contradictoire avec l'esprit même de la statistique textuelle. Il s'agit plus précisément de post-codifier les évènements recueillis dans l'enquête Information et Vie quotidienne (IVQ). Pour cela, nous avons effectué une classification (CDH) sur le tableau lexical

entier croisant les réponses à la question « Avez-vous vécu d'autres événements marquants ? » avec le lexique du corpus. On retient une partition en cinq classes au sein de laquelle on observe une première dichotomie entre des événements de nature collective (guerre d'Algérie, Mai 1968, etc.) et un ensemble de classes qui évoquent des événements de nature individuelle : décès, maladie, accident et violence (Figure 4). Nous avons ajouté à ces cinq classes deux classes supplémentaires : une classe intitulée « Refus » regroupant toutes les réponses qui marquent une volonté de l'enquêté de ne pas détailler l'événement marquant à l'enquêteur (tout en ayant donné une réponse affirmative à la question « Avez-vous connu un autre événement marquant ? ») ; une classe « Autre » au sein de laquelle nous avons regroupé les réponses non classées[4]. Nous avons ensuite cherché à affiner cette typologie en précisant les acteurs éventuels impliqués dans les événements. Par exemple, au sein de la classe « Maladie » (classe 2), nous avons filtré au moyen d'un vocabulaire familial (père, mère, frère, sœur, tante, ami, etc.) et constitué 4 sous-modalités distinctes selon les personnes concernées.

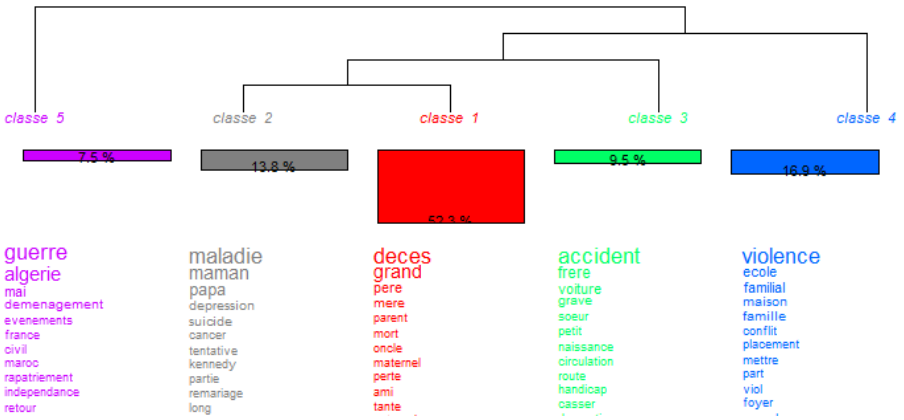


Figure 4 : Typologie des événements marquants de l'enfance
Source : Enquête IVQ, Iramuteq (classification Méthode Reinert)

Nous avons procédé de la même manière pour la classe « violence » en distinguant cette fois les personnes concernées par l'événement et son auteur éventuel. Nous obtenons finalement une typologie construite sur les questionnements ouverts et fermés, composée de 43 items (Baril, Guérin-Pace, 2016), qui pourrait être réutilisée pour d'autres enquêtes nationales.
En conclusion, ces différentes applications sur des corpus variés d'enquêtes

[4] Près de 90 % des 3167 réponses à cette question sont classées.

de la statistique publique permettent de mettre en évidence la diversité des apports des méthodes de la statistique textuelle. Aujourd'hui, de plus en plus d'enquêtes nationales abordent des thématiques sensibles (violences, précarité, illettrisme, etc.). Le recours à un questionnement ouvert s'avère ainsi indispensable en permettant au chercheur d'objectiver sa démarche. Les méthodes de la statistique textuelle se révèlent incontournables dans cette perspective.

Références

Baril E., Guérin-Pace F. (2016). Compétences à l'écrit des adultes et événements marquants de l'enfance : le traitement de l'enquête Information et vie quotidienne à l'aide des méthodes de la statistique textuelle, *Economie et statistique,* n°490, pp. 17-36.

Guérin-Pace F. (2009). Illettrismes et parcours individuels, *Economie et statistique*, n°424-425.

Brennetot A., Emsellem K, Guérin-Pace F., Garnier B. (2013). Dire l'Europe à travers le monde. Les mots des étudiants à travers l'enquête EuroBraodMap, Cybergeo : European Journal of Geography.

Guérin-Pace F., Collomb P. (1998). Les contours du mot environnement : Enseignements de la statistique textuelle, *L'Espace Géographique*, n°1, pp. 41-52.

Guérin-Pace F. (1997). La statistique textuelle : un outil exploratoire en sciences sociales, *Population*, n°4, pp. 865-888.

Laborde, C., Lelièvre, E., Vivier, G. (2007). Trajectoires et événements marquants, comment dire sa vie : Une analyse des faits et des perceptions biographiques. *Population*, vol. 62,(3), pp. 567-585.

ssoc vol. 17: 507–13.

Scheurwegs E., Luyckx K., Luyten L., Goethals B. and Daelemans W. (2017). Assigning Clinical Codes with Data-Driven Concept Representation on Dutch Clinical Free Text. *J Biomed Inform* vol. 69: 118–27.

Vlug A. E., van der Lei J., Mosseveld B.M., van Wijk M.A., van der Linden P.D., Sturkenboom M.C., and van Bemmel J.H. (1999). Postmarketing Surveillance Based on Electronic Patient Records: The IPCI Project. *Methods Inf Med* 38 (4/5): 339–44.

Wermter J. and Hahn U. (2004). Really, Is Medical Sublanguage That Different? Experimental Counter-Evidence from Tagging Medical and Newspaper Corpora. In Fieschi M., Coiera E. and Li Y.-C.L. *Proc. of the 11th World Congress on Medical Informatics (MEDINFO 2004)*, pp. 560–64.

Annotation-based Digital Text Corpora Analysis within the TXM Platform

Serge Heiden
Université de Lyon, ENS de Lyon, IHRIM – UMR5317, CNRS – slh@ens-lyon.fr

Abstract
This paper presents new developments in the TXM textual corpora analysis platform (http://textometrie.org) towards direct text annotation functionalities. Some annotations are related to a web based external historic ontology called SyMoGIH and others to co-reference information between words or to word properties like part of speech or lemma.
The paper discusses the methodological stakes of unifying in a single framework the production and the analysis those annotations with the traditional ones already available in TXM corresponding to the XML markup of the text sources and to the linguistic annotations automatically added to texts by NLP tools.

Keywords: textometry, TXM, digital text representation, XML, TEI, annotation, ontology, co-reference, part of speech, digital hermeneutic circle.

1. Introduction

TXM (Heiden, 2010) is a software platform offering textual corpora analysis tools. It is delivered as a standard desktop application for Windows, Mac and Linux and as a web portal server application (http://textometrie.org).
Its analysis tools combine qualitative types of tools like word lists, concordancing or text edition navigation (close reading) with synthetic quantitative types of tools like factorial analysis, clustering, keywords or statistical co-occurrence analysis (distant reading).
To be able to work on texts, the platform imports first the corpus sources to build a rich internal representation of texts through the following general workflow:

a) first the "base text" of each text is established: this operation implements "digital philology" principles and consists of decoding information in the various formats of the source documents[5] to

[5] TXM can analyze three main types of corpora : corpora of *written texts*, possibly including paginated editions including images of facsimiles ; *record transcriptions* corpora, possibly time synchronized with the audio or video source ;

decide primarily where are the text limits, internal structures boundaries and words and punctuations of the text. Its result is represented in a pivot XML format especially designed for TXM called "XML-TEI TXM" and extending the standard encoding recommendations of the Text Encoding Initiative consortium (TEI Consortium, 2017) ;

b) then, natural language processing (NLP) tools are optionally applied to the base text to automatically add linguistic information like sentence boundaries, grammatical category (pos = part of speech) and lemma of words by eg TreeTagger (Schmid, 1994), etc. As NLP tools generally don't take XML format as input, the pivot representation is first converted to raw text for NLP processing and results are added back into the XML-TEI TXM representation ;
c) finally a specialized representation of texts is built into TXM for efficient execution of its tools (by indexing for search engines and text edition rendering).

From the point of view of TXM, NLP tools results in b) are seen as *automatic* annotations added to the initial XML-TEI TXM representation of texts built in a), and the XML tags of the initial XML-TEI TXM representation in a) can be seen as *manual* annotations added to the base text (or raw text), typically philologically edited with the help of specialized XML editors (like Oxygen XML Editor[6]) outside of TXM when the source is in XML format, or as *automatic* annotations added by TXM when converting from some other format into XML-TEI TXM. All TXM tools apply indiscriminately to all types of annotation regardless of their origin (automatic or manual).

Thus, TXM implements a traditional workflow combining a "text source encoding and annotation" step to an "application of analysis tools to annotated texts" step. The text analysis tools use text annotations (for example word pos) to offer their services and produce their results (for example the concordance of all infinitive verbs). The workflow is unidirectional and the whole of it must be passed through again completely if any annotation needs to be corrected. To add or correct annotations, the user has to edit the sources or the annotations outside of TXM. For example word properties can be exported from the XML-TEI TXM representation, edited in a spreadsheet and inserted back into the texts before re-import[7].

---

and parallel *multilingual* corpora aligned at the level of a textual structure such as the sentence or the paragraph.

[6] https://www.oxygenxml.com

[7] see for example this tutorial based on TXM macros: https://groupes.renater.fr/wiki/txm-users/public/tutoriel_correction_mots.

This paper introduces new services developed in TXM to annotate directly texts from within the results view of specific tools for a better integration of philological and analytic work.

2. Annotation services in TXM

The new annotation services concern both adding and correcting information and all the annotations edited are meant for further exploitation by usual TXM tools.

2.1. SyMoGIH annotation by concordance

The first new service, developed in partnership with the LARHRA research laboratory in history[8], is based on the annotation of concordance pivots: any sequence of words composing the pivots can be annotated with any semantic category[9] coming from the SyMoGIH[10] historical ontology framework (Beretta, 2015). In this architecture, the SyMoGIH web platform hosts the ontology of historic facts and knowledge, and concordances provide the user interface to link identifiers of those data to text spans for further analysis. As an illustration, see figure 1 the annotation of the "Faculté de droit d'Aix" entity (of id CoAc13562) in unverified OCRed texts of the "Bulletin administratif de l'Instruction publique" corpus[11].

TXM internal management of those annotations is equivalent to a re-import of the current pivot representation of the annotated texts. After re-import (after saving annotations) the new annotations are available for all TXM tools to work on like any original "annotation" of the texts (internal structures and their properties, word properties, etc.).

2.2. URS annotation in text edition

The second new service is based on manual annotation of word sequences inside text editions with elements of a Unit-Relation-Schema (URS) annotation model. URS type annotations are designed to encode discourse entities like co-reference chains in texts (Schnedecker, Glikman, & Landragin, 2017). In a URS model, Units or entities have any number of properties and can be linked together by the two other annotation types: Relations, having any number of properties (1-to-1 relation type), and Schemas, having any

---

[8] http://larhra.ish-lyon.cnrs.fr

[9] pivots can also optionally be annotated with simple keywords or with key-value pairs, managed by TXM in a local repository.

[10] http://symogih.org/?lang=en

[11] see the Bibliothèque historique de l'éducation (BHE) project: http://www.persee.fr/collection/bhe

number of properties (1-to-n relation type). Any types and properties of units, schemas, and relationships are definable in the annotation model before and during annotation. The types and properties are chosen by the user, they are not limited to co-reference chains.

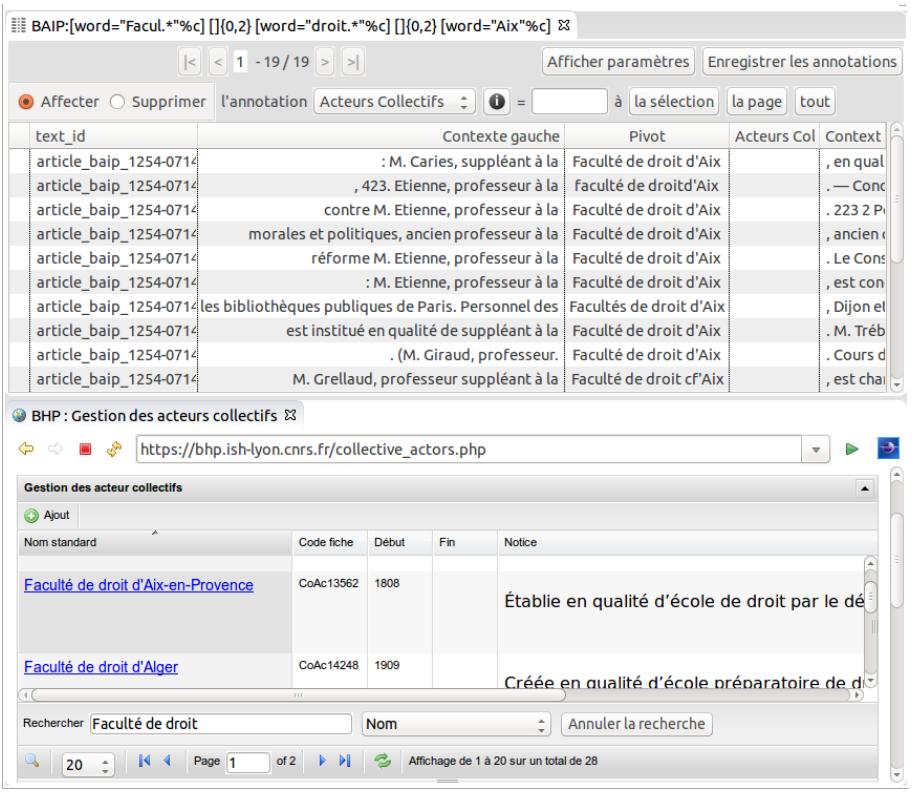


Figure 1: TXM screenshot of a Concordance of a "Faculté de droit d'Aix" word sequence pattern to annotate (top) and of browsing SyMoGIH semantic categories to use for the annotation (bottom).

The original URS model has been designed and developed in the Glozz (Widlöcher & Mathet, 2009) and Analec (Landragin, Poibeau, & Victorri, 2012) software. It is being integrated into TXM through the text edition reading tool for a project funded by the French National Research Agency (ANR) called DEMOCRAT[12].

As an illustration, see figure 2 the annotation of the "ses loix" word sequence

[12] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-15-CE38-0008

with a unit of type MENTION, of "GN.POS" grammatical category and "les lois de la divinité" referent, in the first chapter of the 1755 edition of *De l'esprit des lois* by Montesquieu. TXM internal management of those annotations can be represented as new XML-TEI stand-off annotations anchored to the word elements of the XML-TEI TXM representation of texts (Grobol, Landragin, & Heiden, 2017).

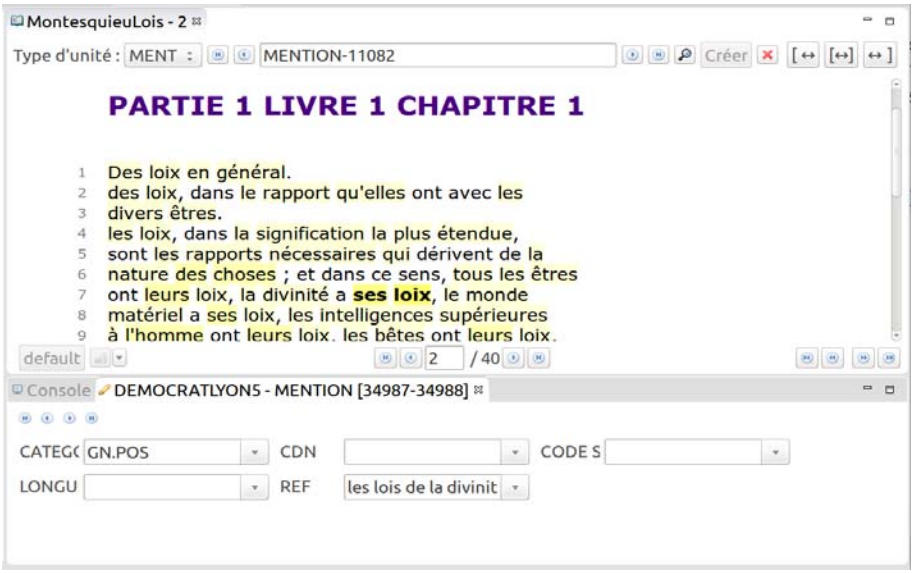


Figure 2: TXM screenshot of the edition of the first page of De l'esprit des lois with units of type MENTION highlighted in yellow and the selected unit in bold (top) and the current values of the properties of the selected unit (bottom).

2.3. Word properties annotation by concordance

The third service will be based on the annotation of concordance pivot words: a word present in the pivots of a concordance will be able to be annotated with properties. The primary goal of that service is to annotate and correct grammatical properties and lemma of word elements of the XML-TEI TXM representation of texts. This development is done for a project co-funded by the ANR and Deutsche Forschungsgemeinschaft (DFG) called PaLaFra[13] <http://palafra.org>.

2.4. Editing XML sources

Finally we are developing the possibility to directly edit the XML sources

[13] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-14-FRAL-0006

from within TXM through an internal XML editor. This editor will eventually be accessed through TXM tools as a "back to source" operation similar to the current "back to text" operation (for example from a concordance line to a text edition page).

3. Discussion

By using a common XML-TEI pivot representation for internal management of corpora for all the annotation services, TXM unifies transcription and annotation activities in a single framework. In this framework, annotations represent manual (user), semi-automatic (machine+user) or automatic (machine) interpretation results used further for analysis and interpretation work. The reflexive nature of the resulting text analysis workflow is schematized in figure 3. Texts are first digitized by OCR, transcribed or converted from digital formats. They are then philologically corrected and established through XML-TEI manual encoding. Then automatically processed by NLP tools while being imported into TXM to produce the TXM internal corpus model. Corpus analysis is then assisted by TXM tools applied to the corpus model. The pivot representation that gathers all annotations produced by annotation tools is figured as the node labeled « Pivot rep. » and the interpretation workflow itself is figured as a digital hermeneutic circle.

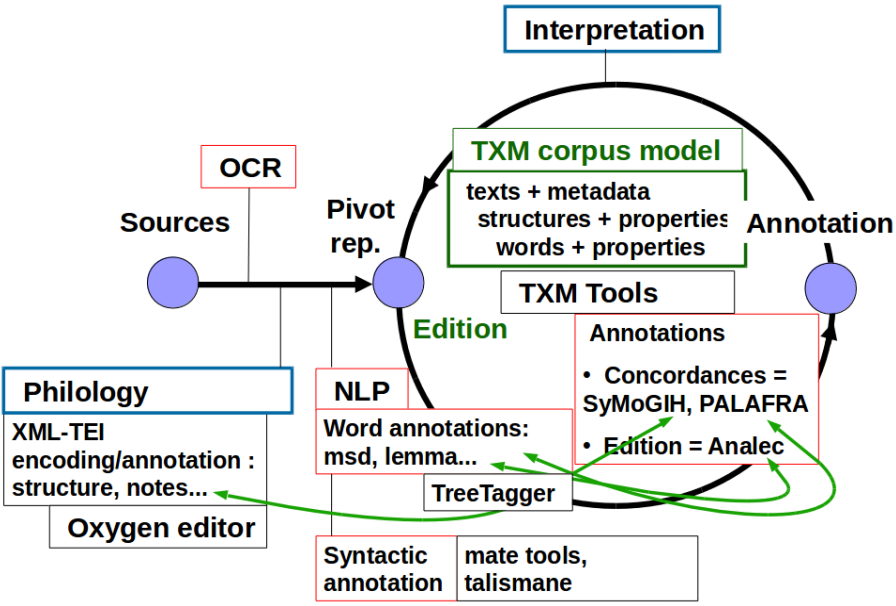


Figure 3: Digital hermeneutic circle integration into TXM.

Legend:
- red box = automatic annotation activity - black box = tool
- blue box = manual annotation activity - green box = TXM corpus data model
- purple disk = data representation - black arrow = activity
- green arrow = annotation equivalence

4. Conclusion

All the new annotation services integrated into TXM are building a comprehensive annotation-based digital text corpora analysis platform. From an epistemological point of view, the integration of different annotation models and tools into the platform should help its users to better define what comes from the source corpus they analyze and what comes from their own or from others interpretation work.

This work was funded by the ANR and the DFG under grant numbers ANR-15-CE38-0008 (DEMOCRAT project) and ANR-14-FRAL-0006 (PaLaFra project).

References

Beretta, F. (2015). Publishing and sharing historical data on the semantic web : the SyMoGIH project – symogih.org. Presented at the Workshop: Semantic Web Applications in the Humanities. Retrieved from https://halshs.archives-ouvertes.fr/halshs-01136533

Grobol, L., Landragin, F., & Heiden, S. (2017). Interoperable annotation of (co)references in the Democrat project. Presented at the Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation. Retrieved from https://hal.archives-ouvertes.fr/hal-01583527/document

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.), *24th Pacific Asia Conference on Language, Information and Computation* (pp. 389–398). Institute for Digital Enhancement of Cognitive Development, Waseda University. Retrieved from http://halshs.archives-ouvertes.fr/halshs-00549764/en/

Landragin, F., Poibeau, T., & Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data (pp. 357–362). Presented at the International Conference on Language Resources and Evaluation (LREC 2012). Retrieved from https://halshs.archives-ouvertes.fr/halshs-00698971/document

Schmid, H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision

Trees. In *Proceedings of the International Conference on New Methods in Language Processing* (Vol. 12).

Schnedecker, C., Glikman, J., & Landragin, F. (2017). Les chaînes de référence : annotation, application et questions théoriques. *Langue française*, (195), 5–16. https://doi.org/10.3917/lf.195.0005

TEI Consortium. (2017). TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium. Retrieved from http://www.tei-c.org/Guidelines/P5

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session posters* (p. 10). Senlis, France, France. Retrieved from https://hal.archives-ouvertes.fr/hal-01011969

Quantifying Translation : an analysis of the conditional perfect in English-French comparable-parallel corpus

Daniel Henkel
Université Paris 8 Vincennes St-Denis – dhenkel@univ-paris8.fr

Abstract
The frequency of the conditional perfect in English and French was observed in an 8-million-word corpus consisting of four 2-million-word comparable and parallel subcorpora, tagged by POS and lemma, and analyzed using regular expressions Intra-linguistically the Wilcoxon-Mann-Whitney test was used to compare authors and translators. Frequencies in source and target texts were evaluated using Spearman's correlation test to identify inter-linguistic influences. Overall, the past conditional in English was found to have a stronger influence in the translation process.

Résumé
La fréquence du conditionnel parfait en anglais et en français a été observée dans un corpus de 8 millions de mots comprenant quatre sous-corpus comparables et parallèles de 2 millions de mots chacun, étiquetés par catégorie grammaticale et par lemme, et analysés par expressions rationnelles (regex). Le test de Wilcoxon-Mann-Whitney a servi pour comparer les auteurs et traducteurs, tandis que la corrélation entre textes-sources et -cibles a été évaluée au moyen du coefficient de corrélation de Spearman. Globalement, l'influence du conditionnel parfait en anglais sur le processus traductionnel paraît plus sensible.

Keywords: corpus, translation, regular expressions, statistical analysis, Wilcoxon-Mann-Whitney, Spearman, conditional perfect

1. Introduction

Since Corpus-based Translation Studies (CBTS) first began to gain momentum around the turn of the 21st century, differences have consistently been shown between corpora of translated English, French and other languages in comparison with untranslated reference corpora in the same languages. The hybrid nature of translated texts is now thus widely

acknowledged as an established fact among specialists[1] in the field so much so that any further proof might seem superfluous. These studies have focused on phenomena such as the use of '*that*' to introduce subordinate clauses (Olohan & Baker, 2000), contractions (Olohan, 2003), manner-of-motion verbs (Cappelle, 2012), existential predications (Loock & Cappelle, 2013) most often in terms of their overall frequency[2]. Such comparisons have provided valuable insights about the languages involved and the translation process. Little consideration has been given so far, however, to the fact that each language-system consists of many individual styles or idiolects which gravitate around a common center, but individually exhibit widely differing characteristics. In other words, while the variation from one author or translator to another is inherent in the very nature of corpus linguistics, this dimension remains absent from the equation in many, if not most, corpus-based translation analyses.

2. Methods

Two important terminological distinctions must be made at the outset. The first is between *ex nihilo*, a.k.a. 'original', English (En0) and French (Fr0), i.e. discourse in each language produced independently of any known prior influence, as opposed to English-translated-from-French (EtrF) and French-translated-from-English (FtrE), which will be used to refer to translations into each language, based on a pre-existing work in the other language, and therefore potentially subject to inter-linguisitic influences. The second distinction is between two sorts of bilingual corpora, 'comparable' and 'parallel'. In keeping with the clarification offered by McEnery & Xiao (2007), the term 'comparable corpus' will hereafter refer to a bilingual corpus consisting of two subcorpora of *ex nihilo* English and French texts, which are therefore not translations of one another, but which share a certain number of common characteristics, whereas the term 'parallel corpus' will designate a

[1] Albeit with some divergence of opinion as to whether such differences are best interpreted as evidence of source-language interference or as consequences of the translation process regardless of the source-language, i.e. characteristics inherent in the 'third code' or 'translationese' (cf. Koppel & Ordan, 2011).

[2] Olohan (2002) apparently subscribes to Stubbs' (2001) view that "*corpus linguistics […] investigates relations between frequency and typicality, and instance and norm. It aims at a theory of the typical*," (while nonetheless encouraging investigation of individual translators' styles in her conclusion), and the predominance of this approach is confirmed again over a decade later by Loock (2013) who observes that "*many studies within the CBTS framework still solely rely on overall quantitative analyses to establish differences between original and translated languages.*"

corpus made up of one sub-corpus of *ex nihilo* works in a source-language and another sub-corpus consisting of the translations of those same works into the target-language.

The corpora used in this study were compiled from public domain works available in electronic format (.epub, .mobi, .html or .txt), the translations of which were also available in electronic format via publicly available sources (primarily Project Gutenberg). Common criteria[3] based on size and date were then used to select 20 works by 20 different authors in En0 and the same number in Fr0, so as to obtain, first of all, two reference sub-corpora comparable in terms of date, size, discourse type and diversity:

Table 1 Summary of characteristics for comparable En0 and Fr0 subcorpora.

	Subcorpus 1 En0 (*n*=20)	**Subcorpus 2 Fr0 (*n*=20)**
Word-counts[4]	**Max.** 199,976 (Collins, *The Moonstone*) **Min.** 59,771 (Mansfield, *The Garden-party*) **Median** 99,558 (Wells, *The War in the Air*) **Total** 2,114,517	**Max.** 192,521 (Zola, *Les trois villes Paris*) **Min.** 62,539 (Rolland, *Les précurseurs*) **Median** 90,873 (Leroux, *La chambre jaune*) **Total** 2,083,787
Dates	**Max.** 1928 (Woolf, *Orlando*) **Min.** 1868 (Collins, *The Moonstone*) **Median** 1901 (Kipling, *Kim*)	**Max.** 1921 (Leblanc, *Les dents du tigre*) **Min.** 1866 (Gaboriau, *L'affaire Lerouge*) **Median** 1901 (Bazin, *Les Oberlé*)

The translations of these works were then compiled into two sub-corpora of EtrF and FtrE, so as to produce an 8m-word 'super-corpus' consisting of four 2m-word sub-corpora, designed to be both comparable and parallel and thereby provide a basis for three types of comparisons:

– between En0 and Fr0, in order to establish benchmark data for each language,

– between EtrF and En0, so as to ascertain whether the linguistic indicator

[3] Whenever several works by the same author were available, preference was given either to the most recent or the one with the highest word-count. In general date was given precedence over size, except in cases where a major difference in word-count was found between works published within a relatively close interval.

[4] Word-counts were estimated using the text editor Geany, after replacing punctuation with whitespaces, given that punctuation has been found to artificially inflate word-counts in French as compared to English.

under investigation, i.e. the conditional perfect, has a similar distribution in EtrF compared to En0, and likewise for FtrE in comparison with Fr0,
– between source- and target-texts, to determine whether correlations exist between the parallel subcorpora (i.e. EtrF~Fr0 and FtrE~En0) which could be taken as evidence of interlinguistic interference.
All of the texts were cleaned of metatext, tagged for POS and Lemma in *TreeTagger*, and interrogated in *TextSTAT* using the following regular expressions to target the conditional perfect.
English (all verbs):
d) (((w|c|sh)ould)|('d)|(might)|(ought))(e?st)?/\S+(\S+/RB[RS]?/\S+)*(to/\S+)?(((ha|')ve|of)/\S+)(\S+/RB[RS]?/\S+)* \S+/V[BHV][ND]/
French (verbs taking AVOIR as an auxiliary, verbs taking ÊTRE, reflexive constructions):
e) \S+/VER:cond/avoir(\S+/ADV/\S+)* \S+/VER:pper
f) \S+/VER:cond/être(\S+/ADV/\S+)* \S+/VER:pper/(r[eé])?(aller|(ad|de|inter|par|pro|sur)?venir|rester|demeurer|(ap|dis)?paraître|naître|mourir|décéder|arriver|partir|tomber|monter|descendre|passer|rentrer|retourner|sortir)
g) ((je/\S+(\S+/ADV/\S+)* m[e']/\S+)|(tu/\S+(\S+/ADV/\S+)* t[e']/\S+)|(nous/\S+(\S+/ADV/\S+)* nous/\S+)|(vous/\S+(\S+/ADV/\S+)* vous/\S+)|(s[e']/\S+))(en|y/\S+)* \S+/VER:cond/être(\S+/ADV/\S+)* \S+/VER:pper/
The results obtained from these queries were converted into frequencies per 1000 words (freq./1k) for each author or translator and analyzed using the Wilcoxon-Mann-Whitney and Spearman tests as described in the following section.

3. Results and analysis

The data collected from each of the subcorpora are presented in the following tables and summarized in Fig. 1.

Table 2a Conditional perfect frequencies in En0

	Cond.Pf. (n=)	Words (n=)	Freq./1k		Cond.Pf. (n=)	Words (n=)	Freq./1k
Buchan	139	102022	1.36	**Lewis**	58	83799	0.69
Burnett	78	84093	0.93	**London**	57	100816	0.57
Collins	326	199976	1.63	**Mansfield**	67	59771	1.12
ConanDoyle	108	105040	1.03	**Reid**	200	94254	2.12
Cox	142	114352	1.24	**Stevenson**	81	70366	1.15
Eliot	319	164456	1.94	**Stoker**	127	161255	0.79

Hardy	254	153076	1.66	**Wallace**	135	101948	1.32
Hope	115	83189	1.38	**Wells**	54	99558	0.54
Joyce	26	69225	0.38	**Wilde**	76	79412	0.96
Kipling	109	107601	1.01	**Woolf**	76	80308	0.95

max: 2.12, **min**: 0.38, **median:** 1.03

Table 2b Conditional perfect frequencies in EtrF

	Cond.Pf. (n=)	Words (n=)	Freq./1k		Cond.Pf. (n=)	Words (n=)	Freq./1k
Tr.Barbusse	48	116179	0.41	**Tr.Leroux**	127	74920	1.7
Tr.Bazin	74	76312	0.97	**Tr.Loti**	15	65837	0.23
Tr.Benoît	41	64301	0.64	**Tr.Massenet**	42	57736	0.73
Tr.Flaubert	125	175678	0.71	**Tr.Maupassant**	45	76070	0.59
Tr.France	66	76830	0.86	**Tr.Mirbeau**	76	101959	0.75
Tr.Gaboriau	335	170870	1.96	**Tr.Proust**	408	198721	2.05
Tr.Gourmont	76	69399	1.1	**Tr.Rolland**	27	65872	0.41
Tr.Hugo	104	125428	0.83	**Tr.Vanderem**	80	95884	0.83
Tr.Huysmans	46	130181	0.35	**Tr.Verne**	89	63760	1.4
Tr.Leblanc	112	128493	0.87	**Tr.Zola**	179	205503	0.87

max: 2.05, **min**: 0.23, **median**: 0.83

Table 2c Conditional perfect frequencies in Fr0

	Cond.Pf. (n=)	Words (n=)	Freq./1k		Cond.Pf. (n=)	Words (n=)	Freq./1k
Barbusse	47	114877	0.41	**Leroux**	78	90873	0.86
Bazin	41	78395	0.52	**Loti**	15	72386	0.21
Benoît	33	67915	0.49	**Massenet**	45	76711	0.59
Flaubert	108	149808	0.72	**Maupassant**	46	75598	0.61
France	20	71998	0.28	**Mirbeau**	59	117035	0.5
Gaboriau	53	120464	0.44	**Proust**	296	170105	1.74
Gourmont	60	73000	0.82	**Rolland**	11	62539	0.18
Hugo	18	118095	0.15	**Vanderem**	44	91476	0.48
Huysmans	22	132824	0.17	**Verne**	50	76890	0.65
Leblanc	47	130277	0.36	**Zola**	141	192521	0.73

max: 1.74, **min**: 0.15, **median**: 0.5

Table 2d Conditional perfect frequencies in FtrE

	Cond.Pf. (n=)	Words (n=)	Freq./1k		Cond.Pf. (n=)	Words (n=)	Freq./1k
Tr.Buchan	69	105082	0.66	**Tr.Lewis**	80	96211	0.83
Tr.Burnett	74	80743	0.83	**Tr.London**	49	86378	0.57
Tr.Collins	138	198988	0.69	**Tr.Mansfield**	82	68674	1.19
Tr.ConanDoyle	119	117280	1.01	**Tr.Reid**	120	93025	1.29
Tr.Cox	194	130967	1.48	**Tr.Stevenson**	64	76757	0.83
Tr.Eliot	120	168125	0.71	**Tr.Stoker**	167	176623	0.95
Tr.Hardy	217	151435	1.43	**Tr.Wallace**	97	87316	1.11
Tr.Hope	99	82966	1.19	**Tr.Wells**	74	108529	0.68
Tr.Joyce	49	72739	0.67	**Tr.Wilde**	63	82430	0.76
Tr.Kipling	68	124885	0.54	**Tr.Woolf**	56	87475	0.64

max: 1.48, **min**: 0.54, **median**: 0.83

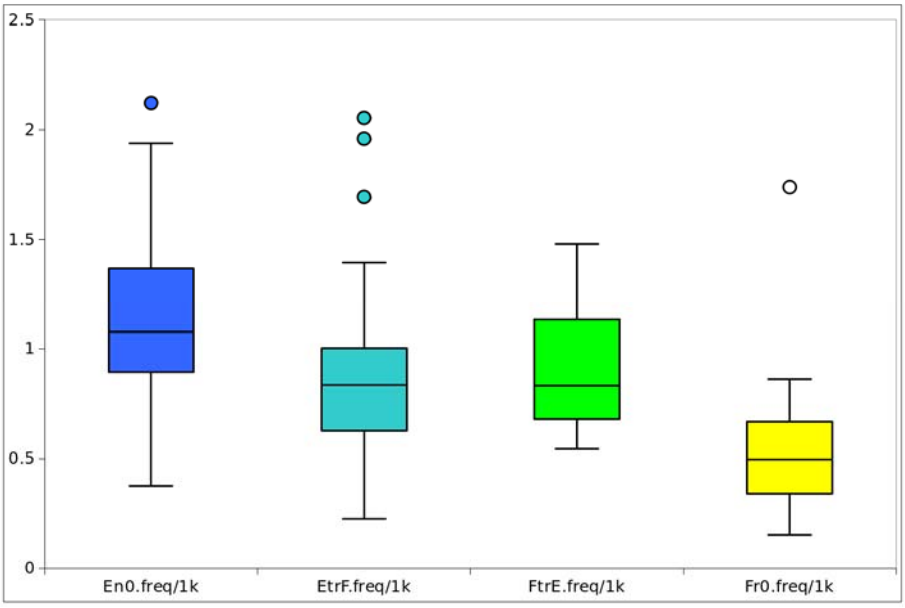


Fig. 1 Distributions of conditional perfect frequencies in En0, EtrF, FtrE and Fr0.

As is readily apparent from Fig. 1, the conditional perfect is used more frequently in En0 than in Fr0, which, aside from one extreme outlier (Proust), is situated below the 1st quartile of En0. EtrF and FtrE (as usual) occupy an

intermediate zone, with practically identical medians (0.83) which are both inferior to Q1 in En0 and superior to Q3 in Fr0. The most striking difference is between authors in Fr0 and translators, who use the conditional perfect almost twice as often in FtrE. As a result, the entire distribution in FtrE is superior to the median for Fr0, with 75% of FtrE (Q2-Q4) in the same range as the top quartile (Q4) of Fr0. Wilcoxon-Mann-Whitney confirms that a similar disparity could hardly occur by chance ($U=337$, $n_1=n_2=20$, $p=0.0002$) and that it is therefore reasonable to infer that – notwithstanding the considerable amount of variation that can be observed from one author or translator to another – FtrE and Fr0 are clearly different with respect to their use of the conditional perfect. Between EtrF and En0, however, the difference is less obvious. Although the interquartile range for EtrF (0.63-1) is noticeably lower than in En0 (0.9-1.37), there is nonetheless a great deal of overlap between the two distributions, and Wilcoxon-Mann-Whitney ($U=135$, $n_1=n_2=20$, $p=0.08$) indicates that the risk of error is too great to say with confidence whether any substantial difference exists between EtrF and En0 in their use of the conditional perfect.

To what extent such differences may be attributed to the influence of the analogous forms in the source-texts can be assessed statistically as illustrated in Fig. 2a and 2b:

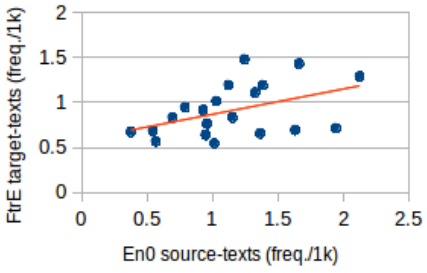


Fig. 2a Frequency of conditional perfect forms in FtrE vs. En0. (ρ=0.47, p=0.036)

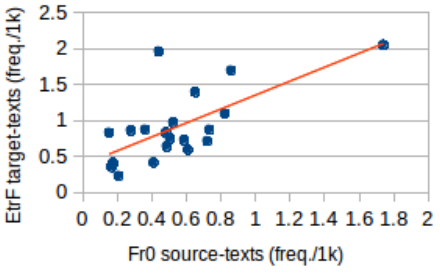


Fig. 2b Frequency of conditional perfect forms in EtrF vs. Fr0. (ρ=0.57, p=0.009)

In both cases, Spearman's[5] correlation test reveals a statistically significant ($p<0.05$) positive correlation ($\varrho=0.57$ for EtrF/Fr0, $\varrho=0.47$ for FtrE/En0) of moderate strength, which somewhat unexpectedly obtains a higher score for

[5] Spearman's was preferred due to the presence of outliers. Pearson's R yields an almost identical result for FtrE/En0, and a somewhat stronger coefficient ($r=0.67$) for EtrF/Fr0, with similar p-values in both cases.

EtrF/Fr0. These correlations of similar strength suggest an intuitively plausible tendency to translate individual instances of the conditional perfect in one language by the analogous form in the other language in both directions and in roughly similar proportions (although this remains to be verified by manual examination of translation segments). Such a hypothesis would help to explain why the medians and interquartile ranges observed in EtrF and FtrE occupy a middle zone between En0 and Fr0, but it does little to account for the greater disparity between FtrE and Fr0 as opposed to EtrF and En0. Other contextual parameters may well be involved, or perhaps the higher frequency of the conditional past in En0 exerts a sort of subliminal effect on translators, who then use it more freely in FtrE with or without a syntactic counterpart in the corresponding En0 segment.

4. Conclusion

These findings demonstrate how quantitative analysis of translated parallel corpora in comparison with untranslated comparable corpora, can be used both to identify disparities between target-texts and the target-language as represented in an *ex nihilo* corpus, and to assess the influence of the source-texts on the target-texts. Such relationships are often asymmetrical: in this case the correlation between the original French conditional perfect and the translations into EtrF is stronger, while the higher frequency of conditional perfect forms in English, though less strongly correlated on a text-to-text basis, nonetheless fosters a style of French-translated-from-English which is markedly different from *ex nihilo* French. While the exact mechanisms involved will require further investigation, the conditional perfect in English appears to exert a stronger influence in the translation process than the corresponding form in French.

References

Hu K. (2016). *Introducing corpus-based translation studies*. Springer.

Koppel M and Ordan N. (2011). Translationese and Its Dialects *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 1318–1326, June 19-24, 2011

Kruger A., Wallmach, K. and Munday J. (Eds.). (2011). *Corpus-based translation studies: Research and applications*. Bloomsbury Publishing.

Loock R. (2013). Close encounters of the third code. In Lefer M.A. and Vogeleer S., eds, *Interference and normalization in genre-controlled multilingual corpora, Belgian Journal of Linguistics* 27: 61-86

Olohan M. (2002). Comparable corpora in translation research. In *LREC Language Resources in Translation Work and Research Workshop Proceedings* pp. 5-9.

Zanettin F. (2013). Corpus methods for descriptive translation studies. *Procedia-Social and Behavioral Sciences*, 95, 20-32.

Hüning Matthias. TextSTAT 2.9c © 2000/2014 Niederländische Philologie, Freie Universität Berlin, http://neon.niederlandistik.fu-berlin.de/en/textstat/

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL https://www.R-project.org/.

Schmid H. TreeTagger, Universitaet Stuttgart, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Extraction of lexical repetitive expressions from complete works of William Shakespeare

Daniel Devatman Hromada
Univesität der Künste, Berlin, Germany – daniel at udk dash berlin dot de

Abstract
Rhetoric tradition has canonized dozens of repetition-involving figures of speech. Our article shows a way how hitherto ignored repetition-involving schemata can be identified by means of translation of so-called "entangled numbers" into backreferencing regular expressions. Each regex is subsequently exposed to all utterances in all works of William Shakespeare, allowing us to pinpoint 3367 instances of 172 distinct repetitive schemata.

Keywords: rhetoric stylometry, figures of speech, repetition, chiasm, entangled numbers, regular expressions, William Shakespeare, non-zipfian distribution

Résumé
On montre, comment peut-on identifier les figures de styles jusqu'ici inconnues. Le but en question est atteint grâce au fait qu'on peut concevoir un certain groupe de figures de style tel un nombre ayant quelques propriétés particulières. Une fois découverte et énumérés, on peut transcrire ces nombres en expressions régulières qui peuvent ensuite être éxposé à un corpus textuel. Dans le cas de notre étude préliminaire, il s'agissait du corpus de William Shakespeare.

Mots clés: stylométrie rhétorique, rfigures de style, répétition, chiasme, expressions régulières, répetition, William Shakespeare

1. Introduction

Masterpieces of litterature and drama abound with repetitions. Rhetorics abounds with repetitions, succesful oratories abound with repetitions. Many a schema and a figure exists which exploits repetition : e.g. a polysyndeton and an anaphore, an anadiplose and an epistrophe, a symploche and an antanaclasis, paranomasis and an antimetabole. And alliterations and paregmenons, and polypoptons, epizeuxiae or even a good old psittacism ?
Many are such schemata, many are such figures. Woe to the one who thinking he knows them all !

Our article presents a way of enumerating of many a new schemata involving one or more repetition of one or more lexical signifiants. The procedure starts with a theoretical insight, that at least certain subset of the set of all such schemata, is easily enumerable. This insight is subsequently transcribed into an algorithm enumerating natural numbers which satisfy following properties. These numbers once identified, they are to be translated into Perl Compatible Regular Expressions exploiting some back-references and negative lookaheads.

1.1. Computational rhetorics and its roots

In literature studies it is fairly common to speak about so-called "rhyme schemes" like AAAA for monorhymes, ABAB for alternate rhyme, ABBA for enclosed rhymes etc.

It is therefore barely surprising that analogic formalisms - that is, formalisms that involve alphabetic indices - have been adopted by scholars aiming to formalize a subgroup of rhetoric figures, known as the group of schemes. For example (Harris et DiMarco, 2009) use a following formalism:

[W]a::: [W]b::: [W]b::: [W]a

to denote the rhetoric figure known as antimetabole. Subsequent studies in automatized chiasm identification pursue a similiar route and often use formulae like ABXBA, ABCBA, ABCXCBA to denote schemata corresponding to utterances such as: "*Drake love loons. Loons love Drake.*", "*All as one. One as all.*" (Hromada, 2011) or "I*n prehistoric times women resembled men, and men resembled women.*" (Dubremetz & Nivre, 2015) .

Table 1: 14 lowest E-numbers, their corresponding alphabetic representations and some corresponding Shakespearean expressions .

E-number	**Alphabetic**	**Example expression**
11	AA	"we split we split " [1]
111	AAA	"we split we split we split "
1111	AAAA	"justice justice justice justice "
1122	AABB	"gross gross fat fat "
1212	ABAB	"to prayers to prayers "

[1] Note that sometimes one single word is attributed the role of a distinct « brick » , sometimes a concatenation of two or even more words assumes such a role. As will be indicated in sections two and three, this behaviour is not a bug, but an anticipated property of our method.

1221	ABBA	"my hearts cheerly cheerly my hearts "
11111	AAAAA	"so so so so so "
11122	AAABB	"great great great pompey pompey "
11212	AABAB	"come come buy come buy "
11221	AABBA	"high day high day freedom freedom high day "
11222	AABBB	"o night o night alack alack alack "
12112	ABAAB	"too vain too too vain "
12121	ABABA	"come hither come hither come "
12122	ABABB	"come buy come buy buy "

1.2. Entangled numbers

The set of entangled numbers (or E-numbers) is a subset of a set of natural numbers (i.e. integers). Entangled numbers are defined as *"words of length n over an alphabet of size 9 that are in standard order and which have the property that every letter that appears in the word is repeated. "* (OEIS, 2016)

Note that the term word, as used in the preceding, as well as in following citations, is used in mathematician's sense, meaning something as « sequence of symbols » : *"A word is in "standard order" if it has the property that whenever a letter i appears, the letter i-1 has already appeared in the word. This implies that all words begin with the letter 1."* (Arndt et Sloane, 2016). Hence, numbers like 22 or 33 are not entangled numbers because they are not in "standard order" and numbers like "12" or "121 "are not entangled because some (or all) of their digits are not repeated. Fourteen smallest (i.e. with the lowest numeric value) entangled numbers and their corresponding alphabetic transcriptions are enumerated in Table 1.

Given that entangled numbers are natural numbers, they can be easily **enumerated** by an incremental algorithm starting at one and iterating towards infinity. Once enumerated (OEIS, 2016), we can bridge the realm of numbers with the realm of text and apply our method.

2. Method

The core idea behind our method can be stated as follows:

Any E-number can be "translated" into a backreference-endowed regular expression.

Concretely speaking, every digit of an E- number can be interpreted as an element or a "brick". In this article, we work only with one type of bricks, those corresponding to sequences which are between two to twenty-three

characters long [2]. Such sequences can correspond to one or multiple lexical units. A first occurence of a novel brick can be represented as a PERL-compatible regular expression (Friedl, 2002 ; Aho, 2014):

(.{2,23})

However, any subsequent repeated occurence of a digit in an E- number is interpreted not as an occurence of the new brick, but rather as a backreference to the brick which was already denoted by the same digit. The very first E- number 11 is therefore NOT to be translated into regex /(.{2,23}) (.{2,23})/. For this would imply existence of two distinct bricks. Rather, the E-number 11 is to be translated into regex:

(.{2,23}) **\1**

wherein the expression **\1** denotes the backreference to the content matched by the regex-brick specified in first parentheses, i.e. brick no.1 .
Hence, the E-number 111 can be easily translated into a regex /(.{2,23}) \1 \1/, 1111 into a regex /(.{2,23}) \1 \1 \1/ etc.
What's more, when we combine the backreference with a negative lookahead operator – traditionally expressed by the formula (?!) - we can make sure that a so-called non-identity principle is also satisfied. That is :

"**Each distinct digit corresponds to distinct content**"

For example, by translating the E-number 121 into the regex

(.{2,23}) (?!\1)(.{2,23}) \1

we can make sure that the content matched by the brick denoted by digit 2 shall be different from the content matched by the brick denoted by digit 1. Thus, a phrase "no no no" shall not be matched by such a regex while an expression "no yes no" shall.
Going somewhat further, an E-number 12321 - which could be understood as an instance of chiasm or antimetabole ABXBA - is to be translated into regex

(.{2,23}) (?!\1)(.{2,23}) (?!\1\2)(.{2,23}) \2 \1

whereby the disjunctive backreference contained in the negative lookahead

[2] These are the only variable parameters of our method.

(?!\1\2) assures that the content matched brick no.3 - corresponing to filler X - shall be different from content matched by the brick representing digit 1 as well as the brick representing digit 2.

3. Corpus & Processing

A digital, unicode-encoded version of Craig's edition of "Complete works of William Shakespeare" has been downloaded from a publicly available Internet source[3] . This corpus contains 17 txt files stored in the sub-folder "comedies", 10 txt files stored in the sub-folder "tragedies" and 10 txt files stored in the sub-folder "historical".

Texts were subsequently split into utterances by interpreting closing tags (e.g. </PERSONA>, </MIRANDA> etc.) as utterance separator. Even more concretely, one can simply consider the slash symbol / to be the utterance separator.

Only two further text-processing steps have been executed during the initialization phase of the experiment hereby presented. Primo, content of each utterance has been put into lowercase. Secundo, non-alphabetic symbols (e.g. dot, comma, exclamation mark etc.) have been replaced by blank spaces. We are aware that such replacement could potentially lead to certain amount of loss of prosody- or pathos- encoding information. However, we consider this step as legitimate because the objective of our experiment was to focus on repetition of lexical units[4].

Pre-processing code once executed, identification of expressions containing diverse types of lexical repetitions is as simple as matching each Shakespearean utterance with each regex.

4. Results

All in all, 3667 instances of a repetitive expressions have been detected in Shakespeare's complete works. These were contained in 2295 distinct utterances and corresponded to 172 distinct schemata. Among these, 71 matched more than one instance: these schemata could thus potentially correspond to a certain cognitive pattern or a habitus in Shakespeare's mind. Table 2 contains summary information concerning 23 schemata matching at least five distinct utterances.

---

[3] http://www.lexically.net/downloads/corpus_linguistics/ShakespearePlaysPlus.zip

[4] Regexes matching repetitions of phonotactic clusters, syllables, or phrases, are also possible. We prefer, however, not to focus on this topic within the limited scope of this conference proposal.

Table 2: Repetitive schemata matching at least 23 distinct utterances present in collected works of William Shakespeare.

Instances	E-number	Example
2332	11	"bestir bestir "
525	1212	"to prayers to prayers "
170	111	"ha ha ha "
100	123123	"cover thy head cover thy head "
48	12121	"come hither come hither come "
35	1221	"fond done done fond"
32	12341234	"let him roar again let him roar again "
32	1122	"with her with her hook on hook on "
30	1111	"great great great great "
23	121212	"come on come on come on "

Another phenomenon may be found noteworthy by a reader interested in purely quantitative aspects of our research. It concerns the relation between the length of the E-number (i.e. the amount of corresponding bricks) and the number of utterances matched by such numbers. In case of trivial repetitions, this relation seems to be plainly Zipfian. For example : Shakespeare's dramas seem contain 2332 duplications (e.g. E=11), 170 triplications (E=111), 30 tetraplications (E=1111), 8 pentaplications (E=11111) two hexaplications (E=111111), one heptaplication (E=1111111) and zero octaplications.

Table 3: Comparison of frequencies of occurrence of schemata of certain length and amount

Digits	Theoretical	Matched
2	1	2332
3	1	170
4	4	622
5	11	91
6	41	211
7	162	56
8	715	86
9	3425	67

It is worth mentioning, however, that generic relation between the length (in digits) of an and the amount of utterances which matches seems not to be Zipfian. As indicated by Table 3, an observed preference for repetitive expressions including two, four, six or eight bricks cannot be explained in terms of number-theoretical distribution of E-numbers themselves.
For example, there exists eleven E-numbers with five digits and fourty-one E-numbers of length six. However, when exposed to Shakespeare corpus, regexes generated from six digits long seem to match 211 utterances while five brick long regexes match only ninety-one of them. Whether this

observed asymmetry is an artefact of our method or whether it is due to *a sort of cognitive bias, a sort of preference for balanced repetitions* within the Poet's mind poses us in front of an argument which we do not dare to tackle here.

4. Conclusion

Insight that certain class of repetition-based schemata can be enumerated allows us to generate myriads hitherto unseen Perl Compatible Regular Expressions[5] which involve back-references and negative lookaheads.

In the end, such regexes have been exposed to corpus containing collected works of William Shakespeare.

Matching all utterances with all regexes generated out of all 4360 E-numbers with less than 10 digits lasted 9555 seconds in case Shakespearean comedies, 6607 seconds in case of tragedies and 6900 seconds in case of historical dramata. All this on one single core of an 1.4 GHz CPU.

This approach allowed us to pinpoint 3667[6] utterances matching at least one among 172 distinct repetitive schemata. 23 among these schemata matched at least 5 distinct utterances, 71 among them matched at least two utterances. This may potentially point to a sort of neurolinguistic habit residing in the opaque sphere between the syntactic and lexical layers.

We believe that at least some among these «figures » could be of certain interest not only for scholars trying to understand inner intricacies of Shakespeare's genius, but also to address more generic topics in fields as distinct as digital humanities, computational rhetorics, discourse stylometry or even more general cognitive sciences.

References

Aho, A. V. (2014). Algorithms for finding patterns in strings. Algorithms and Complexity, 1:255.

Arndt, J., Sloane, N. J. A. (2016). Counting words that are in "standard order". The on-line encyclopedia of integer sequences. https://oeis.org/A278984/a278984.txt.

Dubremetz, M., Nivre, J. (2015). Rhetorical figure detection: the case of chiasmus. On Computational Linguistics for Literature, page 23.

[5] We remind the reader that PCREs are much more powerful than so-called regular grammars. For example, regular grammars are unable to backreference, while for PCREs, backreferencing is a completely legal act.

[6] See https://refused.science/rhethorics/shakespeare-regex/matches.csv (Licenced under CC BY-NC-SA) for list of all matched utterances, including the information about the respective entangled numbers, theater pieces, genres (comedy / tragedy / drama) and the dramatis personae.

Friedl, J. E. F. (2002). Mastering regular expressions. O'Reilly Media, Inc.

Harris, R., DiMarco Ch. (2009). Constructing a rhetorical figuration ontology. In Persuasive Technology and Digital Behaviour Intervention Symposium, pages 47–52. Citeseer.

Hromada, D. D. (2011). Initial experiments with multilingual extraction of rhetoric figures by means of PERL-compatible regular expressions. In RANLP Student Research Workshop, pages 85–90.

OEIS (2016). List of words of length n over an alphabet of size 9 that are in standard order and which have the property that every letter is repeated at least once. https://oeis.org/A273978

Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale)

Olivier Kraif, Julie Sorba
Univ. Grenoble Alpes, LIDILEM
olivier.kraif@univ-grenoble-alpes.fr; julie.sorba@univ-grenoble-alpes.fr

Abstract
In this paper, we aim to test if the classifications of the phraseological units based on recurring trees and ngram methods are functional in order to separate novel genres one from another. Our results confirm that these two methods are relevant for the expressions relative to space and time into our corpora.

Résumé
Notre objectif est de tester les classifications des phraséologismes, opérées par les méthodes des ALR et des SR, dans le but de distinguer des sous-genres romanesques les uns des autres. Dans nos corpus, nos résultats confirment la pertinence de ces classifications pour les deux champs de l'espace et du temps.

Keywords: ngram, recurring trees, novel genres, phraseology

1. Introduction

Notre étude, qui s'inscrit dans le cadre de l'analyse exploratoire des données textuelles, concerne des romans français contemporains rassemblés dans le cadre du projet ANR-DFG PhraseoRom. Ce corpus (plus de 110 millions de mots pour le français) est partitionné en plusieurs sous-corpus correspondant à différents sous-genres littéraires (policier, science-fiction, fantasy, roman historique, roman sentimental, littérature générale). Notre objectif est de caractériser ces genres et sous-genres textuels par les unités phraséologiques spécifiques qu'ils contiennent. À l'instar de Boyer, nous postulons que « chaque genre comprend un certain nombre de sous-ensembles, des séries fondées sur la réutilisation de composantes identiques » (1992, p.91). Dans la mesure où la phraséologie étendue s'intéresse à tout ce qui est « préfabriqué » dans les séquences lexicales, elle constitue donc un point d'entrée privilégié pour mettre en évidence ces « séries ».
Pour cette étude, nous retenons spécifiquement 4 sous-genres : les romans de

science-fiction (SF), les romans policiers (POL), les romans historiques (HIST) et les romans de littérature dite blanche ou générale (GEN). La fouille des textes utilise la technique de repérage des Arbres Lexicosyntaxiques Récurrents (ou ALR, Kraif & Diwersy, 2012 ; Kraif, 2016) dont la validité a déjà été montrée par le repérage d'unités phraséologiques spécifiques dans les textes scientifiques (Tutin & Kraif, 2016). Nous proposons en outre de comparer ici cette technique d'extraction avec celle des segments répétés (Salem, 1987), les ALR ayant montré une meilleure prise en compte de la variabilité syntaxique pour le repérage des routines, mais s'avérant parfois défaillants pour identifier des segments figés en surface, du fait du modèle dépendanciel employé.

Dans des travaux antérieurs, nous avons montré comment les ALR permettaient de repérer des motifs récurrents construits autour d'expressions spécifiques fortement liées à la composante thématique des sous-genres en question : c'était le cas pour « scène de crime » dans POL (Kraif, Novakova & Sorba, 2016). Ici, nous nous concentrons sur des expressions moins directement liées aux univers de référence des sous-genres (le crime, l'amour, la science, etc.), afin de mettre en évidence des traits moins prévisibles. C'est pourquoi, nous avons choisi de sélectionner les séquences – bien souvent adverbiales – liées à l'expression du temps et de l'espace.

Nous allons désormais présenter les résultats obtenus dans des travaux antérieurs (partie 2), puis décrire notre méthodologie expérimentale (partie 3). Enfin, nous exposerons et discuterons nos observations (partie 4) avant de proposer des conclusions et perspectives à notre étude (partie 5).

2. Travaux antérieurs

Lefer, Bestgen & Grabar (2016) s'appuient sur une extraction de n-grammes de 2 à 4 mots pour caractériser 3 genres textuels : des débats parlementaires européens, des éditoriaux de presse et des articles scientifiques. Ces auteurs utilisent une méthode d'AFC pour identifier les expressions les plus typiques et en tirent des observations contrastives concernant l'expression de la certitude et de l'opinion. De notre côté, nous avons analysé des contrastes génériques sur un plan qualitatif, en identifiant des ALR dans des corpus de romans policiers et de science-fiction, en nous fondant sur des mesures de spécificité (Kraif, Novakova & Sorba, 2016). Nous avons également utilisé l'extraction des ALR pour classer automatiquement, dans une approche supervisée, des sous-corpus POL, SF et GEN (Chambre & Kraif, 2017). Ces travaux préliminaires ont montré que les ALR donnaient de meilleurs résultats que les autres catégories de traits (ponctuation, morphosyntaxe, lexique), et permettaient de classer correctement 98% des textes du corpus à partir d'une sélection de traits discriminants. La plupart de ces traits

appartenaient à des champs lexicaux précis, liés aux univers de référence propres à chaque sous-genre, comme ceux du 'téléphone' (*le numéro de portable, passer un coup de fil, etc.*) ou de la 'voiture' (*à travers le pare-brise, démarrer en trombe, etc.*) pour POL. De plus, des expressions temporelles (p.ex. pour POL *à huit heures, vingt et une heure, au bout de X minutes*) et des indications spatiales très variées (p.ex. pour SF *par la voie, dans le territoire, dans la sphère, dans l'espace, la zone de*) ont été mises en évidence.
Nous proposons ici un prolongement de cette expérimentation, d'une part, en étudiant les expressions spatiales et temporelles, et d'autre part, en ajoutant le sous-genre des romans historiques (HIST), afin de déterminer si ces classes d'expression sont suffisantes pour différencier les quatre sous-genres (POL, SF, GEN, HIST).

3. Méthodologie

Pour chaque sous-genre, notre corpus comporte un échantillon d'environ 8 millions de mots, correspondant à environ 70 œuvres d'une quarantaine d'auteurs (cf. Tableau 1). Ces œuvres sont toutes postérieures à 1950, et la majorité d'entre elles ont été publiées pour la première fois après 2000. La classification des œuvres en genre a été effectuée a priori selon des critères éditoriaux, en fonction des collections de publication.

	Auteurs	Romans	Taille
POL	46	69	8 008 395
SF	36	75	8 001 582
HIST	38	70	8 015 933
GEN	46	69	8 008 395

Tableau 1 : Constitution du corpus

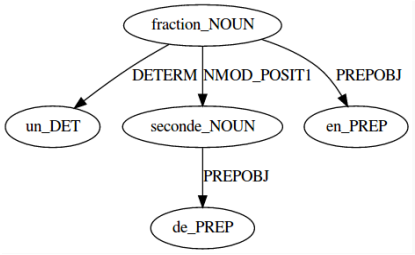


Figure 1 : ALR représentant l'expression en une fraction de seconde

Pour identifier les expressions phraséologiques caractéristiques des différents sous-genres, nous utilisons deux méthodes de repérage :
- la méthode des ALR : nos corpus étant analysés en dépendances avec XIP

(Aït-Mokhtar et al., 2002), ces ALR sont des sous-arbres respectant des critères de fréquence (ici ≥ 10 occurrences), de dispersion (ici ≥ 10 auteurs différents, appartenant à au moins 3 sous-genres différents) et de taille (ici ≥ 3 nœuds et ≤ 8 nœuds). En outre, lors de la recherche de ces ALR, une mesure d'association est calculée afin de ne retenir que les nœuds significativement associés avec le reste de l'arbre. La figure 1 montre un exemple d'ALR correspondant à l'expression *en une fraction de seconde*.

- la méthode des segments répétés (ou SR, Salem, 1987) : nous avons appliqué les mêmes critères de dispersion et de taille (≥ 3 et ≤ 8), afin de comparer les deux méthodes *in fine*. Les SR sont constitués de séquences de lemmes (obtenus avec XIP), et non de formes fléchies. Cette dernière méthode est plus simple à mettre en œuvre et nécessite peu de ressources linguistiques, bien qu'elle pose des problèmes d'explosion combinatoire (cf. partie 4).

Dans un second temps, nous appliquons un filtrage par mots-clés afin de ne retenir que les séquences liées aux deux sous-domaines étudiés, à savoir l'expression du temps et de l'espace. Les mots-clés pour l'espace sont des noms de lieux, d'espaces naturels, de description géographique, des mesures de distance, des adverbes de lieu, sélectionnés après un premier sondage des ALR extraits :

- Mots-clés ESPACE : cave, salon, hôpital, immeuble, bâtiment, camp, restaurant, village, route, rue, quai, chaussée, terrasse, ministère, parc, bureau, carlingue, maison, toit, chambre, hôtel, palais, rez-de-chaussée, entrée, pont, escalier, chemin, place, salle, jardin, seuil, cour, couloir, colline, sentier, sol, rive, rivage, plage, rivière, mont, montagne, mer, océan, lac, bois, forêt, espace, endroit, coin, pays, continent, frontière, direction, cap, sud, est, nord, ouest, confins, mètre, kilomètre, année-lumière, hectare, acre, loin, proche, près de, au bord de, orée, distance.

Les mots-clés pour le temps désignent des moments de la journée et de l'année, des unités de mesure et des découpages conventionnels de période (noms, adverbes et locutions adverbiales) :

- Mots-clés TEMPS : matin, soir, soirée, après-midi, nuit, jour, temps, fois, moment, instant, toujours, jamais, parfois, souvent, autrefois, jadis, tôt, tard, longtemps, brièvement, immédiatement, subitement, tout à coup, tout de suite, aujourd'hui, demain, hier, lendemain, maintenant, heure, minute, seconde, journée, semaine, mois, an, année, décennie, siècle, millénaire, printemps, été, automne, hiver.

Ces listes ne prétendent pas être exhaustives et le filtrage opéré produit à la fois du silence et du bruit, du fait des ambiguïtés. Celles-ci demeurent toutefois marginales (d'après un sondage manuel, le bruit est inférieur à 10 %).

Pour identifier les ensembles de traits pertinents du point de vue des sous-genres, nous injectons ces expressions (ALR ou SR) dans un système de classification automatique. De la sorte, nous visons un double objectif : d'une

part, vérifier que nos classes constituées a priori sont cohérentes et corrélées à des critères objectivables ; d'autre part, identifier ces critères sous la forme d'ensemble de traits discriminants pour la classification.

4. Résultats et discussion

Dans une première étape, nous avons extrait les 6000 ALR les plus fréquents sur l'ensemble du corpus. En effectuant une classification sur ces traits, avec un modèle SVM optimisé par SMO (avec la plate-forme Weka, Eide et al. 2016), on obtient, dans une évaluation croisée à 10 plis, une précision de 74 % (123 sur 166), avec un Kappa de 0,65, ce qui correspond à un très bon accord avec la classification de référence. La matrice de confusion (cf. Tableau 2) montre que les deux genres les mieux classés sont SF (93,1 %) et POL (79,5 %). Le genre GEN obtient la précision la plus faible (64%) avec des confusions fréquentes avec POL et HIST ; HIST est de son côté fréquemment confondu avec GEN.

L'examen des ALR les plus discriminants montre, comme on pouvait s'y attendre, la forte présence de certains thèmes dans POL, HIST et SF (la voiture, le crime, le téléphone pour POL ; la guerre, la religion pour HIST ; l'univers spatial et les artefacts technologiques pour SF) et l'absence de traits saillants dans GEN.

4.1 Sélection des traits TEMPS+ESPACE

Lorsqu'on sélectionne les traits liés à l'expression du temps seul (environ un millier), on obtient une dégradation par rapport aux résultats précédents, avec une précision globale de 48,8 % et un Kappa de 0,31 signifiant un accord faible entre la classification a priori et la classification automatique. Les expressions spatiales, de leur côté (on en obtient 1560, mais nous avons retenu les 1000 plus fréquentes afin de disposer de résultats comparables), obtiennent des résultats un peu meilleurs, toutefois moins bons que les traits non filtrés : la précision est de 59,6 %, avec un Kappa de 0,46 correspondant à un accord modéré.

Quand on sélectionne conjointement les ALR de TEMPS et ESPACE, on obtient une légère amélioration par rapport à la classification avec ESPACE seul : 61,4 % (102 instances bien classées sur 166), avec un Kappa assez bon de 0,48. La matrice de confusion (cf. tableau 2) montre que POL obtient la meilleure précision (69%) et GEN la moins bonne (55,9 %).

Si on sélectionne les traits les plus discriminants (attributs *SfcSubsetEval* avec méthode *BestFirst* dans Weka), on obtient un ensemble de 54 attributs. On peut évaluer, de manière indicative, le pouvoir classificateur de ces attributs sur notre corpus en les réinjectant dans une classification par SMO : on obtient alors une précision globale très légèrement supérieure (62 %), mais il

est intéressant de noter que les genres marqués POL, SF et HIST sont très bien classés sur la base de ces traits (précision de 85,7% pour HIST, 84 % pour SF, 75,7 % pour POL) avec une dégradation forte pour GEN (43,4%), comme le montre la matrice de confusion ci-dessous (tableau 2).

Tableau 1 : Matrices de confusion pour les classifications avec (1) tous les traits, (2) les ALR filtrés (TEMPS+ESPACE) et (3) les ALR sélectionnés

	(1) Tous les traits (6000 ALR plus fréquents)				(2) TEMPS+ESPACE (2571 traits filtrés)				(3) TEMPS+ESPACE Sélection de 54 traits			
	SF	POL	GEN	HIST	SF	POL	GEN	HIST	SF	POL	GEN	HIST
SF	27	2	2	5	18	5	6	7	21	2	13	0
POL	1	35	9	1	5	29	12	0	3	28	15	0
GEN	1	5	32	8	3	3	33	7	1	6	36	3
HIST	0	2	7	29	3	5	8	22	0	1	19	18

L'examen détaillé des 54 traits sélectionnés révèle plusieurs points saillants :
- d'une manière générale, les ALR relatifs à l'espace sont très largement majoritaires avec 33/54 contre 17/54 pour le temps, après élimination du bruit (4/54).
- si on considère les traits spécifiques à HIST, les expressions spatiales désignent surtout des lieux de pouvoir (*la place forte, de son palais, salle du palais, salle du château, pénétrer dans la grande salle*) et la mer (*sur la mer, de la mer*), tandis que les expressions temporelles font référence à une temporalité longue (*au bout de quelques mois*, *règne de X années*, *avoir le temps*) et à des datations absolues ou relative (*du N^e^ siècle, venir le lendemain, à trois heures de l'après-midi*).
- pour POL, en revanche, les expressions temporelles indiquent des datations horaires (*à 8 heures, 21 heures*) et des durées courtes (*une vingtaine de secondes*). Les expressions spatiales, nombreuses, indiquent des pièces et des espaces intérieurs (*de la salle de bain, vers la salle de bain, entrer dans le bureau, vers le bureau, dans le coin*), des lieux urbains (*aller à l'hôtel, passer à l'hôpital, à l'hôpital*), et des localisations vagues (*dans le coin* au sens de « dans les parages »).
- pour SF, les expressions temporelles sont plus nombreuses (7/18) que dans les autres sous-genres. Elles font référence à des durées extrêmes par leur longueur (*milliers d'années, de mille ans*) ou leur brièveté (*une fraction de seconde, un centième de seconde*). Pour l'espace, on trouve des expressions de distances chiffrées *(dizaines de mètres, centaine de mètres, plusieurs centaines de mètres*), des références attendues à l'espace intersidéral (*dans l'espace, à travers l'espace, être dans l'espace, voyager dans l'espace, flotter dans l'espace*), à l'*espace-*

temps et des expressions avec *sol* (*sur le sol, sous-sol*).
- pour GEN : la seule expression spécifique apparaissant dans les traits sélectionnés est *chemin de traverse*.

4.2 Comparaison avec les segments répétés

Nous n'avons pas réussi à extraire la totalité des SR de 3 à 8 mots pour l'ensemble du corpus, du fait des problèmes d'explosion combinatoire (environ 40 000 000 SR générés pour 100 textes du corpus). Nous avons donc retenu les SR contenant les mots-clés sélectionnés pour TEMPS et ESPACE, en conservant les 1000 SR les plus fréquents afin d'avoir des ensembles de traits comparables aux ALR filtrés. On obtient de meilleurs résultats que pour les ALR, avec une précision de 66,7 % pour ESPACE et 58,3 % pour TEMPS contre respectivement 59,6 % et 48,8 %. Pour TEMPS+ESPACE, on constate une certaine dégradation, avec une précision qui tombe à 64,1 %. À ce stade de nos observations, il nous est difficile d'interpréter ces résultats quantitatifs car la sélection du meilleur ensemble de traits pour ESPACE donne peu ou prou les mêmes expressions qu'avec les ALR :
le chambre de, le cour de, à le cour, dans le espace, le salle de bain, de le espace, dans son bureau, de le immeuble, le maison et, à le hôtel de, centaine de mètre, sur le bureau, sur le place de, le palais de, dans le grand salle, de bureau de, de le salle de bain, sur son bureau, cour de France, en route pour, dans mon bureau, dans tout le direction, un dizaine de mètre, de son pays, à le rue, dans le sous-sol, quitter le salle, dans un restaurant, sur le rivage, mètre plus bas, vers le bureau, route vers le, dizaine de mètre de, un kilomètre de, à ministère de, dans le espace et, de un montagne, le espace et le.
Les deux méthodes donnent donc des résultats convergents en termes qualitatifs en extrayant les mêmes expressions. Néanmoins, des investigations complémentaires seront nécessaires pour interpréter correctement le fait que les SR obtiennent de meilleurs résultats quantitatifs.

5. Conclusion et perspectives

Cette étude confirme que les expressions phraséologiques constituent de bons descripteurs pour la classification en sous-genre (Chambre & Kraif, 2017). En effet, même si les résultats obtenus ici à partir du sous-ensemble constitué des expressions spatiales et temporelles sont sensiblement inférieurs à ceux obtenus à partir de traits plus directement liés aux univers de référence de chaque sous-genre (61.4 % /vs/ 98 %), ces expressions moins riches sur le plan informatif permettent cependant de classer les romans dans les sous-genres marqués POL, SF et HIST de manière satisfaisante. En revanche, pour la catégorie des romans généraux (GEN), elles ne sont pas discriminantes. Notre méthode permet aussi de dégager des spécificités

génériques propres à ces deux champs ESPACE et TEMPS (lieux de pouvoir dans HIST /vs/ intérieur et lieux urbains dans POL ; durées et distances extrêmes dans SF). Enfin, à partir de cette sélection d'expressions spatio-temporelles, la méthode des segments répétés produit une classification en sous-genres plus précise que celle des ALR. Ce point, difficile à interpréter à partir de nos premières observations qualitatives, nécessite une étude plus approfondie. Ces résultats nous incitent à poursuivre l'exploration d'autres champs lexicaux en marge des univers de référence de chaque sous-genre, afin, d'une part, d'affiner notre méthodologie et, d'autre part, de cibler les éléments au cœur de la phraséologie.

Références

Aït-Mokhtar S., Chanod J.-P. and Roux C. (2002). Robustness beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8:121-144.

Boyer A.-M. (1992). *La paralittérature*. Presses Universitaires de France.

Chambre J. et Kraif O. (2017). Identification de traits spécifiques du roman policier et de science fiction. Communication présentée aux *Journées Internationales de la Linguistique de Corpus - JLC2017*, Grenoble, 05.07.2017.

Eibe F., Hall M. A. and Witten I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition.

Kraif O., Novakova I. et Sorba J. (2016). Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction. *Lidil*, 53 : 143-159.

Kraif O. et Diwersy S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. *Actes de la conférence TALN 2012*, pp. 399-406.

Lefer M.-A., Bestgen Y. et Grabar N. (2016). Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, pp. 555-563.

Tutin A. et Kraif O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. *Lidil*, 53 : 119-141.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Klincksieck.

Les phrases de Marcel Proust

Cyril Labbé[1], Dominique Labbé[2]
1 Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, F-38000 Grenoble France
(cyril.labbe@imag.fr)
2 Univ. Grenoble Alpes, PACTE (dominique.labbe@umrpacte.fr)

Abstract
Analysis of sentence lengths in Marcel Proust's *A la recherche du temps perdu*. Counting standards and the various available measures are presented. For most of his reading time, the reader of this novel is confronted with very long and syntactically-complex sentences. A comparison with other writers shows that these sentences are atypical but not unique and that some of their characteristics can be observed in a number of other works, some of which are cited in the *Recherche du temps perdu.*

Résumé
Analyse des longueurs de phrases dans *A la recherche du temps perdu* de Marcel Proust. Présentation des normes de dépouillement et des différentes mesures possibles. Durant la majorité de sa lecture, le lecteur se trouve confronté à des phrases très longues et syntaxiquement complexes. Une comparaison avec un large panel d'écrivains montre qu'il s'agit d'un phénomène exceptionnel mais pas unique et que certaines caractéristiques se retrouvent dans quelques œuvres dont certaines sont citées dans la *Recherche du temps perdu*.

Keywords: lexicometry - stylometry - sentence length – French literature - Proust

1. Introduction

Les phrases de Marcel Proust (1871-1922) sont-elles exceptionnelles ? La question a été surtout traitée sous l'angle qualitatif (notamment Curtius 1970). Il existe quelques estimations quantitatives (Bureau 1976, Brunet 1981, Milly 1986), avec des résultats divergents pour des raisons qui seront explicitées au début de cette communication. Mais surtout, nous présentons une comparaison statistique avec d'autres écrivains qui permettra de juger de l'exceptionnalité de la phrase proustienne.

L'analyse des phrases soulève plusieurs des problèmes auxquels est confrontée la lexicométrie (statistique appliquée au langage). En premier lieu, ici, il y a le choix de l'édition de référence. En effet, pour la *Recherche du temps*

perdu, ce choix existe et introduit une légère incertitude concernant la ponctuation de l'oeuvre (discussion dans Ferré 1957 et Serça 2010), spécialement pour les trois derniers volumes. Nous nous sommes tenus au principe général selon lequel fait foi l'ultime version révisée par l'auteur ou, à défaut, la plus proche de sa mort. Il s'agit ici de l'édition originale chez Gallimard (annexe 1). De plus, cette édition originale s'impose puisqu'elle est dans le domaine public et peut être communiquée librement aux chercheurs soucieux de reproduire nos résultats et d'aller plus loin dans cette analyse.

2. Le mot et la phrase

Le mot est défini comme l'occurrence d'un vocable, c'est-à-dire une entrée dans le lexique de la langue française selon la norme présentée par Muller 1963. Cette norme est fondée notamment sur la nomenclature de Hatzfeld et al. 1898. Son implémentation est décrite dans Labbé 1990. Par exemple, "aujourd'hui", "parce que" ou "Saint-Loup" sont des mots uniques et non deux "formes graphiques". Il y a 1 449 "parce que" dans la *Recherche*, soit plus d'un mot pour mille ; et 787 fois "Saint-Loup" (l'un des principaux personnages du roman). A l'inverse, les formes graphiques "le", "la", "les" ont deux entrées (pronom ou article) ; "du" ou "des" sont la contraction de deux entrées du lexique - préposition "de" et article "le". En fonction de la norme retenue (vocable ou formes graphiques), le nombre de mots dans un texte peut varier de près 10%. Selon cette "norme Muller", la *Recherche* compte 1 327 859 mots (N dans la suite) et 21 836 vocables différents.

Quant à la phrase, il y a un accord général pour la définir comme l'empan de texte dont le premier mot comporte une majuscule initiale et qui se trouve compris entre deux ponctuations majeures. Les ponctuations majeures sont le point, les points d'interrogation et d'exclamation, les points de suspension. Cependant, aucun de ces 4 signes typographiques ne marque automatiquement une fin de phrase :

- le point dans « M. Verdurin » ne termine pas une phrase même s'il est suivi d'un mot à majuscule initiale. Il y a dans la *Recherche* 3 152 « monsieur » écrits "M.". C'est le deuxième substantif le plus fréquent dans la *Recherche* (juste derrière "Mme"), soit 2,4 pour mille mots. Ce point "non-terminal" se retrouve dans les initiales que Proust utilise pour "anonymiser" certains noms (Mme X.) ou derrière des abréviations (etc.).
- dans la *Recherche*, plus de trois points d'interrogation sur 10 sont internes à la phrase (721).
- il y a 1 201 points d'exclamation internes à la phrase et 190 points de suspension également dans cette situation. Proust a plusieurs fois déclaré son hostilité envers ces derniers mais il les utilise parfois. Par exemple : « La duchesse émit très fort, mais sans articuler : « C'est l'... i Eon l... b... frère à

Robert. » (*la Prisonnière*).
Cette rapide discussion permet de comprendre la solution adoptée : un automate détermine les fins de phrase et, en cas de doute, l'opérateur choisit : fin de phrase ou ponctuation interne ? A condition que l'opérateur suive toujours la même norme, le dépouillement est fait sans erreur et, surtout, les résultats obtenus sur un auteur sont comparables à ceux de tous les autres. Ce recensement établit le nombre de phrases de la *Recherche* (voir tableau en annexe). P = 37 336 phrases. Comment caractériser ces phrases en fonction de leurs longueurs ?

3. Les indices statistiques usuels.
Les P phrases sont rangées par longueur croissante, dans des classes d'intervalles égaux (ici 1 mots). Par exemple, la première classe (1 mot, généralement une exclamation) contient 124 phrases, soit 0,37% du total. L'effectif de chaque classe est ainsi recensé et son poids relatif est calculé. Ce recensement fournit les informations suivantes :
- **Etendue** de la distribution : 1 à 931 mots. La plus longue phrase est celle sur les homosexuels au début de *Sodome et Gomorrhe*. Les phrases de la *Recherche* ne sont pas réparties uniformément sur cet intervalle. La seconde plus longue – celle sur les chambres au début de *Combray* – compte 542 mots ; la troisième (le salon des Verdurins dans la *Prisonnière*) : 430 ; la quatrième (l'église de Combray) : 399. Ensuite, il n'y a plus de "trou" important dans l'étalement des longueurs.
- Le **mode** est la classe la plus peuplée, ou longueur de phrase que le lecteur a le plus de chance de rencontrer : 11 mots. Il y a donc, dans la *Recherche*, une prédominance des phrases courtes et syntaxiquement simples. Il en est ainsi dans la plupart des textes en français.
- La **médiane** est la valeur de la variable pour l'individu du milieu ou individu "médian". Dans les P phrases rangées par longueurs, l'individu médian est celui qui occupe la place (P+1)/2. Lorsque l'effectif total de la population (P) est pair, la médiane est la moyenne des valeurs de la variable pour les 2 individus situés de part et d'autre. Dans un texte étendu comme la *Recherche*, la médiane se trouve dans une classe dont l'effectif est assez élevé. Dans ce cas, la valeur est interpolée en divisant l'intervalle de la classe où se situe l'individu médian par l'effectif de cette classe. Dans la *Recherche*, ce calcul aboutit à une médiane de 26,28 mots. Etant donné que la variable "longueur de phrase" ne prend que des valeurs entières, les décimales indiquent le sens de l'arrondi et la position de la borne. La longueur médiane des phrases de la *Recherche* est donc de 26 mots. Ou encore la moitié des phrases ont une longueur inférieure ou égale à 26 mots et l'autre moitié une longueur supérieure à 26.

- La **moyenne** (N/P) : 35,57 mots. A cet indice est associée une déviation "standard" des valeurs de la variable autour de la moyenne (écart-type) : racine carrée de la variance (moyenne des carrés des écarts de chaque valeur de la variable à la moyenne arithmétique). L'écart type de la longueur des phrases de la *Recherche* est de 31,42 mots.
La **dispersion** des valeurs autour de la moyenne mesurée par le coefficient de variation relative : rapport de l'écart-type à la moyenne arithmétique (ici 89%). Etant donné l'effectif considéré (37 336 phrases), si les valeurs de la variable "longueur de phrase" étaient distribuées normalement autour de la moyenne (cas d'une population homogène), ce coefficient serait d'environ 4%. Autrement dit, les observations sont extrêmement dispersées. Dans ce cas, la moyenne n'est pas représentative de la série et, en particulier, il n'est pas possible de considérer que cette moyenne se situe à peu près "au milieu" de la population. Dès que la dispersion relative approche les 50% de la moyenne, celle-ci est située dans la partie basse de l'étendue de la distribution qui est fortement asymétrique. Le profil de la distribution des longueurs de phrases dans la *Recherche* est donné par la figure 1 dans laquelle l'effectif relatif de chaque classe est représenté par la hauteur du bâton correspondant (histogramme).

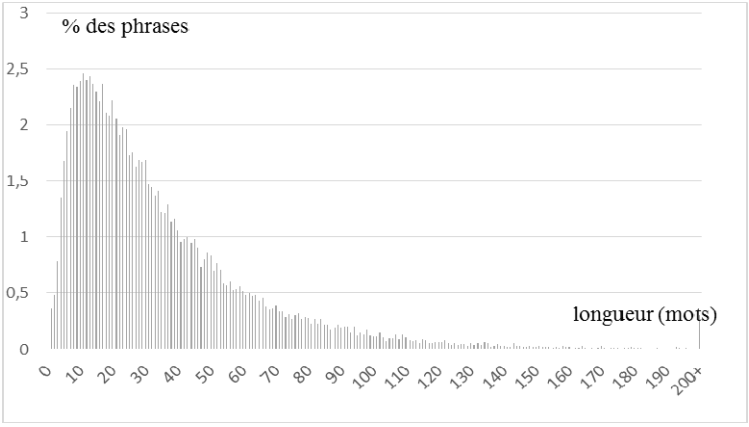


Figure 1. Histogramme de la distribution des longueurs phrases

D'une part, le graphique s'interrompt à la classe 200+ mots et le bâton pour cette classe – à l'extrême-droite du graphique - correspond aux 96 phrases longues de 200 mots et plus (0,3% du total des phrases mais 2,1% de la surface du texte). Le graphique complet est encore plus étalé sur la droite, la grande masse des phrases apparaissant serrées sur la gauche… D'autre part, le bâton le plus haut correspond au mode principal (11 mots) mais l'on observe de nombreux modes secondaires (17, 20, 24, etc.) : plusieurs

populations sont donc mélangées. La plupart des phénomènes sociaux présentent des caractéristiques semblables et, en premier lieu, la distribution des revenus ou des patrimoines. Dans de pareils cas, l'analyse ne se contente pas des valeurs centrales. Elle se centre sur la distribution du caractère étudié (ici la surface du texte) au sein de la population (ici les phrases).

4. L'inégal partage de la surface du texte entre les phrases

Ce renversement de perspective présente un avantage : la surface de texte correspond grosso-modo à la durée de la lecture. Deux méthodes sont possibles pour l'évaluer.

4.1 Quantile et médiale

Les phrases étant classées par longueurs croissantes, la surface du texte qu'elles couvrent est découpée en masses égales (tableau 1).

Tableau 1. Partage de la surface du texte en fonction de la longueur des phrases

Surface divisée en quantiles	Longueur (mots)	% des phrases (cumulé)
Premier décile	18.58	33,8
Deuxième décile	26.70	49,6
Premier quartile	29.53	54,5
Troisième décile	33.30	60,6
Quatrième décile	41.35	70,1
Deuxième quartile (médiale)	**49.93**	**77,5**
Sixième décile	60.20	84,6
Septième décile	72.93	89,7
Dernier quartile	81.13	92,3
Huitième décile	90.57	94,2
Neuvième décile	121.00	97,8

Dans ce tableau, le premier décile est la borne supérieure de l'intervalle comprenant les phrases les plus courtes couvrant en tout 10% de la surface du texte et la borne inférieure du 2e décile. Il indique que les phrases de longueurs inférieures ou égales à 18 mots couvrent 10% du texte et représentent plus du tiers du total des phrases (33,8%). Le lecteur n'y passe au mieux qu'un dixième du temps de la lecture. Or c'est au-dessus de cette longueur que l'on commence à rencontrer des phrases syntaxiquement complexes. Autrement dit, au mieux, le lecteur de la *Recherche* se trouve face à des phrases simples pendant un dixième de sa lecture (ou il est face à des phrases plus ou moins complexes pendant les neuf dixièmes !)

A l'opposé, 2,2% des phrases (700) comptent plus de 121 mots (9e décile). Elles couvrent également 10% du texte, c'est-à-dire la même surface que le tiers évoqué ci-dessus. Cela signifie que le lecteur de la *Recherche* passe (au moins) autant de temps à lire des phrases très longues – dont la construction est nécessairement complexe -, qu'il n'en consacre à la masse des phrases les

plus brèves et structurellement simples.
Dans cette perspective, la valeur centrale la plus caractéristique est la longueur de la phrase qu'il faut atteindre pour avoir lu la moitié du texte. Pour éviter les confusions, cette seconde médiane est appelée **médiale** (Ml). Elle correspond à la borne haute du cinquième décile (ou du deuxième quartile). Dans la *Recherche*, elle est égale à 49,93 mots, soit 50 mots. Le tableau indique que 77,5% des phrases (près de 8 sur 10) sont inférieures à cette médiale. Autrement dit, le lecteur de la *Recherche* passe au moins la moitié de son temps confronté à des phrases de 50 mots et plus, ce dont la plupart d'entre eux n'ont guère l'habitude. Malgré le talent de l'écrivain, c'est évidemment cela que les lecteurs retiennent.

4.2 Mesure de l'inégalité
Deuxième méthode, un indice unique mesure l'inégale répartition de la surface du texte entre les phrases (en fonction de leurs longueurs). Deux calculs sont proposés :
- le rapport entre la médiane (26,28) et la médiale (49,93) soit 0,90. Autrement dit la médiale est de 90% supérieure à la médiane (pour des comparaisons avec d'autres écrivains, voir l'annexe 2). Cet écart considérable suffit à attester la prédominance des phrases longues dans la *Recherche*.
- le second calcul est utilisé en science économique pour étudier la distribution des revenus ou des patrimoines. Il s'agit de l'indice de Gini qui mesure l'écart entre la situation réelle et celle qui serait observée en cas d'égale répartition du caractère (ici la surface du texte) entre les individus (les phrases) composant le livre. En cas d'équirépartition, toutes les phrases de la *Recherche* auraient la longueur moyenne (≈ 36 mots). Pour chaque centile, on calcule la proportion de la surface de texte couverte et l'écart par rapport à ce que serait cette surface dans l'hypothèse d'équirépartition. L'indice de Gini est la somme de ces écarts. Ici, il est égal à 55,4%. Autrement dit, dans la *Recherche*, les longueurs de phrases s'écartent de plus de 55% de ce qui serait constaté dans une population homogène.
Le "diagramme de Gini" permet de visualiser cette situation. Les phrases étant rangées par longueurs croissantes, on compte le nombre qu'il faut lire pour atteindre 1% de la surface (premier centile), puis 2%, etc. jusqu'à 100%. Les valeurs observées pour chaque centile sont reportées sur la figure 2 où la diagonale représente l'hypothèse d'équirépartition. L'indice de Gini est la surface comprise entre la diagonale et la courbe. Deux auteurs contemporains, et importants pour M. Proust, sont ajoutés sur le diagramme afin d'en illustrer les propriétés.

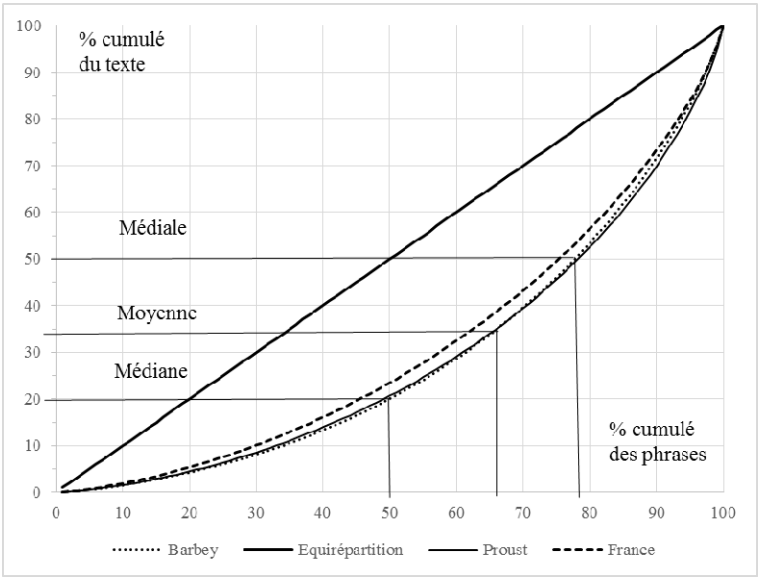


Figure 2 Diagramme de concentration (Gini) de la surface de la Recherche sur les phrases longues, comparée à celle de J. Barbey d'Aurevilly et de A. France.

Ce diagramme permet de comprendre pourquoi la médiane ou la moyenne rendent mal compte des distributions fortement asymétriques comme les longueurs de phrase. Par exemple, les deux tiers des phrases ont des longueurs inférieures à la moyenne et pourtant ces phrases ne couvrent qu'à peine plus d'un tiers du texte (34,5%).

La figure 2 montre également que, si les phrases de la *Recherche* sont singulières par rapport à certains écrivains du XIXe - à commencer par A. France qui aurait fourni le modèle de Bergotte (Levaillant 1952) –, elles semblent très proches de quelques livres comme *Une vieille maîtresse* (1851) de Barbey d'Aurévilly, écrivain que Proust cite à plusieurs reprises (Rogers 2000). C'est la dernière question abordée dans cette communication.

5. Singularité de Proust ?

Pour juger de cette singularité : à qui le comparer ? Et comment décider si les écarts constatés sont statistiquement significatifs ?

Premièrement, il faut comparer Proust à lui-même. Un de ses ouvrages se trouve dans le domaine public : *Les Plaisirs et les jours* (1896) dont les valeurs centrales sont indiquées en première ligne dans le tableau 2.

Tableau 2. Caractéristiques des phrases des Plaisirs et les jours comparés à la Recherche

	Etendue	Mode	Médiane	Moyenne	Médiale	Me/Ml	Gini
Plaisirs et jours	1-250	7	21,30	27,87	37,16	0,754	0,542
Recherche	1-931	11	26,28	35,57	49,93	0,900	0,554

Toutes ces valeurs sont significativement inférieures à celles observées dans la *Recherche*. Cependant, l'indice de Gini indique que le jeune Proust avait déjà tendance à concentrer une proportion importante du texte dans les phrases longues.

Deuxièmement, il faut comparer Proust aux auteurs qu'il cite explicitement ou par allusion, non seulement dans la *Recherche* (Nathan 1968) mais aussi dans ses autres œuvres et dans sa correspondance (Chantal 1967). Dans la *Recherche*, Racine et Mme de Sévigné sont les plus cités, puis en seconde position : Balzac et Saint-Simon ; en troisième : Chateaubriand, Hugo, Molière, Musset, Sand et Vigny. La singularité des phrases théâtrales (Labbé & Labbé 2010) ne permet pas de comparer la *Recherche* (qui est un roman) avec les pièces produites par Molière, Hugo, Musset, Racine ou Vigny.

Enfin, il faut le comparer aux autres romanciers contemporains : ont été ajoutés les principaux écrivains du XIXe et du début du XXe - comme Bourget, Giraudoux, Flaubert, Maupassant, Zola – et quelques auteurs moins connus mais singulièrement proches de Proust.

L'annexe 2 présente un échantillon des résultats. Chaque écrivain est singulier et parfois les indices peuvent varier selon ses oeuvres. La *Recherche* se situe dans la partie haute pour tous les indices et notamment pour la propension à concentrer une proportion importante du texte dans les phrases les plus longues (Gini). Cependant, on observe des caractéristiques supérieures à celle de Proust dans quelques œuvres - Huysmans (*A rebours*), les frères Goncourt (*Mme Gervaisais*) - ou proches dans Barbey d'Aurevilly, mais aussi dans les *Lettres* de Mme de Sévigné ou les *Mémoires* de Saint-Simon.

6. Conclusions

Lorsque, dans une population – ici les phrases d'un texte -, un caractère (la surface de ce texte) est très inégalement réparti, la moyenne et la dispersion standard sont de peu d'utilité. L'indice statistique le plus éclairant est la seconde médiane ou médiale. Pour mesurer le degré de dispersion de la série autour de cette valeur centrale, de nombreux indices sont concevables, notamment les rapports entre quantiles extrêmes. Cependant, le rapport entre médiane et médiale, ou l'indice de Gini paraissent les plus aptes à donner une indication de la concentration du caractère sur une proportion

plus ou moins restreinte de la population totale.
Ces indices montrent que, durant la majorité du temps, le lecteur de la *Recherche* se trouve confronté à des phrases très longues (50 mots et plus) et syntaxiquement complexes. Ils confirment que M. Proust a une propension à concentrer une proportion importante du récit dans les phrases les plus longues.
Ces conclusions ont été acquises grâce à un dépouillement rigoureux, à des indices statistiques adaptés et à une vaste base de textes traités selon les mêmes procédures. A ce prix, la statistique lexicale peut être une auxiliaire utile de l'analyse littéraire.
Enfin, dans une œuvre littéraire, il n'existe pas un type de phrase unique mais plusieurs qui ont chacun leurs particularités lexicales et stylistiques (Monière et al. 2008 ; Labbé & Labbé 2010). Une prochaine publication présentera ces types de phrases avec leurs singularités lexicales, stylistiques et thématiques. Elle répondra aussi à une question pendante : comment déterminer que les écarts entre œuvres et auteurs sont ou non significatifs ?

References

Brunet E. (1981). La phrase de Proust. Longueur et rythme. *Travaux du cercle linguistique de Nice*, p. 97-117.

Bureau C. (1976). Marcel Proust ou le temps retrouvé par la phrase. *Linguistique fonctionnelle et stylistique objective*. Paris : PUF, p. 178-231.

Curtius E.-R. (1971). Etude de lilas. Le rythme des phrases. In Tadié J.-Y. (dir.). *Lectures de Proust.* Paris : A. Colin.

Milly J. (1975). *La phrase de Proust. Des phrases de Bergotte aux phrases de Vinteuil*. Paris : Larousse.

Ferré A. (1957). La ponctuation de M. Proust. *Bulletin de la Société des Amis de Marcel Proust*, 7, p 171-192.

Hatzfeld A., Darmeisteter A., Thomas A. (1898). *Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours*. Paris : Delagrave.

Labbé C., Labbé D. (2010). Ce que disent leurs phrases. In Bolasco S., Chiari I., Giuliano L. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto. Vol 1, p. 297-307.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Levaillant J. (1952). Note sur le personnage de Bergotte. *Revue des sciences humaines*. Janvier-Mars 1952, p 33-48.

Milly J. (1986*). La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine.

Monière D., Labbé C. & Labbé D. (2008). Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, p. 43-69.

Muller C. (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Nathan J. (1969). *Citations, références et allusions de Marcel Proust dans* A la recherche du temps perdu. Paris : Nizet (Première édition : 1953).

Rogers B. (2000). *Proust et Barbey d'Aurevilly. Le dessous des cartes*. Paris : Champion.

Serça I. (2010). *Les coutures apparentes de la Recherche. Proust et la ponctuation*. Paris : Champion.

Annexe 1 Corpus A la Recherche du temps perdu (Marcel Proust. Paris Gallimard 1919-1927)

Livre	Longueur	Vocabulaire	N phrases
Combray	79 906	6 502	1 727
Un amour de Swann	84 142	5 859	2 226
Noms de pays : le nom	19 434	2 823	374
Du côté de chez Swann (1919)	**183 482**	**9 347**	**4 327**
Autour de Mme Swann	91 451	6 532	2 511
Noms de pays : le pays	134 192	8 283	3 334
A l'ombre des jeunes filles en fleur (1919)	**225 643**	**10 396**	**5 845**
Le côté de Guermantes 1	75 494	6 281	1 903
Le côté de Guermantes 2, chapitre 1	84 354	6 368	2 781
Le côté de Guermantes 2, chapitre 2	89 727	6 707	2 700
Le côté de Guermantes (1920-21)	**249 575**	**6 707**	**7 384**
Sodome et Gomorrhe	13 512	2 476	271
Sodome et Gomorrhe 2, chapitre 1	30 699	3 779	2 082
Sodome et Gomorrhe 2, chapitre 2	117 774	7 822	3 056
Sodome et Gomorrhe 2, chapitre 3	57 603	5 311	1 811
Sodome et Gomorrhe 2, chapitre 4	8 137	1 373	250
Sodome et Gomorrhe (1921-22)	**227 725**	**10 972**	**7 470**
La prisonnière (1923)	173 409	9 062	5 124
La fugitive (1925)	115 866	6 456	3 255
Le temps retrouvé (1927)	152 159	8 708	3 931
Dernier volume (posthume)	**441 434**	**13 518**	**12 310**
Total général (*A la recherche du temps perdu)*	**1 327 859**	**21 837**	**37 336**

Annexe 2 Longueur des phrases chez quelques écrivains antérieurs ou contemporains de Proust

	Etendue	Mode	Médiane	Moyenne	Médiale	Me/Ml	Gini
Recherche	**931**	**11**	**26,28**	**35,57**	**49,93**	**0,900**	**0,554**
Balzac	391	10	17,27	21,88	29,00	0,680	0,511
Barbey d'A. (*Chevalier*)	192	7	21,92	29,4	43,00	**0,964**	**0,557**
Barrès	195	8	17,86	21,94	28,59	0,601	0,497
Bourget	201	7	16,62	21,34	29,58	0,780	0,539
Chateaubriand (*Mémoires*)	195	22	24,46	28,5	34,28	0,401	0,437
Daudet	203	5	13,14	17,84	25,26	0,923	0,549
Dumas	243	7	14,90	20,28	29,00	0,947	0,567
Flaubert	231	7	13,75	18,37	25,24	0,837	0,528
France	394	8	15,79	19,98	26,06	0,651	0,504
Gautier*	282	18	27,11	33,07	41,90	0,546	0,493
Giraudoux*	466	4	18,60	25,77	37,76	1,031	0,580
Goncourt (*Gervaisais*)	**670**	**8**	**24,17**	**34,05**	**51,47**	**1,130**	**0,597**
Goncourt (*Journal*)	373	3	19,80	25,37	37,62	**0,900**	**0,580**
Hugo*	828	6	11,39	16,89	23,68	1,079	0,561
Huysmans (*A rebours*)	254	**28**	**44,24**	**51,49**	**65,82**	0,488	**0,557**
Maupassant*	168	6	14,44	18,98	26,39	0,828	0,542
Musset*	197	16	19,56	23,82	29,57	0,512	0,485
Nerval*	136	12	19,93	24,21	31,27	0,569	0,499
Saint-Simon	361	18	27,89	34,15	44,14	0,523	0,506
Sand (*Champi*)	117	21	22,11	26,19	32,56	0,473	0,477
Sévigné (*Lettres*)	307	11	25,72	31,99	40,96	0,593	0,490
Stendhal	235	18	20,18	23,92	29,79	0,477	0,463
Vigny*	315	17	20,82	27,47	37,41	0,797	0,538
Zola	153	8	15,80	19,91	25,66	0,624	0,491

* Uniquement les romans

Verso un dizionario *corpus-based* del lessico dei beni culturali: procedure di estrazione del lemmario

Ludovica Lanini[1], María Carlota Nicolás Martínez [2]
[1] Università degli Studi di Roma La Sapienza– ludovica.lanini@uniroma1.it
[2] Università degli Studi di Firenze – cnicolas@unifi.it

Abstract
The vocabulary of Italian cultural heritage has become a crucial object of interest for different categories of users from a number of countries. However, there are no satisfactory multilingual lexical resources available. The present work moves in that direction. The aim of the paper is twofold: on the one hand, it describes the LBC database, a resource for developing a multilingual electronic dictionary of cultural heritage terms, made up of comparable corpora from nine languages; on the other hand, a corpus-based method for building a comprehensive headword list is proposed.

Keywords: electronic lexicography, multilingual lexical resources, corpus linguistics

1. Introduzione

Di fronte a un interesse crescente, a livello internazionale, per il lessico italiano dei beni culturali, emerge oggi l'esigenza, da parte di diverse categorie di utenti, di risorse elettroniche multilingui relative al patrimonio culturale; nonostante ciò, allo stato attuale, non sono disponibili strumenti multilingui adeguati. Il progetto LBC (Lessico dei Beni Culturali) si propone di affrontare il problema, sviluppando una banca dati testuale comprendente *corpora* specialistici e comparabili per nove lingue (cinese, francese, inglese, italiano, portoghese, russo, spagnolo, tedesco, turco). Fine ultimo è la creazione di un dizionario multilingue del lessico dei beni culturali a base testuale, che abbia come principali utenti studiosi del settore, ma anche traduttori e operatori turistici. L'approccio *corpus-based* viene applicato sin dal processo di definizione del lemmario, focus specifico del contributo.

2. La Banca dati LBC

La Bd-LBC (Banca dati LBC) è un *database* testuale multilingue progettato per essere rappresentativo del lessico dei beni culturali: per il suo disegno si è considerato l'italiano quale punto di partenza, ma si è pensato anche al valore aggiunto derivante dalla possibilità di stabilire relazioni tra le diverse lingue. L'italiano viene scelto come punto di riferimento in virtù della sua

centralità nello sviluppo storico del lessico dei beni culturali; molti testi non italiani relativi a tale dominio hanno inoltre lo sguardo rivolto proprio verso le tecniche e i monumenti realizzati in Italia. La prima fase di lavoro, dedicata alla raccolta dei materiali, è partita dunque dai testi italiani che sono alla base della storia dell'arte e dalle relative traduzioni, ma anche da opere in altre lingue, applicando una metodologia di studio che facesse leva sulle potenziali sinergie plurilingui. Per dare fondamento alla struttura del corpus (Cresti et Panunzi 2013:57), la rappresentatività della risorsa è stata definita fin dall'inizio attraverso dei criteri di campionamento dei testi (Billero et Nicolás 2017: 208): «la rilevanza storico-culturale dell'opera dell'ambito specifico di studio (ad es. testi di Vitruvio o Leonardo); la diffusione internazionale di un'opera relazionata con l'ambito di studio (es. libri di Vasari); il prestigio dato a livello internazionale al patrimonio italiano da parte di un'opera (es. testi di Stendhal o Ruskin); la specificità dell'argomento in rapporto alla storia dell'arte italiana ed in particolare della Toscana (es. Burckhardt) ». Si è in questo modo delimitato un nucleo di testi di base condivisi tra lingue, tale da rendere il corpus parzialmente parallelo, cui si sono aggiunti via via testi peculiari per ogni lingua.

La progettazione del *database* ha previsto inoltre una macrostruttura omogenea per i diversi *corpora*, che condividono i metadati associati a ogni testo, a partire dai quali viene generato automaticamente un nome di file univoco. Per quanto riguarda la microstruttura, la regola fondamentale è stata quella di rispettare il testo originale, mantenendo eventuali note, divisione in capitoli e tratti ortografici arcaici. Seguendo tali regole strutturali, ogni squadra di lavoro, specificamente rivolta a una delle lingue, ha avviato lo sviluppo dei singoli *corpora* (Corpus LBC-francese, Corpus LBC-inglese, etc.), sottoposti a un'operazione di validazione della digitalizzazione da parte di professori e studenti competenti nelle diverse lingue. La banca dati, così disegnata, presenta un'omogeneità in grado di favorire il lavoro lessicografico: la forte coesione strutturale tra *corpora* permette infatti di operare davvero in parallelo.

Tra gli obiettivi del progetto vi è anche quello di implementare strumenti informatici di gestione e interrogazione dei *corpora,* che consentano ai membri del gruppo di effettuare ricerche ed estrarre dati sull'uso lessicale, fondamentali per lo svolgimento del lavoro lessicografico. Si è dunque realizzato un software online, per ora accessibile ai soli membri dell'unità di ricerca, ma in prospettiva disponibile anche per gli utenti, che consenta la consultazione dei *corpora*, sia in chiave monolingue che multilingue. Nella ricerca di soluzioni per l'implementazione di un'installazione del *corpus* su apposito server Internet, si è optato per l'ultima release di NoSketchEngine, versione *open source* di Sketch Engine.

3. Il dizionario LBC: processo di definizione del lemmario

La banca dati, così elaborata, si pone quale risorsa di base per lo sviluppo di un dizionario elettronico multilingue del lessico dei beni culturali, che possa risultare strumento utile soprattutto in ambito traduttivo e turistico. In vista della particolare utenza e applicazione, l'intento è quello di fornire una risorsa lessicografica che presenti le seguenti caratteristiche:

- trattamento dei lemmi più "problematici" del dominio, con inclusione a lemma di nomi propri ed espressioni multiparola, categorie lessicali generalmente assenti dalle risorse, tuttavia di particolare rilevanza in virtù delle difficoltà traduttive e del forte carico culturale;
- attenzione per l'aspetto più prettamente pratico e referenziale del lessico della cultura, con apertura a quelle voci di arti e mestieri tradizionalmente trascurate dalla lessicografia italiana, nonché interesse rivolto alle persone, alle opere e ai luoghi fisici della storia culturale, più che al carattere teorico e mentale (Harris, 2003) ed estetico generale (De Mauro, 1971) che ha a lungo connotato il lessico artistico, in particolare quello della critica d'arte;
- inclusione non solo di nomi, ma anche di verbi, di norma esclusi dalle risorse terminologiche, qui ritenuti di interesse per rendere conto di tecniche e pratiche;
- impianto *corpus-based*, non solo per la selezione, descrizione e traduzione dei lemmi, con individuazione degli equivalenti a partire dall'analisi di concordanze bilingui, ma anche per l'offerta all'utente, entro la scheda lessicografica, di esempi e citazioni testuali reali.

L'approccio *corpus-based* viene adottato sin dal processo di definizione del lemmario, sviluppato a partire dal *corpus* LBC-italiano.

Il metodo proposto prevede la combinazione di tre ordini eterogenei di dati: dato lessicografico; dato testuale quantitativo; dato testuale qualitativo. Il dato di origine lessicografica, assunto sullo sfondo a *frame* di riferimento, viene dunque incrociato con il dato testuale, tanto di livello quantitativo -*keyword* e liste di frequenza- quanto di livello qualitativo -prodotto di ricerche mirate su *corpus* e di osservazione dei contesti.

Per quanto riguarda le risorse adottate, la fonte lessicografica scelta è il *Grande Dizionario Italiano dell'Uso* (De Mauro, 2007), la più estesa risorsa lessicografica esistente per la lingua italiana, mentre alla banca dati LBC viene affiancato, quale corpus generale di riferimento, il corpus *Paisà* (www.corpusitaliano.it), costruito nel 2010 tramite *web-crawling* e raccolta mirata di documenti da specifici siti web, per un totale di 250 milioni di *token*, inteso come rappresentativo della lingua e cultura comune contemporanea (Lyding *et al.*, 2014). Indirettamente, viene assunto come corpus di riferimento anche itTenten16, il *corpus* per la lingua italiana implementato in *Sketch Engine*, interamente raccolto tramite *web-crawling* nel 2016

(5.864.495.700 *token)*. Riguardo agli strumenti impiegati, l'adozione di un *software* di *corpus management* e *query* all'avanguardia come *Sketch Engine* (www.sketchengine.co.uk) risulta infatti cruciale per il processo di lavoro, descritto di seguito nel dettaglio.

3.1 Fasi di lavoro

La prima operazione è consistita nell'estrazione dal corpus LBC di una lista di parole chiave (2000), applicando la funzione *keywords* di *Sketch Engine*: le *keyword* vengono ordinate in base al *keyness score*, dato dal rapporto tra la frequenza normalizzata della parola nel *focus corpus* (LBC) e la sua frequenza normalizzata in un corpus generale (itTenten16), previa applicazione di una costante, denominata *simple math parameter*[1] (Kilgariff *et al.*, 2014).

Alla lista delle *keyword* è stata affiancata la lista di matrice lessicografica, estratta dal Gradit selezionando l'insieme dei lemmi etichettati con marca [TS] (tecnico-specifico) per arte, pittura, scultura e architettura, per un totale di 2515 lemmi, di cui molti (370) multiparola. In maniera inattesa, dal confronto tra le due liste emergono solo 24 coincidenze.

Risultando poco pulita, la lista delle *keyword* è stata sottoposta a uno spoglio manuale, che ha ridotto i 2000 lemmi candidati a 219, primo vero lemmario di base (comprendente nomi propri come *Mantegna*, arcaismi come *fregiatura*, tecnicismi come *nicchia*).

Si è proceduto a questo punto a una serie di confronti, a partire dalla lista di frequenza lemmatizzata del corpus LBC, come sintetizzato in *Tabella 1*. L'incrocio con la lista del Gradit ha restituito 272 lemmi comuni, di cui 235 sono stati accolti previo controllo. Il lavoro di confronto con il *corpus* generale *Paisà* ha seguito invece due linee di sviluppo: lo studio dei lemmi caratterizzati da più alta differenza di frequenza relativa con peso maggiore in LBC (i primi 600), da cui sono emersi 77 lemmi di interesse (*figura*, *Firenze*, *Raffaello*) e lo spoglio dei lemmi presenti in LBC ma non in *Paisà*, che ha permesso di individuarne 62 (tecnicismi come *scalea* e *imbasamento*, numerosi arcaismi e varianti arcaiche come *scarpellino*, *Florenzia*, *Buonarruoto*).

L'insieme delle voci della lista Gradit assenti in LBC (ben 2243) è stato inoltre sottoposto a un esame puntuale, che ha portato ad aggiungere al lemmario 1629 lemmi[2]. Il corpus LBC è in effetti in fase di sviluppo, per cui molte aree

[1] A seconda dei bisogni dell'utente e della natura dei *corpora*, la costante può essere modificata per restituire una lista con candidati a frequenza maggiore o minore, con 100 come valore consigliato per ottenere parole del vocabolario *core* e rumore minimo, qui applicato.

[2] Non si sono accolti: lemmi astratti, propri della critica d'arte (*asemanticità*); lemmi riferiti a movimenti e tendenze generali (*astrattismo*); aggettivi o avverbi. Si

di interesse (per esempio il dominio dell'arte contemporanea) non risultano ancora adeguatamente rappresentate: la lista del Gradit può offrire in questa direzione materiali utili, in attesa dell'ampliamento del corpus.
Dalla convergenza dei lemmi accolti è stato così possibile arrivare alla definizione di un primo lemmario, per un totale di 2147 lemmi.

Tabella 1

Risorse	**Lemmi**	**Lemmi di interesse**	**Lemmi estratti**	**Lemmi accolti**
Lista LBC	8388	Lemmi comuni	272	235
Lista Gradit	2515			
Lista LBC	8388	Lemmi con differenza di frequenza relativa significativa	600	77
Lista Paisà	1032178	Lemmi presenti in LBC assenti in Paisà	1139	62
Lista *keywords* <LBC	2000	Tutti	2000	219
Lista Gradit	2515	Lemmi assenti in LBC	2243	1629
			TOT.	2222 (-75 lemmi ripetuti) = **2147**

Il confronto con *Paisà*, in particolare, ha permesso di individuare una serie di lemmi di interesse, non rappresentati nella lista lessicografica: nomi propri (*Raffaello*, *Firenze*), tecnicismi (*travata*), molti lemmi comuni, spesso semanticamente polivalenti e dotati di accezione specifica (*figura*, *opera*), ma anche arcaismi (*reliquiere*) e varianti grafiche arcaiche (*trivertino*), ritenuti utili in vista della lettura e traduzione di testi.

3.2 Estrazione di lemmi multiparola

---

sono invece accettati: lemmi che siano forme derivate di lemmi rappresentati in LBC (*aggrottescare*, con *grottesca* presente in LBC); lemmi relativi all'arte contemporanea (*acrilico*); unità multiparola ritenute di interesse (*arco acuto*); verbi ritenuti di interesse (*festonare*).

Il lavoro di analisi testuale si è finora limitato a soli lemmi semplici; l'intento è tuttavia quello di includere nel dizionario, come entrate autonome, anche lemmi multiparola. Si tratta di entità problematiche, per le quali sono state proposte innumerevoli denominazioni, classificazioni e criteri di identificazione. Relativamente alla lingua italiana, si è parlato di *lessemi complessi* (Voghera, 1994), *polirematiche* (Voghera, 2004, De Mauro, 2005), *espressioni multiparola* (Masini, 2009). Alla tradizione anglosassone appartiene il termine *multiword expression* (MWE), iperonimo utilizzabile per descrivere una serie di entità generalmente distinte sul piano teorico, ma accomunate da proprietà quali restrizione combinatoria, alto grado di lessicalizzazione e convenzionalità, non-composizionalità e opacità semantica. Tale approccio generalizzante risulta presente anche entro la tradizione lessicografica, poco interessata del resto a distinzioni di ordine teorico (*multiword lexical units* in Zgusta, 1971). La decisione è quella di adottare una terminologia che sia in linea con tale tradizione, scegliendo la denominazione *lemmi multiparola*, e abbracciando una definizione generale come quella proposta da Calzolari *et al.* (2002:1934): «word combinations characterised by different degrees of fixedness and idiomaticity that act as a single unit at some level of linguistic analysis, such as idioms, collocations, preferred combinations». La caratteristica essenziale dei lemmi multiparola è quella di comportarsi come parole semplici: in ambito lessicografico, dunque, la scelta non può essere che quella di trattarli al pari di lemmi semplici, introducendoli come entrate autonome. Tale scelta risulta operativa anche nel Gradit: lo spoglio della lista lessicografica ha permesso già dunque di accogliere un buon numero di lemmi multiparola (303). Per quanto riguarda il dato testuale, un primo tentativo è consistito nell'applicazione della già citata funzione *keywords* fornita da Sketch Engine. La funzione permette in effetti di estrarre anche una lista di *term*, intesi come *noun-phrase* dotati di un buon indice di *keyness*, ma la scarsa qualità dei risultati ottenuti (solo 2 candidati su 2000 sono risultati accettabili) e il fatto che la lista includa solo sintagmi nominali e non verbali ha spinto ad adottare soluzioni alternative.

Si è deciso dunque di partire da un nucleo di lemmi semplici, per estrarne i *word sketch*. Si tratta della funzione più caratteristica del software, vera e propria sintesi *corpus-based* del comportamento collocazionale di una parola: i collocati, selezionati e ordinati in base a un indice di associazione (logDice), vengono mostrati entro categorie, i *gramrel*, corrispondenti a specifici *pattern* definiti da una *sketch grammar* soggiacente, scritta in CQL. Il nucleo dei lemmi di partenza (643) è risultato dalla convergenza di tre sottoinsiemi, scelti in quanto fortemente rappresentativi del dominio ma anche potenzialmente produttivi in termini combinatori: la lista *keyword* (in una versione ripulita, ma ancora comprendente aggettivi e tutti i nomi propri); i

lemmi comuni tra lista LBC e lista Gradit; i lemmi base delle unità multiparola della lista Gradit accolte nel lemmario. Per la generazione degli *sketch* si è applicata inoltre una versione personalizzata della *sketch grammar*, tarata sul dominio di studio: la modifica e l'aggiunta di regole ha permesso di focalizzarsi su *pattern* di specifico interesse, non previsti dalla grammatica del software al momento del lavoro, tra cui in particolare combinazioni includenti nomi propri, sia come basi (*Accademia fiorentina*) che come collocati (*fabrica di San Piero*). Dall'analisi degli *sketch* sono emersi 57 lemmi multiparola, di cui addirittura 46 assenti nella lista del Gradit (80.7%). Le ragioni dell'assenza possono essere diverse: a volte i lemmi risultano collocati all'interno delle entrate, altre volte è in gioco una differente valutazione in merito alla segmentazione delle unità multiparola (per cui, per esempio, *a olio* viene collocato a lemma, mentre *colore a olio* no). È possibile tuttavia individuare numerosi lemmi che si pongono come decisamente nuovi rispetto alla lista "ufficiale" proveniente dal Gradit: tra questi, ancora una volta, nomi propri (*Colonna di Traiano*) e varianti grafiche (*volta a botta*), ma anche nomi comuni con collocati nomi propri (*marmo di Carrara*) e tecnicismi del lessico artistico utili a colmare lacune negli elenchi di oggetti o tecniche risultanti incompleti nella lista del Gradit (*nicchia bislunga*, *nicchia piana*, *nicchia quadra*, da aggiungere alla serie *nicchia a edicola*, *nicchia a tabernacolo*, *nicchia angolare*, *nicchia finta*). I lemmi multiparola estratti dal corpus sono confluiti nel lemmario definitivo, giunto così a un totale di 2204 lemmi, di cui 360 multiparola. Il lavoro sui lemmi multiparola è comunque in corso d'opera e l'intento è quello di portare avanti la ricerca su *corpus* in tale direzione.

4. Conclusioni e lavoro futuro

Il metodo descritto ha portato allo sviluppo di un lemmario per un dizionario del lessico dei beni culturali comprendente 2204 lemmi. L'approccio *corpus-based* ha permesso di individuare una serie di lemmi assenti nella fonte lessicografica adottata, per un totale di 287 nuovi lemmi.

In futuro, il metodo illustrato si vedrà sottoposto a validazione, con l'applicazione ad altre lingue della banca dati LBC. Parallelamente, il processo di lavoro per il lemmario italiano verrà comunque portato avanti, in due direzioni: per l'aggiunta di ulteriori lemmi multiparola -per esempio, tramite estrazione di liste di frequenza di n-grammi- e per la sua crescita in concomitanza con quella del *corpus* stesso. Anche i lemmi estratti dai *corpora* delle altre lingue e non presenti nel lemmario italiano potranno inoltre contribuire al suo arricchimento.

References

Billero R., Nicolás Martínez M. C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In *CHIMERA Romance Corpora and Linguistic Studies*, Madrid, UAM, 4.2, pp. 203-16.

Calzolari N., Fillmore C. *et al*. (2002). Towards best practice for multiword expressions in computational lexicons. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Canary Islands, pp. 1934-40.

Cresti, E., Panunzi, A. 2013. *Introduzione ai corpora dell'italiano*. Bologna, Il mulino.

De Mauro T. (1971). *Senso e significato: studi di semantica teorica e storica.* Adriatica

De Mauro T. (2005). *La fabbrica delle parole: il lessico e problemi di lessicologia*. Torino, UTET

De Mauro T. (2007). *Grande Dizionario Italiano dell'uso*. Torino, UTET.

Harris R. (2003). *The necessity of Artspeak. The language of the arts in the Western tradition.* London, Continuum.

Kilgariff A., Jakubíček M. *et al.* (2014). Finding terms in Corpora for many languages with the Sketch Engine. In *Proceedings of the Demonstrations at the 14th Conference of European Chapter of the Association for Computational Linguistics*, Sweden, April 2014, pp. 53-56

Lyding, V., Stemle, E. *et al.* (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, April 2014.* pp. 36-43.

Masini F. (2009). Combinazioni di parole e parole sintagmatiche. In E. Lombardi Vallauri e L.Mereu, *Spazi linguistici,* Roma, Bulzoni, pp. 191-209

Voghera M. (1994). Lessemi complessi: percorsi di lessicalizzazione a confronto. *Lingua e stile* 28, pp. 185-214.

Voghera M. (2004). Le polirematiche. La formazione delle parole. In M. Grossmann e F. Rainer. Tübingen, Max Niemeyer Verlag, pp. 56-68.

Zgusta L. (1971). *Manual of Lexicography.* Prague-The Hague-Paris, Academia/Mouton.

"The grief that doesn't speak": Text Mining and Brain Structure

Daniela Laricchiuta[1], Francesca Greco[2], Fabrizio Piras[3], Barbara Cordella[4], Debora Cutuli[5], Eleonora Picerni[6], Francesca Assogna[7], Carlo Lai[8], Gianfranco Spalletta[9], Laura Petrosini[10]

[1]Sapienza University, IRCCS Fondazione Santa Lucia – daniela.laricchiuta@uniroma1.it
[2]Sapienza University of Rome, Prisma S.r.l. – francesca.greco@uniroma1.it
[3]IRCCS Fondazione Santa Lucia - f.piras@hsantalucia.it
[4] Sapienza University of Rome – barbara.cordella@uniroma1.it
[5]Sapienza University of Rome, IRCCS Fondazione Santa Lucia – debora_cutuli@yahoo.it
[6]Sapienza University, IRCCS Fondazione Santa Lucia – eleonora.picerni@uniroma1.it
[7]IRCCS Fondazione Santa Lucia - f.assogna@hsantalucia.it
[8]Sapienza University of Rome – carlo.lai@uniroma1.it
[9]IRCCS Fondazione Santa Lucia – g.spalletta@hsantalucia.it
[10]Sapienza University of Rome, IRCCS Fondazione S. Lucia – laura.petrosini@uniroma1.it

Abstract

Contemporary neurosciences have shown that emotions, thought and language involve the functioning of connected brain areas, which allow the recognition and expression of one's own feelings. The scope of this pilot study is to investigate the link among the verbal expression of emotional experiences (assessed with the Toronto Structured Interview for Alexithymia - TSIA -), the linguistic structure and the brain structure. To this aim, 9 healthy adult subjects of both sexes were interviewed by means of the TSIA and the cortical and subcortical structural measures were detected. The TSIA transcripts were analysed by using a cluster analysis and, subsequently, a correspondence analysis, and the values of factors were correlated with cortical and subcortical structural measures as well as TSIA scores, evidencing significant associations. The study highlighted that in healthy subjects it is possible to identify a link between the manner in which people express their experiences, recognize and use their emotions and the brain structural correlates.

Abstract

Le neuroscienze contemporanee hanno evidenziato come le emozioni, il pensiero e il linguaggio coinvolgono il funzionamento di aree cerebrali differenti connesse tra loro, le quali consentono il riconoscimento e l'espressione dei propri sentimenti. Scopo di questo studio pilota è di indagare il nesso tra l'espressione verbale delle proprie esperienze emotive (valutata con la Toronto Structured Interview for Alexithymia – TSIA -), la

struttura del linguaggio utilizzato e la struttura cerebrale. A questo scopo 9 soggetti sani di entrambi i sessi sono stati intervistati con la TSIA e sono state rilevate le misure strutturali corticali e sottocorticali. Le interviste sono state sottoposte ad analisi dei cluster e successivamente ad analisi delle corrispondenze e i valori dei fattori sono stati correlati con le misure strutturali corticali e sottocorticali e con i punteggi della TSIA. I risultati evidenziano associazioni significative, che mettono in luce come in soggetti sani sia possibile individuare un nesso tra il modo in cui le persone raccontano le proprie esperienze, la loro capacità di riconoscere e utilizzare le loro emozioni e la loro struttura cerebrale.

Keywords: Text mining, brain imaging, alexithymia, TSIA

1. Introduction

According to the multiple code theory, the emotional information is represented in verbal, non-verbal symbolic and non-verbal sub-symbolic multiple systems (Bucci, 1997). The verbal system is a communication and reflection code through which the emotional, private and subjective experience can be shared with others. It refers to the capacity of language to direct and regulate ourselves, activate imagination and emotions, stimulate actions and control them. The multiple channels of the non-verbal systems include representations and proceedings related to implicit elaboration associated with visceral, somatic, sensory and motor modalities. While in the non-verbal symbolic system the information is processed in images, in the sub-symbolic one, rapid and complex computations are carried out in an implicit continuous path. These computations contribute to recognize slight facial expressions modifications, identify body movement or vocal quality changes, and distinguish visceral states. The multiple code theory is in line with the contemporary neurosciences (LeDoux, 2012; Damasio et Carvalho, 2013), suggesting that in the presence of the affective experience it is possible to discriminate between emotions and feelings. Emotions occur at a physiological and motor-expressive level, involving bodily systems and subcortical and cortical somato-sensory brain areas. Feelings are based on complex symbolization and cognitive processes related to the functioning of prefrontal and associative cortices. The integration of emotions and feelings, as well as of the verbal and non-verbal systems, depends on the so-called referential processes, which transform the non-verbal symbolic and sub-symbolic materials into words, and *vice versa*.

The referential processes are the core factors contributing to the development, maintenance and promotion of health, since a deficit in these processes generates dysfunctional conditions and pathologies, characterized

by multifactorial and bio-psycho-social etiology as well as marked somatization. Among these dysfunctional conditions, alexithymia is a psychological construct represented by impairment in cognitive-emotional and affective processing (Bagby et al., 1994). It describes people with deficiencies in identifying or describing subjective emotions or feelings, difficulty in distinguishing between bodily sensations of emotional arousal and feelings, and limited affect-related fantasy and imagery. People with alexithymic traits have a tendency to focus on facts without affective involvement rather than inner experiences, exhibiting a "concrete and reality-based cognitive style". They often avoid social situations, seem cold, show a lack of intimacy and warmth and are insecurely attached to others. Although alexithymia is not a psychological disorder *per se*, it is associated with a low quality of life and enhanced risk of psychological impairment and it is present in a broad spectrum of psychosomatic disorders (Taylor et Bagby, 2004). Neuroimaging studies have indicated that people with high alexithymic traits show less activation in brain areas associated with emotional awareness and volumetric variations in brain areas associated with emotional and somato-sensory and sensory-motor processing (Laricchiuta et al., 2015a, and see for a literature review Laricchiuta et al., 2015b). Therefore, the aim of this pilot study is to investigate a complex bio-psycho-social pattern of verbal expression of the emotional experiences, alexithymia levels and brain structure.

2. Data collection and analysis

A sample of 9 (males=5) healthy adult subjects of both sexes was recruited for the pilot study at the IRCCS Fondazione Santa Lucia, Rome. Participants were selected according to the following inclusion criteria: age between 18 and 70 years and suitability for structural Magnetic Resonance Imaging (MRI) scanning. Exclusion criteria included the suspicion of cognitive impairment or dementia; the subjective complaint of memory difficulties or of any other cognitive deficit, regardless of interference with daily activities; major medical illnesses; current or reported psychiatric or neurological disorders; known or suspected history of alcoholism or drug dependence and abuse; and MRI evidence of focal parenchymal abnormalities or cerebro-vascular diseases.

To assess the cortical and subcortical structural measures, participants underwent an imaging protocol that included standard clinical sequences (FLAIR, DP-T2-weighted) and a volumetric whole-brain 3D high-resolution T1-weighted sequence, performed with a 3 T Allegra MR imager, with a standard quadrature head coil. Volumetric whole-brain T1-weighted images were obtained in the sagittal plane using a Modified Driven Equilibrium

Fourier Transform (MDEFT) sequence (Echo Time/Repetition Time -TE/TR- = 2.4/7.92 ms, flip angle 15°, voxel size 1 x 1 x 1 mm3). All planar sequence acquisitions were obtained in the plane of the anterior-posterior commissure line. For the volumetric measures, T1-weighted images were processed and examined using the SPM8 software, specifically the VBM8 toolbox running in Matlab 2007b. For the cortical thickness, FreeSurfer imaging analysis suite (v5.1.0) was used for cortical reconstruction of the whole brain. The segmented, normalized, modulated and smoothed images were used for analyses. Then, participants were interviewed by using the Toronto Structured Interview for Alexithymia (TSIA, Bagby et al., 2006; Italian version Caretti et al., 2011), composed of 24 items referred to four factors of the alexithymia construct: the Difficulty in Identifying Feelings (DIF); the Difficulty in Describing Feelings (DDF); the Externally Oriented Thinking (EOT); and the Imaginal Processes (IP). Each item is assessed by a specific open-ended question and its response is 3-point scored (coded '0', '1', or '2'). The sum of scores results in a total score that ranges from 0 (low alexithymia levels) to 48 (high alexithymia levels). The transcripts of the TSIA responses were used to evaluate the linguistic structure by means of a multivariate analysis. Namely, the nine TSIA transcripts resulted in a medium size corpus of 62.792 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the hapax percentage (TTR= 0,10; Hapax= 51,0%). According to the large size of the corpus both lexical indicators highlighted its richness and indicated the possibility to proceed with the analysis. First, data were cleaned and pre-processed with the software T-Lab (Lancia, 2017) and keywords selected. In particular, we used lemmas as keywords instead of type, filtering out the lemma of the high rank of frequency and those of the low rank of frequency lower to 9 occurrences (for keyword election see Cordella et al., 2014; Greco, 2016). Then, on the context units per keywords matrix, we performed a cluster analysis with a bisecting k-means algorithm (Savaresi et Boley, 2004) limited to ten partitions, excluding all the context units that do not have at least two keywords co-occurrence. To finalize the text mining a correspondence analysis (Lebart et Salem, 1994) on the keywords per clusters matrix was made in order to explore the relationship between clusters and to identify the latent dimensions setting the interviews.

Then parametric associations between TSIA scores and regional volumes or cortical thickness, and between lexical scores (resulted by the correspondence analysis) and TSIA scores or brain structural measures were calculated by means of Pearson's correlations in order to identify the possible direction and extent of the linear relationship between the variables.

3. Main results

The results of the cluster analysis show that the 369 keywords selected allow for the classification of 96.8% of the corpus. According to the theoretical framework (Cordella et al., 2014) we choose the solution with four cluster. The correspondence analysis detected three latent dimensions. In table 1, we can appreciate how the clusters are placed in the factorial space produced by three factors. The first factor represents the experience that could be personal (negative pole) or social (positive pole); the second factor reflects the thought that could be made on feelings (negative pole) or on a rational reasoning (positive pole); and the third factor represents the aim of the thinking process that could lead to make a speculation (negative pole) or a choice (positive pole).

Table 1 – Cluster coordinates on factors (the percentage of explained inertia is reported between brackets under each factor).

Cluster	Label	CU classified	Factor 1	Factor 2	Factor 3
			Experience	Thought	Aim
1	Think	34,24%		Reasoning	Speculate
			0,03	0,46	-0,64
2	Feel	18,09%	Personal	Feeling	Speculate
			-0,13	-0,65	-0,18
3	Relationship	23,15%	Social		Choice
			0,52	0,08	0,24
4	Remember	24,51%	Personal	Reasoning	Choice
			-0,73	0,23	0,30

CU = context units classified.

The four clusters are of different sizes and reflect the general approach to the emotional experience solicited by the TSIA. The first cluster reflects the reasoning on the life event and hypothesis resulting in a rationale thinking process; the second cluster highlights the capacity to reflect on the experience of personal feelings; the third cluster represents the relationships characterizing social life; and the fourth cluster gets back to memories, reasoning on personal choices that were made (table 2).

Table 2 – Cluster (the percentage of context units classified in the cluster is reported between brackets).

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Think		Feel		Relationship		Remember	
keyword	CU	keyword	CU	keyword	CU	keyword	CU
pensare	104	sentire	100	persona	163	immaginare	84
vedere	71	proprio	77	parlare	163	vedere	71
scrivere	45	riuscire	77	sentimento	149	prendere	64
amico	38	momento	67	provare	116	mettere	61
chiedere	32	capire	66	persone	114	positivo	39
chiamare	26	situazione	51	capire	100	casa	31
trovare	26	vivere	50	cercare	80	bello	28
tempo	21	piacere	50	situazione	71	problemi	27
ragazzo	20	succedere	33	amico	70	tornare	27
diverso	20	rabbia	32	trovare	45	portare	24

CU = context units classified in the cluster.

The correlation coefficient between the lexical scores of the three factors (resulted from the correspondence analysis) and the TSIA scores, as well as the brain volumes and the cortical thickness are reported in table 3. Namely, DIF, DDF, EOT and total alexithymia scores were positively associated with the second factor. At neurobiological level, the first factor was negatively associated with volumes of right caudate and thickness of right medial orbitofrontal cortex, as well as positively associated with the thickness of right lateral occipital cortex. Finally, the third factor was negatively associated with volumes of middle anterior, central and middle posterior cerebral cortices, as well as with thickness of right postcentral cortex and left posterior cingulate cortex. Conversely, the third factor was positively associated with thickness of the right posterior cingulate cortex. Finally, the IP scores (TSIA) were positively correlated ($r = 0.72$; $p = 0.03$) with the left entorhinal cortical thickness values.

4. Discussion

Although this is a pilot study and it is not possible to generalise the findings, the present data suggest that the methodology proposed (in order to identify the connections among verbal expression, alexithymia levels and brain structure) seems to be promising for a deeper understanding of the bio-psycho-linguistic connections. In fact, results indicate that high alexithymia scores were associated with a thought modality characterized by a rational (and not emotional) reasoning. Furthermore, the tendency to be engaged in

personal (not social) experience was associated with large volumes of right caudate and thickness of right medial orbitofrontal cortex.

Table 3 – Correlation coefficients between lexical factors and TSIA scores as well as cerebral structure values.

Variables	Factor 1	Factor 2	Factor 3
Difficulty Identifying Feelings (TSIA Factor 1)		.83	
Difficulty Describing Feelings (TSIA Factor 2)		.71	
Externally Oriented Style of Thinking, (TSIA Factor 3)		.68	
TSIA Total score		.77	
Right-Caudate	-.71		
Right Hemisphere Medialorbitofrontal Thickness	-.78		
Right Hemisphere Lateraloccipital Thickness	.70		
Left Hemisphere Posteriorcingulate Thickness			-.71
Mid Posterior Cortical Cortex			-.75
Mid Anterior Cortical Cortex			-.75
Central Cortical Cortex			-.78
Right Hemisphere Postcentral Thickness			-.78
Right Hemisphere Posteriorcingulate Thickness			.68

In the table are reported only the correlation coefficients with a p<0,05.

Conversely, the tendency to be engaged in social (not personal) experiences was associated to great thickness of right lateral occipital cortex. The speculative thinking processes (negative pole of the third factor) was associated with large volumes of middle anterior, central and middle posterior cerebral cortex, as well as with great thickness of right postcentral cortex and left posterior cingulate cortex. Finally, thinking processes related to a choice was associated with great thickness of the right posterior cingulate cortex.

Overall the study indicates that the organizational factors of thought and language, conveying also the emotional meaning of the text, are related to the structure of cerebral areas involved in somato-sensory associative processes (postcentral and lateral occipital cortices), in emotional awareness (entorhinal and posterior cingulate cortices), and in emotional control and feelings (orbitofrontal cortex). Just such functions are compromised in the presence of high levels of alexithymia, because an altered referential process can lead to somato-sensorially perceive but not-verbally express the emotions.

Furthermore, in the present study most of associations were found between first and third factor (resulted from the correspondence analysis) and the macro-structural measures in the right brain hemisphere, totally fitting the

proposal of Bucci (1997) that suggests the right hemisphere as the neurophysiological substratum underlying the processing of emotional information and referential process. On this vein, alexithymia may be considered an embodiment process related to altered perception of physiological correlates (viscero- and somato-motor responses) of the emotional activation resulting in a deficit in the emotional awareness. In fact, a dysfunctional referential process can lead to a lack of words for the emotions, up to being without symbols for the somatic states.

References

Bagby R.M., Parker J.D. and Taylor G.J. (1994). The twenty-item Toronto Alexithymia Scale--I. Item selection and cross-validation of the factor structure. *J Psychosom Res*. 38(1):23-32.

Bagby R.M., Taylor G.J., Parker J.D. and Dickens S.E. (2006). The development of the Toronto Structured Interview for Alexithymia: item selection, factor structure, reliability and concurrent validity. *Psychother Psychosom*. 75(1):25-39.

Bucci W. (1997). Psychoanalysis and cognitive science: A multiple code theory. Guilford Press.

Caretti V., Porcelli P., Solano L., Schimmenti A., Taylor G.J. and Bagby R.M. (2011). Reliability and validity of the Toronto Structured Interview for Alexithymia in a mixed clinical and nonclinical sample from Italy. *Psychiatry Research,* 187:432-436.

Cordella B., Greco F. and Raso A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In Nee E., Daube M., Valette M. and Fleury S., editors, *Actes JADT 2014 (12es Journées internationales d'Analyse Statistque des Données Textuelles, Paris, France, Juin 3-6, 2014),* pp. 173-184.

Damasio A. and Carvalho G.B. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nat Rev Neurosci*., 14(2):143-152.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Lancia F. (2017). *User's Manual: Tools for text analysis*. *T-Lab version Plus 2017*.

Laricchiuta D., Petrosini L., Picerni E., Cutuli D., Iorio M., Chiapponi C., Caltagirone C., Piras F. and Spalletta G. (2015a). The embodied emotion in cerebellum: a neuroimaging study of alexithymia. *Brain Struct Funct.,* 220(4):2275-2287.

Laricchiuta D., Lai C. and Petrosini L. (2015b). Alexithymia: From Neurobiological Basis to Clinical Implications. In Bryant M.L., editor, *Handbook on Emotion Regulation: Processes, Cognitive Effects and Social Consequences,* Nova Science Publishers.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

LeDoux J. (2012). Rethinking the emotional brain. *Neuron,* 73(4): 653-676.

Savaresi S. M. and Boley D. L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis,* 8(4): 345-362.

Taylor G.J. and Bagby R.M. (2004). New trends in alexithymia research. *Psychother Psychosom.,* 73(2):68-77.

Icone gay: tra processi di normalizzazione e di resistenza. Ricostruire la semantica degli hashtag

Gevisa La Rocca[1], Cirus Rinaldi[2]
[1] Kore University of Enna – gevisa.larocca@unikore.it
[2]University of Palermo – cirus.rinaldi@unipa.it

Abstract 1
The mediatization of emotions emerges as an affordance of social media, the study of which involves paying attention to digital practices and the formation of the sense of public affection, of connected audiences expressing their participation through expressions of sentiment. This happens both for the great events and for the daily demonstrations of support or of its negation. Here we choose to analyze the tweets in which the fans express their opinions on the participation in the reality shows of their "icons": Vladimir Luxuria and Cristiano Malgioglio. To reconstruct the hashtag semantics we use: the NodeXL software for network analysis and Iramuteq for the extraction of lexical worlds.

Abstract 2
La mediatizzazione delle emozioni emerge come una affordance dei social media, il cui studio implica il porre attenzione alle pratiche digitali e alla formazione del senso dell'affetto pubblico, dei pubblici connessi che esprimono la loro partecipazione attraverso le espressioni del sentimento. Questo accade tanto per i grandi avvenimenti che per le quotidiane manifestazioni di sostegno o della sua negazione. Qui si sceglie di analizzare i tweets in cui i fans si esprimono in merito alla partecipazione ai reality show delle loro «icône» : Vladimir Luxuria e Cristiano Malgioglio. Per ricostruire la semantica degli hashtag si usa il software NodeXL per la newtork analysis e Iramuteq per l'estrazione dei mondi lessicali.

Keywords: LGTB, rappresentazione sociale, semantica degli hashtag, network analysis.

1. Introduzione

All'interno della più generale questione della rappresentazione dei tipi gay (Dyer, 1992), la riflessione teorica ha rilevato la complessità che si cela dietro una facciata che è solo apparentemente semplice. In particolare, la costruzione di tipi e icone LGBT offre codificazioni simboliche ampie e sovrasignificazioni persino contradditorie, dal momento che la produzione

culturale subculturale e comunitaria gay non soltanto ha da sempre articolato temi e significati già esistenti (Hall, 1996), ma è stata a sua volta incorporata in rappresentazioni eteronormative. Guardando alle principali trasformazioni sociali in atto relative ai processi di acquisizione della cittadinanza nei contesti neo-liberisti e alle istanze di «normalizzazione» tuttora in discussione all'interno della sfera pubblica e del movimentismo LGBTQI[1] (matrimonio egualitario, coppie non-bianche, dis-abilità e coscrizione militare), bisogna prestare attenzione ai processi di costruzione simbolica di (nuove e inedite) maschilità complici assimilate all'interno della produzione del consumo neo-liberista[2]. Il neo-liberismo pone la propria enfasi sulla libertà individuale e sui diritti, e sulla regolazione degli interventi centrali dello Stato: una delle principali modalità di *governance* neo-liberista consiste nelle diverse tecniche di normalizzazione, attraverso le quali vengono individuate e riprodotte norme di comportamento specifiche indicate dai governi liberisti che i "cittadini" dovranno interiorizzare a fini auto-regolativi[3]. Di particolare interesse per la presente trattazione, i modelli di cittadinanza sostenuti sono finalizzati a rafforzare le dimensioni normative di genere e sessualità (matrimonio, coscrizione militare, etc.); la rivendicazione di «eguaglianza» si è fortemente intrecciata con la negazione e il distanziamento da parte degli omosessuali dalle tradizionali rappresentazioni che li associavano a individui «amorali», «inferiori», «subordinati», e «peccaminosi». La *tentazione di essere normali* sta producendo effetti paradossali sulla visibilità dei gay, perché se da un lato assistiamo a quanto viene definito *de-omosessualizzazione* o *eterosessualizzazione* delle omosessualità[4], dall'altro invece, osserviamo un rafforzamento delle dicotomie e delle gerarchie di genere che circoscrivono fortemente l'omosessualità attraverso l'interiorizzazione di un'economia eteronormativa

[1] Per un approfondimento si rinvia a Warner, M. (1999). *The trouble with the normal. Sex, politics and the ethics of queer life*, New York, The Free Press.

[2] Cfr. Duggan, L. (2003). *The Twilight of Equality?: Neoliberalism, Cultural Politics, and the Attack on Democracy*, Boston, Beacon Press e Chasin, A. (2000). *Selling Out: The Gay and Lesbian Movement Goes to Market*, New York, St. Martin's Press.

[3] Richardson, D. (2004). *Locating sexualities: from here to normality*, «Sexualities», 7, 4, 391- 411.

[4] Bersani, L. (1998). *Homos* (1995), tr. It. *Homos*, Milano, Pratiche Editrice, 38, 131. Sui processi di eterosessualizzazione, si consideri la critica fornita in Ingraham, C. (2012), *Atti innaturali: disciplinare l'eterosessualità*, in C. Rinaldi (ed.). *Alterazioni. Introduzione alle sociologie delle omosessualità*, Milano-Udine, Mimesis, 97-119.

del desiderio[5]. Il presente lavoro intendere problematizzare le forme di tipizzazione LGBT a partire dall'analisi delle differenti rappresentazioni che hanno avuto le partecipazioni ai reality show di personaggi pubblici, dichiaratamente parte della comunità LGBT, come Vladimir Luxuria e Cristiano Malgioglio.

2. Processi di normalizzazione e di resistenza

I processi di assimilazione eteronormativa da parte delle comunità gay, e la conseguente costruzione normalizzante di alcune soggettività LGBT, hanno portato a nuove gerarchie di desiderabilità all'interno delle medesime comunità LGBT. Uno dei principali effetti è la depoliticizzazione e la privatizzazione della cultura movimentistica LGBT, ancorata nella riproduzione e nella protezione del domestico (la famiglia/le famiglie e la nazione o il Nuovo Ordine Globale), nella riproduzione iperbolica della dicotomia di genere normativa, nel consumo di tali rappresentazioni e nella partecipazione attiva al nuovo ordine globale. A corroborare questa prospettiva, i processi politici globali in atto (sia di tipo securitario che di valorizzazione di alcune soggettività) – insieme alla globalizzazione delle maschilità nelle forme della loro mercificazione e delle immagini veicolate dai mass media transnazionali – sembrano fondarsi sulla definizione anche di standard omonormativi e omonazionalisti che non soltanto si contrappongono al nemico da combattere (la figura ambigua dell'indecoroso, del terrorista o dell'immigrato) ma che si discostano, nel contempo, da tutte quelle corporeità terroristiche che lo standard nazionalista individua come eccessive/eccedenti (a meno che anche esse non scelgano di diventare docili). Uno degli effetti principali delle nuove forme di normalizzazione, è un diretto rafforzamento degli assetti di genere normativi, un rafforzamento della maschilità e dei suoi indicatori in termini generali e, soprattutto, una diretta complicità nella marginalizzazione degli omosessuali effeminati non soltanto nella società più vasta a ma anche all'interno delle comunità gay.

La costruzione delle nuove maschilità si fonda principalmente sull'epurazione del femminile e dell'effeminatezza, attraverso una serie di contesti e pratiche – dalle rappresentazioni nei mass media, al marketing della «pink economy» sino agli annunci personali nei social network, nelle app per incontri – che attestano quanto le maschilità omosessuali egemoni e normative considerino accettabili e giustificabili le condotte e gli

[5] Martino, W. (2006). *Straight-acting masculinities: normalization and gender hierarchies in gay men's lives*, in Kendall, C., Martino, W. (eds.), *Gendered outcasts and sexual outlaws. Sexual oppression and gender hierarchies in queer men's lives*, The Haworth Press, New York, 35-60.

atteggiamenti effeminofobi. La storica associazione dell'omosessualità con l'effeminatezza e con il femminile di fatto non stimola un ripensamento della maschilità omosessuale (e delle forme di strutturazione della maschilità in generale), ma la fa arroccare su posizioni difensive e distanzianti e su forme contro-reattive e contro-culturali che, di fatto, riproducono la norma. La paradossalità della maschilità omosessuale consiste nella ricerca dell'affrancamento culturale attraverso la proposta di una maschilità vera, senza metterne in discussione però misoginia, patriarcato e checcofobia. Una maggiore visibilità delle omosessualità implica un costo da pagare, nello specifico una differenza che si dissolve in eguaglianza e nella costituzione di un blocco egemonico in grado di rendere meno percettivamente visibili le divisioni di genere del patriarcato.
Se pensiamo in termini relazionali le maschilità (eterosessuali e omosessuali), una loro ibridazione giustificata su base consumistica, se permette alle prime di apparire in modo meno tradizionale e rigido e alle seconde di acquisire in risorse simboliche della tradizione virilista, tuttavia riproduce –attraverso entrambe – l'illusione della scomparsa del (dividendo del) patriarcato. Dal momento che entrambe le forme di soggettivazione maschile partecipano in modo selettivo e differenziale all'accaparramento e all'appannaggio di risorse materiali e simboliche egemoniche, esse non saranno definibili esclusivamente in termini oppositivi, piuttosto si alimenteranno e rafforzeranno in maniera reciproca. Questa polarizzazione virilista non soltanto perpetra forme di annichilimento nei confronti della manifestazione della maschilità frocia ma impedisce, di fatto, anche la eventuale costituzione pubblica di maschilità eterosessuali al di fuori delle rappresentazioni eteronormative. Tuttavia, il costo da pagare perché la maschilità gay sia accettata socialmente è la negazione del suo carattere sessuale, della sua diversità in termini di desideri erotici, di pratiche, della materialità della sua espressione sessuale. L'assimilazione prevede, di conseguenza, la riproduzione dei sistemi di divisione di genere, di classe e di età presenti nella società più vasta, attraverso la riproduzione di modelli di consumo esistenti all'interno del più vasto ordine globale.

3. La mediatizzazione e gli hashtag

Con l'introduzione del termine e del concetto di mediatizzazione ci si riferisce a quel processo in base al quale le istituzioni sociali e culturali e le modalità di interazione sono cambiate, cambiano e cambieranno come conseguenza della crescita dell'influenza dei media, tenendo conto, però, delle circostanze, ovvero di come mutano la cultura e la società. Si tratta di quel costante contatto comunicativo con gli altri, la cui esplorazione avviene in modi del tutto inediti (Cardoso, 2008; Boccia Artieri, 2012; Colombo, 2013),

che trasforma la condizione del vissuto in un nuovo orizzonte di senso sociale (Boccia Artieri *et al.*, 2017) producendo una metamorfosi delle relazioni sociali. Si tratta di quell'insieme di pratiche o di *habitus* caratterizzate da una regolarità dell'agire in relazione a specifici bisogni, che porta con sé un intero mondo di capacità, vincoli e potere (Couldry, 2012).
Non è difficile accettare che l'uso di un mezzo di comunicazione per un periodo di tempo continuato determini il carattere stesso della conoscenza da comunicarsi, e che a causa della sua stessa pervasività porti all'emergere di una nuova civiltà, ovvero renda possibile lo strutturarsi di una forma particolare con cui si manifesta la vita materiale, sociale e spirituale di un popolo (Innis, 1951; Ong, 1982; La Rocca, 2017). L'aspetto sociale dei social media è, ormai, evidente da sé; essi sono parte di una società in cui svolgono una pluralità di funzioni di intermediazione (Colombo, 2003; 2013), essi, infatti, sembrano essere pensati per rendere possibili le collaborazioni partecipative, cioè dal basso, e allo stesso tempo, alimentano una socievolezza di simmeliana memoria, che è riassumibile nelle caratteristiche oggi individuate da Peter Dahlgren (2009) nella *talkative society*.
Si tratta di considerare *emoticons*, *emoji*, *hashtag*, commenti, rimandi, foto, link, video, tutti quegli strumenti che consentono di ricollocare il testo nelle intenzioni enunciative di chi lo ha posto in essere o condiviso. Sono fenomeni di cui si occupa con più interesse la *discourse analysis*, ma che non possono essere ignorati se l'obiettivo è un'analisi del contenuto digitale dei *new e social media*. Questo per una ovvia considerazione, com'è possibile limitare l'osservazione solo allo scritto e non estenderlo ai suoi elementi accessori, se l'obiettivo è conoscere il senso di quanto viene detto a proposito di un dato argomento o fenomeno in rete? Solo in questo modo l'analisi del contenuto si apre alla possibilità di includere il linguaggio utilizzato come una meta risorsa tecnologizzata. È chiaro che intesa in questo modo l'analisi del contenuto si avvicina più alla *discourse-ethnografic* (Androutsopoulos, 2010; 2011) piuttosto che a un'analisi delle frequenze; perché innanzitutto ricostruire i percorsi e le emozioni di un *topic* online non è semplice e questo è dovuto alla struttura della grammatica e della sintassi della costruzione dei messaggi, alla commistione linguistica che quindi richiede nel momento dell'analisi un software capace di elaborare testi in più lingue, ma anche di ricodificare le *emoticons* e di valutare in relazione a esse le intenzioni del testo. Si tratta di sviluppare un approccio di *analisi del contenuto multimodale*, indicando con questo termine come anche in questo settore sia necessario attuare ciò che è avvenuto nello studio del discorso (Jewitt, 2014; Kress, van Leeuwen, 2001), dove si presta attenzione al modo in cui il linguaggio interagisce con altri sistemi semiotici; sostituendo al "linguaggio", la costruzione del contenuto, che inevitabilmente interagisce con gli altri sistemi

semiotici. In questo caso è chiaro che un approccio in cui è il ricercatore a svolgere manualmente tutte queste operazioni o a ricodificare le espressioni riconducendole a categorie di senso condivise diventa la soluzione più opportuna. Ma appare altresì veritiero che un lavoro di questo tipo, a meno di ricevere un grosso finanziamento a supporto della ricerca, rimane sempre legato a un numero limitato di osservazioni.

3.1. Per una ricostruzione della semantica degli hashtag

Si sceglie quindi di lavorare in una duplice ottica: da un lato si scaricano i tweets relativi alla partecipazione ai reality show quali: *L'isola dei famosi* e *Grande Fratello VIP* per i due soggetti individuati, e si ricostruisce il network dei topic, dall'altro si esplorano mediante l'analisi lessicale i contenuti dei tweets individuati come perni connettori.

Il ventaglio dei sentimenti che gli individui agganciano a questi messaggi è ampio e variegato e attiene alla natura umana, è possibile interpretarlo seguendo l'impostazione data da Korina Giaxoglou e Katrin Döveling (2018) nella loro *special issue* dedicata alla mediatizzazione delle emozioni sui social media. La mediatizzazione delle emozioni emerge come una *effordance* dei social media il cui studio implica il porre attenzione alle pratiche digitali e alla formazione del senso dell'affetto pubblico, dei pubblici connessi (Boyd, 2010) che esprimono la loro partecipazione attraverso le espressioni del sentimento (Papacharissi, 2016). Finora, gli studi hanno esaminato la formazione e la veicolazione dei sentimenti in rete legandoli a grandi avvenimenti sociali, momenti storici che prefigurano cambiamenti epocali. Si tratta, senza dubbio, di storie di connessione ed espressione, dove gli *hashtag* servono come significanti vuoti che invitano ad una identificazione ideologica a vasto orientamento polisemico (Colleoni, 2013; Papacharissi, 2016). I *post* promossi dai singoli, individuati e volti a sostenere o denigrare, quelle che qui sono state individuate come icône gay, si sostanziano o forse meglio si foraggiano sicuramente di un senso emotivo che è personale.

Come per Zizi Papacharissi (2016) anche le nostre interpretazioni, in questo contesto, sono guidate dalla comprensione dell'affetto come una forma di intensità pre-emotiva soggettivamente sperimentata e connessa a processi di premeditazione o anticipazione di eventi prima del loro verificarsi. Ci sono radici emotive legate alla percezione e sperimentazione dell'affetto che provengono da contesti socioculturali cui gli individui appartengono, in questo senso le emozioni mediatizzate sono delle forme espressive di culture più profonde. *Per realizzare l'analisi dei network degli hashtag si utilizza NodeXL, per l'analisi lessicale Iramuteq.*

References

Androutsopoulos J. (2010). Localising the global on the participatory web: Vernacular spectacles as local responses to global media flows. In Coupland, N. (ed.), *Handbook of Language and Globalization* (203-231). Oxford: Wiley-Blackwell.

Androutsopoulos J. (2011). From variation to heteroglossia in the study of computer-mediated discourse. In Thurlow, C. e Mroczek, K. (eds.), *Digital Discourse: Language in the New Media* (277–298). London: Oxford University Press.

Boccia Artieri G. (2012). *Stati di connessione*. Milano: Franco Angeli.

Boccia Artieri G., Gemini L., Pasquali F., Carlo S., Farci M., Pedroni M. (2017). *Fenomenologia dei social network. Presenza, relazioni e consumi mediali degli italiani online.* Milano: Guerini.

Boyd D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Papacharissi, Z. (ed.), *A Networked Self: Identity, Community, and Culture on*

Social Network Sites (39-58). New York: Routledge.

Cardoso G. (2008). Preference for online social interaction: A theory of problematic Internet use and psychosocial wellbeing. *Communication Research*, 30: 625-648.

Chasin A. (2000). *Selling Out: The Gay and Lesbian Movement Goes to Market.* New York: St. Martin's Press.

Colleoni E. (2013). Beyond the differences: The use of empty signifiers as organizing device in the #occupy movement. In *Proc. Of Material Participation: Technology, the Environment and Everyday Publics*, Università di Milano, Maggio.

Colombo F. (2003). *Introduzione allo studio dei media.* Roma: Carocci.

Colombo F. (2013). *Il potere socievole: storia e critica dei social media.* Milano: Bruno Mondadori

Couldry N. (2012). *Media, Society, World: Social Theory and Digital Media Practice.* Cambridge: Polity.

Dahalgren P. (2009), *Media and Political Engagement. Citizens, Comunication, and Democracy*. New York: Cambridge University Press.

Duggan L. (2003). *The Twilight of Equality?: Neoliberalism, Cultural Politics, and the Attack on Democracy.* Boston: Beacon Press.

Giaxoglou K., Döveling K. (2018). Mediatization of emotion on social media: forms and norms in digital mourning practices. *Social Media + Society*, pp. 1-4.

Ingraham C. (2012). Atti innaturali: disciplinare l'eterosessualità. In Rinaldi, C. (Ed.), *Alterazioni. Introduzione alle sociologie delle omosessualità* (97-119). Milano-Udine: Mimesis.

Innis H. A. (1951). *Empire and communication.* Oxford: University of Oxford Press.

Jewitt C. (2014). *The Routledge Handbook of Multimodal Analysis.* London: Routledge.

Kress G., Van Leeuwen T. J. (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication.* London: Arnold.

La Rocca G. (2017). Cantami o diva. La bazzecola del dicunt in rete. *Sociologia della comunicazione*, 53: 56-74.

Martino W. (2006). *Straight-acting masculinities: normalization and gender hierarchies in gay men's lives.* In Kendall, C., Martino, W. (Eds.), *Gendered outcasts and sexual outlaws. Sexual oppression and gender hierarchies in queer men's lives* (35-60). New York: The Haworth Press.

Ong W. J. (1982). *Orality and Literacy. The Tecnologizing of the Ward.* London: Routledge.

Papacharissi Z. (2016). Affective publics and structures of storytelling: Sentiment, events and mediality. *Information, Communication & Society*, 19: 307–324.

Puar J. (2007). *Terrorist Assemblages: Homonationalism in Queer Times.* Durham, N.C.: Duke University Press.

Richardson D. (2004). Locating sexualities: from here to normality. *Sexualities*, 7(4): 391- 411.

Rinaldi C. (2012). Globalizzazione della maschilità e maschilizzazione dei processi globali. In G.E.M. Scichilone (Ed.), *L'era globale: linguaggi, paradigmi, culture politiche* (173-189). Milano: Franco Angeli.

Rinaldi C. (2007). De-gener(azioni): riflessioni per una sociologia del transgenderismo. In Antosa, S. (a cura di), *Spazio e identità queer* (127-148). Omosapiens vol. 2. Roma: Carocci.

Looking for *topics*: a brief review

Ludovic Lebart[1]
[1]Telecom-ParisTech – ludovic@lebart.fr

Abstract
This paper presents a brief review of several endeavors to identify latent variables (axes or clusters). When dealing with textual data, these latent variables (clusters or axes) are sometimes designated *ex ante* by the term "topic". The first attempts to identify interpretable latent variables dates back to factor analysis at the beginning of last century. Recent years have witnessed a series of algorithmic attempts such as non-negative matrix factorization (NMF) or Latent Dirichlet Allocation (LDA). In the meantime, latent variables are also identified through several hybridizations and synergies of principal axes methods and clustering techniques. A single medium-sized classical corpus (Shakespeare's 154 Sonnets) will serve as a benchmark to sketch and compare in a compact way some characteristic features of several methods.

Keywords: Topic Modelling, NMF, LDA, Correspondence Analysis, Factor Analysis, Clustering.

1. Introduction

There is a profusion of new disciplines around the industrial applications involving texts, with subsequent proliferations of tools and disparities of terminologies. There are also disparities in the attitude towards the texts, sometimes influenced by the availability and the user-friendliness of software. The problems entailed by huge sets of newsgroup or tweets are quite different from those encountered when dealing with literature, political discourses, psychological surveys. Because they are well known, translated in almost every language, deeply studied and commented, we will use Shakespeare's Sonnets as a benchmark to briefly compare the ability of several techniques to recognize topics in a corpus.

2. An outline of the contents of Shakespeare's sonnets

The 154 sonnets of William Shakespeare deal with themes such as love, friendship, effects of time, beauty, treason, lust, death. Note that the definition of topics in Text Mining is a pragmatic one, and may also recover the concepts of theme and motif.

2.1. Theme, Topic, Motif

Usually, a topic is an objective explanation of the subject matter, whereas a theme represents the deeper underlying message. A motif is simply a recurring idea or pattern used to reinforce the main theme. Schematically, topics answer the questions:, "What's the story about? Who? What? How?" and themes answer: "Why was the story written?". Topics in literature are easier to identify than themes.
Three main contiguous series of sonnets are generally recognized as three dominant themes:
Sonnets 1 to 17: (*Procreation*). These sonnets celebrate the beauty of a young man who is urged by the poet to marry so as to perpetuate that beauty.
Sonnets 18 to 126: (*Young man*). This longest sequence concerns the same young man (not definitively identified), the destructive effect of time, the force of love, friendship and poetry.
Sonnets 127 to 154: (*Dark Lady*).These sonnets are mostly addressed to a dark haired woman, not without some irony and cynicism (the two last sonnets 153 and 154 are specific epigrams in an ancient style; they should deserve in fact a specific category).

2.2. Eight themes derived from expert commentaries

The themes *Young man* and *Dark lady* could contain five sub-themes. While the first theme (*Procreation*) remains untouched, the new *Young man* and *Dark lady* themes will comprise only those sonnets which were not assigned to the five new categories below (*Absence, Storm, Rivalry, Death, Eternal poetry*).

Table 1. List of eight a priori *themes/topics with the corresponding sonnets numbers*

Procreation	1 - 17
YoungMan	20-25, 33-38, 40-42, 46, 47, 49, 53-55, 59-60,62-70, 75-77, 88-106, 108-112, 115-125,
DarkLady	127-136, 139, 140, 143-146, 153,154
Absence	26-32, 39, 43-45, 48, 50-52, 56-58, 61, 113-114
Storm	141,142,147-152
Rivalry	78-87
Death	71-74
Etern_poetry	18,19,81

The partition of sonnets given in Table 1 is inspired by the works of Alden (1913) and Paterson (2010) but not explicitly mentioned by these authors. Figure 1 shows however that, after a blind correspondence analysis ignoring these themes, most of their locations are statistically significant on the principal plane of visualization.

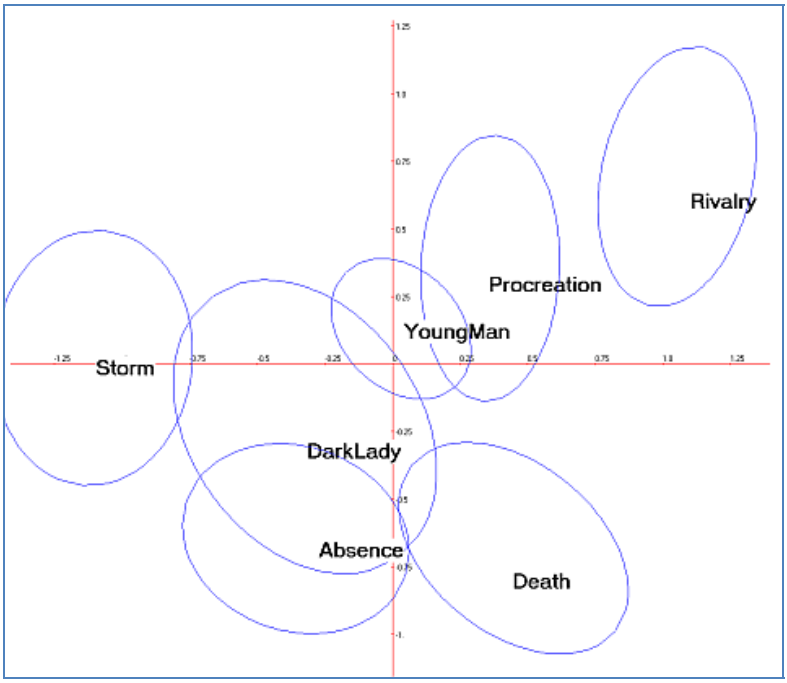


Figure 1. Locations of 7 themes/topics in the principal plane of the correspondence analysis of the lexical table (154 sonnets x 173 words, [min. frequency = 10]) as supplementary categorical variables. Conservative bootstrap confidence ellipses [drawing with replacement of sonnets] show the significant distances between several pairs of a priori *themes. Note that the theme "Eternal Poetry", too much overlapping with others, is missing in this graphical display.*

Evidently, the following attempts to find topics into the corpus of sonnets will ignore that *a priori* partition into themes. We do not expect either to retrieve automatically these themes.
However, the knowledge of these themes issued from literary criticism can provide us with a template for reading and interpreting the results more easily. Note that statistical tools mainly based on frequencies detect almost indifferently topics, themes or motifs.

3. Six selected methods for topic research

Among the six procedures selected in the present paper, three (RFA, FCA, LOA, LSA) make use of the Singular Values Decomposition (SVD). The remaining two methods (NMF and LDA), less geometrical, involve a specific model and more complex algorithms.
RFA (Rotated Factor Analysis) is historically the first attempt to identify

unobserved "latent factors" (Thurstone, 1947, after the pioneering papers of Spearman,1904, and of Garnet, 1919). RFA involves SVD in one of the most popular algorithms known as Principal Factor Analysis. In this case, the *topics* are the words characterizing each kept factors. Initially conceived for numerical values, it could be adapted to sparse frequency tables. [**R** packages 'psych' and 'GPArotation'].

FCA (Fragmented Correspondence Analysis), in the vein of *ALCESTE* methodology (Reinert, 1986), is based on the CA of fragments of texts [in our case 7 consecutive lines, i.e.: half a sonnet), cf. Lebart (2012). The principal axes of CA serve to cluster these fragments (hybrid clustering using Hierarchical Classification –Ward criterion – and k-means). At the end of the process, the *topics* are defined by the series of words that characterize each cluster (software 'DtmVic').

LOA (LOgarithmic Analysis) (Kazmierczak, 1985) is similar to Spectral Mapping (Lewi, 1976) thanks to a difference of weighting. Both methods, like CA, comply with the principle of distributional equivalence (stability of the results vis-à-vis fusions of similar columns or rows). Applied to contingency or frequency tables, LOA often produces results similar to those of CA, with less sensitivity towards outliers as a consequence of the logarithmic shrinkage. A clustering of sonnets (similar to that of FCA) is then performed. The *topics* are then the words characterizing each cluster (software: 'DtmVic').

LSA (Latent Semantic Analysis (or Indexing), Deerwester *et al.*, 1990) which is basically a SVD of the matrix of Tf.Idf coefficients (Term frequency x Inverse of document frequency). A clustering of sonnets (similar to that of FCA and LOA) is then performed. The *topics* are then the words characterizing each cluster. (**R** package 'lsa' [Fridolin Wild] and 'DtmVic')

In the domain of text analysis, the two following methods belong more specifically to the field of "Topic Modelling".

NMF (non-negative matrix factorization) starts with an equation that reminds Singular Values Decomposition (SVD): Decomposition of a data matrix **A** as the product of two matrices of lower rank, **B** and **C**: **A** = **B** **C**. The marked difference lies in a constraint of positivity of the coefficients of **B** and **C** (those of **A** being already supposed > 0) (Lee and Seung, 1999; Berry *et al*, 2007; after Paatero &Tapper, 1994. See also Gaujoux, 2010). In the topic modeling context, the main output of NMF is a set of topics characterized by list of words (software 'scikit-learn' [Python] by Grisel O., Buitinck L., Yau C.K; In: Pedregosa *et al.*, 2011).

LDA (Latent Dirichlet Allocation) (Blei et al., 2003; Griffiths et al., 2007) is a generative statistical model (involving unobserved topics, words, and document) devised to uncover the underlying semantic structure of a

collection of texts (documents, supposed to be a mixture of a small number of topics). The method is based on a hierarchical Bayesian analysis of the texts. (package **R**: 'topicmodels', and software 'scikit-learn' [Python]).
At this stage, we have limited our investigation to six techniques out of a great number of approaches likely to identify topics. Among these approaches let us mention the direct use of CA without fragmentation of the texts, the techniques of clustering (used in FCA and LOA) which contain many more methods and variants, the already mentioned *Alceste* methodology (Reinert, 1986). The present piece of research evidently needs to be extended. In fact, each method involves also a series of parameters (threshold of frequency for the words; preprocessing options such as lemmatization/stop words; size of fragments or context units, number of iterations). The following experiment limited to six methods will be tersely summarized. A thorough investigation would need many more pages.

4. Excerpts from the list of 49 topics (limited to two topics per method)
The number of topics detected by each of the six selected methods varies between six and ten. Only two topics are printed below for each method.
4.1 Rotated Factor Analysis (Rotation Oblimin). *(2 topics out of 6)*
RFA1 eyes see bright lies best form say days
RFA2 beauty false old face black now truth seem
4.2 FCA (Fragmented Correspondence Analysis) *(2 topics out of 7)*
FCA1 beauty truth muse age youth praise old eyes glass long seen lies false time days
FCA2 night day bright see look sight
4.3 Logarithmic Analysis (Spectral mapping) *(2 topics out of 8)*
LOA1 summer away youth sweet state hand seen age rich beauty time hold nature death
LOA2 pen decay men live earth verse muse once life hours make give gentle death
4.4 Latent Semantic Analysis *(2 topics out of 8)*
LSA1 time heart beauty more one eyes eye now myself art still sweet world
LSA2 end grace leave words lie spirit change shame self could ever decay write
4.5 NMF topics *(2 topics out of 10)*
NMF0: love true new hate sweet dear say prove lest things best like ill let know fair soul
NMF1: beauty fair praise art eyes old days truth sweet false summer nature brow black live
4.6 Latent Dirichlet Allocation LDA *(2 topics out of 10)*
LDA0 summer worse praise nature making time like increase flower let copy

rich year die LDA1 sing sweets summer hear love music eyes bear single confounds prove shade eternal.

5. A synthesis of produced topics

How to compare the complete lists of topics, since neither the order of topics, nor the order of words within a topic are meaningful? We deal here with real 'bags of words' exemplified by the excerpts of lines in section 4. We will add the eight *a priori* themes defined in table 1. Each *a priori* theme corresponds to a subset of sonnets. That subset will be described by its characteristics words. We can then perform a clustering of these 57 topics/themes (49 + 8). The technique of additive trees (Sattath and Tversky, 1977; Huson and Bryant, 2006) seems to be the most powerful tool for synthesizing in compact form these 57 topics/themes (figure 2). Let us recall one important property of additive trees: the real distance between two points can be read directly on the tree as the shortest path between the two points.

Ideally, we expect to find a tree with as many branches as there are real topics in the corpus, each branch of the additive tree being characterized by seven labels: six labels corresponding to the six methods briefly described above, plus one label corresponding to one *a priori* theme. Such situation occurs when each method has uncovered the same real topics. The observed configuration is not that good, but we can distinguish between six and nine main branches, which is probably the order of magnitude of the number of different topics. We note also that several different methods often participate in the same branch, which suggest that that branch correspond to a real topic discovered by almost all the six methods. Let us mention that a similar additive tree performed on the 49 topics (not involving the eight *a priori* themes) produces approximately the same branches. Thus, the eight *a priori* themes can be considered here as illustrative elements, serving only as potential identifiers of the branches.

It is remarkable that the eight *a priori* themes (boxed labels) are well distributed over the whole of Figure 2. If we except the branch of the tree located in the upper right part of the display, on the right of the label "Young man", all the main branches have as a counterpart one of the *a priori* themes. As an example of interpretation of figure 2, the branch in the lower center part of figure 2: [NMF7, LOA4, RFA3, LDA7, LSA5] is clearly closely linked to the *a priori* topic named *Rivalry* (see section 2.2) (concurrence of five methods out of six). Most of the branches of the additive tree could be interpreted likewise. The upper right branch identified by none of the *a priori* themes may represent an unforeseen topic. More research and an expertise in Elizabethan poetry are required to confirm that we are dealing here with an undetected new theme. To conclude, we can only observe that each of the

involved method, be it ancient or modern, may contribute to detect topics… and that exploratory tools are essential to visualize the complexity of the process and assess the obtained results.

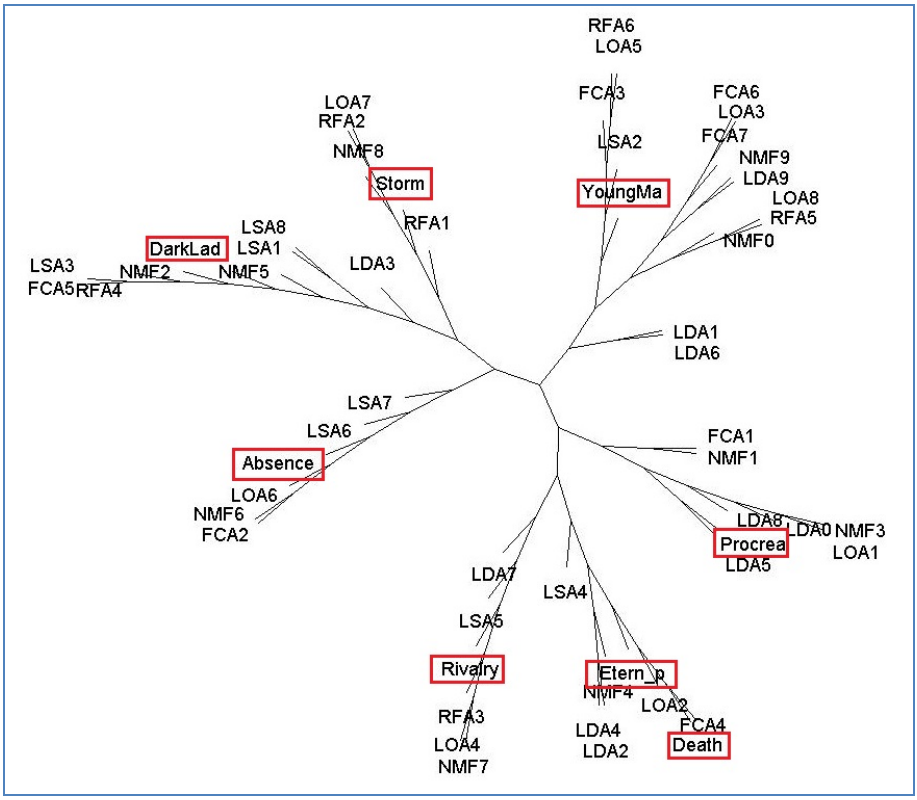


Figure 2. Additive Tree describing the links between the 49 topics provided by the 6 selected methods and the 8 a priori themes. The identifiers are those of section 4 for the 6 selected methods. The 3 first letters indicate the method, followed by the index of the produced topic. The distance between two topics is the chi-square distance between their lexical profiles. Threshold of frequencies for words: 2. The boxed identifiers of the a priori themes are those (possibly shortened) of table 1.

References

Alden, R. M. (1913). *Sonnets and a Lover's Complaint*. New York: Macmillan.

Berry M.W., Browne M., Langville Amy N., Pauca V.P., and Plemmons R.J. (2007). "Algorithms and applications for approximate nonnegative matrix factorization". In: *Computational Statistics & Data Analysis* 52.1: 155-173.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993—1022.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K.and Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6): 391-407.

Garnett J.-C. (1919). General ability, cleverness and purpose. *British J. of Psych,* 9, 345-366.

Griffiths T.,L., Steyvers M., and Tenenbaum J.,B. (2007). Topics in Semantic Representation. *Psychological Review*, 114, 2, 211-244.

Huson D. H., Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution,* 23 (2): 254 - 267. Software available from www.splitstree.org.

Kazmierczak J.-B. (1985). Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.*, 33, (1), 13-24.

Lee D.D. and Seung H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature,* 401: 788-791.

Lebart L. (2012). Articulation entre exploration et inférence. In : *JADT_2012*. Dister A., Longree D., Purnelle G., Editors. Presse Universitaire de Liège.

Lewi P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim. Forsch. in: *Drug Res.* 26, 1295-1300.

Paterson D. (2010). *Reading Shakespeare Sonnets.* Faber & Faber Ltd. London.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* , 12, 2825-2830.

Reinert, M. (1986). Un logiciel d'analyse lexicale: [ALCESTE]. *Cahiers de l'Analyse des Données*, 4, 471–484.

Sattath S. and Tversky A. (1977). Additive similarity trees. *Psychometrika,* vol. (42), 3: 319-345.

Shakespeare, W. (1901). *Poems and sonnets: Booklover's Edition*. Ed. The University Society and Israel Gollancz. New York: University Society Press. Shakespeare Online. Dec. 2017.

Spearman C. (1904). General intelligence, objectively determined and measured. *Amer. Journal of Psychology,* 15, 201-293.

Gaujoux R.*et al.* (2010). A flexible R package for nonnegative matrix factorization. In: BMC Bioinformatics 11.1 (2010): 367.

Thurstone L. L. (1947). *Multiple Factor Analysis.* The Univ. of Chicago Press, Chicago.

Analyse Diachronique de Corpus : le cas du poker

Gaël Lejeune[1], Lichao Zhu[2]
[1] STIH, Sorbonne Université – gael.lejeune@sorbonne-universite.fr
[2] LLSHS, Université Paris XIII – lichao.zhu@univ-paris13.fr

Abstract
In this paper we will investigate a diachronic corpus. We want to highlight how people's mentalities evolve regarding the gambling especially the poker game and how the evolution is correlated with the way that the game is considered in press articles. We study plain or metaphorical meanings of the terms in question by using clustering and statistical methods in order to detect changes of meanings in a relatively large period of time.

Résumé
Dans cet article nous nous intéressons à l'étude diachronique de corpus de presse dans le but d'illustrer des évolutions dans la vision de la société sur les jeux d'argent et de hasard ainsi que sur les joueurs. Nous utilisons des méthodes de statistique textuelles et de *clustering* pour détecter les grandes tendances visibles sur noter échelle de temps en nous focalisant sur le poker . Nous montrons que si le regain de popularité du jeu de poker se traduit par un traitement médiatique plus important, les métaphores exploitant la notion de poker restent très fréquentes.

Keywords: analyse diachronique, corpus, jeux d'argent et de hasard

1. Introduction

L'analyse diachronique de corpus opère sur un champ assez large. Nous pouvons en juger par exemple en observant les nombreux travaux sur l'évolution des langues, travaux qui passionnent aussi bien la communauté scientifique (Dediu & de Boer 2016) que les médias si l'on se fie par exemple à l'intérêt renouvelé porté par ceux-ci sur l'évolution des dictionnaires. Dans le champ purement scientifique, les intérêts dans le domaine embrassent tous les niveaux de l'analyse linguistique même si la morphologie (Macaulay 2017) et le lexique (la néologie par exemple chez Gérard et al. 2014). La sémantique est un autre aspect des études diachroniques notamment pour étudier les représentations mentales des locuteurs (Hamilton et al. 2016). Le travail présenté ici s'intéresse à une autre catégorie de représentations mentales qui est l'image que certaines activités ludiques peuvent prendre au cours du temps. Nous nous intéressons ici à un jeu d'argent et de hasard qui

a connu une sorte de nouvelle jeunesse ces dernières années : le jeu de poker. Dans ce travail, nous nous inspirons de l'analyse de l'usage du lexique dans (Hamilton et al. 2016), nous souhaitons examiner l'évolution de l'usage d'un mot, d'un terme particulier au cours du temps. Ce travail, même si notre ambition est moins large, peut se rattacher aux études sur la néologie sémantique (Sablayrolles 2002) ou néosémie (Rastier et Valette 2009). Pour illustrer l'intérêt que représente le poker en tant que phénomène de société, nous pouvons considérer le retentissement autour du *Moneymaker Effect*[1] ou encore cette citation du journal Le Monde daté du 22 janvier 2007 qui illustre le changement d'image de ce jeu: « *Considéré il y a encore peu de temps comme un jeu sulfureux se jouant dans les arrière-salles de bars louches ou dans des appartements huppés à l'abri des regards indiscrets, le poker fait une entrée en force à la télévision* ». En particulier, dans sa variante à la mode Texas Hold'Em, le poker est redevenu un jeu dont on parle et dont on parle plutôt positivement. Notre objectif est d'une part de mesurer à quel point ce regain d'attention a pu se traduire par une amélioration de l'image du jeu de poker en général. D'autre part, il s'agit de voir dans quelle mesure les usages métaphoriques du terme poker, plutôt connotés "négativement" (poker menteur, coup de poker[2]…) ont pu évoluer conjointement à cette plus grande popularité du jeu lui même. Dans la section 2 nous présenterons le corpus que nous avons constitué pour cette étude. Puis, nous proposerons dans les deux sections suivantes une analyse statistiques des prédicats puis une analyse sous forme de *clustering*. Enfin, nous présenterons nos conclusions et perspectives.

2. Présentation de notre corpus d'étude

De manière à pouvoir s'affranchir des variations de choix éditoriaux entre journaux, nous nous avons souhaité nous concentrer sur une seule publication. Nous avons choisi le Monde ce qui nous permettais d'exploiter des articles dont la publication s'étalent sur 30 ans : 1988-2017. Pour la partie 1988-2005 nous avons utilisé le corpus du monde distribué par ELRA[3], nous avons restreint aux textes contenant le terme poker. Pour les années 2006 à 2017 nous avons extrait d'Europresse[4] les articles qui comportait le terme poker. Dans les deux cas nous avons considéré toutes les variantes possibles dans la casse. Nous avons ainsi obtenu 3528 textes dont la répartition dans le

[1] Par exemple : http://www.slate.com/articles/news_and_politics/explainer/2011/06/the_moneymaker_effect.html

[2] Dans le sport par exemple, on remarque des contextes de « tentative désespérée », « dernière chance » ...

[3] http://catalog.elra.info/product_info.php?products_id=438&language=fr

[4] http://www.europresse.com/fr/

temps est présentée Figure 1. Nous pouvons observer que le nombre d'articles a connu une chute entre 2005 et 2006. Ceci semble être dû au fait que nous passions à ce moment précis d'une étude du corpus complet du monde tel qu'existant auprès d'ELRA à une étude fondée sur la base Europresse. De fait, sur nos critères de recherche, la base Europresse ne totalise que 47 articles pour 2003 (contre 129 dans le corpus ELRA), 62 pour 2004 (contre 117) et 67 articles pour 2005 (contre 117). Les contraintes respectives d'utilisation de ces deux sources de données nous ont interdit de pouvoir disposer d'un corpus dont la constitution soit constante. Nous nous sommes efforcés de s'affranchir de ce biais en adaptant notre méthodologie (notamment le *clustering*).

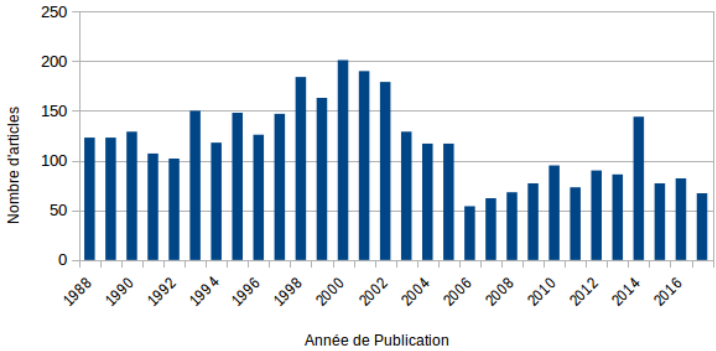


Figure 1 : Répartition du nombre d'articles par année

Nous avons 4353 occurrences du terme recherché, leur répartition est instructive (Figure 2) : la très grande majorité des articles (2834/3528 soit 80,33%) ne comporte qu'une seule occurrence. Nous pensons que ceci est le reflet de deux tendances. D'une part le sujet de l'article est rarement le poker pour lui même, il est question d'un personnage qui par ailleurs joue au poker par exemple. D'autre part, cette rareté de la répétition révèle un usage massivement métaphorique, en effet comme l'a montré (Lejeune 2013) une métaphore perd de sa force en étant répétée. Si un terme est répété, il est très probable qu'il soit employé dans son sens plein. Si cette observation était faite sur des noms de maladies infectieuses, il nous semble que ceci est avant tout lié au genre de texte et que cela s'applique également ici. Si nous allons un peu plus loin, nous pouvons faire l'hypothèse que la métaphore peut être filée, mais qu'elle est rare dans les articles expositifs. D'autre part, dans le cas peu probable d'une métaphore filée, les conventions stylistiques impliquent de changer le terme employé, le journaliste utilisera plutôt des termes du même champ lexical.

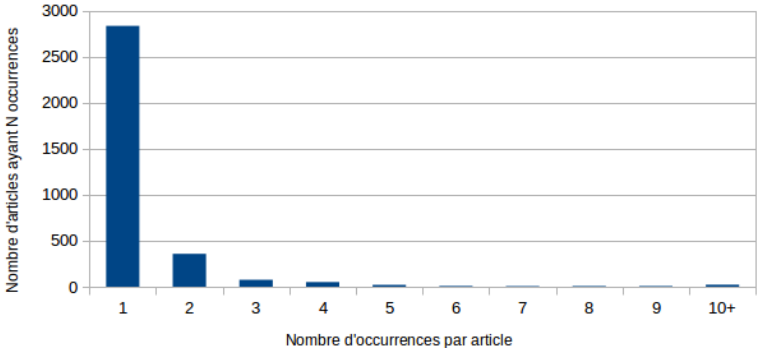


Figure 2 : Répartition des d'articles selon le nombre d'occurrences du terme « poker »

La répartition des articles entre ceux qui comportent une et une seule occurrence et ceux qui en comportent plusieurs montre des variations importantes dans le temps (Figure 3). Si l'on observe des périodes de 5 ans, on peut se rendre compte que le nombre d'articles comprenant plusieurs occurrences de "poker" représente 15% des articles sélectionnés sur la période 1988-1992, se pourcentage descend à 10% jusqu'en en 2003 puis remonte progressivement pour finalement rester au-dessus de 20% à partir de 2004-2008 avec une pointe à 30% pour les périodes 2007-2011 à 2009-2013.

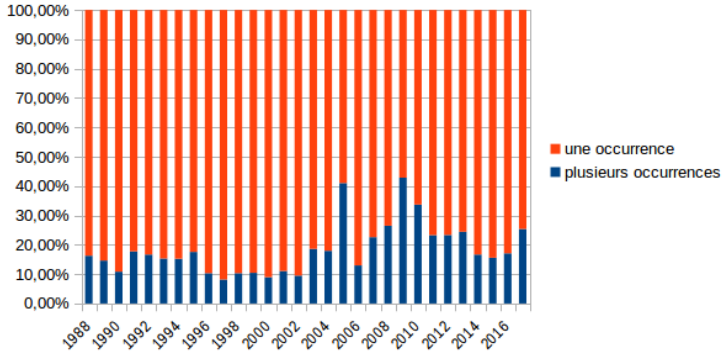


Figure 3 : Répartition par année des articles selon le nombre d'occurrences

3. Prédicats et séquences figées

Dans la théorie linguistique lexique-grammaire de M. Gross (1975) et de G. Gross (2012), les prédicats sont considérés comme les noyaux d'une phrase capables de disposer d'arguments, grâce à leurs propriétés transformationnelles et distributionnelles. Parmi les apports de cette théorie figurent le « schéma d'arguments » et les « prédicats appropriés ». Nous relevons dans notre corpus les contextes gauches et droits des séquences

figées « partie de poker » et « coup de poker » afin de distinguer leurs emplois métaphoriques et non métaphoriques. Ce travail est fait en étudiant le premier verbe précédant ou suivant l'expression (sans remonter au-delà d'une phrase). Nous montrons dans les tableaux 1 et 2 les 20 verbes les plus fréquents pour chaque contexte se trouvent le plus fréquemment dans ces contextes (20 dans les contextes gauches, 20 dans les contextes droits).

Tableau 1 : Effectif des verbes dans le contexte gauche de "[partie|coup] de poker"

être (76)	jouer (62)	faire (15)	tenter (14)	gagner (11)
avoir (11)	ressembler (10)	prendre (9)	tenir (8)	lancer (8)
perdre (7)	voir (6)	partir (6)	engager (6)	agir (5)
réussir (4)	livrer (4)	remporter (3)	organiser (3)	mener(2)

Tableau 2 : Effectif des verbes dans le contexte droit de "[partie|coup] de poker"

être (98)	avoir (75)	jouer (16)	pouvoir (13)	devoir (8)
gagner (7)	engager (7)	venir (6)	livrer (6)	faire (5)
vouloir (4)	voir (4)	tenter (4)	tenir (4)	réussir (4)
prendre (4)	monter (4)	bluffer (4)	aller (4)	retrouver (3)

Hormis les verbes « être » et « avoir » qui sont susceptibles d'être des verbes auxiliaires ou semi-auxiliaires, pour les autres verbes on peut se trouver dans trois cas de figure :

h) Verbe support
i) Prédicat approprié : le sens littéral de l'expression peut être activé
j) Prédicat non approprié : le sens métaphorique de l'expression est activé

Le cas des verbes support n'est pas pertinent pour notre étude. Pour le second cas, nous observons que le verbe jouer, prédicat approprié pour les deux séquences décrites, est très souvent lié à un usage métaphorique. Dans le troisième cas, de loin le plus fréquent. Les verbes « tenter », « s'engager », « réussir », « mener », « lancer » voire « remporter » ne sont pas tout à fait congruents avec le sens premier de la séquence, c'est-à-dire qu'ils ne sont pas des prédicats appropriés au sens propre du jeu de poker. Des occurrences de ces verbes dans le corpus confirment cette intuition :

*Il leur fallait **lancer** la partie de poker que Bonn et Paris s'apprêtent à jouer sur le GATT (1993)*

*les enjeux de la partie de poker qui **s'engagera** mercredi à la mi-journée lorsque l'ambassadeur[...] (2017)*

*[ils] avaient pu croire un moment que leur coup de poker allait **réussir**. (1989)*

*[Celui qui] est davantage connu pour ses coups de poker financiers continue à **mener** sa stratégie (2015)*
*Elle venait de **remporter** la partie de poker menteur qui constitue l'essentiel des premiers hectomètres. (1995)*

4. Étude des champs lexicaux par *clustering*

Si les séquences « partie de poker » et « coup de poker » sont ambiguës dans le sens où elles figurent dans des champs lexicaux différents, on peut se demander ce qu'il en est des champs lexicaux du terme « poker » en général. Pour étudier cette question, nous avons réalisé un *clustering* de notre corpus. Nous avons utilisé l'implantation des k-moyennes (*K*-means) de la bibliothèque Python scikit-learn. Nous avons fixé le nombre de *clusters* K à 10[5] et le nombre maximal d'itérations à 400, la mesure des poids est le tf-idf. Nous avons extrait tous les n-grammes de mots avec n allant de 1 à 3 puis seulement nous avons utilisé une stop-list. De sorte que, par exemple, « de » n'était pas gardé en tant que tel mais que nous le retrouvions dans « coup de poker » ou « loi de Robien ». Nous avons tout d'abord travaillé sur le corpus lemmatisé puis nous avons observé que les résultats étaient semblables sans lemmatisation, nous avons donc supprimé ce pré-traitement. Nous allons maintenant décrire chaque *cluster* en donnant la proportion du corpus qu'il couvre ainsi que les 10 termes les plus significatifs.

Cluster 0, « sport et poker 1 » : 3,1 % (club, football, équipe, Ligue, France, championnat, saison, joueurs, OM, Marseille) Ce *cluster* comporte deux volets : l'un sur les « coups de poker » dans les championnats de football et l'autre où il est question des championnats de poker eux mêmes.
Cluster 1, « politique » : 18,79 % (ministre, président, politique, gouvernement, pays, ,État, premier ministre, premier, États, faire). Un *cluster* autour de l'action politique, notamment au niveau européen. Un exemple intéressant de métaphore (filée) ici : « M. Erdogan remet tout en jeu, comme un joueur de **poker** fait tapis »
Cluster 2, « fourre-tout » : 38,01 % (être, bien, film, vie, entre, Jean, monde, France, temps, homme) Le seul de nos clusters qui n'ait pas d'unité ni de tendance thématique, ici les expressions contenant poker sont pour moitié métaphoriques.
Cluster 3, « culture_1 » : 5,13 % (film, Booker Prize, roman, prix, livres, livre, littéraire, base, prix littéraire, attribué). Ce *cluster* rassemble les livres ayant trait au poker, les expressions liées sont prises dans leur sens littéral

[5] Selon la méthode du coude (*elbow method*), la valeur optimale se situait entre 9 et 12.

(l'expression « coup de poker » y est quasi absente).
Cluster 4 « finance »: 4,2 % (Vivendi, marché, groupe, Bourse, marches, actionnaires, titres, taux, millions, fonds, terme, milliards, prix) Il se caractérise uniquement par des thématiques associées au domaine de finance et notamment aux coups de poker boursiers.
Cluster 5 « sport et poker 2»: 5,04 % (Coupe, match, équipe, joueurs, France, club, football, finale, francs, PSG). Nous avons ici un *cluster* sur le sport où environ la moitié des articles concernent toutefois le poker lui même.
Cluster 6 « industrie du poker »: 12,96 % (jeux, paris, ligne, marché, milliards, euros, millions, Internet, dollars, Bourse) Ici nous avons tout ce qui est lié à l'industrie du poker et notamment à l'essor des jeux d'argent sur Internet (dont le poker a été un fer de lance).
Cluster 7 « sport »: 3,26 % (Tour, numéros, France, coureur, étape, peloton, course, équipe, Tour de France, maillot) Nous avons ici des usages, massivement métaphoriques, dans le domaine du sport (principalement le cyclisme). Un exemple avec le terme spécialisé *flop* : « [P.A.Bosse] avait trouvé cette image [...] : Si on compare le 1500m au poker, il a un *flop* d'avance. »
Cluster 8 « culture_2 » : 7,14 % (blues, musique, CD, rock, John Lee Hooker, jazz, album, guitare, musiciens, scène) Un usage métaphorique dans le domaine de la musique avec des expressions telles que « poker face », « poker perdant »...
Cluster 9 « culture_3 » : 2,38 % (Dracula, Bram Stoker, vampire, roman, film, fantastique, Christie, Coppola, comte, Frankenstein) Le *cluster* 3 était centré sur le domaine littéraire, ici il est question de cinéma et particulièrement des personnalités liées au poker. L'usage y est surtout littéral.

Pour ce qui est de la répartition temporelle, il est très intéressant de noter que le cluster 6 (l'industrie du poker) devient le second plus important derrière le cluster 2 (à partir de 2005 (popularisation des jeux d'argent sur Internet) et plus encore à partir de 2010 (légalisation des paris en ligne). Le *cluster* 0 (sport et poker) devient plus important à partir de 2004 d'autant qu'en son sein la thématique poker y est alors largement majoritaire.

5. Conclusion
Nous avons proposé dans cet article une étude diachronique d'articles de presse contenant le mot « poker ». Notre hypothèse initiale était que ce terme était souvent employé dans des expressions métaphoriques et que le regain de popularité de ce jeu depuis quelques années avait du amener une plus grande proportion d'usage littéral. Nous avons observé que dans plus de 80 % des cas, le terme poker n'apparaissait qu'une fois dans les textes. Nous

avons montré que ceci était dû à un usage principalement métaphorique, on ne répète pas une métaphore, mais aussi au fait que le poker est rarement le sujet central de l'article. Cette tendance change quelque peu à partir de 2005, le poker devenant lié à des championnats et des retransmissions télévisuelles plutôt qu'à des tripots et des casinos. Enfin, nous avons montré que les usages métaphoriques relevaient très majoritairement de 3 domaines : la finance, la politique et le sport.

References

Dediu D. and de Boer B. (2016)., Language evolution needs its own journal , Journal of Language Evolution, Volume 1, Issue 1, 1 January 2016, Pages 1–6

Gérard C., Falk I., and Bernhard D. (2014). Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? Actes du 4e Congrès mondial de linguistique française (CMLF 2014), Berlin, Pages 2627-2646

Gross, M. (1975). Méthodes en syntaxe: régime des constructions complétives, Hermann.

Gross, G. (2012). Manuel d'analyse linguistique: Approche sémantico-syntaxique du lexique, Presses Universitaires du Septentrion.

Hamilton W.L., Leskovec J., and Jurafsky D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In Proc. Of the Association for Computational Linguistics Conference (ACL) 2016

Lejeune G. (2013) Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel, Thèse de doctorat en Informatique de l'Université de Caen

Macaulay, M. & Salmons. (2017). Synchrony and diachrony in Menominee derivational morphology, J. Morphology 27: 179

Rastier, F., Valette, M. (2009) « De la polysémie à la néosémie », Le français moderne, S. Mejri, éd.,

La problématique du mot, 77, 97-116.

Sablayrolles, F. (2002). « Fondements théoriques des difficultés pratiques du traitement des

néologismes », Revue française de linguistique appliquée, VII-1, pp. 97-111.

Approche textométrique des variations du sens

Julien Longhi[1], André Salem[2]
[1]Université de Cergy-Pontoise, France – julien.longhi@u-cergy.fr
[2]Université de la Sorbonne nouvelle, France – salem@msh-paris.fr

Abstract
The use of textometric methods relies on the hypotheses, firtly, that stable units exist (forms, lemmas or their graphical approximations) and, secondly, that occurrences of these forms can be retrieved from different parts of a corpus. Once automatic counting performed, more sophisticated textometric methods can be employed to focus on textual variations (repeated segments, collocations, etc.) that occur around the same unit but in different contexts found within the corpus. This approach leads to the identification of semantic variations with relation to the context of each occurrence as highlighted through automatic segmentation. We will illustrate this by using examples of repeated segments within the corpus that contain the N-gram /enemy / taken from a widely-studied chronological text series.

Résumé
Pour pouvoir mettre en œuvre les méthodes de la textométrie, il est indispensable de postuler, dans un premier temps, l'existence d'unités stables (formes, lemmes ou leurs approximations graphiques), dont on recensera ensuite les occurrences dans les différentes parties du corpus étudié. Une fois les dépouillements automatiques réalisés, il est cependant possible d'utiliser des méthodes textométriques plus élaborées pour accéder aux variations textuelles (segments, répétés, cooccurrences, etc.) qui peuvent se réaliser autour d'une même forme dans chacun des contextes particuliers du corpus. Cette démarche permet d'accéder au repérage de variations sémantiques qui se rapportent à chacune des occurrences des formes produites par la segmentation automatique. Nous illustrons notre démarche à l'aide d'exemples prélevés dans les parties d'une série textuelle chronologique largement étudiée, des segments répétés du corpus qui contiennent le N-gram /ennemi/.

Keywords:.unité textométrique, sémantique, variation du sens

1. Introduction

Notre étude s'inscrit dans une perspective de prise en compte des dynamiques du sens à l'œuvre dans les discours, qui tiendrait compte de la

variation, de l'hétérogénéité, ou encore de l'articulation entre topologie textuelle et discursive, sens et profilage. Le sens se construit dans différents champs où il est susceptible de paraître, et s'analyse « par le contexte, sous forme d'indices de position liés aux modalités de sa mise en place dans le champ » (Cadiot et Visetti, 2011), la caractérisation sémantique se faisant alors sur la base de la composition et décomposition des profils disponibles. L'automatisation du dépouillement de vastes corpus de textes, à des fins textométriques, nécessite au contraire que le repérage des unités de décompte puisse être confié à des machines. Pour pouvoir mettre en œuvre les méthodes de la textométrie, il est indispensable de postuler, dans un premier temps, l'existence d'unités stables (lexèmes, lemmes ou leurs approximations graphiques), dont on recensera ensuite les occurrences dans différentes parties du texte. Cette manière de faire permet d'étudier la répartition de chacune des unités dans un corpus ou encore de rapprocher les différents contextes qui contiennent chaque unité textométrique. Ces simplifications, incontournables dans le premier temps de l'analyse, nous éloignent de l'étude du sens de chacune des occurrences que l'on peut élaborer dans chaque contexte particulier. Cependant, une fois les premiers dépouillements automatiques réalisés, il est possible d'utiliser des méthodes textométriques plus élaborées pour accéder aux variations textuelles qui peuvent se réaliser autour d'une même forme dans le corpus (segments répétés, cooccurrences, etc.). C'est ce croisement de perspectives et ce va-et-vient entre approche empirique et théorisation sémantique, que nous souhaitons mettre à l'épreuve dans la présente étude.

2. Application au corpus *Duchesne*

Pour illustrer notre démarche, nous appliquons ces méthodes à l'étude de la ventilation, dans les différentes parties d'une série textuelle chronologique largement étudiée, des segments répétés du corpus qui contiennent le N-gram */ennemi/*.

2.1. Rappels sur l'analyse de la série chronologique Duchesne

La série chronologique *Père Duchesne* a déjà fait l'objet de nombreuses analyses textométriques[1]. Nous avons montré, en particulier, que les

[1] Le corpus Père Duchesne est constitué par la réunion d'un ensemble de livraisons du journal *Le Père Duchesne* de Jacques-René Hébert, parues entre 1793 et 1794. Pour une description plus avancée de ce corpus, on consultera, par exemple (Salem, 1988).

Les analyses dont nous rendons compte ci-dessous, ont été effectuées à l'aide du logiciel Lexico5. Cedric Lamalle, William Martinez, Serge Fleury ont largement

typologies réalisées à partir d'une partition de ce corpus en huit périodes, correspondant chacune à un mois de parution, mettaient en évidence un renouvellement lexical fortement lié à l'évolution dans le temps. On peut vérifier, sur la figure 1, que les parties correspondant aux périodes successives de parution sont proches sur les facteurs issus de l'analyse du tableau (8 parties x 1420 formes dont la fréquence dépasse dix occurrences)[2].
La méthode des ***segments répétés*** permet de repérer toutes les occurrences de suite de formes graphiques qui apparaissent plusieurs fois dans un corpus de textes (Lafon et Salem, 1983 ; Salem, 1986). Pour la présente étude, nous avons constitué un ensemble d'unités textuelles qui contient outre les formes graphiques *ennemi* et *ennemis*, tous les segments répétés qui contiennent l'une ou l'autre de ces formes. On a projeté sur la figure 1, en qualité d'éléments supplémentaires, cet ensemble de segments. La position sur ce graphique des différents segments montre que ces unités ne sont pas employées de manière uniforme tout au long des périodes.

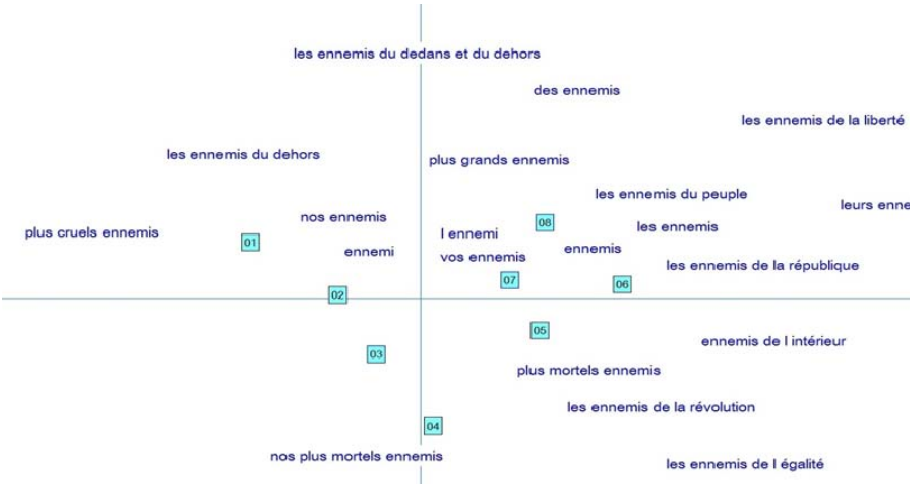


Figure 1 : Duchesne. Les segments contenant la séquence ennemi sur le plan des deux premiers facteurs issus de l'analyse de tableau 8 parties x 1420 formes (F>=10)

Guide de lecture pour la figure 1 : La figure fournit la représentation des huit parties du corpus ***Duchesne***, sur les deux premiers axes issus d'une Analyse

contribué au développement des fonctionnalités de ce logiciel. Les auteurs tiennent à les en remercier.

[2] Ce phénomène connu sous le nom *d'effet Guttman*, a été largement décrit par Guttman (1941, 1946, 1950), Benzecri (1973) et Van Rijckevorsel (1987).

des correspondances, réalisée sur l'ensemble des formes dont la fréquence dépasse 10 occurrences. Les segments répétés du corpus contenant la séquence de caractères /*ennemi* / ont été projetés sur ce même plan, en tant qu'éléments supplémentaires. La figure a été allégée des segments redondants (ex : segments contenus dans des segments plus longs). Certains des éléments superposés par l'analyse ont été très légèrement déplacés à fin de rendre la figure plus lisible.

Ainsi par exemple, le segment *plus cruels ennemis* trouve toutes ses occurrences au début du corpus alors que celles du segment *ennemis de la liberté* sont plutôt concentrées vers la fin.
L'analyse des projections des différents segments qui contiennent le n-gram /ennemi/ va nous permettre de dégager des contextes dont la distribution diffère fortement entre le début et la fin de la période temporelle couverte par le corpus.

2.2. L'évolution du contexte de la forme ennemi(s)

On peut estimer que le contenu sémantique de la forme *ennemi(s)* conserve une valeur relativement stable tout au long des périodes couvertes par le corpus que nous étudions. Le chercheur confronté à l'analyse de ces textes retrouvera sans peine, lors de l'examen de chacune des occurrences du terme, les principaux traits sémantiques décrits dans un dictionnaire de langue à propos de ce lexème (opposé, hostile, etc.). Cependant, l'analyse de ces mêmes contextes montre qu'il en va tout autrement pour ce qui concerne les référents auxquels la forme renvoie, dans chaque période particulière. Aux *plus cruels ennemis*, *plus mortels ennemis*, *ennemis du dehors* (les puissances étrangères, les expatriés), des périodes du début, succèdent bientôt *les ennemis du dedans et du dehors,* expressions qui peuvent s'analyser comme une dénonciation du fait que les *ennemis du dehors* ne constituent pas le seul danger et qui opère donc une modification manifeste du référent de départ. Par la suite la mention des *ennemis de l'intérieur* complètera la notion d'*ennemis du dedans*. Il faut noter que les *ennemis de l'intérieur* sont de plus en plus souvent précédés de l'article défini *les* qui les désigne comme une réalité dont l'existence est présupposée (elle n'est plus à démontrer).
Progressivement, *nos ennemis*, deviennent *vos ennemis*, puis *les ennemis*. Dans la dernière période les ennemis, désormais désignés, de manière préférentielle, au pluriel, ne sont plus qualifiés par leur localisation ou par leur rapport aux destinataires du message (*nos/vos ennemis*) mais par des valeurs supposée communes auxquelles ils sont censés s'opposer : *ennemis du peuple, ennemis de la république, ennemis de la révolution, ennemis de la liberté, ennemis de l'égalité.*

3. La sémantique de *ennemi(s)*

Les variations constatées montrent que la forme *ennemi(s)* prend différents sens selon les contextes dans lesquels elle s'inscrit, en ce qu'ils sont associés à des référents distincts. Plutôt que de représenter le sens comme la somme des cooccurrences constatées, nous souhaitons analyser ces valeurs comme un sous-ensemble prélevé sur un ensemble de valeurs acquises. Les espaces sémantiques déterminés et caractérisés par l'analyse statistique jouent un rôle fondamental qui, au-delà des synonymies, ou des polysémies, se renouvellent « en étant confronté aux textes – ce qui impliquerait de prêter attention à d'autres corrélations » (Visetti 2004 : 11). La description sémantique que nous proposons s'inscrit dans le champ de la sémantique lexicale[3], du côté des approches qui envisagent la construction des référents comme extrinsèque. Cependant, alors que ces approches mobilisent en général des analyses phrastiques, et travaillent sur des exemples forgés, nous introduisons une perspective statistique qui précède la représentation du sens. La description de l'objet *ennemi(s)* n'est pas séparée des rapports que l'on entretient avec lui, et sa description suppose une prise en compte différenciée de ses propriétés extrinsèques (relatives à ces rapports), et de ses propriétés intrinsèques (supposées stables et indépendantes).

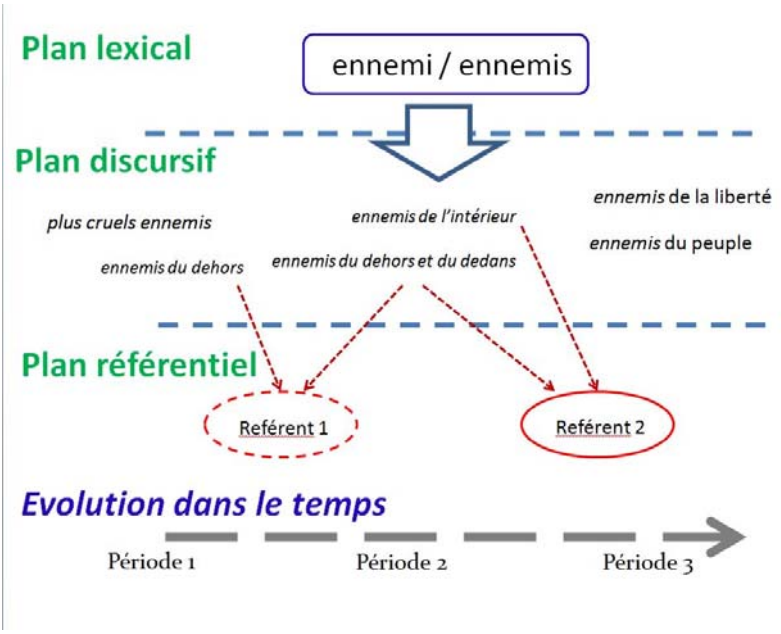


Figure 2 : Niveaux et unités d'analyse

[3] Cadiot et Némo (1997 : 127-128)

L'intérêt de cette démonstration textométrique est pour nous de fournir des résultats concrets et matériels pour l'analyse des sens d'une unité lexicale. Ceci a plusieurs conséquences pour la mise en œuvre d'une sémantique soucieuse de l'exploitation des constats empiriques :

1) la représentation des variations du sens en contexte nous a permis d'identifier la manière donc les propriétés sont introduites et attribuées dans le corpus. Le référent change au fil du temps, puisque les *ennemis*, initialement définis comme *du dehors*, et introduits par *nos*, deviennent *vos ennemis*, et se présentent finalement sous la forme *ennemi(s) de* + N. Le besoin d'être déterminé par un complément du nom, ou son équivalent, qui indique avec quoi le terme « relatif » se trouve mis en relation », cette complémentation explicitant « ainsi la référence identitaire » (Steuckardt, 2008).
2) L'évolution dans le corpus au fil du temps permet de rendre compte de la dynamique sémantique à l'œuvre, laquelle rend compte diachroniquement des évolutions de sens. La textométrie permet ainsi de saisir les processus, et donc de donner du sens à la dimension potentiellement « hétéroclite » des propriétés des référents.

Ainsi, au plan linguistique, le passage du référent 1 ou référent 2 se fait par l'intermédiaire d'une transformation des propriétés de *ennemi(s)* : défini de manière situationnelle (*du dehors*) et relative (*nos*, *nos plus cruels*), il acquiert des propriétés plus polémiques (*vos*, *du dedans et du dehors*), pour s'intégrer ensuite dans un processus discursif qui construit le référent (ennemi de + N : *ennemi de la liberté* ; *ennemi du peuple*), par l'introduction de termes à fort charge axiologique. Le référent introduit alors un point de vue, qui n'est pas strictement géographique ou institutionnel, mais aussi politique et idéologique. L'approche statistique dévoile, en outre, que c'est le pluriel qui est prioritairement mobilisé.

3. Conclusion

De manière désormais classique, les méthodes de la textométrie permettent de mettre en évidence les variations du vocabulaire qui surviennent au cours des périodes successives d'une même série textuelle chronologique. Dans la présente étude, nous avons appliqué les méthodes d'analyse statistique multidimensionnelle (AFC) à l'étude d'un ensemble particulier, celui des segments répétés réunis sur la base du fait qu'ils contenaient tous une même unité graphique (en l'occurrence, le n-gram */ennemi/*).

La confrontation des segments ainsi sélectionnés nous permet d'observer des variations autour des formes graphiques *ennemi* et *ennemis*. L'analyse de ces

variations dans le temps nous conduit à distinguer des référents qui varient en fonction des périodes réunies dans le corpus.
Au-delà des séries textuelles chronologiques, la méthode que nous avons présentée est susceptible de recevoir des applications dans l'étude de nombreux types de corpus. L'extraction semi-automatique des unités dont les contextes varient fortement en fonction des parties d'un corpus textuelle peut également être envisagée.

References

Benzécri J-P. and coll. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie*. Dunod.

Cadiot P. and Nemo F. (1997). Propriétés extrinsèques en sémantique lexicale. *Journal of French Language Studies*, 7(2): 127-146.

Cadiot P. and Visetti Y.-M. (2001). *Pour une théorie des formes sémantiques*.PUF.

Guttman L. (1941). The quantification of a class of attributes: a theory and method of a scale construction. In P. Horst, *The prediction of personal adjustment*, SSCR New York.

Lafon P. and Salem A. (1983). L'Inventaire des segments répétés d'un texte. *Mots. Les langages du politique,* 6 : 161-177.

Lamalle C, Martinez W, Fleury S, and Salem A. (2002). *Les dix premiers pas avec Lexico3. Outils lexicométrique*s. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Longhi J. (2008). *Objets discursifs et doxa. Essai de sémantique discursive*. L'Harmattan, coll. « Sémantiques ».

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Honoré Champion, coll. « Lettres numériques ».

Salem A. (1987). *Pratique des segments répétés*. Klincksieck.

Salem A. (1988). Approches du temps lexical. *Mots. Les langages du politique,* 17 : 105-143.

Steuckardt A. (2008). Les ennemis selon *L'Ami du peuple*, ou la catégorisation identitaire par contraste. *Mots. Les langages du politique* [En ligne], 69 | 2002. http://journals.openedition.org/mots/10023

Van Rijckevorsel J. (1987). *The application of fuzzy coding and horseshoes in multiple correspondances analysis.* DSWO Press.

Visetti Y.-M. (2004). Le Continu en sémantique : une question de formes. *Texto ! juin 2004.* http://www.revue-texto.net/Inedits/Visetti/Visetti_Continu.html

ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables

Laurent Vanni[1], Damon Mayaffre, Dominique Longree[2]
[1]UMR 7320 : Bases, Corpus, Langage - prenom.nom@unice.fr
[2]L.A.S.L.A. - prenom.nom@uliege.be

Abstract 1
This contribution confronts ADT and Machine learning. The extraction of statistical key-passages is undertaken following several calculations implemented using the Hyperbase software. An evaluation of these calculations according to the filters applied (taking into account only positive specificities, only substantives, etc.) is given.
The extraction of key passages obtained by deep learning - passages that have the best recognition rate at the time of a prediction - is then proposed. The hypothesis is that deep learning is of course sensitive to the linguistic units on which the computation of the key statistical sentences are based, but also sensitive to phenomena other than frequency and other complex linguistic observables that the ADT has more difficulty taking into account - as would be the case with underlying patterns (Mellet et Longrée, 2009). If this hypothesis is confirmed, it would on the one hand permit better understanding of the *black box* of deep learning algorithms and on the other hand to offer the ADT community a new point of view.

Abstract 2
Cette contribution confronte ADT et Deep learning. L'extraction de passages-clefs statistiques est d'abord proposée selon plusieurs calculs implémentés dans le logiciel Hyperbase. Une évaluation de ces calculs en fonction des filtres appliqués (prise en compte des spécificités positives seulement, prise en compte de substantifs seulement, etc) est donnée. L'extraction de passages-clefs obtenus par deep learning - c'est-à-dire des passages qui ont le meilleur taux de reconnaissance au moment d'une prédiction - est ensuite proposée. L'hypothèse est que le deep learning est bien sûr sensible aux unités linguistes sur lesquelles le calcul des phrases-clefs statistiques se fondent, mais sensible également à d'autres phénomènes que fréquentiels et d'autres observables linguistiques complexes que l'ADT a plus de mal à prendre en compte - comme le seraient des motifs sous-jacents (Mellet et Longrée, 2009). Si cette hypothèse se confirmait, elle permettrait d'une part de mieux appréhender la *boîte noire* des algorithmes de deep learning et d'autre part d'offrir à la communauté ADT de nouveaux points de vue.

Keywords: ADT, deep learning, phrase-clef, motif, spécificités, nouveaux observables

1. Introduction

Pour des raisons techniques avant tout, l'ADT s'est constituée à partir des années 1960 autour du token, c'est-à-dire du mot graphico-informatique. Depuis lors, la discipline n'a cessé de varier et d'élargir ses observables, convaincue que le token seul rendait difficilement compte du texte dans sa complexité linguistique. Ainsi la tokenisation en particules graphiques élémentaires reste l'acte informatique premier des traitements textométriques, et le calcul des *spécificités* lexicales reste l'entrée statistique privilégiée de nos parcours interprétatifs. Cependant, la recherche d'unités phraséologiques élargies et complexes, caractérisantes et structurantes des textes, est devenue le programme d'une discipline désormais adulte. Historiquement, dès 1987, le calcul des *segments répétés* (Salem, 1987) ou les n-grams a représenté une avancée puisque les segments significatifs du texte, de taille indéterminée, étaient automatiquement repérés ; et aujourd'hui la détection automatique, non supervisée, de *motifs* (Mellet et Longrée, 2009; Quiniou et al., 2012; Mellet et Longrée, 2012; Longrée et Mellet, 2013) - objets linguistiques complexes à empans variables et discontinus - apparait un enjeu décisif. C'est dans cette perspective que cette contribution travaille et met à l'épreuve l'idée de *passages-clefs* du texte, tels qu'ils sont implémentés dans les deux versions d'Hyperbase (locale développée par Etienne Brunet et web développée par Laurent Vanni) que l'UMR Bases, Corpus, Langage produit en collaboration avec le LASLA. La démonstration se fait en deux temps. D'abord, nous proposons une extraction statistique de \textit{passages-clefs}, avec évaluation de leur pertinence interprétative sur un corpus français et un corpus latin. Ensuite une confrontation méthodologique avec le deep learning est mise en œuvre puisque le traitement deep learning attribue, après apprentissage, les passages de texte à leur auteur avec un taux de réussite éprouvé : par déconvolution nous repérons alors au sein de ces passages les *zones d'activation*, en soupçonnant qu'il s'agit, d'un point de vue linguistique, de motifs remarquables.

2. Les passages-clefs en ADT

2.1. Terminologie

Si nous préférons le terme de *passage-clef* à celui de *phrase-clef* c'est que les traitements ici présentés n'ont pas de modèle syntaxique, et que la ponctuation forte qui délimite habituellement la phrase est un jalon utile mais non-nécessaire à nos traitements. La notion de passage a été fortement

théorisée par (Rastier, 2007) dans un article éponyme et désigne une « grandeur » du texte dont la valeur textuelle c'est-à-dire interprétative est patente. Un passage est donc un morceau de texte jugé suffisamment parlant, notamment par sa taille qui gagne à dépasser le mot, le segment voire la phrase, pour prétendre rendre compte d'un texte. Le passage-clef, quant à lui, s'appuie sur la définition rastirienne mais est une unité de surcroit textométrique ; c'est-à-dire une unité dont la pertinence est calculable et l'extraction automatique.

2.2. Implémentations

Les logociels ADT comme Hyperbase, Dtm-Vic, Iramuteq implémentent des calculs et l'extraction de passages-clefs. Dans tous les cas, les calculs proposés reposent sur l'examen des mots spécifiques (Lafon, 1984) : grosso modo, plus un passage concentre de spécificités, plus ce passage est jugé remarquable. Nous présentons ici deux types d'approche sur des passages arbitrairement constitués de 50 mots : un calcul naïf et sans filtre dans lequel tous les mots du passage sont considérés et un calcul filtré par nos connaissances linguistiques (sélection a priori des mots à considérer). Une évaluation de ces deux types d'approche est ensuite donnée.

2.3. Calcul sans filtre

Dans le cadre des études contrastives habituelles en ADT, l'indice de spécificité de chaque mot (Lafon, 1984) est sommé, qu'il soit positif ou négatif en postulant que si les mots positifs (les mots sur-utilisés par un auteur par exemple) doivent promouvoir le passage, il est légitime que les mots négatifs (les mots sous-utilisés par un auteur) doivent l'handicaper. Chaque passage du corpus se trouve ainsi doté d'un super-indice de spécificité et Hyperbase fait remonter en bon ordre les passages les plus caractéristiques des textes comparés.

Ainsi pour le français, sur le corpus de la présidentielle française 2017, le passage-clef le plus fortement d'E. Macron (versus les autres candidats) est le suivant :

[...] nous croyons dans l'innovation, dans la transformation écologique et environnementale, parce que nous voulons réconcilier cette perspective et l'ambition de nos agriculteurs, parce que nous croyons dans la transformation digitale, parce que nous sommes pour une société de l'innovation, parce que nous voulons […]

Quoique naïf, le calcul apparait performant puisque l'interprétabilité socio-linguistique de ce passage est évidente : de fait Macron s'est fait élire sur un discours dynamique (*voulons* , *innovation* (deux fois), *transformation* (deux fois), *digitale*) et un discours rassembleur susceptible de transcender le clivage gauche/droite (*nous* (5 fois), *réconcilier*).

2.4. Calcul filtré

Par connaissances linguistiques et statistiques, le calcul peut être raffiné. Par exemple, seules les spécificités positives – et parmi elles, les spécificités les plus fortes – peuvent être considérées au motif qu'un objet s'identifie mieux par ses qualités que par ses défauts. Ensuite, les mots outils (conjonctions, déterminants) peuvent être écartés : ils présentent le double inconvénient d'avoir de très hautes fréquences (potentiellement déterminante pour le calcul des spécificités) et d'être peu parlants d'un point de vue sémantico-thématique. Et encore, la catégorie grammaticale peut être choisie : par exemple seuls les noms propres et communs, parfois plus chargés de sens, sont pris en compte. Ainsi pour le latin un passage-clef de Jules César, contrasté à de nombreux auteurs contenus dans la base du LASLA, est le suivant :

[...] partes Galliae uenire audere quas Caesar possideret neque exercitum sine magno commeatu atque molimento in unum locum contrahereposse sibi autem mirum uideri quid in sua Gallia quam bello uicisset aut Caesari aut omnino populo Romano negotii esset his responsis ad Caesarem relatis iterum ad eum Caesar […]

De fait, ce passage de la Guerre des Gaules peut être effectivement considéré comme très représentatif de l'œuvre de César. On relève des noms propres connus (*Galliae*, *Caesar*, *Gallia*) ou des noms communs correspondant à la réalité militaire du moment (*bello*, *commeatu*). Toutefois la méthode ne permet pas de repérer des structures caractéristiques de la langue et du style de César, comme par exemple une proposition participiale marquant la transition entre épisodes dans une négociation : *His responsis ad Caesarem relatis*, « Ces réponses ayant été rapportées à César ».

2.5. Evaluation

Calcul naïf ou calcul élaboré : nous récapitulons quelques performances. Dans un corpus contrastif, nous calculons le score de super-spécificité de chaque passage en fonction des différents auteurs comparés (Tableau 1). Par exemple pour le français, sans aucun filtre 58% des passages du corpus de la présidentielle sont attribués justement à leur auteur ; et en ne considérant que les spécificités positives, le score descend à 52%. A l'opposé, en imposant le double filtre de la catégorique grammaticale (seulement les substantifs) et de l'indice de spécificité (seulement les spécificités positives) nous élevons le taux de bonne attribution à 89% pour le français et 82% pour le corpus latin du LASLA.

Tableau 1. Taux d'attribution ADT et taux de prédiction deep learning

	Passages-clefs - ADT				Deeplearning
	sans filtre	$z > 0$	substantifs	substantifs et $z > 0$	
français	58%	52%	88%	89%	90%
latin	69%	62%	84%	82%	85%

3. Deep learning : à la recherche de nouveaux marqueurs linguistiques

3.1. Convolution et déconvolution, les principes

Le découpage du texte en segments de taille fixe est une méthode qui peut aussi être utilisée pour entraîner un réseau de neurones. Chaque segment devient alors une image d'un texte que le réseau va utiliser pour apprendre (Ducoffe et al., 2016) et faire ensuite de la prédiction. Sur nos deux corpus de référence (français et latin), les taux de précision convergent rapidement et atteignent le même niveau que ceux obtenus avec l'ADT (Figure 1). Si nous connaissons les paramètres à faire varier pour optimiser la détection des passages-clefs ADT, ceux issus du deep learning sont complètement non supervisés et découverts automatiquement par le réseau. L'idée des réseaux à convolution est de proposer un modèle capable de faire automatiquement une abstraction performante des données.[1] La convolution utilise pour cela un mécanisme de filtres qui va lire le texte avec une fenêtre coulissante pour extraire à chaque fois une partie de la matière linguistique présente dans la fenêtre (Figure 2). Avec des centaines de filtres de tailles différentes, le texte est lu en utilisant tous les empans linguistiques possibles et le mécanisme de back-propagation[2] finit par accorder un certain poids à certains éléments du texte qui le pousse à prendre la bonne décision. Le deep learning est souvent considéré comme une boîte noire faute de pouvoir mettre en évidence précisément ces éléments. Nous avons donc ici concentré nos efforts sur la déconvolution. Ce mécanisme utilisé notamment en analyse d'images permet de démêler le réseau et de lui redonner une forme interprétable par l'humain. Notre modèle est composé d'une couche de pré-apprentissage (Mikolov et al., 2013) pour la représentation des mots en vecteurs, d'une couche de convolution (Kim, 2014), un maxpooling pour compresser l'information et enfin un réseau classique de perceptron à une couche cachée pour la classification (Figure 2). La déconvoltution est en fait une simple copie partielle de ce réseau (jusqu'à la convolution) à laquelle on ajoute à la fin une transposée de la convolution. On copie bien sûr le poids de chaque neurone

[1] L'abstraction des données peut être considérée comme les saillances lexicales d'un texte qui lui donnent une identité propre

[2] \Correction de l'erreur à chaque phase d'apprentissage.

après l'entraînement dans cette copie de réseau et on obtient un nouveau réseau dont la couche de sortie correspond au résultat de chaque filtre de la convolution. Une simple somme de ces filtres pour chaque mot nous donne un indice d'activation du mot dans son contexte. Au final nous observons ici des zones de texte s'activer plus ou moins suivant l'importance que leurs a accordée le réseau.

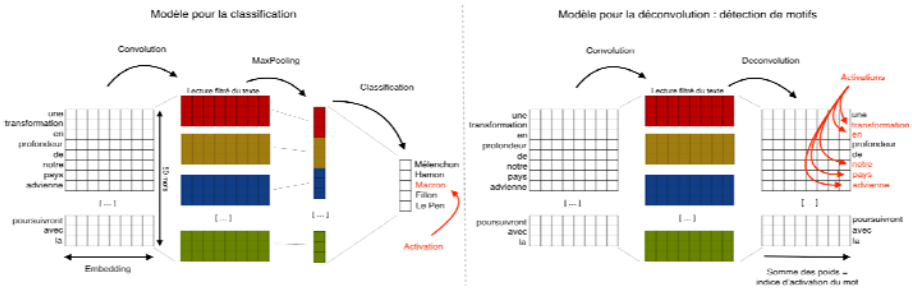


Figure 2. Convolution et déconvolution d'un passage du discours d'E. Macron

3.1. Résultats et perspectives

A la lecture des résultats, nous voyons que le modèle identifie, sans surprise, des mots que le traitement statistique avait calculés comme spécifiques. Mais pas seulement. Certaines zones éclairées par le réseau semblent relever d'une nouvelle forme de lecture du texte. Nous pouvons illustrer ce constat avec un extrait des vœux d'E. Macron le 31 décembre 2017:

[...] une transformation en profondeur de notre pays advienne à l'école pour nos enfants , au travail pour l' ensemble de nos concitoyens , pour le climat , pour le quotidien de chacune et chacun d' entre vous . Ces transformations profondes ont commencé et se poursuivront avec la [...]

Dans ce passage, les mots *transformation* et *notre*, fortement spécifique de Macron, sont activés : ici il n'y a pas de plus-value heuristique par rapport à l'ADT. De même, le segment répété *chacune et chacun*, très spécifique, est repéré par le réseau. Mais il y a aussi les mots *pays* et *advienne* qui ne sont pas statistiquement spécifique de Macron et qui ont pourtant fortement contribué à la reconnaissance du passage. Si l'on regarde maintenant les activations autour de ces mots ciblés, on voit que c'est une expression formée de plusieurs mots, pas forcément contigus, qui est repérée par le réseau. Il semble donc que le deep learning ait identifié des structures phraséologiques ou motifs linguistiques sensibles aux occurrences et à leur organisation syntagmatique. Plus loin, la visualisation du passage dans son ensemble met au jour une topologie textuelle ou un rythme auxquels le deep a été sensible (Figure 3).

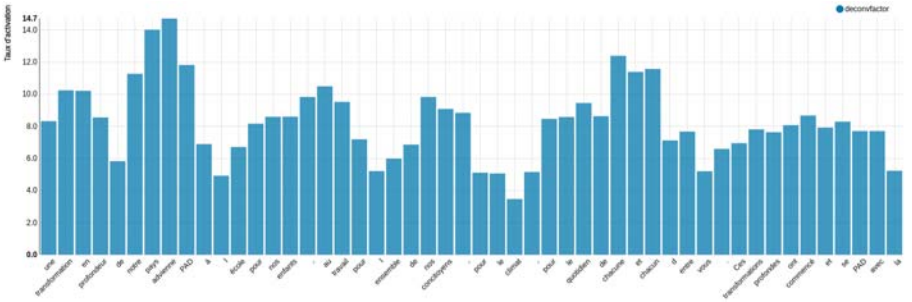


Figure 3. Déconvolution : observation de la topologie d'un passage

3. Conclusion

L'ADT et le deep learning ne sont peut-être pas des continents étrangers l'un à l'autre (Lebart, 1997). Cette contribution en croisant approche statistique et réseau de neurones nous a permis d'identifier des passage-clefs et peut-être des motifs susceptibles de nourrir nos traitements textuels. Si les observables qui ont présidé à la détection de passages-clefs par l'ADT (les spécificités lexicales) sont connus et éprouvés, les zones d'activation du deep learning semblent relever de nouveaux observables linguistiques. Rappelons que la matière linguistique et la topologie des passages ne sauraient renvoyer au hasard : les zones d'activations permettent d'obtenir des taux de reconnaissance de plus de 90% sur le discours politique français et de 85% sur le corpus du LASLA ; soit des taux équivalents ou supérieurs aux taux obtenus par le calcul statistique des passages-clefs. Reste désormais à améliorer le modèle et à en comprendre tous les aboutissants mathématiques comme linguistiques. La première amélioration que l'on se propose désormais d'implémenter est l'injection d'informations morphosyntaxiques dans le réseau afin de mettre à l'épreuve des motifs linguistiques toujours plus complexes.

References

Ducoffe, M., Precioso, F., Arthur, A., Mayaffre, D., Lavigne, F., et Vanni, L. (2016). Machine learning under the light of phraseology expertise : use case of presidential speeches, de Gaulle - Hollande (1958-2016). *Actes de JADT* 2016, pages 155–168.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP*, pages 1746–1751.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris, Slatkine-Champion.

Lebart, L. (1997). Réseaux de neurones et analyse des correspondances. *Modulad*, (INRIA Paris), 18, pages 21–37.

Longrée, D. et Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Etendre le champ de la phrase´ologie de la langue au discours. *Langages* 189, pages 65–79.

Mellet, S. et Longre´e, D. (2009). Syntactical motifs and textual structures. *Belgian Journal of Linguistics* 23, pages 161–173.

Mellet, S. et Longrée, D. (2012). Légitimité d'une unité textométrique : le motif. *Actes de JADT 2012*, pages 715–728.

Mikolov, T., Chen, K., Corrado, G., et Dean, J. (2013). Efficient estimation of word representations in

vector space. *ArXiv* : 1301.3781.

Quiniou, S., Cellier, P., Charnois, T., et Legallois, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. *Actes de JADT 2012*.

Rastier, F. (2007). Passages. *Corpus* 6, pages 25–54.

Salem, A. (1987). *Pratique des segments répétés. essai de statistique textuelle*. Paris : Klincksieck.

Déconstruction et reconstruction de corpus... À la recherche de la pertinence et du contexte

Lucie Loubère
Lerass Université de Toulouse – lucie.loubere@iut-tlse3.fr

Abstract
Faced with corpora of large sets of texts, we propose a method of selection, based on the identification of segments of texts relevant to a topic by successive classification, then recomposition of the corpus with all the texts having at least one relevant segment . This approach makes it possible to preserve the contextualizations and narrative discourses surrounding a theme while excluding off-topic texts.

Résumé
Face aux corpus constitués de grands ensembles de textes, nous proposons une méthode de sélection, basée sur l'identification de segments de textes pertinents à une thématique par classification successive, puis recomposition du corpus avec l'intégralité des textes ayant au moins un segment pertinent. Cette démarche permet ainsi de conserver les contextualisations et discours narratifs entourant une thématique tout en excluant les textes hors-sujet.

Keywords: Big corpus, Reinert classification, Iramuteq

1. Introduction

La multiplication d'outils d'extraction de contenus numériques ou l'abonnement des universités aux bases de données de presse, sont autant de raisons favorisant la création de corpus de grande taille. À ces facilités grandissantes s'opposent de nouvelles difficultés. L'hétérogénéité des contenus mis à disposition par une communauté, les algorithmes de recherche de bases de données, ou simplement les limites d'ambiguïté de requêtes génèrent de nombreux bruits à nos corpus. Nous proposerons ici une méthode s'appuyant sur une identification de contenu par classification successive (Ratinaud et Marchand, 2015), puis une régénération du corpus par concaténation de l'intégralité des articles contenant au moins un segment de texte (ST) dan le matériel identifié comme pertinent.

2. Problématique

La sélection de corpus par classifications successives, en utilisant comme

unité le segment de texte, permet d'obtenir un sous corpus pertinent avec une thématique (Loubère, 2014; Ratinaud et Marchand, 2015). Cependant, lorsque le corpus de départ est constitué de textes au contenu narratif structuré et délimité (article de presse, blog, argumentaires dans une concertation…) ce processus peut supprimer les éléments périphériques au thème étudié. Ces contenus restent portant pertinents pour la compréhension de l'objet d'étude, mais peuvent être classés avec le bruit des textes hors sujet dès les premières étapes de sélection. L'objectif de cettte méthode est donc d'exclure le bruit de textes hors-sujets tout en conservant le contexte d'évocation de la thématique principale.

3. Méthodologie

Le processus proposé ici se décompose en trois étapes :

k) Numérotation des textes par un identifiant en méthadonnée
l) Extractions des segments de textes propres à notre thématique par classifications successives. Cette étape repose sur la classification hiérarchique descendante (CHD) de type Reinert (Reinert, 1983) proposée par le logiciel Iramuteq (Ratinaud, 2009). En permettant de faire émerger les mondes lexicaux, ce traitement nous permet de sélectionner les segments concernant notre thématique, puis de les re-soumettre à une CHD afin de préciser le corpus. Cette étape est reconduite jusqu'à avoir une classification dont toutes les classes concernent la thématique étudiée.
m) Re-composition du corpus par concaténation des articles apparaissant au moins une fois dans l'extraction finale de l'étape 2

4. Exemple empirique

Dans les parties qui suivront, nous présenterons une mise en application de cette méthode sur un corpus utilisé lors de notre thèse (Loubère, 2018). Il est constitué d'une extraction d'article de presses quotidiennes nationales (libération, l'humanité, le monde, la croix, le figaro) portant sur la thématique du numérique éducatif du 01/01/2000 au 31/12/2014. Afin de couvrir le plus d'informations possible la requête exécutée sur la base de donnée d'Europresse retournait tous les articles contenant au moins un terme éducatif dans la liste : collège, lycée, école, éducation et au moins un terme numérique dans la liste : numérique, informatique, multimédia, TICE.

4.1. Les classifications successives

Cette extraction retourna 18 804 articles, auxquels nous avons retiré 875 doublons. LE corpus exploité ici est donc constitué de 17 929 articles représentant 450 815 segments de textes, sur lesquels nous avons apposé en

méthadonnée le numéro de l'article source. Nous allons présenter ici les classifications successives

Nous avons effectué une CHD de 20 classes en phase 1 et un minimum de 1000 ST par classe, nous obtenons 16 classes représentant 99,72 % du corpus. Le résultat obtenu est présenté sur le dendrogramme en illustration 1

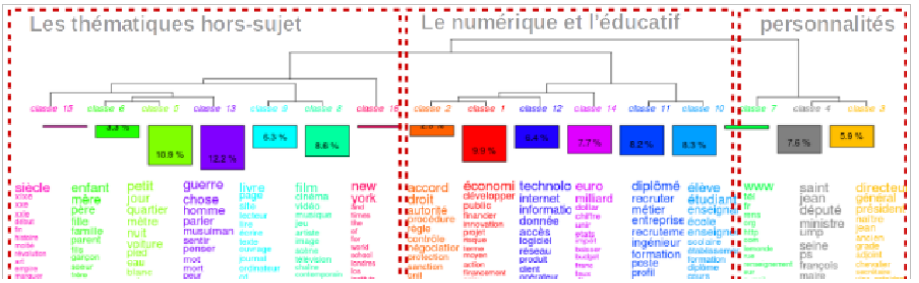


Illustration 1 : dendrogramme de la première CHD

Ce premier découpage montre une séparation en 3 blocs. Le premier est composé des personnalités publiques, le second est composé par des thématiques extérieures à notre sujet. En effet, de nombreux articles contiennent les termes de notre requête sans être pour autant dans le domaine éducatif (ou numérique). Ainsi, les classes 9 et 8 regroupent les actualités ou dossiers portant dans le domaine de la culture. Nous citerons comme exemple non exhaustif d'article de ce domaine un article du journal Le monde commentant les sorties cinématographiques dans lequel nous relèverons « les enfants privés d'école jouant dans les rues », et pour un autre film « les décors numériques ». Nous retrouvons sur le même principe les classes 6, 5 et 13 traitant des conflits armés détruisant les lycées et relatant une infériorité numérique.Enfin, le troisième bloc présente une classe centrée sur le numérique (classe 12), deux classes centrées sur l'éducatif (11 et 10) et deux classes sur l'aspect législatif et économique (classes 1 et 2). Afin de pouvoir affiner ces thématiques et les possibles interactions, nous avons choisi de conserver le bloc entier, soit les segments composant les classes 1, 2, 10, 11, 12 et 14. L'export précédent nous a permis d'obtenir 194 966 segments de texte sur lesquels nous avons effectué une deuxième CHD de 15 classes en phase 1 et seuil minimal de 100 ST. Nous obtenons 14 classes portant sur 99,97 % des segments. Le résultat est présenté en illustration 2.

Ce deuxième découpage reprend une structure en trois groupes. Ici, nous relevons le contexte économique du marché du numérique (classe 14, 5 et 6).

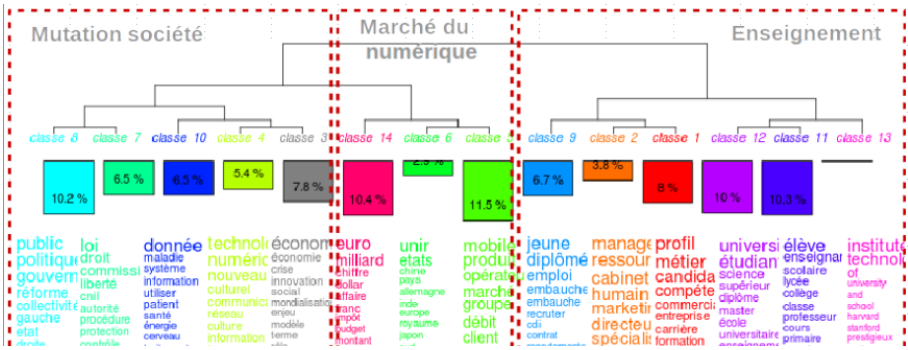


Illustration 2 : dendrogramme de la deuxième CHD

Le second bloc (classe 4, 3, 7, 8, 10) est constitué des différents discours témoins de la numérisation de la société. Le troisième groupe séparé du reste du corpus par le premier facteur est centré sur-le-champ éducatif. Les trois premières classes à se détacher partagent un discours sur l'après-formation et le recrutement (classes 9, 2 et 1). La classe 11 constituant 10,3 % du corpus est entrée sur l'éducation primaire et secondaire, alors que la classe 12 porte sur l'enseignement supérieur et la recherche. Notre étude portant sur le système scolaire secondaire, nous ne conserverons que la classe 11 pour l'étape suivante.

L'export de cette dernière constitue un corpus de 20 167 segments de texte sur lesquels nous avons effectué une CHD de 15 classes en phase 1 et un minimum de 100 ST par classe. Nous obtenons 8 classes rapportant 99,22 % des segments. Ce dendrogramme, structuré en deux blocs, nous montre une séparation entre un discours centré sur l'aspect structurel de l'éducation (classes 8, 6, 4, 3) et celui traitant de l'enseignement (classes 2, 1, 5, 7).

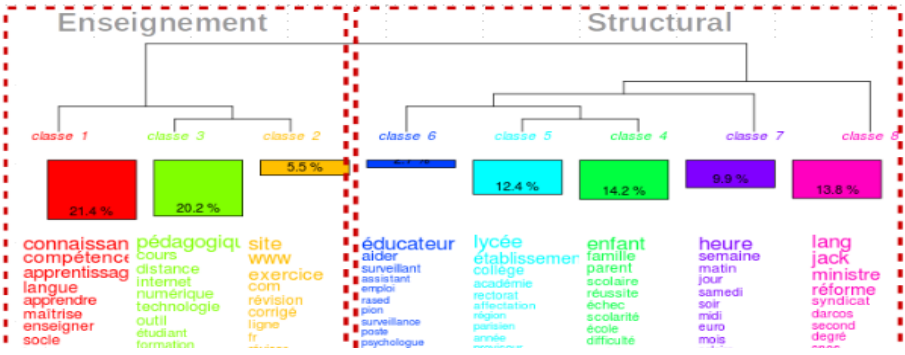


Illustration 3: dendrogramme de la troisième CHD

Dans la partie structurale nous retrouverons les segments de texte traitant des réformes sous un angle gouvernemental (classe 8), suivie de tout le discours se regroupant des aspects temporels, comme le temps de travail mais également les rythmes scolaires (classe 6). La classe 3 constitue un discours sociologique sur l'éducation, nous y retrouvons de nombreuses statistiques étudiant les répartitions sociales dans les différents cursus. Enfin, la classe 4 traite des établissements scolaires dans leurs diversités.
Les autres classes portent toutes sur le domaine pédagogique : la classe 7 concerne les contenus d'enseignement. La classe 5 traite de la mise en place d'outil numérique parascolaire (jeux éducatifs, fiche de révision) alors que la classe 2 est centrée sur la mise en place de formations à distance. Enfin, la classe 1 est le discours portant sur le numérique dans l'éducation, les mots clés employés dans notre requête y sont tous surreprésentés. Nous ne conserverons dons que les segments composant cette classe.
L'extraction de cette dernière classe nous permet d'obtenir 2072 segments sur lesquels nous avons effectué une CHD de 20 classes en phase 1 avec un seuil de 100 ST par classe. Cette classification nous a montré une réelle stabilité de la thématique. En effet, les 8 classes exposée portent chacune sur un aspect du numérique éducatif.

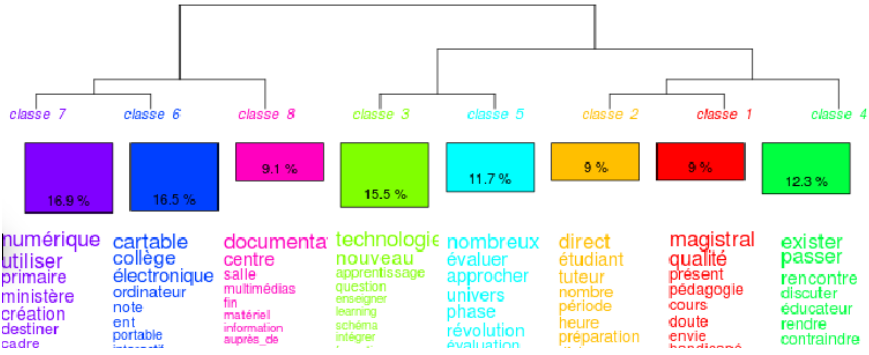


Illustration 4 : dendrogramme de la quatrième CHD

4.2. Classification du corpus recomposé

Le corpus recomposé des 2902 articles contenant au moins un segment de texte dans la classe 1 de la troisième CHD est constitué de 72460 segments. Une CHD de 20 classes en phase 1 et un minimum de 800 ST par classe nous donne le dendrograme suivant :

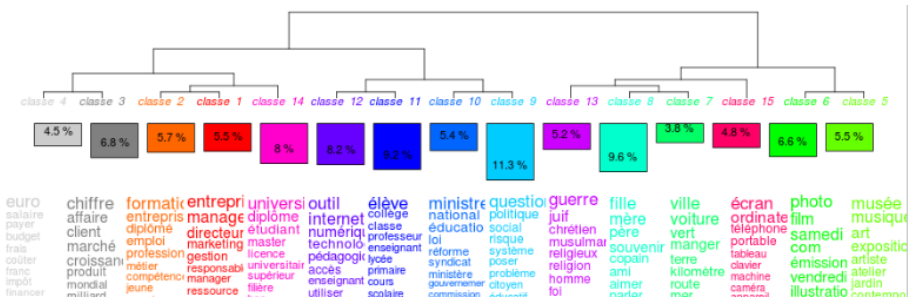


Illustration 5 : dendrogramme de la CHD sur le corpus recomposé

Nous y retrouvons donc au-delà de discours sur l'utilisation du numérique dans les établissements, un discours sur l'économie reflétant le marché du numérique éducatif et les frais engendrés par les dotations des établissements. Un discours à la frontière de la culture et de l'éducation, avec les formations de ces domaines empreinte de numérique. Mais également un discours sur l'actualité géopolitique mondiale contextualisant des initiatives où le numérique apporte des solutions éducatives lors de ségrégation ethniques, ou éloignements géographiques. Tous ces mondes lexicaux constituent des éléments du discours social sur notre sujet, qu'une étude réduite aux segments ciblés lors des CHD successives ne permettrait pas d'explorer.

5. Conclusion

Le principe des CHD successives, s'il nous permet d'accéder finement aux segments contenant le discours sur le numérique éducatif, nous éloigne d'une compréhension globale du sujet. En effet, interroger les bases de données de presse sur une longue période et une sélection de presse généraliste apporte une quantité importante de documents hors contexte. Ces données portent des éléments contextuels communs avec les articles traitant de notre sujet (personnalités politiques, discours économique…), la proximité lexicale des segments de ces champs structure les classes de discours communes aux articles portant sur notre sujet ou non. Cette hétérogénéité associée à l'insécurité d'un grand ensemble (Geffroy et Lafon, 1982) nous empêchant une connaissance du corpus antérieure à l'analyse lexicométrique conduit « à tracer un peu trop vite une autoroute » (Geffroy et Lafon, 1982, p. 140) jusqu'à notre classe 1 finale. Ce phénomène questionne la constitution d'un corpus sur une dimension architextuelle, alors même que l'outil de classification utilisé ici joue sur un niveau intertextuel et cotextuel (Rastier, 2015), rapprochant des passages de textes en fonction de leur structure lexicale. La présence de textes aux sujets hétéroclites fait ressortir de façon

précoce des thématiques indépendamment de leur hypothétique poids dans le corpus qu'aurait constitué une sélection de textes centrés sur notre sujet. Ainsi, les segments traitant de sujets de politique générale ou exposant le contexte social d'un pays dans les articles traitant du numérique éducatif sont classés avec ceux des articles hors sujets. Cette difficulté éloigne le chercheur de la compréhension d'un discours. La démarche que nous venons de présenter nous permet de se rapprocher d'un positionnement de textomètre (Pincemin, 2012), sélectionnant les segments pertinent par une démarche inductive, mais en conservant l'unité sématique du texte dans la construction du corpus final.

Bibliography

Geffroy, A., & Lafon, P. (1982). L'insécurité dans les grands ensembles. Aperçu critique sur le vocabulaire français de 1789 à nos jours d'Etienne Brunet. Mots, 5(1), 129-141.

Loubère, L. (2014). Le traitement des TICE dans les discours politiques et dans la presse. In Présenté à 12èmes Journées internationales d'Analyse statistique des Données Textuelles.

Pincemin, B. (2012). Sémantique interprétative et textométrie. Texto! Textes et Cultures, 17(3), 1-21.

Rastier, F. (2015). Arts et sciences du texte. Paris: Presses universitaires de France.

Ratinaud, P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. Consulté à l'adresse http://www.iramuteq.org

Ratinaud, P., & Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). Mots. Les langages du politique, (2), 57-77.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données, 8(2), 187-198.

L'apport du *corpus-maquette* à la mise en évidence des niveaux descriptifs de la chronologie du sens. Essai sur une Série Textuelle Chronologique du *Monde diplomatique* (1990-2008).

Heba Metwally
Université d'Alexandrie, Égypte – heba.metwally77@gmail.com

Abstract
Chronological corpora and particularly time series (Lebart et Salem 1994) organize the textual data in corpora according to their natural sequence in time. Today, scholars are interfacing increasingly with chronological corpora following the democratization of access to big data. The lexicometry develops into stylometry, textometry and logometry. And statistical data analysis integrates the observation of co-occurrential systems and lexical networks in their complexity. This improves the analysis of semantic contents according to their localisation in the semantic strata. This contribution aims to enhance the description of the chronology of meaning. The study is based on a corpus of more than 5000 articles (ca 11 millions of tokens) published in the *Monde diplomatique* between January 1990 and December 2008.
To analyze big chronological corpora we propose a *scale model* of the chronological corpus by compressing the initial corpus to its most frequents nouns. The compression procedure is duplicated in the four sub-corpuses of relevant semantic stability. We obtain two descriptive levels of chronology: the synthetic level of dominant contents and the analytical level of the four chronological phases of meaning. The two levels are intended to respond to different investigations on time and meaning. Working on sets of *scale models* that are either connected horizontally (chronological sequence) or vertically (the synthetic perspective clarified by an analytic perspective) enlarges our field of observation and deepens our understanding of chronological data in particular and the unfolding of text in general.

Keywords: chronological corpus – logometry – logogenesis – clustering – method Reinert – corpus semantics – media analysis

Résumé
Les corpus chronologiques et a fortiori les Séries Textuelles Chronologiques (Lebart et Salem, 1994) organisent les données textuelles dans le corpus selon leur enchaînement naturel dans le temps. La banalisation des corpus textuels et l'accès facilité et accéléré au big data multiplient les corpus chronologiques, puisque finalement toute production textuelle s'étale dans le temps. La lexicométrie – au sens classique – doublée de la stylométrie, de la textométrie voire de la logométrie, et la statistique occurrentielle enrichie par un outillage cooccurrentiel (Viprey, 1997), (Mayaffre, 2014), la voie est ouverte aujourd'hui à une observation améliorée des contenus sémantiques qui gagnent en visibilité grâce aux tentatives parfois incontrôlées de leur objectivation. Cette contribution a pour objectif de contribuer à la description de la chronologie des contenus sémantiques. On s'appuie sur un corpus d'articles du MD (1990-2008). On compte plus de 5000 articles et plus de 11 millions d'occurrences. On propose pour cela le recours à un *corpus-maquette*, une compression du corpus chronologique intégral à partir des noms les plus fréquents. Cette démarche de compression est reproductible dans les sous-corpus des périodes de stabilité sémantique. On obtient deux niveaux descriptifs de la chronologie, à savoir le niveau global, synthétique des contenus dominants et le niveau subordonné, analytique des sens particuliers des phases transitoires du discours. Les deux niveaux infèrent sur un questionnement différent sur le temps en multipliant les pistes d'interrogation et en articulant le niveau synthétique et son niveau analytique.

Mots-clés: corpus chronologique – logométrie – logogénétique – classification – méthode Reinert – sémantique de corpus – Analyse de discours médiatique

1. Introduction

Dans la tradition lexicométrique, les STC (Séries Textuelles Chronologiques) problématisent les investigations sur le temps[1]. Ce type de corpus est né, dans les études à caractère historique, du questionnement sur le changement dans le discours au fil du temps. Et les travaux d'André

[1] « Nous appelons séries textuelles chronologiques ces corpus homogènes constitués par des textes produits dans des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables. » (Lebart et Salem, 1994 : 217)

Salem[2] témoignent de l'intérêt porté à la description des corpus textuels chronologiques. Pour ce faire, André Salem généralise les STC, décrit la particularité des sorties machines des analyses statistiques qu'elles produisent (AFC ; calcul de spécificités), introduit la notion de «temps lexical », et conçoit une gamme de calculs visant, dans un premier temps, la « mise en évidence et la mesure du stock lexical au cours du temps » (Salem, 1988 : 118) et, dans un second temps, la caractérisation des périodes dans une STC. Plus généralement, la particularité des STC est de concilier la linéarité du texte, du temps et la sérialité du corpus. Si tous les corpus sont partitionnés en séries pour permettre la comparaison, ces séries ont l'avantage de conserver l'ordre naturel des textes qui s'échelonnent – sans conflit – dans le corpus et dans le temps.

Aujourd'hui, le champ des observables est constamment élargi grâce à l'évolution des outils informatiques et au progrès de la tokenisation pour embrasser progressivement des niveaux descriptifs textuels que le chercheur filtre ou articule à sa guise. La lexicométrie est enrichie et mise à jour par la textométrie et la logométrie dont le projet est de dépasser la lexie vers les textes, le discours et le sens. Le sens est objectivable grâce à la formalisation de la cooccurrence, et à son baptême comme unité minimale de contextualisation, i.e de sens (Mayaffre, 2008). Dès lors, la statistique occurrentielle se double de la statistique cooccurrentielle. La cooccurrence devient unité de décompte généralisée à laquelle s'applique les calculs statistiques traditionnels (Brunet, 2012). Des applications d'ADT de tradition benzécriste se développent pour appréhender les réseaux lexicaux dans leur complexité. La *cooccurrence généralisée* (Viprey, 1997, 2005, 2006) se donne une visée exploratoire et la méthode Alceste (Reinert, 1983, 1993) procède à la démarche classificatoire des réseaux lexicaux structurants des textes. C'est dans ce cadre des progrès de la méthodologie et de la technologie qu'une *sémantique de corpus* (Rastier, 2011) est envisageable.

Ce champ d'investigation intéresse naturellement les études chronologiques qui peuvent désormais observer le mouvement des contenus sémantiques dans le temps pour comprendre l'impact du temps dans la thématisation d'une Série Textuelle Chronologique[3]. Pour l'objectivation des *fonds*

[2] Cf. (Salem, 1988, 1991, 1993, 1994)

[3] Ce point précisément constitue la problématique de notre thèse de doctorat intitulée « Les thèmes et le temps dans *Le Monde diplomatique* (1990-2008) », soutenue le 11 décembre 2017 à l'Université Côté d'Azur (UCA) à Nice.

sémantiques[4] du discours, on sollicite la méthode Alceste implémentée dans le logiciel libre Iramuteq (Ratinaud et Marchand, 2012) qui s'articule à Hyperbase. Pour une visualisation améliorée des topics du discours, on propose de recourir à une *maquette* du corpus et de ses sous-corpus. Au sens propre, la *maquette* est une représentation en trois dimensions, à échelle réduite qui reste fidèle dans ses proportions. Ici, dans le cas des corpus textuels, la *maquette* est une compression du corpus intégral qui se réduit à ses noms les plus fréquents. A partir d'une STC du *Monde diplomatique* (1990-2008), cette contribution se donne deux objectifs. Dans un premier temps, elle vise à mettre en exergue les deux niveaux descriptifs complémentaires de la chronologie du sens : chronologie des contenus dominants (3.) et la *logogénétique* (4.) tout en relevant l'intérêt de étude conjointe de ces deux niveaux. Dans un second temps, il s'agit également de mettre à l'épreuve notre proposition de la *maquette*. On recherche une visualisation améliorée des contenus sémantiques structurants grâce au recours à une *maquette*, reproduction grossière et fidèle des textes dont l'usage spécifique sera illustré dans les lignes suivantes.

2. Du corpus intégral à la *maquette* du sens et du temps

Le choix du *Monde diplomatique* pour l'étude de l'évolution du sens s'appuie sur la richesse et la stabilité de son contenu. La période couverte par cette étude marque un moment historique important, à savoir le monde après la chute du Mur de Berlin. En plus, cette période se caractérise par une continuité éditoriale[5]. Bref, nous avons affaire à un discours stable, sans complexe qui à l'examen multidimensionnel épouse un schéma évolutif classique sans ruptures[6]. On estime que la stabilité du discours est un facteur indispensable à l'étude de l'évolution, celle-ci reposant principalement sur la continuité.

La finalité de ce travail, à savoir l'étude de la chronologie du sens d'un gros corpus textuel, préside à la conception de la *maquette*. La taille du corpus

[4] Les *fonds sémantiques* sont les *isotopies* ou les macrostructures sémantiques sur lesquelles se détachent les *formes sémantiques* que sont les *thèmes*. Cf. (Rastier, 2011 : 24)

[5] Il s'agit du mandat d'Ignacio Ramonet qui est directeur de la publication de janvier 1990 à mars 2008.

[6] Par examen multidimensionnel on entend l'AFC de la distance entre les textes qui dans le cas des données sérielles reproduit une forme parabolique baptisée parabole Guttman et qui est symbolique du mouvement linéaire des données ordonnées dans le temps. Cf. (Salem, 1991).

intégral excédant 11 millions d'occurrences (voir ci-dessous Tableau 1) pose immédiatement le problème de son interprétation comme il nous confronte à la difficulté de l'appréhension des *fonds sémantiques* structurants du corpus. En ADT, les chercheurs procèdent assez souvent pour des raisons pratiques à des sélections au sein de la population statistique étudiée. A notre tour, on propose un mode de réduction qui se fonde sur la finalité herméneutique et perpétue la pratique d'une sémantique interne. On pose ici – sans généraliser – que le discours médiatique par sa vocation informative et sa référence au monde structure son contenu d'une manière privilégiée autour des noms. La classe nominale (noms communs et noms propres) est la classe grammaticale la plus importante dans le corpus ; elle couvre 28,9% de la surface du corpus. Elle connaît également une stabilité distributionnelle au fil de la STC. L'importance numérique absolue et la distribution équilibrée attestent le critère de la représentativité statistique[7]. Aussi une comparaison avec d'autres corpus[8] entre les listes des lemmes les plus fréquents triés par catégorie grammaticale confirme le pouvoir caractérisant de la classe nominale en général et des noms propres en particulier. On s'appuie donc sur la classe nominale et l'argument fréquentiel pour réduire le corpus intégral à ses 380 noms les plus fréquents. La démarche laisse intacts les partitions du corpus et l'enchaînement des textes pour respecter la structure séquentielle des textes et la conception chronologique du corpus. L'une et l'autre garantissent au corpus textuel son authenticité ; seul leur maintien autorise l'examen de l'hypothèse de travail présidant à la conception du corpus textuel. Pour expliquer un peu ce travail philologique simple dans son principe, la démarche consiste à mettre un cache sur tout le texte à l'exclusion des 380 noms les plus fréquents. Cette procédure est à reprendre dans les sous-corpus de stabilité sémantique. Celle-ci se laisse mesurer d'une manière endogène à l'aide du calcul de la distance entre les textes à partir de la forme minimale de signification thématique, la cooccurrence. La distance intertextuelle calculée sur les cooccurrences au sein des noms de la *maquette* donne à voir quatre périodes qui fondent les quatre sous-corpus, ceux-ci réduits à leur tour à des *maquettes*. Cette périodisation endogène fonde le

[7] Dans notre travail doctoral (Metwally, 2017), nous avons étudié les contenus des classes de fréquences du corpus intégral pour une compréhension de la hiérarchie numérique du lexique. Aussi avons-nous analysé la structure grammaticale des données et leur distribution dans la STC.

[8] (Labbé et Monière, 2003); (Mayaffre, 2004).

temps sémantique[9] selon lequel on remodèle le corpus intégral et sa *maquette*. Le tableau 1 (ci-dessous) synthétise la structure lexicale du corpus, des sous-corpus et de leurs *maquettes*. Celles-ci couvrent chacune approximativement 9,8% de la surface leurs corpus originaux respectifs. Cette stabilité de représentativité numérique autorise la comparaison entre les données.

Tableau 1: Tableau synthétique de la structure lexicale du corpus, des sous-corpus et de leurs maquettes

corpus et sous-corpus	1990-1993	1994-1997	1998-2001	2002-2008	1990-2008
taille (N=occurrences)	2697013	2402434	2552998	3765908	11418356
vocabulaire (V=mots)	67989	67571	70954	86032	140690
maquette et *sous-maquettes* (V=noms)	307	282	290	375	380
maquette et *sous-maquettes* (taille)	266439	218643	229119	382298	1115311

On obtient donc finalement un dispositif complexe à deux niveaux : le niveau global des contenus sémantiques de l'ensemble de l'empan chronologique étudié dont on peut étudier la dynamique (3.); et le niveau analytique, d'ordonnancement chronologique, des phases sémantiques stables et qui permet et l'observation du mouvement des contenus sémantiques et la confrontation avec le niveau global synthétique (4.). L'étude des *fonds sémantiques* est concevable en mobilisant la statistique cooccurrentielle qui met en évidence les structures sémantiques pertinentes. A l'issue de la CHD appliqué à la *maquette* et ses *sous-maquettes*, sont observables les mondes lexicaux stabilisés (Reinert, 1993, 2008) du sens global et de ses phases transitoires (voir les dendrogrammes Fig. 1, 3, 4).

3. La dynamique des contenus dominants

La démarche habituelle dans les études chronologiques repose d'abord sur une étude statique première du sens global pour procéder ensuite à une vue dynamisée. Les vues statiques relèvent d'un artifice méthodologique provisoire destiné à mettre en évidence les contenus sémantiques stabilisés

[9] On s'est permis de parler de *temps sémantique* à la suite du *temps lexical* d'André Salem (1988). Le *temps sémantique* est le rythme selon lequel s'organisent dans le temps les contenus sémantiques et que mesure ici la distance intertextuelle calculée sur la cooccurrence.

au bout d'un mouvement dynamique. La saisie du sens global répond au questionnement sur les contenus dominants, consensuels d'une période à l'autre, qui survivent au cours de 19 ans de production d'articles. Pour l'analyse de la structure sémantique de la *maquette*, on donne à Iramuteq la *maquette* globale, où les 380 noms les plus fréquents s'organisent sur l'axe syntagmatique selon l'ordre de leur apparition, et dont les partitions assurent au corpus une structure chronologique adaptée au *temps sémantique* du corpus. Une fois Iramuteq mobilisé, il se met à découper le texte en segments de textes paramétrables. Le choix de l'étendue des segments de textes (ST) est capital, car ce sont les ST qui constituent les énoncés analysés et classés par la méthode Alceste. Pour ces unités de contexte on a estimé la succession de 10 noms dans le *corpus-maquette* comme l'équivalent dans le corpus intégral de la fenêtre contextuelle de 33 mots[10]. On vise par là un espace intermédiaire entre la phrase et le paragraphe. Une fois Alceste activé, il procède à une CHD qui croise les ST et les noms pour effectuer un classement partant du caractère lexical prédominant des ST.

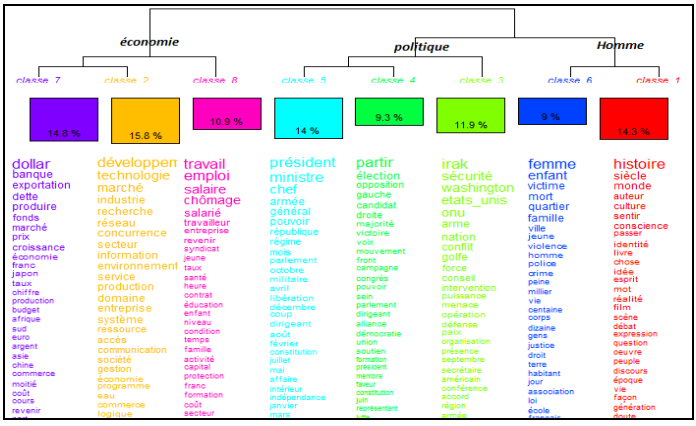


Figure 1 : Les mondes lexicaux de la maquette (1990-2008)[11]

On impose à l'algorithme un paramétrage exigeant qui nous garantit une grille de lecture assez riche. Avec 15 classes demandées à l'issue de la phase

[10] Cette estimation repose sur le pourcentage de la classe nominale dans l'ensemble du corpus (28,9%). Voir (Metwally, 2017).

[11] Dans ces listes, on peut repérer quelques verbes (partir, produire, revenir, sentir, passer). Il s'agit d'une erreur due à une lemmatisation effectuée par Iramuteq malgré les tentatives de dissuasion. Il s'agit plutôt de substantifs (parti, produit, revenu, sens, passé).

1, 8 se trouvent stabilisées (Figure 1). Les sorties machines de la CHD sont multiples. La représentation en dendrogramme correspond au classement stricto sensu ; et elle est enrichie d'informations supplémentaires qui mettent en valeur la CHD. On commence par l'identification rapide de la structure sémantique du discours et de la hiérarchie de l'information.

Le dendrogramme, par sa logique binaire de représentation, oppose les contenus économiques, les plus importants avec 41,5% des ST classés, aux contenus non-économiques. Ceux-ci distinguent les thématiques politiques (35,2% des ST classés) et les thèmes de l'Homme (23,3% des ST classés), thématiques socio-culturelles qui traitent de sujets historiques et culturels et de questions sociétales. Suivant la logique hiérarchique descendante de la classification, des classes spécialisées se stabilisent pour mieux caractériser les trois domaines sémantiques identifiés. Au sein des classes économiques se spécialise une classe socio-économique dédiée aux questions de l'emploi et du travail (classe 8 ; « emploi », « travail », « chômage », « salaire », « syndicat ») ; celle-ci se distingue des deux classes de la macro-économie qui traitent de l'économie domestique (classe 2), de la machine économique des pays (« développement », « industrie », « concurrence », « secteur »), et l'économie mondiale (classe 7) qui couvre les questions des finances et de la performance économique des pays sur le marché mondial (« dollar », « banque », « dette », « prix », « croissance »). Attachés à la même branche des thèmes politiques, les mondes lexicaux de l'Homme connaissent une variation qui différencie les questions philosophiques et/ou idéologiques sur l'histoire et la culture (classe 1 ; « histoire », « siècle », « monde », « culture », « sens », « conscience », « passé ») du quotidien des êtres humains dans ce monde (classe 6 ; « femme », « enfant », « victime », « quartier », « violence », « police », « vie », « école »). Si l'analyse du sens passe nécessairement par la suspension provisoire de la structure sérielle du corpus, l'interrogation des partitions de la *maquette* sur leur part aux classes lexicales restitue la temporalité définitoire du corpus. Une projection des classes dans les périodes de stabilité sémantique met en évidence la dynamique des classes, la thématisation de chaque période pour permettre finalement d'inférer sur l'évolution du sens.

Les classes lexicales poursuivent différentes tendances au cours du temps. Les thèmes du pouvoir (classes 4 et 5) est un axe informatif important qui ne subit guère de variations quantitatives. La classe des politiques internationales (classe 3) connait un pic positif exceptionnel dans la dernière période.

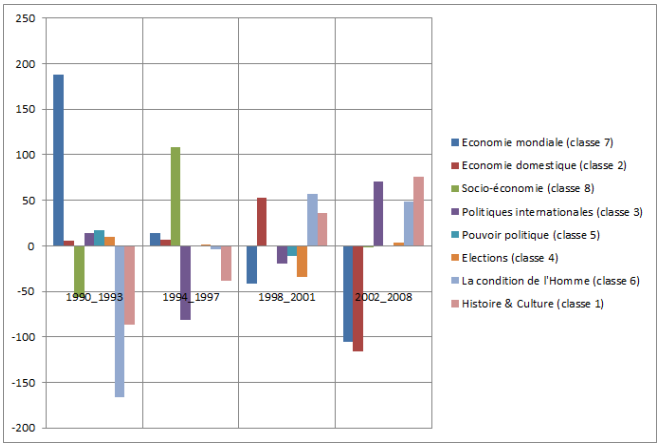


Figure 2 : Périodes et classes de la maquette (écarts en Chi2)

Ce sont les contenus économiques et socio-historiques qui sont traversés par deux logiques évolutives opposées. L'ordonnancement des bâtons positifs met en relief les pics positifs importants et exclusifs de deux classes économiques dans les deux premières périodes. Cette importance s'évanouit progressivement. Dans la dernière période les déficits les plus importants sont ceux des classes économiques. Face à la régression des contenus économiques, la progression est réservée aux contenus socio-historiques (classes 1 et 6). Il s'ensuit une couleur thématique changeante d'une période à l'autre. Les contenus économiques qui marquent les 19 ans qui ont suivi la chute du Mur de Berlin proviennent majoritairement des deux premières périodes, tandis que les deux périodes suivantes connaissent des centres d'intérêt socio-historiques qui se mêlent dans la troisième période à des thèmes économiques et dans la dernière période aux événements globaux de politiques internationales. A l'œil nu, l'histogramme de la dynamique du sens global se laisse diviser en deux moments évolutifs distincts et asymétriques. Sur le plan quantitatif, le sur-emploi de la première moitié de la série n'est jamais égalé par un sur-emploi pareil dans la deuxième moitié. Sur le plan qualitatif, les contenus majoritaires de la première partie sont des contenus techniques et relèvent de l'axe informatif le plus important, un axe technique qui relève des visions macro. Par contre, les contenus dominants de la deuxième moitié de la série sont plus variés et traduisent un intérêt croissant aux sujets philosophiques et humanistes. Un mouvement général semble déplacer le focus de l'ordre mondial vers les hommes et le sens de leur vie dans le monde.

La description de la chronologie du sens touche à ses limites. Car les contenus dominants qu'on observe ici sont précisément les contenus consensuels, ceux qui trouvent toujours leur expression d'une période à l'autre selon un dosage qui leur garantit finalement la supériorité quantitative. Le mouvement dynamique de ces contenus revient donc à une interrogation sur leurs périodes spécifiques. Ceci dit, on pose que la dynamique des contenus dominants repose nécessairement sur les sens particuliers de ces périodes. L'étude du niveau subordonné de la génétique du discours (tout de suite ci-dessous) est certes instructive pour une analyse plus détaillée de la spécificité sémantique de chaque période. L'étude de la formation du sens nous renseigne également sur le rapport entre le sens particulier, temporaire et le sens général, dominant. Elle est indispensable pour compléter et éclairer nos observations sur l'évolution.

4. La *logogénétique* ou la génétique du discours

Le mot *logogénétique* reprend le mot anglais *logogenesis* dont Halliday (1994) explicite la signification et l'intérêt en termes suivants :

> *"It is helpful to have a term for this general phenomenon – i.e. the creation of meaning in the course of the unfolding of text. We shall call it logogenesis, with 'logos' in its original sense of 'discourse' (see Halliday & Matthiessen, 1999: 18; Matthiessen, 2002b). Since logogenesis is the creation of meaning in the course of the unfolding of a text, it is concerned with patterns that appear gradually in the course of this unfolding; and the gradual appearance of patterns is, of course, not limited to single texts but is rather a property of texts in general instantiating the system of language."* (Halliday, 1994 : 601)

La *logogénétique* ou la génétique du discours permet de renouer avec les modèles linguistiques qui traversent le texte et contribuent à sa formation. Concrètement ici, on voit dans l'observation et la confrontation ordonnée dans le temps des CHD des quatre *sous-maquettes* un grand intérêt pour rétablir les modèles sémantiques propres des périodes de stabilité sémantique et qui fondent le mouvement général du sens et sa stabilisation au niveau global au cours du temps. On reprend les mêmes paramètres utilisés pour la CHD de la *maquette* globale dans les quatre *sous-maquettes* pour obtenir les dendrogrammes ci-dessous (Fig. 3, 4). Un examen attentif de la structure interne des *sous-maquettes* du sens est susceptible d'offrir des grilles de lectures analytiques des contenus dominants, de leur dynamique et de leur formation. On ne saura pas épuiser la valeur heuristique de ces

dendrogrammes. Et on se contente de souligner l'apport principal de cette démarche à la description du sens sans prétendre effectuer une analyse fouillée du sens. Celle-ci devrait reposer sur une étude systématique des réseaux lexicaux ce qui dépasse l'objectif de cette contribution.

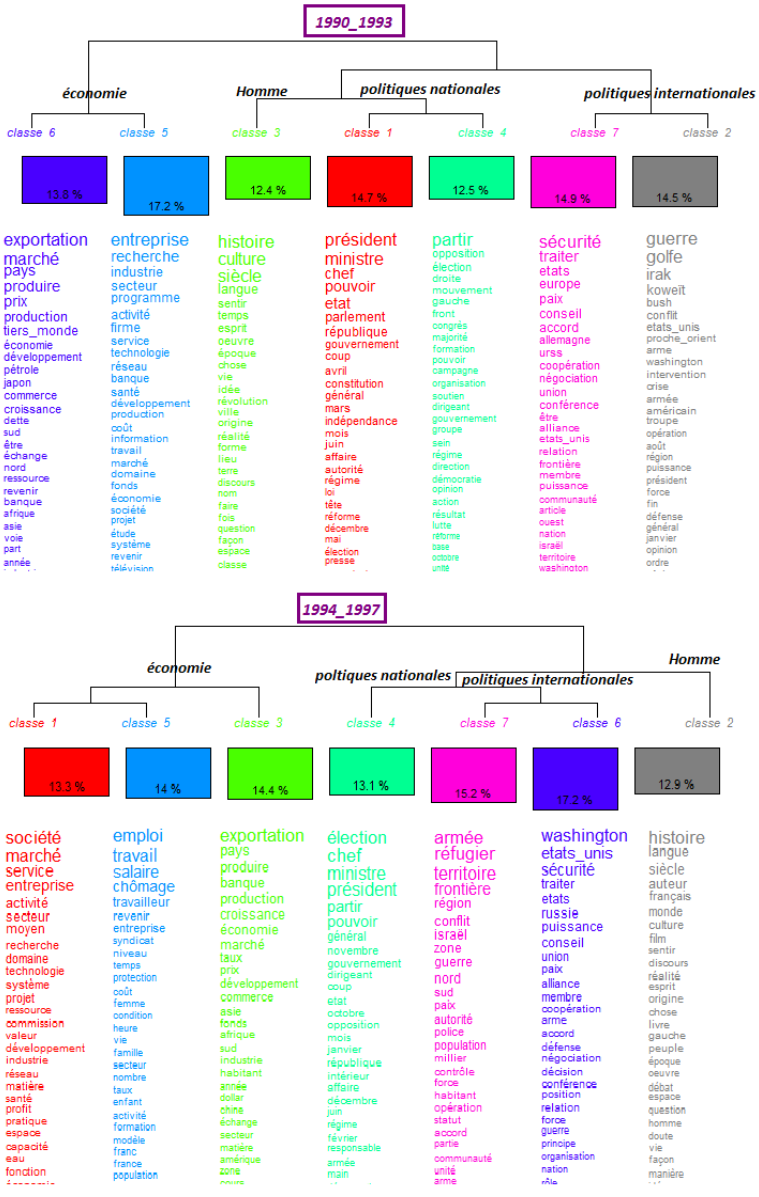


Figure 3 : Les mondes lexicaux des deux premières périodes

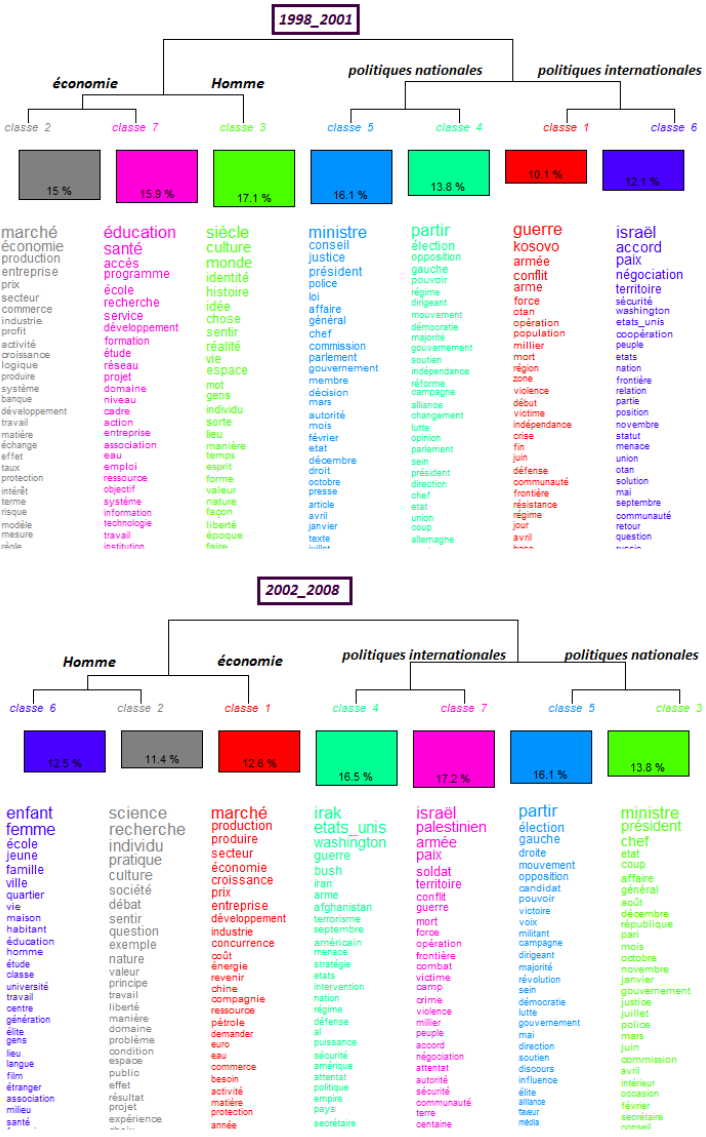


Figure 4 : Les mondes lexicaux des deux dernières périodes

La première remarque à souligner est la permanence des fondamentaux du discours et le nombre fixe de mondes lexicaux qui se stabilisent d'une période à l'autre. Cette stabilité de la structure sémantique ratifie la pertinence de l'étude de l'évolution. Celle-ci s'effectue nécessairement au sein d'un environnement stable. Observons l'évolution de la hiérarchie de

l'information d'une période à l'autre. Le graphique ci-dessous (Figure 5) rend compte de l'importance de chaque domaine sémantique au sein des ST classés. La comparaison est instructive d'une période à l'autre, et entre le niveau des *sous-maquettes* et le niveau supérieur de la *maquette* globale.

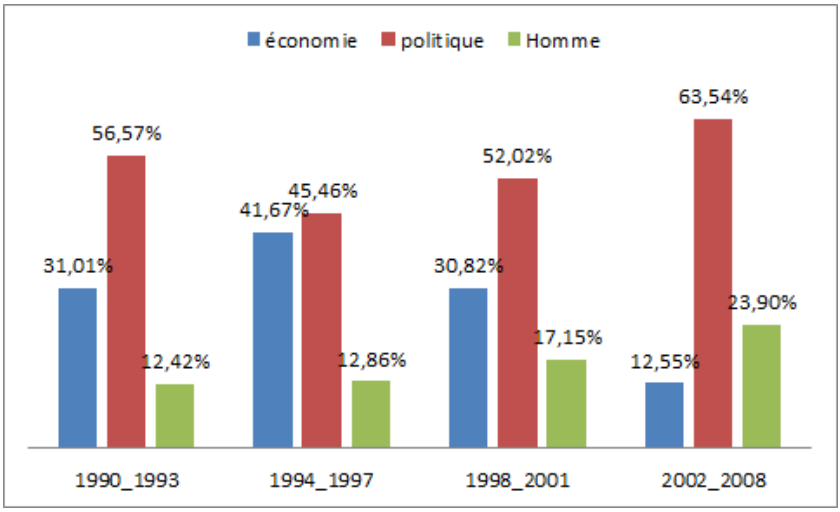


Figure 5 : L'évolution de l'importance des fondamentaux du discours au cours du temps (en pourcentages)

Quelle que soit la période, les contenus politiques restent les plus dominants. A l'examen de la répartition interne des classes politiques on note l'importance des classes de politiques internationales qui sont constamment au nombre de deux (Fig. 3, 4) par opposition au niveau global qui ne connaît qu'une seule classe (Fig. 1, classe 3). C'est l'ampleur des classes de politiques internationales dans les *sous-maquettes* qui fait la supériorité des thématiques politiques. Et pourtant, ce n'est pas le cas au niveau global. Ceci est dû principalement à la nature conjoncturelle des événements internationaux : les guerres américaines de la première et dernière période, les questions sécuritaires d'actualité en Europe après la chute du mur de Berlin, la guerre de Kosovo dans la troisième période, le conflit israélo-palestinien avec ses variantes et ses flux et reflux au cours du temps (voir le contenu des classes lexicales, Fig. 3, 4). Tant d'événements spécifiques de certaines périodes et qui ne parviennent pas tous à se stabiliser au niveau global pour caractériser les 19 ans. D'où la prédominance des contenus politiques dans les *sous-maquettes* et leur recul au niveau global.

Par contre, les contenus économiques connaissent une tendance inverse. Au niveau global, ils occupent le sommet de la pyramide hiérarchique avec trois

classes. Au niveau subordonné des *sous-maquettes*, ils viennent en deuxième rang pour passer dans la dernière période au troisième rang. Le nombre de leurs classes fluctue entre trois et un. Ce qui est curieux est que la variété maximale du nombre des classes économiques finit par se stabiliser au niveau global. À la différence des thématiques de politiques internationales, les thématiques économiques connaissent des prolongements plus pérennes. Il suffit d'observer les dendrogrammes des *sous-maquettes* pour localiser dans le temps les sources des trois classes économiques de la *maquette* globale.

Comme le montre bien l'évolution de la hiérarchie de l'information (Fig. 5), les thèmes socio-historiques continuent à s'amplifier pour dépasser les thématiques économiques dans la dernière période. Ce constat est bien compatible avec la dynamique du sens global (Fig. 2) où on a observé les déficits record des thèmes économiques et le sur-emploi significatif des classes socio-historiques. Notons également que ces dernières croissent quantitativement et qualitativement. C'est exclusivement dans la dernière période qu'on a affaire à deux classes socio-historiques. Dans cette dernière période la classe 6 caractérisée par « enfant » et « femme » ressemble à la classe 6 de la *maquette* globale (Fig. 1), tandis que la classe voisine (classe 2) lexicalisée par « science », « recherche », « individu », « pratique » n'a pas d'équivalent lexical au niveau global. Il s'agit de contenus émergents qui ne trouvent pas de précédents dans la STC. Le vocabulaire de la classe 2 se situe à mi-chemin entre le sociétal et le social. Le ST le plus caractéristique de la classe nous éclaire sur sa particularité rhétorique. A l'occasion du Sommet G8 2007 dont le thème est 'croissance et responsabilité', le *MD* lance un tract appelant à une révolution culturelle généralisée. On élargit la fenêtre de l'observation au-delà des limites du ST[12] pour améliorer l'identification du contenu sémantique:[13]

> *« A quand, là encore, la lancée d'initiatives mondiales de la part de quelques pays courageux – on attend la France – pour prendre à contre-pied la vieille tentation d'inféoder la recherche aux désignations*

[12] Tandis que le ST se limite à la succession de 10 noms parmi les 380 noms les plus fréquents du corpus, la lecture ne s'arrête pas aux frontières des ST mais elle en part. Selon Rastier (2007), le passage - *îlot de pertinence* – « n'a pas de bornes fixes et son empan dépend évidemment du point de vue qui a déterminé sa sélection » (p. 31). Notre paramétrage cible le paragraphe, i.e la période, qui relève du niveau mésotextuel, lieu de l'observation et de l'objectivation des thèmes. Et la lecture poursuit sur l'axe syntagmatique le développement d'un thème d'un ST à l'autre.

[13] Sont mis en rouge uniquement les noms spécifiques de la classe 2.

d'objectifs par quelques manipulateurs, et pour lancer les chercheurs, au contraire, à l'assaut des nouvelles questions vitales : telles, en sciences humaines, les formes de légitimité anthropologique, politique et démocratique qui conviendraient à une société-monde en formation ; telle, en sciences technologiques, la rupture nécessaire avec les grands systèmes énergivores, laquelle permettrait demain aux sociétés – locales, urbaines, régionales – d'assurer leur autonomie alimentaire et énergétique sans se désengager de la conversation mondiale autorisée par la circulation instantanée des données ? Bref, le pire des réflexes de solidarité défensive ne parvient plus à occulter les questions désormais immédiatement planétaires : celle qu'on ne tergiversera plus à nommer simplement la nature, ce support de la vie terrestre devenu poste de résistance principal pour le mirage de la valeur argent ; celle de la culture, aussi bien identitaire et artistique que scientifique, et qui constitue – au moins à l'égal de la production matérielle désormais technologisée – un vaste univers d'activités essentielles, dont la logique ouverte ne peut être inféodée au rendement de type industriel ou financier sans péril pour l'humanité civilisée, et pour sa pluralité démocratique ; et enfin la question cruciale des sociétés plus autonomes par rapport au tourbillon techno-chrématistique, et qui seront dans l'avenir autant de sources d'emplois plus stables, d'activités moins gaspilleuses d'énergie et moins polluantes, et aussi de conversations politique plus proches des citoyens. » (Août 2007)

Le ST le plus spécifique fait partie d'un passage qui fait appel à une révolution culturelle généralisée. Celle-ci se charge de poser les questions sociétales et civilisationnelles les plus urgentes et de promouvoir les alternatives-solutions. La révolution est celle de la culture scientifique. Est urgente une refonte de la pensée dominante et unique dans tous les domaines. Tout est à réinventer : des théories de référence pour une société-monde autre que la mondialisation, des théories économiques au service des sociétés et des hommes, d'autres technologies bioéthiques qui respectent la nature, ceci pour rester fidèle à la culture démocratique. Ce passage donne une idée sur la couleur sémantique de cette classe exclusive de la dernière période et qui échappe au sens global. D'une manière générale, les contenus socio-historiques connaissent un tournant qualitatif au cours du temps. Sur les dendrogrammes (Fig. 3, 4) on identifie leur emplacement libre entre les thèmes politiques et les thèmes économiques d'une période à l'autre. Dans les deux premières périodes, les questionnements sur l'histoire et la condition

de l'Homme sont mobilisés par la situation politique, tandis que les contenus économiques régressifs des deux dernières périodes attirent les thèmes socio-historiques.

5. Conclusion

Rapporter la structure sémantique des *sous-maquettes* à la dynamique des contenus dominants nous éclaire sur la formation du sens global et sur sa logique. Autrement dit, la dynamisation du sens global par la projection des classes lexicales sur la chronologie constitue un niveau intermédiaire entre le niveau des *sous-maquettes*, celui des phases sémantiques stables et de leurs sens particuliers d'un côté et le niveau synthétique du sens qui finalement se stabilise au niveau global après l'accumulation des sens particuliers. Ce qu'on voulait illustrer ici c'est ponctuellement l'intérêt du recours à une *maquette*, réduction raisonnée du corpus à ses noms les plus fréquents, modèle à échelle réduite repris dans les sous-corpus de stabilité sémantique. Cet usage couplé à une statistique cooccurrentielle ciblant les réseaux lexicaux structurants permet un accès rapide aux *fonds sémantiques*, condition première pour pratiquer une sémantique de corpus. La *maquette* balise une sémantique de corpus qui va *du global au local* (Rastier 2001). Plus concrètement, si la cooccurrence est l'interprétant minimal saisi au sein du passage (Rastier 2007), on lui a assigné la mission de mesurer le *temps sémantique* pour déterminer les phases de stabilité sémantique où l'on peut observer les mondes lexicaux stabilisés (Reinert 1993, 2008). Ceux-ci sont les interprétants maximaux objectivables au niveau de la *maquette* et des *sous-maquettes*. La *maquette* telle qu'on la conçoit ne renvoie pas à un modèle généralisable mais à un usage généralisable. Un usage qui pour chaque corpus contribue à la reconstitution de son modèle sémantique quelle que soit sa spécificité et à réaliser la vocation de sa conception. Ici, dans le cas des corpus chronologiques, la *maquette* réconcilie l'étude du sens et l'étude du temps. Tandis que la première passe par délinéarisation du texte et la capture de la structure non-séquentielle du texte, la seconde poursuit l'organisation séquentielle des textes. La *maquette* en tant que dispositif destiné à un usage prédéfini intègre l'étude du non-séquentiel dans le séquentiel et efface le faux contraste entre eux.

Références

Brunet E. (2008). Les séquences (suite). *JADT 2008*.

Brunet E. (2012). Nouveau traitement des cooccurrences dans Hyperbase. *Corpus (11)*.

Halliday M. A. (1994). *Introduction to Functional Grammar*. London : Edward Arnold.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris : Dunod.

Mayaffre D. (2008a). Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la cooccurrence. *JADT 2008*.

Mayaffre D. (2008b). De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. *Sémantique & synatxe (9)*.

Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse des Données co(n)textuelles. Parcours coocurrentiels dans le discours présidentiel français (1958-2014). *JADT 2014*.

Metwally H. (2017), *Les thèmes et le temps dans Le Monde diplomatique (1990-2008)*. Thèse de doctorat, Université Côté d'Azur.

Rastier F. (2001). *Arts et sciences du texte*. PUF.

Rastier F. (2007). Passages. *Corpus (6),* pp. 25-54.

Rastier F. (2011). *La mesure et le grain*. *Sémantique de corpus.* Paris : Champion.

Ratinaud P. et Marchand P. (2012). Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRAMUTEQ. *JADT 2012*.

Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données. 8(2),* pp. 187-198.

Reinert M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société* (66), pp. 5-39.

Salem A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots (17).* pp. 105-143.

Salem A. (1991). Les séries textuelles chronologiques. *Histoire & Mesure, VI (1/2)*. pp. 149-175.

Salem A. (1993). De travailleurs à salariés. Repères pour une évolution du vocabulaire syndical (1970-1993). *Mots(63)*. pp. 74-83.

Salem A. (1994). La lexicométrie chronologique. Dans *Actes du colloque de lexicologie politique 'Langages de la Révolution'*. Paris : Klincksieck.

Viprey J.-M. (2005). Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus. Dans A. Condamines, *Sémantique et corpus*. Paris : Lavoisier.

Viprey J.-M. (2006). Structure non-séquentielle des textes. *Langages* (183).

Séries textuelles homogènes

Jun Miao [1], André Salem [2]
[1] Université Lumière de Lyon 2, France – miaojun@miaojun.net
[2] Université de la Sorbonne nouvelle - Paris 3, France – salem@msh-paris.fr

Abstract
Textometric methods, widely used for the study of large corpora, are applied here to a set of small texts, which, however, present homogeneous characteristics. Our study focuses on a chronological textual series consisting of reports of successive congresses of the CCP (Chinese Communist Party) during the period 1982-2017. The textometrical methods are firstly used to highlight the changes occurred during the 2017 congress.
Secondly, we apply these same methods to the subcorpora consisting of a collection of fragments, automatically extracted from each congress and related to the same *topic*. This subcorpora thereby constituted make it possible to observe, with greater efficiency, the contextual variations that occur over time around the same *type*. The method can be extended to any corpora consisting of fragment systems that present a certain level of homogeneity among them.

Keywords: Textual series, Chinese political speeches, homogeneous subcorpora

Résumé
Nous appliquons ici des méthodes textométriques, largement utilisées pour l'étude de vastes corpus, à des ensembles de textes dont la taille est réduite mais qui présentent de fortes caractéristiques d'homogénéité. Notre étude porte sur une série textuelle chronologique constituée par les rapports successifs des congrès du PCC (Parti Communiste Chinois) durant les années 1982-2017.
Les méthodes de la *veille textuelle textométrique* sont d'abord mises en œuvre pour mettre en évidence les changements survenus lors du congrès de 2017. Dans un deuxième temps, nous appliquons ces mêmes méthodes à des sous-corpus, constitués par la réunion de fragments extraits de chacun des congrès et relatifs à un même *thème*. Les sous-corpus ainsi constitués permettent d'observer avec une efficacité accrue des variations contextuelles qui surviennent au fil du temps autour d'une même forme-pôle. La méthode peut être appliquée à tout corpus constitué de systèmes de fragments présentant une certaine homogénéité entre eux.

Mots-clés: Séries textuelles, discours politique chinois, sous-corpus homogène.

1. Introduction[1]

Le développement des capacités textométriques permet désormais d'explorer avec profit des ensembles de textes extrêmement vastes et souvent variés. Nous avons, cependant, insisté, avec d'autres, sur l'intérêt qu'il y a à appliquer ces mêmes méthodes à des corpus constitués par la réunion de productions textuelles présentant de fortes caractéristiques d'homogénéité et forcément plus réduites de ce fait (Salem 1991). Au delà des séries chronologiques, auxquelles nous empruntons nos exemples, la démarche que nous présentons peut être appliquée à différents types de corpus.

Depuis quelques décennies, le Congrès national du Parti communiste chinois (PCC) a lieu une fois tous les cinq ans. Il constitue la plus haute instance de ce Parti, dans laquelle sont annoncées les décisions importantes[2]. Dans la dernière décennie, les commentaires et les analyses quantitatives, portant sur les textes de congrès du PCC, plus ou moins appuyés sur des méthodes d'analyse statistiques, se sont multipliés dans la presse et sur différents sites de l'Internet.

Le corpus que nous étudions est constitué d'un ensemble des textes produits lors des congrès du PCC, entre 1982 et 2017. Pour des raisons que nous analysons, les textes produits durant cette dernière période présentent une grande homogénéité, tant du point de vue de leur taille que de celui des thèmes qu'ils abordent et du style qu'ils emploient. Nous commençons par étudier de manière classique la série chronologique ***PCC1982-2017*** divisée en ***congrès*** afin de mettre en évidence des variations dans l'emploi du vocabulaire. Nous proposerons ensuite une méthode qui permet, selon nous, d'étudier au plus près les variations du contexte immédiat d'un terme donné.

2. Analyse chronologique de la série PCC1982-2017

Le corpus ainsi constitué compte au total 115 1338 occurrences pour 7365

---

[1] Les analyses dont nous rendons compte ci-dessous, ont été effectuées à l'aide du logiciel *Lexico5*. Cedric Lamalle, William Martinez, Serge Fleury ont largement contribué au développement des fonctionnalités de ce logiciel. Les auteurs tiennent à les en remercier.

[2] L'article de Salem et Wu (2008) constitue une étude chronologique portant sur l'intégralité des congrès du PCC survenus depuis sa fondation 1921 jusqu'à l'année 2012. Au-delà des évolutions chronologiques qu'elle avait permis de mettre à jour, cette étude montre le caractère hétérogène de la forme ***congrès*** considérée sur une échelle aussi large.

formes différentes[3]. La division en congrès amène une partition du corpus en huit parties. Les longueurs des parties, pour chaque congrès, s'échelonnent entre 2 400 et 2 900 occurrences. La forme de fréquence maximale est toujours la forme**的** (de, DE1), dont on peut vérifier la forte diminution au fil des congrès[4].

2.1 Le congrès 2017

Lorsque survient un nouveau congrès qui complète une série chronologique pré-existante, la *méthode des spécificités* permet de répondre à la question : *Quelles sont les principales évolutions lexicales survenues lors du dernier congrès de la série* ? C'est une opération de veille lexicale. Le calcul des spécificités appliqué au congrès de 2017 signale des spécificités positives, dont le contenu revêt un caractère nettement *lexical* : 时代 (shídài, *ère, S +24*), 治理 (zhìlǐ, *gérer, S +21*), 生态 (shēngtài, *écologie, S +15*), 梦 (mèng, *rêve, S +14*)[5]. A l'inverse, les formes de spécificités négatives, pour cette même période, sont plutôt des *formes grammaticales*, telles que 的 (de, DE1, *S -38*) , 这 (zhe, *ce, S -22*), 地 (de, DE2, *S -14*).

Le même calcul appliqué aux segments répétés du corpus permet de préciser les modifications survenues lors ce même congrès. La mise en vedette du terme 新 时代 (xīn shídài, *nouvelle ère*), employé 36 fois lors du congrès de 2017, a été largement commentée par les analystes qui se sont penchés sur ce texte[6]. Le recensement systématique des *segments* fortement spécifiques pour cette même période permet de mettre en évidence des séquences répétées dont certaines ont pu échapper aux commentateurs et qui constituent également des néologismes par rapport aux congrès précédents : 新 时代

[3] La séquence textuelle continue des textes chinois, composés de caractères juxtaposés (*scriptio continua,* dans laquelle les mots ne sont pas séparés par des espaces), a été soumise à un segmenteur automatique NLPIR (Zhang, 2016), très largement utilisé dans le monde sinophone, afin d'être segmentée en *mots graphiques*.

[4] Nous expliquons dans une étude parallèle comment cette diminution progressive peut être mise en rapport avec l'évolution du style d'écriture.

[5] Dans nos exemples, la forme *native* chinoise est suivie de sa transcription en *pinyin*, puis d'un équivalent français (lequel ne peut prétendre au statut de traduction satisfaisante pour chacune des occurrences du terme). Un *coefficient de spécificité*, positive ou négative, de forme S +/- xx indique enfin le degré de spécificité de la forme dans la partie du texte considérée.

[6] De nombreux articles publiés à cette occasion ont explicitement mentionné la fréquence (36 occurrences) de la formule**新 时代** (xīn shídài, *nouvelle ère*) ex : Vandnepitte (2017). D'autres sites ont proposé aux internautes de classer congrès par fréquence d'apparition de plusieurs termes répétés dans chaque congrès (Qian, 2017).

中国 特色 社会主义, (*le socialisme à la chinoise dans la nouvelle ère*, 13 occ., *S +12*) , **治理 体系**. (*le système de gouvernance*, 13 occ., *S +12*). Plus remarquable à nos yeux, certaines expressions extrêmement courantes dans les périodes précédentes ont complètement disparu du texte du dernier congrès. Tel est le cas, par exemple pour des segments comme :

有 中国 特色, (*posséder des caractéristiques chinoises*, 0 occ., *S -7*),

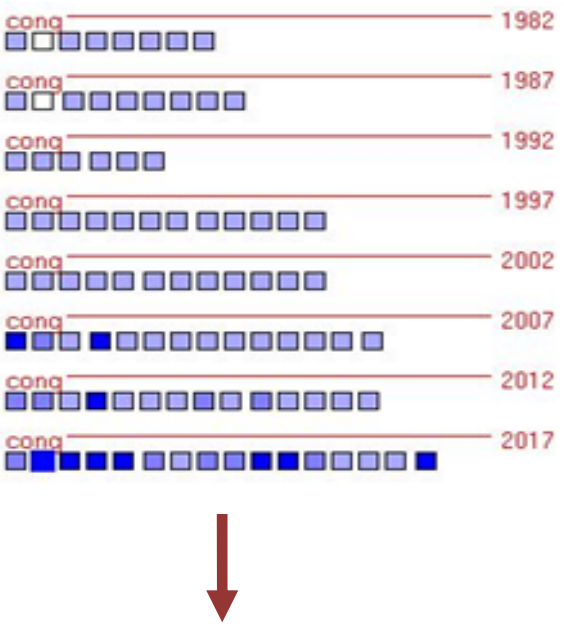
有 中国 特色 社会主义 (*avoir un socialisme à la chinoise*, 0 occ., *S -5*).

L'analyse des spécificités permet également de localiser des parties du texte dans lesquelles le renouvellement lexical se révèle particulièrement important. Sur la figure 1, une carte des sections a été établie pour chacun des congrès, divisé en chapitres. Les sections apparaissent d'autant plus sombres qu'elles renferment de nombreuses occurrences de termes spécifiques pour le dernier congrès. La représentation permet de vérifier que le renouvellement ne se fait pas de manière uniforme, dans le dernier congrès. Une partie du vocabulaire spécifique du congrès de 2017, était déjà largement présente dans les deux congrès précédents. La carte permet en outre de localiser précisément les chapitres du dernier congrès qui font le plus fortement l'objet d'un renouvellement lexical.

La figure 2 ci-dessous permet d'apprécier l'évolution du vocabulaire survenue dans la dernière période en combinant une représentation factorielle sur l'ensemble des congrès et les spécificités calculées pour le dernier congrès. Une analyse réalisée sur les huit congrès met en évidence la progressivité des changements lexicaux. On a projeté en qualité d'éléments supplémentaires les formes *spécifiques positives* de la dernière partie. Ce type de représentation peut être articulé avec les cartes de section, présentées ci-dessus pour illustrer les changements lexicaux.

3. Utiliser la structure des documents

Dans chacun des textes de l'édition originale des congrès, des repères éditoriaux (intertitres, numérotation de sous-parties, etc.) permettent d'effectuer un découpage en unités plus petites que nous appellerons *chapitres*. Chaque chapitre correspond à l'évocation d'un thème particulier (développement économique, perspectives internationales, état des forces armées, etc.). Lors de chacun des congrès, ces thèmes sont abordés tour à tour, souvent dans un ordre similaire qui peut conduire à proposer une description globale de l'ordonnancement de ces textes de congrès.



Guide de lecture:
A gauche, on trouve une *carte des sections* réalisée à partir d'un découpage en *chapitres*. Chaque ligne regroupe les chapitres relatifs à un même congrès.
Les carrés les plus foncés correspondent aux chapitres les plus chargés en formes spécifiques dans le dernier congrès (S+>10).
En bas : le texte du deuxième chapitre du dernier congrès, qui figure au dessous de la carte est signalé comme *particulièrement chargé en formes spécifiques*.

$<chap1=19-00># <para=19-05>同志 们 : ¶ # <para=19-06>现在 , 我 代表 第十八 届 中央 委员会 向 大会 作 报告 . ¶ # <para=19-07>中国共产党 第十九 次 全国 代表大会 , 是 在 全面 建成 小康 社会 决胜 阶段 * 中国 特色 社会主义 进入 新 时代 的 关键 时期 召开 的 一 次 十分 重要 的 大会 . ¶ # <para=19-08>大会 的 主题 是 : 不 忘 初心 , 牢记 使命 , 高举 中国 特色 社会主义 伟大 旗帜 , 决胜 全面 建成 小康 社会 , 夺取 新 时代 中国 特色 社会主义 伟大 胜利 , 为 实现 中华民族 伟大 复兴 的 中国 梦 不懈 奋斗 /... /. ¶.

Figure 1 : Repérage des portions caractéristiques pour le dernier congrès (2017)

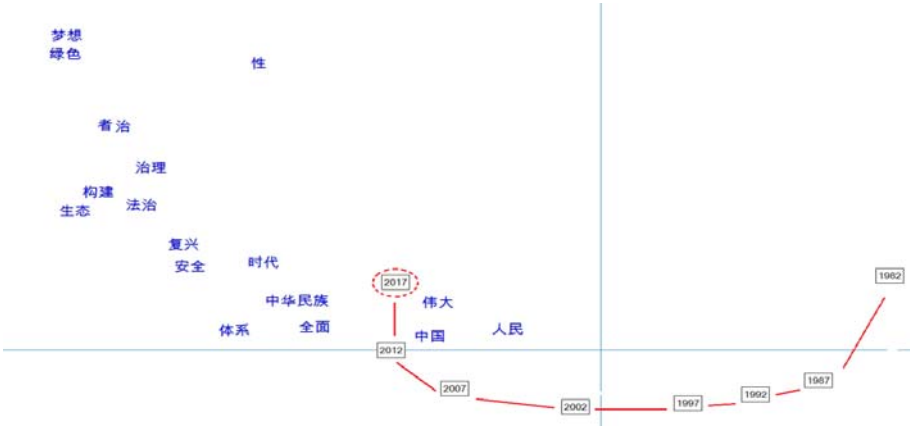
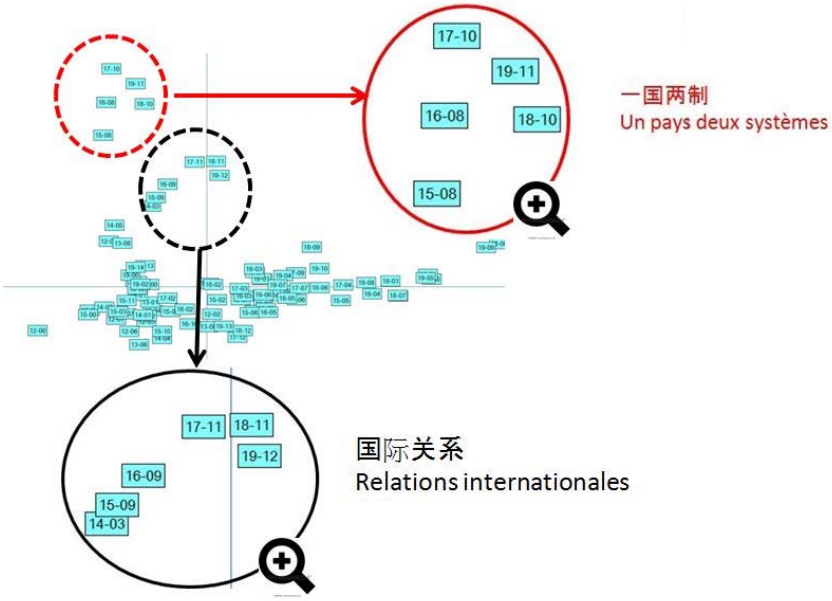


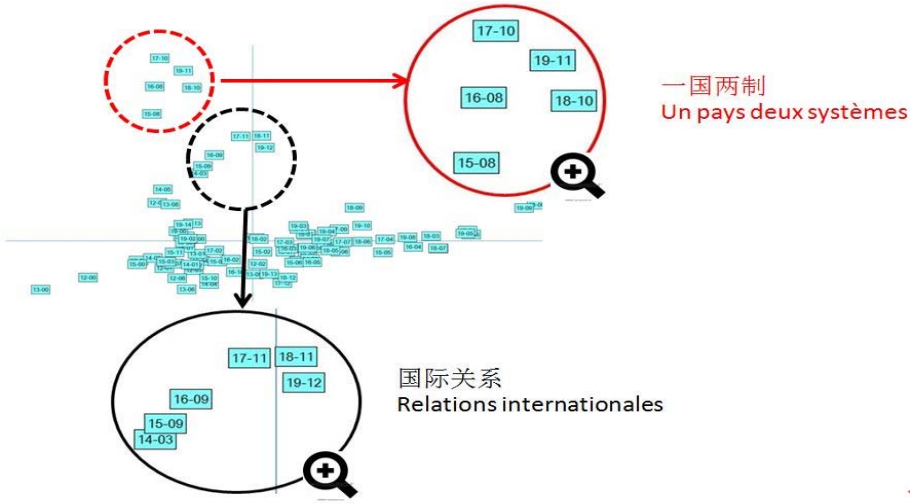
Figure 2 : Spécificités positives du congrès 2017 mises en évidence dans l'AFC

Guide de lecture: Sur la figure 2, les différents congrès s'échelonnent dans le temps selon une parabole. Cet échelonnement résulte d'un renouvellement important du vocabulaire au fil des congrès. Les formes les plus spécifiques pour le dernier congrès ont été projetées en qualité d'éléments supplémentaires.



3.1 Analyse en chapitres

Lorsqu'on soumet à des analyses typologiques, le même corpus divisé, cette fois, en chapitres, on constate que les chapitres correspondant aux mêmes thèmes, mais appartenant à différents congrès, ont une forte tendance à se regrouper, du fait qu'ils emploient des vocabulaires proches. La structure chronologique, mise en évidence par l'analyse en congrès s'efface, dans ce cas, devant une typologie d'ordre *thématique*. La figure 3 montre les résultats d'une Analyse factorielle des correspondances effectuée à partir du corpus ***PCC1982-2017*** divisé cette fois en 89 chapitres. Sur cette figure, les identificateurs des chapitres sont constituées de deux parties**.** Le premier nombre indique le numéro du congrès dont le chapitre est extrait. Le second, l'ordre du chapitre à l'intérieur du congrès. Comme on peut le vérifier sur cette figure les chapitres correspondant à un même thème ont tendance à se regrouper fortement.



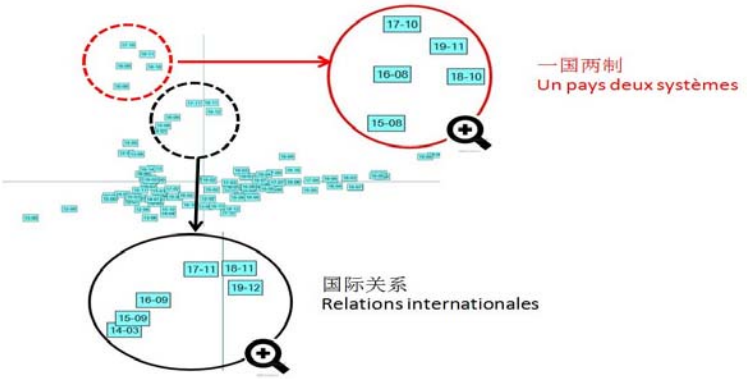


Figure 3 : Analyse factorielle des correspondances sur le corpus divisé en chapitres

A titre d'exemple, nous avons agrandi les portions du graphique qui correspondent à deux groupes thématiques :

a) le groupe *un pays deux systèmes* qui correspond à une orientation politique constante du PCC, réaffirmée à chaque congrès ;
b) un groupe de chapitres correspondant à l'analyse des *relations internationales*, qui constitue également un moment incontournable pour chaque congrès, à partir du 14ème.

3.2 Le sous-corpus thématique « un pays deux systèmes »

L'étape suivante consiste à réitérer ces mêmes analyses à partir de sous-corpus réduits, rassemblant les seules chapitres relatifs à une même thématique. Les analyses textométriques effectuées sur ces *sous-corpus homogènes* débouchent sur des résultats particulièrement lisibles. Lors de l'analyse de ce type de corpus, la dimension chronologique revient au premier plan. Le sous-corpus qui rassemble les passages relatifs au thème *un pays, deux systèmes* ne compte que deux milles occurrences, sur l'ensemble des congrès. L'analyse des formes qui apparaissent spécifiquement dans les contextes de ce terme, montre cependant une nette évolution du contexte immédiat de ce terme. Le Congrès de 1987, présente la formule comme un *principe à mettre en œuvre*. Dans les congrès suivants, on voit apparaître les verbes *maintenir* et *continuer* (2002) puis *mettre en œuvre sans faille* (2007). En 2017, il s'agit d'appliquer *intégralement et avec précision le principe un pays, deux systèmes*. La figure 4 montre une projection des différents segments qui contiennent l'expression sur l'analyse réalisée à partir du sous-corpus[7].

[7] Le graphique a été légèrement modifié pour permettre une plus grande lisibilité. Les segments redondants ont été écartés; les points superposés ont été légèrement déplacés.

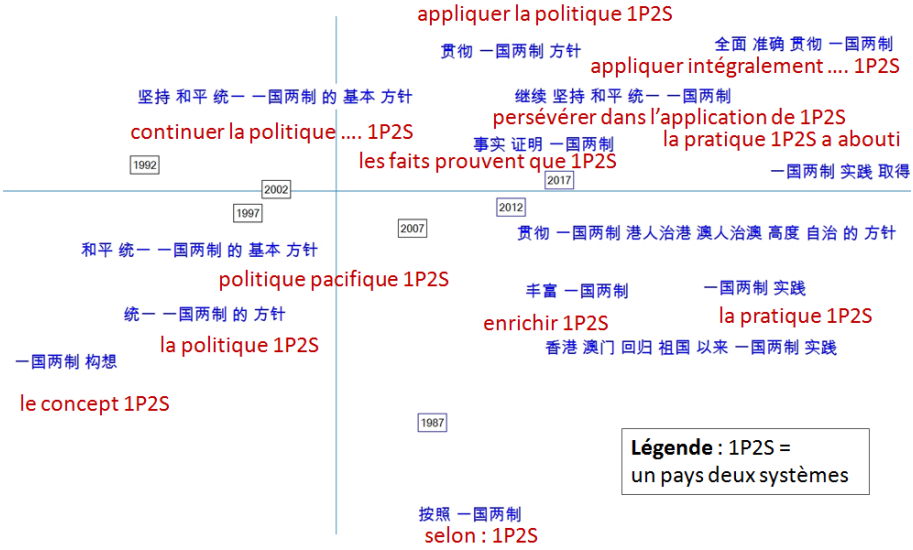


Figure 4 : Variations lexicales autour de l'expression : un pays deux systèmes

4. Conclusion

Nos expériences nous amènent à conclure que l'analyse textométrique opérée à partir de regroupements de fragments homogènes, prélevés autour d'un même *thème* durant les années couvertes par une série chronologique conduit à des résultats dont l'interprétation se révèle particulièrement aisée. La grande homogénéité lexicale des fragments rapprochés permet alors d'observer des variations très fines. Elle compense largement la taille réduite du corpus, peu favorable, a priori, dans le cas d'études textométriques.

Au delà des applications aux seules *séries textuelles chronologiques*, la méthode pourra être utilisée pour toute sorte de corpus, dans une large variété de langues, à la condition qu'il soit possible de distinguer des sous ensembles thématiques homogènes

Références

Miao J. (2012). *Approches textométriques de la notion de style du traducteur - Analyses d'un corpus parallèle français-chinois : Jean-Christophe de Romain Rolland et ses trois traductions chinoises*. Thèse doctorale dirigée sous la direction de M. André Salem, Paris 3.

Qian G. (2017). 中共历届党代大会报告语象分析 (Analyses lexicales des rapports de tous les congrès du Parti communiste chinois).*Lianhe Zaobao* du 19 novembre 2017.

Salem A. (1991). Les séries textuelles chronologiques. *Histoire & Mesure Année*, Vol. (6) : 149-175.

Salem A., Wu Li-Chi. (2008). Essai de textométrie politique chinoise. In André Salem et Serge Fleury, éditeurs, *Lexicometrica – Explorations textométriques,* Vol. (1). URL : http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm (consulté le 5 février 2017).

Vandepitte M. (2017). Quatre choses à savoir sur la Chine – dans le cadre du XIXème congrès du Parti. Traduit par Anne Meert en français du néerlandais. *Investig'Action* du 15 novembre 2017. URL : goo.gl/8fgSkq (consulté le 25 novembre 2017).

Logiciels utilisés :

Zhang H.P. (2017). Segmenteur automatique chinois NLPIR. URL : http://www.nlpir.org/

Salem A. (2017). L'outil d'analyse textométrique *Lexico 5.* URL : http://www.lexi-co.com/index.html

TaLTaC in ENEAGRID Infrastructure

Silvio Migliori[1], Andrea Quintiliani[1], Daniela Alderuccio[1],
Fiorenzo Ambrosino[1], Antonio Colavincenzo[1], Marialuisa Mongelli[1],
Samuele Pierattini[1], Giovanni Ponti[1] Sergio Bolasco[2],
Francesco Baiocchi[3], Giovanni De Gasperis[4],
[1]ENEA DTE-ICT – silvio.migliori@enea.it, [2] Sapienza Università di Roma,
[3] Staff TaLTac - info@taltac.it, [4] Dip. DISIM Università dell'Aquila

Abstract
The aim of this joint ENEA-TaLTaC project is to enable the TaLTaC User Community and the Digital Humanists to have remote access to the TaLTaC software through the ENEAGRID Infrastructure. ENEA's research activities on the integration of Language Technologies (Multilingual Text Mining Software and Lexical Resources) in the ENEA distributed digital infrastructure provide a "community Cloud" approach in a digital collaborative environment and on an integrated platform of tools and digital resources, for the sharing of knowledge and analysis of textual corpora in Economic and Social Sciences and e-Humanities. Access to the TaLTac software in Windows and Linux version will exploit the high computational capacity (800 Teraflops) of the e-infrastructure, to which users access as a single virtual supercomputer.

Riassunto
Obiettivo del progetto congiunto ENEA-TaLTaC è consentire alla comunità degli utenti TaLTaC e ai ricercatori nelle Digital Humanities l'accesso remoto al software TaLTaC attraverso l'infrastruttura digitale ENEAGRID. Le attività di ricerca dell'ENEA sull'integrazione delle tecnologie linguistiche (software di Text Mining per testi multilingue e risorse lessicali) in ENEAGRID forniscono un approccio "community Cloud" in un ambiente collaborativo digitale e su una piattaforma integrata di strumenti e risorse digitali, per la condivisione delle conoscenze e l'analisi di corpora testuali in Scienze Economiche e Sociali ed e-Humanities. L'accesso al software TaLTac in versione Windows e Linux sfrutterà l'elevata capacità computazionale (800 Teraflops) dell'infrastruttura di calcolo, a cui gli utenti accedono come ad un unico supercomputer virtuale.

Keywords: Text Mining Software, Cloud Computing, Digital-Humanities, Socio-Economic Sciences, Big Data.

1. Introduction

"TaLTaC in CLOUD" is a joint ENEA-TaLTaC project for the set-up of an ICT portal on the ENEA distributed e-Infrastructure[1] (Ponti *et al*., 2014), hosting TaLTaC Software (Bolasco *et al*., 2016, 2017). Users will access TaLTaC software (Windows and Linux versions) in a remote and ubiquitous way, and the computational power (800 Teraflops) of ICT ENEA distributed resources, as a single supercomputer. The aim of this joint ENEA-TaLTaC project is to enable the TaLTaC User Community and Digital Humanists to have remote access to TaLTaC software through ENEAGRID Infrastructure, integrating ICT inside Digital Cultural Research.

ENEAGRID offers a digital collaborative environment and an integrated platform of tools and resources assisting research collaborations, for sharing knowledge and digital resources and for storing textual data. In this virtual environment, TaLTaC software evolves from a stand-alone uniprocessor software toward a multiprocessor design, integrated in an ICT research e-infrastructure. Furthermore, it evolves towards implementing ancient language lexical and semantic knowledge and e-resources, facing research needs and implementing solutions also for Digital Humanities communities.

2. TaLTaC Software

The TaLTaC software package, conceived at the beginning of the 2000s, has been progressively developed to date in three major releases: T1 (2001), T2 (2005) and T3 (2016); it is widespread among the text analysis community in Italy and abroad with over 1000 licenses, including two hundred entities between university departments, research institutions and other organizations.

The 2018 release of the software, T3, implemented the following priority objectives: *i*) the processing of big data (around of a billion words), achieving the independence from the dimensions of the text corpora, limited only by hardware resources; *ii*) the automatic extraction on multiple layers of results from text parsing (tokenization): layer zero (text in the original version), layer 1 (recognition of words with automatic corrections of the accents), layer 2 (pre-recognition of most common Named Entities), layer 3 (reconstruction of pre-defined multiwords); *iii)* computing speed, taking advantage of the power of the multi-core processing readily available on current computers

[1] The ENEAGRID infrastructure is based on several software components which interact with each other to offer an integrated distributed system. The ENEAGRID infrastructure allows access to all these resources as a single virtual system, with an integrated computational availability of about 16000 cores, provided by several multiplatform systems.

(personal or cloud).
Table 1 shows the processing times of three parsing, up to layer 2, for larger corpora on PC (1-core and 8-cores) and on ENEAGRID. Preliminary results on ENEAGRID (1core-CRESCO) show that with increasing corpus size there is an even greater saving of time.
TALTAC was installed in ENEAGRID infrastructure, but the computational capabilities of the HPC system are not yet exploited because the current version of the software does not support multi-core. Therefore, the present ENEAGRID capabilities allow only multi-users access and computation; future versions of the software will be tested for multi-core capabilities to exploit the real power of ENEA ICT High Performance Computing.

Table 1. Preliminary results of processing times of three parsing on PC and on ENEAGRID.

				MAC i7 (7th generation)			ENEAGRID
		tokens	size of file	1 core	8 cores	8core /1core	1 core (CRESCO)
		millions	*GB*	*in minutes*		*in %*	*in minutes*
1	"*La Repubblica*" (100 th Artic.)	74	0,41	3,4	1,1	0,33	3,5
2	"*La Repubblica*" (400 th Artic.)	284	1,55	13,0	3,8	0,29	13,2
3	Italian and French Press	535	2,89	37,4	8,8	0,24	41,3
4	Various Press Collection	1.138	6,18	88,2	14,0	0,16	54,7

For the characteristics of the technological architecture of the TaLTaC3 platform, see previous works (Bolasco *et al*. 2016, 2017), that can be summarized here as: a1) *HTML 5 for the GUI* and *jQuery* with its derived Javascript frameworks to encapsulate the GUI user interaction functions for the MAC and Cloud solution; a2) Windows native *DotNET* desktop application; b) *JSON* (*JavaScript Object Notation*): as an inter-module language standard, with a structured and agile format for data exchange in client/server applications; c) *Python* / *PyPy*: advanced script/compiled programming language, mostly used for textual data analysis and natural language processing at the CORE back end; d) *No-SQL*: high performance key/value data structure storage server *Redis* adopted for vocabularies/linguistic resources persistence; e) *RESTful*: interface standard for data exchange over the HTTP web protocol; f) *MULTI-PROCESSING*: exploiting in the best possible way multi-core hardware, distributing processing power among different CPU cores.
The choice of the Python language allowed to develop a cross-platform computational core running on Windows, Linux, macOS. In particular, the overall system of software processes runs smoothly over a linux-based cloud computing facility, like the ENEAGRID. Furthermore, the Python code

compiled through the 64bit *PyPy* just-in-time-compiler allows very efficient macro operations over a large set of data, stored as hash dictionaries, so that the upper limits of performance and capacity is only due to the physical limit of the host machine, in terms of RAM and number of cores and OS kernel scheduler. In our test each node in the ENEAGRID infrastructure hosted a single Redis instance and a number of 24 logic cores, with 16GB of RAM.

3. ENEAGRID Infrastructure

ENEA activities are supported by its ICT infrastructure, providing advanced services as High Performance Computing (HPC), Cloud and Big Data services, communication and collaboration tools. Advanced ICT services are based on ENEA research and development activities in the domains of HPC, of high performance networking and data management, including the integration of large experimental facilities, with a special attention to public services and industrial applications. As far as High Performance Computing is concerned, ENEA manages and develops ENEAGRID, a computing infrastructure distributed over 6 ENEA research centers for a total of about 16000 cores and a peak computing power of 800 Tflops.

HPC clusters are mostly based on conventional Intel Xeon cpu with the addition of some accelerated systems as Intel Xeon/PHI and Nvidia GPU. Storage resources includes RAID systems for a total of 1.8 PB, in SAN/Switched and SRP/Infiniband configuration. Data are made available by distributed and high performances files systems (AFS and GPFS).

ENEA Portici Center has become one of the most important italian HPC center in 2008 with the project CRESCO - Computational RESearch Center for COmplex Systems. CRESCO HPC clusters are used in many of the main ENEA research and developments activities, such as energy, atmosphere and sea modeling, bioinformatics, material science, critical infrastructures analysis, fission and fusion nuclear science and technology, complex systems simulation. CRESCO clusters have provided in 2015 and 2016 more than 40 million core hours each year to ENEA researchers and technologists and to their external partners (external users account for about 30% of the total machine time).

CRESCO6, the new HPC cluster recently installed in Portici in the framework of the 2015 ENEA-CINECA agreement, provides a peak computing power of 700 Tflops and is based on the new 24 cores Intel SkyLake cpu. Its nodes will be connected by the new Intel OmniPath high performance network, providing a 100 Gbps bandwidth.

ENEA ICT department provides also general purpose communication, elaboration and collaboration tools and services as Network management, E-Mail, Video Conferencing and Voip services, Cloud Computing and Storage.

A friendly user access to scientific and technical applications (as Ansys, Comsol, Nastran, Fluent) is provided by dedicated web portals (Virtual laboratories) relying on optimized remote data access tools as NX technology.

4. TaLTaC in ENEAGRID Infrastructure

4.1 Software Installation and Access on ENEA e-Infrastructure

The software TaLTaC is available on Windows and Linux through ENEAGRID via AFS in a geographically distributed file system, which allows remote access to each computing node of the HPC CRESCO systems and Cloud infrastructure from anywhere in the world.

This provides three capabilities: i) data mining, sharing and storage; ii) ICT services necessary for the efficient use of HPC resources, collaborative work, visualization and data analysis; iii) the implementation of software and its settings for future data processing and analysis. Moreover, the availability of the software on the ENEA ICT infrastructure can benefit of the advantages of AFS such as scalability, redundance, backup and so on.

Through the ACL rules it can be possible to manage the accessibility of the software to the community of users in compliance of the license policies that will be put in place. The following two options are provided for TaLTaC running: the first one is to use the applications installed in the windows system and the second one is to use FARO2 – Fast Access to Remote Objects (the general purpose interface for hardware and software capabilities by web access) to directly access the applications installed in the Linux environment and that refer to the data in AFS.

4.1.1. TaLTaC2 (Windows) on Remote Desktop Access

The software TaLTaC2 is available on "Windows Server 2012 R2" by remote desktop access to a virtual machine that can be reached by the ThinLinc general-purpose and intuitive interface. All the users involved in the project activities can access the server but only the person in charge of developing and installing the application can obtain administrator privileges. For this reason, AFS authentication is always required. Every TaLTaC2 user with AFS credentials can access ENEAGRID to run the software and to manage data on AFS own areas via web and from any remote location. In the AFS environment, an assigned disk area with a large memory capacity is defined. This area is mainly used for storage and sharing of large amounts of data (less than 200 MB) (analysis, reports and documents) that come from running the software on a single processor, in serial mode, or for future parallel data mining applications.

4.1.2. TaLTaC3 (Linux) on CRESCO System

On the CRESCO systems, that is accessible from ENEAGRID infrastructure, TaLTaC3 is available on CentOS Linux nodes and then it is possible to leverage the overall computing power dedicated to the activities of TaLTaC and Digital Humanists communities. Every user can start own work session allocating a node with a reserved *Redis* instance and as many computing core as needed.

Performance improvements are obtainable through the parallelization so that a single user can use the full capacity of the assigned node, in terms of number of computing cores. The TaLTaC3 package is automatically started as the user login to the node by a shell script. The opensource Mozilla Firefox web browser makes the user interface in the current beta version. The access to the TaLTaC3 portal use the ThinLinc remote desktop visualization technology that allows an almost transparent remote session on the HPC system, including the graphical user interface, thanks to the built-in features such as load-balancing, accelerated graphics and platform-specific optimisations. Input and output data can be accessed through the ENEAGRID filesystems and therefore easily uploaded and downloaded.

4.2 Case Studies

ENEA distributed infrastructure (and cloud services) enables the management of research process in Economic-Social Sciences and Digital Humanities, providing technology solutions and tools to academic departments and research institutes: building and analyzing collections to generate new intellectual products or cultural patterns, data or research processes, building teaching resources, enabling collaborative working and interdisciplinary knowledge transfer.

4.2.1. TaLTaC User Community

The current (2018) community of TaLTaC over the years aggregated users from the computer laboratories of automatic analysis of texts and text mining, also carried out within the institutional courses of bachelor and magistral degrees, plus Ph.D. students from doctoral degree courses at the universities of Rome "La Sapienza" and "Tor Vergata", of Padua, Modena, Pisa, Naples and Calabria (a total estimate of over 1300 students over the last eight years); furthermore, there is another set of users that subscribed to specific tutorial courses dedicated to TaLTaC (more than 60 courses for a total number of 750 tutorial participants).

A call about the opportunity of using "remotely" the software via the ENEA distributed computing facilities, received the manifestation of interest by 40 departments and other research institutes.

4.2.2. Digital Humanities Community as TaLTaC user

In collaboration with academic experts, ENEA focused on Digital Humanities projects in Text Mining & Analysis in Ancient Writings Systems of the Near East and used TaLTaC2 to perform quantitative linguistic analysis in cuneiform corpora (transliterated into latin alphabet) (Ponti *et al*., 2017).

Cuneiform was used by a number of cultures in the ancient Near East to write 15 languages over 3,000 years. The cuneiform corpus was estimated to be larger than the corpus of Latin texts but only about 1/10 of the extant cuneiform texts have been read even once in modern times. This huge cuneiform corpus and the restricted number of experts lead to the use of Text Mining and Analysis, clustering algorithms, social network analysis in the TIGRIS Virtual Lab for Digital Assiriology[2], a virtual research environment implemented in ENEA research e-infrastructure. In TIGRIS V-Lab researchers perform basic tasks to extract knowledge from cuneiform corpora. (i.e. dictionaries extraction with word list of toponyms, chrononyms, theonyms, personal names, grammatical and semantic tagging, concordances, corpora annotation, lexicon building, grammar writing, etc.).

5. Conclusions

Researchers and their collaborators will use computational resources in ENEAGRID to perform their work regardless of the location of the specific machine or of the employed hardware/ software platform.

ENEAGRID offers computation and storage resources and services in a ubiquitous and remote way. It integrates a cloud computing environment and exports: a) remote software (i.e. TaLTaC); b) Virtual Labs: thematic areas accessible via web, where researchers can find set of software (and documentation regarding specific research areas); c) remote storage facilities (with OpenAFS file system). In this virtual environment, TaLTaC software evolves from a uniprocessor software toward a multiprocessor design, integrated in an ICT research e-infrastructure.

This project leads to the TaLTaC evolution from a stand-alone software (allowing Text Mining & Analysis to search for linguistic constructions in textual corpora, showing results in a table or concordance list) to a software "*always and anywhere on*", that also can be accessed, providing an interface where you can visualize results, create interpretative models, collaborate with others, combine different textual representations and storing data, co-developing research practices. Furthermore, this project reflects the shift

[2] TIGRIS - Toward Integration of e-tools in GRId Infrastructure for e-aSsyriology http://www.afs.enea.it/project/tigris/indexOpen.php - http://www.laboratorivirtuali.enea.it/it/prime-pagine/ctigris

from the individual-researcher-approach to a collaborative research community-approach, leading to a community-driven software design, tailor-made on specific research community needs and to Community Cloud Computing. This interdisciplinary knowledge transfer enables creating/activating new knowledge from big (cultural and socio-economic) data, both in modern and ancient languages.

References

Bolasco, S., Baiocchi, F., Canzonetti, A., De Gasperis, G. (2016). "TaLTaC3.0, un software multi-lessicale e uni-testuale ad architettura web", in D. Mayaffre, C. Poudat, L. Vanni, V. Magri, P. Follette (eds.), *Proceedings of JADT 2016*, CNRS University Nice Sophia Antipolis, Volume I, pp. 225-235.

Bolasco S., De Gasperis G. (2017). "TaLTaC 3.0 A Web Multilevel Platform for Textual Big Data in the Social Sciences" in C. Lauro, E. Amaturo, M.G. Grassia, B. Aragona, M. Marino. (eds.) *Data Science and Social Research - Epistemology, Methods, Technology and Applications* (series: Studies in Classification, Data Analysis, and Knowledge Organization) Springer Publ., pp. 97-103.

Ponti G., Palombi F., Abate D., Ambrosino F., Aprea G., Bastianelli T., Beone F., Bertini R., Bracco G., Caporicci M., Calosso B., Chinnici M., Colavincenzo A., Cucurullo A., Dangelo P., De Rosa M., De Michele P., Funel A., Furini G., Giammattei D., Giusepponi S., Guadagni R., Guarnieri G., Italiano A., Magagnino S., Mariano A., Mencuccini G., Mercuri C., Migliori S., Ornelli P., Pecoraro S., Perozziello A., Pierattini S., Podda S., Poggi F., Quintiliani A., Rocchi A., Sciò C., Simoni F., Vita A. (2014) "The Role of Medium Size Facilities in the HPC Ecosystem: The Case of the New CRESCO4 Cluster Integrated in the ENEAGRID Infrastructure". In: *Proceedings of the International Conference on High Performance Computing and Simulation, HPCS* (2014), ISBN: 978-1-4799-5160-4.

Ponti G., Alderuccio, D., Mencuccini, G., Rocchi, A., Migliori, S., Bracco, G., Negri Scafa, P. (2017) "Data Mining Tools and GRID Infrastructure for Text Analysis" in "Private and State in the Ancient Near East" Proceedings of the *58th Rencontre Assyriologique Internationale*, Leiden 16-20 July 2012, edited by R. De Boer and J.G. Dercksen, Eisensbrauns Inc. - LCCN 2017032823 (print) | LCCN 2017034599 (ebook) | ISBN 9781575067858 (ePDF) | ISBN 9781575067841.

ENEAGRID http://www.ict.enea.it/it/hpc -
Laboratori Virtuali http://www.ict.enea.it/it/laboratori-virtualixxx/virtual-labs
TIGRIS Virtual Lab http://www.afs.enea.it/project/tigris/indexOpen.php
TaLTaC: www.taltac.it

The dimensions of Gender in the International Review of Sociology. A lexicometric approach to the analysis of the publications in the last twenty years

Isabella Mingo, Mariella Nocenzi
Sapienza University of Rome – isabella.mingo@uniroma1.it; mariella.nocenzi@uniroma1.it

Abstract 1 (in English)
The Social Sciences and, specifically, the sociological research has progressively assumed the gender factor as one of the strategic keys to understand contemporary phenomena. In fact, as a variable for socio-statistical analysis or as a characterizing trait of individual identity, it is a decisive factor in the interpretation of the deep social transformations and it inspires the self-reflection of the sociologists about the analytical tools of their discipline. The contribution proposes, through a lexicometric approach, an analysis of the articles published in the last two decades by the oldest Journal of Sociology, published by Routledge. The main aim is to highlight the different ways in which gender issues are declined in the international sociological researches presented in the repertoire of the International Review of Sociology and to outline, both on the lexical level and on the topic level, the changes occurred over time.

Abstract 2 (in French, Italian or Spanish)
Le scienze sociali e, nello specifico, la ricerca sociologica hanno progressivamente assunto il fattore del genere come una delle più strategiche chiavi di lettura dei fenomeni contemporanei. Si tratta, infatti, di un fattore che, quale variabile per l'analisi socio-statistica o come tratto caratterizzante dell'identità individuale, si rivela dirimente nell'interpretazione delle profonde trasformazioni sociali in atto e spunto per un'autoriflessione degli stessi sociologi sugli strumenti di analisi della loro disciplina. Il contributo propone, mediante un approccio lessico-metrico, un'analisi degli articoli pubblicati nelle ultime due decadi dalla più antica rivista di sociologia, edita da Routledge, con l'obiettivo di evidenziare i diversi modi con cui il concetto di genere viene declinato nelle ricerche sociologiche internazionali presentate nel repertorio dell'International Review of Sociology e di delineare, sia sul piano lessicale che su quello delle tematiche, i cambiamenti intervenuti nel corso del tempo.
Keywords: Gender, International Review of Sociology, Lexicometric Analysis, Textual Analysis, Social Change, Sociological Analysis

1. Introduction and the hypothesis of the paper

From 1955, when in a relevant paper the American scholar John Money (et al., 1955) coined the term of *gender* for the definition of "those things that a person says or does to disclose himself or herself as having the status of boy or man, girl or woman", the social sciences have developed entire subfields and a wide range of topics to analyse it with a variety of research methods.

Sociologists, in particular, had outlined specific theoretical approaches and had led many detailed studies to understand firstly what gender is and the difference with sex. They had shared that if the meaning of *sex* is the *biological* classification based on body parts, *gender*, on the other hand, is the *social* classification based on one's identity, presentation of self, behavior, and interaction with others. Sociologists, hence, view *gender* as a learned behavior and a culturally produced identity, and, for these reasons, they define it as a "social" category. It has always been a very relevant category for the critical analysis of the social construction because one of the most important social structures is the status and one of the most strategic statuses is just gender.

In the last decades, the sociological theories and researches based on *gender* are become more and more widespread, articulated, integrated with other subfields of sociology and of the other social sciences. One of the most representative indicator of this research development and specialization is not only the common recognition and, then, institution of the sociology of the gender as a subfield of the sociology, but the most frequent use of *gender* as reference concept for all the other sociological theoretical approaches to the analysis of the social system. The same sociology of gender has studied many topics, with multiple research methods, including identity, social interaction, power and oppression, and the interaction with race, class, culture, religion, and sexuality, among others.

This paper aims to observe and, if possible, to interpret this progressive diffusion and specialization in the use of gender as a theoretical and research category through the publications of the *International Review of Sociology,* a sociological journal, edited by Routledge with a worldwide online and paper diffusion, during the last two decades. This journal, the oldest review in the field of sociology in Europe, founded by René Worms in 1893 in Paris, still maintains – as the "Aims and scope of the Review" state – «the traditional orientation of the journal as well as of the world's first international academic organization of sociology which started as an association of contributors to *International Review of Sociology*: it assumes that sociology is not conceived apart from economics, history, demography, anthropology and social psychology. Rather, sociology is a science which aims to discover the links between the various areas of social activity and not just a set of empty formulas. Thus, *International Review of Sociology* provides a medium through

which up-to-date results of interdisciplinary research can be spread across disciplines as well as across continents and cultures»[1].
The Authors proposes to highlight the different ways in which *gender issues* are declined in the international sociological researches, through an analysis of the articles published in the last two decades (1997-2017) in *International Review of Sociology.* We consider the last two decades of publication not only because of the best accessibility to the *International Review of Sociology* catalogue. For the sociology, indeed, the recent gender studies and researches have registered a deeper specialization in terms of connection with other disciplines, unusual application of the gender approach to some social phenomena, exploration of new research frontiers (multiple gender identities, gender sensitive data arrangement, the non-alignment statuses of sex and gender et similia).

2. Data and Methods

The analysis of the International Review of Sociology papers was carried out mainly through a lexicometric approach, integrated with hermeneutic analysis useful both in the first and in the last phase of the study. The first phase has regarded the collection of the corpus, while the last one has concerned the interpretation of the results obtained from quantitative and automatic procedures. The lexicometric analyses, supported by the software IRaMuTeQ[2], were carried out to extract the most relevant forms/lemma and to apply some exploratory techniques for identifying the main lexical-textual dimensions, the relationships between some keywords, the recurring topics, and possible differences over the time analysed.

2.1. The Corpus: Selection Criteria and Preliminary Analysis

The texts analyzed in this study have been collected from the archive of the *International Review of Sociology*, considering the papers published from 1997 to 2017.
In the first stage, they have been extracted all the papers which propose the term *gender* in title, abstract, body text and/or key words).They were 235, distributed over the past 20 years, as shown in Table 1.
Then, they have been selected only those papers which present a relevant

[1] See at the International Review of Sociology web site, page "Aims and scope", https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=cirs20.

[2] IRaMuTeQ is a open software, distributed under license GNU GPL, based on R statistical software and on Python language. It has now reached version 0.7 alpha 2 and it is still under development (Ratinaud, 2009).

reference to *gender* as theoretical or empirical category – and not only as a composing part of a title of some sources, a statistical variable, or synonym – in order to outline meaningful remarks for the aims of each article. This selection has been supported by a hermeneutic analysis, based on careful reading of the papers to evaluate the centrality of the gender issues in their hypotheses and theses, as in the implementation of the theoretical and/or empirical methodologies. They resulted 67, distributed over the past 20 years, as shown in Table 1.

Table 1 - Extracted and Selected Papers

	Extracted Papers (EP)	Selected Papers (SP)	SP/EP%
1997-1999	19	2	10,53
2000-2002	18	3	16,67
2003-2005	22	3	13,64
2006-2008	21	3	14,29
2009-2011	45	20	44,44
2012-2014	55	15	27,27
2015-2017	55	21	38,18
Total	235	67	28,51

The incidence of the selected papers on the extracted ones (SP/EP%) highlights the increased relevance of the term *gender* over time: it is used more and more often as analytic category in sociological research, rather than as a synonym or to indicate only a demographic characteristic of individuals.
The corpus, submitted to the subsequent analyzes, includes therefore 67 selected papers, and has the following lexicometric measurements: dimension N=495470, word types V=21680; Type/token ratio TTR= 4,38%; Hapax/V= 41,56%; Hapax/N=1,82%.
These characteristics show that the corpus can be considered sufficiently large for a quantitative approach analysis (Bolasco, 1999, p.203).

2.2. Strategy of Analysis

The analyzes on the corpus, carried out with IRaMuTeQ, will be the following:

1- Lexicon Analysis: exploration of the lexicon used in the corpus and identification of theme-words/lemma;
2- Analysis of the specific lexicon: individuation of specific words/lemma by time and by author/authors gender;
3- Correspondence Analysis: extraction of lexical dimensions starting from the Aggregated Lessical Table (ALT) Lemma/Texts (Lebart, Salem 1994),

in which the texts were identified according to the different years of publication (Y = 1997 ..., 2017;) and the gender of the author/authors (G = 1-Female; 2-Male; 3-Male and Female)

4- Cluster Analysis: identification of main topics through descending hierarchical analysis (Reinart 1983) applied to the Binary Lexical Table (BLT), Text segments / Lemma.

5- Similarity Analysis: description of the clusters obtained in point 4), through graphic representation starting from the proximity matrix between forms or lemmas.

References

Bolasco S. (1999). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*, Roma, Carocci

Lerbart L., Salem S. (1994). *Statistique textuelle*, Paris, Dunod.

Money, John; Hampson, Joan G; Hampson, John (1955). "An Examination of Some Basic Sexual Concepts: The Evidence of Human Hermaphroditism". *Bull. Johns Hopkins Hosp*. Johns Hopkins University. 97 (4), pp. 301–19.

Ratinaud, P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. http://www.iramuteq.org.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse Des Données*, 8, 187–198.

The Rhythm of Epic Verse in Portuguese From the 16th to the 21st Century

Adiel Mittmann, Alckmar Luiz dos Santos
Universidade Federal de Santa Catarina (Florianópolis, Brazil)
adiel@mittmann.net.br, alckmar@gmail.com

Abstract
The verses of most epic poems in Portuguese have been written following the example of the Italian *endecasillabo*: a verse whose last stressed syllable is the tenth, which usually means, in both Italian and Portuguese, that most verses have a total of eleven syllables. In addition to the tenth, other syllables may be stressed within the verse as well, and the specific distributions of stressed and unstressed syllables make up different rhythmic patterns. In this article, we investigate how such patterns were used in six epic poems written in Portuguese, ranging from the 16th to the 21st century, for a total of 52,412 verses. In order to analyze such a large amount of verses, we used Aoidos, an automatic scansion tool for Portuguese. By using supervised and unsupervised machine learning, we show that, though the influence of earlier poets (especially Camões) is ever present, poets favor different rhythmic patterns, which can be regarded as their rhythmic signature.
Keywords: Epic poetry, Portuguese, Scansion.

Résumé
Les vers de la plupart des épopées en portugais ont été écrits à l'instar de l'*endecasillabo* italien : un vers dont la dernière syllabe accentuée est la dixième, ce qui signifie généralement, en italien et en portugais, que la plupart des vers ont onze syllabes au total. En plus de la dixième, des autres syllabes peuvent aussi être accentuées dans ce vers, chaque combinaison de syllabes accentuées et non accentuées représentant un standard rythmique. Dans cet article, nous examinons comment ces standards ont été utilisés dans six épopées écrites en portugais, du XVI^{ème} au XXI^{ème} siècles, dans un total de 52.412 vers. Pour analyser une telle quantité de vers, nous avons employé Aoidos, un outil automatique de scansion pour le portugais. En utilisant des apprentissages supervisés et non-supervisés, nous concluons que, encore que l'influence de poètes précédents (surtout celle de Camões) se fasse toujours remarquer, chaque poète préfère de différents standards rythmiques, qui peuvent être considérés comme sa signature rythmique.
Mots-clés: Epopée, Portugais, Scansion.

1. Introduction

Poets are frequently compared to one another, but over the centuries rarely have such comparisons been made objectively, especially with respect to verse structures. When critics state that a poet has followed the steps of another too closely and has therefore produced unoriginal and derivative work, they can seldom rely on objective facts. Works such as that of Chociay (1994), who manually analyzed and tabulated more than 1,500 verses, are not the rule, but the exception. It is indeed a tedious and tiresome task for any human to carry out; but looking at a great amount of text from afar and extracting relevant information from it constitutes a core element of *distant reading* (Moretti, 2013).

Table 1: Poems included in the corpus. The code is derived from the poem's title.

Code	Author	Born in	Poem	Year	Verses
L	Luís de Camões	Portugal	*Os Lusíadas*	1572	8,816
M	Francisco de Sá de	Portugal	*Malaca*	1634	10,656
C	Santa Rita Durão	Brazil	*Caramuru*	1781	6,672
A	Fagundes Varela	Brazil	*Anchieta*	1875	8,484
B	Carlos Alberto Nunes	Brazil	*Os Brasileidas*	1938	8,504
F	José Carlos de Souza	Brazil	*Famagusta*	2016	9,280
					52,412

In this article, we turn our attention to the verse most commonly used in epic poetry in Portuguese, the *decassílabo*, which was borrowed from Italian[1]. It is the verse used by Dante in his *Divina Commedia* and by Petrarch in his *Canzoniere*. Stressed syllables are distributed in the verse according to certain rules; in particular, the 10th syllable (which defines the length of the verse) must always be stressed. Other syllables may also be stressed, producing many possible rhythmic patterns—which are, both in Portuguese and Italian, required to have their 6th or, less commonly, their 4th syllable stressed (Versace, 2014). We identify such patterns by indicating the syllabic positions that are stressed within a given verse, so that a pattern like 3-6-10 means that the 3rd, 6th and 10th syllables are stressed.

We are interested in tracking which rhythmic patterns poets have favored

[1] In both Italian and Portuguese, this kind of verse always has its 10th syllable stressed and typically has a total of eleven syllables, since most words in both languages have a stress on the penult. However, in Italian this verse is called *endecasillabo* because of the total number of syllables, whereas the Portuguese term *decassílabo* emphasizes the fact that the 10th is the last stressed syllable in the verse.

over the centuries and whether such patterns are characteristic to each poet. For this purpose, we have assembled a corpus consisting of six poems, whose publication dates range from the 16th to the 21st century, for a total of 52,412 verses (about 300,000 words). In order to analyze such an amount of verses, we have used our automatic scansion tool, Aoidos (Mittmann et al., 2016), which is capable of scanning thousands of verses in a few seconds and producing rhythmic information. The next section describes the corpus we used in our experiments; Section 3 reports the results obtained with our analyses; finally, Section 4 presents our conclusions and discusses future work.

2. Corpus

The poems chosen to compose the corpus for this article are summarized in Table 1. We adopted two criteria in order to select these poems. Firstly, we searched for an important—and thus well known—or exemplary epic poem in each century, from the 16th up to the present. Secondly, we required trustful and reliable digital editions; in one case (17th century), we produced a digital edition especially for this article, since no suitable candidate was found.

Camões' poem *Os Lusíadas* is by far the most important epic poem ever written in Portuguese. Its influence can be felt, for instance, even in 20th-century lyrical poets such as Jorge de Lima. Meneses' *Malaca Conquistada* and Durão's *Caramuru* follow very closely the Camonean model: they use identical rhyme schemes, they have a similar argument and they celebrate a protagonist in like manner. Nevertheless, we would like to investigate whether the two authors innovated with respect to rhythm, even though they kept the overall model of the Camonean epic. These three poems in our corpus were written by Portuguese citizens (Durão was born in colonial Brazil and died before the country's independence), while the remaining three poems were written by Brazilian poets.

	1	2	3	4	5	6	7	8	9	10	11
1-4-6-10	**Ce-**	ssem	do	**sá-**	bio	**Gre-**	go e	do	Troi-	**a-**	no
6-10	As	na-	ve-	ga-	ções	**gran-**	des	que	fi-	**ze-**	ram;
1-6-10	**Ca-**	le-	se	de A-	le-	**xan-**	dro e	de	Tra-	**ja-**	no
2-	A	**fa-**	ma	das	vi-	**tó-**	rias	que	ti-	**ve-**	ram;

6-10											
2-4-6-10	Que eu	**can-**	to o	**pei-**	to i-	**lus-**	tre	Lu-	si-	**ta-**	no,
2-4-6-10	A	**quem**	Nep-	**tu-**	no e	**Mar-**	te o-	be-	de-	**ce-**	ram:
1-3-6-8-10	**Ce-**	sse	**tu-**	do o	que a	**Mu-**	sa an-	**ti-**	ga	**can-**	ta,
1-4-6-10	**Que ou-**	tro	va-	**lor**	mai-	**s al-**	to	se a-	le-	**van-**	ta.

Figure 1: Scansion produced by Aoidos.

Fagundes Varela's *Anchieta*, a romantic piece of the 19th century, would not be, at a first glance, an epic poem, since its subject is the telling of New Testament stories to Brazilian Indians by priest José de Anchieta. However, as historian Maria Aparecida Ribeiro and others remark, *Anchieta* is a kind of "religious epopee" (Ribeiro, 2003), which drives our attention to the Romantic effort to renew the ancient models inherited from Classical or Neoclassical literature (although it clearly returns to the Greek epic model, as it does not adopt regular sized stanzas). Despite some important differences in the narrative logic, the verses reproduce the most important invariants of the genre: the honoring of a protagonist (Anchieta) and the use of the *decassílabo* (blank ones, in this case). As for Carlos Alberto Nunes' *Os Brasileidas*, this poem also presents some invariants that characterize the traditional epic poem: blank *decassílabo* verses; several cantos, beginning with the proposition; the intention of celebrating an individual hero, in this case Antônio Raposo Tavares, a 17th-century Brazilian trailblazer. In addition to the absence of rhymes, in order to emphasize the differences in relation to the Camonean epic style, there is no regular stanza division in each one of the nine cantos (ten, if we consider the epilogue), as in *Anchieta*, although they may vary significantly, from seven up to sixty five or more verses. Finally, regarding *Famagusta*, by José Carlos de Souza Teixeira, one quickly notices that it is a curious combination of traditional epic elements from different ages. In addition to the epic intention of celebrating an historical event and

some sort of heroic action, its formal elements are, to say the least, very heterogeneous. For instance, it takes the Camonean eight verse stanza but adopts a different rhyme scheme, resulting no more in the well-known *ottava rima* (ABABABCC), but in the medieval Sicilian stanza called *strambotto romagnuolo* (ABABCCDD), scarcely used in Brazilian literature[2].

3. Analysis

In order to analyze the corpus, we used Aoidos, an automatic scansion tool for Portuguese (Mittmann et al., 2016), much like Métromètre (Beaudouin and Yvon, 2004) and Anamètre (Delente and Renault, 2015) for French. Starting from the written word, Aoidos produces a phonetic transcription for each verse and then applies many rules (such as elision or syncope) to produce a series of alternative scansion. By examining the poem as a whole, the system then selects the most appropriate alternative and, by applying a set of heuristics, proposes a rhythmic pattern for each verse. The scansions generated by Aoidos have been manually verified to be correct in 99.0% of cases (Mittmann, 2016). Figure 1 shows the output produced by the system for the 3rd stanza of Camões' *Os Lusíadas*.

Table 2: Rhythmic pattern usage (%) for each poem.

Poem	3-6-10	2-6-10	2-4-6-10	2-6-8-10	3-6-8-10	2-4-6-8-10	1-3-6-10	1-4-6-10	4-6-10	2-4-8-10	1-3-6-8-10	1-4-6-8-10	4-6-8-10	1-4-8-10	4-8-10	1-6-10	1-6-8-10
L	10.3	9.0	**15.2**	7.7	7.6	11.0	6.2	7.9	6.2	1.1	4.5	5.0	4.1	0.5	0.4	1.3	1.0
M	11.9	9.6	**12.2**	7.1	8.2	9.5	5.2	5.7	5.5	6.5	3.8	3.9	3.6	2.2	1.8	1.1	0.8
C	10.3	7.1	**11.2**	7.6	8.5	9.1	7.7	6.1	3.6	8.1	5.8	4.6	2.9	3.6	2.2	0.4	0.8
A	**14.4**	11.3	7.3	9.4	9.0	4.0	6.1	7.2	5.4	6.5	4.2	3.3	2.5	4.6	1.2	1.6	1.3
B	**14.0**	13.2	7.7	9.6	7.3	5.1	6.8	5.1	5.2	4.6	3.5	2.7	3.5	3.4	3.1	2.1	1.4
F	13.8	**16.2**	11.7	10.5	9.0	8.0	7.0	4.0	4.5	0.2	4.3	3.1	2.7	0.1	0.1	1.8	1.2

A total of 42 different rhythmic patterns were found among all 6 poems. Table 2 shows how frequently patterns with an average usage of at least 1% were employed in each poem. In each row, the bold number indicates the pattern most favored by that row's poem. Although some patterns, such as

[2] The Brazilian-born baroque poet Manoel Botelho de Oliveira did use this stanza in some madrigals written in Spanish, such as this one: Si Cupido me inflama, / Si desdeñas mi empleo; / En amorosa llama, / En nieve desdeñosa el Etna veo, / Con amor, y tibieza / Tenemos su firmeza, / Y en disonancia breve / Suspiro fuego yo, tu brotas nieve.

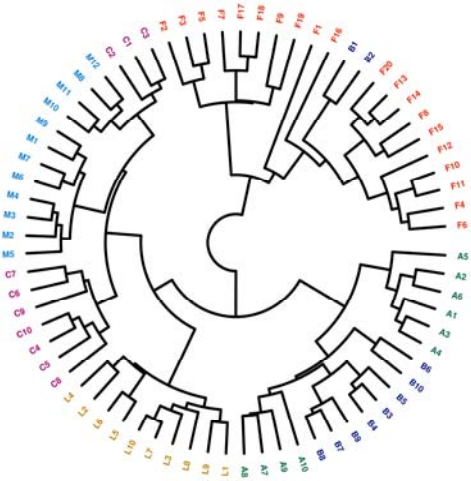


Figure 2: Dendrogram built from all cantos of all poems.

3-6-8-10 and 1-3-6-10 remain more or less constant, many others display a wide range of relative usage: pattern 2-6-10 ranges from 7.1% to 16.2%, and pattern 1-4-8-10 from 0.1% to 3.1%. Whereas Camões (L) does seem to set the tone for the following poems, there are clear differences when one considers patterns such as 2-4-6-10 and 2-4-8-10. In fact, pairs such as *Malaca* (M) and *Caramuru* (C) or *Anchieta* (A) and *Os Brasileidas* are more similar between themselves than Camões' *Os Lusíadas* (L) is to any other poem. By looking at numbers from one century to the next, twice a change of more than 5% can be seen: from *Caramuru* (C) to *Anchieta* (A) there was a decrease of 5.1% for the pattern 2-4-6-8-10, and from *Os Lusíadas* (L) to *Malaca* (M) the pattern 2-4-8-10 increased in usage by 5.4%.

An interesting question arises at this point: do smaller parts of the poems reflect the overall distribution shown in Table 2? In other words, given a smaller part of a poem, could we tell from which work it was taken simply by looking at its rhythmic signature? To answer this question, we divided each poem into its cantos, for a total of 72 divisions, with an average of 727.9 verses per canto. We then extracted the usage frequency of the rhythmic patterns, thus producing a feature vector for each canto. By iteratively clustering such vectors, we obtained the dendrogram shown in Figure 2; complete linkage was used. Each canto in the figure is indicated by a letter (the poem code) and a number (the canto number within the poem). Cantos from the same poem are also displayed with the same color. The closer to the center that two branches link together, the more different the cantos they contain are. We can immediately see that, in general, cantos that belong to the same poem are located next to each other. All cantos of Camões' *Os*

Lusíadas (L), in particular, are tightly grouped in their own branch. It is also interesting to note that, except for *Famagusta* (F), whenever a smaller group of cantos from the same poem were placed far from the larger group of cantos, there is a certain order: it was the first three cantos of *Caramuru* (C) were separated; the last four of *Anchieta* (A); and the first two of *Os Brasileidas* (B). Two cantos from *Famagusta* (F1 and F16) are only linked with other nodes at a great distance; this stems from the fact that these two cantos are the shortest ones in all of the corpus: the first canto has only 24 verses, the sixteenth 112. Such small amounts of verses produce poor feature vectors.

In order to further investigate how well the cantos reflect the poems, we employed a nearest centroid classifier. In this case, each of the 72 feature vectors (the rhythmic signatures of the cantos) was labeled with the poem they belong to. We then used stratified k-fold cross validation, with k = 4 and 100 repetitions to assess the classifier's performance. The mean precision obtained was 96.5%, mean recall 95.9% and mean F1 score 95.5%; the mean accuracy was 95.6%. This means that, given a sample of 54 cantos (because k = 4), the classifier guesses the right poem for the other 18 cantos in about 96% of the cases.

4. Conclusion

The frequency with which poets employ certain patterns of stressed and unstressed syllables in their verses can be regarded as a rhythmic signature—at least in epic poems, the subject of this article. In this work, we have subjected 72 individual cantos to a hierarchical clustering technique (Figure 2), which shows that rhythmic patterns do reflect an author's preferences (unconscious as they might be). Furthermore, a nearest centroid classifier obtained a mean accuracy of 95.6%, which is also evidence for the existence of a rhythmic signature. This kind of analysis is possible thanks to automatic scansion systems, such as Aoidos, which allow a large amount of verses (more than 50,000 in this case) to be scanned and analyzed.

Although Camões, whose poem *Os Lusíadas* is the oldest in our corpus, has influenced newer generations of poets, this article shows that, at least rhythmically, each poet in our corpus took their own path. In fact, Camões' verses are the ones most easily distinguished from the others (see Figure 2). Lesser-known poems, such as *Malaca* or *Os Brasileidas*, have not failed to produce rhythmic signatures that, in most cases, set them apart from other works. In addition to the rhythmic signature, we would like to investigate, in the future, additional features that could be extracted from verses and used in stylometric analyses. In particular, the *decassílabo* usually falls into one of two categories: either the 6th syllable has the dominant stress or—less commonly—the 4th; in the former case, the verse is *heroic*; in the latter,

Sapphic. A verse whose rhythmic pattern includes the 6th syllable, but not the 4th, is heroic; but one that includes both the 6th and the 4th could be either heroic or Sapphic. It would be interesting to resolve this ambiguity and evaluate how well these categories characterize a poet's style.

Although this article has only considered epic poems, there is no reason to believe that rhythmic signatures are limited to this genre. In the future, we would like to explore how well the approach shown here fares when applied to other verses and other genres.

Acknowledgments

For the nearest centroid classifier we employed Scikit-learn (Pedregosa et al., 2011). For the dendrogram, we used Dendextend (Galili, 2015) and Circlize (Gu et al., 2014).

References

Beaudouin, Valérie and Yvon, François (2004). "Contribution de la métrique à la stylométrie". 7èmes Journées internationales d'Analyse statistique des Données Textuelles. (2004), pp. 107–118.

Chociay, Rogério (1994). A Identidade Formal do Decassílabo em "O Uraguai". *Revista de Letras* 34, 229–243.

Delente, Éliane and Renault, Richard (2015). Projet Anamètre : Le calcul du mètre des vers complexes. *Langages* 3.199, 125–148.

Galili, Tal (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31 (22), 3718–3720.

Gu, Zuguang et al. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics* 30 (19), 2811–2812.

Mittmann, Adiel (2016). "Escansão Automático de Versos em Português". PhD thesis. Universidade Federal de Santa Catarina.

Mittmann, Adiel, Wangenheim, Aldo von, and Luiz dos Santos, Alckmar (2016). "Aoidos: A System for the Automatic Scansion of Poetry Written in Portuguese". 17th International Conference on Intelligent Text Processing and Computational Linguistics. (2016).

Moretti, Franco (2013). *Distant reading*. London: Verso.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

Ribeiro, Maria Aparecida (2003). Anchieta no Brasil: Que Memória? *História Revista* 8, 21–56.

Versace, Stefano (2014). A Bracketed Grid account of the Italian *endecasillabo* meter. *Lingua* 143, 1–19.

Le vocabulaire des campagnes électorales

Denis Monière[1], Dominique Labbé[2]
[1] Université de Montréal (denis.moniere@umontreal.ca)
[2] PACTE CNRS - Université de Grenoble (dominique.labbe@umrpacte.fr)

Abstract
After having done a first presidential term, V. Giscard d'Estaing, F. Mitterrand, J. Chirac and N. Sarkozy were candidates for a second term. In this study, their electoral speeches are compared with their presidential ones drawing attention to the specific nature of the vocabulary used. It would appear that this calculation is mainly biased by grammatical categories and word frequency. We present modifications of the classical formulae which make it possible to neutralize the influence of grammatical categories and, at least partially, that of word frequency. Electoral discourse privileges the verb over the name, as such speech is more personalized than governmental discourse, it focuses on the country and its inhabitants, the rest of the world being pushed into the background. Finally, in recent years, the polemical dimension is becoming predominant.

Résumé
Après un premier mandat présidentiel, V. Giscard d'Estaing, F. Mitterrand, J. Chirac et N. Sarkozy ont été candidats à un deuxième mandat. On compare leurs discours électoraux avec leurs discours présidentiels à l'aide des spécificités du vocabulaire. Il apparaît que ces spécificités dépendent surtout des catégories grammaticales et des effectifs des mots. On présente des modifications du calcul classique qui permettent de neutraliser l'influence des catégories grammaticales et, au moins partiellement, celle des fréquences. Le discours électoral privilégie le verbe au détriment du nom, il est plus personnalisé que le discours au pouvoir, il se centre sur le pays et ses habitants, le reste du monde passant au second plan. Enfin, ces dernières années, la dimension polémique devient prédominante.

Keywords: lexicometry ; political discourse ; French presidential campaigns ; specific vocabulary ; spécificités du vocabulaire.

1. Introduction

Le discours électoral diffère-t-il du discours de gouvernement et en quoi ? La réponse est difficile car il faut neutraliser l'effet des personnalités et des conjonctures pour isoler l'effet sur le discours des choix stratégiques du

locuteur. L'idéal serait de pouvoir étudier les mêmes hommes à peu près simultanément dans les deux positions de gouvernant puis de candidat. Le corpus des discours des présidents français depuis 1958 remplit ces deux conditions (présentation du corpus dans Arnold et al 2016). En effet, pour 5 présidents (C. de Gaulle, V. Giscard d'Estaing, F. Mitterrand, J. Chirac et N. Sarkozy), ce corpus contient leurs interventions lorsqu'ils étaient présidents et leurs discours de campagne pour leur réélection. Certes, en 1965, de Gaulle n'a pratiquement pas fait campagne (Labbé 2005), mais ses successeurs ne l'ont pas imité en 1981, 1988, 2002 et 2012 (corpus en annexe).
Pour comparer ces corpus, le calcul des "spécificités" semble l'outil le plus adapté (Lafon 1980 et 1984). Il rapporte le vocabulaire d'un sous-ensemble de textes (sous-corpus) à un corpus de référence. Mais il se heurte à une double difficulté : la spécificité éventuelle d'un vocable est liée à sa catégorie grammaticale et à sa fréquence d'emploi (Labbé, Labbé 1994 ; Monière et al. 2005), comme nous allons le vérifier d'abord avec le cas de Sarkozy en 2012 (Sur cette campagne : Labbé, Monière 2013). Dès lors, la mesure des spécificités doit neutraliser, autant que possible, ces deux inconvénients.

2. Les catégories grammaticales du discours électoral

Le discours présidentiel de Sarkozy s'étend de son investiture (16 mai 2007) au 12 février 2012 (annonce de sa candidature). La campagne s'étend jusqu'au soir du second tour (6 mai 2012). Le corpus complet (P) compte 1074 interventions, soit au total 3 221 259 mots avec 21 602 vocables différents. A partir de sa déclaration de candidature, Sarkozy est intervenu 110 fois (sous-corpus E), soit 369 808 mots et un vocabulaire de 8 511 vocables différents. Ces interventions sont d'abord marquées par un net changement de style (tableau1).

Tableau 1. Densités des catégories grammaticales dans les interventions de Sarkozy lors de la campagne de 2012 comparées à ses interventions comme président 2007-2012 (en ‰)

Catégories	P-E (Corpus-Sous corpus)	E Sous corpus	(P-E)/P	Indice
Verbes	159.2	169.4	+6.4	+
Futurs	7.0	7.2	+1.6	+
Conditionnels	3.2	2.8	-11.2	-
Présents	82.9	89.3	+7.7	+
Imparfaits	6.4	6.4	-0.2	≈
Passés simple	0.6	0.3	-55.2	-
Participes passés	20.8	23.8	+14.6	+
Participes présents	2.1	2.1	+2.9	≈
Infinitifs	36.3	37.6	+3.6	+
Noms propres	27.9	23.0	-17.3	-
Substantifs	178.4	176.0	-1.3	-
Adjectifs	54.0	46.6	-13.7	-

Adj. participe passé	5.2	4.5	-13.1	-
Pronoms	124.3	132.6	+6.7	+
Pronoms personnels	65.4	69.6	+6.5	+
Déterminants	181.6	182.5	+0.5	+
Articles	131.9	128.1	-2.9	-
Nombres	18.7	20.9	+11.9	+
Possessifs	14.5	17.0	+17.3	+
Démonstratifs	7.6	7.8	+2.7	+
Indéfinis	8.9	8.7	-2.4	-
Adverbes	67.1	68.9	+2.7	+
Prépositions	150.1	145.6	-3.0	-
Coordinations	29.1	25.4	-12.7	-
Subordination	25.9	27.9	+8.0	+

Dans le discours présidentiel, on rencontre 159 verbes en moyenne pour 1 000 mots ; dans les discours électoraux, cette proportion passe à 169‰, soit une augmentation de +6,4%, ce qui est un écart significatif avec moins de une chance sur 10 000 de se tromper (signe + en dernière colonne). Les lignes suivantes donnent le détail des temps et des modes. Le recul le plus significatif concerne le conditionnel (le discours électoral ne doit pas connaître le doute). En revanche, le participe passé connait l'augmentation la plus forte (le président sortant peut difficilement éviter de défendre sa gestion).

Les pronoms, les adverbes et les conjonctions de subordination évoluent dans le même sens que les verbes. Ils sont réunis dans le "groupe du verbe". A l'inverse, les substantifs, adjectifs, articles et prépositions suivent la tendance inverse : groupe du nom. Le tableau 2 donne les densités des deux groupes chez les 4 présidents.

Tableau 2. Densités des groupes du verbe et du nom (en ‰) dans les discours électoraux (E) comparés aux discours présidentiels (P-E).

Catégories	P-E (Corpus-Sous corpus)	E Sous corpus	(P-E)/E	Indice
Sarkozy (2007-2012)				
Groupe du verbe	376.6	398.9	+5.9	+
Groupe du nom	621.1	599.2	-3.5	-
Giscard d'Estaing (1974-1981)				
Groupe du verbe	351.5	392.5	+11.7	+
Groupe du nom	646.1	604.5	-6.4-	-
Mitterrand (1981-1988)				
Groupe du verbe	386.4	427.1	+10.5	+
Groupe du nom	611.0	569.8	-6.7	-
Chirac (1995-2002)				
Groupe du verbe	329.5	333.2	+1.1	+
Groupe du nom	668.8	665.1	-0,6	-

Chez tous les présidents en campagne, il se produit une augmentation du groupe du verbe et un recul de celui des noms. Statistiquement, ces mouvements sont significatifs (avec ◎ = 1%). L'écart le plus fort est observé chez Giscard d'Estaing puis chez Mitterrand. Cependant, Chirac tranche sur les autres avec une densité du verbe beaucoup plus faible et une campagne présidentielle presqu'aussi distanciée que ses interventions lors de son premier mandat, marqué par une cohabitation de 5 ans (1997-2002) avec un Premier ministre socialiste (Jospin). Dans son discours électoral, la densité des verbes augmente nettement (+3,6%) mais se trouve en partie compensée par un recul des pronoms, ce qui accentue le caractère dépersonnalisé des propos de Chirac à l'opposé des trois autres.
En conséquence, pour les 4 présidents, les principaux verbes apparaissent en spécificités positives du discours électoral et il ne s'en trouve que quelques-uns en spécificités négatives. Il en est de même pour les pronoms et les adverbes. La situation inverse se constate pour les adjectifs, les substantifs, etc. Autrement dit, si un mot appartient à une catégorie sous-employée dans le sous-corpus (par rapport à sa densité d'utilisation dans le corpus entier), ce vocable a toute chance d'apparaître dans les spécificités négatives (et positives dans le cas inverse). Il est possible de neutraliser ce biais.

3. Neutralisation de la catégorie grammaticale

Le calcul standard est le suivant. Soit :
- le corpus de référence (*P*) long de N_p mots ;
- le sous-corpus *E* long N_e mots dont on recherche les spécificités par rapport à *P* ;
- un vocable *i* avec F_{ip} occurrences dans *P* et F_{ie} dans *E*. Si sa répartition est uniforme, ce vocable apparaîtra $E_{ie(u)}$ fois dans le sous-corpus *E* :

$$E_{ie(u)} = F_{ip} * U \text{ avec } U = \frac{N_e}{N_p} = \frac{369\,808}{3\,223\,570} = 0.113 \quad (1)$$

La probabilité pour que le vocable *i* soit observé F_{ie} fois dans *E* suit une loi hypergéométrique de paramètres F_{ip}, F_{ie}, N_e, N_p :

$$P(X = F_{ie}) = \frac{\begin{bmatrix} F_{ip} \\ F_{ie} \end{bmatrix} \begin{bmatrix} N_p - F_{ip} \\ N_e - F_{ie} \end{bmatrix}}{\begin{bmatrix} N_p \\ N_e \end{bmatrix}} \quad (2)$$

L'indice de spécificité (S) est la somme des probabilités – calculées avec (2) –

de survenue des *J* valeurs entières de *X* variant de 0 à F_{ie} {*X*=0 ; *X*= F_{ie}} :

$$S = P(X \leq F_{ie}) = \sum_{j=0}^{j=Fie} P(X = j) \qquad (3)$$

Si au seuil◎◎, F_{ie} excède $E_{ie(u)}$, le vocable est « spécifique plus » (S+) ; S- dans le cas contraire. Avec ce calcul, la plus grande partie des verbes usuels de Sarkozy apparaissent donc en S+ de sa campagne électorale et la majorité des substantifs en S-, parce que, dans ses discours électoraux, la première catégorie est privilégiée par rapport au discours de gouvernement où elle est moins utilisée (à l'inverse des substantifs). Pour corriger ce biais, le calcul prend en compte les catégories grammaticales (g). La modification est présentée dans : Monière, Labbé, Labbé 2005 ; Mayaffre 2006 et Monière, Labbé 2012.
Soit : N_{ge} et N_{gp} le nombre de mots appartenant à la catégorie grammaticale *G* respectivement dans le sous-corpus *E* et le corpus entier *P*. Les formules (1) et (2) deviennent :

$$E_{ie(u)} = F_{ip} * U \; avec \; U = \frac{N_{ge}}{N_{gp}} \qquad (4)$$

$$P(X = F_{ie}) = \frac{\begin{bmatrix} F_{ip} \\ F_{ie} \end{bmatrix} \begin{bmatrix} N_{gp} - F_{ip} \\ N_{ge} - F_{ie} \end{bmatrix}}{\begin{bmatrix} N_{gp} \\ N_{ge} \end{bmatrix}} \qquad (5)$$

Les formules (4) et (5) appliquées aux 4 corpus aboutissent à un équilibre relatif, au sein de chaque catégorie, entre les S+ et les S- (tableau 4). Ces formules neutralisent donc la liaison entre spécificités et densité des catégories grammaticales. Comme indiqué dans Monière & Labbé 2012, cette modification change drastiquement la liste des "mots spécifiques" mais elle laisse subsister la liaison entre spécificité et fréquence.

4. Questions de seuils

Le calcul porte sur une minorité du vocabulaire et il est asymétrique. En effet, avec ◎ = 1% :
- l'effectif minimal pour être S+ est de 5 occurrences ("seuil de spécificité positive"), toutes dans les discours électoraux (*E*) et à condition que $E_{ie(u)} < .5$, ce qui signifie que $N_{ge} < 0.10 N_{gp}$. Par construction le calcul élimine donc tous les vocables d'effectifs inférieurs à 5. Dans le corpus Sarkozy, cela représente

plus de la moitié du vocabulaire (54 % des vocables). Autrement dit, seulement 46% du vocabulaire peut être S+ ;
- le "seuil de spécificité négative" correspond à la situation suivante : un vocable *i* absent de *E* ($F_{ie} = 0$) alors qu'on en attend au moins 5 ($E_{ie(u)} \geq 5$). En pratique, cela signifie que son effectif dans *P* est égal ou supérieur à 5*1/*U*, soit ici 40. Autrement dit, pour le discours électoral de Sarkozy, 83% du vocabulaire de *P* ne peut apparaître en S-.
Dès lors, les vocables dont les effectifs dans *P* sont compris entre 5 et 39 peuvent être S+ mais pas S- dans *E*. On s'attend donc à ce qu'il y ait plus de vocables S+ que S-.

5. Liaison entre spécificité et fréquence

9 876 vocables apparaissent 5 fois ou plus dans *P*. Si ce corpus était homogène (hypothèse nulle H_0), une distribution normale des vocables laisserait attendre - avec◎◎ = 1% - environ 100 vocables spécifiques. Le tableau 3 compare les résultats observés et attendus (avec H_0).

Tableau 3. Effectifs des vocables classés par catégories grammaticales et par spécificités

	Effectifs ($F_{iP} \geq 5$)	H_0	S+	S-	Total S
Verbes	1 540	15	176	143	319
Mots à majuscule	1 501	15	112	142	254
Substantifs	4 175	42	455	468	923
Adjectifs	2 065	21	140	115	255
Pronoms	52	1	18	13	31
Adverbes	411	4	20	57	77
Déterminants	72	1	21	12	33
Prépositions & conjonc.	60	1	21	9	30
Total	9 876	100	963	959	1 922

Il y a donc vingt fois plus de vocables spécifiques que n'en laisse attendre H_0 (répartition homogène des mots entre corpus et sous-corpus). A priori, cela signifie simplement que discours électoral et discours de gouvernement sont fortement contrastés. En fait, ce décalage provient essentiellement des vocables les plus fréquents (tableau 4 et Figure 1).

Tableau 4. Proportion des vocables spécifiques de E dans l'ensemble du vocabulaire (P) classé par fréquence absolues.

Classe de fréquence (*P*)	Vocables spécifiques de *E* dans la classe	Total vocables de *P* dans la classe	Proportion des vocables de *P* spécifiques de *E*
5-9	64	2 759	2,3
10-14	68	1 237	5,5
15-19	55	757	7,3
20-29	89	987	9,0

30-49	143	997	14,3
50-99	317	1 054	30,1
100-199	332	799	41,6
200-499	398	686	58,0
500+	473	640	73,9
Total	1 939	9 916	19,6

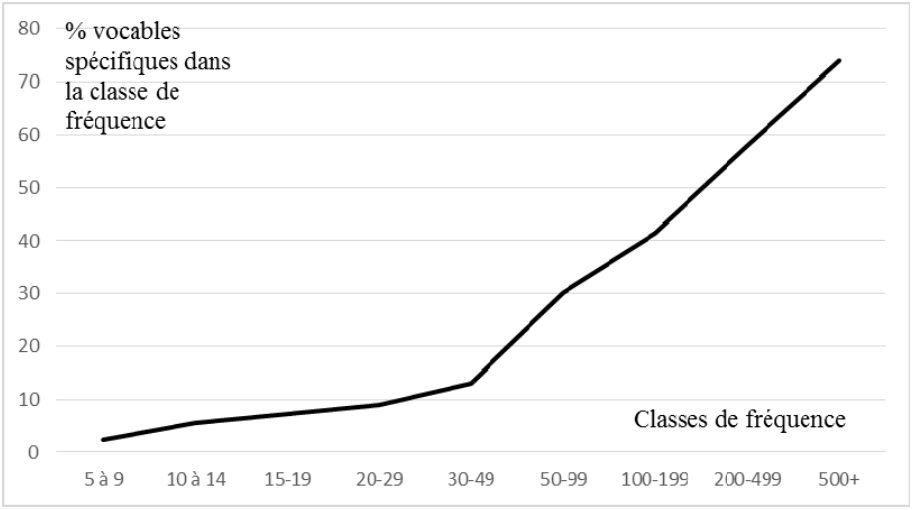


Figure 1. Liaison entre la spécificité et la fréquence

Au-dessus du seuil de spécificité positive (ici 40), la proportion de vocables spécifiques est directement corrélée avec la fréquence : la courbe suit la diagonale du tableau et le coefficient de détermination de Y par X est égal à 0,997, ce qui indique une liaison rigide et linéaire. Il en est toujours ainsi : plus un vocable donné est fréquent dans un corpus, plus il a de chances d'être "spécifique" à l'une quelconque des parties de ce corpus. Cette dépendance peut être interprétée de deux manières. D'une part, l'essentiel des choix thématiques seraient véhiculés par les vocables les plus fréquents et la variation dans leurs fréquences d'emploi seraient la principale manifestation de ces choix. Cependant, dès que le corpus atteint une certaine longueur, l'observateur se trouve noyé dans des listes qui contiennent la plus grande part du vocabulaire usuel, ce qui en rend l'interprétation difficile. D'autre part et à l'inverse, on peut penser que le raisonnement probabiliste - qui sous-tend ce calcul - doit être adapté à cette liaison manifeste entre spécificité et fréquence.

6. Neutralisation de la liaison entre fréquence et spécificité

Les limites des classes de fréquence du tableau 5 et de la figure 1 ont été fixées selon une échelle proche d'une progression géométrique, ce qui assure

aux classes des effectifs sinon égaux du moins suffisamment proches et importants. Ceci correspond à une particularité dite "loi de Zipf" - ou "Zipf-Mandelbrot" - selon laquelle le nombre d'occurrences d'un mot dans un texte est lié à son rang dans la distribution des fréquences (Zipf 1935 ; Mandelbrot 1957).

Dès que le corpus atteint une longueur suffisante (au moins un demi-million de mots) et que le sous-corpus est égal à au moins d'un dixième du corpus, on peut découper le vocabulaire en quelques classes de fréquence. Pour un corpus de la dimension de celui de Sarkozy (et des trois autres présidents), trois classes suffisent : vocables "rares" (inférieurs à 100 occurrences) ; "fréquents" (de 100 à moins de 500) ; "très fréquents" (500 et plus). Dans ces trois classes, les vocables sont classés par catégorie grammaticale puis en fonction de leur indice de spécificité et, dans chacune des classes, seuls les plus caractéristiques sont retenus. Le tableau 5 donne les 5% les plus caractéristiques du discours électoral de Sarkozy comparé à son discours présidentiel, pour trois catégories grammaticales.

Tableau 5. Spécificités les plus remarquables du discours électoral de Sarkozy par rapport à son discours présidentiel (par catégories grammaticales en trois classes de fréquence)

	<100	100 – 499	500+
Vocables significativement sur-employés :			
Verbes :	voler, cotiser, détester, casser, éduquer, suspendre, démolir	adresser, bénéficier, apprendre, souffrir, supprimer, régulariser	dire, vouloir, parler, vivre, proposer, changer, respecter, défendre
Mots à majuscule	Mélenchon, Le Pen,	François, Polynésie, Hollande, Schengen, TVA	France, Français, Corse
Substantifs	honte, rassemblement, héritier, socialiste, colère, délit, amalgame	jeunesse, souffrance, gauche, destin, erreur, étranger, salaire, outremer	travail, entreprise, droit, république, vie, emploi, ami, enfant, territoire, peuple,
Vocables significativement sous-employés :			
Verbes	admirer, illustrer, expérimenter, inaugurer,	progresser, témoigner, évoquer, marquer, associer	être, devoir, savoir, comprendre, trouver, attendre, remercier, essayer
Mots à majuscule	Bush, Poutine, Roumanie, Quatar	Russie, Inde, Iran, Barroso	Afrique, G20, Méditerranée, Merkel, Paris, Chine
Substantifs	refondation, coalition, scientifique, lycéen	processus, visite, équipe, conférence, planète, gouvernance, alliance,	pays, monsieur, président, état, ministre, politique, gouvernement, question

Chez Sarkozy, le discours électoral est affaire de volonté, il se centre sur le pays, ses habitants mais aussi l'adversaire – la gauche, Hollande - dont il dénonce les amalgames et les erreurs. Les spécificités négatives indiquent que le discours électoral n'est pas affaire de devoir ou de connaissance ; il "oublie" le reste du monde et ses dirigeants, les institutions du pays comme le gouvernement et les ministres, etc.

7. Conclusions

Lorsqu'un président entre en campagne, il doit descendre dans l'arène et adopter un discours de combat qui se caractérise avant tout par une augmentation de la densité des verbes, une forte personnalisation et un recul de la place accordée aux substantifs et aux adjectifs. Ces caractéristiques se retrouvent dans les discours électoraux des Premiers ministres canadiens (Monière, Labbé 2010). Cependant, en campagne ces derniers insistent sur le "nous" car, dans un système parlementaire, il s'agit de faire élire une majorité de députés, alors que les présidents français privilégient le "je"… Enfin, ces dernières années en Amérique du nord comme en France, la forte présence de la construction négative et la désignation des adversaires (noms propres) soulignent le caractère polémique du discours électoral.

Le calcul des spécificités – tel qu'il est utilisé en analyse des données textuelles – enregistre la catégorie grammaticale du vocable analysé et sa fréquence d'emploi et non pas les choix thématiques du locuteur. La neutralisation de la catégorie grammaticale est aisée si les mots ont été étiquetés. En revanche, l'effet de la fréquence est susceptible de plusieurs interprétations. Toutefois, si l'on souhaite ne pas être enseveli sous les listes produites par le calcul classique, la solution réside dans le classement des vocables en classes de fréquence –selon une échelle géométrique - et, au sein de chacune de ces classes, dans la sélection des vocables les plus singuliers. A ce prix, les singularités d'un sous-corpus peuvent être identifiées sans avoir à effectuer des tris discutables dans des listes trop longues.

References

Arnold E., Labbé C. & Monière D. (2016). *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble : Laboratoire d'Informatique de Grenoble, 2016.

Labbé C., Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble : CERAT, décembre 1994. Reproduit dans *Lexicometrica*, 3, 2001.

Labbé D., Monière D. (2010). Quelle est la spécificité des discours électoraux? Le cas de Stephen Harper. *Canadian Journal of Political Science*, 43:1, p. 69–86.

Labbé D., Monière D. (2013). *La campagne présidentielle de 2012. Votez pour*

moi ! Paris : l'Harmattan.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, p. 127-165.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris : Slatkine-Champion.

Mandelbrot B. (1957). Étude de la loi d'Estoup et de Zipf Fréquences des mots dans le discours. Apostel L et al. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

Mayaffre D. (2006). Faut-il pondérer les spécificités lexicales par la composition grammaticale des textes ? Tests logométriques appliqués au discours présidentiel sous la Vème République. Condé C., Viprey J.-M. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon : Presses universitaires de Franche Comté, II, p. 677-685.

Monière D., Labbé C., Labbé D. (2005). Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, 4, p.79-104.

Monière D., Labbé D. (2012). Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs. Dister A. et al. (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, p.737-751.

Zipf G. K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Faire émerger les traces d'une pratique imitative dans la presse de tranchées à l'aide des outils textométriques

Cyrielle Montrichard
ELLIADD, UBFC – cyrielle.montrichard@edu.univ-fcomte.fr

Abstract
The main goal of this paper is to show how textometric tools can help to reveal the imitative usage of genres. During the Great War, soldiers must not criticize the hierarchy or the governement. Trench press is written by and for French soldiers in which we can find a great number of media and literary genres. Plus, we assume that writers use a number of discursive schemes to implicitly tell their point of view on the war, the governement and the « sacred union » discours which has become the mainstream speech in the public space in the early begining of the war. Therefore a corpus of this press seems to be the perfect place to search the notion of imitative usage of genres. To put into perspective the results given by the textometric tools we use a sample corpus from the national french press.

Résumé
L'objectif de cette contribution est d'interroger la pratique imitative des genres médiatiques et littéraires. Pour ce faire, nous mobilisons un corpus de presse de tranchées dans lequel se déploient de nombreux genres et sous-genres. Portant notre attention tout particulièrement sur les genres des dépêches et du roman-feuilleton nous montrons, en comparant ce corpus à un corpus échantillon de textes parus dans la presse quotidienne nationale en quoi la presse de tranchées copie les genres instaurés dans la presse civile. La seconde partie interroge le corpus au niveau syntagmatique pour tenter de faire émerger les registres ludiques et satiriques ayant court dans cette presse.

Keywords : presse écrite, genre, pratique imitative, première guerre mondiale, presse de tranchées.

1. Introduction

La presse de tranchées est un type de document né pendant la première guerre mondiale. Cette presse a la particularité d'être écrite *par* et *pour* les combattants (Audoin-Rouzeau, 1986). La censure ainsi que le discours

doxique d'union sacrée tenant place dans l'espace publique durant la période du conflit ne permettent pas aux locuteurs d'exprimer ouvertement leur opinion (Forcade, 2016). L'objectif de cette communication est de montrer comment émergent les registres ludiques et satiriques dans la presse de tranchées à travers l'inscription de discours dans des genres faisant écho à la matrice générique médiatique et littéraire.

Comment repérer à l'aide des outils textométriques les traces discursives d'une pratique imitative des genres médiatiques et littéraires dans la presse de tranchées ?

Cette communication vise à interroger la « pratique imitative » c'est-à-dire les « différentes formes ou genres qui permettent à un auteur de produire un texte (T2) attribué, sérieusement ou non, et de manière plus ou moins explicite, au modèle dont il s'est inspiré (T1) » (Aron, 2013). Pour ce faire, nous avons réuni en corpus cinq titres de presse de tranchées au format XML-TEI pour plus de 500 000 occurrences permettant une analyse du discours outillée.

À l'aide des outils textométriques et de la plateforme TXM (Heiden *et al.*, 2010) nous proposons de montrer comment les textes s'inscrivent et reprennent les codes établis des genres médiatiques et littéraires. Ensuite, nous proposons des pistes d'analyse visant à faire émerger le registre ludique ou satirique usité par les rédacteurs pour détourner le genre.

2. Contexte de la recherche et présentation du corpus

Notre étude propose d'investir la notion de pratique imitative. Cette dernière est proche de l'hypertextualité et de l'imitation (Genette, 1982) c'est-à-dire la reproduction d'un style, d'une manière. En analyse du discours, D. Maingueneau (1984) a investi la notion de pastiche, confirmant que celui-ci peut s'opérer sur un genre. Mais le pastiche pour G. Genette (1982) est associé principalement à une fonction ludique et dans le cadre de notre étude, la question entre registre satirique et registre ludique reste ouverte, c'est pourquoi nous nous cantonnerons donc à la notion de « pratique imitative ». Il n'existe, à notre connaissance, pas de travaux visant à interroger la pratique imitative en analyse du discours outillée. Xavier Garnerin (2009), pasticheur, tente de déterminer les méthodes des pasticheurs qui se situent selon lui « entre analyse et intuition » ce qui dénote toute la difficulté pour le chercheur à mettre au jour de façon systématique les liens unissant un texte T2 imitant un texte T1. Nous proposons de mettre à l'épreuve les outils textométriques pour tenter de percevoir la pratique imitative des genres.

Notre corpus se compose de cinq titres de presse de tranchées parus entre 1915 et 1918. Nous avons mis en place des variables permettant d'investir les

genres et les sous-genres (Rastier et Malrieu, 2002). La variable genre scinde le corpus en deux parties : le genre littéraire (287184 occurrences, 747 articles) et le genre médiatique (216534 occurrences pour 1005 articles).
Afin d'opérer une étude fine, nous avons aussi catégorisé les textes en sous-genre permettant ainsi de distinguer les romans-feuilletons, les nouvelles, les poèmes, etc., au sein du genre littéraire et les brèves, filets, dépêches, échos, faits divers, etc. dans le genre médiatique. L'espace de la contribution ne nous permet pas d'analyser chacun de ces sous-genres de façon particulière, c'est pourquoi nous concentrons notre étude sur un sous-genre littéraire, le roman-feuilleton et un sous-genre médiatique, la dépêche. Afin de mettre en perspective les résultats obtenus, nous avons constitué un corpus échantillon donnant à voir 38 dépêches parus entre 1915 et 1918 dans deux quotidiens nationaux (*Le Petit Journal* et *Le Matin*) et trois romans-feuilletons[1]. Ce corpus échantillon sera principalement mis à profit pour observer les constructions syntaxiques et la place des catégories morphosyntaxiques dans les deux sous-genres. Ainsi, la taille des effectifs n'est pas déterminante.

3. L'ancrage dans les moules discursifs médiatiques et littéraires

Dans cette partie, nous montrons comment les textes reprennent les codes établis dans la presse et dans la littérature à travers l'étude des catégories morphosyntaxiques et du lexique.

3.1. Les catégories morphosyntaxiques

Le graphique AFC ci-dessous donne à voir la distribution des catégories morphosyntaxiques (point-ligne en bleu) dans le sous-corpus du genre littéraire partitionné en sous-genres (point-colonne en rouge). On remarque, dans cette représentation graphique, que l'axe 1 contribue pour 60,63% à la structure du graphique. Cet axe semble structuré par le temps des verbes. En effet, à gauche du graphique on trouve les verbes au présent et au futur alors qu'à droite, on retrouve les temps du passé (passé simple, imparfait). On remarque que le roman-feuilleton se situe du côté des verbes au passé, respectant ainsi les caractéristiques du genre usant des temps du récit.
De plus, si l'on regarde la distribution des verbes en pourcentage dans la presse de tranchées et la Presse Quotidienne Nationale (PQN), on repère la proximité dans les temps employés.

[1] *Entre deux âmes* (1912) de Delly paru dans *L'Echo de Paris, Le Château noir* (1914) et *Confitou* (1916) de Gaston Leroux parus dans *Le Matin.*

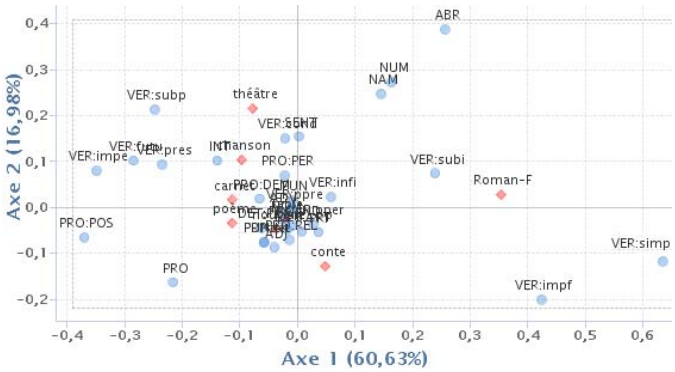


Figure 1. AFC des catégories morphosyntaxique du sous-corpus littéraire partitionné en sous-genre dans le corpus de presse de tranchées.

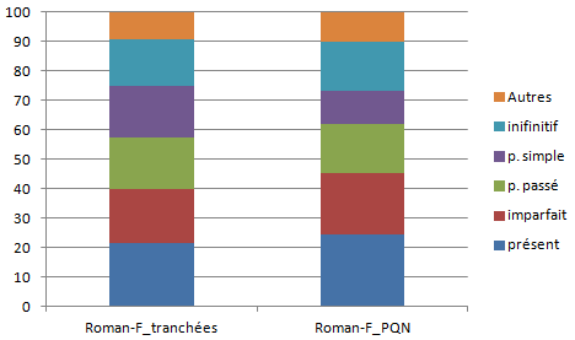


Figure 2. Graphique représentant pour cent verbes les temps utilisés dans les romans feuilletons parus dans la presse de tranchées (à gauche) et ceux parus dans la PQN (à droite)

Du côté du genre médiatique, le calcul des spécificités sur les catégories morphosyntaxiques indique que les dépêches dévoilent un score positif pour les noms communs (2) alors que les adverbes et les pronoms personnels sont en sous-emploi (respectivement des scores de -5,4 et -8,7). Ces résultats sont à mettre en lien direct avec les caractéristiques de la dépêche :

> [..] l'auteur de la dépêche se plie à un modèle de représentation qui doit faire l'économie des ressources stylistiques propres au littéraire : ni dialogue, ni focalisation interne, ni commentaire sur l'évènement rapporté. (Kalifa *et al.*, 2011 : 738)

On comprend ainsi le sous-emploi des adverbes et des pronoms personnels, souvent usités pour introduire un commentaire, alors que l'objectivation de l'information et l'effacement énonciatif préfèrent les catégories nominales

aux catégories verbales (Rabatel, 2004). D'ailleurs, on observe sur le graphique ci-dessous une proximité dans l'emploi des catégories morphosyntaxiques entre les dépêches de la presse de tranchées et celles de la PQN.

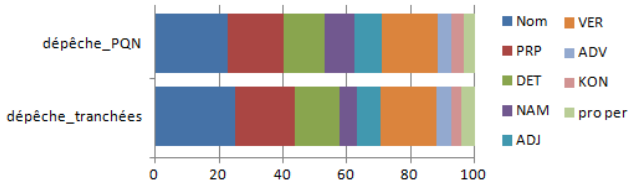


Figure 3. Graphique qui montre la proportion des grandes catégories morphosyntaxiques utilisées dans les dépêches parues dans la presse de tranchées (en bas) et celles parues dans la PQN (en haut)

L'observation de la ventilation des catégories morphosyntaxiques laisse entrevoir que presse civile et presse de tranchées usent des mêmes catégories morphosyntaxiques selon les genres.

3.2. Le lexique et les segments répétés

Dans la presse du début XXème, la dépêche débute souvent par une ligne indiquant le lieu et le jour de l'évènement. Les dépêches de notre corpus de presse de tranchées suivent cette règle et reprennent cette mise en scène de l'information. On le voit à travers de nombreux noms de lieux en spécificité positive comme : « Londres » (4,9), « Paris » (4,2), « Berlin » (2,3), etc. Les dépêches de la PQN confirment cette tendance avec une moyenne de 4 noms de lieux par article.

> L'escamotage de l'auteur passe d'abord par la mise au point d'un système d'énonciation à double détente : soit la source de l'évènement est indiquée – renvoyant toujours à un point de vue neutre – soit l'évènement est rapporté directement, sans mention manifeste de la source. (Kalifa *et al.*, 2011 : 738)

Les combattants improvisés journalistes mentionnent souvent une source que l'on peut percevoir à travers le suremploi des formes graphiques « communiqué » (score de 16,5) ou « dépêche » (score de 2). De plus, lorsque l'on s'intéresse aux segments répétés, on remarque que 7 dépêches de l'*Argonnaute* débutent par « Communiqué officiel de l'intérieur téléphoné par […] ». Du côté de la presse civile on retrouve les formes « dépêche » et « annonce » justifiant respectivement de 9 et 6 occurrences ainsi qu' « Havas » (17 occurrences). Pour le roman-feuilleton dans la presse de

tranchées, on repère des termes indiquant là aussi le respect de la mise en scène du roman en « chapitre » (score de 49) et le format feuilleton avec les termes « suite » (score de 37,4) et « suivre » (22,4).

4. Repérer la pratique imitative

À ce stade de notre étude, nous avons montré la proximité entre presse de tranchées et PQN mais ni l'étude des catégories grammaticales ni l'étude lexicale n'a permis de mettre au jour les registres ludiques et/ou satiriques signes d'une imitation et non d'une inscription dans le genre. Fort de ce constat, il apparaît nécessaire d'effectuer des recherches qui soient plus larges que celles du lemme mais plus précise que celles menées jusqu'alors sur les catégories morphosyntaxiques. Dès lors, une recherche au niveau syntagmatique semble s'imposer.

4.1. Constructions syntaxiques en suremploi pour les dépêches

Nous avons effectué des recherches pour obtenir les constructions syntaxiques enchaînant deux catégories morphosyntaxiques sur l'ensemble du corpus partitionné en sous-genre. Les résultats des premiers syntagmes en spécificité positive confirment ce que nous avons déjà pu voir : la catégorie préposition suivie d'un nom propre présente un score de +10,3 et un retour au texte confirme qu'il s'agit de la présentation du lieu de l'évènement (« à Londres », « de Paris », etc.). Aussi, on trouve une construction syntaxique qui induit une construction passive (verbe au présent suivi d'un verbe au participe passé) indiquant encore l'effacement énonciatif (Rabatel, 2004). Dans la liste des spécificités positives nous trouvons la combinaison nom suivi d'adjectif (score de +2,3). La liste éditée donne à voir 74 syntagmes. Quatorze d'entre eux (soit 19%) ont attiré notre attention de part, soit l'invraisemblance du dire (« homme volant », « provision inépuisable »), soit parce que leur présence ne fait pas sens dans le genre dans lequel ils se déploient (« bicyclette usagée », « cellules nerveuses », « chauffage central », « crayon ennemi »). À noter le syntagme « agence Ivile » jouant de l'homonymie avec « agence civile ». Le retour au texte permet de mieux comprendre l'usage de ces syntagmes par les rédacteurs jouant souvent sur le double sens des mots.

> Plusieurs saucisses boches (de Francfort) ont été capturées à la devanture d'un charcutier par un audacieux **homme volant**. (*Argonnaute,* 15 mars 1916)

Le syntagme « saucisses boches » peut renvoyer en 1916 à deux signifiés : le produit de charcuterie ou le projectile ennemi. C'est sur cette ambiguïté qu'est basée l'énoncé accentuée par la présence du nom « charcutier » et du

participe passé « capturées » qui indique chacun une possibilité d'interprétation différente. Enfin, l'« homme volant » peut être entendu comme un briguant ayant dérobé de la charcuterie où un homme ayant la capacité de voler dans les airs et ayant capturé les projectiles ennemis avant l'impact. Cet exemple dévoile comment les rédacteurs par un registre ludique créent de la connivence avec le lecteur qui partage les mêmes références. Un autre exemple permet d'introduire l'idée d'un registre satirique avec la critique du discours dominant dans l'espace publique.

> […]Paris, 31 avril
> […]Rue du Paon-Blanc (14h.) Paris gronde. Le régime a vécu. Vive la révolution ! Les bains de la Samaritaine sont en état de siège. Le syndicat de la Grande Presse n'autorise plus que la parution d'un bulletin relatant le Communiqué. La censure s'est tranchée la gorge avec ses ciseaux. **L'héroïsme sacré** fait battre les cœurs.[…] C'est **l'union sacrée**. Concierges, locataires et propriétaires s'embrassent aux portes des immeubles. (*Rigolboche,* 10 mai 1917)

L'article remet ici en cause la censure, les festivités parisiennes et fait également écho aux désaccords entre les propriétaires et les locataires mobilisés remettant ainsi en cause le discours d'union sacrée tout en réinvestissant ses dires (Authier-Revuz, 1984). La recherche de syntagmes nous permet donc d'entrer dans le corpus au niveau du texte et de percevoir ce qui dans les articles semblent détourner le genre à des fins ludiques et satiriques.

4.2. Construction syntaxique en suremploi pour le roman-feuilleton

Le roman-feuilleton tient une place importante dans la presse du XIXème siècle et du XXème (Kalifa *et al,* 2011). Le conflit ne modifie pas la place de cette fiction.

> La guerre pénètre très rapidement dans le « rez-de-chaussée », et le roman-feuilleton, sous la forme de récits patriotiques, se mue en instrument destiné à entretenir et intensifier la mobilisation de la population en faveur de l'effort de guerre. (Erbs, 2016 : 740)

Voici ce qui est donné à lire aux combattants qui reçoivent et lisent la presse civile (Gilles, 2013). Nous avons, comme pour les dépêches tenter d'effectuer une recherche sur les syntagmes de deux occurrences à travers les spécificités selon les catégories grammaticales. Ces recherches n'ont pas été fructueuses pour le roman feuilleton. Nous avons donc étendu la recherche à trois

occurrences. La construction syntagmatique « verbe au passé simple + déterminant + nom » avec un score de +52 a attiré notre attention. Sur les 130 syntagmes, 24 nous ont interpellés, soit 14% d'entre eux. D'abord, nous avons repéré des syntagmes qui semblent construits sur des expressions figées mais où l'un des termes a été modifié comme « **fouilla** l'horizon » ou « **coupa** la pipe ». Nous avons aussi repéré des syntagmes qui ne semblent pas faire sens comme « revêtit l'ampleur » ou « trancha les jours ».

> Alors une colère terrible parut animer l'Armada toute entière. Proue baissée, les navires foncèrent sur le pirate boche ... Cependant une première torpille alla frôler par bâbord le vaisseau amiral ; une deuxième, lancée trop haut, **coupa la pipe** du commandant qui flegmatiquement, sortit d'un étui une cigarette qu'il ajusta au tuyau mutilé de sa pipe. […] (« Krotufex », *Rigolboche* 10/12/1917)

La torpille coupe littéralement la pipe du commandant alors qu'on aurait pu s'attendre à ce que ce dernier *casse sa pipe* dans un tel contexte. Cela renvoie au registre ludique avec le jeu sur l'expression figée mais certainement aussi au registre satirique offrant ici une critique des romans-feuilletons patriotiques décrivant des batailles sanglantes sans jamais que le héros ne succombe. En étudiant les mêmes syntagmes dans le sous-corpus roman-feuilleton dans la PQN, on repère la présence abondante de noms renvoyant à une partie du corps (« leva les yeux », « prit la main », « secoua la tête », « tendit la main ») : sur les dix premiers syntagmes six ont cette caractéristique. On observe également la présence du corps dans ces syntagmes dans la presse de tranchées mais ceux-ci semblent une fois encore surréalistes et usés à des fins ludiques, copiant le genre en le détournant : « cala les joues », « déchaussa son pied », « frotta la mandibule », « tomba le torse », etc.

5. Conclusion

Notre contribution avait pour objectif d'investir la pratique imitative avec les outils textométriques sur un corpus singulier de presse de tranchées mis en perspective avec un corpus échantillon issu de la PQN. Nous avons pu montrer dans un premier temps comment les genres sont imités en reprenant les codes établis dans la presse civile. Pour faire émerger les traces d'une pratique imitative, il nous a semblé nécessaire d'interroger le corpus, à l'aide du logiciel textométrique TXM, au niveau syntagmatique. Cette recherche a, dans le cas de notre étude, permis de faire émerger les registres ludiques et satiriques ayant court dans la presse de tranchées. Cette presse est un lieu

énonciatif où l'implicite et la connivence tiennent une place importante au vue de la censure mais aussi des liens particuliers qui unissent lecteurs et rédacteurs. Il serait intéressant de voir si, en usant de la même méthodologie, sur des textes et des genres différents, des résultats similaires peuvent être observés.

Références

Aron, P. (2013). Le pastiche et la parodie, instruments de mesure des échanges littéraires internationaux. In Gauvin, L. dir., *Littératures francophones : Parodies, pastiches, réécritures*. ENS Éditions.

Audoin-Rouzeau, S. (1986). *14-18, les combattants des tranchées : à travers leurs journaux*. A. Colin.

Authier-Revuz, J. (1984). Hétérogénéité(s) énonciatives. *Langages,* vol.(73) : 98-111.

Erbs, D. (2016). *Le roman-feuilleton français et le serial britannique pendant le premier conflit mondial, 1912-1920.* (thèse de doctorat).

Forcade, O. (2016). *La censure en France pendant la Grande guerre.* Fayard.

Garnerin, X (2009). Le pastiche, entre intuition et analyse. *Modèles linguistiques*, vol.(60): 77-91.

Genette, G. (1982). *Palimpsestes.* Seuil.

Gilles, B. (2013). *Lectures de poilus: livres et journaux dans les tranchées, 1914-1918*. Ed. Autrement.

Heiden, S., Magué, J-P. and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio B. *et al.* editors, *Proc. of JADT 2010 (10th International Conference on the Statistical Analysis of Textual Data)*, pp. 1021-1032.

Kalifa, D., Régnier, P., Thérenty, M.-E. *et al.* (2011). *La civilisation du journal : histoire culturelle et littéraire de la presse française au XIXème siècle.* Nouveau monde éditions.

Maingueneau, D. (1984). *Genèses du discours.* Madraga.

Malrieu, D & Rastier, F. (2002). Genres et variations morphosyntaxiques. *Traitement automatique des langues* vol.(42) : 548-577.

Rabatel, A. (2004). Effacement énonciatif et effets argumentatifs indirects dans l'incipit du Mort qu'il faut de Semprun. *Semen*, vol.(17) : 111-148.

Evolución diacrónica de la terminología y la fraseología jurídico-administrativa en los Estatutos de autonomía de Catalunya de 1932, 1979 y 2006

Albert Morales Moreno
Università Ca' Foscari Venezia / Université de Genève – albert.morales@unige.ch

Abstract
During the first half of 2017, research was carried out at the Institut de Lingüística Aplicada of the Universitat Pompeu Fabra thanks to a grant from the Generalitat de Catalunya's Institut d'Estudis de l'Autogovern in order to study diachronically the Statutes of Autonomy of Catalonia (EAC acronym, in Spanish) approved in 1932, 1979 and 2006.
As in other countries and traditions, the negotiation of such an important law is a challenge in the historical moment in which it occurs, both in legal and political terms (see Abelló (2007) for the EAC of 1932, Sobrequés (2010) for the 1979 EAC and Serrano (2008) for the 2006 EAC).
We take lexicometrics as an analytical methodology and the communicative theory of terminology (Cabré, 1999) as the grounds for our research to study the use of legal and administrative terminology with respect to the assignment of competences from a diachronic approach. Specifically, we are interested in combining the study of repeated segments and the study of specificities to identify the terms, positions and key institutions of each EAC, as well as the use of some locutions between 1932 and 2006 in Catalan statutory discourse.

Resumen
Durante la primera mitad de 2017, se desarrolló una investigación en el Institut de Lingüística Aplicada de la Universitat Pompeu Fabra para el Institut d'Estudis de l'Autogovern de la Generalitat de Catalunya (EAC) para estudiar diacrónicamente los diferentes Estatutos de Autonomía de Cataluña (EAC), aprobados en 1932, 1979 y 2006.
Al igual que en otros países y tradiciones, la negociación de los proyectos de regulación de esta escala es un reto en el momento histórico en que ocurre, tanto en términos legales y políticos (Abelló (2007) para el EAC de 1932, Sobrequés (2010) para la de 1979 y Serrano (2008) para el de 2006Partimos de la lexicometría como metodología analítica y de la teoría comunicativa de la terminología (Cabré, 1999) para estudiar el uso de la terminología jurídica y administrativa con respecto a la asignación de competencias y materiales a

partir de un enfoque diacrónico. En concreto, nos interesa combinar el estudio de segmentos repetidos con el estudio de especificidades para identificar los términos, cargos e instituciones clave de cada EAC, así como el uso de algunas locuciones entre 1932 y 2006 en el discurso estatutario catalán.
Keywords: discourse analysis, legal discourse, Catalan statute of autonomy, repeated segments, terminology, diachronic analysis

1. Introducción

El presente artículo presenta un estudio enmarcado dentro de un proyecto más amplio de análisis diacrónico de la redacción normativa en catalán. En dicha investigación, realizada gracias a una financiación posdoctoral del Institut d'Estudis de l'Autogovern de la Generalitat de Catalunya, se han estudiado los Estatutos de Autonomía de Catalunya (EAC) de 1932, 1979 y 2006 se han llevado a cabo estudios lexicológicos, estadísticos, terminológicos, traductológicos y pragmáticos de los distintos EAC.

En esta en concreto, nos hemos centrado a estudiar, desde un punto de vista terminológico, los segmentos repetidos para evaluar si esta es una estrategia válida para identificar la evolución de la fraseología especializada relativa a un ámbito especializado como el Derecho a través del estudio de los segmentos repetidos específicos de cada EAC. Asimismo, nos proponemos comparar dichas unidades para ver cuál ha sido la evolución, desde un punto de vista diacrónico.

Así pues, después de un exhaustivo estudio lexicométrico del corpus, hemos seleccionado unidades terminológicas especializadas (UTE) relativas al ámbito jurídico-administrativo que contribuyen a establecer las competencias de Catalunya en los diferentes EAC, con términos como *competència/es*, *correspon* o *atribucioó/ons*.

Para dicho análisis, hemos partido de los índices estadísticos que ha arrojado la exploración lexicométrica desarrollada con *Lexico3.6* y como marco teórico hemos empleado la Teoría Comunicativa de la Terminología (Cabré *et al.* 1999).

2. Los EAC de 1932, 1979 y 2006

En primer lugar, cabe definir el estatuto de autonomía como una unidad relativa al ámbito del derecho constitucional que se define como la "norma institucional básica de las comunidades autónomas" (*Diccionario del español jurídico* (DEJ), Real Academia Española).

Numerosos juristas reconocen funcionalmente al EA de las comunidades como "equivalente a la constitución de un estado miembro de una federación, porque regula las instituciones autonómicas, establece las

competencias que deben tener y no puede ser modificado por ninguna otra ley, ni autonómica ni estatal: sólo puede reformarse por el procedimiento que el mismo Estatuto prevé, característica propia de las constituciones y no de las leyes" (Albertí, *et al.* 2002:111). El Estatuto, pues, "tiene rango de ley orgánica estatal, forma parte del bloque de la constitucionalidad y está sometido a unos procedimientos agravados de aprobación y reforma, y sus previsiones disfrutan de unas garantías reforzadas que no proporciona la legislación ordinaria" (Pons y Pla 2007:187). En Cataluña, a principios del siglo XX, con la Mancomunitat, comienza la recuperación del autogobierno. En el marco de dicha institución, se redacta un primer proyecto de Estatuto de autonomía, aunque este no se llega a debatir, "porque el 27 de febrero de 1919 se suspendían las sesiones parlamentarias como consecuencia de la huelga de la *Canadenca*" (Fontana 2014:327). Debido al desarrollo histórico convulso de los años posteriores y de la dictadura de Miguel Primo de Rivera, los proyectos autonomistas se paralizan. Hay que esperar a 1931, la República, para que se redacte el primer EAC. Aquel texto se debate en las Cortes en mayo de 1932. Abelló afirma que aquel texto prevé "la inserción de Cataluña en una república federal" (2007:35) y lo define como "moderado" (2007:44). A pesar de los recortes que sufre, "se convirtió en una herramienta útil, que, con la reconquista de las instituciones catalanas de autogobierno, facultaría una legislación propia, a pesar de que esta fuera limitada" (Abelló 2007:187). La Generalitat de Catalunya asume las competencias durante poco tiempo, y el 6 de octubre de 1936 el EAC 1932 se suspende parcialmente; con la llegada de las tropas franquistas a Cataluña, Franco aprueba la ley de derogación del EAC el 5 de abril de 1938. Con la dictadura de Franco, el Estado se concibe desde una óptica recentralizadora y, como ya se ha señalado, se abole la autonomía de las comunidades. Hay que esperar hasta la muerte del dictador, el 20 de noviembre de 1975, para que, según Sobrequés (2010: 11), España y Cataluña iniciaran el proceso que tenía que cambiar su historia: la Transición. Durante esta, se sella el pacto constitucional de 1978 (la Constitución entra en vigor el 29 de diciembre de ese año) y se construyen los cimientos jurídicos del Estado autonómico con un ordenamiento que, a través de los estatutos de autonomía –al menos desde un enfoque teórico–, se da a los gobiernos autonómicos bastante autogobierno. El proyecto de redacción comienza el el 8 de septiembre de 1978 y su texto final se aprueba en referéndum el 25 de octubre de 1979.

A principios del siglo XXI, sin embargo, un sector considerable del espectro social y político catalán percibe el EAC 1979 como un modelo sin recorrido (la conocida como *doctrina Argullol*, que supone releer de manera menos centralista la CE), pero rápidamente se comprueba "hay un número importante de competencias que, a pesar de ser incluidas en el Estatuto de

autonomía, no han sido objeto de desarrollo legislativo" (BOPC 2002:89). Por ese motivo, tras las elecciones autonómicas de 2003, la coalición tripartita integrada por PSC, ERC e ICV-EUiA inicia en 2004 la tramitación parlamentaria para la reforma estatutaria. Ello implica una primera negociación para que se aprobara en el Parlament de Catalunya el 30 de septiembre de 2005, y una segunda negociación para aprobarlo en las Cortes Generales (en esa segunda fase, tal y como se expone en Morales (2015), se producen los cambios más significativos).

El texto final se aprueba en sede parlamentaria el 10 de mayo de 2006, día en el que el Pleno del Senado aprueba el nuevo estatuto con 128 votos a favor, 125 en contra y 6 abstenciones. El 31 de julio de 2006, Federico Trillo-Figueroa y Martínez-Conde (junto con 98 diputados más del PP) presenta el 31 de julio de 2006 un recurso de inconstitucionalidad contra la mayoría de artículos del nuevo Estatuto (Bosch 2013: 44) porque, entre otras razones, "aplicaba el término nación en Cataluña, imponía el catalán, establecía una serie de derechos y deberes que restringían las libertades de los ciudadanos de Cataluña […] y cuestionaba la unidad de España" (Segura 2013: 217-218). El 28 de junio de 2010, el Tribunal Constitucional hace pública parte de la sentencia 31/2010 sobre la constitucionalidad del Estatuto, que declara inconstitucionales algunas de las partes del EAC 2006. Según numerosos politólogos e historiadores, esa fecha es clave para la historia política contemporánea porque "fue el día de la ruptura sentimental con España, el día en que [muchos catalanes] se convencieron de que Cataluña y los ciudadanos de Cataluña no tenían cabida en España" (Segura 2013: 32) y para muchos ciudadanos supuso el salto del autonomismo al independentismo, sin pasar por el nacionalismo (Segura 2013: 241).

El corpus constituido es, pues, representativo para estudiar diacrónicamente la evolución del discurso estatutario en lengua catalana a través de los diferentes Estatutos aprobados a lo largo de la Historia. Para concluir, cabe añadir que según André Salem (1991:149) este corpus se considera una "serie textual cronológica", puesto que son textos lingüística y pragmáticamente comparables de un arco temporal que permite extraer conclusiones sobre la evolución del discurso estatutario en lengua catalana de los últimos ochenta años.

3. Marco teórico y metodológico

Desde la restauración de las instituciones de autogobierno, ha habido numerosas iniciativas, tanto públicas como privadas, de modernización del discurso normativo catalán. Cabe destacar el trabajo del Grupo de Estudios de Técnica Legislativa (GRETEL), de la Dirección General de Política Lingüística, del TERMCAT, de la Escuela de Administración Pública de

Cataluña o del Parlament de Catalunya. El modelo que se sigue es el de Québec, adoptando –y adaptando– las directrices de Spar y Schwab *Rédaction des lois: rendez-vous du droit et de la culture*. Según Montolío, se aprovecha para renovar dicha tradición:

> Un caso especial lo constituyen las otras lenguas oficiales del Estado español (gallego, vasco y catalán). Para estas tres lenguas, la renovación del lenguaje jurídico ha venido impulsada por una motivación adicional: la voluntad de recrear una tradición jurídica truncada tras cuarenta años de prohibición. Entre ellas, cabe destacar la renovación del lenguaje jurídico catalán.
> (Montolío y Albertí 2012:99)

Por ese motivo, los criterios y principios de la que parte la normalización del lenguaje jurídico catalán son el de economía, el de claridad y el de precisión en la expresión (DGPL 1999: 7). La falta de estudios lingüísticos exhaustivos de un componente esencial del discurso normativo catalán como es su Estatuto de autonomía, ha motivado este trabajo. Este trabajo nace de la necesidad de analizar combinando la estadística textual y el análisis del discurso, con una perspectiva diacrónica, los diferentes EAC que ha habido en vigor hasta la fecha, a partir de una disciplina consolidada: la Lingüística de Corpus.

De acuerdo con la revisión presentada en Morales (2015:101-175), se han empleado dichas metodologías para estudiar textos similares. Para garantizar una selección de las unidades análisis objetiva, pertinente y representativa basada en criterios estadísticos, nuestro trabajo se desarrolla a partir de la lexicometría. Dicha escuela ha permitido caracterizar, entre otros, el vocabulario de personajes sociopolíticos, y de movimientos sociales e históricos.

Dentro de la lexicometría, nuestra aproximación parte de una aproximación lexicométrica formalista, puesto que nuestra unidad básica de análisis es la *forma*. Posteriormente, hemos normalizado el texto (a partir de metodologías como las de Arnold (2008:110) y Menuet (2006:157)) para corregir las formas con errores (gramaticales o de escritura) y evitar que haya conteos duplicados debido a diferencias mínimas en la ortotipografía. Por último, hemos insertado en nuestro corpus las marcas estructurales requeridas por *Lexico3.6* para identificar los diferentes EAC.

De las múltiples funcionalidades que incluye el programa, han arrojado resultados especialmente interesantes el estudio de las concordancias, de los segmentos repetidos y de especificidades.

Tras la primera exploración lexicométrica, hemos analizado algunos términos

clave identificados con el análisis de segmentos repetidos para ver si nos permite caracterizar la fraseología y terminologías propias del ámbito.

4. Análisis

El corpus analizado presenta las principales características lexicométricas siguientes[1]:

Identificador	Documento	Ocurrencias		Formas	Hápax
01_1932	EAC 1932	4.242	(7,7 %)	1.009	606
02_1979	EAC 1979	10.580	(19,3 %)	1.766	935
03_2006	EAC 2006	40.011	(73,0 %)	3.457	1.546
		54.833	(100 %)	4.226	1.804

Esta parte del análisis se centra en analizar los ya señalados segmentos repetidos (SR), es decir, las secuencias de formas repetidas con una frecuencia superior a 5.

La exploración lexicométrica ha arrojado 2.398 segmentos repetidos, pero nos centraremos en algunos de los más significativos. Su distribución en relación con su longitud es:

Longitud	Secuencias	Ejemplos
2	1282	*de Barcelona*
		les llibertats
		la coordinació
3	660	*de la Constitució*
		de seguretat pública
		en aquest Estatut
4	281	*les lleis de Catalunya*
		a les Corts Generals
		el president o presidenta
5	98	*de conformitat amb les lleis*
		els poders públics han de
6	31	*correspon a la generalitat la competència*
		d ' acord amb allò que
7	23	*sens perjudici d ' allò que disposa*
		el president o presidenta de la generalitat
8	10	*els poders públics han de vetllar per la*
		impost sobre la renda de les persones físiques

[1] Debido a las diferencias de tamaño obvias, aplicamos, gracias a la profesora Arjuna Tuzzi, técnicas de análisis estadístico que las tienen en cuenta a la hora de hacer los cálculos de representatividad y selección esperados, a partir de, entre otros Tuzzi (2003:128-129) o Van Gijsels, Speelman, y Geeraerts (2005:1).

Longitud	Secuencias	Ejemplos
9	7	*en una votació final sobre el conjunt del text*
		en el diari oficial de la generalitat de Catalunya
10	4	*correspon a la generalitat la competència exclusiva en matèria de*
11	11	*de l ' apartat 1 de l ' article 149 de*
		la carta dels drets i els deures dels ciutadans de Catalunya

De las 20 más frecuentes, por ejemplo, solo cinco tenían interés para nuestro estudio lingüístico en tanto que unidades con semántica plena, como *la Generalitat, de Catalunya* o *la competència*.

Además de aislar segmentos como *de les quals* (10), *els altres* (23), *la resta* (17), *les quals* (18) o *en el termini* (25) o *la seva* (57) –que podrían ser interesantes para investigaciones estilométricas o de atribución de autoría–, a continuación analizamos algunas de las unidades con una frecuencia superior.

El sistema ha permitido identificar, por ejemplo, algunos sintagmas relativos a cargos e instituciones previstos estatutariamente, como *les Corts* (46) (y *les Corts Generals* (33)), *Poder Judicial* (46), la *Comissió Mixta d'Afers Econòmics i Fiscals Estat-Generalitat* (14), l'*Agència Tributària de Catalunya* (10), el *Consell de Justícia de Catalunya* (19), *el Govern* (50), *el President* (38), el *President o Presidenta de la Generalitat* (26), *la Unió Europea* (31) i *el Parlament de Catalunya* (24). Ha dado buenos resultados, pues, para identificar sintagmas relativos a unidades muy lexicalizadas como cargos o instituciones.

Uno de los SR más frecuentes es *correspon a la Generalitat*. Dicho segmento presenta la distribución siguiente en el corpus:

SR: correspon a la Generalitat	**EAC 1932**	**EAC 1979**	**EAC 2006**
FA	1	9	144
FR (x10000)	4,4	8,5	36,0

Su uso es, como se constata, paradigmático del EAC 2006 (E+11) (presenta especificidad negativa en los EAC 1932 (E-05) y EAC 1979 (E-07)) y, tal y como se expone en Morales (2018, en prensa), el ámbito de la atribución competencial (de la que el segmento repetido es una de las expresiones lingüísticas más características, al menos en la redacción estatutaria contemporánea) es de las que más singularidades presenta en el EAC 2006 y que más cambios ha presentado en el corpus estudiado desde un punto de vista diacrónico.

Otro de los SR más frecuentes (105 ocurrencias) es *la Constitució*, que se reparte de la manera siguiente:

SR: la Constitució	**EAC 1932**	**EAC 1979**	**EAC 2006**
FA	17	42	46
FR (x10000)	40,1	39,7	11,5

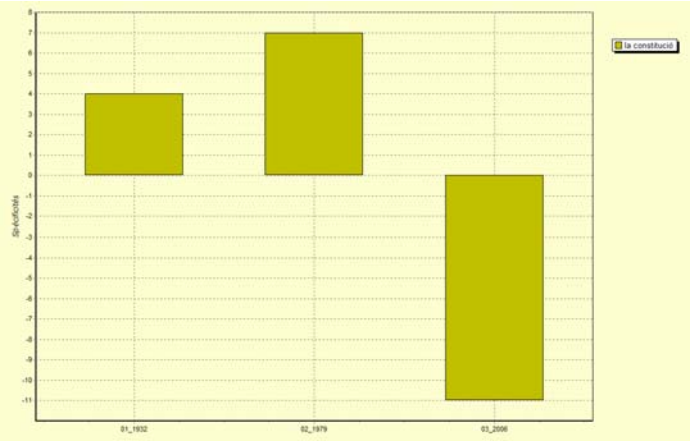
Partie : 01_1932, Nombre de contextes : 17

ix en regió autònoma dintre de l ' estat espanyol , de conformitat amb la constitució de la república i el present estatut . el seu org
tèria civil , exceptuant allò que disposa l ' article 15 , número 1 de la constitució , i l ' administrativa que li sigui plenament atr
ibunal de garanties constitucionals , d ' acord amb l ' article 121 de la constitució . el tribunal de garanties constitucionals , si h
que està disposat en els números , 4 , 10 , i 16 de l ' article 14 de la constitució , queden reservats a l ' estat tots els serveis de
a fluvial , no contradient allò que està disposat en l ' article 14 de la constitució . les mancomunitats hidrogràfiques el radi d ' ac
te a les lleis socials , feta en el paràgraf 1r . de l ' article 15 de la constitució . § article 13 . la generalitat de catalunya pren
agrària , salvat el que disposa el paràgraf 5è . de l ' article 15 de la constitució i la reserva sobre lleis socials consignada en el
, exceptuant el que està disposat en el número 7è de l ' article 15 de la constitució . § e) l ' establiment i ordenació de centres de
res públiques de catalunya , salvat allò que disposa l ' article 15 de la constitució . § b) els serveis forestals , els agronòmics i
lunya § article 5 . de conformitat amb el que preveu l ' article 15 de la constitució , la generalitat executarà la legislació de l ' e
reserà a catalunya , tenint en compte allò que ordena l ' article 20 de la constitució , una junta de seguretat formada per representant
les proves i els requisits que , de conformitat amb l ' article 49 de la constitució , estableixi l ' estat per a l ' expedició de tí
oportuns , sempre subjectant - se a allò que disposa l ' article 50 de la constitució , amb independència de les institucions docents i
garanties , en l ' ordre civil i el criminal , per les infraccions de la constitució , de l ' estatut i de les lleis . § article 15 .
en la militar i la de l ' armada , de conformitat amb els preceptes de la constitució i de les lleis processals i orgàniques de l ' est
n el funcionament d ' aquests organismes , d ' acord amb l ' estatut i la constitució . el parlament , que exercirà les funcions legisl
ri català . § article 3 . els drets individuals són els establerts per la constitució de la república espanyola . la generalitat de cat

Partie : 02_1979, Nombre de contextes : 42

de catalunya són titulars dels drets i deures fonamentals establerts a la constitució . § 2 . correspon a la generalitat , com a poder
al seu autogovern , es constitueix en comunitat autònoma d ' acord amb la constitució i amb el present estatut , que és la seva norma i
unya ordenaran el funcionament d ' aquestes institucions d ' acord amb la constitució i el present estatut . § capítol primer . el parl
i les altres funcions que els encarrega directament l ' article 104 de la constitució i les que els atribueixi la llei orgànica que el
el ministeri fiscal en les funcions especificades a l ' article 126 de la constitució i en els termes que disposin les lleis procesals
e les altres facultats previstes a l ' apartat 2 de l ' article 129 de la constitució . § títol quart . reforma de l ' estatut § articl
er ús de les facultats previstes a l ' apartat 1 de l ' article 130 de la constitució , i podrà fomentar mitjançant una legislació adeq
ulades , sens perjudici d ' allò que disposen els articles 36 i 139 de la constitució . § 24 . fundacions i associacions de caràcter do
estat , de conformitat amb allò que preveuen els articles 137 i 141 de la constitució . § article 6 § 1 . als efectes del present estat
respectant l ' autonomia que els reconeixen els articles 140 i 142 de la constitució i d ' acord amb l ' article 9 . 8 d ' aquest esta
' allò que disposa el número 10 de l ' apartat 1 de l ' article 149 de la constitució . § 23 . col·legis professionals i exercici de le
38 , 131 i els números 11 i 13 de l ' apartat 1 de l ' article 149 de la constitució , la competència exclusiva de les matèries següen
' allò que disposa el número 15 de l ' apartat 1 de l ' article 149 de la constitució . les acadèmies que tinguin llur seu central a ca
' allò que disposa el número 16 de l ' apartat 1 de l ' article 149 de la constitució . § 20 . establiment i ordenació de centres de co
' allò que disposa el número 18 de l ' apartat 1 de l ' article 149 de la constitució . alteracions dels termes municipals i denominaci
e l ' article 92 i el número 18 de l ' apartat 1 de l ' article 149 de la constitució , i correspon a l ' estat l ' autorització de llu
ue disposen els números 20 i 21 de l ' apartat 1 de l ' article 149 de la constitució . centres de contractació i terminals de càrrega
a què es refereix el número 21 de l ' apartat 1 de l ' article 149 de la constitució , sens perjudici de l ' execució directa que es r
b allò que disposa el número 23 de l ' apartat 1 de l ' article 149 de la constitució . § 11 . higiene , tenint en compte allò que disp
allò que estableix el número 25 de l ' apartat 1 de l ' article 149 de la constitució . § 17 . pesca en aigües interiors , cria i recol
' allò que disposa el número 28 de l ' apartat 1 de l ' article 149 de la constitució . § 6 . arxius , biblioteques , museus , hemerote

En la mayoría de ocasiones se trata de contextos que hacen referencia a un artículo concreto de la CE 1978. Son fórmulas que sirven para restringir el alcance estatutario y establecer una remisión con la Carta Magna española. Es interesante señalar que el análisis de especificidades denota un uso específico positivo de dicho SR en los EAC 1932 (E+04) y EAC 1979 (E+07):



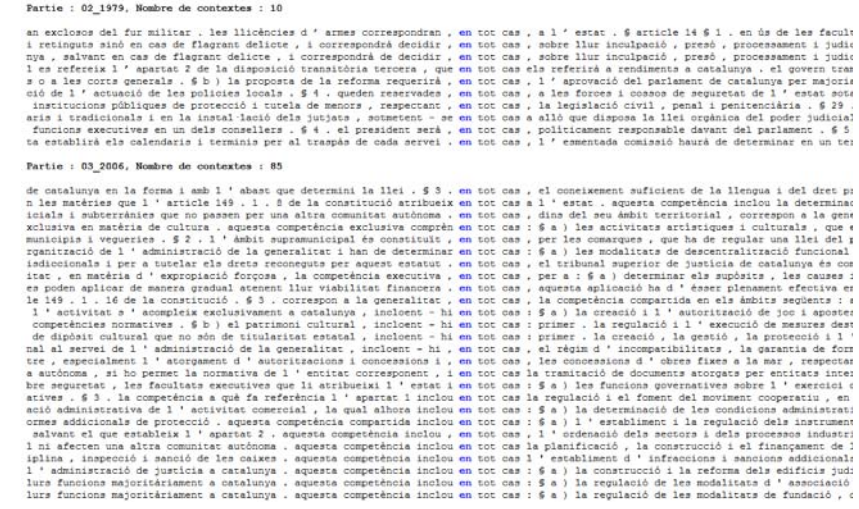
Otras remisiones legislativas que hemos identificado gracias al estudio de los segmentos repetidos han sido *aquest Estatut* (96), *l'article 149 (de la Constitució)* (26) o *el Títol V* del mismo EAC (12).

Al tratarse de un corpus legislativo, el análisis también ha permitido identificar como SR numerosas unidades pertenecientes al lenguaje jurídico-administrativo que se rigen según el patrón *determinante + sustantivo* o *sustantivo + adjetivo*, como *l'article, l'estatut, la legislació, una llei, llei orgànica, administracions públiques, l'administració, aquest article, comunitat autònoma, de catalunya, de seguretat, del règim jurídic, disposició addicional, domini públic, dret civil, el control, el foment, el règim, els àmbits, els articles, els deures, els mecanismes, els principis, els procediments, els processos, la llei, la llengua, la majoria, la normativa, la propietat, la salut, les activitats, les actuacions, les administracions, les administracions públiques, les comunitats, les empreses, les entitats, les iniciatives, les matèries, les normes, les organitzacions, les polítiques, les universitats, llei del parlament, polítiques públiques, règim jurídic, serveis públics, serveis socials, tributs estatals* y *una llei del parlament*.

El aspecto en el que el presente estudio ha proporcionado resultados más interesantes es, sin lugar a dudas, el relativo a las locuciones más empleadas en alguno de los EAC, y que en algunos casos se usan de manera especializada. Algunas de las unidades que hemos estudiado en profundidad han sido *en matèria de/d'*, *si escau*, *d'acord amb*, *en tot cas*, *en els termes que* o *sens perjudici*.

El SR *en tot cas* presenta especificidad en el EAC 2006. Su uso es especifico positivo del EAC 2006 (E+05) y negativo de los EAC 1932 (E-04) y EAC 1979 (E-03). Sus 95 ocurrencias se distribuyen de la manera siguiente:

SR: en tot cas	**EAC 1932**	**EAC 1979**	**EAC 2006**
FA	–	10	85
FR (x10000)	–	9,5	21,2
Esp	E-04	E-03	E+05



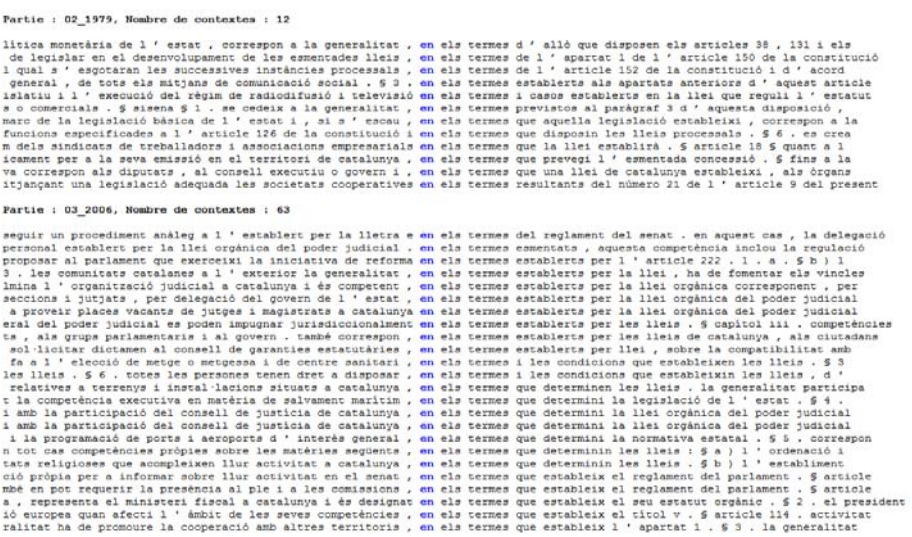
En la tesis (Morales 2015:398-400), se comprobó que esta es una cláusula bastante usada en el discurso estatutario catalán contemporáneo y describimos los usos de dicha cláusula. El libro de estilo del Parlament, una referencia básica para la redacción estatutaria contemporánea, la define así:

> en tot cas
>
> Locució adverbial, equivalent a *en qualsevol cas*, que es pot emprar amb valor concessiu o amb el sentit de 'en tots els casos'. Quan té aquest sentit, per raons de claredat i precisió, és preferible substituir-la per *sempre* o *en tots els casos* o, si escau, prescindir-ne.
>
> (SAL 2014:272)

Otra cláusula identificada con el análisis de SR es *en els termes*, que se distribuye en el corpus de la manera siguiente:

SR: en els termes	**EAC 1932**	**EAC 1979**	**EAC 2006**
FA	–	12	63
FR (x10000)	–	11,3	15,7
Esp	E-04	–	E+03

El análisis de especificidades indica que su uso es característico positivo en el EAC 2006, mientras que en los otros no presenta especificidad (EAC 1979) o bien presenta especificidad negativa (E-04, en el EAC 1932). Al leer detalladamente las concordancias, se comprueba que aparece sobre todo en contextos como *en els termes que disposin/determini/estableix* o similares (*en els termes establerts*…):



Cabe señalar que el EAC 1979 presenta más variedad en relación con el uso de esta cláusula (las 12 ocurrencias presentan 12 realizaciones diferentes), mientras que el EAC 2006 se constata menos variación; de los 63 contextos en los que aparece, las que acumulan más ocurrencias son:

- *en els termes que estableix/estableixen/estableixi/estableixin + [les lleis, la legislació…]*: 41

- *en els termes que determinin/determinen + [la llei orgànica, la legislació…]*: 7

Se comprueba, pues, una fijación más alta. Habría que analizar corpus más grandes para verificar esta hipótesis, pero esta tendencia a tener un discurso estatutario más fijado en el EAC 2006 parece confirmarse. Hemos constatado, sin embargo, que en la mayoría de segmentos repetidos se observa un comportamiento lingüístico diferente entre los EAC 1932 y 1979 por un lado, y el EAC 2006 por el otro. Por lo tanto, estos resultados confirman la hipótesis planteada inicialmente y confirmada con el estudio de distancia intertextual llevado a cabo por la Dra. Arjuna Tuzzi (Università degli Studi di Padova).

Otro de los segmentos identificados que equivale a una locución es *sens perjudici*, que presenta la distribución siguiente en el corpus:

SR: sens perjudici	**EAC 1932**	**EAC 1979**	**EAC 2006**
FA	1	28	23
FR (x10000)	2.4	26.5	5.7
Esp	–	E+09	E-06

Algunas de sus concordancias son:

Ya hemos visto en el apartado dedicado al pronombre *allò* que, en algunos casos, este SR forma parte de la locución *sens perjudici d'allò que*. Carles Viver Pi-Sunyer afirma que (2007:37) el uso de dicha cláusula está relacionado con la técnica legislativa que se expone a continuación:

> L'Estatut d'Andalusia i les propostes de Canàries i de Castella la Manxa apliquen la mateixa tècnica que l'Estatut de Catalunya, malgrat que en alguns casos no totes les submatèries que en l'Estatut de Catalunya es consideren exclusives tenen la mateixa consideració en els altres tres. Per contra, els estatuts o projectes d'estatut de la Comunitat Valenciana, d'Aragó, de les Illes Balears i de Castella i Lleó no identifiquen submatèries exclusives dins d'àmbits materials en què l'Estat fins ara ha pogut dictar bases, però, en canvi, com hem vist, en d'altres casos declaren exclusives «sens perjudici» competències bàsiques estatals, àmbits en els quals es clar que l'Estat pot establir bases perquè així ho diu expressament la Constitució.
> (Viver Pi-Sunyer 2007:37)

Aunque hemos constatado que aparece en el EAC 2006 en 23 ocasiones, la bibliografía indica que al redactar dicho Estatuto se produjo una innovación en la técnica legislativa relacionada con el uso de la cláusula en cuestión (*sens perjudici*), tal y como afirma Ernest Benach:

> Em sembla que [l'EAC 2006] és important, per «la seva nova tècnica legislativa d'assumpció de competències, que renuncia a la clàusula del "sens perjudici" i opta per la definició casuística i detallada, dins de cada àmbit competencial, de submatèries o perfils competencials». I hi afegeixo jo que a ningú no el podrà sorprendre que, després de vint-i-cinc anys de patir els perjudicis del «sens perjudici», els redactors de la Proposta del nou Estatut hagin optat per una tècnica legislativa moderna que precisa amb claredat l'abast de les competències de la Generalitat.
> (Benach 2006:20)

Es un cambio, pues, que se comprueba que es fruto de la modernización del discurso legislativo en redacción estatutaria para obtener en el EAC 2006 un blindaje competencial más amplio del que se había conseguido con el EAC 1979.

5. Conclusiones

El estudio presentado, como ya se ha señalado, se enmarca dentro de un proyecto de investigación postdoctoral más amplio realizado durante la primera mitad del año 2017 en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra gracias a la financiación del Institut d'Estudis de l'Autogovern de la Generalitat de Catalunya. En dicho estudio hemos llevado a cabo varios análisis lingüísticos (riqueza léxica, distancia intertextual, especificidades…) de un corpus de discurso jurídico en lengua catalana integrado por los Estatutos de autonomía de Catalunya aprobados en 1932, 1979 y 2006.

Como ya se ha señalado, se han analizado los segmentos repetidos (*SR*) que genera el análisis lexicométrico de *Lexico3.6*. Puesto que los resultados que generaba eran 2.398 y muchas de las unidades no eran representativas para, desde el punto de vista del análisis del discurso, estudiar la evolución del discurso normativo, se ha optado por analizar cualitativamente algunos de los *SR* que presentan especificidad en alguno de los subcorpus. Además, el estudio ha permitido identificar las unidades léxicas y terminológicas más empleadas en la redacción estatutaria en catalán, así como las instituciones y cargos que se regulan en dicho EAC.

Hemos identificado que, en el caso de *Correspon a la Generalitat* es un SR específico del EAC 2006 que se ha convertido, como ya se ha analizado en Morales (2018, en prensa) en una de las estructuras formulaicas más empleadas en la redacción de leyes en catalán. Asimismo, hemos identificado que, mientras en el EAC 2006 el sintagma *la Constitución* presenta especificidad negativa, en los otros dos EAC estudiados sí que se emplea por encima de las veces esperadas estadísticamente. Habrá que realizar investigaciones más amplias para entender dicha evolución en la redacción estatutaria en catalán.

El ámbito en el que la presente investigación ha resultado útil ha sido en la identificación de locuciones, que en algunos casos se emplean como unidades de conocimiento especializado (UCE, en terminología de Cabré (1999)). Las más características, en positivo, del EAC 2006 son *en tot cas* y *en els termes que*, mientras que *sens perjudici* se tendía a utilizar más en la redacción del EAC 1979. En la bibliografía hemos identificado las motivaciones de dichos cambios.

Así pues, este estudio ha permitido identificar, cruzando dos análisis

lexicométricos obtenidos con *Lexico3.6* (el de segmentos repetidos y el de especificidades), algunas unidades lingüísticas (locuciones, términos y unidades poliléxicas del discurso estatutario y jurídico-administrativo, así como cargos e instituciones) que han presentado evolución en el discurso normativo catalán en el periodo 1932-2006. En futuras investigaciones, ampliaremos el estudio de este tipo de *n-grams* y ampliarlo a unidades fraseológicas y estructuras formulaicas, porque parece que podrían aportar resultados interesantes para describir el discurso estatutario catalán desde una aproximación cronológica.

Bibliografía

[BOE] Boletín Oficial del Estado. *Constitución española*. Madrid: Agencia Estatal Boletín Oficial del Estado, 1978.

[BOPC] Butlletí Oficial del Parlament de Catalunya. "Moció 187/VI del Parlament de Catalunya, sobre l'exercici de l'autogovern." *Butlletí Oficial del Parlament de Catalunya*. 366. Barcelona: Parlament de Catalunya, 2002. 89.

[DGPL] Direcció General de Política Lingüística. *Criteris de traducció de textos normatius del castellà al català*. Barcelona: Generalitat de Catalunya. Departament de Cultura, 1999.

[SAL] Serveis d'Assessorament Lingüístic. *Llibre d'estil de les lleis i altres textos del Parlament de Catalunya*. Barcelona: Parlament de Catalunya, 2014.

Abelló Güell, Teresa. *El debat estatutari del 1932*. Barcelona: Parlament de Catalunya, 2007.

Albertí, Enoch, et al. *Manual de dret públic de Catalunya*. Barcelona: Generalitat de Catalunya. Institut d'Estudis Autonòmics, 2002.

Arnold, Edward. "Le sens des mots chez Tony Blair (people et Europe)." *JADT 2008: actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon, 12-14 mars 2008: proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14, 2008.* Eds. Heiden, Serge, Bénédicte Pincemin and Liliane Vosghanian. Lió: Presses Universitaires de Lyon, 2008. 109-19.

Benach, Ernest. *L'Estatut: una aposta democràtica i moderna: Barcelona, 7 de novembre de 2005*. Barcelona: Parlament de Catalunya, 2006.

Bosch, Jaume. *De l'Estatut a l'autodeterminació: esquerra nacional, crisi econòmica, independència i Països Catalans*. Barcelona: Base, 2013.

Cabré Castellví, M. Teresa. *La terminología. Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.* Sèrie Monografies, 3. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 1999.

Fontana, Josep. *La formació d'una identitat. Una història de Catalunya*. Vic:

Eumo Editorial, 2014.

Lamalle, Cédric, et al. *Manuel d'utilisation. Lexico3 (Version 3.41 - Février 2003)*. París: SYLED–CLA2T. Université de la Sorbonne nouvelle–Paris 3, 2003.

Menuet, Laëtitia. "Le discours sur l'espace judiciaire européen: analyse du discours et sémantique argumentative." Université de Nantes, 2006.

Montolío, Estrella, and Enoch Albertí. *Hacia la modernización del discurso jurídico: contribuciones a la I Jornada sobre la Modernización del Discurso Jurídico Español*. Barcelona: Publicacions i Edicions de la Universitat de Barcelona, 2012.

Morales Moreno, Albert. "Estudi lexicomètric del procés de redacció de l'Estatut d'Autonomia de Catalunya (2006)." Tesi doctoral no publicada. Universitat Pompeu Fabra, 2015.

Pons, Eva, and Anna M. Pla. "La llengua en el procés de reforma de l'Estatut d'autonomia de Catalunya." *Revista de Llengua i Dret*.47 (2007): 183-226.

Real Academia Española. Consejo General del Poder Judicial. "[DEJ] Diccionario del español jurídico." Madrid.

Salem, André. "Les séries textuelles chronologiques (1)." *Histoire et mesure*.VI-1/2 (1991): 149-75.

Salem, André, M. Teresa Cabré, and Lydia Romeu. *Vocabulari de la lexicometria: català, castellà, francès*. Barcelona: Centre de Lexicometria, Divisió de Ciències Humanes i Socials, 1990.

Segura, Antoni. *Crònica del catalanisme: de l'autonomia a la independència*. Barcelona: Angle Editorial, 2013.

Sobrequés, Jaume. *L'Estatut de la Transició: l'Estatut de Sau (1978-1979)*. Barcelona: Parlament de Catalunya, 2010.

Tuzzi, Arjuna. *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Roma: Carocci, 2003.

van Gijsel, Sofie, Dirk Speelman, and Dirk Geeraerts. "A Variationist, Corpus Linguistic Analysis of Lexical Richness." *Proceedings of the Corpus Linguistics 2005 Conference, July 14-17, Birmingham, UK* 1.1 (2005): 1-16.

Viver Pi-Sunyer, Carles. "Les competències de la Generalitat a l'Estatut de 2006: objectius, tècniques emprades, criteris d'interpretació i comparació amb els altres estatuts reformats." *La distribució de competències en el nou Estatut*. Eds. Viver i Pi-Sunyer, Carles, et al. Barcelona: Institut d'Estudis Autonòmics, 2007. 13-52.

Comment penser la recherche d'un signe pour une plateforme multilingue et multimodale français écrit / langue des signes française ?

Cédric Moreau
Grhapes EA 7287 - INS HEA - UPL – cedric.moreau@inshea.fr

Abstract 1 (in English)
This article examines the access to the signs in French Sign Language (LSF) within a corpus taken from the collaborative platform Ocelles, from a multilingual French bijective/LSF perspective. There is currently no monolingual dictionary in SL, so deaf users must necessarily master the written language of the country to access SL contents. Most of the available tools are based on a hypothetical conceptual relationship of equivalence between the signs of SL and the words of the dominant vocal languages. This approach originates in works that ask deaf speakers to translate a lexema outside the context of the spoken language into the signed language. This corpus is subsequently used for an inventory of minimal pairs, in which configurations, locations and movements are widely represented. This approach is thus the anchorage point for a phonological hypothesis of SL in which the previous equivalence 'sign – word' is dominant and decisive in the conception of dictionaries. This study lies within a completely different paradigm, that of the semiotical model which stems from the description of a typology and the identification of the three main transfer structures (size and form, situational, and personal). According to Cuxac, the signer can thus 'make visible' the experience by relying on the maximal resemblance sequence of signs/experience, or use the lexical unit without resemblance with the referent. This model, which is also integrative, therefore takes into account the diachronic link existing within language under the influence of pressures between transfer structures and lexical units. The morphemic approach to the study of lexical units is in this case legitimate since their compositionality does not rely on strict phonology but, in the first place, on complex morphology. First of all, we shall present our paradigm and the origins of the Ocelles multilingual and multimodal platform (written, oral, and signed languages), out of which our French written/LSF corpus is built. We will then describe a process likely to enable users to search for an LSF signifier and to relate this result to that of the corresponding written French signifier.

Abstract 2 (in French)
Cet article interroge l'accès aux signes de la langue des signes française (LSF) d'un corpus dans une perspective multilingue bijective français / LSF à partir de la plateforme collaborative Ocelles. Actuellement il n'existe pas de dictionnaire monolingue en LS, les utilisateurs sourds doivent donc nécessairement maîtriser la langue écrite du pays pour accéder à un contenu en LS. La plupart des outils à disposition s'appuient sur une hypothétique relation d'équivalence conceptuelle entre les signes des LS et les mots des langues vocales dominantes. Cette démarche prend sa source dans des travaux qui interrogent les locuteurs sourds en leur demandant de traduire un lexème hors contexte de la langue vocale en langue signée. Ce corpus est ensuite utilisé dans l'élaboration d'un inventaire de paires minimales, dans lequel les configurations, leurs emplacements et leurs mouvements sont largement représentés. Cette approche est ainsi le point d'encrage d'une hypothèse phonologique des LS dans laquelle l'équivalence « signe – mot » précédente est dominante et déterminante dans l'élaboration de dictionnaires. Notre étude s'inscrit dans un tout autre paradigme, celui du modèle sémiologique qui prend ses origines dans la description d'une typologie et de la mise en évidence des trois structures de transfert principales (de taille et de forme, situationnel et personnel). Selon Cuxac, le signeur peut ainsi « donner à voir » l'expérience en s'appuyant sur la ressemblance maximale séquence de signes/expérience, ou utiliser l'unité lexicale sans ressemblance avec le référent. Ce modèle, également intégratif, prend donc en considération le lien diachronique qui existe au sein de la langue sous l'influence de pressions entre structures de transferts et unités lexicales. L'approche morphémique pour l'étude des unités lexicales est dans ce cas légitime, leur compositionnalité ne relevant pas d'une phonologie au sens strict mais bien, en premier lieu, d'une morphologie complexe.
Nous exposerons tout d'abord notre paradigme et les origines de la plateforme multilingue et multimodale (langues écrites, orales et signées) Ocelles sur laquelle notre corpus français écrit / LSF se constitue. Nous décrirons ensuite un processus susceptible de permettre aux utilisateurs la recherche d'un signifiant en LSF et de lier ce résultat à celui du signifiant en français écrit correspondant.

Keywords: Collaborative platform, Multilingualism, Multi-modality, French Sign Language, LSF, deaf, Signs research, Semiological model, Ocelles

1. Introduction

Lorsqu'un locuteur de la langue des signes souhaite accéder à une ressource dans sa langue, notamment pour rechercher une définition dans un

dictionnaire de langue des signes (LS), il est confronté à deux obstacles. Le premier repose sur le fait que très peu d'outils présentés comme étant des dictionnaires numériques de langue des signes ne sont que des lexiques. Parmi 105 sites répertoriés sur le web, une majorité utilise le qualificatif « dictionnaire », or seulement 17 d'entre eux présentent des définitions écrites. Parmi ces 17, uniquement 7 donnent des définitions en LS. La quantité de dictionnaires en LS est donc extrêmement faible. De plus le nombre de définitions ne dépasse pas 5 000, nous sommes ainsi très éloignés des 135 000 proposées par le dictionnaire Larousse en ligne (Moreau, 2012). Le second obstacle porte sur la difficulté, pour l'utilisateur sourd d'accéder aux contenus mêmes d'un dictionnaire de ce type. En effet, nous avons constaté que dans la grande majorité des cas, les entrées proposées sont étroitement liées à la connaissance de la langue écrite du pays. Un prérequis nécessaire est donc la maîtrise de cette langue, ce qui constitue un obstacle majeur pour les personnes sourdes qui ont la LS pour langue première et la langue écrite, souvent mal maîtrisée, comme langue seconde. Parmi les 7 sites précédemment évoqués, seulement 2 offrent une entrée via les paramètres linguistiques de la LS (Moreau, 2012).

Cette question prend également un écho particulier lorsque nous interrogeons le mode de transmission des LS. Il ne s'agit pas d'un mode de transmission héréditaire, puisqu'environ 95 % des sourds ont des parents entendants qui, pour la majorité, ne pratiquent pas la LS. L'apprentissage de la langue a donc lieu dans des contextes variés, à tout âge, souvent sans la référence fixe d'un adulte proche.

Le pourcentage restant (environ 5 %) est donc constitué de sourds de parents sourds. Des parents qui, eux-mêmes pour la plupart, font partie de la catégorie précédente issus de familles entendantes. Seule 0,02 % de la population sourde signante est en effet composée d'une généalogie comptant trois générations successives de sourds signeurs. La norme d'apprentissage des LS ne peut donc pas être comparée à celles des entendants (Cuxac et Pizzuto, 2010).

En outre, la langue des signes française (LSF), marquée par plus d'un siècle d'interdiction comme langue d'enseignement, n'est reconnue comme langue de la République que depuis 2005. C'est dans ce contexte qu'est né le projet collaboratif multilingue et multimodale Ocelles[1], qui ambitionne de définir tous les concepts, dans tous les champs de la connaissance et dans toutes les langues (écrites, orales ou signées) (Moreau, 2017).

[1] https://ocelles.inshea.fr Projet sous l'égide et avec l'aide de la Délégation générale à la langue française et aux langues de France (DGLFLF) et du ministère de l'Éducation nationale.

2. Affrontement de deux paradigmes

2.1. Une hypothèse phonologique des LS

Susan Goldin-Meadow a mis en évidence, à partir d'une étude basée sur la communication préscolaire, entre petits enfants sourds et leur entourage entendant, la création de gestes appelés « home signs » (Goldin-Meadow et Mylander, 1991) (Goldin-Meadow, 2003). Pour tenter de rentrer en communication avec leur entourage, ces enfants les réalisent dans l'univers perceptivo-pratique. Ces productions permettent de faire l'hypothèse de stabilisations conceptuelles pré linguistiques, à la différence des productions d'enfants entendants du même âge, pour lesquels le lien entre la langue et ces savoirs perceptivo-pratiques n'existe pas. Une fois scolarisé, ces enfants entrent ensuite en contact avec une langue des signes institutionnalisée. Selon Golwin-Meadow dans la mesure où les formes signifiantes des langues des signes institutionnalisées ont un statut phonologique, les composants des « home signs » de l'enfant perdraient alors leur statut de morphèmes pour devenir des équivalents de phonèmes. Cette hypothèse peut être envisagée comme point de départ à l'affrontement de deux paradigmes. L'iconicité est alors comparée à de la gestuelle co-verbale illustrative, reléguée au rang de pantomime en dehors de tout phénomène linguistique.

C'est dans ce paradigme que s'inscrivent la plupart des « dictionnaires » de langues des signes actuellement. Leurs entrées sont majoritairement définies à partir d'une hypothétique équivalence conceptuelle entre les mots des langues vocales dominantes et celles des unités lexématiques (UL) des langues signées. (Fusellier-Souza, 2006). L'origine de cette méthodologie prend racine dans des travaux qui interrogent les locuteurs sourds en leur demandant de traduire un lexème hors contexte de la langue vocale en langue signée. Ce corpus est ensuite utilisé dans l'élaboration d'un inventaire de paires minimales, dans lequel les configurations (formes de la main), leurs emplacements et leurs mouvements sont largement représentés (Klima et Bellugi, 1979).

2.1. Une hypothèse morphémique des LS

Notre travail s'inscrit dans un tout autre paradigme dans lequel la conséquence de la surdité n'est plus un simple effet de changement de canal. La possibilité de dire et de montrer étant le seul fait du canal visuo-gestuel a conféré aux langues des signes une architecture différente de celle des langues vocales.

Selon Cuxac (Cuxac, 2000), deux stratégies discursives d'énonciations coexistent en LSF. Le signeur via le canal visuo-gestuel, choisit de dire sans montrer ou bien de dire en montrant. Il peut ainsi « donner à voir » l'expérience en s'appuyant sur la ressemblance maximale séquence de

signes/expérience, ou utiliser l'UL sans ressemblance avec le référent. Le modèle sémiologique (Cuxac et Pizzuto, 2010) prend ses origines dans la description d'une typologie et dans la mise en évidence des trois structures de transfert principales :

- les volumes des entités (transferts de taille et de forme (TTF)),
- les déplacements d'actants par rapport à des locatifs stables, à l'image d'un environnement en quatre dimensions (les trois de l'espace et le temps) recréé devant le locuteur (transferts situationnels (TS)),
- l'entité souhaitée par le locuteur, qui devient alors cette entité (transferts personnels (TP))

(Cuxac, 2000; Sallandre, 2003). Des expériences imaginaires ou réelles sont ainsi anamorphosées par le locuteur.

Le modèle sémiologique, prend donc en considération le lien diachronique qui existe au sein de la langue sous l'influence de pressions entre structures de transferts et UL. Lien qui se retrouve parfois dans l'étymologie de certaines des UL. L'approche morphémique pour l'étude des UL est dans ce cas légitime, leur compositionnalité ne relevant pas d'une phonologie au sens strict mais bien, en premier lieu, d'une morphologie complexe.

Lors de la réalisation d'un signe (transfert ou UL), tout le corps du locuteur prend une valeur sémantique via une organisation des éléments morphémiques qui le composent (regard expression faciale, posture, orientation du visage, configuration, le mouvement, l'emplacement (Stokoe et al., 1965), l'orientation (Friedman, 1977; Liddell, 1980; Moody, 1980; Yau, 1992).

3. Éléments prégnants dans la recherche d'un signe pour une plateforme multilingue et multimodale français écrit / LSF

3.1. Contexte d'une recherche d'un signe dans un corpus bilingue langue écrite/LS

Le projet collaboratif Ocelles permet de relier au fil des contributions, des définitions de concepts à plusieurs signifiants qu'ils soient sous formes textuelles, orales ou signées. Les entrées ne sont pas contraintes par la langue d'origine et l'architecture se déploie au fur à mesure des contributions des usagers. L'entrée textuelle peut donc prendre la forme, d'un mot ou d'une expression dans le cas où l'origine du dépôt provient d'une structure de transfert de la langue des signes. La réflexion actuelle porte donc sur le type d'indexation possible des signes indispensable au processus de recherche d'un signe dans le cadre d'un corpus bilingue langue écrite/LS.

3.2. Automatisation de l'indexation

L'indexation d'un signe se fait via l'entrée textuelle correspondante. Il n'existe pas aujourd'hui d'indexation automatique de corpus collaboratif dynamique de signes des LS qui pourrait servir de base pour un moteur de recherche d'une UL ou d'un transfert directement à partir des paramètres linguistiques des LS. La nature même du signal vidéo, très complexe à analyser ne permet pas l'indexation automatique. Outre les pertes d'informations tridimensionnelles liées aux projections de l'espace 3D à celui 2D de la vidéo, ce travail nécessiterait des outils fins d'analyses et de reconnaissances, des différents composants corporels, intervenant en parallèle, à des échelles spatiales et temporelles très différentes, mis au point pour des langues vocales, linéaires et mono source mais par pour les LS (Braffort et Dalle, 2012).

3.3. Situation actuelle et limite

Aujourd'hui l'entrée à partir des paramètres linguistiques des signes des LS se fait majoritairement à partir de la configuration. Sur les 105 sites répertoriés qui proposent des signes en LS seuls 18 offrent une possibilité d'accéder directement à un signe à partir des paramètres linguistiques de la langue des signes, sans recours à une langue écrite. Sur ces 18, 17 proposent une entrée à partir de la configuration (le nombre de ces entrées manuelles varie d'ailleurs de 9 à 211 en fonction des sites), 6 proposent une entrée à partir du mouvement, 10 à partir de l'emplacement et 1 pour la symétrie, l'image labiale et la mimique faciale (Moreau, 2012). Cette indexation phonologique des LS, avec un tel écart dans le nombre envisageable de configurations de 9 à 211 par exemple, interroge la gestion de l'erreur potentielle du locuteur qui recherche un signe qu'il aurait perçu en discours (ce qui est le cas dans la majorité des cas, compte tenu du caractère oral des LS). En outre, sur un choix entre 211 configurations, le locuteur a une chance sur 211 de choisir la bonne ou 210 risques sur 211 de se tromper…

3.4. Description et critères de recherche

L'indexation ne peut donc reposer uniquement sur une approche strictement phonologique et doit tenir compte de la gestion des erreurs possibles. Notre hypothèse repose sur une prégnance pour le locuteur de certaines unités linguistiques dans une approche morphémique mises en jeux lors de la formulation d'un signe (Moreau, 2012).

Notre approche est fondée sur le principe d'une indexation collaborative qui permet de rendre compte des perceptions des locuteurs. Le principe est basé sur le processus suivant :

- prise en compte du ou des type(s) de transfert(s) utilisé(s)

(TS / TP / TTF) dans la réalisation d'un signe à moins que l'unité lexématique puisse éventuellement trouver son origine dans l'un de ces transferts,

- itération dans le choix d'images clés à partir desquelles repose une description des unités linguistiques prégnantes (Thom, 1988),
- une description plus fine des unités retenues est ensuite proposée

Si aujourd'hui les structures linguistiques ne peuvent être admises comme familières à l'ensemble des contributeurs, leur prise en compte ne peut être ignorée. Deux approches sont envisagées. Une première inhérente à l'objectif premier de la plateforme, repose sur la proposition d'une définition de ces concepts afin de familiariser progressivement les locuteurs à leurs usages. Une succession d'anamorphoses possibles de plus en plus précises est ensuite proposée. Cette approche est cohérente avec l'utilisation de n'importe quel outil pour lequel un minimum de prérequis sont nécessaires, à l'image de l'alphabet pour un dictionnaire. Une seconde approche repose sur la prise en compte de ces lacunes en inscrivant le processus dans un continuum, qui permet une possible contribution basée sur la sélection puis la description d'images représentatives du signe du point de vue de l'usager. C'est donc l'ensemble des descriptions macro-microscopiques, de chaque contributeur qui sert de base à la pondération des unités linguistiques prégnantes. Ces données seront ensuite réutilisées comme critère de recherche d'un signe.

Conclusion ADT et visualisation, pour une nouvelle lecture des corpus Les débats de 2ème tour des Présidentielles (1974-2017)

Jean Moscarola[1], Boris Moscarola[2]
1 Université Savoie Mont Blanc, 2 Le Sphinx-Développement

Abstract 1
The progress of textual data analysis leads from a statistical and lexical description of corpora to their semantic analysis. The software thus offers the qualitative researchers the opportunity to feed their interpretations on the basis of substitutes that summarize them or to code them automatically. Finally, data visualization offers the reader an experience of the corpus creating the conditions for a critical control. This approach is illustrated on the analysis of the 2nd round debate in the presidential election conducted with DataViv the new Sphinx module.

Abstract 2
Les progrès de l'analyse de données textuelles conduisent d'une description statistique et lexicale des corpus à leur analyse sémantique. Les logiciels offrent ainsi au chercheur qualitatif la possibilité de nourrir leurs interprétations sur la base de substituts qui les résument ou de les coder automatiquement. Enfin la datavisualisation offre au lecteur une expérience du corpus créant les conditions d'un contrôle critique. Cette approche est illustrée sur l'analyse des débats de 2ème tour à l'élection présidentielle effectué avec DataViv le nouveau module de Sphinx.
Keywords: Analyse de discours, statistique lexicale, analyse sémantique, data visulaisation, logiciel Sphinx

1. Introduction

L'ADT, née d'une rencontre entre la recherche littéraire et la statistique, passe de l'étude de grandes œuvres à celle des médias de masse et de la communication politique. Avec le big data et le web sémantique elle s'enrichie des nouveaux outils de l'IA en abordant tous types de corpus. Dans les sciences humaines, l'analyse de contenu s'est développée à l'articulation de la recherche qualitative pure et des méthodes quantitatives mais sans rapport explicite avec l'ADT. Ce papier s'adresse aux chercheurs et chargés d'étude qualitative qui restent réticents à l'usage des outils de l'ADT. Il s'appuie sur l'étude du corpus des débats de 2ème tour à l'élection

présidentielle et utilise la nouvelle application Dataviv de Sphinx pour illustrer une nouvelle expérience de lecture.

2. Les méthodes et les techniques

1.1 Des humanités numériques à l'intelligence artificielle

L'outil informatique a depuis longtemps été utilisé pour informatiser les grands corpus de la littérature (Frantext). C'est ainsi qu'apparaissent dans les années 60 les humanités numériques (Burdick) et l'utilisation de la statistique pour caractériser le style de grands auteurs ou leur attribuer des œuvres anonymes (Muller). Puis dans les années 70 des statisticiens fondent le courant français de l'analyse de données textuelle qui trouve un écho avec le structuralisme et l'analyse de discours (Beaudouin). Dans les années 60 aux Etats Unis une autre voie était ouverte avec la construction de thésaurus informatisés (Stone) utilisés pour coder le contenu des media de masse.

Ces approches sont à l'origine des techniques que nous allons exposer. Elles sont enrichies dans les années 2000 par les progrès de l'ingénierie linguistique, et du traitement automatique des langues (Veronis).

2.1 Analyse de données textuelle

L'examen statistique des textes a évolué du décompte des mots à l'étude de leurs associations. Dans la tradition des concordanciers, la voie est ouverte à la recherche des segments répétés (Lebart), émaillant les discours politiques (Marchand) ou publicitaires (Floch). L'informatique graphique, les cartes cognitives (Eden) et les nuages de mots donnent une représentation visuelle de ces concordances. L'influence des contextes et la recherche des spécificités lexicales complète des descriptions globales (Brunet, Lebart)

Les méthodes d'analyses factorielles (Benzecri) font la synthèse entre la rigidité des segments répétés et le désordre des nuages de mot. En dégageant des d'affinités entre termes fréquemment associés, elles offrent une analyse structurale des textes popularisée par les cartes factorielles disposant les univers lexicaux révélateurs des thèmes du texte. A l'analyste d'en faire une lecture sémiotique.

De manière duale à la mise en évidence des univers lexicaux, Reinert propose le regroupement des unités de signification (réponses, phrases ou séquence de mots…) pour créer une partition à partir de plusieurs analyses factorielles utilisées pour progressivement définir des classes homogènes. Cette méthode, mise en oeuvre avec le logiciel ALCESTE qui lui a donné son nom, a été reprise et enrichie par d'autres logiciels (IRAMUTEC, SPHINX).

On retrouve des approches voisines chez les anglo-saxon. 'L'analyse sémantique latente' (Landauer) déplace l'attention de l'observation des cooccurrences vers la recherche de dimension latentes mesurées par les axes

factoriels. La théorie du cadrage (Frame Analysis) formulée par Goffman interprète l'usage de certains mots clés et leurs relations comme des « conceptualisations diffuses » Ces cadres sont une manière d'interpréter les univers lexicaux.

2.2 Linguistique

A l'origine les logiciels ne repéraient que les formes graphiques (séquence de lettre ne comportant aucun séparateur) sans parvenir à différencier singulier et pluriel ou les différentes flexions d'un même verbe.
La lemmatisation a représenté un grand progrès en remplaçant les différentes graphies d'un mot par son lemme : L'infinitif pour les verbes, le masculin singulier pour les noms et adjectifs. Puis l'analyse des propriétés morphosyntaxiques conduit à distinguer les 'mots pleins' selon leur statut grammatical. Les substantifs, donnent les objets des textes ou des discours, les adjectifs les appréciations et opinions, les verbes renvoient aux actions. La recherche des syntagmes permet d'identifier les expressions propres au domaine, formes les plus expressives des concordances (Mayaffre).

2.3 Sémantique

La sémantique s'intéresse au sens en passant du niveau des signifiants à celui des signifiés.
Malgré leur intérêt théorique, les travaux de linguistique générale n'ont pu déboucher sur les applications qui marquent, avec la linguistique de corpus, le véritable essor de l'analyse sémantique.
L'idée est de modéliser les connaissances de domaines particuliers comme des signifiés définis par l'ensemble des signifiants qui s'y rattachent (Saussure).
Dès les année 60, « General Inquirer» développe à Harward des ressources informatiques permettant de coder automatiquement le contenu des médias. Ces dictionnaires sont toujours accessibles. WordNet® grande base de données lexicales de l'anglais développée par l'université de Princeton généralise cette approche en améliorant l'efficacité des dictionnaires par l'usage de réseaux sémantiques. WordNet peut être considéré comme un thésaurus généralisé reflet des corpus sur lesquels il est construit. Ces idées sont reprises par les moteurs sémantiques.
Dans les années 2000, l'ingénierie linguistique et le traitement automatique des langues Normier) dépasse l'approche purement lexicale en spécifiant les thésaurus (Da Silva), par des ontologies(Grubert) et réseaux sémantiques(Godard). Le thésaurus définit l'arborescence des catégories conceptuelles : les signifiés. Les ontologies sont constituées de la liste des mots qui documentent ces catégories : les signifiant. Les réseaux sémantiques

précisent l'affectation des termes aux catégories du thésaurus en fonctions des liens constatés à partir de corpus de référence : les référents.
Avec l'essor des réseaux sociaux il devenait enfin primordial enfin d'appréhender la tonalité de messages susceptibles de faire ou défaire les réputations. Ainsi dans les années 2010 apparaissent des applications de traitement automatique des langues pour synthétiser les avis et les opinions du web. Elles ont acquis leur notoriété sous l'appellation de 'sentiment analysis' ou 'd'opinion mining' (Thelwall). Ces analyses complètent la reconnaissance des catégories du thésaurus en évaluant les textes selon leur orientation positive ou négative mesurée sur une échelle assimilable à une mesure de l'opinion.
L'Analyse de Données textuelles a ainsi évolué d'une approche descriptive statistique et lexicale à une approche sémantique fondée sur une modélisation des connaissances. Rendue très accessible par les logiciels (Boughzala) , elle présente une ressource pour la recherche qualitative ce que nous allons illustrer sur un exemple de corpus politique.

3. Contributions de l'ADT à l'analyse de corpus.

3.1 l'exemple des débats de 2ème tour

L'analyse des discours politiques est un classique de l'ADT (Marchand, Mayaffre). Leurs transcriptions analysées à différents niveaux, (les locuteurs, les tours de paroles ou les phrases) sont traitées comme des données pour révéler le style, les structures lexicales, les idées et les opinions qui les caractérisent. Le corpus des 7 débats de deuxième tour couvre de 1974 à 2017, 43 ans de vie politique. Il est analysé à l'adresse suivante **https://www.sphinxonline.net/debats/1974-2017/analyse.htm**, qui présente de manière détaillée ce dont nous donnons qu'un aperçu dans cet article. Notre but est d'illustrer les méthodes qui viennent d'être évoquées et de discuter leur pertinence pour la recherche qualitative. Le lecteur est invité à en faire lui-même l'expérience plus riche que l'aperçu qui suit :
-Les propos des candidats sont précis : les articles définis sont présents dans 2 phrases sur 3. Les embrayeurs 'je' et 'vous' sont utilisés de manière plus fréquente que 'nous'
-Les expressions « premier ministre », « assemblée nationale » « pouvoir d'achat », « général de gaulle » « milliard d'euro » dominent sur l'ensemble de la période.
-La carte des univers lexicaux montre une opposition entre l'évocation de la vie politique d'une part et les termes de l'économie et de la société d'autre part.
-Sur les 11 thèmes identifiés par la classification automatique, les thèmes 'Gouvernement-Majorité', 'Pays, Français', 'Année Nucléaire', 'Entreprise

Salarié' arrivent en tête.
-Les principaux concepts reconnus par le thésaurus de l'application utilisée[1] sont « Vote » « Civilisation » « Emploi et salaire » « Politique fiscale » « Citoyenneté »…
-La tonalité des propos est neutre pour la moitié des interventions, pour le reste les prises de position positives sont un peu plus fréquentes.

La référence aux candidats et aux périodes complète la description globale.
-A chacun son style : Jospin Royal et Mitterrand se distinguent par l'usage de 'je' ; Chirac par le 'nous' plus collectif et Marine Le Pen interpelle son débateur (vous) à moins qu'elle ne s'adresse à l'audience. Macron fait preuve de l'usage le mieux équilibré.
-Les mots clés sur représentés dans chaque période marquent bien le changement de siècle : 'politique', 'gouvernement' 'problème' au XXème, 'entreprise' 'emploi' 'européen' au XXIème.
-Les catégories thématiques de la classification lexicale sont associées à des groupes de candidats : Sarkozy, Royal et Hollande développent les thèmes 'Entreprise, Salarié', 'Loi', 'Crise Priorité' 'Pouvoir Président'. Mitterrand et Giscard d'Estaing, 'Socialiste Communiste', 'Gouvernement Majorité', Macron et Le Pen 'Chômage, Emploi', 'Français, Pays'
-Enfin les concepts de l'analyse sémantique distinguent nettement les périodes : 'Vote' 'Civilisation' 'Degré de libéralisme' au XXème, et 'emploi', citoyenneté' 'politique fiscale' au XXIème

3.2 Contribution à l'analyse qualitative pure

Ces résultats plus abondamment décrits dans l'application en ligne peuvent être utilisées dans l'esprit de la recherche qualitative pure dès lors qu'on les envisage dans une démarche descriptive et exploratoire dont la valeur réside que dans la capacité du chercheur à les lire et à les d'interpréter (Moscarola). Les mots clé, nuages, cartes, les classifications et les concepts proposés par les logiciels sont des substituts du corpus. Ils portent la trace des modèles mentaux (Johnson-Laird) et des représentations et l'influences sociales dont parlent la théorie des actes de langage (Austin) et la sociolinguistique. L'ADT permet d'en faire une sorte de radioscopie et de mieux les comprendre. Elle offre aussi la possibilité d'une lecture distanciée échappant au risque de récursivité (Dumez) ou donnant la possibilité de le contrôler. En effet les substituts lexicaux ou sémantiques sur lesquels le chercheur fonde ses interprétations peuvent être communiqués pour exposer la lecture qu'il en fait à la critique d'une discussion basée sur des éléments partagés.

[1] Thésaurus Larousse (Péchon 1994) intégré à SphinxIQ2

3.2 Contribution à l'analyse de contenu

L'ADT peut également être vue comme une modalité de l'analyse de contenu traditionnelle (Belerson, Bardin). Elle s'en distingue par l'automatisme d'une 'lecture artificielle' identifiant des catégories établies statistiquement par apprentissage ou à partir d'un thésaurus. On retrouve ainsi l'approche inductive conduisant à interpréter à postériori les structures révélées par les analyses factorielles ou à reconnaître dans le corpus les concepts du thésaurus.

Chaque unité de signification peut ainsi être codée dans une variable 'mesurant' le sens et utilisable selon les procédures classiques de l'analyse quantitative. Dans notre exemple on peut ainsi chercher les éléments lexicaux ou sémantique explicatif ou discriminant les appartenances politique des candidats…

3.3 Retour au texte et 'data visualisation'

Le recours à l'ADT lexicale ou sémantique comporte deux risques majeurs malgré son intérêts pratique et scientifique : le risque d'erreur systématique auquel expose la lecture par une machine et le risque de réduction abusive imposé par les choix du chercheur, qu'il s'agisse de sa problématique ou des résultats qu'il choisit de communiquer.

Le premier risque peut être évité par le retour au texte et une lecture de vérification. C'est la seule manière pour le chercheur et son lecteur de contrôler le sens des éléments lexicaux ou la pertinence des concepts et évaluations identifiés par les moteurs sémantiques ? Cette possibilité apparaît avec les hypertextes. Elle est d'autant plus nécessaire, qu'avec l'aide des infographies (nuages de mots, cartes) les représentations deviennent de plus en plus parlantes.

Les méthodes dites de navigation facilitent ce retour au texte et peuvent être enrichies par les entrées provenant des codifications lexicales et sémantiques ou par les éléments des représentations visuelles. La navigation lexicale généralisée dans l'esprit de la datavisualisation (Faulx Briole) donne ainsi au lecteur la possibilité d'accéder directement aux verbatims associés aux mots d'un nuage ou d'une carte, aux catégories d'une classification automatique ou aux concepts et appréciations d'une analyse sémantique. Par exemples à quel verbatim correspond l'usage des mots 'gens' ou 'français', sont-ils plutôt de gauche ou de droite, à quoi correspond le concept 'citoyenneté' et est-il daté par un époque ou spécifique à certains candidats ? Retour au texte, mais au contexte aussi.

L'analyse des discours politique a été pionnière dans ce domaine. Le Monde publie le 15-03-2012 une infographie dynamique donnant accès aux discours

de campagne des candidats (Véronis). L'observatoire du discours politique (Mayaffre) en est un autre exemple. Il permet à partir d'un nuage de mots synthétisant le contenu des discours, d'en détailler les significations par du verbatim et d'en spécifier l'usage selon les différents candidats.

Avec ce type d'application le chercheur qualitatif peut compléter la communication de ses résultats et de ses interprétations en donnant accès au corpus par l'expérience d'une navigation interactive proposée au lecteur. Il peut ainsi vérifier les interprétations de l'auteur et les prolonger par ses propres conjectures. C'est ce que nous proposons à l'adresse : **https://www.sphinxonline.net/debats/1974-2017/analyse.htm** Y sont présentés les substituts et synthèses qui conduisent à conclure à une profonde transformation du débat politique amorcée au tournant du siècle. Ces tendances peuvent être expérimentées par le lecteur pour nourrir une discussion critique ou susciter de nouvelles explorations et conjecture. Le logiciel utilisé permet ainsi de produire des résultats et en même temps de donner la possibilité au lecteur de les discuter. C'est le propre de la démarche scientifique.

Bibliographie

BARDIN L. (1977) *L'Analyse de Contenu* PUF

BEAUDOUIN V. (2016) *Retour aux origines de la statistique textuelle : Benzécri et l'école française de l'analyse de données* JADT 2016

BENZECRI JP. (1992) *Correspondance Analysis Handbook* Marcel Decker Inc. 1992

BERELSON, B (1952). *Content Analysis in Communication Research*. Glencoe: Free Press..

BOUGHZALA Y., HERVE H., MOSCAROLA J. (2014) *Sphinx Quali : un nouvel outil d'analyses textuelles et sémantiques* JADT Université de Paris

BRUNET E. (2016) *Apports des technologies modernes à l'histoire littéraire* HAL

BURDICK A., DRUCKER J & ali. (2012) *Digital humanities* MIT Press

DA SILVA L. (2006) *Thésaurus et systèmes de traitement automatique de la langue,* Documentation et bibliothèque

DUPUY, P.-O. & MARCHAND, P., (2016) *Les débats de l'entre-deux-tours de l'élection présidentielle française (1974-2012)* Mots. Les langages du politique,

EDEN C. (1988). "Cognitive mapping". *European Journal of Operational Research*

FAULX-BRIOLE A. (2017) *Datavisualisation et tableaux de bord interactifs* Solution Business

FLOCH J.M.(1988), *The contribution of structural semiotics to the design of a hypermarket, International* Journal of Research in Marketing, 4, 3, Semiotics and Marketing

GOFFMANN E. *Frame analysis: An essay on the organization of experience* Harper and Row 1974

GRUBER T. (1992) *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* In: International Journal Human-Computer Studies

JOHNSON-LAIRD, P N. (1983) *Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness*. Harvard University Press

LANDAUER, T. K., FOLTZ, P. W., & LAHAM, D. (1998) *An introduction to latent semantic analysis*. *In Discourse processes,* Routledge

LEBART L SALEM A. (1988) *Analyse de données textuelles* DUNOD

MARCHAND, P. (2016). *Les représentations sociales dans le champ des médias*. In G. Lo Monaco, S.

MAYAFFRE D. (2005) *Analyse du discours politique et Logométrie : point de vue pratique et théorique* Langage et société N° 114

MAYAFFRE D. (2014) *Plaidoyer en faveur de l'analyse de données c(n)textuelle. Parcours coocurrentiels dans le discours présidentiel français.* Actes JADT Nice

MOSCAROLA J. (2018) *Faire parler les données. Editions* EMS

MULLER C. (1979,). *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Slatkine

NORMIER B. (2007). L'apport des technologies linguistiques au traitement et à la valorisation de l'information textuelle. ADBS.

REINERT A., (1983), *Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte"* Les cahiers de l'analyse des données, Tome 8, N°2, pp. 187198.

STONE D.C. DUNPHY, M.S. SMITH, . M. OGILVIE. (1966) *The General Inquirer: a computer Approach to Content Analysis* MIT Press

THELWALL M. (2017) *Sentiment Analysis for Smal and Big Data*.SAGE

VERONIS J. (2014) *Le traitement automatique des corpus oraux. In Traitement automatique des langues*. Hermes

A conversation analysis of interactions in personal finance forums

Maurizio Naldi
University of Rome Tor Vergata– maurizio.naldi@uniroma2.it

Abstract 1
Interactions on a personal finance forum are investigated as a conversation, with post submitters acting as speakers. The presence of dominant positions is analysed through concentration indices. Patterns in replies are analysed through the graph of replies and the distribution of reply times.
Keywords: Personal finance; Conversation analysis; Concentration indices.

1. Introduction

Decisions concerning personal finance are often taken by individuals not just on the basis of factual information (e.g., company's official financial statements or information about past performance of funds), but also considering the opinions of other individuals. Nowadays personal finance forums on the Internet have often replaced friends and professionals in that role. In those forums the interaction occurs among people who typically do not know one another personally and know very few personal information (if any) about other participants. Anyway, they often create online communities that can bring value to all participants [1]. Examples of such forums are SavingAdvice (http://www.savingadvice.com/forums/) or Money Talk (http://www.money-talk.org/board.html).

The actual influence of such forums on individuals' decisions has been investigated in sev- eral papers, considering, e.g., how the level of activity on forums impacts on stock trading levels [2], how participation in such forums pushes towards a more risky-seeking behaviour [3], or introducing an agents-based model to determine how individual competences evolve due to the interaction [4]. It has been observed that such forums may be employed by more aggressive participants to manipulate more inexperienced ones [5], establishing a dominance over the forum. In addition to being undesirable for ethical reasons, such an influence is often contrary to the very same rules of the forum. Here we investigate the subject by adopting a different approach from the semantic anal- ysis of [5]. In particular, we investigate the presence of imbalances in the online discussion and the dynamics of the interaction between participants. The rationale is that partici- pants wishing to manipulate others would try to take control of the discussion by posting

more frequently and being more reactive.
For that purpose we employ two datasets extracted from the two most popular personal finance threads on the *SavingAdvice* website. For the purpose of the analysis the thread is represented as the sequence of participants taking turns, with dates and times of each post attached.
We conduct a conversation analysis, wishing to assess if: 1) there are any dominant par- ticipants (in particular the thread starter); 2) repetitive patterns appear such as sustained monologues or sparring matches between two participants; 3) replies occur on a short-time scale.
The paper provides the following contributions:

- through the use of concentration indices we find out that, though no dominance exist, the top 4 speakers submit over 60% of the posts (Section 3);
- both recurring reply sequences and monologues appear (Section 4);
- reply times can be modelled by a lognormal distribution, with 50% of the posts being submitted no longer than 14 or 23 minutes (for the two datasets respectively) after the last one (Section 4).

2. Datasets

We consider the two most popular threads on the SavingAdvice website. The topics are the following, where we indicate an identifying short name between parentheses:

1. Should struggling families tithe? (Struggling) African-American Personal Finance Gurus (Guru)

The main characteristics of those datasets are reported in Table 1. For each thread we identify the set of speakers $S = \{s_1, s_2, \ldots, s_n\}$, i.e., the individuals who submit posts. We identify also the set of posts $P = \{p_1, p_2, \ldots, p_m\}$ and a function $F : P \rightarrow S$, that assigns each post to its submitter. For each speaker we can therefore compute the number of posts submitted by him/her. If we use the indicator function $\mathbb{1}(\cdot)$, the number of posts submitted by the generic speaker s_i is

$$N_i = \sum_{j=1}^{m} \mathbb{1}_{F(p_j)=s_i}. \quad (1)$$

Table 1: Datasets

Struggling	jpg7n16	25	155
Guru	james.hendrickson	18	104

3. Dominance in a thread

In this section we wish to examine if some dominance emerges in a thread. We adopt concentration indices borrowed from the field of industrial economics.We analyse domi- nance by considering the frequency of posts: an individual (or a group of individuals) is dominant if it submits most of the posts. We first examine how posts are distributed by looking at the rank-size plot: after ranking speakers by the number of posts they submit, the frequency of posts is plotted vs the rank of the speaker. In Figure 1, we see that a linear relationship appears between log N(i) and the rank *i*, so that a power law $N_{(i)} = k/i^{\alpha}$ (a.k.a. a generalized Zipf law) may be assumed to apply roughly, where *k* is a normalizing constant and α is the Zipf exponent (see, e.g., [6]), measuring the slope of the log-linear curve, hence the imbalances between the contributions of all the speakers. By performing a linear regression, we get a rough estimate of α, reported in Table 2.

Table 2: Concentration measures

Struggling	0.2545	0.1220	61.94%
Guru	0.2501	0.1396	67.31%

As more general indices to assess dominance position we borrow two from Industrial Economics: the Hirschman-Herfindahl Index (HHI) [7, 8, 9], and the CR4 [10, 11]. For a market where n companies operate, whose market shares are v_1, v_2, …, v_n the HHI is

$$\mathrm{HHI} = \sum_{i=1}^{n} v_i^2. \tag{2}$$

The HHI satisfies the inequality $1/n \leq \mathrm{HHI} \leq 1$, where the lowest value corresponds to the case of no concentration (perfect equidistribution of the market) and the highest value represents the case of monopoly. Therefore, the larger the HHI the larger the concentration. Instead, the CR_4 measures the percentage of the whole market owned by the top four companies: similarly, the higher the CR_4, the heavier the concentration.

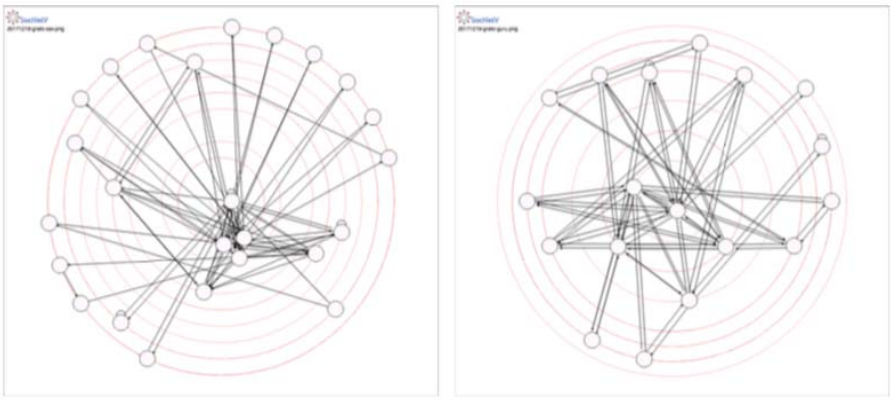
In our case, the fraction of posts submitted by a speaker can be considered as his/her market share, so that the HHI can be redefined as

$$\mathrm{HHI} = \sum_{i=1}^{n} \left(\frac{N_i}{m} \right)^2 = \frac{\sum_{i=1}^{n} N_i^2}{m^2}. \tag{3}$$

Instead, the CR4 is defined as

$$\mathrm{CR}_4 = \frac{\sum_{i=1}^{4} N_i}{\sum_{i=1}^{n} N_i} \tag{4}$$

For our datasets we get the results reported in Table 2. According to the guidelines provided by the U.S. Department of Justice, the point of demarcation between unconcentrated and moderately concentrated markets is set as HHI = 0.15 [12]. Since the values in Table 2 are below that threshold, we cannot conclude that there is a significant concentration phenomenon. However, the CR_4 index shows that the top 4 speakers submit more than 60% of all the posts. Delving deeper into the top 4, we also see the most frequent speaker typically contributes around 1/4 of the overall number of posts, which represents a major influence. In the Struggling dataset, the most frequent speaker is the thread originator itself (with 22.6% of posts), while that's not true in the Guru dataset, where the the most frequent speaker contributes 26.9% of posts and the originator just 2.88%.



(a) Struggling dataset (b) Guru dataset

Fig. 1: Rank-size plot

4. Replies

After examining dominance, we turn to interactions. In this section we analyse the pattern of replies, looking for recurrences in the sequence of

replies and examining the time elapsed before a post is replied to.
We build a graph representing how speakers reply to each other. We consider each post as a reply to the previous one. We build the replies graph by setting a link from a node A to a node B if the speaker represented by node A has replied at least once in the thread to a post submitted by the speaker represented by node B. The resulting graphs are shown in Figure 2, which is ordered from the core to the periphery in order of decreasing degree of the nodes, laid out on concentric rings. Here the degree of a node represents the number of speaker to which it replies. In both cases an inner core of most connected nodes appear, which represent the speakers replying to most other speakers. Reply patterns emerge as bidirectional links (couples of speakers who reply to each other). Loops represent monologues instead, i.e., speakers submitting two or more posts in a row.

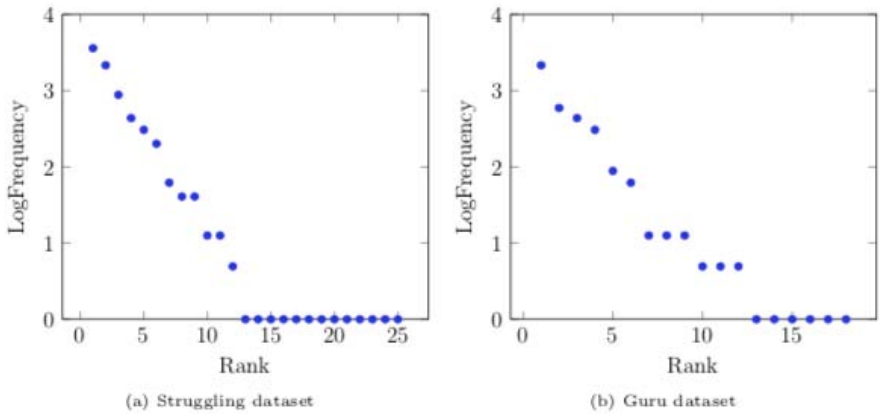


Fig. 2: Replies graph

Further, we are interested in how fast the interactions are between contributors to the thread. We define the reply time as the time elapsing between a post and the subsequent one. The main statistics of the reply time are reported in Table 3. In both dataset the mean reply time is around 1 hour, but 50% of the replies take place within either 14 minutes (Guru dataset) or 23 minutes (Struggling dataset), i.e., with a much smaller turnaround. There is therefore a significant skewness to the right.
A more complete view of the variety of reply times is obtained if we model the probability density function. In Figure 3, we report the curves obtained through a Gaussian kernel estimator, an exponential model, and a lognormal model (whose parameters have been estimated by the method of moments). By applying the Anderson-Darling test, we find out that the exponential

hypothesis is rejected at the 5% significance level, while the lognormal one is not rejected, with a p-value as high as 0.72 for the Struggling dataset and 0.076 for the Guru dataset.

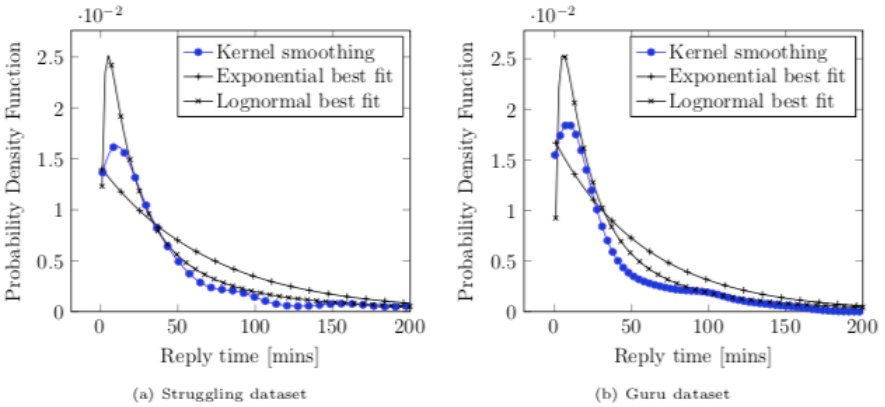


Fig. 3: Reply time

Table 3: Reply time statistics (in minutes)

Struggling	70.5	23	156.2	254.7
Guru	58.9	14	112.7	406.7

5. Conclusions

We have analysed two major threads within a personal finance forum as a conversation between submitters acting as speakers, searching for dominance and interaction patterns. Though no significant concentration exists, the top four speakers submit over 60% of the posts. Patterns of interaction emerge as the presence of several couples of speakers who reply to each other, several monologues, and short reply times (with 50% being below 14 and 23 minutes for the two datasets, though a significant distribution tail is present).

References

[1] Arthur Armstrong and John Hagel. The real value of online communities. Knowledge and communities, 74(3):85–95, 2000.

[2] Robert Tumarkin and Robert F Whitelaw. News or noise? internet postings and stock prices. Financial Analysts Journal, 57(3):41–51, 2001.

[3] Rui Zhu, Utpal M Dholakia, Xinlei Chen, and Ren´e Algesheimer. Does online community participation foster risky financial behavior? Journal of

Marketing Research, 49(3):394–407, 2012.
[4] Loretta Mastroeni, Pierluigi Vellucci, and Maurizio Naldi. Individual Competence Evolution under Equality Bias. In 2017 European Modelling Symposium (EMS), Nov 2017.
[5] John Campbell and Dubravka Cecez-Kecmanovic. Communicative practices in an on- line financial forum during abnormal stock market behavior. Information & management, 48(1):37–52, 2011.
[6] Maurizio Naldi and Claudia Salaris. Rank-size distribution of teletraffic and customers over a wide area network. Transactions on Emerging Telecommunications Technologies, 17(4):415–421, 2006.
[7] Stephen A Rhoades. The Herfindahl-Hirschman Index. Fed. Res. Bull., 79:188, 1993.
[8] Maurizio Naldi. Concentration indices and Zipf's law. Economics Letters, 78(3):329–334,
2003.
[9] Maurizio Naldi and Marta Flamini. Censoring and Distortion in the Hirschman–Herfindahl
Index Computation. Economic Papers: A journal of applied economics and policy, 2017.
[10] I Pavic, F Galetic, and Damir Piplica. Similarities and Differences between the CR and HHI as an Indicator of Market Concentration and Market Power. British Journal of Economics, Management and Trade, 13(1):1–8, 2016.
[11] Maurizio Naldi and Marta Flamini. Correlation and concordance between the CR4 index and the Herfindahl-Hirschman index. SSRN Working paper series, 2014.
[12] The U.S. Department of Justice and the Federal Trade Commission. Horizontal Merger Guidelines, 19 August 2010.

Analisi testuale, rumore semantico e peculiarità morfosintattiche: problemi e strategie di pretrattamento di corpora speciali.

Stefano Nobile
Sapienza Università di Roma – stefano.nobile@uniroma1.it

Abstract 1
The proliferation of text analysis techniques has made possible the combined use of different software, directed each time to specific needs for analysis and research. However, the opportunities offered by the different software do not mitigate a fundamental problem, inherent in the characteristics of some peculiar corpora. Perfectly suited for analysis on texts written accurately and based on a supervised style, however these software can not reduce some issues. Among these, one of the most common concerns the morphosyntactic rules of the language with its semantic noise. Problems of "noise", such as that generated in spontaneous conversations, require many precautions for the preparation of the corpus. This situation is exaggerated with Twitter, whose ease of access and messaging download has produced analysis that is not always adequately supported from the theoretical point of view. Poems and songs present a similar problem. In these kinds of corpora the problem derives from the structure of this style of communication, which in using some rhetorical expedients accentuates the critical mass generated by some words. What strategies are possible to adequately prepare the corpora to be analysed in these two particular situations? The contribution proposes some strategies on how to operate in these particular conditions, highlighting the advantages on the empirical level but also the effects on the theoretical one.

Abstract 2
La moltiplicazione delle tecniche di analisi testuale ha reso possibile l'uso combinato di software diversi, piegati di volta in volta a singole esigenze di analisi e ricerca. Tuttavia, l'ampiezza di opportunità offerte dai diversi software non attenua un problema di fondo, insito nelle caratteristiche stesse di alcuni corpora peculiari. Perfettamente adatti ad analisi su testi redatti accuratamente e improntati a uno stile sorvegliato, questi software non riescono tuttavia a togliere l'utente dall'impaccio nel quale può trovarsi in alcune circostanze. Tra queste, una delle più comuni riguarda le regole morfosintattiche della lingua di riferimento e quindi portatrice di quote elevate di rumore semantico. Problemi di "rumore", come quello generato

nelle conversazioni spontanee, richiedono al ricercatore una serie di accorgimenti per la preparazione del corpus che tengano conto della necessità di evitare di ottenere dati fortemente distorti. Questo discorso si esaspera con Twitter, la cui facilità d'accesso e download dei messaggi è da qualche tempo foriero di analisi non sempre adeguatamente sostenute dal punto di vista teorico. A questi casi si aggiunge quello di corpora altrettanto peculiari come quelli delle poesie e delle canzoni. In corpora di questo tipo il problema deriva dal costrutto stesso di questo genere comunicativo, che nel servirsi di alcuni espedienti retorici accentua la massa critica generata da alcune parole, andando così a incidere, tra l'altro, sul calcolo di alcuni parametri rilevanti e rendendo meno leggibili i risultati. Quali strategie sono dunque possibili al ricercatore per preparare adeguatamente i corpora da analizzare in queste due situazioni particolari? Il contributo che si intende presentare vuole avanzare alcune proposte su come operare in queste particolari condizioni, evidenziando i vantaggi sul piano empirico ma anche le ricadute su quello teorico soggiacente agli obiettivi stessi che analisi su corpora di questo genere possono porsi.

Keywords: rumore semantico, poesia, canzone, retorica, pretrattamento del corpus, costruttivismo vs. realismo.

1. Rumore semantico e corpora testuali peculiari

La moltiplicazione delle tecniche di analisi testuale ha reso possibile, ai ricercatori interessati a lavorare in questo ambito, l'uso – anche combinato – di diversi software, ciascuno con le proprie peculiarità in risposta alle differenti esigenze di analisi e ricerca. Tuttavia, l'ampiezza di opportunità offerte dai tanti software in commercio (T-Lab, Taltac, Spad-T, R, eccetera) non attenua un problema di fondo, insito nelle caratteristiche stesse di alcuni corpora peculiari: quello delle distorsioni imputabili al rumore semantico generato sia da elementi irrilevanti dal punto di vista contenutistico, sia da ridondanze che alterano i rapporti di forza tra parole.

Perfettamente adatti ad analisi testuali su testi redatti accuratamente e improntati a uno stile sorvegliato come può essere quello delle testate giornalistiche o di materiali di tipo istituzionale, questi software non riescono tuttavia a togliere l'utente dall'impaccio nel quale può trovarsi in alcune circostanze che, più o meno in concomitanza con la diffusione dei social network, hanno cominciato ad essere egemoniche quanto a produzioni di testi sul web. Tra queste circostanze, una delle più comuni riguarda quella che si potrebbe definire oralità scritta, poco o per nulla accorta alle regole morfosintattiche della lingua di riferimento e quindi portatrice di quote elevate di rumore semantico, qui inteso come forma leggibile e trattabile di

testo. Problemi di "rumore" come quello generato nelle conversazioni spontanee, rinvenibili – nelle forme più disparate – in rete, richiedono al ricercatore una serie di accorgimenti per la preparazione del corpus che tengano conto della necessità di evitare di ottenere dati fortemente distorti. Vale a dire che le forme linguistiche contratte (*cmq, nn, xké*), gli elementi espressivi tesi a restituire i toni del parlato (*belloooo*, *bravaaaa*), i segni grafici del tutto peculiari (Ã, Ã², ðŸ, Ã¹, Ã©, ðŸ), le ridondanze, i *retweet*, il testo non in formato Ascii, gli hashtag, i collegamenti multimediali, il linguaggio di *markup*, sono addendi di una somma che dà come risultato una proliferazione di rumore semantico, ai cui effetti si aggiungono quelli derivanti dalle distorsioni imputabili agli indici prodotti (ricercatezza ed estensione lessicale) nonché alle misure del corpus (occorrenze, forme grafiche, hapax). Questo discorso si esaspera con *Twitter*, la cui facilità d'accesso e download dei messaggi è da qualche tempo foriero di analisi non sempre adeguatamente sostenute dal punto di vista teorico (Ebner, Altmann e Softic, 2011). Accade infatti sempre più spesso che «l'elevato grado di automatismo delle procedure e la forte tendenza alla modellizzazione statistica possono esporre *l'analisi testuale* a stili di ricerca segnati da un'ingenua rincorsa dell'oggettività tramite l'estremizzazione ossessiva del calcolo numerico applicato ai testi, con la conseguente grave perdita del ruolo del contesto» (Tipaldo, 2014: 191; corsivo aggiunto). La necessità di contrarre il testo in 120 caratteri (raddoppiati soltanto a partire dal novembre 2017, ma la sostanza non cambia) determina infatti negli utenti l'inclinazione a trovare soluzioni – a volte convenzionali, altre volte originali – per poter ridurre il testo entro i limiti prefissati, così come si faceva quando gli sms avevano set limitati di caratteri ed erano relativamente dispendiosi. Da qui, la produzione di una quantità considerevole di rumore semantico che rende difficilmente trattabili i dati testuali "naturali". Ai casi appena passati in rassegna – oggi largamente diffusi – si aggiunge quello di corpora altrettanto peculiari, ma del tutto diversi, come quelli delle poesie e delle canzoni (Nobile, 2012). In testi di questa natura, il problema deriva dal costrutto stesso di questi generi della comunicazione. Essi, infatti, nel momento in cui si servono di alcuni espedienti retorici (l'anadiplosi, l'epanalessi, il poliptoto, l'anafora, l'epanadiplosi e altri ancora), accentuano la massa critica generata da alcune parole. Ciò finisce con l'incidere sul calcolo di alcuni parametri rilevanti (specificità tipiche ed esclusive, estensione lessicale, ricercatezza lessicale, rango delle singole parole, confronto con i lessici peculiari, eccetera), rendendo meno leggibili i risultati.

Un caso assai frequente, qui portato al parossismo, è il seguente: nella canzone, alcune parole, per necessità squisitamente ritmiche oppure per enfatizzare l'effetto-tormentone, vengono ripetute ossessivamente. È quanto

accade – per fare un solo esempio, dati i margini ridotti entro i quali deve rimanere questo contributo – con la canzone *Pino (fratello di Paolo)*, nella quale la parola *Pino* compare addirittura 60 volte nel giro di pochi secondi, andando ineluttabilmente a gonfiare tutte le modalità delle variabili (artista, decennio di pubblicazione, macro e microgenere musicale, sesso) a cui questa singola canzone è collegata (Nobile, 2012). Per l'uso delle figure retoriche vale un discorso analogo. Tra le tante possiamo prendere l'anafora a titolo esemplificativo. L'anafora è una figura retorica che consiste nella ripetizione di una o più parole all'inizio di una frase o di un verso. Per quanto essa sia rintracciabile anche nella prosa, è nella poesia e nella canzone che essa ottimizza le proprie potenzialità espressive. Tra lo sterminato numero di esempi che potremmo scegliere, uno è quello di *Vai in Africa, Celestino!*, un brano che il cantautore Francesco De Gregori ha pubblicato nel 2005: *pezzi di stella, pezzi di costellazione / pezzi d'amore eterno, pezzi di stagione / pezzi di ceramica, pezzi di vetro / pezzi di occhi che si guardano indietro / pezzi di carne, pezzi di carbone / pezzi di sorriso, pezzi di canzone / pezzi di parola, pezzi di parlamento / pezzi di pioggia, pezzi di fuoco spento*. In questo caso, è la parola *pezzi* a comparire un considerevole numero di volte grazie, appunto, all'espediente retorico dell'anafora. Non diverso, ovviamente, è il caso della letteratura, per il quale – a titolo esemplificativo – possiamo scomodare il celeberrimo III canto (canto e canzone, appunto…) dell'Inferno dantesco: *Per me si va ne la città dolente / per me si va ne l'etterno dolore / per me si va tra la perduta gente*. La poesia e la canzone, dunque, possono presentare delle caratteristiche strutturali che vanno a incidere sul *text mining* operabile dai diversi software, nella misura in cui forniscono informazioni numeriche alterate. Quantunque la ridondanza di alcuni termini non implichi necessariamente lo stravolgimento dell'asse sintagmatico (Bolasco, 2005), ossia della possibilità di ricostruire il senso del testo in ragione di un criterio di adiacenza delle parole all'interno dei contesti elementari, essa può compromettere il senso espresso dai dati relativi alla frequenza delle parole piene, alle peculiarità (sia quelle endogene, esprimibili in termini di specificità, sia quelle esogene, traducibili in termini di linguaggio peculiare) e alla numerosità di forme grafiche. Quali strategie sono dunque possibili al ricercatore per preparare adeguatamente i corpora da analizzare in queste due situazioni particolari, ossia profluvio di segni grafici e parole ripetute? Certamente non è sufficiente ripulire ortograficamente il testo né espungere da esso tutti quei segni, come le emoticons o la sintassi comunicativa propria di *Twitter*, che vanno a interferire su molti parametri d'analisi. Né d'altronde si può "addomesticare" il corpus fino al punto da stravolgerne l'aspetto precipuo, ossia la spontaneità del simil parlato del primo caso e la struttura morfosintattica e retorica del secondo.

2. Strategie di pre-trattamento del corpus

Le soluzioni ai tipi di problemi testé esposti variano a seconda della natura del problema, delle competenze informatiche dell'utente e della prospettiva analitica assunta dal ricercatore e dipenderanno dalla combinazione tra queste tre dimensioni. Vediamole. La pulizia dei caratteri di testi naturali dipende in larga misura dalle competenze informatiche dell'utente, al netto delle potenzialità dei software utilizzati. Ad oggi, un utente privo di abilità informatiche avanzate non è in grado di fare un lavoro di pulizia impeccabile su corpora testuali molto "sporchi" come sono quelli che provengono da *Twitter*. Se da un lato gli potrà essere d'aiuto una elevata quota di pazienza per utilizzare un correttore ortografico che ripulisca il testo dagli errori di battitura tipici di testi "naturali", e quindi non supervisionati, dall'altro dovrà necessariamente scontrarsi con la ridda di caratteri speciali che sono stati richiamati in precedenza. Le soluzioni a disposizione sono tre: il livello base consiste nella sostituzione manuale e in blocco di tutti i segni grafici da correggere, facendo attenzione – nell'uso di un normale *word processor* – alle maiuscole e alle minuscole. Si tratta di un'operazione tanto più lunga e faticosa quanto più lungo, complesso e ricco di rimandi ipertestuali è il corpus da ripulire. In alcuni casi, esistono software come Taltac che possiedono al loro interno una funzione di rimozione di alcuni caratteri particolari. Una seconda soluzione è quella di programmare delle macro (o, alternativamente, di usare programmi esterni) che risolvano lo stesso tipo di problema. La soluzione è più efficace dal punto di vista del risultato finale, ma altrettanto impegnativa da quello delle competenze e del tempo richiesti. La terza soluzione è, sulla carta, quella in grado di ottimizzare meglio il rapporto costi/benefici. Si tratterebbe, in questo caso, di sfruttare le potenzialità di programmi di ricerca che si sono dati come obiettivo proprio quello della pulizia di testi originati nel web e utilizzati per analisi testuali. Vanno in questa direzione progetti come *Readability* o *CleanEval* (Baroni et al., 2008), che tuttavia presentano a loro volta due ordini di problemi: uno legato ai costi; l'altro alla effettiva possibilità d'accesso. Entrambi, peraltro, evidenziano problemi di flessibilità rispetto ai diversi formati di corpora da elaborare (Claridge, 2007; Petri e Tavosanis, 2009). La questione del trattamento di corpora che devono la loro peculiarità alla struttura soggiacente, pur non presentando problemi rilevanti di ordine informatico, è più complessa e implica scelte decisive da parte del ricercatore. Il ricercatore dovrà infatti operare delle scelte di carattere gnoseologico e teorico rispetto ai fini che si pone, ben sapendo che le decisioni che prenderà avranno inevitabili ricadute sul piano delle risultanze empiriche. In altri termini, il ricercatore che impatta con materiale testuale che non nasce in forma di

prosa, ma di verso, si trova sostanzialmente a dover operare una scelta tra una rappresentazione fedele, "fotografica", delle caratteristiche del corpus esaminato e quella che invece tiene conto delle ridondanze e di tutti quegli elementi che possono contribuire a gonfiare alcuni parametri del corpus, a partire dal conteggio di forme grafiche e a finire con gli hapax. Nel primo caso gli esiti dell'analisi subiranno l'impatto non solo di quegli elementi retorici e morfosintattici che possono caratterizzare la forma-canzone o la forma-poesia, ma soprattutto del ritornello. Accettare questa prospettiva significa assumere alcune sezioni di testo – nonché gli elementi di esso che contribuiscono a ispessire alcuni termini per via delle scelte operate sui versi dagli autori – come elementi che, proprio perché ripetuti, meritano di svettare in termini parametrici dall'analisi del corpus stesso. Possiamo dire che in un caso come questo i risultati siano ingannevoli? Dipende, appunto, dalla prospettiva che si intende assumere. Una rappresentazione iperrealistica ci porta a scegliere la prima formula, quella del massimo rigore filologico, dello zelo assoluto: a un certo ammontare di parole, seppur ripetute a iosa, deve corrispondere il reale valore di frequenza delle parole stesse, con tutto ciò che questo implica in termini di relazioni tra parole, di frequenze e di individuazione di topics all'interno del corpus. All'opposto, il ricercatore potrebbe avere delle ottime ragioni per propendere per una prospettiva costruttivista, in virtù della quale il dato viene forgiato in ragione non già della frequenza effettiva delle parole – con le ridondanze che alcuni corpora si portano dietro per le ragioni già esposte – bensì del testo spurgato dagli elementi ridondanti. Un esempio che dovrebbe rendere palmare le implicazioni e la differenza esistente tra le due opzioni può essere tratto da un recente lavoro sui testi della canzone italiana che costituisce un aggiornamento in una direzione più spintamente sociolinguistica di un mio lavoro precedente (Nobile, 2012). Dal corpus[1] che raccoglie i testi degli artisti che sono riusciti a piazzare uno o più dischi nei primi sessanta posti delle classifiche di vendita tra gli anni '60 del Novecento e il 2016 selezioniamo i due che hanno fatto registrare il maggior numero di ingressi[2]: Mina (170 canzoni) e Renato Zero (177). Da ciascuno dei due corpora andiamo a estrarre, previa lemmatizzazione e normalizzazione del testo, le parole piene. A questo punto possiamo assegnare il rango a ciascuna di esse in base al numero di occorrenze nella prima e nella seconda situazione: quella nella

[1] Il corpus è costituito dai testi di 5940 canzoni, che hanno sviluppato 1.321.994 occorrenze, 43.855 forme grafiche diverse, 22.160 parole piene e 1.905 hapax.

[2] Per il criteri di campionamento, si veda Nobile, 2012: 51-53 o anche Nobile, *L'italiano della canzone dagli anni sessanta a oggi. Una prospettiva sociolinguistica*, in corso di pubblicazione.

quale il testo è riportato pedissequamente così come viene cantato (quindi con tutti gli elementi di ridondanza di cui si è parlato) e quella in cui esso è stato invece ripulito da questi elementi che determinano una consistente ripetizione, imputabile appunto alla struttura della canzone, di alcuni termini[3]. Il confronto tra i due ranghi, operato rispetto ai due diversi artisti, suggerisce l'uso del coefficiente di cograduazione di Spearman (ϱ). I valori ricavati dai due confronti forniscono risultati di indubbio interesse: nel caso di Mina, il valore del ϱ di Spearman è di 0,61; in quello di Renato Zero di 0,68. Questa informazione, da sola, ci fonisce un'indicazione su quanto la pulizia del testo e il rumore semantico generato dalle ridondanze possa produrre conseguenze più che tangibili nella strutturazione dei dati da elaborare: una parola che ha basso rango ha più probabilità di essere selezionata tra le parole chiave, di comparire come termine specifico di un certo sottoinsieme, di emergere come parola capace di differenziarsi in ragione del rango che essa occupa in dizionari di riferimento (De Mauro et al., 1993) e, quindi, di ergersi a indicatore della peculiarità linguistica di un determinato locutore o di una certa unità aggregata di analisi. Così, nel corpus di Mina la parola *specchio*, una volta sacrificati i ritornelli, arriva a uno scarto di rango di 165 posizioni e la parola *rabbia* perde 100 posizioni nei due diversi trattamenti del corpus. Analogamente, nel corpus di Renato Zero la parola *identikit* perde 226 posizioni a seconda che il corpus sia ripulito dalle ridondanze oppure no: essa si trova in una sola canzone (*Io uguale io*), ripetuta un'infinità di volte. Stesso discorso con la parola *fame*, che perde 183 posizioni: essa, pur essendo – al contrario di *identikit* – del tutto trasversale nel canzoniere del cantautore romano, ricorre un consistente numero di volte come tormentone della canzone *C'è fame*.

3. Conclusioni

In queste pagine si è visto che alla facilità di accesso a una quantità ciclopica di materiale testuale rinvenibile sul web non corrisponde una altrettanto disinvolta possibilità di analisi dello stesso. Da una parte, infatti, questo materiale incorpora le caratteristiche tipiche del linguaggio cosiddetto naturale e, in quanto tale, va incontro non soltanto ai comuni problemi di *machine learning* e di *text mining* (i più comuni dei quali sono riscontrabili, per esempio, nei traduttori automatici o nei programmi di riconoscimento vocale), ma anche a quelli creati dal sovradosaggio di elementi sempre più

[3] La pulizia del testo espunto dai versi duplicati è stata realizzata utilizzando una funzione del programma Excel (dati, rimuovi duplicati) tenendo fissi i riferimenti alle singole canzoni e ai diversi autori, in modo da evitare la rimozione di versi duplicati a prescindere dai due parametri di riferimento testé indicati.

diffusi come emoticons, caratteri speciali, eccetera. A questi problemi se ne possono aggiungere altri, annoverabili nell'ambito della poesia e della canzone, che rendono necessaria una fase particolarmente accurata e meditata del pre-trattamento dei testi stessi, prima che questi vengano sottoposti ad analisi. Nell'articolo si è cercato di mostrare come le scelte di ordine gnoseologico compiute a monte dal ricercatore abbiano, nel caso delle forme linguistiche peculiari di cui si è parlato, ricadute rilevanti sulle stesse risultanze empiriche. In più, le operazioni di tipo lessicometrico su materiale testuale con forte rumore semantico rischiano, se non adeguatamente supportate da una pulizia – tutt'altro che agile – del corpus spesso, di produrre risultati in cui la quota di rumore semantico rischia di essere addirittura superiore a quella del testo vettore di effettivo significato (Nobile, 2016).

Riferimenti bibliografici

Baroni M., Chantree F., Kilgarriff A. and Sharoff S. (2008). Cleaneval: A competition for cleaning webpages. *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC)* (pp. 638-643). Elda.

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica, 7*, pp. 17-53.

Chiari I. (2007). *Introduzione alla linguistica computazionale.* Laterza.

Claridge C. (2007). Constructing a corpus from the web: message boards. In M. Hundt, N. Nesselhauf, and C. Biewer, *Corpus Linguistics and the Web* (pp. 87-108). Rodopi.

De Mauro T., Mancini F., Vedovelli M. and Voghera M. (1993). *Lessico di frequenza dell'italiano parlato.* EtasLibri.

Ebner M., Altmann T. and Softic S. (2011). @twitter analysis of #edmedia10 – is the #informationstream usable for the #mass. *Form@re, 11* (74), pp. 36-45.

Lancia F. (2004). *Strumenti per l'analisi dei testi.* FrancoAngeli.

Nobile S. (2012). *Mezzo secolo di canzoni italiane. Una prospettiva sociologica (1960-2010).* Roma: Carocci.

Nobile S. (2016). Consenso e dissenso. Le reazioni degli elettori ai post dei candidati. In Morcellini M., Faggiano M.P. and Nobile S. (a cura di), *Dinamica Capitale. Traiettorie di ricerca sulle amministrative 2016* (pp. 115-138). Maggioli.

Pandolfini V. (2017). *Il sociologo e l'algoritmo. l'analisi dei dati testuali al tempo di Internet*, FrancoAngeli.

Petri S. and Tavosanis M. (2009). Building a Corpus of Italian Web Forums: Standard Encoding Issues and Linguistic Features. *JLCL, 24* (1), 115-128.

Tipaldo G. (2014). *L'analisi del contenuto e i mass media.* Il Mulino.

L'individu dans le(s) groupe(s) : focus group et partitionnement du corpus

Daniel Pélissier
Université Toulouse 1 Capitole - daniel2.pelissier@ut-capitole.fr

Abstract
Lexicometric analyzes of the focus groups depend in particular on the choice of partitioning of the corpus by researcher. After having proposed a typology of possible partitioning, we present the results of an experiment of one of these approaches on a corpus of ten focus groups. These analyzes highlight some contributions and limitations of lexicometry compared to conversational analysis.

Résumé
Les analyses lexicométriques des focus groups dépendent notamment des choix de partitionnement du corpus par le chercheur. Après avoir proposé une typologie des partitionnements possibles, nous présentons les résultats d'une expérimentation d'une de ces approches sur un corpus de dix focus groups. Ces analyses mettent en évidence certains apports et limites de la lexicométrie par rapport à l'analyse conversationnelle.

Keywords: Focus groups, partitioning, individual, group.
Mots clefs : Focus groups, partitionnement, individu, groupe.

1. Introduction

La lexicométrie a étudié d'abord des discours écrits (articles de journaux, discours politiques, etc.) et des réponses à des questions ouvertes (Lebart et Salem, 1988) puis s'est intéressée aux conversations orales retranscrites (Rouré et Reinert, 1993; Bonneau et Dister, 2010). L'analyse de ces dernières est en effet plus délicate en raison de textes en général plus courts, de syntaxes particulières. Les focus groups appartiennent à cette famille de données en posant le problème particulier du nombre important de participants. Selon certains auteurs, ce type de données est difficile à analyser avec des logiciels de lexicométrie (Duchesne et Haegel, 2014).

Pourtant, l'analyse lexicométrique a été utilisée dans plusieurs études (Guerrero et al., 2009; Grésillon et al., 2012; Hulin, 2013; Bengough et al., 2015; Brangier et al., 2015) et des articles méthodologiques ont analysé l'efficacité des traitements lexicométriques (Dransfield et al., 2004; Peyrat-

Guillard et al., 2014).
Ainsi, la possibilité de traiter les focus groups par la lexicométrie est établie. Cependant, les apports spécifiques d'une approche quantitative sont à préciser dans un domaine dominé par les approches qualitatives dont l'analyse conversationnelle. Par exemple, le lien entre focus groups et représentations sociales est mis en avant (Jovchelovitch, 2004) et la classification descendante hiérarchique (CDH) de Reinert (1983) forme des mondes lexicaux (Ratinaud et Marchand, 2015) dont la nature est proche des représentations sociales. Nous insisterons, dans cet article, sur la place de l'individu dans le(s) groupe(s), problématique que la lexicométrie permet d'approcher par un jeu de variables adapté. Mais cette analyse suppose de préparer le corpus avec des méthodes spécifiques.
Nous présenterons ainsi une typologie des méthodes de préparation d'un corpus de focus groups en complétant les analyses de Peyrat-Guillard et al. (2014) et en mettant en exergue celles centrées sur l'individu. Puis, nous analyserons les résultats de l'expérimentation d'une de ces méthodes en montrant en quoi elle permet une compréhension des discours de l'individu dans le(s) groupe(s).

2. Typologie des partitionnements d'un corpus de focus groups

Avant de commencer le traitement lexicométrique de focus groups, le corpus exige une préparation spécifique. En effet, certaines décisions de partitionnement détermineront notamment les méthodes lexicométriques employables et les analyses possibles.
Les textes des modérateurs sont souvent supprimés du focus groups (Guerrero et al., 2009 ; Peyrat-Guillard et al., 2014) car ses interventions, dans le cadre d'un focus group servent à fluidifier les échanges sans les orienter. Cependant, il peut être conseillé de comparer les résultats avec ou sans les interventions du modérateur (Peyrat-Guillard et al., 2014).
La deuxième question porte sur la partition du corpus issu du focus group. Plusieurs méthodes existent. Une première possibilité est d'analyser le focus group comme une entité sans prendre en compte les échanges entre les individus. Soit chaque focus group constitue un texte sans distinction d'individu (Dransfield et al., 2004) ; l'argument avancé par les utilisateurs de cette méthode est de faciliter les analyses statistiques mais cela n'est pas une évidence, le nombre de segments étant stable. Soit le focus group est partitionné en thèmes à partir d'une analyse de contenu (Bengough et al., 2015) ; cette approche permet de comparer par exemple les résultats d'une analyse thématique avec celle proposée au chercheur par la lexicométrie. La deuxième famille de partition est celle qui souhaite conserver les échanges du focus group. Soit la partition peut être centrée sur les individus, dite

decrowded (Peyrat-Guillard et al., 2014) ; les textes des interventions de chaque individu sont alors rassemblés (Guerrero et al., 2009). Soit chaque intervention est considérée comme un texte, approche dite *crowded* (Peyrat-Guillard et al., 2014).
Chacune de ces méthodes a des avantages et des inconvénients. Nous ne pensons pas qu'une partition soit à privilégier mais que la décision dépend des analyses envisagées par le chercheur selon sa problématique. Dans cet article, nous nous centrerons sur la deuxième famille qui permet d'étudier l'individu dans le(s) groupe(s) et pas seulement les thèmes abordés.

3. Résultats de l'expérimentation du partitionnement par locuteur

Nous avons pu expérimenter ces méthodes de partition d'un corpus de focus groups à partir d'une recherche que nous avons menée auprès de jeunes diplômés de l'enseignement supérieur (niveaux bac+3 et bac+5). Les discussions des focus groups concernaient la communication numérique de recrutement des banques et ces jeunes diplômés échangeaient sur les dispositifs utilisés par les entreprises pour recruter. Nous avons animé puis restranscrit10 focus groups de 6 à 7 personnes soit 67 locuteurs au total.

3.1. Préparation du corpus et partitionnement

Une fois les textes préparés (anonymisation, intégration des noms propres (BNP, Facebook, etc.) au dictionnaire, adaptation du dictionnaire selon les spécificités du discours, etc.), nous avons décidé de supprimer les interventions du chercheur car elles restaient neutres par rapport aux discours des jeunes diplômés que nous souhaitions analyser.
Nous avons alors créé une partition par tours de parole selon ce principe : (variables entre crochets)
[Groupe1, Ingénieurs , NUM1, 18ans, masc]: il y a des choses marquantes, il y a un site web où on n'a pas beaucoup d'informations et un autre site où il y a beaucoup d'informations.
[Groupe1, Ingénieurs , NUM2, 20ans, masc]: je suis d'accord avec toi.
En effet, nous souhaitions repérer des discours individuels dans les focus groups et pouvoir associer des variables de profil à un locuteur.
Les variables utilisées (tableau 1) ont été déterminées selon nos hypothèses de recherche et leur accessibilité puis ont été associées par un script automatique à chaque intervention de locuteur.

Tableau 1. Variables du focus groups associées aux locuteurs.

Num	Code variable	Valeur	Source	Description
1	num	1, 2, 3, etc.		Numéro de chaque intervenant
2	formation	3IL : école		Désignation du

Num	Code variable	Valeur	Source	Description
		d'ingénieur LPB : licence professionnelle banque		groupe
3	groupe	1, 2, 3, etc. 10 groupes au total		Numéro du groupe
4	sexe	M, F		
5	participation	TA, PA, A TA : très actif A : actif PA : pas actif	Statistiques SONAL selon le nombre d'interventions	Indicateur quantitatif de la participation de chaque intervenant
6	initial	STS, IUT	Données organisme de formation	Formation initiale des intervenants

Le corpus se présentait ainsi de cette façon pour être utilisé dans Iramuteq (Ratinaud, 2009) :
**** *num_44 *formation_LPB *groupe_1 *sexe_M *participation_ A *initial_STS moi je veux bien commencer. Quand je suis allé sur le site de la SG, … Les caractéristiques du corpus obtenu et traité à l'aide du logiciel Iramuteq sont alors les suivantes : 1876 textes allant d'une seule forme (Oui par exemple) pour les plus courts à 126 formes ou 280 occurrences pour le plus long, 40404 occurrences et 2094 formes au total, 21,54 occurrences par texte en moyenne, les hapax représentent 41,26% des formes. Chaque texte correspond alors à une intervention d'un locuteur dans un focus group.

3.2. Choix méthodologiques

Si la CDH de Reinert est la plus souvent citée dans la littérature (Duchesne et al., 2010 ; Gresillon et al., 2012; Hulin, 2013; Peyrat-Guillard et al., 2014; Brangier et al., 2015; Freitas et Luis, 2015, etc.) d'autres techniques sont impliquées comme l'analyse factorielle (Dransfield et al., 2004; Guerrero et al., 2009) ou plus rarement l'analyse de similitude (Bengough et al., 2015). Notre choix de la classification de Reinert est lié à nos hypothèses de recherche qui associent les discours de ces jeunes diplômés aux représentations sociales. Or, la CDH de Reinert (1983) favorise le repérage de représentations sociales (Ratinaud et Marchand, 2015). Nous avons effectué plusieurs CDH simples sur segments de texte en faisant varier le nombre de classes demandées, le nombre minimum de segments par classe. Nous avons choisi de retenir les formes dont la fréquence est supérieure à 3 (soit 687 formes dans ce cas) pour centrer le traitement sur les formes les plus présentes. Au terme de ces simulations, nous avons retenu une CDH qui présente 15 classes avec un taux de segments classés de 83,63%.

3.3. Exemple d'utilisation de variables, groupes et degré de participation

Chaque intervention ayant été associée à des variables de contexte, la méthode choisie permet de vérifier le lien existant entre les groupes et chaque classe repérée. Ainsi, pour ce corpus de focus groups, la classe 1 (Chi²=20,82, recherche d'emploi) et la classe 12 (Chi²=16,76, articles de journaux) sont associées aux étudiants de 3IL. La classe 7 (Chi²=32,17, Dupuy) et la classe 13 (Chi²=11,44, avantages et valeurs) sont plutôt liées au groupe des licences banques (fig. 1).

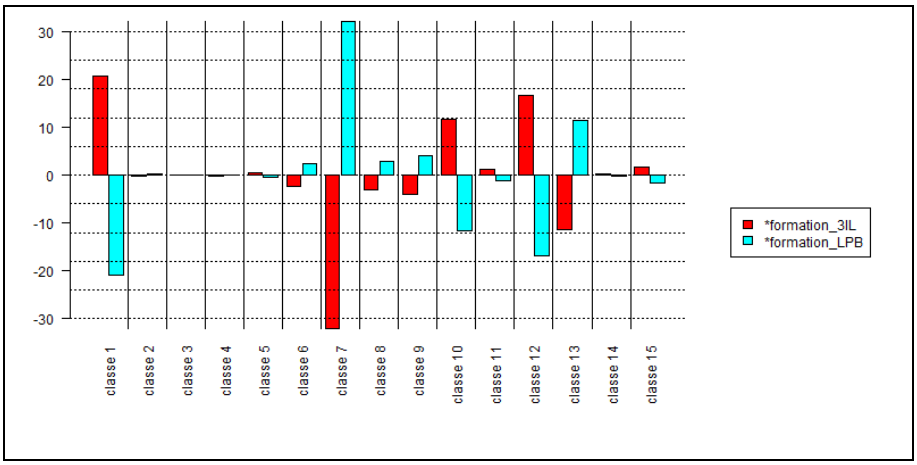


Figure 1. Chi² par classe pour la variable 'formation'.

De même, la variable sur la participation (tableau 1 et fig 2.) a permis d'associer certaines classes avec cette caractéristique. Les résultats de la CDH permettent ainsi de poser une hypothèse sur le degré de consensus entourant une représentation sociale.

eff. s.t.	eff. total	pourcentage	chi2	Type	forme
13	61	21.31	4.19		*participation_PA

Figure 2. Association de la classe 8 avec la variable participation.

En effet, la classe 8 sur la taille de l'organisation est associée aux locuteurs qui ont peu participé globalement (Variable PA (Peu Actif), Chi²=4,19 ; fig. 2) comme pour la classe 3 (mobilité). Les discussions sur la recherche d'emploi (classe 1), la banque Dupuy (classe 7) ou les classements des sites internet et témoignages sont dominées par les locuteurs les plus actifs (Variable TA (Très actif) : Chi²=5,69 pour la classe 1 et Variable A (Actif) : Chi²=7,51 pour la classe 7). Elles peuvent être perçues comme plus conflictuelles ou engagées. Les échanges sur la taille ont ainsi laissé plus de places aux locuteurs peu

actifs avec des discussions plus consensuelles moins conflictuelles que pour des représentations moins stabilisées. Cette hypothèse renvoie alors à la structure possible de cette représentation sociale construite autour d'un noyau central stable qui exigerait des études complémentaires pour être confirmée.

3.4. Repérage de discours individuels par l'analyse factorielle de correspondance (AFC)

Le partitionnement effectué permet aussi de repérer des individus dont les discours sont différents (fig. 3) grâce à une AFC réalisée à la suite d'une CDH de Reinert. Dans ce cas, deux individus se détachent principalement : 17 et 37. Le retour au texte permet de confirmer ce repérage. L'autre intérêt est aussi de souligner des regroupements d'individus différents de leur rattachement à un focus groups. L'AFC, en mettant en évidence des ensembles de locuteurs, propose une approche qui dépasse la frontière de chaque focus groups pour proposer une analyse de l'individu dans les groupes.

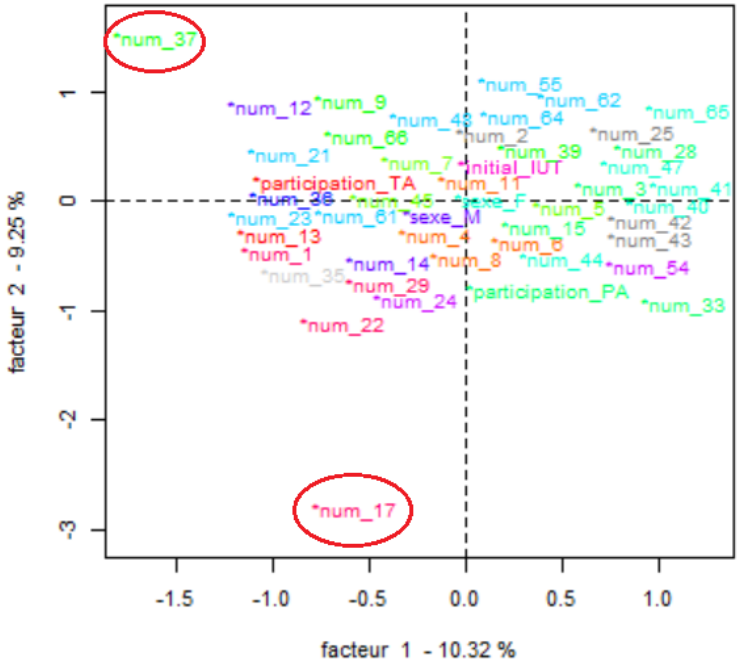


Figure 3. AFC à partir de la CDH présentant les variables (F1/F2, 19,57 % de l'inertie).

4. Conclusion

Les méthodes lexicométriques utilisées pour analyser des focus groups

dépendent notamment de la partition du corpus effectuée en amont. Dans notre recherche, l'association de variables à chaque intervention de locuteur a permis de repérer des sous-groupes d'individus à l'intérieur des focus groups, des discours d'individus isolés ou des sous-groupes associés à plusieurs focus groups qui n'apparaissaient pas de façon évidente pendant les échanges. Cette approche a cependant certaines limites. D'abord, la procédure automatisée d'association des variables utilisée dans cette expérimentation ne permet pas de repérer l'évolution des thèmes pendant la discussion, une variable repérant les tours de paroles aurait alors été nécessaire. Ensuite, le repérage des individus s'est fait sur une AFC qui explique une faible part de la variance (19,57 %) et les causes de la singularité des discours est ainsi difficile à associer à la CDH. Enfin, d'autres méthodes auraient pu être investies (analyse des antiprofils, spécificités, similitudes, etc.).

Sans remplacer l'analyse conversationnelle qui apporte des nuances spécifiques, certaines méthodes lexicométriques peuvent ainsi permettre de comprendre le corpus différemment et compléter la compréhension de ce type de données riches et profondes en dépassant notamment la frontière de chaque focus groups et faciliter une approche transversale du sens.

Remerciements : merci à Pascal Marchand, Pierre Ratinaud et Lucie Loubère pour leur initiation à la lexicométrie et à Iramuteq.

References

Bengough, T., Bovet E., Bécherraz C., Schlegel S., Burnand B., et Pidoux, V. (2015). Swiss family physicians' perceptions and attitudes towards knowledge translation practices. *BMC Family Practice*, décembre: 1–12.

Bonneau, J., and Dister, A. (2010). Logométrie et modélisation des interactions discursives, l'exemple des entretiens semi-directifs. *Journées internationales d'Analyse statistique de Données Textuelles*, pp. 253–264.

Brangier, E., Barcenilla, J., Bornet, C., Roussel, B., Vivian, R., and Bost, A. (2015). Prospective ergonomics in the ideation of hydrogen energy usages. In *Proceedings 19th Triennial Congress of the IEA*. Melbourne, pp. 1–2.

Dransfield, E., Morrot, G., Martin, J.-F., and Ngapo, T.-M. (2004). The application of a text clustering statistical analysis to aid the interpretation of focus group interviews. *Food Quality and Preference*, 15(4): 477–488.

Duchesne, S., and Haegel, F. (2014). *L'entretien collectif*. Armand Colin. Paris.

Duschesne, S., Florence Haegel, Elizabeth FRAZER, Virginie Van Ingelgom, and Guillaume Garcia, André-Paul Frognier. (2010). Europe between integration and globalisation social differences and national frames in the analysis of focus groups conducted in France, francophone Belgium and the United Kingdom. *Politique Européenne*, 30(1): 67–105.

Freitas, E. A. M., and Luis, M. A. V. (2015). Perception of students about alcohol consumption and illicit drugs. *Acta Paul Enferm.*, 28(5): 408–414.

Gresillon, E., and Marianne Cohen, Julien Lefour, Lydie Goeldner et Laurent Simon. (2012). Les trames vertes et bleues habitantes : un cheminement entre pratiques et représentations. L'exemple de la ville de Paris (France). *Développement Durable et Territoires*, 3: 2-17.

Guerrero, L., Guàrdia, M., and Xicola, J. (2009). Consumer-driven definition of traditional food products and innovation in traditional foods. A qualitative cross-cultural study. *Appetite*, 52(2): 345–354.

Hulin, T. (2013). Enseigner l'activité « écriture collaborative ». *Tic&société*, 7(1): 89–116.

Jovchelovitch, S. (2004). Contextualiser les focus groups : comprendre les groupes et les cultures dans la recherche sur les représentations. *Bulletin de Psychologie*, 57(3): 245–261.

Lebart, L., and Salem, A. (1988). *Analyse statistique des données textuelles*. Dunod. Paris.

Peyrat-Guillard, D., Lancelot Miltgen, C., et Welcomer, S. (2014). Analysing conversational data with computer-aided content analysis: The importance of data partitioning. *Journées internationales d'Analyse statistique des Données Textuelles*, pp. 519–530.

Pélissier, D. (2016), Pourquoi et comment utiliser la lexicométrie pour l'analyse de focus groups ?, *Présence numérique des organisations*, 11/07/2016.

Ratinaud, P. (2009). Iramuteq. Lerass.

Ratinaud, P., and Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les Langages du Politique*, 108(2): 57–77.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de L'analyse Des Données*, 8(2): 187–198.

Rouré, H., and Reinert, M. (1993). Analyse d'un entretien à l'aide d'une méthode d'analyse lexicale. *Journées internationales d'Analyse statistique de Données Textuelles*. ENST, Paris, pp. 418-42

Using the First Axis of a Correspondence Analysis as an Analytical Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts

Bénédicte Pincemin[1], Céline Guillot-Barbance[2], Alexei Lavrentiev[3]
Univ. Lyon, CNRS, IHRIM UMR5317 - benedicte *dot* pincemin *at* ens-lyon *dot* fr; celine *dot* guillot *at* ens-lyon *dot* fr; alexei *dot* lavrentev *at* ens-lyon *dot* fr

Abstract
Our corpus of medieval French texts is divided into 59 discourse units (DUs) which cross text genres and spoken *vs* non spoken text chunks (as tagged with *q* and *sp* TEI tags). A correspondence analysis (CA) performed on selected POS tags indicates orality as the main dimension of variation across DUs. We then design several methodological paths to investigate this gradient as computed by the CA first axis. Bootstrap is used to check the stability of observations; gradient-ordered barplots provide both a synthetic and analytic view of the correlation of any variable with the gradient; a way is also found to characterize the gradient poles (here, more-oral or less-oral poles) not only with the POS used for the CA analysis, but also with words, in order to get a more precise and lexical description. This methodology could be transposed to other data with a potential gradient structure.

Keywords: textometry, Old French, represented speech, spoken genres, methodology, correspondence analysis, 1D model, data visualization, XML TEI, TXM software, DtmVic software.

1. Linguistic issue and preparation of textual data

We investigate spoken language features of Medieval French in a corpus composed of 137 texts (4 million tokens), taken from the Base de français médiéval[1]. The corpus is annotated with part-of-speech (POS) tags at the word level; speech quotation chunks and speech turns are marked up using TEI XML tags at an intermediate level between sentences and paragraphs; and every text can be situated in a 32-genre typology (Guillot et al., 2017). Our hypothesis is that the features of orality may be related to text chunks representing speech, and also to text genres, as for instance some text genres

[1] Base de français médieval: http://bfm.ens-lyon.fr

are intended for oral performance. In order to perform a textometric analysis (Lebart et al. 1998) on our XML-TEI annotated data, we use the TXM open-source corpus analysis platform (Heiden, 2010; Heiden et al., 2010)[2].

We divide our corpus into 59 discourse units (DUs) obtained by splitting every genre into parts which represent speech on the one hand, and the remaining parts on the other hand (some text genres have no spoken passages). Discourse unit labels, like *q_rbrefLn* for instance, combine four pieces of information: (i) the first letter is either *q* for quoted speech chunks, *sp* for speech turns, or *z* for remaining (non oral) chunks; (ii) then we have the short name of the text genre (here, *rbref* means "récit bref", i. e. short narrative); (iii) the uppercase letter stands for the domain[3]; (iv) the last character indicates whether this DU is represented in our corpus by one (1), two (2) or more (*n*) texts. We linguistically represent our texts with the POS tags[4] they use[5]. The reliability of POS tags was measured in a previous study (Guillot et al., 2015) for a subset of 7 texts in which tags had been manually checked. For the present analysis, we eliminate low-frequency POS tags (freq. < 1 500), which include many high error rate tags and do not carry much weight into the quantitative analysis. For the remaining high error rate tags (with more than 25% wrong assignments), we measure their influence on the correspondence analysis (CA) by checking their contribution to the first axis. Then we remove the proper nouns category (NOMpro) which shows both high error rate and high contribution to the first axis (14.66 %).

A new correspondence analysis enables two additional improvements from a linguistic perspective. We remove compound determiners (DETcom, PRE.DETcom, like *ledit*) as they emerged at the end of the 13th century, so that they introduce a singular and substantial diachronic effect (high contributions on the first axis). Moreover, the second axis describes mainly the association between psalms (z_psautierRn) and possessive adjectives (ADJpos): this corresponds to very specific phrases with some distinctive nouns (*la meie aneme*, *li miens Deus*, *la tue misericorde*), and the adjective is equivalent to a possessive determiner in other contexts, so we merge the two categories (DETADJpos). We finally get a contingency table crossing 59 DUs with 33 POS tags to explore with a CA.

---

[2] Textometry Project and TXM software: http://textometrie.org

[3] There are 6 domains: literature (L), education (D for "didactique"), religion (R), history (H), law (J for "juridique"), practical acts (P).

[4] We use the Cattex2009 tagset, designed for Old French: http://bfm.ens-lyon.fr/spip.php?article176.

[5] We exclude punctuations, editorial markup and foreign words. CQL query: [fropos!="PON.*|ETR|OUT|RED"]

2. Linguistic and methodological results from correspondence analysis

Our study reveals that the first axis can in fact be interpreted as an orality gradient. The factorial map (Fig. 1) shows z_ DUs on the left hand side of the first axis, opposed to q_ and sp_ DUs on the right hand side. Some genres intended for oral performance go to the right with speech chunks (especially plays –dramatiqueL, dramatiqueR), whereas genres related to written processing (especially practical acts (P): charters, etc.) go to the left with out-of-speech chunks. As this opposition matches the first axis, orality appears as the first contrastive dimension for Old French (as regards POS frequencies), as it is in Biber's experiences with English (Biber, 1988), with the same kind of linguistic features (Table 1). Then, as a second result, DUs can be sorted according to their degree of orality, from "less oral" to "more oral" (see Appendix[6]). Peculiar positions (for didactic dialogs or psalms for instance) can be explained by a formal use of language given by the rules of the genre. The linguistic analysis of the DU gradient is detailed in (Guillot-Barbance et al., 2017)[7].

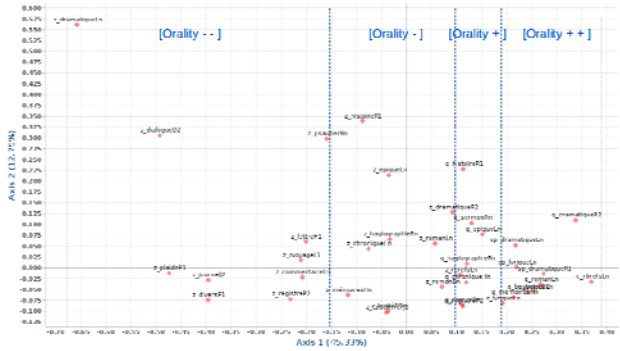


Figure 1. CA map of the 59 DUs (TXM). 21 DUs with low representation quality (cosine squared to 1 × 2 plane < 0.3) and no significant contribution to this plane (ctrb1 < 2 % & ctrb2 < 2 %) have been filtered out (macro CAfilter.groovy), so that the figure is clearer.

[6] Appendix is available online as a related file of this paper in HAL archive: https://halshs.archives-ouvertes.fr/halshs-01759219

[7] Improvements made to the statistical processing in 2018 (management of the second axis with ADJpos and DETpos merging, confidence ellipses) strengthen the linguistic interpretation published in 2017, no significant change is observed on gradient given by the first axis, according to the four zones defined by the analysis, except for a few points which are not related to this axis (low cosine squared).

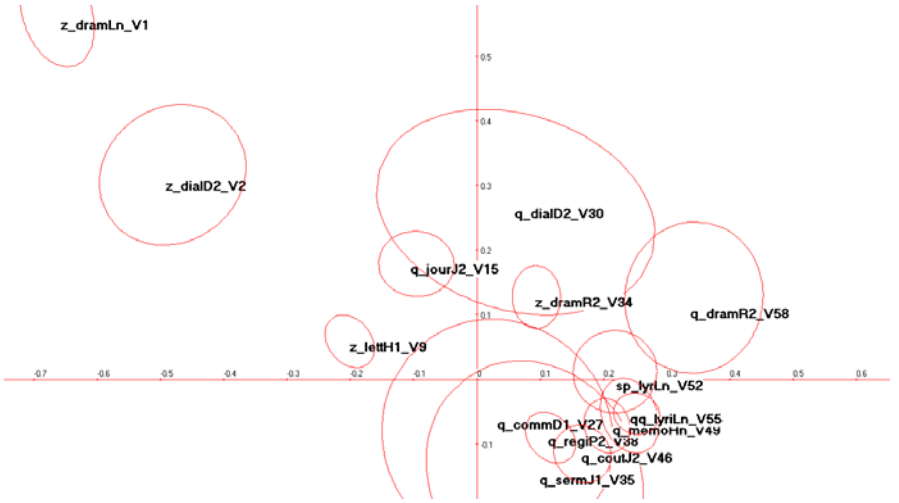


Figure 2. CA map of the 17 DUs with the largest confidence ellipses (DtmVic). The two largest ones (q_proverbesD2, q_lapidaireD2) couldn't be drawn; the following three largest ones (q_commentaireD1, q_dialogueD2, q_sermentJ1) show that these DU positions cannot be interpreted; then other smaller ellipses indicate that the 54 remaining DU positions on axes #1 and 2 are stable.

Table 1. The eight POS with the highest contributions on the first axis, for both sides.

"Less oral" pole		***"More oral" pole***	
PRE	preposition	PROper	personal pronoun
NOMcom	common noun	ADVgen	general adverb
PRE.DETdef	preposition + definite determiner	ADVneg	negative adverb
VERppe	past participle	VERcjg	finite verb
DETdef	definite determiner	PROadv	adverbial pronoun (*en*, *y*)
DETcar	cardinal determiner	DETADJpos	possessive determiner or adjective
VERppa	present participle	CONsub	subordinating conjunction
CONcoo	coordinating conjunction	VERinf	infinitive verb

A bootstrap validation (Dupuis & Lebart, 2008, Lebart & Piron, 2016) is applied to evaluate the stability of DU positions on the first axis (Figure 2). Sizes of ellipses in the 1×2 map are correlated to sizes of DUs: the fewer the words there are in the DU, the less data the statistics process, and the greater is the confidence ellipse (Table 1). Only five DUs are ascribed a big ellipse which shows their uncertain position (Figure 2): all of them are DUs from about ten words to about a hundred words, which are DUs for very singular linguistic usages, and are neither representative nor relevant for this overall linguistic analysis. The orality gradient is then confirmed throughout a

statistic validation on our data.
The 2D factorial map provides a synthetic and efficient visualization. The second axis display reveals that the "more oral" pole is more compact, more consistent, than the "less oral" pole, which is more heterogeneous (the cosine squared values corroborate this). But what we want to stress in this methodological paper, is that the main linguistic result is uniquely provided by the interpretation of the first axis. Benzécri has illustrated the same kind of approach by using a 1D CA to reveal the hierarchy of characters in Racine's *Phèdre* (1981 : 68). This method emphasizes the analytic power of CA, which separates the data (by the mathematical means of Singular Value Decomposition) into "deep" components (factors), just as a prism breaks light up into its constituent spectral colors. Despite its main use as a 2D illustration of a corpus structure in the textual data analysis field, CA is much more than a suggestive visualization or a quick sketch.

3. Complementary tools to analyse 1D gradient in textual data

We now test new means to gain insight into the causation of this gradient in our data.

3.1. Gradient-ordered barplot

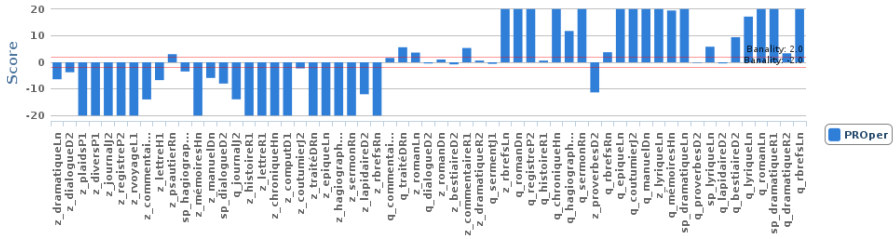


Figure 3. Gradient-ordered specificity barplot for Personal Pronoun, as example of a POS which is correlated to the first axis. For readability reasons, the height of specificity bars is limited to 20.

The first method we propose is to visualize the evolution of POS frequencies according to the orality gradient using a specificity bar-plot chart where the DU order on the x-axis is given by the DU order on the first CA axis: this display visually reveals how much a POS is correlated with speech or non speech features, and details its affinity with each DU. For instance, personal pronouns are typical for the more-oral pole: this is displayed as a rising profile (Figure 3), and one can easily find out which DU have an outlying use of this POS. Whereas a POS like adjectives (Figure 4), which is not correlated to the orality gradient, gets a chart with no overall pattern.

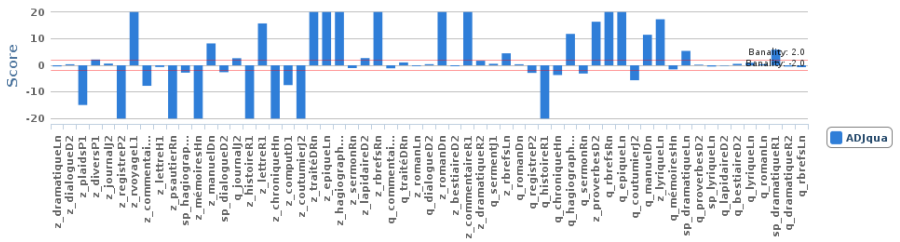


Figure 4. Gradient-ordered specificity barplot for adjectives, as example of a POS which is not correlated to the first axis. For readability reasons, the height of specificity bars is limited to 20.

3.2. Back-to-text close reading by getting representative words for each side of the first axis

The second methodological innovation concerns obtaining lexical information about orality characteristics in our texts. We select two sets of DUs based on their cosine squared scores for the first CA axis in order to represent the more-oral ($cos^2 1 > 0.4$) and less-oral ($cos^2 1 > 0.35$) poles (Table 2). The cos^2 thresholds are adjusted to get two balanced sets with enough different DUs to get an adequate representativeness. Then, a specificity computation, which statistically characterizes the distribution of words into these two sets, reveals lexical features for more oral and less oral poles, showing typical words as they can be read in texts. Light is thus shed on the quantitative result throughqualitative observations.

Table 2. Representative DUs

Less-oral pole	***More-oral pole***
z_journalJ2	q_romanLn
z_plaidsP1	sp_dramatiqueR1
z_commentaireD1	q_rbrefLn
z_diversP1	q_bestiaireD2
z_registreP2	sp_dramatiqueLn
z_lettreH1	q_lyriqueLn
z_dialogueD2	z_lyriqueLn
z_rvoyageL1	q_chroniqueHn
	sp_lyriqueLn
	q_hagiographieRn
	q_romanDn
	q_mémoiresHn

Table 3a. Adjectives typical for the less-oral subcorpus

fropos	word	F	f_z	S+_z
ADJqua	ladite	709	709	275
ADJqua	presens	432	430	162
ADJqua	Cedit	331	331	128
ADJqua	maistre	344	329	105
ADJqua	Saint	530	442	89
ADJqua	oudit	207	207	80
ADJqua	present	195	191	67
ADJqua	certeinnes	122	122	47
ADJqua	frans	239	205	46
ADJqua	VIIJ	105	105	41
ADJqua	feu	132	126	40
ADJqua	Petit	144	131	36
ADJqua	parisis	92	92	36
ADJqua	sains	242	192	33
ADJqua	Porel	68	68	26
ADJqua	GRACE	67	67	26
ADJqua	Saincte	90	83	24
ADJqua	royaulx	73	70	23
ADJqua	Perrin	64	63	23
ADJqua	yceulx	68	66	23

Table 3b. Adjectives typical for the more-oral subcorpus

fropos	word	F	f_q&sp	S+_q&sp
ADJqua	grant	3288	2594	129
ADJqua	bele	344	344	79
ADJqua	granz	288	281	54
ADJqua	biax	197	197	45
ADJqua	Biax	196	196	45
ADJqua	bone	198	195	40
ADJqua	voir	224	214	36
ADJqua	douce	128	128	29
ADJqua	Biaus	127	127	29
ADJqua	biaus	125	125	29
ADJqua	biau	122	122	28
ADJqua	sage	204	189	27
ADJqua	Bele	111	111	25
ADJqua	mal	166	156	24
ADJqua	Biau	98	98	22
ADJqua	mortel	97	97	22
ADJqua	bel	250	217	21
ADJqua	boen	88	88	20
ADJqua	meillor	87	87	20
ADJqua	las	100	98	20

Our example sheds light on the uses of adjective: whereas adjectives are not related to the orality gradient as a category (Figure 4), they have strong associations at a lexical level (Table 3). Represented speech makes much use

of terms of address introducing speech turns (*bel*, *douz* – and their formal variants: *biaus*, *biax*, etc.), and evaluative adjectives (*grant*, *mal*, *boen*). For the less-oral pole, there are more POS tagging errors; adjectives are more diverse and often associated with a subset of DUs, for instance *present*, *saint*, *maistre* are typical of two texts.

4. Conclusion

In this contribution, we have shown several ways to take into account the limits of real data, especially textual data: managing the POS tags reliability (§1), validation process to identify where data is lacking (§2), refining morphosyntatic based analysis with lexical information (§3). But our main objective is to establish a methodology in order to reveal and study any gradient-like deep structuration of data. A simple seriation (as illustrated in Dupuis & Lebart, 2008) could provide the same results for the first step, as it generates the same ordered view of the data. But CA gives much more information, qualifying the relation of each variable to the gradient with indicators like contributions and cosines squared. Interpretation can go further: CA coordinates are controlled with bootstrap and confidence ellipses, gradient-ordered barplot visualizations are efficient to analyse in detail the relationship of any individual variable to the overall gradient, and the gradient poles can be illustrated by words, which add a concrete and textual account for the deep structure. Thus, on our corpus of French medieval texts, we discover that orality is the main contrastive dimension and that it characterizes represented speech as well as text genres. The methodology could be applied to other data, and is already entirely implemented using tools freely available to the scientific community.

This research has benefited from the PaLaFra ANR-DFG project (ANR-14-FRAL-0006), for corpus extension and POS evaluation. We are also very grateful to Ludovic Lebart, for his inspiring comments on a preliminary presentation of this research, and for DtmVic software, which has evolved in order to take into account the quantitative particularities of our data.

References

Benzécri J.-P. et al. (1981). *Pratique de l'Analyse des données, tome 3. Linguistique & lexicologie.* Dunod, Bordas, Paris.

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press.

Dupuis F., Lebart L. (2008). Visualisation, validation et sériation. Application à un corpus de textes médiévaux. In Heiden S. and Pincemin B., eds, *Actes JADT 2008*, Presses univ. de Lyon: 433-444.

Guillot C., Heiden S., Lavrentiev A., Pincemin B. (2015). L'oral représenté

dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique. In Kragh K. J. and Lindschouw J., eds, *Les variations diasystémiques et leurs interdépendances dans les langues romanes -Actes du Colloque DIA II*, Éd. de linguistique et de philologie, Strasbourg : 15-28.

Guillot-Barbance C., Pincemin B., Lavrentiev A. (2017). Représentation de l'oral en français médiéval et genres textuels, *Langages*, 208: 53-68.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al., eds, *PACLIC24*, Waseda Univ., Sendai : 389-398.

Heiden S., Magué J.-Ph., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Bolasco S. et al., eds, *Statistical Analysis of Textual Data -Proceedings of JADT 2010*, Edizioni Univ. di Lettere Economia Diritto, Rome : 1021-1031.

Lebart L., Piron M. (2016). *Pratique de l'Analyse de Données Numériques et Textuelles avec Dtm-Vic*. L2C, http://www.dtmvic.com.

Lebart L., Salem A., Berry L. (1998). *Exploring Textual Data*. Kluwer academic pub., Boston.

Explorer les désaccords dans les fils de discussion du Wikipédia francophone

Céline Poudat
Université Côte d'Azur, CNRS, BCL, France – poudat@unice.fr

Abstract
This article concentrates on the exploration of French Wikipedia talk pages, with a focus on conflicts. We developed a typology of speech acts expressing disagreement, including direct and explicit forms (*je ne suis pas d'accord* / *je suis en désaccord*) as well as indirect acts, which are besides the most widespread. Disagreement is indeed a negative reaction that may threaten the face of the addressee. For this reason, disagreements are rather expressed indirectly in order to protect faces in interaction. A subset of the Wikiconflits corpus (Poudat et al., 2016) was annotated according to the typology and we carried on a primary exploration of the data using statistical methods.

Résumé
Cette étude se concentre sur l'exploration de l'encyclopédie Wikipédia, l'un des plus gros succès du Web 2.0, et spécifiquement sur l'exploration de ses discussions éditoriales, avec un intérêt particulier pour les conflits. Nous nous intéressons aux actes de langage exprimant le désaccord, de son expression la plus directe et la plus explicite (*je ne suis pas d'accord* / *je suis en désaccord*) à ses formes les plus indirectes, et d'ailleurs les plus usuelles ; le désaccord est effectivement plutôt exprimé de manière indirecte pour préserver sa face et celle de l'autre. Nous présentons la typologie que nous avons développée et nous l'appliquons à un sous-ensemble du corpus Wikiconflits que nous avons développé (Poudat et al., 2016). Le corpus annoté est ensuite exploré avec les méthodes de l'ADT et nous restituons certaines de ses caractéristiques.

Keywords: Wikipedia, CMC corpora, Conflicts, Disagreements, Pragmatics, Semantic Annotation, Text statistics

1. Introduction

Cette étude se concentre sur l'exploration de l'un des plus gros succès du Web 2.0 : l'encyclopédie Wikipédia, qui rassemble des milliers de contributeurs à travers le monde, mais qui demeure paradoxalement peu observée par les études de linguistique, certainement du fait de la complexité

de l'objet, qui multiplie les versions, les types de pages et les genres textuels. Nous nous intéressons spécifiquement aux fils des pages de discussion du Wikipédia francophone, avec un intérêt particulier pour les conflits. Plutôt abordés par les sciences sociales (cf. Kittur et Kraut, 2008, 2010; Auray et al., 2009, Sumi et al., 2011, Borra et al., 2014), les conflits dans Wikipédia ont été peu décrits d'un point de vue linguistique. Nous proposons de les décrire au moyen d'une annotation en actes de langage, en distinguant entre marqueurs du (dés)accord et marqueurs du conflit : si tout désaccord ne tourne pas au conflit, un conflit nait souvent d'un désaccord. Deux entreprises d'annotation des interactions conflictuelles de Wikipédia ont été menées ces dernières années (Bender et al., 2011, Fershke et al., 2012), mais elles ne portaient pas sur le français, et se positionnaient dans un cadre distinct. La présente communication se concentre spécifiquement sur l'exploration des marqueurs du désaccord dans Wikipédia, de son expression la plus directe et la plus explicite (*je ne suis pas d'accord* / *je suis en désaccord*) à ses formes les plus indirectes, et d'ailleurs les plus usuelles ; le désaccord est effectivement plutôt exprimé de manière indirecte pour préserver sa face et celle de l'autre. Après avoir présenté le corpus de travail (2.), nous décrirons la typologie exploratoire que nous avons développée et les marqueurs que nous avons annotés manuellement (3.). Nous présenterons enfin certaines des régularités observées (4.).

2. Wikiconflits : pages et fils conflictuels

Le corpus de travail sur lequel se fonde notre étude comprend un sous-ensemble du corpus Wikiconflits (Poudat et al., 2016), à savoir l'ensemble des discussions autour de six articles ayant été identifiés par Wikipédia comme conflictuels : *Igor et Grichka Bogdanoff*, *Chiropratique*, *Éolienne*, *Histoire de la logique*, *Psychanalyse* et *Quotient intellectuel*. La conflictualité de chaque fil a été évaluée et annotée avec une variable à trois modalités : si les fils non conflictuels sont catégorisés **C0**, **C1** signale la présence d'un désaccord et **C2** la présence d'un conflit sur le fil.

Tableau 1 : Corpus de travail

page	tokens	messages	Fils C0	Fils C1	Fils C2
Bogdanoff	73864	493	30	16	20
Chiropratique	29919	226	5	3	12
Éolienne	13454	152	2	7	0
Histoire de la logique	3358	46	4	2	0
Psychanalyse	102338	878	54	39	34
Quotient intellectuel	20059	170	10	20	12

Désaccords et conflits sont deux formes d'affrontement verbal, à cette différence que le **désaccord** est un acte **réactif** qui exprime une réaction négative relative à une assertion préalablement exprimée (Kerbrat-Orecchioni, 2016) tandis que le **conflit** est un acte **agressif**, qui implique la présence d'au moins une séquence *attaque-réplique* caractérisée par l'usage de marqueurs de violence verbale et d'actes de langage agressifs pour la face de l'allocutaire (Poudat et Ho-Dac, 2018). Ces définitions doivent être précisées relativement au genre très particulier qu'incarne la discussion Wikipédia, qui a pour fonction majeure de permettre aux rédacteurs de l'article de se coordonner et de clarifier leurs éventuels différends. L'article encyclopédique est ainsi le premier terrain de coopération entre les contributeurs, la discussion faisant plutôt office de coulisses de la rédaction – beaucoup d'utilisateurs réguliers de Wikipédia méconnaissent d'ailleurs l'existence de ces discussions. En d'autres termes, l'article est le **genre premier**, la discussion faisant figure de **genre lié** ou **non autonome**. Les désaccords et les conflits que l'on y observe s'adossent ainsi sur l'article, ce qui nous a amenée par exemple à observer qu'un désaccord pouvait porter sur un passage de l'article, considéré dans ce cas comme une assertion contestable. De la même manière, un conflit peut prendre sa source au cours de la rédaction de l'article, via une suppression ou un retour en arrière litigieux, qui pourra donner lieu à l'écriture d'une réplique agressive sur la page de discussion. Notons que nous écartons de notre étude les conflits non verbaux et autres guerres d'édition, largement observés par les sciences sociales.

Les fils catégorisés **C1** portent la trace verbale d'un désaccord tandis que les fils étiquetés **C2** contiennent au moins une attaque manifeste de la face de l'un des contributeurs du fil. Cette annotation ne va bien sûr pas de soi et nous a souvent demandé d'arbitrer entre le contenu du message et son positionnement dans le fil d'interaction. Un message peut ainsi exprimer un désaccord ou être agressif sans recevoir de réponse, tandis qu'un contributeur peut être en désaccord avec un point de vue existant qui n'est pas pour autant celui de l'un de ses co-énonciateurs. Nous n'avons retenu que les désaccords ou les attaques orientés vers le(s) co-énonciateur(s) / co-rédacteurs(s), en ce sens qu'un passage très agressif envers un tiers auteur ou article par exemple, ne sera pas été considéré comme conflictuel.

3. Le désaccord comme acte de langage : types et marqueurs

Nous nous sommes ensuite concentrée sur l'annotation manuelle des actes de langage exprimant le désaccord en développant une typologie adaptée aux caractéristiques du corpus de travail. Le désaccord étant un acte exprimant une réaction négative, il est potentiellement menaçant pour la face de l'allocutaire auquel il s'adresse. C'est pourquoi il est généralement exprimé de manière indirecte. Les chiffres sont éloquents dans notre corpus : 82% des actes exprimant le désaccord relevés sont indirects, tandis que près de la moitié des désaccords exprimés directement sont adoucis ou minimisés.

Les deux grands types d'expression indirecte du désaccord les plus récurrents que nous avons observés consistent à (i) recourir à la concession pour mettre en scène un accord partiel et (ii) exprimer son désaccord en se posant explicitement comme source évaluative (*personnellement, je ne pense pas que… ; j'avoue ne pas comprendre, etc.*). Comme nous le signalons dans le tableau 2, nous avons choisi d'annoter les concessions accompagnées d'un accord explicite comme « *Ok, mais des solutions existent (développement de pales furtives absorbant les ondes radars)* » (discussion Éolienne), ce qui explique peut-être pourquoi au final nous n'en obtenons qu'un petit nombre (9 occ.). L'expression du désaccord indirect semble privilégier significativement les actes secondaires de l'**incompréhension** (48 occ.) et de l'**expression d'une opinion** (29 occ.). À titre de comparaison, nous avons systématiquement annoté les manifestations d'accord explicites rencontrées. Contrairement au désaccord, l'accord est dans notre culture un acte positif pour la face de l'allocutaire. Peu employé de manière indirecte, il est plutôt intensifié qu'atténué (*je suis **tout à fait** d'accord*). On relève 57 actes d'accord explicite dans le corpus ; à titre de comparaison, on rencontre trois fois plus de formes exprimant un désaccord, ce qui est probablement dû à la dimension conflictuelle du corpus. Il nous faut enfin souligner que plus des deux tiers des 270 fils de discussion considérés ne contenaient aucune des formes observées, ce qui n'est pas surprenant : un quart des fils ne contiennent qu'un seul message tandis que nous avons conservé les fils catégorisés harmonieux à titre de contraste.

Tableau 2 : Typologie du désaccord

Attributs	Valeurs	Exemples
polarité	**accord**	*je suis d'accord*
	désaccord	*Je suis contre l'avis de X*
type	**explicite**	**Accord explicite** : *je suis d'accord, je suis pour X, favorable à X, tout à fait de votre avis, je suis de ton avis, OK pour X…* **Désaccord explicite** : *pas d'accord, en désaccord, je ne suis pas favorable, je suis contre, totalement contre*
	implicite	Voir acte **indirect**.

atténuation	**oui / non**	**Atténuation d'un accord explicite** : je suis ***assez*** d'accord **Atténuation d'un désaccord explicite** : Nous sommes en désaccord ***(mineur)*** sur un point ***(mineur)***
indirect	**non**	
Concéder	**concession**	Seuls les actes d'accord explicite accompagnés d'une concession ont été retenus. ***D'accord*** *pour refuser le paragraphe ajouté à partir d'arkiv ;* ***en revanche*** *la suppression de la participation d'AR à la mission ne me semblait pas déraisonnable* (discussion Bogdanoff)
Se poser comme source évaluative	**avis**	« Personnellement, je pense que non », je ne crois pas, je ne pense pas… mots-clés : personnellement, pense, crois, trouve
	émotion	***émotion*** *(rare dans le corpus pour exprimer le désaccord)* ***j'ai été personnellement choqué*** *par les affirmations gratuites comme "de gauche/de droite" dès le début de l'article, que je pense tout à fait intempestives et parfaitement corrélées à la hauteur du QI du contributeur et aux théories raciales de Rushton,* (discussion QI)
	doute	*Je doute de la pertinence de ce passage dans cet article.* mots-clés : certain, sûr, doute
	Incompré-hension	*Je ne vois pas bien quel rapport ta source a avec ce constat.* (discussion Psychanalyse) *Encore une fois, je ne comprends pas le problème.*
	assertion négative forte	Ce n'est **pas du tout** une question de vocabulaire secondaire (discussion Bogdanoff)

4. Analyses

Le corpus annoté a ensuite été soumis à différentes méthodes de l'analyse de données textuelles afin d'explorer ses caractéristiques et de mettre en évidence les relations entre les types de désaccord et la situation du fil, harmonieuse, dissonante ou conflictuelle. Comme le montre la Figure 1, les fils identifiés comme lieux d'un désaccord (C1) sont ceux qui contiennent le nombre le plus significatif de marqueurs d'accord et de désaccord. Au contraire, les fils identifiés comme conflictuels contiennent significativement moins de marques d'accord explicite et de marques de désaccord. Nous voilà donc rassurée par la cohérence de notre annotation.

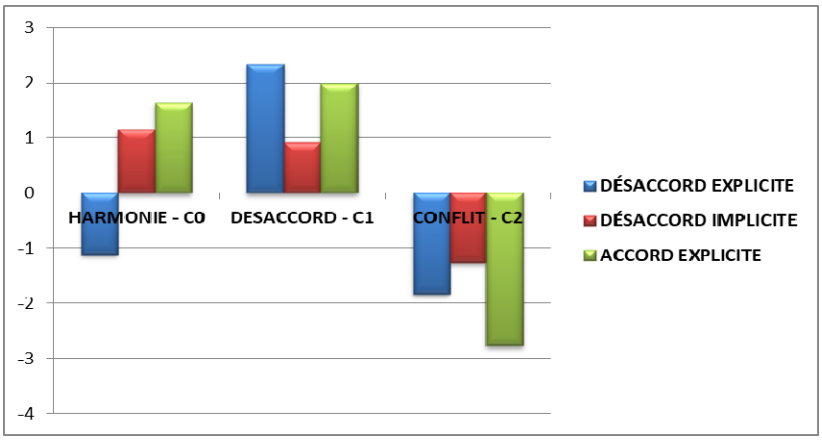


Figure 1 : Ventilation des types d'accord et de désaccord d'un type de fil à l'autre (données Hyperbase Web)

Afin d'évaluer plus précisément la structure de l'ensemble des annotations apposées sur les textes, nous avons réalisé une Analyse en Composantes Principales (ACP) sur la table des décomptes d'annotations en prenant le fil de discussion comme unité textuelle. Nous avons dû procéder à certains ajustements, (i) en écartant les fils qui ne contenaient aucune annotation ; (ii) en isolant certaines variables trop marginales (*i.e.* 2 occ. de la valeur *émotion*) et (iii) en distinguant entre les observations restantes celles qui seront utilisées comme variables actives ou comme variables supplémentaires. Ainsi, les variables ayant le trait *atténuation* ont été intégrées à titre illustratif. Au total, l'ACP a été réalisé sur un ensemble de taille restreinte, à savoir 98 fils * 8 variables actives (et 13 variables supplémentaires). De manière intéressante, l'ACP met en évidence la présence d'un **facteur taille**, c'est-à-dire que **toutes les observations sont corrélées positivement entre elles** et se regroupent donc du même côté du premier axe factoriel. Certains fils de discussion ont des valeurs fortes pour toutes les variables, tandis que d'autres ont des valeurs faibles pour toutes les variables.

Si l'on s'intéresse aux facteurs 2 et 3 (Figure 2) sur lesquels on projette le degré de conflictualité et les pages du corpus à titre illustratif, on observe une **opposition entre accord et désaccord**, et dans une moindre mesure entre **explicite et implicite** sur le facteur 2. Accords et actes explicites seraient du côté de l'harmonie et du désaccord tandis que les désaccords en général et les désaccords indirects en particulier seraient plus caractéristiques du conflit. Cette dernière remarque, qui devra être éprouvée et confirmée sur des jeux de données plus importants, nous semble intéressante : est-ce que les marqueurs implicites du désaccord vont de pair avec les marqueurs du conflit ? Y a-t-il une corrélation négative entre expression explicite du

désaccord et attaques personnelles ?

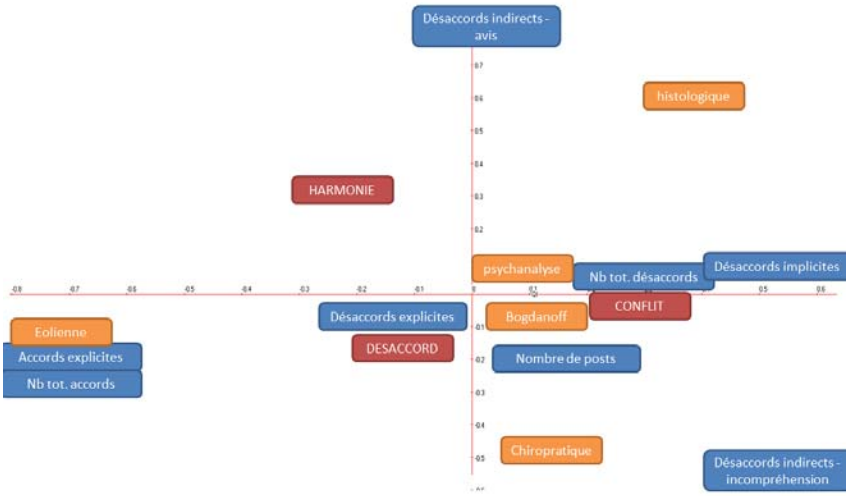


*Figure 2 : Facteurs 2 et 3 de l'ACP – 98 fils * 8 variables actives – Dtm-vic*

5. Conclusion et perspectives

Nous avons ainsi proposé une première typologie des actes exprimant le désaccord en français ; cette typologie a été développée dans le cadre d'un projet plus général d'exploration des conflits dans Wikipédia. Une seconde typologie, centrée sur les marqueurs de violence verbale et supposément caractéristique du conflit, est en cours de développement et viendra faire système avec la typologie du désaccord pour mettre en évidence les caractéristiques des interactions conflictuelles dans Wikipédia et dans les CMC.

En ce qui concerne l'annotation présentée, un guide est actuellement en cours de rédaction ; chaque marqueur sera validé et évalué au moyen d'un kappa de Cohen. La typologie est encore en cours d'amélioration ; ainsi une troisième forme d'expression indirecte du désaccord que nous avions observée consiste à le neutraliser en déplaçant le focus sur une proposition ou une suggestion, *i.e.* un acte de langage positif (*ne vaudrait-il pas mieux… ? Il faudrait peut-être d'abord définir ce qu'on entend par..)*. Ce type de séquence, plus complexe à identifier car plus ambigu, est en cours d'intégration.

Enfin, reste à mettre en œuvre des parcours interprétatifs adaptés pour explorer ce type de données annotées avec nos méthodes ADT ; c'est aussi l'une des pistes que nous poursuivons ces dernières années, dans nos travaux (Poudat et Landragin, 2017) et dans le cadre du consortium CORLI.

Références

Auray, N., Hurault-Plantet, M., Poudat, C., & Jacquemin, B. (2009). La négociation des points de vue : une cartographie sociale des conflits et des querelles dans le Wikipédia francophone. In *Réseaux* 2/2009, n° 154: 15-50.

Bender E.M., Morgan J.T., Oxley M., Zachry M., Hutchinson B., Marin, A., Ostendorf, M. (2011). Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. In *Proceedings of the Workshop on Languages in Social Media* (pp. 48–57). Stroudsburg, PA, USA: Association for Computational Linguistics.

Borra E., Weltevrede E., Ciuccarelli P., Kaltenbrunner A., Laniado D., Magni G., Venturini T. (2014). Contropedia - the Analysis and Visualization of Controversies in Wikipedia Articles. In *Proceedings of The International Symposium on Open Collaboration* (pp. 34:1–34:1). New York, NY, USA.

Ferschke O., Gurevych I., Chebotar Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 777–786). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kerbrat-Orecchioni, C. (2016). Le désaccord, réaction « non préférée » ? Le cas des débats présidentiels. *Cahiers de praxématique*, (67).

Poudat C. et Ho-Dac L.-M. (2018). Désaccords et conflits dans le Wikipédia francophone. In *Travaux linguistiques du Cerlico*, Presses Universitaires de Rennes (sous presse).

Poudat C. et Landragin F. (2017). *Explorer un corpus textuel. Méthodes – Pratiques – Outils*. Collection Champs linguistiques, De Boeck, Louvain-la-Neuve.

Poudat C., Grabar N., Paloque-Berges C., Chanier T. et Kun J. (2017). Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In Wigham, C.R & Ledegen, G., *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Collection Humanités numériques. Paris : L'Harmattan, pp. 19-36.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., & Kertész, J. (2011). Edit wars in Wikipedia. In: Proceedings of the ACM WebSci'11, Koblenz, Germany. pp. 1–3.

Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes

Matthieu Quignard[1], Serge Heiden[2], Frédéric Landragin[3], Matthieu Decorde[2]
[1]ICAR, CNRS, University of Lyon – matthieu.quignard@ens-lyon.fr
[2]IHRIM, ENS Lyon, CNRS, University of Lyon – {slh,matthieu.decorde}@ens-lyon.fr
[3]Lattice, CNRS, ENS Paris, University Sorbonne Nouvelle, PSL Research University, USPC – frederic.landragin@ens.fr

Abstract
In this article we present a set of measures – some of which can lead to specific visualisations – with the objective to enrich the possibilities of exploration and exploitation of annotated data, and in particular coreference chains. We first present a specific use of the well-known concordancer, which is here adapted to present the elements of a coreference chain. We then present a histogram generator that allows for example to display the distribution of the various coreference chains of a text, given a value from the annotated properties. Finally, we present what we call progress diagrams, whose purpose is to display the progress of each chain throughout the text. We conclude on the interest of these (interactive) modes of visualization in order to make the annotation phase more controlled and more effective.

Résumé
Nous présentons dans cet article un ensemble de mesures – dont certaines peuvent amener à des visualisations spécifiques – dont l'objectif est d'enrichir les possibilités d'exploration et d'exploitation des données annotées, en particulier quand il s'agit de chaînes de coréférences. Nous présentons tout d'abord une utilisation adaptée de l'outil bien connu qu'est le concordancier, en n'affichant que les maillons d'une chaîne choisie. Puis nous montrons un générateur d'histogramme qui permet par exemple d'afficher la répartition des chaînes de coréférences d'un texte à partir d'une propriété annotée. Nous montrons enfin ce que nous appelons des diagrammes de progression, dont le but est d'afficher les avancées au fur et à mesure du texte des chaînes de coréférences qu'il contient. Nous concluons sur l'intérêt de ces modes (interactifs) de visualisation pour rendre la phase d'annotation plus maîtrisée et plus efficace.
Keywords: coreference chain, corpus annotation, annotation tool, visualisation tool, exploration tool, statistical analysis of textual data.

1. Introduction

The manual annotation of a textual corpus with referring expressions (Charolles, 2002) and coreference chains (Schnedecker, 1997, Landragin & Schnedecker, 2014) requires adapted tools. A coreference chain can cover the whole text; it is therefore a linguistic object for which the existing means of visualization and exploration are few and often perfectible. The MMAX2 tool (Müller & Strube, 2006) allows for visualizing the links between referring expressions using arrows which link markables. The GLOZZ tool (Mathet & Wildlöcher, 2009) offers several means of visualization: with arrows like MMAX2, or with a specific marking in the margin or the middle of the text. The ANALEC tool (Landragin *et al*., 2012) and its specific extension for coreference chains (Landragin, 2016) proposes a graphic metaphor based on the succession of coloured dots. This allows the analyst to configure visual parameters, for instance the colour which can be linked to any of the annotated properties. This type of visualization makes it possible to see at a glance the structural differences between the different reference chains of a text. That must be useful to the analyst, in addition to manual explorations and finer linguistic analyses.

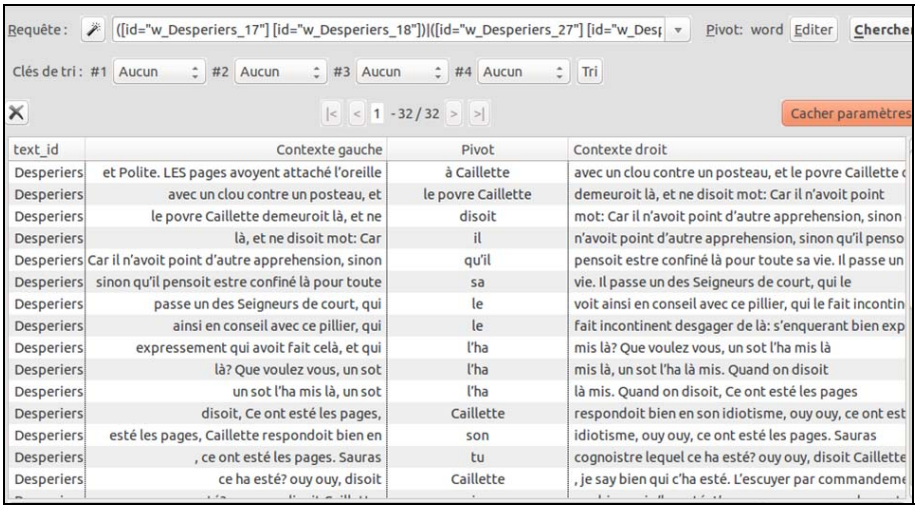
2. Linguistic objects and methodology

In the continuity of previous works (Heiden, 2010; Landragin, 2016), we present here a set of measures – some of which can lead to specific visualisations – with the objective to enrich the possibilities of exploration and exploitation of annotated data. We focus in particular on annotations which concern discursive phenomena like coreference, i.e., annotations which are necessarily described within two levels: 1. markable, group of contiguous words to which is assigned some labels, using for instance a feature structure; 2. set of markables, or links between markables, as is it the case for any chain of annotations: anaphoric chains, textual organizers chains, textual structure elements chains, etc. A feature structure can also be assigned at level 2, i.e., to the set or to the links.

3. A concordancer adapted to annotations chains

As a first visualization mode, we reuse the very classic concordancer to display the elements which constitute a coreference chain. The use of such a visualization tool, which is well established in the community of corpus exploration (Poudat & Landragin, 2017), seemed natural for visualizing chains of annotations. The last version of TXM (Heiden, 2010) thus includes a concordancer which makes it possible to display in a column all the elements (e.g. referring expressions) of a chain (e.g. coreference chain), with left and right contexts for each elements. Compared to MMAX2 (Müller & Strube,

2006) and GLOZZ (Mathet & Wildlöcher, 2009) visualisation choices, i.e. arrows linking marquables which are displayed directly on the text, this concordancer has the advantage of regrouping all the relevant information in a small graphic space.



text_id	Contexte gauche	Pivot	Contexte droit
Desperiers	et Polite. LES pages avoyent attaché l'oreille	à Caillette	avec un clou contre un posteau, et le povre Caillette
Desperiers	avec un clou contre un posteau, et	le povre Caillette	demeuroit là, et ne disoit mot: Car il n'avoit point
Desperiers	le povre Caillette demeuroit là, et ne	disoit	mot: Car il n'avoit point d'autre apprehension, sinon
Desperiers	là, et ne disoit mot: Car	il	n'avoit point d'autre apprehension, sinon qu'il penso
Desperiers	Car il n'avoit point d'autre apprehension, sinon	qu'il	pensoit estre confiné là pour toute sa vie. Il passe un
Desperiers	sinon qu'il pensoit estre confiné là pour toute	sa	vie. Il passe un des Seigneurs de court, qui le
Desperiers	passe un des Seigneurs de court, qui	le	voit ainsi en conseil avec ce pillier, qui le fait incontin
Desperiers	ainsi en conseil avec ce pillier, qui	le	fait incontinent desgager de là: s'enquerant bien exp
Desperiers	expressement qui avoit fait celà, et qui	l'ha	mis là? Que voulez vous, un sot l'ha mis là
Desperiers	là? Que voulez vous, un sot	l'ha	mis là, un sot l'ha là mis. Quand on disoit
Desperiers	un sot l'ha mis là, un sot	l'ha	là mis. Quand on disoit, Ce ont esté les pages
Desperiers	disoit, Ce ont esté les pages,	Caillette	respondoit bien en son idiotisme, ouy ouy, ce ont est
Desperiers	esté les pages, Caillette respondoit bien en	son	idiotisme, ouy ouy, ce ont esté les pages. Sauras
Desperiers	, ce ont esté les pages. Sauras	tu	cognoistre lequel ce ha esté? ouy ouy, disoit Caillette
Desperiers	ce ha esté? ouy ouy, disoit	Caillette	, je say bien qui c'ha esté. L'escuyer par commandeme

Fig 1: Concordancer with the elements of a coreference chain, dedicated to a character named "Caillette".

Fig. 1 shows the list of all referring expression to the character 'Caillette'. Sorted in the textual order, the concordancer shows the alternation of the use of proper nouns, pronouns, possessives, etc. This concordancer may also be sorted along a given property of the marquable, e.g. its POS label. This representation may then be exploited to see whether the POS annotation is consistent or not.

4. Histograms for visualising distributions of annotations chains

A second mode of visualization, also very traditional, is the histogram (bar plot). The user can select one or several properties – the determination of the referring expressions, for instance, or the type of referent – and launch calculations on their occurrences: cross-counts, correlation computation and so on. TXM now includes a histogram generator, which allows for example to display the distribution of coreference chains throughout the text, as well as the distribution of chains according to the number of referring expressions they include. These calculations and their associated visualizations provide TXM with integrated functionalities which required in other state-of-the-art tools the development of scripts, in order to export the relevant data and exploit them in an external tool like a spreadsheet.

Figure 2 compares the distribution of grammatical categories of referring expressions in three texts. Although all texts are all encyclopedical ones, the Discourse from Bossuet shows a particular profile, with a high number of proper nouns (GN.NAM).

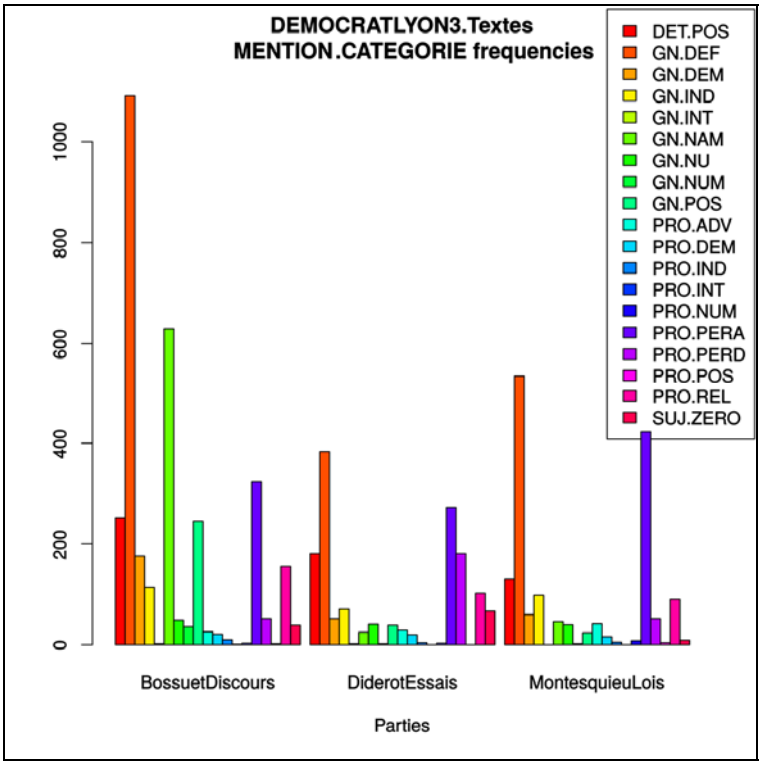


Fig 2: Comparative barplots of grammatical categories usage by reference units in three texts: Bossuet, "Discours sur l'histoire universelle" (1681), Diderot, "Essais sur la peinture" (1759-1766), Montesquieu, "Esprit des lois" (1755).

5. Progression charts for annotations chains

A third (new) mode of visualization consists to graphically show the progress of each chain throughout the text. The principle is simple, but the possibilities of exploration and exploitation of the generated graph are numerous. In a two-dimensional chart the abscissa of which represents the linearity of the text, chains are displayed point by point (cf. Fig. 3): each occurrence of a referring expression increases by one notch the ordinate of the corresponding point. The resulting broken lines are all ascending but can considerably vary in their areas of progression and flat areas.

When they are visualized simultaneously, it is possible to detect the parts of

the text where several referents are competitors, or on the contrary those where several referents appear alternately. Zooming (in and out) as well as focussing features allows for visualizing the characteristics of each point, thus enriching the exploration possibilities of these progression chart and the underlying coreference chains.

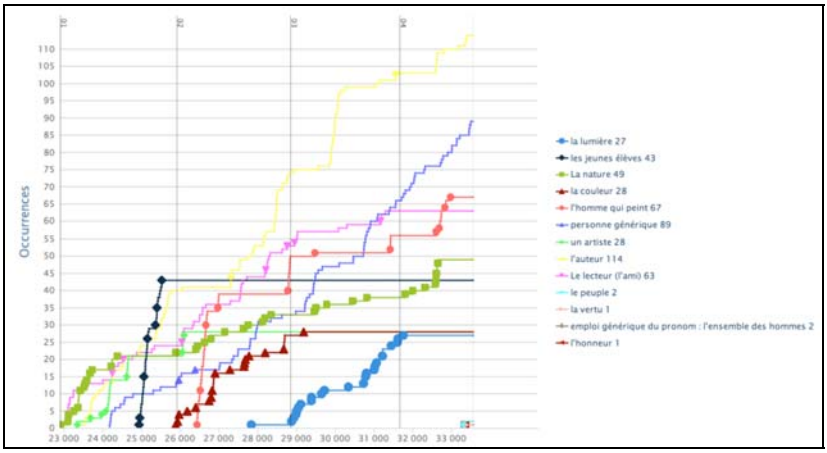


Fig 3: Progression graph of the main coreference chains at the beginning of "Essais sur la peinture" from Denis Diderot. The dots highlighted with symbols correspond to referring expressions with low accessibility.

6. Discussion

The common points of these new visualization modes is not only to propose visual representations which are easy to understand (and possibly interactive, when it is possible to modify on the fly one of the properties), to allow the visualization of these representations directly in TXM, with no need to export annotated data and to use external tools, but also to facilitate the detection by the analyst of intruders, outliers and deviant examples. For instance potential annotation errors: it can be the case for a referring expression which has nothing to do in the currently visualised chain. It may be a peak or a suspect flat in one of the generated histograms. It may be a zone with a very high slope (or a very long flat) in a progression diagram. In all three cases, the analyst can directly access the suspicious annotation, in order to verify it and of course to modify it. The integration of the measurements and their visualizations in TXM allows this immediate return to the corpus annotation phase. This is particularly effective when the corpus is being annotated manually.

7. Conclusion and future works

One can say that it is by annotating that we can see the mistakes we make, but we still need appropriate tools to detect these errors. With the new possibilities of interaction that we propose here, we hope that we are taking a significant step in this direction. The first tests which we have carried out demonstrated the relevance of our approach.

References

Charolles M. (2002). *La référence et les expressions référentielles en français*. Ophrys, Paris, France.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 389-398, available at halshs.archives-ouvertes.fr/halshs-00549764.

Landragin F. (2016). Conception d'un outil de visualisation et d'exploration de chaînes de coréférences. *Statistical Analysis of Textual Data – Proceedings of 13th International Conference Journées d'Analyse statistique des Données Textuelles (JADT 2016)*, Nice, France, pp. 109-120.

Landragin F., Poibeau T. and Victorri B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. *Proceedings of LREC 2012*, Istanbul, Turkey, pp. 357-362.

Landragin F. and Schnedecker C., editors (2014). *Les chaînes de référence*. Volume 195 of the *Langages* journal, Armand Colin, Paris, France.

Müller C. and Strube M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun S., Kohn K. and Mukherjee J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Peter Lang, Frankfurt, Germany.

Poudat, C. and Landragin, F. (2017). Explorer un corpus textuel : méthodes, pratiques, outils. Champs Linguistiques. De Boeck Supérieur : Louvain-la-Neuve.

Schnedecker C. (1997). *Nom propre et chaîne de référence*. Klincksieck, Paris, France.

Widlöcher A. and Mathet Y. (2012). The Glozz platform: a corpus annotation and mining tool. In Concolato C. and Schmitz P, editors, *Proceedings of the ACM Symposium on Document Engineering (DocEng'12)*, Paris, France, pp. 171-180.

Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ

Pierre Ratinaud
LERASS, Université de Toulouse – ratinaud@univ-tlse2.fr

Abstract
This work presents a proposal to improve the accuracy and the speed of execution of the divisive hierarchical clustering (DHC) algorithm used by the Reinert method implemented in the IRaMuTeQ free software. The DHC of the Reinert method is a serie of bi-partitions on a presence / absence matrix that intersects text segments and words. In the original version of this algorithm, after each partition, the largest of the remaining classes is selected to be split. We propose to replace the selection mode of the classes to be partitioned by a criteria of homogeneity. The complete rewriting of this part of the IRaMuTeQ code has also been an opportunity to improve its speed by implementing part of the code in C ++ and paralleling the procedure. An experiment carried out on 6 corpora shows that the new algorithm based on these principles is indeed more precise and faster.

Résumé
Ce travail présente une proposition d'amélioration de la précision et de la vitesse d'exécution de l'algorithme de classification hiérarchique descendante (CHD) utilisé par la méthode Reinert implémentée dans le logiciel libre IRaMuTeQ. La CHD de la méthode Reinert est une série de bi-partitions de matrices de présence / absence qui croise des segments de texte et des formes. Dans la version originale de cet algorithme, après chaque partition, la plus grande des classes restantes est sélectionnée pour être à son tour coupée en deux. Nous proposons de remplacer le mode de sélection des classes à partitionner par un critère d'homogénéité. La ré-écriture complète de cette partie du code d'IRaMuTeQ a également été l'occasion d'une amélioration de sa célérité par l'implémentation d'une partie du code en C++ et la parallélisation de la procédure. Une expérimentation menée sur 6 corpus permet de constater que le nouvel algorithme reposant sur ces principes est effectivement plus précis et plus rapide.

Keywords: méthode Reinert, classification hiérarchique descendante, IraMuTeQ, précision

1. Introduction

La méthode Reinert a pour objectif de faire émerger les différentes thématiques qui traversent un corpus textuel. Sa plus grande originalité est sûrement l'algorithme de classification hiérarchique descendante (CHD) proposé par Reinert (1983). Après avoir rappelé les différentes étapes de ce type d'analyse, nous proposerons une modification de cet algorithme de classification dans l'objectif d'améliorer la précision de l'ensemble de la procédure. Le changement proposé concerne le critère de sélection des sous-matrices après chacune des partitions. La description de cette nouvelle procédure est complétée par une expérimentation sur 6 corpus en français et en anglais permettant de comparer la nouvelle version de l'algorithme avec l'ancienne. Les résultats que nous présentons attestent effectivement d'une augmentation de la précision de l'algorithme, dont la ré-écriture à également permis une augmentation de la vitesse d'exécution. Avant d'entamer cette présentation, il nous semble toutefois nécessaire de rappeler que la CHD n'est pas la seule particularité de la méthode Reinert.

2. Des corpus aux matrices

Une autre originalité de cette procédure est l'unité utilisée dans la classification. Dans la plupart des situations, la classification ne porte pas sur les textes dans leur ensemble, mais sur une granularité inférieure. Les unités classées sont des segments de texte. Dans le logiciel IRaMuTeQ (Ratinaud, 2014; Ratinaud & Marchand, 2012), la taille de ces segments est fixée par défaut à 40 occurrences et leur découpage tient compte de la ponctuation. La règle de découpage essaie donc de proposer des unités de taille homogène (autour de 40 occurrences) et de respecter le découpage « naturel » des textes marqué par la ponctuation. Une seconde originalité qu'il convient de préciser est la distinction opérée entre formes pleines et mots outils. Dans ces analyses, la plupart du temps, seules les formes pleines (verbes, adverbes, adjectifs et substantifs) sont considérées. Les corpus peuvent alors être représentés sous la forme de matrices qui croisent les segments de texte et les formes pleines. Les cellules de ces matrices marquent la présence ou l'absence des formes dans les segments en codant 1 la présence et 0 l'absence. Le tableau 1 présente une telle matrice pour un corpus composé de 10 segments de texte (notés i_1 à i_{10}) et de 9 formes (notées j_1 à j_9).

Tableau 1 : Exemple d'une matrice croisant des segments de texte (en ligne) et les formes (en colonne)

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
I_1	1	1	1	1	0	0	0	0	0
I_2	0	0	0	0	1	1	1	1	1
I_3	0	0	1	0	1	0	1	0	0
I_4	1	0	1	0	1	0	0	0	1
I_5	0	0	1	0	1	0	1	0	0
I_6	1	1	1	1	0	0	0	0	1
I_7	0	0	0	0	1	1	1	1	0
I_8	1	0	1	0	1	0	0	0	0
I_9	0	0	1	0	1	0	1	0	1
I_{10}	0	0	1	0	1	0	1	0	0

La matrice présentée dans le tableau 1 est un exemple très simplifié de ce qu'il se passe dans la réalité. Les matrices générées sur des corpus textuels sont beaucoup plus grandes et beaucoup plus « creuses » (la proportion de 1 est très faible dans la matrice). Nous noterons N le nombre total de 1 dans la matrice. L'objectif de la classification est de proposer une réorganisation de cette matrice en sous-groupes de segments qui maximisent les propriétés suivantes :

n) Les segments regroupés doivent être homogènes entre eux : la méthode doit réunir les segments de texte qui se ressemblent, c'est-à-dire les segments qui ont tendance à contenir les mêmes mots.

o) Les ensembles doivent être hétérogènes entre eux : les groupes de segments constitués doivent être les plus différents possibles.

L'illustration 1 propose un découpage de la matrice présentée dans le Tableau 1 en 4 classes qui respectent ces critères.

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
I_1	1	1	1	1	0	0	0	0	0
I_6	1	1	1	1	0	0	0	0	1

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
I_8	1	0	1	0	1	0	0	0	0
I_4	1	0	1	0	1	0	0	0	1

Illustration 1 : Découpage de la matrice du Tableau 1 en 4 classes

La « qualité » de cette solution peut être déterminée par le calcul du chi2/N du tableau réduit (Reinert, 1983).

Dans cet exemple, la solution optimale serait obtenue en séparant les lignes i_6, i_4, i_2 et i_9 de leur classe d'appartenance pour les laisser former leur propre classe. La solution à 8 classes obtenue résumerait alors l'intégralité de l'information contenue dans la matrice du Tableau 1.

Tableau 2 : Tableau réduit de la classification de l'illustration 1

	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
Σ [i_1,i_6]	2	2	2	2	0	0	0	0	1
Σ [i_4,i_8]	2	0	2	0	2	0	0	0	1
Σ [i_9,i_3,i_5,i_{10}]	0	0	4	0	4	0	4	0	1
Σ [i_2,i_7]	0	0	0	0	2	2	2	2	1

3. La CHD de la méthode Reinert

Rappelons que la méthode permettant de construire automatiquement ces classes s'appuie sur une série de bi-partitions reposant chacune sur une analyse factorielle des correspondances (AFC). La première coupure est obtenue en cherchant le long du premier facteur de cette AFC les deux sous-matrices qui maximisent le chi2/N du tableau réduit. La partition produite est améliorée en inversant chacune des lignes du tableau d'une classe à l'autre et en recalculant le chi2/N du tableau réduit. Toutes les inversions qui augmentent la valeur du chi2/N sont conservées. Cette étape boucle jusqu'à ce que plus aucune inversion n'augmente cette valeur. Une dernière étape consiste à retirer les formes (les colonnes) statistiquement sous-représentées dans les matrices (sur la base d'un chi2).

Cette procédure (bi-partition de la matrice, inversion des lignes, suppression des colonnes) constitue une des partitions de la CHD. La CHD dans son ensemble réalisera cette procédure autant de fois que nécessaire pour atteindre le nombre de classes terminales paramétré. Il faut n-1 partition(s) pour constituer n classe(s) terminale(s). Après chacune de ces partitions, dans sa formulation d'origine, l'algorithme sélectionne la plus grande des classes constituées (c'est-à-dire celle qui contient le plus de lignes) pour lui faire à son tour subir une partition.

Le tableau 3 présente, de façon très caricaturale, une matrice pour laquelle cette démarche ne conduit pas à un résultat satisfaisant. Si nous soumettions cette matrice à la CHD précédemment décrite, la première partition conduirait à la création d'une classe constituée des lignes i_1, i_2 et i_3 (notée [i_1,i_2,i_3]) et d'une autre constituée des lignes i_4 et i_5 (notée [i_4,i_5]). La première de ces classes étant la plus grande, elle serait sélectionnée pour, à son tour, subir une partition. Or, il est évident ici qu'il n'y a plus aucune information à extraire de cette matrice, les lignes étant toutes identiques. Seule la séparation des lignes i_4 et i_5 est, dans cet exemple, susceptible d'augmenter la qualité du résultat. Pour cela, il aurait donc fallu sélectionner la classe restante la plus hétérogène ([i_4,i_5]) plutôt que de sélectionner la plus grande ([i_1,i_2,i_3]).

Tableau 3 : Une matrice problématique

	J_1	J_2	J_3	J_4	J_5
i_1	1	1	0	0	0
i_2	1	1	0	0	0
i_3	1	1	0	0	0
i_4	0	0	1	1	0
i_5	0	0	0	1	1

Il convient donc de percevoir que dans la version actuellement disponible de cette méthode, l'algorithme de classification fait l'hypothèse que la matrice la plus grande est également la plus hétérogène.
Nous pensons que certains corpus ne respectent pas cette propriété et qu'il est tout à fait possible qu'à différents moments d'une classification, la plus grande des matrices restantes ne soit pas la plus hétérogène.

4. Une nouvelle solution pour l'enchaînement des partitions

Il apparaît alors pertinent de pouvoir tester, après chacune des phases de partition, l'homogénéité des matrices restantes de façon à sélectionner la plus hétérogène. Comme le calcul de l'analyse factorielle des correspondances nécessaire à chaque partition permet de déterminer le chi2 de la matrice dans son ensemble, nous avons utilisé cette propriété pour revoir le déroulement de l'algorithme. Dans cette nouvelle version, après chaque partition, l'AFC et le chi2 des deux matrices générées sont calculés a priori. Pour chacune de ces matrices, nous déterminons un indice d'homogénéité qui tient compte du chi2 de la matrice, de sa taille et du nombre total de formes. Ce critère relève de la formule suivante :

$$c = \chi^2 * \left(\frac{N}{Nbre\ de\ lignes * Nbre\ de\ colonnes} \right)$$

Il s'agit donc de multiplier le chi2 de la matrice par le ratio de 1 qu'elle contient.
Cette méthode permet de ne plus supposer que la matrice la plus grande est la plus hétérogène mais de tester cette hétérogénéité. Elle a pour désavantage de nécessiter le calcul systématique de l'AFC sur pratiquement toutes les matrices produites. Sans autre modification, cette procédure serait beaucoup plus lente que la version précédente de l'algorithme. Dans l'objectif d'accélérer ces analyses, la ré-écriture théorique de l'algorithme s'est accompagnée d'une recherche de gain de performances qui a ici suivi deux directions :

- Les parties les plus gourmandes en calcul ont été ré-écrites en C++ par l'intermédiaire des packages Rccp (Eddelbuettel et al., 2017) et RcppEigen (Bates, Eddelbuettel, Francois, & Yixuan, 2017) de R. Les parties concernées sont la recherche de la partition qui maximise le Chi2/N après l'AFC et le reclassement des lignes.
- Ces deux parties étant une suite de calculs de chi2 sur la base d'une seule matrice, il a été possible de les paralléliser pour profiter de la nature multi-coeur de la plupart des processeurs modernes. Les calculs sont donc potentiellement distribués aux différents cœurs/threads de la machine par l'intermédiaire des packages Parallel et DoParallel (Calaway, Microsoft Corporation, Weston, & Tenenbaum, 2017) de R.

Ces changements ont en fait nécessité la réécriture complète de l'algorithme de la méthode Reinert dans IRaMuTeQ.

5. Expérimentation

De façon à tester les bénéfices apportés par cette nouvelle procédure, en termes de précision et de rapidité, une expérimentation sur 6 corpus différents a été réalisée. Nous avons associé à des corpus de grandes tailles (les plus susceptibles de présenter des disproportions dans les thématiques qu'ils contiennent) un corpus de taille plus réduite. Les caractéristiques de ces corpus sont présentées dans le Tableau 4.

Tableau 4 : description des corpus utilisés dans l'expérimentation

	dataconf	20newsgroup	lemondefr	ssm	AN2011	LRU
origine	résumés de communications scientifiques	listes de discussion	articles du journal Le Monde en ligne	« same sex marriage » : articles de la presse américaine et anglaise	Compte-rendus des débats à l'assemblée nationale en 2011	articles de la presse quotidienne française sur la loi LRU
langue	anglais	anglais	français	anglais	français	français
textes	18464	18794	10500	6091	261	100
segments	47548	156619	158901	155509	242274	1682
occurrences	1427658	5365593	5682301	5560179	8587277	59870
formes actives	7273	9535	9645	9501	9603	1833
freq . Min.	4	27	23	20	12	2

Le corpus dataconf correspond à des titres et à des résumés de conférences du domaine de l'informatique, il est uniquement en anglais. 20Newsgroup[1] est un corpus également en anglais qui réunit 20 listes de discussions sur des thématiques très diverses (Lang, 1995). lemondefr est un corpus d'articles du

[1] http://qwone.com/~jason/20Newsgroups/

site web du monde en ligne[2], il est en français. Ssm, pour « same sex marriage », est un corpus d'articles de presse américaine et anglaise sur la thématique du mariage entre personnes de même sexe. Il a été constitué par Nathalie Paton. AN2011 correspond à l'année 2011 de la retranscription des débats à l'assemblée nationale française (Ratinaud & Marchand, 2015). Enfin, le corpus noté LRU regroupe 100 articles de la presse quotidienne française sur la thématique de la loi liberté et responsabilité des universités.

L'expérimentation consiste donc à faire subir les deux versions de l'algorithme de classification aux matrices extraites de ces corpus et à comparer la qualité des résultats obtenus. Le nombre de classes terminales a été fixé à 100 pour les "gros" corpus et à 30 pour le "petit". Dans un cas, l'algorithme utilisera le critère de taille pour sélectionner les matrices à partitionner et dans l'autre il utilisera le critère d'homogénéité. Les résultats se présentent sous la forme de graphiques qui montrent l'évolution de la quantité d'information extraite après chacune des partitions. La valeur renvoyée est celle du Chi2/N du tableau réduit des classes. Dans les graphiques de l'illustration 2, les courbes rouges représentent les valeurs obtenues avec l'ancienne version de l'algorithme (notée Reinert) et les courbes bleues les valeurs obtenues avec la nouvelle version (notée Reinert++). Une valeur supérieure correspond à une meilleure qualité de la partition. Le graphique en barres présente le pourcentage d'augmentation ou de diminution de la qualité de la partition du nouvel algorithme en prenant l'ancien comme référence. Les barres vertes signalent une augmentation de la qualité et les barres rouges une diminution. Pour la nouvelle version de l'algorithme, 6 cœurs ont été alloués à la procédure[3].

Ces résultats montrent assez clairement que la nouvelle version de l'algorithme augmente dans la majorité des cas la précision de la classification. Ils permettent également de percevoir que ce gain de qualité est lié à la distribution des thématiques dans les corpus. Tous les corpus ne profitent donc pas de cette évolution de la même façon. Il faut également noter que sur le corpus LRU, il n'y a pratiquement pas de différences entre les deux méthodes. La perte de précision de 1 à 3 % à différents moments de la classification sur ce corpus est tout à fait négligeable et doit être attribuée à des différences d'arrondis entre le code en R et le code en C++. À l'opposé, certains corpus, comme 20newsgroup, présentent des gains de précision qui peuvent atteindre 15 %.

[2] http://www.lemonde.fr

[3] Ces tests ont été réalisés sur un macbook pro 11,3 équipé d'un processeur intel i7-4960HQ

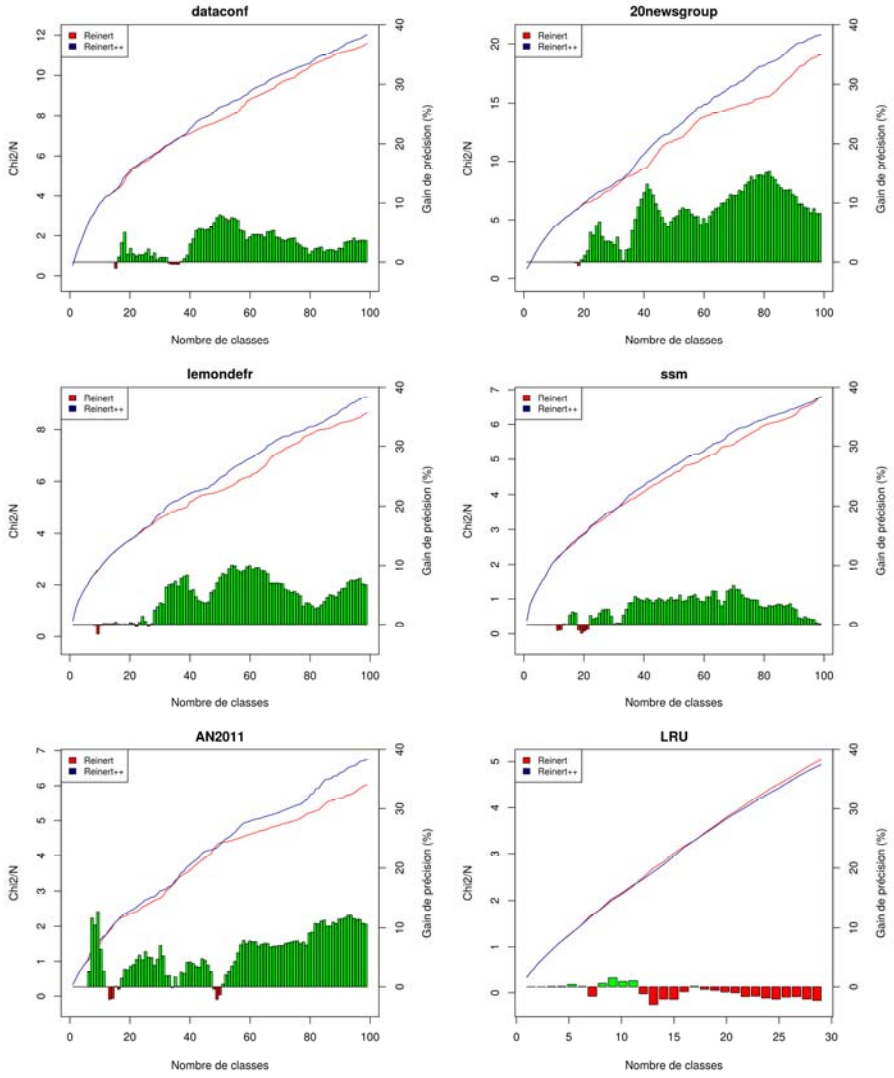


Illustration 2 : Comparaison des résultats entre l'ancienne version (Reinert) et la nouvelle version (Reinert++) de l'algorithme de classification

L'illustration 3 montre que sur les corpus conséquents, le gain de performances introduit par le passage au C++ et à la parallélisation est compris entre un facteur 4 et un facteur 6. Autrement dit, ce nouvel algorithme est jusqu'à 6 fois plus rapide sur la machine sur laquelle ces calculs ont été réalisés.

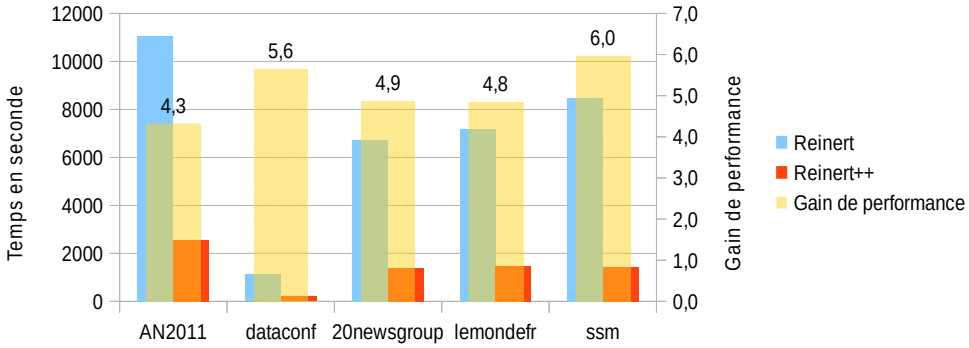


Illustration 3 : comparaison des temps d'analyse entre l'ancienne version (Reinert) et la nouvelle version (Reinert++) de l'algorithme

6. Conclusion

Dans ce travail, nous proposons une nouvelle formalisation de la procédure de classification hiérarchique descendante de la méthode Reinert. Partant de l'hypothèse que dans certains corpus et à certains moments de ces classifications, la classe la plus hétérogène n'est pas forcément la plus grande, nous proposons de substituer le critère du choix de l'enchaînement des matrices d'un critère de taille à un critère d'homogénéité. Les résultats d'une expérimentation sur 6 corpus montrent que les corpus volumineux profitent effectivement de ce changement. Ces résultats sont aussi une invitation à continuer les investigations sur cette méthode. Cette procédure sera implémentée dans la prochaine version du logiciel IRaMuTeQ. L'utilisation du critère d'homogénéité sera optionnelle, de façon à permettre aux utilisateurs de revenir à l'ancienne version.

Bibliographie

Bates, D., Eddelbuettel, D., Francois, R., and Yixuan, Q. (2017). RcppEigen: « Rcpp » Integration for the « Eigen » Templated Linear Algebra Library (Version 0.3.3.3.1). Consulté à l'adresse https://cran.r-project.org/web/packages/RcppEigen/index.html

Calaway, R., Microsoft Corporation, Weston, S., and Tenenbaum, D. (2017). doParallel: Foreach Parallel Adaptor for the « parallel » Package (Version 1.0.11). Consulté à l'adresse https://cran.r-project.org/web/packages/doParallel/index.html

Eddelbuettel, D., Francois, R., Allaire, J. J., Ushey, K., Kou, Q., Russell, N., … Chambers, J. (2017). Rcpp: Seamless R and C++ Integration (Version 0.12.14). Consulté à l'adresse https://cran.r-

project.org/web/packages/Rcpp/index.html

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning* (p. 331-339).

Ratinaud, P. (2014). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Version 0.7 alpha 2) [Windows, GNU/Linux, Mac OS X]. Consulté à l'adresse http://www.iramuteq.org

Ratinaud, P., and Marchand, P. (2012). Application de la méthode ALCESTE à de « gros » corpus et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRaMuTeQ. In *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012)* (p. 835-844). Liège, Belgique. Consulté à l'adresse http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Ratinaud,%20Pierre%20et%20al.%20-%20Application%20de%20la%20methode%20Alceste.pdf

Ratinaud, P., and Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, 2015(108), 57-77.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, VIII(2), 187-198.

Reinert, M. (1990). ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologique*, (26), 24-54.

Il parametro della *frequenza* tra paradossi e antinomie: il caso dell'*italiano scolastico*

Luisa Revelli
Università della Valle d'Aosta– l.revelli@univda.it

Abstract
Emblem of a formal register, the linguistic variety proposed as a model in the Italian school system ever since National Unity is characterized by a lasting artificiality and a strong unwillingness to innovate, even within a frame of progressive slow changes along its historical development. That's why lexical frequencies recorded for "Scholastic Italian" can appear as inherently inconsistent, contrasting with basic vocabulary, even contradictory compared with other apparently similar Italian varieties. Consequently, to study their configuration it's necessary to adopt analysis models capable to interpret quantitative data (volume figures) in the light of the complexity of paradigmatic relations between concurring solutions and of the composite connections between number and type of meanings exhibited in current use. By taking in consideration as a case study *Scholastic Italian* used by teachers during the first 150 years of the national school system, and starting from the data collected by the diachronic corpus of CoDiSV, the contribution aims at verifying opportunities and criticalities of lexicometric analysis applied to such a linguistic variety, that is addressed to an unsophisticated audience, yet characterized by a specialized point of view; of high aspirations, but influenced by educational needs; constantly evolving and yet always recalcitrant to the solicitations of the contemporary language.

Riassunto
Emblema di un canone 'antiparlato', la varietà linguistica proposta a modello nella scuola italiana a partire dall'Unità nazionale, pur presentando in diacronia evidenti tratti evolutivi, si caratterizza per una duratura tendenza all'artificiosità e per una marcata refrattarietà all'innovazione. Le frequenze lessicali documentate nell'*italiano scolastico* possono, per queste ragioni, risultare discordanti in rapporto a quelle del *vocabolario di base*, presentarsi come intrinsecamente poco coerenti, contraddittorie rispetto alle evidenze rintracciabili in varietà d'italiano apparentemente affini: lo studio delle loro configurazioni richiede, pertanto, modelli di analisi capaci di interpretare i dati quantitativi alla luce della complessità delle relazioni paradigmatiche tra le potenziali soluzioni concorrenti nonché dei compositi rapporti tra numero

e tipologia delle accezioni testimoniate nei concreti impieghi contestuali. Assumendo l'*italiano scolastico* proposto dagli insegnanti nei primi centocinquant'anni di scuola nazionale a caso di studio, a partire dai dati ricavati dal corpus diacronico del CoDiSV, il contributo si prefigge allora di verificare opportunità e criticità poste dall'applicazione di parametri lessicometrici a una varietà linguistica al contempo rivolta a un pubblico ingenuo e connotata in prospettiva specialistica, di aspirazione elevata ma condizionata da esigenze didascaliche, in costante evoluzione e ciò malgrado costantemente recalcitrante rispetto alle sollecitazioni della lingua viva e coeva.

Parole-chiave: italiano scolastico; frequenza lessicale; lessicometria; vocabolario di base.

1. Introduzione

Al contempo rivolto a un pubblico ingenuo e connotato in prospettiva specialistica, di aspirazione elevata ma condizionato da esigenze didascaliche, in costante evoluzione e ciò malgrado costantemente recalcitrante rispetto alle sollecitazioni della lingua viva e coeva, l'*italiano scolastico* (d'ora in poi IS) proposto dagli insegnanti nei primi centocinquant'anni di scuola nazionale sembra costituire un buon banco di prova per far emergere le zone di criticità derivanti dall'applicazione di parametri lessicometrici a varietà linguistiche poligenetiche e costituzionalmente disomogenee[1]. Nell'IS, in effetti, un ideale di ricchezza espressiva perseguito attraverso una marcata ostilità nei confronti di ogni forma di ridondanza, ripetizione o generalità delle espressioni spinge verso un'ostentata e ricercata *variatio*, ma la contemporanea esigenza di alfabetizzare i giovani allievi orientandoli a privilegiare specifici membri di serie sinonimiche ritenuti maggiormente corretti, appropriati o esornativi tende, di fatto e in opposta direzione, a ridurre la gamma delle possibilità espressive disponibili. La necessità di veicolare attraverso la lingua i saperi disciplinari rende, d'altra parte, necessario l'uso di metalinguaggi, tecnicismi e accezioni semantiche che sembrano destabilizzare ulteriormente il serbatoio lessicale di riferimento dell'IS allontanandolo significativamente dal vocabolario di base della lingua italiana. In che misura e in che termini questo avvenga realmente è quanto ci si propone di verificare qui di seguito, integrando i dati lessicometrici e quantitativi disponibili con alcune

[1] Per un inquadramento delle caratteristiche, stabili ed evolutive, dell'IS si rimanda a De Blasi 1993, Cortelazzo 1995, Benedetti G. e Serianni L. (2009), Revelli 2013.

riflessioni di natura qualitativa. Relativamente all'IS, la base lessicale presa a riferimento è costituita da un lessico di frequenza elaborato da chi scrive (Revelli 2013) a partire da un corpus iniziale di 830 quaderni di scuola elementare redatti in area valdostana nel periodo compreso tra la fine del XIX e i primi anni del XXI secolo. I 2.022 termini che compongono il vocabolario di base sono stati individuati dopo che una selezione bilanciata dei documenti, ripartiti in subcorpora cronologici ventennali, è stata sottoposta a trattamento computazionale con lo scopo di identificare la dimensione della variazione diacronica nei canoni linguistici proposti a modello da parte degli insegnanti[2]. A fianco delle concordanze, è stato così ricavato in prima battuta un vocabolario composto da 152.151 occorrenze (*tokens*), ricondotte a 18.898 forme (*types*) e 11.751 lemmi[3]. Un'ulteriore selezione ha poi dato luogo all'identificazione dei 2.022 sostantivi, aggettivi e verbi considerati pancronici perché stabilmente assestati nel *vocabolario di base dell'italiano scolastico* (d'ora in poi VoBIS), in quanto testimoniati con più di cinque occorrenze in almeno quattro dei sei repertori cronologici o in tre non consecutivi. Il termine di paragone è costituito dall'edizione 2016 del *Nuovo Vocabolario di base della lingua italiana* (d'ora in poi NVdB) di Isabella Chiari e Tullio de Mauro[4], che ripartisce le circa 7.000 parole statisticamente più frequenti e accessibili ai parlanti italiani del XXI secolo nei tre serbatoi del *lessico fondamentale* (FO, circa 2000 parole ad altissima frequenza usate nell'86% dei discorsi e dei testi), del *lessico ad alto uso* (AU, circa 3.000 parole di uso frequente che coprono il 6% delle occorrenze) e del *lessico di alta disponibilità* (AD, circa 2000 parole "usate solo in alcuni contesti ma comprensibili da tutti i parlanti e percepite come aventi una disponibilità pari o perfino superiore alle parole di maggior uso"). La scelta di fare riferimento a tale base, che comprende al proprio interno anche le frequenze relative alle varietà parlate e si colloca temporalmente in un periodo successivo a quello considerato per il lessico scolastico, risponde all'esigenza di verificare se e in che misura il modello scritto offerto da quest'ultimo possa aver inciso sulla configurazione dei successivi usi concreti.

[2] Le tipologie testuali prese in considerazione sono costituite dalle consegne degli esercizi, dai titoli dei componimenti, da dettati, interventi correttivi, valutazioni e giudizi documentati nei quaderni degli alunni.

[3] Il vocabolario e le concordanze del corpus sono stati ricavati, previa annotazione e lemmatizzazione, tramite il software T-LAB, ideato e distribuito da Franco Lancia. Per un approfondimento a proposito dei principi adottati e la metodologia seguita si rimanda a Revelli 2013.

[4] https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

2. Vocabolari di base *a confronto*: le *frequenze* nel NVdB e nel VoBIS

La comparazione del serbatoio lessicale dei due repertori presi a confronto consente di compiere, in prima battuta, alcune osservazioni generali: dei 2022 lemmi del VoBIS, 1784 trovano riscontro nel NVdB, spartendosi per il 53% nel serbatoio del *lessico* FO, per il 26% in quello di AU e per il 9% in quello di AD. Senza entrare qui nel merito delle convergenze che accomunano i due vocabolari, sembra comunque opportuno segnalare che dietro molti esempi di apparente coincidenza delle distribuzioni di frequenza si celano in realtà difformità significative, prevalentemente indotte dalla tendenza dell'IS al restringimento o in alcuni casi anche alla rideterminazione semantica: fra le molte parole che assumono specifici sensi scolastici (ad es. *diario, interrogazione*, *nota*, *pensierino*, *voto*), alcune perdono del tutto l'ancoramento ai significati di cui sono dotati nella lingua comune, come accaduto a *tema*, passato a identificare non più un soggetto o argomento da trattare, ma invece il prodotto di una specifica tipologia testuale.

Per ciò che concerne le 238 parole assenti nel NVdB (12%), esse possono essere raggruppate in categorie utili a mettere fuoco diverse criticità relative all'applicazione del parametro della frequenza comparativamente applicato.

Un primo, corposo gruppo che risulta esclusivo dell'IS è costituito da logonimi caratteristici della nomenclatura metalinguistica dell'apparato scolastico, del tipo *alfabetico*, *apostrofo*, *coniugazione*, *preposizione*, ecc. Osserviamo che, malgrado il loro potenziale polisemico, molti di questi – come *coniugare*, *derivato*, *imperfetto*, *possessivo*, *primitivo* - raggiungono nell'ambito dell'IS frequenze molto elevate nel loro esclusivo ruolo di etichette destinate alla riflessione metalinguistica[5]: la rappresentatività quantitativa non implica quindi un contatto degli allievi con le diverse accezioni di cui quegli stessi termini possono essere portatori, ma corrisponde invece a un'insistita specializzazione motivata da esigenze didascaliche. Un secondo gruppo è costituito da termini tipici dei contesti d'insegnamento della letto-scrittura: si tratta principalmente di sostantivi che fanno riferimento a referenti concreti ma di scarsa prominenza nella quotidianità, la cui forma scritta guida e richiede la conoscenza di convenzioni controintuitive eppure fondamentali per la corretta codifica e decodifica ortografica. Citiamo a titolo di esempio parole come *acquaio*, *acquavite* e *acqueo*, evidentemente introdotte non per stringente necessità tematica quanto invece con scopo di consolidamento delle corrette rappresentazioni grafematiche.

A scopi didattici legati agli insegnamenti disciplinari o più genericamente a

[5] Ad esempio, *dimostrativo* - sempre preceduto da *aggettivo* o *pronome* - non entra mai in combinazione con *atto*, *gesto*, ecc.

scelte tematiche caratteristiche del contesto educativo sono da imputare le alte frequenze di diversi termini relativi all'ambito storico-geografico (*legione*, *vetta*), di voci descrittive dell'universo naturale (*arto*, *astro*) e della vita rurale (*semina*, *vendemmia*); di serie di verbi (*castigare*, *disobbedire*) di aggettivi (*diligente*, *ordinato*) e di sostantivi astratti (*umiltà*, *penitenza*) appartenenti al formulario tipico dell'educazione civica o morale e a quello della valutazione scolastica. A differenza del NVdB, per la sua impostazione diacronica il lemmario del VoBIS trova, d'altra parte, rappresentati numerosi arcaismi: si tratta in alcuni casi di varianti formali oggi dismesse (ad es. *annunziare* per *annunciare*,) o dispreferite (*ubbidire* per *obbedire*); di termini relativi a referenti che i cambiamenti sociali dell'ultimo cinquantennio hanno reso superflui o anacronistici (*manto*, *ricamatrice*); di membri di coppie o serie sinonimiche superati o formali, che soltanto in ambito scolastico sono o sono stati più a lungo privilegiati rispetto a concorrenti avvertiti dai parlanti come più attuali (*persuadere* per *convincere*)[6].

Proseguendo con le mancate corrispondenze nei due repertori, se l'assenza nel NVdB di voci scolastiche un po' leziose come *diletto*, *garbato*, *vezzo* e *soave* risulta scontata, stupisce invece la mancata inclusione di termini che appaiono stabili nel tempo e di diffusione panitaliana: è il caso di zoonimi come *bue*, *elefante*, *formica*; di nomi di frutti usualmente presenti sulle tavole degli italiani come *fragola*, *noce* e *uva*; di nomi concreti d'uso comune come *carezza*, *martello*, *ombrello*. La mancanza di riscontri nel NVdB per termini di questo tipo può essere solo in parte interpretato in una dimensione propriamente sociolinguistica: pur essendo vero che - dato il pubblico cui si orienta - l'IS fa più frequente riferimento a temi e referenti della cultura materiale ed esperienziale di quanto non accada nelle varietà linguistiche rivolte a e prodotte da parlanti adulti, è altrettanto vero che in linea teorica tutti i vocaboli, a maggior ragione se accolti e veicolati dalla scuola, dovrebbero rientrare in quel patrimonio di «parole che può accaderci di non dire né tanto meno di scrivere mai o quasi mai, ma legate a oggetti, fatti, esperienze ben noti a tutte le persone adulte nella vita quotidiana» (De Mauro 1980: 148). Ci aspetteremmo quindi di trovare riscontri almeno all'interno di quel serbatoio di parole di AD di cui tuttavia lo stesso De Mauro ha in più occasioni dichiarato la natura sfuggente, non statistica ma

[6] Ad es. *bambagia*, *cagionare*, *figliolo*, *focolare*, *garzone*, *uscio.* Proprio nell'ambito di quest'ultima categoria il serbatoio dell'IS si differenzia d'altra parte in modo evidente da quello del lessico corrente, privilegiando sistematicamente soluzioni assenti nel NVdB, a scapito di quelle invece lì documentate e in molti casi dotate di marca d'uso FO (ad es. *appetito* per *fame, ardere* per *bruciare*, *sciupare* per *rovinare*, ecc.).

congetturale[7]. E, in effetti, probabilmente neppure le analisi quantitative più imponenti e minuziose possono aspirare ad azzerare inevitabili fattori di imprevedibilità e accidentalità della frequenza. Nel caso qui preso a campione, che relativamente all'IS non dispone di un corpus di partenza di dimensioni del tutto soddisfacenti, lacune relative a termini rispetto ai quali ci si aspetterebbe di avere riscontri si verificano anche capovolgendo la prospettiva e quindi partendo dal lemmario del NVdB: pure ampliando l'orizzonte all'intero vocabolario del corpus, a risultare mancanti non sono soltanto termini marcati come AD, ma anche parole fondamentali che sono, sì, probabilmente note ai bambini, ma non compaiono nel campione preso in esame per ragioni meramente accidentali.[8] Certamente motivate e intenzionali sono invece specifiche tipologie di omissioni facilmente identificabili come specifiche dell'IS: si tratta di neologismi e prestiti di lusso, che i modelli dei maestri – forse in alcuni casi anche per ragioni ortografiche - tendono a respingere quand'anche ormai stabilmente acclimatati nell'italiano standard (*jeans*, *quiz*, *smog*); di termini riferiti a concetti ritenuti sconvenienti per un pubblico acerbo (*aborto*, *droga*, *sesso*); di voci gergali, espressioni volgari, insulti, improperi (*coglione*, *culo*, *ruttare*); di appellativi discriminatori (*ebreo*, *nano*, *negro*) ma anche di parole prudenzialmente evitate perché avvertite come potenzialmente faziose, propagandistiche o almeno ideologicamente e politicamente orientate: su quest'ultimo aspetto, che incarna l'intimità dei rapporti tra lessico, scuola, clima sociale e temperie culturale non è tuttavia possibile compiere generalizzazioni, perché gli indizi relativi alle diverse caratterizzazioni assunte dal fenomeno nel corso dei tempi, anche molto recenti, richiedono di essere intercettati sulle frequenze basse o inesistenti, piuttosto che su quelle elevate del lessico di base.

3. Conclusioni e prospettive

Come ci si è proposti di evidenziare, l'esame quali-quantitativo dell'IS conferma che, pur presentando in diacronia tratti di ammodernamento, il modello linguistico proposto dagli insegnanti risulta caratterizzato dallo

[7] Nella *Prefazione* al NVdB è specificato che le parole di AD "sono state ricavate partendo dalla lista di 2.300 parole di alta disponibilità del vecchio VdB e sottoponendola a gruppi di studenti e studentesse universitari per eliminare le parole non più avvertite come di maggior uso e per accogliere invece nuove parole avvertite come di alta disponibilità".

[8] Esemplificativo dei margini di casualità può essere il caso degli etnici, che mancano in alcuni casi al CoDiSV (ad es. *cinese*, *iugoslavo*) che pure ne documenta moltissimi altri almeno apparentemente di analoga diffusione (ad es. *giapponese*, *inglese*).

stabile impiego di termini estranei al vocabolario di base e dal parallelo evitamento di termini correnti, ritenuti inadeguati o sconvenienti o più semplicemente logorati da un uso reputato eccessivo. Lo studio dei dati consente, poi, di rilevare un'abbondante presenza di logonimi ed etichette tipici o esclusivi del metalinguaggio didattico e grammaticale, l'uso di hapax spesso confinati nell'ambito di occasionali specifiche tipologie esercitative ma per il loro ruolo strategico didatticamente irrinunciabili nonché per il ricorso a un formulario al cui interno termini correnti assumono tramite fenomeni di rideterminazione semantica accezioni differenti da quelle consuete, specializzandosi in relazione a compiti e routines comunicative tipici del contesto educativo. Le frequenze lessicali documentate nelle varietà dell'IS si presentano in parte, per queste ragioni, come intrinsecamente poco coerenti, discordanti in rapporto a quelle del vocabolario di base, contraddittorie rispetto alle evidenze rintracciabili in varietà d'italiano apparentemente affini: lo studio delle loro configurazioni richiede, pertanto, modelli di analisi capaci di interpretare i dati quantitativi alla luce della complessità delle relazioni paradigmatiche tra le potenziali soluzioni concorrenti nonché dei compositi rapporti tra numero e tipologia delle accezioni testimoniate nei concreti impieghi contestuali. In questa direzione, in parte già esplorata in particolare negli studi di taglio psicolinguistico e glottodidattico dedicati ai processi della comprensione e alla leggibilità dei testi, sembra che un raffronto comparativo tra il lessico dell'IS e quello del VdB condotto in modo sistematico su corpora cronologicamente armonizzati possa fornire ulteriori linee di ricerca in almeno due specifici ambiti d'indagine.

Un primo, di prospettiva più propriamente acquisizionale, andrebbe finalizzato a verificare gli effettivi esiti della protratta esposizione in età scolare alla percentuale di parole dell'IS che risulta estranea al vocabolario di base: in questa direzione, tenuto conto della natura incrementale e adattiva degli apprendimenti lessicali ma anche dell'effetto di evanescenza che la mancata pratica può esercitare sulle competenze possedute, si potrebbe tentare di rispondere a domande del tipo: quanto incide effettivamente l'insistenza con cui un termine è presente nell'input offerto nell'ambito dell'IS sul suo effettivo impiego nei domini da questo distinti e successivamente sperimentati? In che misura la concettualizzazione relativa una determinata accezione di un termine veicolata dall'insegnamento può condizionare (positivamente o negativamente) la successiva acquisizione di significati ulteriori e diversi per quello stesso termine? In che termini le soluzioni preferenziali e le scelte paradigmatiche proposte dall'IS risultano vincenti, almeno a livello di competenza ricettiva, nella concorrenza con le analisi statistiche che i parlanti sperimentano su altre varietà e in contesti potenzialmente più pregnanti? E in questo senso, quanto può essere

percepito come autorevole, significativo, dotato di rilevanza comunicativa il modello lessicale scolastico in un Paese in cui l'italiano è diventato lingua materna per la gran parte dei cittadini e la concorrenza di input – non soltanto lessicale - proveniente da fonti alternative alla scuola appare quantitativamente strabordante?

Un secondo ambito d'indagine, al precedente correlato ma di prospettiva principalmente lessicografica, potrebbe invece essere indirizzato ad esplorare l'ipotesi che una parte del *vocabolario scolastico di base* possa essere considerata denominatore comune delle competenze lessicali possedute dai parlanti adulti alfabetizzati, e venire impiegata soprattutto come punto di riferimento per la definizione del *vocabolario di alta disponibilità*. In questo senso, le oggettive difficoltà di identificazione di quelle "parole che riteniamo più comuni, psicologicamente e culturalmente, ma che poi hanno in realtà una frequenza minima, vicina a zero, soprattutto nell'uso scritto" (De Mauro 2004: 142) potrebbero essere in parte superate facendo riferimento a quella porzione di bagaglio lessicale condiviso e acquisito, se non attraverso altri canali, per il tramite dell'IS: seppure statisticamente poco rilevanti nelle produzioni adulte, i termini a chiunque familiari perché proposti con frequenze elevate e funzioni significative nell'italiano per i bambini – ad esempio i termini tipicamente indicati sugli alfabetieri (*oca*), usualmente utilizzati per l'insegnamento delle particolarità ortografiche (*camoscio*), presenti nelle denominazioni più diffuse di giochi e tipologie esercitative (*cruciverba*), in fiabe e racconti (*carrozza*), corrispondenti a discipline (*geografia*) o *routines* scolastiche (*giustificazione*) – potrebbero probabilmente superare qualunque prova di elicitazione sui parlanti e quindi, seppur difficilmente rintracciabili nel lessico adulto, essere selezionate per entrare nel *vocabolario di base* con attribuzione della marca AD.

Anche in questo caso, certamente, per evitare insidie e ambiguità semantiche andrebbero individuati dispositivi utili ad accertare la fenomenologia delle accezioni effettivamente attive nonché a verificare e interpretare criticamente le relazioni intercorrenti tra la frequenza dell'input lessicale (e semantico) in ingresso e la frequenza dell'output lessicale (e semantico) fattuale ma anche potenziale, in un modello descrittivo che – nel contemplare un'interazione dialettica, dinamica e comparativa tra le dimensioni della *ricettività*, *produttività* e *disponibilità* e attribuendo i giusti pesi a quella delicata e complessa combinazione di *quantità* e *qualità* che De Mauro (1994: 97) felicemente ebbe modo di battezzare *binomio indispensabile* – consenta di distinguere gli autentici dai solo apparenti paradossi della frequenza.

Riferimenti bibliografici

Benedetti G. e Serianni L. (2009). *Scritti sui banchi. L'italiano a scuola fra alunni e insegnanti.* Roma, Carocci.

Chiari I. e De Mauro T. (2012). The new basic vocabulary of Italian: problems and methods. *Rivista di statistica applicata / Italian Journal of Applied Statistics*, vol. 22 (1): 21-35.

Cortelazzo M. (1995). *Un'ipotesi per la storia dell'italiano scolastico*. In Antonelli, Q. & Becchi E. curatori, *Scritture bambine*, Roma-Bari, Laterza: 237-252.

De Blasi N. (1993). *L'italiano nella scuola.* In Serianni, L. e Trifone, P. curatori, *Storia della lingua italiana*, vol. I "I luoghi della codificazione". Torino, Einaudi: 383–423.

De Mauro T. (1980). *Guida all'uso delle parole.* Editori Riuniti, Roma 1980.

De Mauro T. (2004). *La cultura degli italiani*. A cura di Francesco Erbani. Roma-Bari, Laterza.

De Mauro T. (2005). *La fabbrica delle parole*. Torino, Utet Libreria.

Revelli L. (2013). *Diacronia dell'italiano scolastico.* Roma, Aracne.

How Twitter emotional sentiments mirror on the Bitcoin transaction network

Piergiorgio Ricci
Tor Vergata University – piergiorgio.ricci@gmail.com

Abstract
Bitcoin represents the first and most popular decentralized cryptocurrency. It was launched in 2008 by Satoshi Nakamoto, the name used by the unknown person or people who designed Bitcoin system and created its original reference implementation. It is based on Blockchain technology that is considered one of the most promising technologies for the future. It is more than an instrument of finance and will likely disrupt many industries from banking to governance in the next years. This research explores a geolocalized subset of Bitcoin blockchain and compares it with Twitter communication related to the topic in order to discover what people living in different geographical areas think about Bitcoin cryptocurrency and to assess potential relationship between characteristics of language adopted by Twitter users in posts containing the key word Bitcoin and the structure of geolocalized blockchain. It also answers a variety of interesting questions about the national use of Bitcoin.

Keywords: Bitcoin, Blockchain, Cryptocurrency, Social Network Analysis, Semantic Analysis.

1. Introduction

Bitcoin cryptocurrency is based on blockchain technology that consists in an open and distributed ledger where all transactions occuring in the system are recorded in a verifiable and permanent way. (Narayanan A., Bonneau J., Felten E., Miller A., and Goldfeder S., 2016) They are organized in blocks which are generated periodically and linked by using cryptography techniques (*SHA256*)(Drainville D., 2012). Each of them needs to be validated by a peer to peer network respecting a specific protocol for validating new blocks. Once stored, data can not be tampered without tampering all subsequent blocks, activity that requires collusion of the network majority. (Nakamoto, 2008) This approach complies with consensus theory, a social theory which holds that social changes and innovation can be reached without conflicts and the social system is fair. In fact, Bitcoin's protocol relies on a strong social consensus among all partecipants of the

system that represent a node of the network and run a software with the aim to improve enforcement of rules they agree on. Bitcoin network is decentralized and it does not require trusting in a third party, such as a bank or a government institution. For sure, it represents a new concept of money (Evans, 2014) and the main purpose of this work is to find out what people living in different geographical areas think about Bitcoin cryptocurrency and to assess potential relationship between characteristics of language used on Twitter posts related to the topic and the structure of geolocalized Bitcoin Blockchain. Research has been conducted to analyze correlations and causalities between social network metrics performed on the geolocalized Bitcoin Transaction Network and Bitcoin emotional signals interecepted by analyzing Twitter users posts grouped by country. In particular, it has been considered important to discover wheter there is a specific kind of communication adopted by Twitter users belonging to a specific country that holds certain transaction network centrality measures. In other words, the core question to be answered has regarded the analysis on existence of correlation between the centrality in the Bitcoin transactions network of a country and characteristics of language used on Twitter Bitcoin posts by their citizens. To achieve this purpose, two datasets reperesenting Bitcoin transactions and Twitter communication related to Bitcoin, have been collected and classified on the basis of geography. Prior research has been focused on economic aspects (Ron D. and Shamir A., 2012) and structural proprieties of Bitcoin transaction network (Lischke et Fabian, 2016) (Fleder M., Kester M. and Pillai S. 2015), but it has rarely considered the existing relationship between transactions and social media communication. This study also answers a variety of interesting questions about the national use of Bitcoin and how Twitter users perceive it through communication signals posted on Twitter microblogging platform. One of the most widely accepted use cases for Bitcoin has to do with payments for digital content (Grinberg R., 2012) and, at present, Bitcoin system is used only by early adopters and innovators among population.

2. Data set

2.1 Bitcoin dataset

In order to analyze and compare the network of Bitcoin transactions and the relative user sentiment on Twitter, two differents dataset have been built by using a serious of Application Program Interface (*API*) available on the web. The first dataset to be extracted has been the Bitcoin transaction network that is publicly available from many free web services (such as Blockchain.info) or by using a Bitcoin client that requires and stores the whole transaction history, known as blockchain (Moser M., 2013). In order to reduce and

manage its complexity, a subset of blockchain, composed by more than 2 million transactions from July
2013 to July 2017 has been collected. They have been imported through the Blockchain Data API service that allows Bitcoin block and transaction payments data query functionality, providing requests for data regarding single block, single transaction and block heights.

Fig. 1 - Example of transaction with multiple inputs and outputs (www.blockchain.info)

Fig.2 - Word Cloud related to USA Twitter dataset

These transactions have been geolocalized by using IPInfo.it web service and stored in a NoSQL database. Geolocalization activity has regarded the discovery of the countries involved in each transaction and it has been carried out by scraping transactions IP addresses (Kaminsky, 2011). Each transaction block contains a set of transactions and is characterized by following attributes: flow identifier, hash transaction, timestamp, origin country, destination country, sender, recevier and total amount (Ober M., Katzenbeisser S. et Hamacher K. 2013). Since each transaction can allow multiple input and output addresses (Reid F. et Harrigan, 2012), they have been decomposed in transaction flows. In order to attach geographical informations to each transaction, the service provides by ipinfo.io website has been used. It offers a web interface where is possible to retrive the origin country of an IP Address provided as input.

2.2 Twitter dataset

A set of tweets from 10 different countries containing the word "Bitcoin" have been collected in order to be analyzed. Sentiment analysis have been conducted using the Software *Condor* (*MIT Center for Collective Intelligence*) that automatically recognizes sentiment in English, Spanish, German, French, Italian and Portuguese and allows tweets fetching restricted to a given geolocation. It also allows to calculate sentiment of posts by using semantic analysis techinques. This dataset is partially misaligned with the first one for technical reasons.

3. Research methodology

Research has been conducted combining social network analysis (*SNA*) and semantic analysis methodology with a particular focus on the relationship among main indicators related to these two fields calculated on the dataset.

3.1. Social Network Analysis

Using a Social Network Analysis approach, several strategies are possible to examine the structure of the Bitcoin transaction network. In order to counduct the analysis some of the most common measures of centrality have been identified. Most of them have been proposed by Freeman (1979) and also analyzed in other Social Network Analysis articles (Batagelj, 2011). In the following subsections they are briefly described.

3.1.1 Degree centrality

This measure is based on the degree that indicates the number of nodes attached directly to a specific node for which it is computed. In the case of directed networks, two different measures of degree centrality can be calculated, defined as indegree and outdegree. The first one is given by the number of ties directed to the node, while outdegree is the number of ties that the node directs to others. In such cases, the degree is the sum of indegree and outdegree. The (weighted) all-degree for the generic node n_i in a directed graph is represented by the following equation:

$$C_D(n_i) = \sum_{t=1}^{N}(a_{t,i} + a_{i,t})$$

where $a_{t,i}$ counts the number of incoming ties and $a_{i,t}$ represent the number of outgoing ties. A node with an high degree centrality is central in the network structure and tend to influence the others.

3.1.2 Closeness Centrality

Closeness centrality indicates the inverse of the distance of a node from all the others in the graph. It is based on the shortest paths that between each couple of nodes in the network. Closeness centrality of node n_i, in a graph with N nodes, is defined as following:

$$C_C(n_i) = \sum_{i=1}^{N} 1/d(n_i, n_j)$$

where, $d(n_i, n_j)$ is the number of edges in the shorterst path linking n_i and n_j. Closeness centrality is normalized as shown below:

$$CC(n_i) = (N - 1)C_c(n_i)$$

This measure can be considered as a proxy of the speed by which a social actor can reach the others.

3.1.3 Betweenness Centrality

This variable considers the shortest paths that connect every other couple of nodes and is higher when a node is more frequently in this subset. For a network with N nodes, the betweenness centrality for node:

$$C_b(n_i) = \sum_{i<k} g_{ik}(n_i)/ g_{ik}$$

where, g_{jk} is the number of the shortest paths linking two nodes in the network and $g_{jk}(n_i)$ is the number of shortest path linking two nodes that go through the node n_i. Social network indicators described above can be used to analyze the structure and the dynamics of the Geographical Bitcoin Network. In particular, once collected the target set of transactions and enriched them with geographical informations, two directed graphs has been modeled. In the first one, identified as Generic Network, each node represents a Bitcoin address owned by a user belonging to a specific country, while each link indicates a transaction of a certain amount (weight of link) occuring between two different addresses, while in the second one, defined as Geographical Network, each node symbolize a country and links act for transactions that can involve single or different countries. All the network metrics used in this study will be explained in the next chapter. They have been performed on the Geographical Network, obtained by merging General Network transactions on geographical basis.

3.2 Semantic Analysis

Semantic analysis of textual data allows to turn text into data for analysis. This is possible applying natural language processing techinques and analytical methods. (Hu X., Tang L., Tang J. and Liu H., 2013). In the following subsections a set of communication indicators will be briefly described.

3.2.1 Sentiment

This indicator describes wheter messages are positive or not. Its value is between 0 in the case of very negative messages and 1 viceversa. It is computed as the average score for the whole text in a message.

3.2.2 Emotionality

This variable expresses the degree of emotion of an individual text fragment and it is involved in sentiment elaboration.

3.2.3 Complexity

It measures the rarity of a word, or the likelihood that a single word will occur in a text. It is higher when a text contains many rare words.

4. Results

The aim of this study has been to find out whether characteristics of Twitter communication related to Bitcoin reflects the GeoBlockchain network structure. Analysis has been conducted combining most important social network centrality metrics, such as Degree Centrality, Closeness Centrality and Betweenness Centrality with some other language indicators measuring the characteristics of the textual data used in Twitter communication. On the one hand, centrality metrics measures the importance, influence or power of a node in the network and are widely applied in social network analysis, on the other, language indicators allow to identify whether communication referred to Bitcoin is positive or not, its emotionality and the complexity of word usage. During analysis, country rankings for each social network indicator has been calculated in order to be correlated with Twitter Sentiment, Complexity and Emotionality national rankings performed on Tweets containing the key word "Bitcoin". Spearman's correlation, computed considering a set of 10 different countries with a high number of transactions and tweets, shows a significative correlation between centrality measures computated in the Geographical blockchain netowork and language on microblogging platform Twitter. In particular, communication of people belonging to most central countries in the Bitcoin network, e.g. Germany and USA, is more complex and less emotional than the one of peripheral country nodes. This is probably due to a more depth knowledge of Bitcoin phenomena in the most innovative countries as shown by their Word clouds. In fact, they tweet more and with a quite technical language (e.g. they speak

about technical aspects such as fork of blockchain), while the others one, for example Spain, appear frightened of the new cryptocurrency's diffusion.

			Emotionality	Degree Centrality
Spearman's Rho	Emotionality	Correlation Coefficient	1,000	-,638*
		Sig. (2-tailed)	.	,047
		N	10	10
	Degree Centrality	Correlation Coefficient	-,638*	1,000
		Sig. (2-tailed)	,047	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

			Complexity	Degree Centrality
Spearman's Rho	Complexity	Correlation Coefficient	1,000	-,693*
		Sig. (2-tailed)	.	,026
		N	10	10
	Degree Centrality	Correlation Coefficient	-,693*	1,000
		Sig. (2-tailed)	,026	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Fig.3 - Spearman's correlations calculated on national rankings of Complexity - Degree Centrality and Emotionality - Degree Centrality

5. Conclusion and future works

The analysis highlights the Bitcoin transactions geographical distribution and shows national differences in its adoption, revealing the major businesses and markets. In particular, the most central countries in Bitcoin transaction network are characterized by a positive and quite complex language, while peripheral countries use a more emotional language and the sentiment of their people about it is fairly variable. This result leads to the interpretation that Twitter emotional sentiments mirror the Bitcoin transaction network and this could be seen as an interesting signal for investors and entrepreneurs interested in the development of new payment systems based on Bitcoin technology and in the choice of the start up country. Main findings of the study could be applied to crypto-payments national regulation as well as to the economic and financial impact assessment of cryptocurrencies and future

works include investigation on the principle barriers to mass adoption of Bitcoin cryptocurrency.

References

De Nooy W.,Mrvar A. and Batagelj V. (2011). *Exploratory social network analysis with pajek* (2nd Ed.). Cambridge University Press.

Freeman L.C. (1979). *Centrality in social networks conceptual clarification*. *Social Networks, 1* , 215–239.

Lischke M. and Fabian B. (2016). *Analyzing the Bitcoin Network: The First Four Years*. MDPI AG.

Nakamoto S. (2008). Bitcoin: *A Peer-to-Peer Electronic Cash System*.

Reid F. and Harrigan M. (2012). *An Analysis of Anonymity in the Bitcoin System*. Springer.

Ober M., Katzenbeisser S. and Hamacher, K. (2013) *Structure and Anonymity of the Bitcoin Transaction Graph*. *Future Internet. MDPI.*

Kaminsky D. (2011). *Black Ops of TCP/IP. Black Hat & Chaos Communication Camp*

Drainville D. (2012). *An Analysis of the Bitcoin Electronic Cash System.* University of Waterloo

Ron D. and Shamir A. (2012). *Quantitative Analysis of the Full Bitcoin Transaction Graph.* IACR Cryptology ePrint Archive

Fleder M., Kester M. and Pillai S. (2015) *Bitcoin Transaction Graph Analysis*

Moser M. (2013) *Anonymity of Bitcoin Transactions.* Munster Bitcoin Conference

Grinberg R. (2012). Bitcoin: An Innovative Alternative Digital Currency. Hastings Sci. & Tech

Hu X., Tang L., Tang J. and Liu H. (2013). *Exploiting social relations for sentiment analysis in microblogging*. In Proceedings of the sixth ACM international conference on Web search and data mining. ACM.

Narayanan A., Bonneau J., Felten E., Miller A., and Goldfeder S. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press

Evans D. (2014) *Economic Aspects of Bitcoin and Other Decentralized Public-Ledger Currency Platforms*. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 685

Analyse de contenu versus méthode Reinert : l'analyse comparée d'un corpus bilingue de discours acadiens et loyalistes du N.-B., Canada

Chantal Richard[1], Sylvia Kasparian[2]
[1]Université du Nouveau-Brunswick, Canada – chantal.richard@unb.ca
[2]Université de Moncton, Nouveau-Brunswick, Canada – sylvia.kasparian@umoncton.ca

Abstract
In this paper we compare two methods of thematic analysis by applying them to the same corpus. Specifically, we will compare the results of the classification of units of contexts using the Reinert method in IRAMUTEQ, with a content analysis (manually coded themes) analyzed using SPHINX in 2012. The bilingual corpus consists of two sub-corpora: speeches at the *Conventions nationales acadiennes* (in French) and centennial commemorative speeches by Loyalists (in English). Our goal is to determine whether the Reinert method of distribution by class confirms, contradicts, or enhances a traditional content or thematic analysis.

Résumé
Cet article compare deux méthodes d'analyse thématique de données textuelles appliquées à un corpus bilingue. Notamment, nous comparons la répartition par classes selon la méthode Reinert, intégrée dans IRAMUTEQ, avec les résultats d'une analyse de contenu (codification manuelle des thèmes) analysés par SPHINX en 2012. Le corpus est constitué de discours acadiens (en français) et de discours loyalistes (en anglais). Cette étude permet de voir dans quelle mesure la méthode Reinert confirme, contredit, ou bonifie l'analyse de contenu traditionnelle pour étudier les mondes lexicaux ou univers de discours de ces deux sous-corpus.

Mots-clés : analyse de contenu, IRAMUTEQ, méthode Reinert, classification hiérarchique descendante.

1. Introduction

Aux JADT 2012, nous avions présenté une analyse de contenu des thèmes principaux d'un corpus bilingue tiré de la base de données Vocabulaires identitaires. Cette base regroupe des discours en français et en anglais qui traitent de l'identité collective de deux peuples diasporiques au Nouveau-Brunswick, Canada : les Acadiens et les Loyalistes. Depuis 2012, la base de

données est passée de 74 à 1525 textes. S'imposait alors une démarche plus efficace – pour cela nous avons choisi la méthode Reinert de classification hiérarchique descendante. Avant d'entamer l'analyse du corpus plus large nous avons voulu comparer la méthode Reinert aux résultats de l'analyse de contenu de 2012 en l'appliquant au corpus original de 74 textes. Cet article permet de voir dans quelle mesure la méthode Reinert bonifie l'analyse de contenu traditionnelle pour étudier les mondes lexicaux de ce corpus.

2. Analyse de contenu et méthode Reinert

Avant de procéder à l'analyse, nous définirons brièvement les deux types d'analyse tout en expliquant notre démarche méthodologique.

2.1 Analyse de contenu

Nous entendons par analyse de contenu une « méthode de classification ou de codification dans diverses catégories des éléments du document analysé pour en faire ressortir les différentes caractéristiques en vue d'en mieux comprendre le sens exact et précis » (L'Écuyer 50). En d'autres mots, une lecture exhaustive du corpus permet de choisir des unités de classification, de générer une catégorisation sous forme de tableaux à être traités statistiquement, et l'interprétation des résultats de l'analyse statistique permet une description des thèmes relevés. C'est la méthodologie utilisée dans notre première étude du corpus à l'aide des logiciels SPHINX et HYPERBASE afin d'extraire les mots-clés des sous-corpus. Ci-dessous (Tableaux 1 et 2) se trouvent les thèmes et quelques mots-clés qui les constituent.

2.2 Méthode Reinert

La méthode Reinert de la classification hiérarchique descendante a été adaptée pour le logiciel IRAMUTEQ et appliquée à notre corpus selon les modalités décrites par Ratinaud et Marchand (2012). Cette méthode consiste à identifier les unités de contexte élémentaires selon l'organisation interne du texte qui a été lemmatisé pour ensuite être réparti par classes en procédant par bipartitions successives. Comme pour l'analyse de contenu, nous avons analysé séparément les sous-corpus par langue. Les classifications obtenues ainsi ont été contrastées avec les premiers résultats obtenus à l'aide de l'analyse de contenu.

3. Corpus

Les 34 discours des conventions nationales acadiennes, prononcés de 1881 à 1890, constituent le corpus acadien de langue française, qui compte 56 368 mots. À cette époque, les Acadiens procédaient à une réorganisation sociale

par le choix de symboles nationaux. Les Loyalistes du Nouveau-Brunswick, pour leur part, sont un groupe d'Américains royalistes ayant fui le pays suite à l'Indépendance pour s'établir au Nouveau-Brunswick où ils fêtent leur centenaire en 1883. Les 40 discours du centenaire des Loyalistes, publiés entre 1882 et 1887, forment le corpus de langue anglaise qui compte 69 610 mots.

4. Analyse

L'analyse contrastive des résultats obtenus par ces deux méthodes d'analyse sont présentés par sous-corpus en affichant en premier le tableau thématique accompagné de quelques mots-clefs générés par l'analyse de contenu, suivi du dendrogramme produit par IRAMUTEQ.

4.1 Corpus des Conventions nationales acadiennes (français)

Tableau 1 : Thèmes et mots-clés extraits du sous-corpus acadien par l'analyse de contenu

Événement rassembleur (symboles)	**Progrès et avenir**	**Références au passé**	**Relations (inter)natio-nales**	**Caractéris-tiques associées au peuple**	**Race, ethnie et culture**	**Religion**
fête convention drapeau adopter distinct monument assemblée tricolore légitime étoile…	*avancement intérêts droits développe-ment sauvegarde surmonter triomphant amélioration combattre…*	*colonie histoire perdu ancêtres origine persécutés misère pères mort larmes souvenir infortune ruine…*	*compatriotes anglais union sympathie ennemi confédération américains fusion puissance Louisiane préjugés…*	*grand bonheur malheur honneur noble, digne devoir, petit courage difficultés persévérance faible. pauvre humble…*	*peuple nation race patrio-tisme sang Acadie patrie âmes usages traits…*	*saint religieuses frères foi patron Dieu Marie Église Assomption chrétien…*

La répartition par classes selon la méthode Reinert effectuée par IRAMUTEQ sépare en premier la classe 6 des autres classes. Cette classe est représentée par un lexique autour du choix d'une fête nationale acadienne, premier objectif de ce grand rassemblement patriotique. Une deuxième partition se fait entre les classes 3 et 4 et les classes 2, 1 et 5. La classe 4 est caractérisée par un lexique de valeurs associées à la religion alors que la classe 3 illustre des valeurs associées à un style de vie traditionnel attaché au passé. Le lien entre les deux est révélateur du fait que pour les Acadiens de l'époque, le style de vie traditionnel est fortement lié à la religion catholique. Si les classes 3 et 4 se réfèrent au passé, les classes 2, 1 et 5 suggèrent plutôt un regard tourné vers l'avenir, notamment dans les domaines des progrès matériel et intellectuel

(classe 2), de la presse francophone (1) et de l'éducation (5).

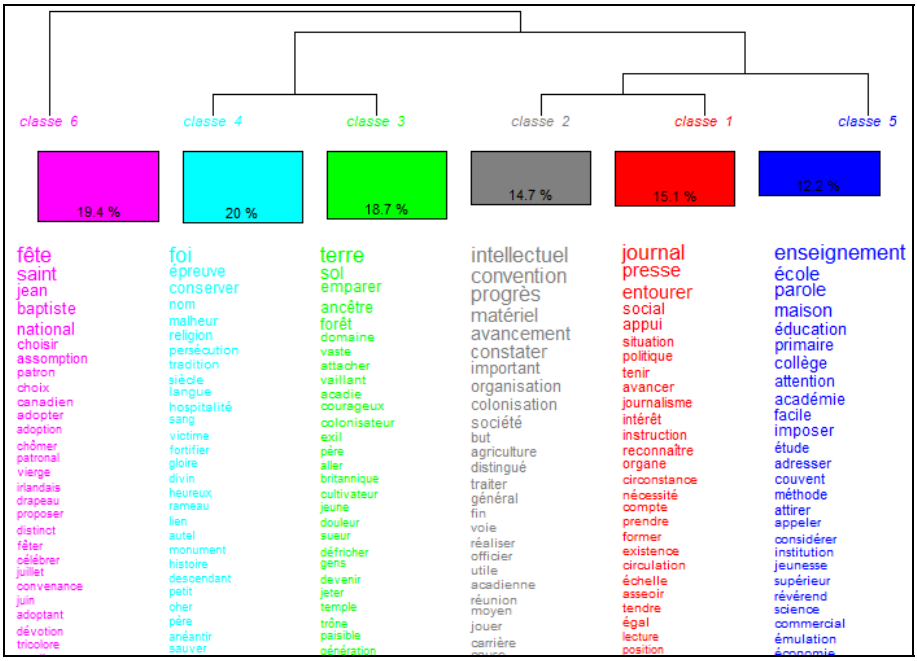


Figure 1 : Dendrogramme CHD1 – phylogram produit par IRAMUTEQ : classification hiérarchique descendante par la méthode Reinert pour le corpus acadien

Quant à la comparaison aux thèmes relevés par l'analyse de contenu traditionnelle (Tableau 1), certains rapprochements sont possibles. La classe 6 partage une quantité importante de formes avec le thème « événement rassembleur » dans l'analyse de contenu, notamment les mots-clés communs aux deux méthodologies : *fête*, *adopter*, *drapeau*, *tricolore* et *distinct*. Il est également possible de rapprocher les classes 3 et 4 des thèmes « Religion » et « Références au passé » du Tableau 1. Ces deux classes contiennent quelques mots présents sous le thème « Caractéristiques associées au peuple » du Tableau 1. Ces classes (2, 1 et 5) partagent une certaine partie de leur lexique avec le thème « Progrès et avenir » du Tableau 1.

Quel est l'apport de la méthode Reinert à notre analyse? Dans ce cas, il est pertinent de s'interroger sur ce qu'elle ne relève pas. Notamment, les catégories de l'analyse de contenu « Relations nationales et internationales » et « Race, ethnie et culture » (bien que certaines formes telles que « sang » et « Acadie » se retrouvent dans les classes 3 et 4). Ces deux thèmes se rapprochent le plus des axes d'intérêt des chercheurs, ce qui suggère une interférence humaine probable. De plus, l'ordre des partitions proposé par IRAMUTEQ, qui sépare la classe 6 et répartit les 5 autres classes entre le

passé et l'avenir, est très révélateur d'un discours paradoxal juxtaposant le progrès social à la préservation d'une identité ancrée dans le passé, ce qui n'était pas ressorti lors de l'analyse de contenu traditionnelle par thèmes.

4.2 Corpus des commémorations centenaires des Loyalistes du N.-B.

Tableau 2 : Thèmes et mots-clés extraits du sous-corpus loyaliste par l'analyse de contenu (HYPERBASE et SPHINX)

Événement rassembleur (commémoration)	**Progrès et avenir**	**Références au passé**	**Relations nationales et internationales**	**Caractéristiques associées au peuple**	**Race, ethnie et culture**	**Religion**
anniversary commemorate War, 1783 forefathers memorial Parrtown Victoria 1883, 18th Institute Regiment…	*advancement building cities commerce development establishment factories harbour hotels industrial…*	*abandoned bitterness choice confiscated defence hardship heroes, duty Israelites rugged struggle…*	*alliance annexation commonwealth constitution Independence monarchy government King Mother protection…*	*active brave brotherhood conservative determination intelligent deserving strength…*	*civil civilized humanity race superior anglo-saxon yanks elevate blood…*	*God bibles bless Christian churches devotion Faith morality temperance …*

Sept classes sont proposées dans le dendrogramme produit par IRAMUTEQ pour le corpus loyaliste en anglais. Une première répartition sépare les classes 3 et 2 de toutes les autres classes. La classe 3 est composée de références militaires à des personnages, des lieux et des dates, et la classe 2 rassemble un lexique désignant des structures associatives responsables de préserver la mémoire. Les deux classes sont caractérisées par un grand nombre de noms propres. La classe 7 se distingue ensuite par ses termes juridiques rattachés à l'empire britannique et ses colonies. Pour sa part, la classe 6 est constituée d'un lexique autour des ressources naturelles et du progrès matériel ou commercial, ce qui suggère une vision de domination de la nature par l'être humain. La classe 1 traite des valeurs morales et religieuses prisées par les Loyalistes. Finalement, les classes 4 et 5 sont très proches, et désignent respectivement les circonstances du départ des Loyalistes des États-Unis par loyauté à la couronne britannique, et la célébration de leur succès en tant que fondateurs d'une nouvelle province (le Nouveau-Brunswick) cent ans plus tard.

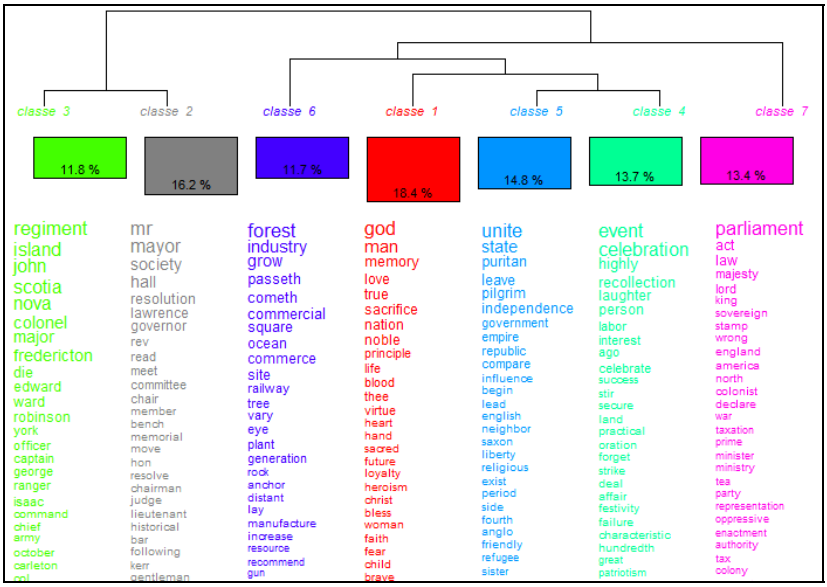


Figure 2 : Dendrogramme CHD1 – phylogram produit par IRAMUTEQ : classification hiérarchique descendante par la méthode Reinert pour le corpus loyaliste

Les classes ainsi obtenues peuvent être comparées aux thèmes du Tableau 2. Par exemple, la classe 1 (valeurs morales et religieuses) partage son lexique avec les thèmes « Religion » et « Caractéristiques associées au peuple ». La classe 4 (circonstances du départ) est très semblable au thème « Références au passé » et la classe 5 (célébration du succès) pourrait également être mise en parallèle avec « Événement rassembleur : commémoration » ainsi que le thème « Race, ethnie et culture », extrait par l'analyse de contenu. Les classes 2 (structures associatives) et 3 (références militaires) peuvent être rapprochées du thème désigné dans le Tableau 2 sous « Événement rassembleur : commémoration ». La classe 7 (empire britannique et ses colonies) se rapproche du thème « Relations nationales et internationales » sans toutefois être identique, et la classe 6 (ressources naturelles et progrès) ressemble au thème « Progrès et avenir », mais avec certaines distinctions, notamment, l'inclusion des mots se référant à la nature dans le thème du progrès matériel. L'originalité de la répartition par classes par IRAMUTEQ se trouve en partie dans la juxtaposition du passé et du présent dans les classes 3 (références militaires du passé) et 2 (associations pour la préservation de la mémoire par des activités commémoratives), ainsi que les classes 5 (célébration du succès) et 4 (circonstances du départ) qui en sont, en quelque sorte, l'écho. De plus, les catégories établies dans l'analyse de contenu se sont avérées incomplètes, et le lexique est réorganisé par la classification hiérarchique descendante. Selon les répartitions de la méthode Reinert, les

termes juridiques (*parliament*, *act*, *law*, etc.) se retrouvent avec les termes se référent à la couronne britannique et ses colonies alors qu'ils n'avaient pas été relevés dans notre étude de 2012. De même, les mots désignant le monde naturel (*forest*, *ocean*, *tree*, etc.) côtoient le lexique du progrès matériel et commercial dans le dendrogramme, ce qui n'était pas intuitif à la lecture humaine, mais fort révélateur. C'est précisément dans ces apparentes contradictions qu'apparaissent les interprétations les plus nuancées, et donc les plus judicieuses d'un corpus textuel.

5. Conclusion

Outre le fait de pouvoir traiter de corpus plus volumineux dans plusieurs langues, quels sont donc les avantages de l'application de la méthode Reinert à notre corpus bilingue? En somme, la répartition par classes nous a amené à réviser et nuancer les résultats de l'analyse de contenu originale. Si les partitions ressemblent parfois aux thèmes relevés en 2012, la méthode Reinert a l'avantage de dévoiler les liens entre les classes par ses partitions graduelles sans égard à la langue, ce qui nous a permis d'observer une répartition temporelle passé/avenir dans le sous-corpus acadien et passé/présent dans le sous-corpus loyaliste. De plus, les unités de contexte ne reposent pas sur des préconçus ou des dictionnaires internes, mais sur une répartition des mondes lexicaux qui respecte l'organisation interne des corpus, ce qui a donné une réorganisation du lexique et l'inclusion de mots qui ne figuraient pas dans l'analyse originale.

C'est justement l'inclusion de ce lexique apparemment paradoxal qui mène à une analyse plus objective et plus fine. Par exemple, le côtoiement de la nature et du progrès matériel dans les discours loyalistes suggère une vision de la domination de la nature par l'être humain et les discours acadiens visent un progrès social, économique et commercial tout en souhaitant préserver une identité ancrée dans le passé. Ainsi, nos observations sur les discours patriotiques des Loyalistes et des Acadiens à la fin du 19e siècle se trouvent considérablement enrichies par la méthode Reinert telle qu'intégrée dans le logiciel IRAMUTEQ.

Note : Cet article a bénéficié d'une subvention Savoir du Conseil de recherches en sciences humaines du Canada. Nous remercions aussi Marc-André Bouchard pour son aide technique.

Bibliographie

Baulac Y. et Moscarola J. SPHINX Solutions d'enquêtes et d'analyses de données. www.lesphinx-developpement.fr.

Brunet É. HYPERBASE Laboratoire UMR 6039 Bases Corpus Langage, Université de NICE-Sophia Antipolis.

http://ancilla.unice.fr/~brunet/pub/logiciels.html.

L'Écuyer R. (1987). L'analyse de contenu : notion et étapes. In Deslauriers, J.-P., editor. Les méthodes de la recherche qualitative. Presses de l'Université du Québec, pp. 49-64.

Ratinaud P. et Marchand P. (2012) Application de la méthode ALCESTE aux « gros » corpus
et stabilité des « mondes lexicaux » : analyse du « CableGate » avec IRAMUTEQ. Dister A., Longrée D., Purnelle G., editors, Actes/Proceedings of JADT 2012. (11e journées internationales d'Analyse statistique de Données Textuelles, pp. 845-857.

Ratinaud P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. http://www.iramuteq.org.

Richard C. et Kasparian S. (2012). Vocabulaire de l'identité nationaliste : analyse lexicale et morphosyntaxique des discours acadiens et loyalistes entre 1881 et 1890 au N.-B., Canada. Dister A., Longrée D., Purnelle G. editors, Actes/Proceedings of JADT 2012. (11e journées internationales d'Analyse statistique de Données Textuelles), pp. 845-857.

Richard C., Bourque D., Brown A., Conrad M., Davies G., Francis C., Huskins B., Kasparian S., Marquis G., Mullally, S. Base de données : Vocabulaires identitaire/Vocabularies of Identity. https://voi.lib.unb.ca

Bridge over the ocean: Histories of social psychology in Europe and North America. An analysis of chronological corpora[1]

Valentina Rizzoli, Arjuna Tuzzi
University of Padova – valentina.rizzoli@phd.unipd.it; arjuna.tuzzi@unipd.it

Abstract
Since the European Association of Social Psychology (EASP - initially called European Association of Experimental Social Psychology) has been established in 1966, what was then considered "European" social psychology has been working to affirm its own identity by presenting a distinctive brand to the rest of the world in general and to North America in particular. This study aims to compare European and U.S. social psychology through the analysis of the papers published by two of the main journals in their field: The Journal of Personality and Social Psychology and the European Journal of Social Psychology. All the abstracts (from the first publication to the last one in 2016) of the two journals papers have been collected. By means of a (lexical) correspondence analysis (SPAD software), the existence of a latent temporal pattern in keywords' occurrences was explored. Furthermore, in order to detect, retrieve and compare the main topics the journals dealt with over time, an analysis implemented by means of Reinert's method was conducted (IRaMuTeQ and R software). Results show that even if there are some typical features that distinguish the "European" from the "American" social psychology some publication trends seem to converge. Results will be discussed also reflecting on the contribution of these methods in studying the history/ies of a discipline.

Keywords: diachronic corpora, chronological textual data, text clustering, correspondence analysis, Reinert's method, history of social psychology

1. Introduction

It is widely spread that what is called "the modern social psychology" came from Europe with the migration of scholars during the second world war,

[1]This study is a new development of a an interdisciplinary research project funded by the University of Padova, fund CPDA145940 (2014) "Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature" (P.I. Arjuna Tuzzi).

and started to develop mainly in the United States. Moscovici and Markova (2006) referred to an American indigenous tradition that compete with a newer Euro-American tradition, not intending to argue that there was a socio-psychological tradition born in Europe and brought to America; but a genuinely American tradition that began with the work of the immigrant Lewin and his new students. While there was a prosperous development of social psychology in U.S., in Europe there were scholars working on social psychology, but there was no European school (Moscovici, 1999). The establishment of the European Association of (Experimental) Social Psychology (EASP - initially EAESP) in 1966 has been fundamental in the development of a "European" social psychology. EASP represented a distinctive brand of the discipline to the rest of the world in general and to North America in particular, by providing a voice for a more "social" social psychology (http://www.easp.eu/about/?). To consider an "American" and a "European" social psychology as two completely separated and counterpoised entities would be wrong since there was a clear influence between them. Moreover, the first EASP meeting, which fostered the birth of EAESP, was an initiative of U.S. scholars (cf. Moscovici and Markova, 2006). By saying "American" social psychology we usually refer to the indigenous U.S. tradition explicated by Floyd Allport's work in 1924, which considers social psychology as part of general psychology and keeps more attention on the "individual". "European" social psychology usually refers to the Euro-American tradition, promoted by the EASP, that regards social psychology as strictly connected to close disciplines such as sociology and anthropology and accords a greater role to social and cultural aspects (http://www.easp.eu/about/?). This contribute consists in an empirical analysis that moves from the study of scientific production. Over time, scientific journals shape the history of a discipline as they include objects, fields of application and methods that contribute to delineate the trajectory of a discipline. Thus, an in-depth understanding of the past and the temporal evolution of a discipline can be achieved by analysing the scientific debate inside relevant scientific journals (Trevisani and Tuzzi, 2015; 2018). We have taken into account the European Journal of Social Psychology (EJSP) and the Journal of Personality and Social Psychology (JPSP). The former is an official publication of the EASP and worldwide represents the association's voice. The JPSP belongs to the American Psychological Association, that represents the most widespread community of psychologists in the United States, and not only: It is an important scientific reference that provides guidelines also in Europe. In terms of visibility and prestige, the JPSP is considered one of the most relevant journals of the field. The main aim is to observe and compare the trajectory of the two Journal publications and to reflect about

what contribution these methods can provide for the study of the history of a discipline. We particularly intend: 1) to portray the temporal pattern of the main concepts debated in the past and covered today by EJSP and JPSP; 2) to detect, retrieve and compare the main topics these journals dealt with over time.

2. Methods

All the available abstracts of the two journals have been included in two corpora and collected from different acknowledged sources compared with the website of the journals. As regards EJSP, a total of 2,559 items was collected, for a period of 46 years, from the very first in 1971, Volume No. 1, Issue No. 1 to the latest of 2016, No. 46, Issue No. 7. Regarding JPSP, an amount of 9,568 item was downloaded, for a period of 52 years, from 1965, Volume No. 1, Issue No. 1 to 2016, No. 111, Issue No. 6. Items without any abstract have been deleted (e.g., editorials, master heads, errata, acknowledgements). The EJSP corpus is composed of 2,195 abstracts, while the JPSP one of 9,536 abstracts.

To improve the homogeneity of the corpora we decided to privilege the British spelling (e.g., we replaced *analyzed* with *analysed)* in EJSP and those American in JPSP. Our corpora have been normalised only replacing uppercase with lowercase letters. The lexicometric measures showed that there is a good redundancy, that is fundamental to work with frequencies (Lebart, Salem, & Berry, 1998; Tuzzi, 2003; Bolasco, 2013).

Multi-words (MW) with frequencies ≥ 5 for the EJSP corpus and ≥ 10 for the JPSP one (it is consistently larger than the former) have been recognised, selected and considered as textual units. We resort to a procedure for automatic information retrieval that permits to recognise repeated informative sequences, e.g., an adjective followed by a noun as in "social psychology", that produce a MW (Pavone, 2010). Two encyclopaedias of social psychology (Manstead et al., 1995; Baumeister & Vohs, 2007) and index of keywords available in the downloading process provided further MWs.

In order to depict the structure of the association between years and words and to establish the existence of a chronological dimension, a (lexical) correspondence analysis (CA) has been conducted on two matrices: 5,784 words over 46 years (rows per columns) for EJSP corpus and 8,349 x 52 for JPSP. To detect a set of relevant topics included in the journals and observe their temporal development, an analysis implemented by means of Reinert's method (1986) has been conducted. Topics can be defined as "lexical worlds" (Reinert, 1993), that are groups of words referring to a class of meaning. The result, performed with a hierarchical descending classification, is a dendrogram that groups units into classes that mirror a similar lexical

context. Textual data were processed with the Taltac2 dedicated software and statistical analyses were conducted with SPAD, Iramuteq and R software packages.

3. Results

By means of CA we can observe the existence of clear-cut temporal dimension in both the corpora (Figure 1). The keywords which mainly contributed to the factorial solution show which concepts typifies each time-span.

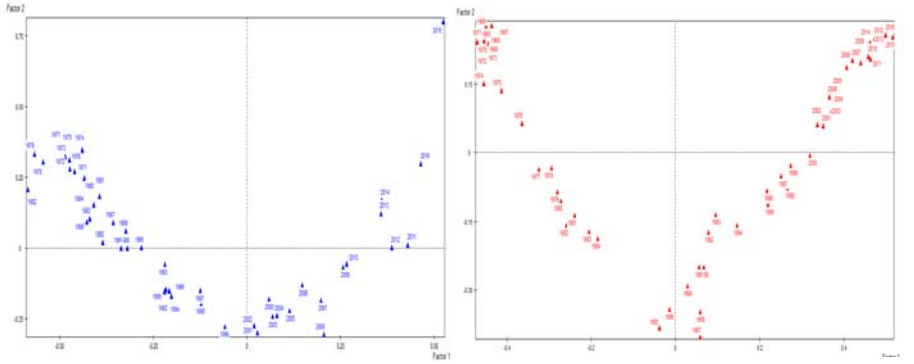


Figure 1 - First factorial plan of Correspondence Analysis of EJSP (left side) and JPSP (right side). Projection of years

In the EJSP (Figure 1, left side) the first period (1971-1990) is strongly characterised by words that refer to the experimental design. This is the period mainly concerned with the study of aggression, risk taking, dissonance, and attribution theory. The keywords of the subsequent period (nineties) seem to be related to social change, which is characterised by the study of social influence, categorization, and words referring to Moscovici and Tajfel's theories (that marked the European production: social representations, minority influence and minimal group paradigm). In the following years (2000s) we can observe that the attention has turned on the self, ingroup/outgroup relations and the social cognition with the study of stereotypes, emotions, motivation, agency/communion, and so on. In recent years (2011-2016) mainly social issues (e.g., gender, migration, environment, religion) and everyday life concerns are highlighted.

As regards the JPSP (Figure 1, right side), in the first decade considered (1965-1976) the main contribution is given by words as reinforcement, verbal reinforcement, conditioning, and so on, that together refer to behaviourism. At the same time, we can observe the occurrence of words pertain to game's theories, conflict/cooperation as well as aggression and dissonance theory.

Also physiological measurements (e.g. heart rate) and experiments (experimental) are visible. The second period includes the last Seventies until the last Eighties. Its distinctive words are masculinity/femininity, and other terms that remind to motivational theories. Moreover, the presence of words related to personality is evident and becomes stronger in the following period, that includes the Nineties. In this period mood, personality, individual differences, memory and the self represent the main contribution. At the same time also issues about gender and women are noteworthy. The last period starts from the 2000s and shows many references to explicit/implicit, and intimate relationships. Moreover, further specific words about positive psychology (life satisfaction, goal pursuit, and so on) and culture (cultural, culture) are relevant.

The analysis conducted by means of Reinert's method enlightens the presence of nine different lexical worlds (79.64% of the abstracts have been classified) in EJSP (Figure 2).

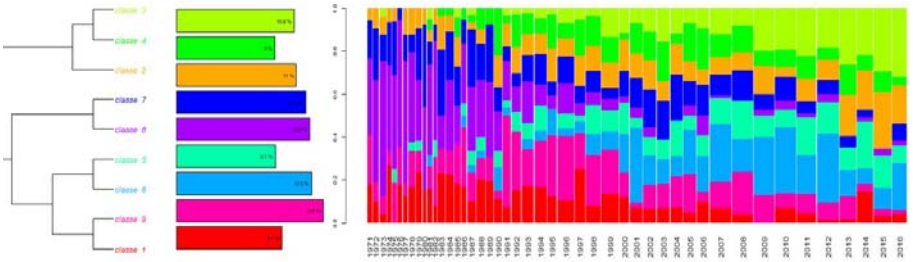


Figure 2 - EJSP classes and their distributions over years – Unsupervised clustering method

Following the classes order from the bottom to the top of Figure 2, a brief outline of their contents is provided below. Class 1 (red) concerns attribution and methodological issues (e.g., method, statistical, model). Class 9 (fuchsia) contains words related to impression formation, categorisation and stereotype. Both these classes show decreasing trends without disappearing. Class 6 (light blue) includes mainly words related to gender studies and implicit measures (e.g., prime, IAT). Class 5 (water blue) concerns moods and regulatory focus theory. These two classes show increasing trends. Class 8 (purple) concerns studies on aggression (in which mainly male/female as subjects involved in an experiment were compared). This class was initially hegemonic in the field and then disappeared along time. Class 7 (blue) includes game theories and studies on cooperation competition and shows a decreasing trend. Class 2 (orange) concerns politics and culture (mainly cross cultural studies) and it is an ever-present topic, as well as Class 4 (green) that concerns the social identity theory and ingroup/outgroup dynamics. Class 3, that concerns the applications of that theory (e.g., migration), shows a clear

increasing trend. As regards JPSP, the analysis shows the presence of eleven clusters (76,08% of the abstracts have been classified - Figure 3). Following the order of the classes from the bottom to the top of Figure 3: Class 7 (light blue) concerns consensus formations and attribution, and seems to be an ever-present topic. Class 6 (water blue) contains processes regarding memory, stereotypes and categorisation and it is particularly recurrent in the nineties and 2000s. Class 3 (grey) contains studies on self, emotion and motivation and shows a clear increasing trend, becoming one of the most relevant topics nowadays. Classes 11 (fuchsia), 10 (lilac), and 1 (red) concern, respectively, studies on aggression and physical measurements, on dissonance and opinion changes, and male and female involved in experimental studies. They were predominant in the first years considered and then disappeared. Class 9 (purple) concerns culture (mainly comparing west and east ones) and politics. It shows an increasing trend although it is not among main topics nowadays. Class 2 (orange) includes words regarding the measurements and their validity (e.g., scale, reliability, test retest) and shows a stable trend. Class 8 (blue) contains words relate to interpersonally differences (based on gender or studied with twin studies). It seems to remain constant even if with a slight decreasing trend. Class 5 (water green) is represented by words concerning health (mental and physical) and how to cope with related problems. Class 4 (green) concerns romantic and couple relationships. Both those classes show increasing trends.

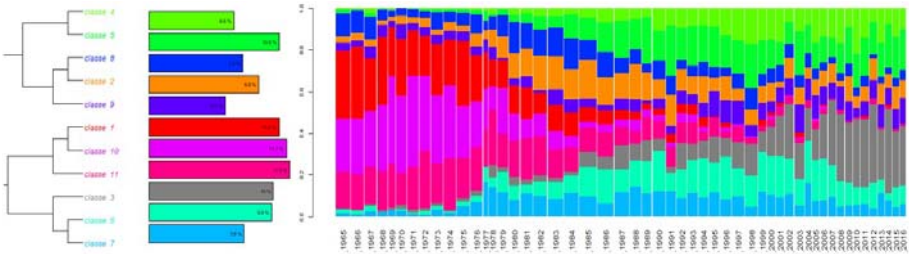


Figure 3 - JPSP classes and their distributions over years – Unsupervised clustering method

4. Discussion and conclusions

The aim of the present study is to compare American and European social psychology offering food for thought on the contribution of the methods used in studying the histories of a discipline. Thanks to these preliminary results we succeeded in highlighting the history of a discipline from the particular point of view of its effective scientific production.

In the first years considered, some similarities among the contents tackled in the two journals can be noticed (e.g., dissonance theory and aggression). The main differentiation that emerged concerns the stronger attention on

individual and personality in JPSP, on the one hand, and the different impact of Tajfel and Moscovici's contributions on the psychology of groups and Moscovici's works on social representations, on the other. This emerged as particularly evident in '80s and '90s. The predominant approach of social cognition seems to be a common feature, as well as methods and research design that mainly refer to the experimental method and topics concerning cross cultural studies and politics. As regards the topics identified, some common trajectories of publication were enlightened. They are, for example, Class 11 in EJSP and 8 in JPSP, concerning studies on aggression that were predominant in the first decades and later decline. Class 1 in EJSP and 7 in JPSP, as regards, studies on attribution. Also, class 2 in EJSP and 9 in JPSP, that are related to culture and politics. Similar contents but different trajectories are shown by Class 9 in EJSP and 6 in JPSP. The main difference between the journals is observed in JPSP Classes concerning personality, health, cope, and romantic and couple relationships (8, 5, 4), and EJSP Classes concerning ingroup/outgroup processes, and intergroup contact and applied concerns (4, 3).

It is worth mentioning the core of the difference between American and European social psychology: the attention on the individual in the American and on the social in the European one. That difference finds its way as a greater attention on social issues in EJSP and individual related studies (e.g. interpersonal relations, personality) in JPSP. Two histories of publications in social psychology have been traced, one North American and the other European. Their typical differentiation is historically well known in the community, but the empirical works that contributed to that debate are less. This is an example of the contribution that quantitative analysis of textual data can provide to the study of the history of a discipline, also known as digital history.

References

Allport, F. (1924). *Social Psychology*. Boston, MA: Houghton Mifflin.

Baumeister, R. F., & Vohs, K. D. (2007). *Encyclopedia of social psychology*. Thousand Oaks, CA: Sage.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Netherlands: Springer. doi:10.1007/978-94-017-1525-6

Manstead, A. S., Hewstone, M. E., Fiske, S. T., Hogg, M. A., Reis, H. T., & Semin, G. R. (1995). *The Blackwell Encyclopedia of Social Psychology*. Blackwell Reference/Blackwell Publishers.

Moscovici, S. (1999). Ringraziamento. In *Laurea Honoris Causa in Psicologia a Serge Moscovici*. Università degli studi di Roma "La Sapienza": Centro Stampa d'Ateneo.

Moscovici, S., & Markova, I. (2006). *The making of modern social psychology*. Cambridge: Polity.

Pavone, P. (2010). Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus. In S. Bolasco, I. Chiari, & L. Giuliano (Eds.) *Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, 9-11 June 2010, Sapienza University of Rome, pp. 131-140. LED.

Ratinaud, P. (2014). Visualisation chronologique des analyses ALCESTE: application à Twitter avec l'exemple du hashtag# mariagepourtous. *Actes des 12es Journées internationales d'Analyse statistique des Données Textuelles*. Paris Sorbonne Nouvelle–Inalco.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.

Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» àtravers l'analyse statistique d'un corpus de récits de cauchemars. *Langage & Société*, 66, 5–39.

Trevisani, M., & Tuzzi, A. (2015). A portrait of JASA: The history of Statistics through analysis of keyword counts in an early scientific journal. *Quality and Quantity*, 49, 1287-1304.

Trevisani, M., & Tuzzi, A. (2018). Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based System*, 146, 29-141

Les « itemsets fréquents » comme descripteurs de documents textuels

Louis Rompré[1], Ismaïl Biskri[2]
[1]Université du Québec à Trois-Rivières – rompre.louis@courrier.uqam.ca
[2] Université du Québec à Trois-Rivières – ismail.biskri@uqtr.ca

Abstract
Automated classification is one of the preferred approaches applied to the problem of organizing information. The classification process is based on identification and evaluation of descriptors which characterize the information. It's usually necessary to discover them following a raw data analysis. Generally, words are considered during this analysis. In this paper, we propose to use frequent itemsets as descriptors. We present how they can be identified and used to define a level of similarity between several texts. The experiments conducted demonstrate the potential of the proposed approach for defining similarity between texts and linking news broadcast on the web.

Résumé
La classification automatisée est une des principales approches appliquées au problème d'organisation de l'information. Le processus de classification repose sur l'identification et l'évaluation de descripteurs qui caractérisent l'information. Il est souvent nécessaire de déduire ces descripteurs à partir d'une analyse des données brutes. Généralement, les mots sont considérés pour mener cette analyse. Dans cet article, nous proposons d'utiliser des itemsets fréquents comme descripteurs. Les expérimentations effectuées démontrent le potentiel de cette approche pour établir un degré de similarité entre différents textes et lier des nouvelles diffusées sur le web.

Keywords: Classification, Itemset fréquent, Descripteur, Document, Texte.

1. Introduction

La digitalisation des documents a facilité la diffusion de l'information. Dès qu'un événement se produit de multiples articles sont rédigés et diffusés sur les différentes plateformes numériques. Plusieurs documents textuels diffusés sur le web sont composés uniquement de quelques centaines de mots. C'est en consultant différents documents, qu'une description riche peut être obtenue. Différents documents peuvent aborder un même sujet et chacun de ces documents est susceptible de contenir de l'information

complémentaire. Toutefois, la quantité de données disponibles et leur manque de structure limitent notre capacité à capturer ces informations d'où la nécessité d'avoir recours à des outils facilitant l'accès à l'information. La classification automatique est l'une des stratégies appliquées au problème d'organisation de l'information. Un processus classificatoire appliqué à des documents textuels, qu'il soit automatisé ou non, organise les documents de sorte que ceux qui partagent des similarités soient regroupés. L'organisation qui en découle peut être utilisée pour orienter, par exemple, la recherche d'information, l'extraction de connaissances, l'aide au résumé, etc.
Plusieurs classifieurs automatiques ont fait l'objet de publications. Comparer ces classifieurs pour déterminer leur performance est une tâche complexe et, surtout, subjective. Un classifieur peut performer avec un ensemble particulier de documents et engendrer des classes bruitées avec un autre ensemble. La pertinence d'une classification est jugée en fonction de l'homogénéité des classes qui en résultent. Ce critère est toutefois relatif. L'examen d'une classe par un intervenant est accompli à partir de ses objectifs de recherche et de ses connaissances du domaine abordé. La qualité recherchée pour un système de classification automatisée est d'être capable de cibler les informations pertinentes à l'intérieur des documents visés et de déterminer comment ces informations peuvent être utilisées pour établir un niveau de similarité entre ces documents. La classification numérique repose sur l'identification et l'évaluation de descripteurs qui permettent de différencier une classe d'une autre. Le choix d'un descripteur aux dépens d'un autre revient à prendre position sur la nature des résultats générés. Il influence le comportement du classifieur car la présence ou l'absence d'un descripteur est un indice permettant de cibler la classe à laquelle appartient un document. Pour la classification textuelle, le mot est souvent utilisé comme descripteur discriminant (McCallum et Nigam, 1998). Lorsque plusieurs mots apparaissent à des fréquences comparables dans deux documents alors ces documents sont considérés comme étant similaires. Toutefois, il est courant que des documents partagent un nombre important de mots et ce même si ces documents traitent de sujets différents. La présence seule de ces mots est donc peu porteuse d'information et son utilité pour établir le niveau de similarité entre des documents est limitée. Néanmoins, les relations qu'entretiennent ces mots avec d'autres peuvent mettre en lumière des particularités propres à certains documents. Il est possible d'utiliser ces relations pour établir le niveau de similarité entre documents.

2. Les règles d'associations

Le développement récent des règles d'association découle des travaux d'Agrawal sur l'extraction de connaissances à partir de données

transactionnelles (Agrawal et al., 1993). Agrawal proposait de dégager des relations entre des items qui cooccurrent dans des transactions commerciales. Par exemple, les clients qui achètent les items *x* et *y* achètent également l'item *z*. Depuis, l'approche a été transposée à d'autres domaines, les règles d'association pouvant être appliquées à divers domaines dans la mesure où le concept de transaction peut y être défini.

Soit T un ensemble de transactions tel que $T = \{t_1, t_2, t_3, \dots, t_n\}$, les éléments qui composent les transactions $t_i \in T$ sont appelés des *items*. Un item est une donnée dont la nature dépend du domaine abordé. Par exemple, les items peuvent correspondre à des descripteurs extraits d'une musique (Rompré et al., 2017), à des descripteurs extraits d'une image (Alghamdi et al., 2014) ou simplement à des mots extraits d'un texte (Zaïane et Antoine, 2002). Ainsi, une transaction peut être définie simplement comme un sous-ensemble de descripteurs.

Soit $I = \{i_1, i_2, i_3, \dots, i_d\}$ un ensemble de d items distincts, chaque sous-ensemble qu'il est possible de générer à partir des items $i_i \in T$ est appelé un itemset. Pour un ensemble I de taille d, le nombre d'itemsets possibles est 2^d (Tan et al., 2002). Le nombre d'itemsets potentiels est exponentiel, en fonction de la taille de I. L'objectif à atteindre lors du processus d'extraction des règles d'association étant de découvrir des relations cachées, il n'y a pas d'indice permettant de cibler les items à considérer. Ainsi, l'espace de recherche équivaut à l'ensemble des itemsets possibles. Même s'il est théoriquement possible de créer 2^d itemsets à partir d'un ensemble de taille d, en pratique plusieurs combinaisons apparaissent peu ou tout simplement pas dans les transactions. Ces combinaisons peuvent, donc, être ignorées. Le support est une mesure qui permet de cibler les itemsets à ignorer. Le support d'un itemset X représente le pourcentage des transactions de T qui contiennent X. Il est noté $S(X)$, et donné par l'équation 3.1 où n équivaut au nombre total de transactions contenues dans T et $\sigma(X)$ au support brut. Le support brut d'un itemset X représente le nombre de transactions de T qui contiennent X. Il est donné par l'équation 3.2.

$$S(X) = \sigma(X) / n \tag{3.1}$$

$$\sigma(X) = |\{t_v \mid X \subseteq t_v, t_v \in T\}| \quad (3.2)$$

Un itemset X est considéré fréquent lorsque son support est supérieur ou égal à un seuil prédéterminé. Soit X et Y deux itemsets fréquents tel que $X \cap Y = \emptyset$, une règle d'association notée $X \rightarrow Y$ traduit une relation de cooccurrence entre ces itemsets. Par convention, le premier terme est appelé l'antécédent tandis que le second est appelé le conséquent. Une règle d'association est jugée de qualité selon une mesure m et un seuil préalablement fixé. Ainsi, une règle d'association $X \rightarrow Y$ est jugée de qualité si $(X \rightarrow Y) \geq seuil$. La quantité de règles générées, leur pertinence de même que leur utilité dépendent fortement des mesures et des seuils minimaux fixés. L'évaluation des mesures d'intérêt des règles d'association a fait l'objet de plusieurs travaux (Le Bras et al., 2010 ; Geng et Hamilton, 2006; Tan et al., 2002). Même s'il existe plusieurs variantes, l'extraction des règles d'association est généralement effectuée à l'aide de l'algorithme Appriori (Agrawal et Srikant, 1994) ou FP-Growth (Han et al., 2000). D'autres algorithmes sont présentés dans (Fournier-Viger et al., 2017). Les deux principales difficultés liées à l'extraction des règles d'association sont la gestion de la mémoire et l'effort computationnel nécessaire à la recherche des itemsets fréquents. Contrôler le nombre d'items à considérer demeure le meilleur moyen de traiter ces difficultés. Depuis deux décennies plusieurs travaux portent sur l'application des règles d'association à des fins de classification (Liu et al., 1998; Zaïane et Antoine, 2002 ; Bahri et Lallich, 2010). Les différents classifieurs qui découlent de ces travaux produisent des résultats qui sont en mesure de rivaliser avec ceux obtenus à l'aide d'autre approches comme les arbres de décision (Mittal et al., 2017). Le principal avantage des classifieurs à base de règles d'association est que les connaissances qu'ils exploitent pour guider le processus classificatoire peuvent être facilement interprétées. Ainsi, un classifieur qui exploite des règles d'association peut être utilisé pour identifier les descripteurs pertinents. Les différentes approches proposées dans la littérature impliquent généralement des règles de la forme $X \rightarrow c$ où X correspond à un ensemble de descripteurs et C à une classe de similarité. Les documents sont considérés comme étant les transactions tandis que les descripteurs (mots clés, fréquence d'apparition des mots, etc.) et les classes sont considérés comme étant les

items. Soit un ensemble de descripteurs $D = \{d_1, d_2, d_3, \ldots, d_i\}$, et un ensemble d'étiquettes $C = \{c_1, c_2, c_3, \ldots, c_k\}$ représentant différentes classes, alors un ensemble de documents peut être représenté de la manière suivante :

$$document_1 = \{d_{10}, d_{12}, d_{16}, d_{20}, d_{22}, d_{26}, c_1\}$$
$$document_2 = \{d_{10}, d_{16}, d_{22}, d_{26}, d_{30}, d_{32}, c_1\}$$
$$document_3 = \{d_{43}, d_{46}, d_{48}, d_{49}, d_{89}, d_{60}, c_2\}$$

Cette forme de représentation implique que les classes de similarité auxquelles appartiennent les documents soient préalablement connues. Un ensemble d'apprentissage est constitué et utilisé pour entraîner le classifieur. Les règles d'association dégagées lors de la phase d'entraînement sont utilisées pour prédire la classe de nouveaux documents. Ce processus demande généralement un effort considérable et les résultats générés dépendent de l'ensemble utilisé pour entraîner le classifieur.

3. Méthodologie

À l'instar des classifieurs à base de règles d'association, notre approche exploite des itemsets fréquents pour décrire les documents. Toutefois, elle ne nécessite pas de phase d'entraînement. Des itemsets fréquents sont extraits de chacun des documents et comparés. Le degré de similarité entre deux documents est fonction du nombre d'itemsets fréquents qu'ils partagent. L'hypothèse derrière cette approche est que lorsque des mots co-occurrent fréquemment au sein des phrases qui composent un texte, alors ces mots sont représentatifs de ce texte. Ainsi, en considérant quelques itemsets fréquents, il est possible de dégager les thèmes spécifiques traités dans les documents. L'approche proposée comporte 4 étapes.

La première étape consiste à segmenter les documents afin de les préparer à l'extraction des itemsets fréquents. Les documents sont traités comme des ensembles de transactions où les phrases constituent les transactions et les mots les items. Le nombre de mots différents susceptibles d'apparaître dans un ensemble de documents textuels est théoriquement de l'ordre de la taille du vocabulaire de la langue d'écriture de ces documents. Le nombre de mots qui composent le français est estimé par l'Office Québécois de la Langue Française à plus de 500 000. Considérant qu'à partir de 500 000 mots il est possible de générer $2^{500\,000}$ itemsets, il est nécessaire d'imposer certaines conditions aux textes en entrée afin de contrôler le nombre de mots. La

diversité d'un lexique augmentant avec la taille d'un texte, nous devons limiter les textes en entrée à quelques milliers de mots.
La deuxième étape consacre la réduction du nombre d'items et donc de l'espace de recherche lors de l'extraction des itemsets fréquents. Certains mots jugés peu porteur d'information sont supprimés des transactions. Une liste de 502 mots vides est utilisée. Les chiffres et les caractères de ponctuation sont également supprimés.
La troisième étape vise à extraire les itemsets fréquents. Cette étape est réalisée à l'aide de l'algorithme Apriori. Un effort est porté afin de dégager un nombre restreint d'itemsets fréquents. La recherche des itemsets fréquents est effectuée de manière itérative. Lors de la première itération, le support minimum est fixé à une valeur élevée. Lorsque le nombre d'itemsets fréquents extraits est inférieur à 10 alors le support minimum est diminué de 0.1. Le processus cesse lorsque le nombre d'itemsets obtenus est supérieur à 10 ou que le support minimum est inférieur à 0.1.
La dernière étape établit le degré de similarité entre les documents. Les itemsets fréquents utilisés pour décrire les documents sont comparés. Plus le nombre d'itemsets partagés par deux documents est grand, plus ces documents sont jugés comme étant similaire.

4. Expérimentation et discussion

Afin d'évaluer l'approche proposée, plusieurs expérimentations ont été effectuées avec une application que nous avons développée en Python. Un corpus formé d'une centaine d'articles tirés de l'actualité et diffusés sur le web a été utilisé. Ce corpus se distingue par le fait qu'il présente les mêmes nouvelles sous l'angle de différentes agences de presse. Il regroupe des articles diffusés sur le web provenant de 6 sources différentes et contenant entre 500 et 1500 mots. Ces articles sont parfaitement adaptés aux conditions de l'approche proposée.
Lors de nos expérimentations, nous avons mesuré le pouvoir discriminant des itemsets fréquents. Nous avons effectué une comparaison entre les classifications produites lorsque les descripteurs sont les itemsets fréquents et les classifications produites lorsque les mots sont les descripteurs. La nature des résultats obtenus suggère que les itemsets fréquents peuvent servir à raffiner la description d'une classe. À titre d'exemple, le mot {avions} est utilisé pour décrire 15% des articles du corpus. Même si ces articles sont associés à l'aviation, ils traitent néanmoins de 4 sujets différents. Nos expérimentations démontrent que l'utilisation des itemsets fréquents comme descripteurs peut servir à décrire plus précisément le contenu de ces articles. Les figures 1 et 2 illustrent respectivement la précision obtenue en considérant des itemsets fréquents et celle obtenue en considérant

uniquement des mots. Il est à noter que lorsque seuls les mots sont considérés, les classes de similarité générées sont moins homogènes. En effet, des articles qui traitent de sujets autres que l'aviation y sont inclus.

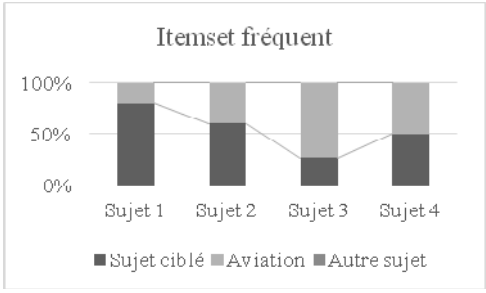


Figure 1 : Précision avec les itemsets fréquents

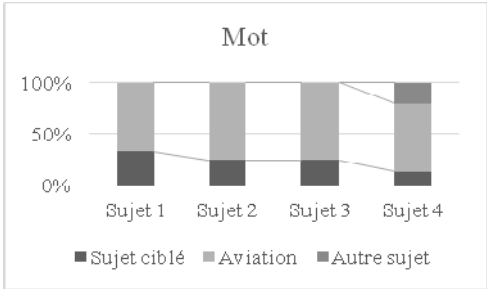


Figure 2 : Précision avec les mots

La figure 3 illustre la matrice de similarité produite pour des articles traitant de la crise nord-coréenne. La première colonne contient l'identifiant de l'article, la seconde indique le sujet abordé tandis que les colonnes suivantes donnent le nombre d'itemsets fréquents partagés par les articles. La diagonale équivaut aux nombres d'itemsets fréquents extraits pour un article. La figure 2 est représentative des résultats observés. Moins de 10 itemsets fréquents ont été extraits pour la moitié de ces articles. Néanmoins, ils ont tous été associés à la même classe.

		ID 29	ID 32	ID 33	ID 35	ID 37	ID 45	ID 46	ID 47
ID 29	Tir missile Corée du Nord	14	4	4	1	5	1	1	1
ID 32	Corée du Nord, spectre guerre	4	6	5	1	4	1	1	1
ID 33	Corée du Nord, spectre guerre	4	5	5	1	4	1	1	1
ID 35	Tir missile Corée du Nord	1	1	1	7	4	1	1	1
ID 37	Tir missile Corée du Nord	5	4	4	4	45	1	1	1
ID 45	Nucléaire : Washington prêt à discuter	1	1	1	1	1	4	1	1
ID 46	Nucléaire : Washington prêt à discuter	1	1	1	1	1	1	11	1
ID 47	Nucléaire : Washington prêt à discuter	1	1	1	1	1	1	1	15

Figure 3 : Matrice de similarité des documents traitant de la crise Nord-coréenne.

Malgré le fait qu'ils traitent du même sujet, certains articles partagent peu d'itemsets fréquents avec les autres articles qui forment la classe. Ceci s'explique par le lexique employé. Il est possible que les performances puissent être améliorées en ajoutant une étape de lemmatisation. Toutefois, certaines relations demeurent difficiles à établir automatiquement. Par exemple, le document 45 contient les itemsets {nucléaire, pyongyang} et {nucléaire, washington} tandis que le document 46 contient les itemsets {nucléaire, corée} et {nucléaire, américaine}. Les résultats présentés constituent uniquement un échantillon des connaissances extraites à l'aide de

l'approche proposée. En plus d'être faciles à interpréter, les itemsets fréquents permettent de décrire plus précisément le contenu des documents que les mots seuls.

5. Conclusion

Nous avons proposé une approche non supervisée pour établir des relations entre des documents textuels. L'approche proposée repose sur l'utilisation d'itemsets fréquents. Ces descripteurs expriment la cooccurrence de mots au sein des phrases qui composent un texte. Les itemsets fréquents ont tendance à être plus discriminant que les mots seuls. Par conséquent, ils peuvent aider à rehausser la description d'une classe. L'un des avantages de la méthode proposée est que les résultats produits sont faciles à interpréter. Les expérimentations effectuées suggèrent que les itemsets fréquents, tels que définis, sont suffisamment informatifs pour servir à établir des liens cohérents entre des documents. Plusieurs débouchés sont envisageables. Entre autres, l'approche proposée pourrait servir comme prétraitement à la navigation entre différents documents, à l'annotation, au filtrage de l'information, etc.

Références

Agrawal, R., Imielinski T., et Swami, A. (1993). Minning association rules between sets of items in large databases, In Proc. of the SIGMOD Conference on Management of Data, pp 207-216.

Agrawal, R., Srikant, R. (1994). Fast Algorithms for Mining Association Rules, In Proc. of the 20th International Conference on Very Large Database, pp. 487-499

Alghamdi, R. A., Taileb, M., et Ameen, M. (2014). A new multimodal fusion method based on association rules mining for image retrieval. In Mediterranean Electrotechnical Conference (MELECON), 2014 17th IEEE (pp. 493-499). IEEE.

Bahri, E., et Lallich, S. (2010). Proposition d'une méthode de classification associative adaptative. 10eme journées Francophones d'Extraction et Gestion des Connaissances, EGC 2010, pp. 501-512.

Fournier-Viger, P., Lin, J. C. W, Vo, B., Chi, T. T., Zhang, J. et Le, H. B. (2017). A survey of itemset mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

Geng, L., et Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), vol. 38, no 3, p. 9.

Han, J., Pei, J., et Yin, Y. (2000). Mining frequent patterns without candidate generation. In ACM sigmod record (Vol. 29, No. 2, pp. 1-12). ACM.

Le Bras, Y., Meyer, P., Lenca, P., et Lallich, S. (2010). Mesure de la robustesse de règles d'association. QDC 2010.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In Knowledge Discovery and Data Mining, pp. 80–86.

McCallum, A., et Nigam, K. (1998). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

Mittal, K., Aggarwal, G., et Mahajan, P. (2017). A comparative study of association rule mining techniques and predictive mining approaches for association classification. International Journal of Advanced Research in Computer Science, 8(9).

Rompré, L, Biskri, I et Meunier, J-G (2017). Using Association Rules Mining for Retrieving Genre-Specific Music Files, In Proc. of FLAIRS 2017, pp. 706-711.

Tan, P. N., Kumar, V., et Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 32-41). ACM.

Zaïane, O. R., et Antonie, M. L. (2002). Classifying text documents by associating terms with text categories. In Australian computer Science communications (Vol. 24, No. 2, pp. 215-222).

Discursive Functions of French Epistemic Adverbs: What can Correspondence Analysis tell us about Genre and Diachronic Variation?

Corinne Rossari, Ljiljana Dolamic, Annalena Hütsch,
Claudia Ricci, Dennis Wandel
University of Neuchâtel – corinne.rossari@unine.ch

Abstract
Our aim is to describe discursive functions of a set of French epistemic adverbs by establishing their combinatory profiles on the basis of their co-occurrence with different connectors. We then compare these profiles using correspondence analysis in order to find evidence of genre and diachronic variation. The use of these adverbs is explored in contexts of informative discourse within two distinctly different genres – contemporary written press and encyclopedic discourse – as well as within two diachronic spans.

Keywords: epistemic adverbs, connectors, co-occurrences, correspondence analysis, genre variation, diachronic variation

1. Introduction

Our aim is to analyze the genre and diachronic variation of discursive functions of French epistemic adverbs (E-ADV). By discursive function we mean the rhetorical aim of the utterance in which the adverb occurs: counter-argument, argument, or conclusion (cf. Roulet et al., 1991). Our paradigm of E-ADVs consists of the following items: *certainement*, *certes*, *peut-être*, *probablement*, *sans doute* and *sûrement*[1]. The functions of these adverbs are explored in contexts of informative discourse within two distinctly different genres: contemporary written press and encyclopedic discourse. The former is represented by three daily newspapers: *Le Monde* (2008, 20 410 766 tokens), *Le Figaro* (2008, 10 795 373 tokens) and *Sud-Ouest* (2002, 29 763 988 tokens). In the latter, we consider two diachronic spans: the 18th century, represented by Diderot & d'Alembert's *Encyclopédie* (DDA, 29 940 181 tokens) and the 21st century, represented by the 2005 edition of *Encyclopédie Universalis* (UNI, 49 859 864 tokens) and by a random sample of the 2015 version of *Wikipédia*

[1] Selection based on Roulet's (1979) paradigm of epistemic assertive adverbs.

(WIKI, 50 396 345 tokens).[2]
We first proceed to an analysis based on the combinatory profile of each E-ADV (section 2) in our corpus of contemporary written press, and then, after having pinpointed what such an analysis can and cannot show, we use a more holistic approach based on correspondence analysis (section 3).

2. Analysis of Combinatory Profiles

In order to identify the discursive functions of the E-ADVs considered here, we searched connectors (C) specifically co-occurring with each of these E-ADVs within a 20-token span. We have chosen a 20-token span rather than a sentence span, because a connector's combinatory profile can go beyond the sentence boundaries. We define connectors as linguistic forms linking segments of discourse. Such a functional category is not part of the tagset of the platform we used. We therefore made our query by searching for three different categories: adverbs, subordinating conjunctions and coordinating conjunctions. We then manually filtered the resulting forms by keeping those which proved to function as a connector.
For all our sub-corpora, each of these adverbs is thus specifically assigned a series of connectors within constructions of the type "E-ADV…C1/C2/Cn" and "C1/C2/Cn…E-ADV", which represent their *discursive combinatory profile*[3]. We call each sequence within a combinatory profile a *discourse movement* as we consider it to have specific, rhetorically motivated discursive aims. These aims (mentioned in section 1) are signaled by the connectors co-occurring specifically with an E-ADV: *néanmoins* and *mais* signal that the utterance preceding them is a counter-argument to the utterance they introduce; *donc* and *finalement* signal that the utterance they introduce is a conclusion; *car* and *parce que* signal that the utterance they introduce is an argument in favor of the utterance preceding them.
The tables below show the discursive combinatory profiles in three sub-corpora of contemporary press (*Le Monde 2008* ; *Le Figaro 2008* ; *Sud-Ouest 2002*). The significance of each co-occurrence of a connector with an E-ADV is calculated using log-likelihood (LL).[4]

[2] All the corpora used were supplied by the platform *BTLC* (*Base Textuelle Lexico-statistique de Cologne*), conceived by Sascha Diwersy (Diwersy, 2014), and were constituted within the French-German projects *Presto* (http://presto.ens-lyon.fr) and *Emolex* (http://emolex.u-grenoble3.fr).

[3] We adapt the term *combinatory profile* used by Blumenthal et al. (2005) and Blumenthal (2008; 2012).

[4] Although LL can be directly calculated on the BTLC platform, we used the platform to extract the corresponding frequencies, and calculated the LL by using R.

Tables 1-3. Log-likelihood scores (threshold: 10.83; all scores equal or above are marked in bold).

***Le Monde* (2008)**	*certainement (385)*		*certes (1943)*		*peut-être (2900)*		*probablement (723)*		*sans doute (2482)*		*sûrement (307)*	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*
car (6 706)	0.08 (3)	0.08 (3)	-0.27 (11)	0 (13)	**49.30** (57)	-0.37 (12)	**18.99** (17)	0.92 (7)	**16.09** (35)	0.02 (17)	1.51 (4)	-0.64 (1)
donc (8 276)	2.09 (6)	2.09 (6)	-2.47 (10)	2.9 (23)	-1.45 (18)	**24.09** (51)	-0.68 (4)	1.43 (9)	0.03 (21)	4.18 (30)	0.10 (3)	0.10 (3)
finalement (1 559)	-1.18 (0)	-0.01 (1)	-1.77 (1)	1.14 (5)	3.60 (9)	**15.43** (15)	-0.01 (1)	0.58 (2)	0.01 (4)	6.96 (10)	-0.94 (0)	6.08 (3)
mais (51 544)	**29.83** (47)	**27.94** (46)	**3248** (979)	-22.55 (52)	**371.45** (423)	**107.28** (284)	10.30 (57)	10.30 (57)	**205.88** (310)	**32.85** (193)	**30.90** (41)	**65.68** (55)
néanmoins (968)	-0.73 (0)	-0.73 (0)	5.84 (6)	**14.22 (9)**	0.02 (3)	-0.23 (2)	0.12 (1)	-1.38 (0)	-0.06 (2)	-0.06 (2)	-0.58 (0)	-0.58 (0)
parce que (2 514)	0.88 (2)	2.81 (3)	1.78 (8)	-4.47 (1)	**62.15** (37)	9.75 (17)	2.03 (4)	-3.58 (0)	**86.58** (41)	0.12 (7)	3.78 (3)	0.07 (1)

***Le Figaro* (2008)**	*certainement (268)*		*certes (1084)*		*peut-être (1851)*		*probablement (441)*		*sans doute (1240)*		*sûrement (211)*	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*
car (3 922)	-0.57 (1)	-0.57 (1)	3.89 (14)	-2.35 (4)	**28.08** (37)	-0.48 (11)	**14.27** (12)	1.95 (6)	8.45 (19)	4.43 (16)	7.53 (6)	-3.08 (0)
donc (4 763)	-1.02 (1)	-1.02 (1)	-0.03 (9)	1.81 (14)	-20.39 (2)	3.16 (24)	-0.22 (3)	6.74 (10)	-0.37 (9)	0.37 (13)	-3.74 (0)	-0.48 (1)
finalement (1 150)	-1.14 (0)	-1.14 (0)	-0.04 (2)	-0.95 (1)	-3.15 (1)	6.52 (10)	0 (1)	0 (1)	1.67 (5)	-1.34 (1)	-0.90 (0)	-0.90 (0)
mais (28 552)	**36.23** (41)	**14.25** (30)	**1757.55** (545)	-1.39 (49)	**245.20** (281)	**93.30** (204)	0.56 (27)	2.38 (31)	**86.88** (151)	**34.59** (117)	**17.23** (27)	**87.11** (52)
néanmoins (580)	1.07 (1)	-0.58 (0)	10.02 (6)	0.49 (2)	0.44 (3)	-3.84 (0)	-0.95 (0)	-0.95 (0)	-2.67 (09	-2.67 (0)	-0.45 (0)	-0.45 (0)
parce que (1 435)	0.10 (1)	-1.43 (0)	-1.65 (1)	-0.30 (2)	**31.90** (22)	-2.25 (2)	6.88 (5)	-0.03 (1)	4.79 (8)	0.14 (4)	0.28 (1)	2.22 (2)

***Ouest Sud* (2002)**	*certainement (1277)*		*certes (2795)*		*peut-être (4950)*		*probablement (812)*		*sans doute (3930)*		*sûrement (684)*	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*
car (12 434)	10.78 (23)	6.53 (20)	2.89 (32)	5.92 (36)	**44.71** (91)	-0.29 (38)	1.34 (10)	-1.35 (4)	**45.13** (78)	-0.26 (30)	6.87 (13)	-0.09 (5)
donc (19 185)	-3.00 (10)	5.72 (27)	-6.45 (22)	0.10 (38)	-1.98 (53)	1.55 (74)	-1.32 (7)	9.77 (22)	-1.61 (42)	9.50 (74)	-0.41 (7)	0.15 (10)
finalement (2 858)	-0.09 (2)	0.11 (3)	3.19 (10)	0.07 (6)	7.35 (19)	3.68 (16)	-0.23 (1)	-0.23 (1)	2.23 (12)	0.72 (10)	-0.08 (1)	0.31 (2)
mais (77 108)	**28.87** (113)	**64.77** (139)	**6962.42** (1778)	-5.72 (118)	**520.54** (682)	**211.15** (513)	10.35 (64)	9.48 (63)	**209.38** (434)	**123.59** (376)	7.09 (52)	**162.80** (130)
néanmoins (1 698)	-0.16 (1)	-0.16 (1)	9.25 (10)	-0.01 (3)	5.39 (12)	-0.54 (4)	0.93 (2)	-1.85 (0)	6.72 (11)	1.20 (7)	0.06 (1)	0.06 (1)
parce que (5 981)	-4.77 (11)	-13.45 (6)	**12.51** (15)	8.48 (13)	**814.50** (135)	**104.89** (33)	**16.58** (13)	-0.57 (2)	**233.04** (108)	0.00 (16)	-1.04 (8)	-0.48 (9)

The data lead to the following observations: (i) Although the E-ADVs belong to the same semantic class, each has its own specific combinatory profile. (ii) Certain E-ADVs share comparable combinatory profiles: *sans doute* and *peut-être* share an almost identical set of specific connectors; more frequently, several E-ADVs essentially only share one or more specific connectors (for instance the connector *mais* for *certainement*, *sûrement*, *peut-être* and *sans doute*). (iii) Certain E-ADVs stand out for their unique combinatory features: *certes* is almost exclusively associated with *mais*, but only with *mais*_R, and with a notably higher log-likelihood score than the other E-ADVs. *Probablement* is also associated with only a few connectors, but with a low log-likelihood score, close to the threshold of 10.83. (iv) There is homogeneity in the significant association for each E-ADV in the three sub-corpora of contemporary press. However, preceding studies – Rossari et al. (2016) and Rossari & Salsmann (2017) – show that the E-ADVs' combinatory profile varies depending on different genres and diachronic periods: contrary to what is observed for the press genre, in DDA and UNI the association *peut-être…mais* is less significant than the association *mais…peut-être*. For instance, in DDA, no significant association *certes…mais* is observed, while the association *sans doute…mais* in the same corpus proves to be highly significant. The analysis of combinatory profiles (based on the significance measure log-likelihood; cf. Blumenthal et al., 2005) allows for one-to-one comparison of the different sequences of the type E-ADV...C and C...E-ADV. Thus, the associations of each E-ADV with each connector can easily be compared across corpora representing different newspapers, but also across different genres and diachronic periods. It is also possible to compare the

associations of different E-ADVs with one or a few connectors. However, this method has certain insufficiencies when it comes to simultaneously comparing all of these variables in a holistic view. This type of analysis of combinatory profiles never takes into account all variables at the same time (e.g. frequencies, log-likelihood scores, paradigm of E-ADVs, paradigm of connectors). Moreover, using a threshold (in our case 10.83) in order to decide whether an association is significant is useful for traditional collocation analysis. But our goal is to also represent the use of each E-ADV in its typical discourse movements in contrast to its non-typical discourse movements. It thus seems counterproductive that all sequences (E-ADV...C/C...E-ADV) which are not statistically significant for certain E-ADVs as such are not taken into account when establishing their combinatory profiles, since these nonsignificant cases play an important role in characterizing the overall use of the E-ADVs and connectors. In order to allow for a holistic approach, we propose to use correspondence analysis (CA) (Greenacre, 2017).

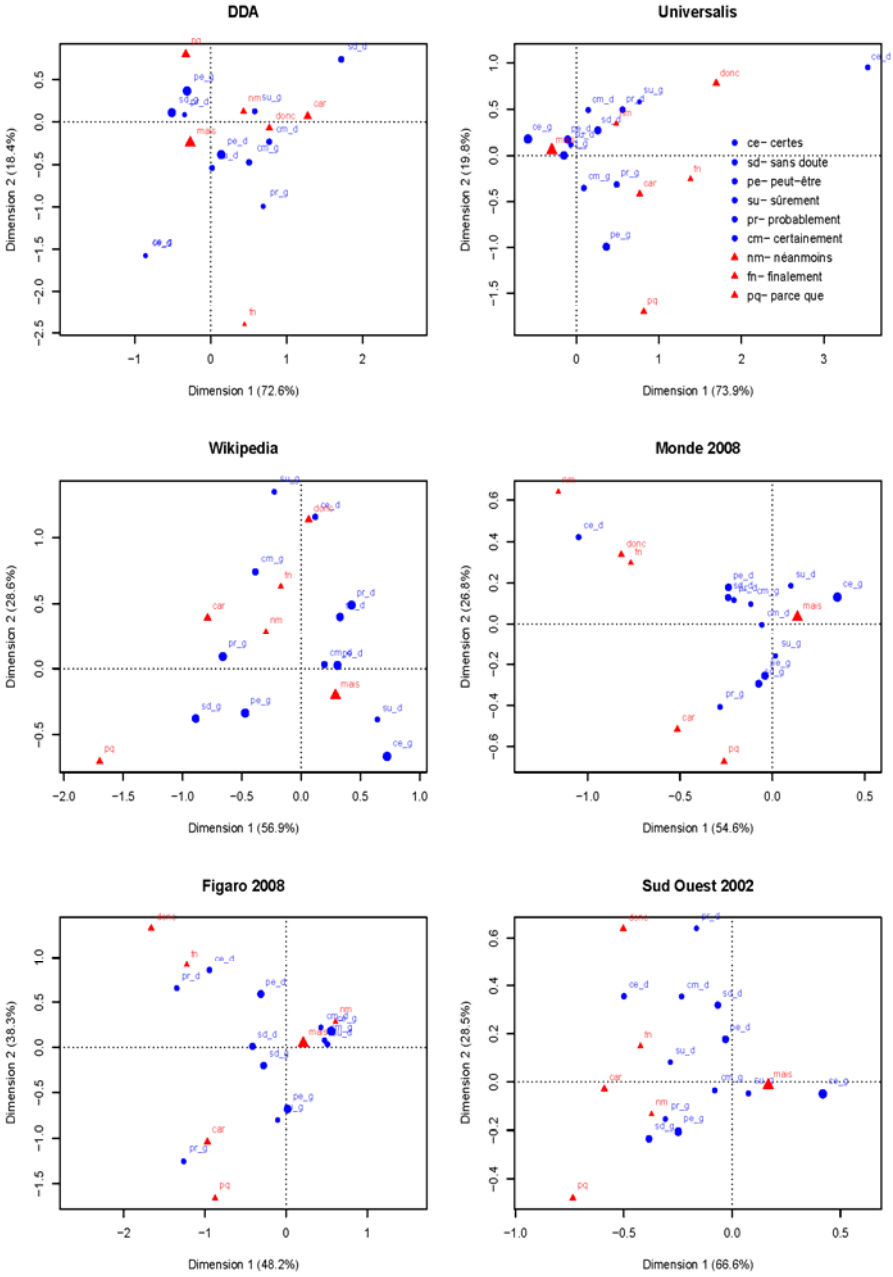
3. Correspondence Analysis (CA)

The correspondence analysis presented in this section was performed using the R software and the package "CA" (Nenadić & Greenacre, 2007). (1) In DDA, representing the 18th century, *certes* has a use which stands out. *Certes* left and right of *mais* differ clearly from all other E-ADVs as to their associations with the connectors. *Certes* is not typically used with any other connector analyzed and, most importantly, its association is not stronger with *mais* on its right than it is with *mais* on its left. Conversely, in all other five sub-corpora (encyclopedic and press corpora), which represent the 21st century, there is an important difference between the use of *certes* right and left of *mais*: while *certes*_L is strongly linked to *mais*, *certes*_R is not. (2) In all six sub-corpora, *mais* appears to be opposed to all other connectors when it comes to its associations with E-ADVs. Its central position appears to be linked to its high frequency, indicating its high contribution to the horizontal axis, this being confirmed by the analysis of the correspondence analysis indicators. (3) An association between *sans doute*_L and *parce que* can be observed in DDA and WIKI, whereas in UNI, the adverb and the connector appear to be in the opposite relation. This behavior indicates variation has to be expected even within the encyclopedic sub-corpus, based on at least two parameters: on the one hand, the diachronic parameter is involved in some discursive uses of E-ADVs, like *certes*_L and *certes*_R showing no difference as to their association with *mais* in DDA, consistently with its different meaning at that time, whereas only *certes*_L is associated with *mais* in all other sub-corpora; on the other hand, some convergence between DDA and

WIKI could be interpreted as showing similarities in writing style. (4) The results of the correspondence analysis show that in all sub-corpora of one particular genre, in most cases, the same E-ADVs are strongly associated with the same connector or group of connectors (*donc* and *finalement* ; *car* and *parce que* ; *mais*); this phenomenon is particularly pronounced in the sub-corpora representing written press. The connector *mais* differs the most from the other connectors in what concerns the strength of its associations. Although *mais* is associated with most E-ADVs, its association appears to be strong with only a few of them in all sub-corpora (*certes*_L being the only constant), while most other connectors have a higher number of strong associations. This indicates that certain discourse movements (such as E-ADV...*car* / *parce que*) seem to be rather regular, whereas *certes...mais* proves to be a special association, although only in the 21st century corpora. (5) The behavior of *néanmoins* in the *Figaro 2008* corpus should be interpreted with caution since the two axes describe only 10% of its variation.

4. Perspectives

Our first attempt to use correspondence analysis to study different discursive movements has provided promising results regarding the genre and diachronic variation of discursive functions of French epistemic adverbs in these cases. We intend to further extend our analysis in three directions: First, we would like to enlarge our corpora to see if this allows to extend the paradigm of connectors, so as to give a better overview of the different discursive movements that exist and to better represent the different discursive functions of the E-ADVs that we have found. It would be especially interesting to cover different diachronic spans of press, allowing for a study of possible changes within this specific genre. Likewise, other text types may be considered in order to better represent possible variation between genres. Second, through the comparative analysis of the discursive combinatory profiles of each E-ADV, we aim to identify regularities concerning the rhetorical purpose of the sequence in which the E-ADV typically occurs by understanding its motivation. For instance, beyond the difference between a counter-argument, an argument, and a conclusion, there is a more fundamental difference between a discourse movement used with the rhetorical aim (i) to present a content as being in the discursive background (when the E-ADV is followed by *mais*), (ii) to introduce a content which the speaker considers to be most relevant (when the E-ADV is introduced by *mais* or *donc*), and (iii) to add evidence to a relevant content (when the E-ADV follows *car* or *parce que*). Third, in order to confirm the reliability and precision of the positions on the correspondence analysis planes, our intention is to apply bootstrap validation (Lebart, 2010).



Figures 1-6. Correspondence analysis scatter plots for the six corpora.

References

Blumenthal P. (2008). Combinatoire des prépositions : approche quantitative. *Langue française*, 157: 37-51.

Blumenthal P. (2012). Particularités combinatoires du français en Afrique : essai méthodologique. *Le français en Afrique*, 27: 55-74.

Blumenthal P., Diwersy S. and Mielebacher, J. (2005). Kombinatorische Wortprofile und Profilkontraste. Berechnungsverfahren und Anwendungen. *Zeitschrift für romanische Philologie*, 121: 49-83.

Diwersy S. (2014). Corpus diachronique de la presse française : base textuelle créée dans le cadre du projet ANR-DFG PRESTO. Institut des Langues Romanes, Université de Cologne.

Greenacre M. J. (2017). *Correspondence analysis in practice*. 3rd ed. Boca Raton: Chapman.

Lebart L. (2010). Validation techniques for textual data analysis. *Statistica Applicata - Italian Journal of Applied Statistics*, 22(1): 37-51.

Nenadić O. and Greenacre M. J. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3): 1-13.

Rossari C., Hutsch A., Ricci C., Salsmann M. and Wandel, D. (2016). Le pouvoir attracteur de *mais* sur le paradigme des adverbes épistémiques : du quantitatif au qualitatif. In Mayaffre D. *et al*. (eds), *Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, II: 819-823.

Rossari C. and Salsmann M. (2017). Étude quantitative des propriétés dialogiques des adverbes épistémiques. *Actes des 9èmes Journées Internationales de la Linguistique de corpus*: 87-93.

Roulet E. (1979). Des modalités implicites intégrées en français contemporain. *Cahiers Ferdinand de Saussure*, 33: 41-76.

Roulet E., Auchlin A., Moeschler J., Schelling M. and Rubattel C. (1991). *L'articulation du discours en français contemporain.* 3rd ed. Bern: Lang.

Misleading information in online propaganda networks

Vanessa Russo[1], Mara Maretti[2],
Lara Fontanella[3], Alice Tontodimamma[4]
[1]D'Annunzio University of Chieti-Pescara – russov1983@gmail.com
[2]D'Annunzio University of Chieti-Pescara – mara.maretti@unich.it
[3]D'Annunzio University of Chieti-Pescara – lara.fontanella@unich.it
[4]D'Annunzio University of Chieti-Pescara – alicetontodimamma@gmail.com

Abstract 1
Nowadays, the spreading of inaccurate, false or misleading information over the digital space is amplified by the increasing use of social networks and social media. In different cases, misleading information can be linked to a propaganda activity aimed at supporting offline organizations. In fact, in such cases, online pages, conveying unintentionally (*misinformation*) or intentionally (*disinformation*) inaccurate information, are embedded into a network system composed by political and ideological advertise. In this paper, we discuss the different structures of online networks linked to some official pages of different political parties. The analyzed networks were identified through Social Network Analysis.

Abstract 2
La diffusione di informazioni inesatte, false o fuorvianti nello spazio digitale è amplificata dal crescente uso di social network e social media. In diversi casi, tali informazioni approssimative e/o fuorvianti possono essere collegate ad un'attività di propaganda volta a supportare organizzazioni offline. Infatti, in questi casi, le pagine online, che trasmettono informazioni non intenzionalmente (*misinformation*) o intenzionalmente (*disinformation*) errate, sono incorporate in un sistema di rete composto da pubblicità politica e ideologica. In questo articolo, discutiamo le diverse strutture delle reti online. Le reti analizzate sono state identificate attraverso la Social Network Analysis.

Keywords: misinformation, disinformation, propaganda activity, Social Network Analysis

1. Background: misinformation and disinformation online

The development of the digital space relates to a new form of web-mediated communication, which can be defined according to the following main features. Web-communication can be thought of as a participative act and is

not part of a broadcast system (McLuhan, 1962) but is a networkcast system. In fact, a web content generates connections, denoted as "Affinity networks" (Rainie and Wellman, 2012; Castells, 2000), based on the sharing of a given content. In this network system, Web-communication yields temporary consensus areas based on alliances between users with respect to the shared contents. Moreover, Web-communication favors a mobilization of skills that generates new paths of social action and collective projects (Levy, 2002). In the digital space, content validity relies on activism and interest of digital users and every opinion "has citizenship rights" (Quattrociocchi and Vicini, 2016; Mocanu, 2015). In this framework, misinformation and disinformation processes share the previous characteristics. Furthermore, the accidental or deliberate propagation of false information is strictly linked to a "loss of disintermediation" (Jenkins, 2006). According to this theory, one of the most important effects of webmediated communication is the loss of traceability of official information sources. In fact, phenomena like Wikipedia, Social Media sites or Blog news produce the culture of unofficial knowledge, creating a virtuous circle of free sources, on the one hand, and a vicious circle of misleading information, on the other hand. Disinformation and misinformation processes can be both related to Fake news and Hate Speeches. "Fake news" or "Junk news" refers to web sources completely invented or simply distorted. In fact, in the digital space, anyone gain access at different information sources and can, also, create information content with low costs and high distribution potential. Furthermore, the fake new propagation process can develop into a viral system, dominated by the high sharing power of different recurring themes. Usually, Hate Speech phenomenon is linked to sharing and commenting fake news. Web 3.0 era is permeated by hatred, mainly directed to immigrants, political parties and homosexual people. Although hater activity concerns specific themes, it becomes a fundamental part in redefining the digital public sphere (Lévy, 2002).

2. Research Design and Methodology

The disinformation and misinformation online phenomena have become a propaganda activity to support offline organizations. In fact, in many cases online fake news and hate speeches are contained within a network system consisting of political and ideological advertising. In particular, this tendency gained attention during Trump's election campaign (Ott, 2017). The Computational Propaganda Research Project, promoted by Oxford University, aims at investigating «how tools like social media bots are used to manipulate public opinion by amplifying or repressing political content, disinformation, hate speech, and junk news». Woolley and Howard (2017),

mapping the computational propaganda in different countries, analyzed tens of millions of posts on seven different social media platforms, referring to elections, political crises and national security incidents. Each case study takes into account qualitative, quantitative and computational evidences collected between 2015 and 2017. In this framework, following a computational approach (Lazer et al., 2009), our research aims at identifying and comparing propaganda policy networks. For this purpose, we investigated the networks in which different political Facebook Like pages are embedded. More specifically, we selected the following Facebook Like pages related to political institutional information: "Ricostruiamo il centro destra" (*Centre-Right wing*), "Di Battista Alessandro" (*Five Star Movement*) e "Partito Democratico" (*Centre-Left wing*). Exploiting Social Network Analysis and focusing the attention on each of the chosen pages, we detected the online networks. The analyzed adjacency matrices were built considering as link the "likes". The analysis was implemented using the free and open source NodeXL extension of the Microsoft Excel spreadsheet (Hansen et al., 2011). For each network, we present the centrality measures, which describe how a particular vertex can be said to be in the "middle" of the network. In particular, betweenness centrality measures how often a given vertex lies on the shortest path between two other vertices. Vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. As pointed out by Hansen et al. (2011), these measures can be thought of as a kind of "bridge" score, a measure of how much removing a node would disrupt the connections between other vertices in the network. Closeness centrality captures the average distance between a vertices and every other vertex in the network. In NodeXL the inverse of the average distance is implemented so that higher closeness values indicate more central vertices. The Eigenvector Centrality network metric takes into consideration not only how many connections a vertex has (i.e., its degree), but also the degree of the vertices that it is connected to. A node with few connections could have a very high eigenvector centrality if those few connections were themselves very well connected. These centrality measures allowed to identify the most relevant nodes of each network. The identified Facebook Like Pages were classified in "official pages" and "junk pages" according to their contents. Junk information is strictly linked to the so-called post-truth politics, meaning a political culture in which truth is no longer significant or relevant and «objective facts are less influential in shaping public opinion than appeals to emotion and personal belief» (Oxford Dictionaries, 2016). In this context, the term junk information refers to fake news, conspiracy theories, hate speeches, misinformation and deliberately misleading disinformation. Accordingly,

Facebook Like pages containing posts, comments or images conveying this kind of information were classified as "junk pages". It is worth noticing how in the identified networks we did not retrieve hybrid forms, that is pages composed of both official and junk contents.

3. Preliminary results

The network built by considering the Facebook Like page "Ricostruiamo il centro destra" is depicted in Figure 1. This social media network, linked to a Centre-Right political view, is composed by 159 nodes, comprising both institutional and junk pages (e.g. "unitaliasenzacomunisti", "SapereEundovere"). Centrality values, provided in Table 1 for the six pages with higher levels of betweenneess centrality, highlight a connection between junk and institutional nodes; furthermore, the influence of junk pages in the network is very outstanding.

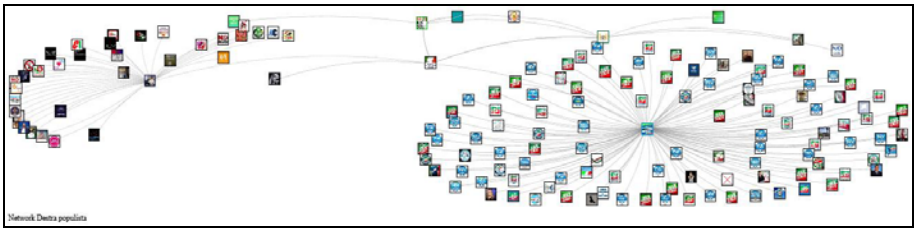


Figure 1. NodeXL social media network diagram of relationships derived from the Facebook Like page "Ricostruiamo il centro destra".

Table 1: Social media network of relationships derived from the Facebook Like page "Ricostruiamo il centro destra": centrality measures for the vertex pages with higher levels of betweenness

Vertex	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
ricostruiamocentrodestra	22644.000	0.004	0.009
unitaliasenzacomunisti	10986.000	0.003	0.009
SapereEundovere	10044.000	0.003	0.000
radionewsinformazionelibera	1087.000	0.002	0.000
italianinonsonorazzistisonostanchi diquestainvasione	777.000	0.002	0.000

A similar situation was detected for the Five Star Movement. This network, represented in Figure 2, is composed by 664 nodes comprising again both institutional and junk pages. In this case, the junk pages are specifically of the Five Star Movement and institutional pages are personal pages of political candidates. The Five Star Movement network shows three big cluster in which the central node (WIlM5s) is a junk page.

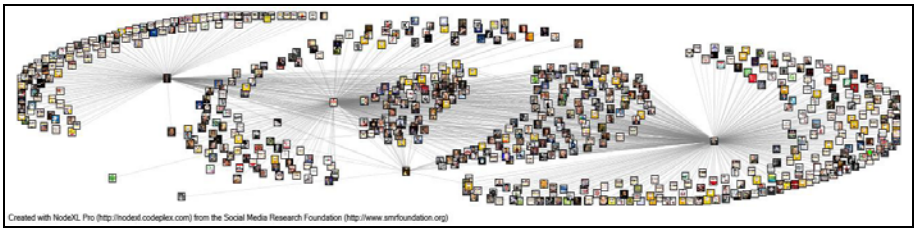


Figure 2. NodeXL social media network diagram of relationships derived from the Facebook Like page "Di Battista Alessandro".

Table 2: Social media network of relationships derived from the Facebook Like page "Di Battista Alessandro": centrality measures for the vertex pages with higher levels of betweenness.

Vertex	Betweenness Centrality	Closeness Centrality	EigenVector Centrality
MassimoEnricoBaroni	281353.000	0.001	0.032
WIlM5s	172430.333	0.001	0.024
sorial.giorgio	143457.000	0.001	0.013
dibattista.alessandro	3405.667	0.001	0.006
pierrecantagallo89	1324.000	0.001	0.001
perchevotarem5s	702.000	0.001	0.003

The social media network of relationships derived from the Facebook Like page "Partito Democratico" does not show the features found out for the previous networks. In fact, the network related to the Centre-Left political party is composed by only institutional propaganda pages.

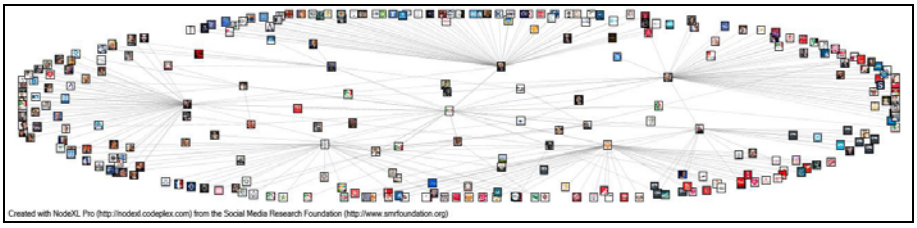


Figure 3. Centrality measures for the social media network of relationships derived from the Facebook Like page "Partito Democratico".

4. Community clusters

The mapping process of propaganda pages resulted into different structures of network. For the classification of these structures, we make use of the model elaborated by Smith et al. (2014) in order to define a taxonomy of social networks derived from conversations within Twitter. The authors defined six types of Networks: polarized crowds, tight crowds, community cluster, brand cluster, broadcast network and support network (see Figure 5).

Table 3: Social media network of relationships derived from the Facebook Like page "Partito Democratico": centrality measures for the vertex pages with a higher level of betweenness

Vertex	Betweenness Centrality	Closeness Centrality	EigenVector Centrality
partitodemocratico.it	46486.100	0.002	0.024
enricoletta.it	28853.657	0.002	0.047
scalfarotto	24167.162	0.002	0.038
giannipittella	23136.533	0.001	0.018
giovanidem	19798.000	0.001	0.011
palazzochigi.it	12633.519	0.001	0.009

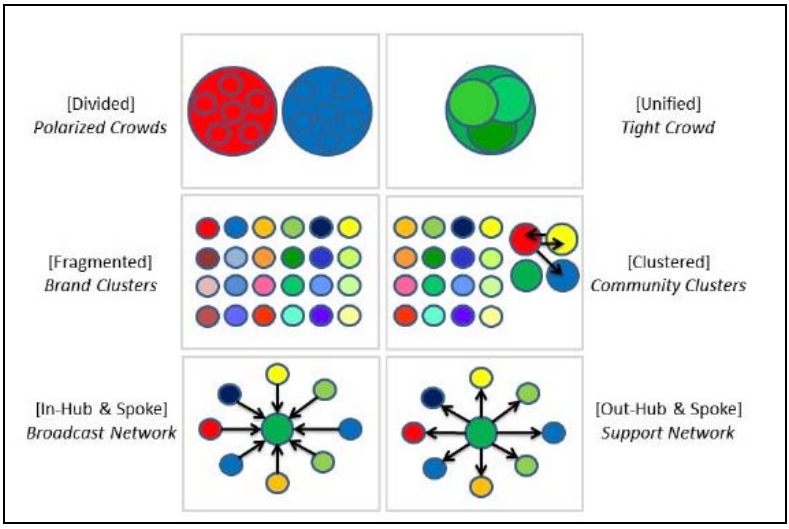


Figure 5: Diagrams of the differences in the six types of social media networks (Smith et al 2014).

In this framework, we can recognize how the Centre-Right wing social media network shows a conformation similar to a mixture of Polarized Crowd and Support Network. On the one hand, the Polarized Crowd model is characterized by two groups, polarized on specific opinions and sharing few connections. On the other hand, the Support Network model consists of a central node that sends information to the peripheral nodes. The Five Star Movement social network adheres more closely to Tight Crowd and Support network structures. The Tight Crowds is composed by highly connected nodes and specific shared themes. Finally, the Democratic Party network reflects the structures of a Community Cluster, which is organized in many cliques that share specific topics of conversation.

4. Conclusions and future works

In this preliminary phase of our research, we considered the network structures related to the online propaganda linked to different political areas. Our analysis allowed to highlight the differences in the networks and to cast the reconstructed networks into the taxonomy proposed by Smith et al. (2014). In addition, in two out of the three analyzed social networks we found out the presence of junk pages contributing to the disinformation and misinformation processes by spreading out fake news and indulging in hate speeches. The cluster structures of those two networks, leading to closed circle of highly polarized information, facilitates the diffusion process of misleading information. Based on these preliminary results, future works will focus on the textual analysis of posts and comments shared on the retrieved junk pages, in order to identify the main discussed topics. To this end, Text mining and machine learning techniques will be exploited.

References

Castells M. (2000). *The Rise of the Network Society*, Blackwell Publishers Oxford

Hansen D. L., Schneiderman B., Smith M. A. (2011). *Analyzing social media networks with NodeXL: insights from a connected world,* Morgan Kaufmann

Jenkins H., (2006). *Fans, Bloggers and Gamers: Exploring partecipatory culture*, New York University Press.

Lazer D., Pentland A., Adamic L., Aral S., Barabási A.L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., Van Alstyne M., (2009). Life in the network: the coming of computational social science, *Science* 323(5915): 721–723

Lévy P. (2002). *Cyberdémocratie. Essai de philosophie politique*, Paris: O. Jacob

McLuhan, M. (1962). *The Gutenberg Galaxy: the making of typographic man*, University of Toronto Press.

Mocanu, D.; Rossi L., Zhang Q., Karsai M., Quattrociocchi W. (2015) Collective attention in the age of (mis)information. *Computers In Human Behavior*, 51, 1198-1204

Ott B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement, *Critical Studies in Media Communication,* 34, (1): 59-68

Oxford Dictionaries (2016). *Word of the Year 2016 Is...*, https://en.oxforddictionaries.com/word-of-the-year/word-of-theyear-2016.

Quattrociocchi W., Vicini A. (2016). *Misinformation. Guida alla società dell'informazione e della credulità*, Franco Angeli.

Rainie L., Wellman B. (2012). *Networked: The New Social Operating System*, MIT Press.

Smith M., Raine L., Shneiderman B., Himelboim I. (2014). Mapping Twitter Topic Network: From polarized Crowds to community Cluster, *Pew*

Research Internet Project, February 20, http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/#

Woolley S. C, Howard P. N. (2017). Computational Propaganda Worldwide: Executive Summar,. *Working Paper 2017.11. Oxford, UK: Project on Computational Propaganda. comprop.oii.ox.ac.uk. 14 pp.*

Topic modeling of Twitter conversations

Eliana Sanandres[1], Camilo Madariaga[2], Raimundo Abello[3]
[1]Universidad del Norte – esanandres@uninorte.edu.co
[2]Universidad del Norte – cmadaria@uninorte.edu.co
[3]Universidad del Norte – rabello@uninorte.edu.co

Abstract
Topic modeling provides a useful method of finding symbolic representations of ongoing social events. It has received special attention from social researchers, particularly among cultural sociologists, in the last decade (DiMaggio et al., 2013; Sanandres and Otalora, 2015). During this time, Twitter has acted as the most common platform for people to share narratives about social events (Himelboim et al., 2013). This study proposes LDA (Latent Dirichlet Allocation) based topic modeling of Twitter conversations to determine what topics are shared on Twitter in relation to social events. The dataset for this study was constructed from public messages posted on Twitter related to the financial crisis of the National University of Colombia. Over an eight-week period, we downloaded all tweets that included the hashtag #crisisUNAL (UNAL is the Spanish acronym of the university) using the Twitter API interface. We analyzed over 45,000 tweets published between 2011 and 2015 using the R package topicmodels to fit the LDA Model in five steps: first, we transformed the tweets into a corpus, which we exported into a document-term matrix; the terms were stemmed and the stop words, punctuation marks, numbers, and terms shorter than three letters were removed. Second, we used the mean term frequency-inverse document frequency (tf-idf) over documents containing this term to select the vocabulary. We only included terms with a tf-idf value of at least 0.1, which is a bit less than the median, to ensure that the most frequent terms were omitted. Third, we defined the number of topics k by estimating the log-likelihood of the model for each topic number starting with 1 though to 300 topics and selected $k = 12$ because it had the highest log-likelihood value (LL = -198000). Fourth, we run the LDA Model for $k = 12$ topics. Fifth, we labeled the $k = 12$ topics previously identified by choosing the top N terms ranked based on the probability of that topic. This article illustrates the strength of topic modeling for analyzing large text corpora and provides a way to study the narratives that people share on Twitter.

Keywords: Topic modeling, LDA, Twitter.

1. Introduction

This article presents a way to analyze large amounts of textual data from Twitter conversations in an efficient and effective way. Specifically, we explain how to capture the narratives that people share on Twitter about social events, reduce their complexity, and provide plausible explanations. This is a research concern that has received special attention among social researchers (Kovanović et al., 2015; Yann et al., 2011; Newman and Block, 2006; Griffiths and Steyvers, 2004), particularly among cultural sociologists, who face the methodological challenge of working qualitatively with large amounts of data (Sanandres and Otalora, 2015; Eyerman et al., 2011; Alexander, 2004). In this paper we propose an LDA (Latent Dirichlet Allocation) based topic model to address this challenge. Topic modeling is a useful approach because the set of terms found within topics index discursive environments or frames that define patterns of association between a focal issue and other constructs (DiMaggio et al., 2013). These patterns of association are to be interpreted as symbolic representations of ongoing social events, which represent claims about the shape of social reality, its causes, and the responsibility for action such causes imply (Alexander, 2004). We applied an LDA-based model to Twitter conversations about the financial crisis of the National University of Colombia to examine how the debate over this crisis was framed on Twitter, from 2011 when it emerged, until 2015. We analyzed over 45,000 tweets and illustrated the strength of topic modeling for the analysis of large text corpora as a way to study narratives shared on Twitter.

2. Background: The financial crisis of the National University of Colombia

Over the last decade, Colombian academics and representatives of the government have recognized that the limitations of their budgets are the major limitation in the response of public universities to the increasing demands of society. To face this problem, the government proposed to reform the entire system of higher education (Ministry of National Education, 2010). The intention was to find new sources of money for higher education, enable more people to attend college, encourage transparency and good governance in the education sector, and improve the quality of higher education. One of the most controversial proposed changes was the opening of the education sector to private investment by for-profit companies (El Espectador, 2011). This was immediately rejected by public universities, who claimed that the proposed reform would lead to a full-scale privatization of the system of higher education (Semana, 2011).

At the public National University of Colombia, the largest higher education

institution in Colombia, some students and professors claimed that the reform offered no clear solution to the financial crisis of the university. They explained that the university had been using a funding model with its sources of support mixed between the state and external resources, claiming that since 2004 this model had borne dwindling state support and ever-increasing costs to be covered by external resources. They showed that government transfers had decreased from 70% in 2004 to 64% in 2013, while the external resources produced from activities such as tuition fees, non-formal education courses, and academic extension services, among others, had increased from 30% to 36% in the same period (National University of Colombia, 2014). This statement reopened the debate on the financial crisis of the National University of Colombia and became a Twitter trending topic with the hashtag #CrisisUnal (UNAL is the Spanish acronym for the name of the university).

3. The financial crisis of the National University of Colombia on Twitter

Here, we investigate how the financial crisis in the National University of Colombia was framed on Twitter. It may be asked why we should care about Twitter conversations on this topic? However, it should be considered that Twitter conversations can offer clues to what the university is thinking and doing about the crisis. A central advantage of using Twitter for analyses is that it covers topics in real time, producing a large amount of data that can be used to look at people's perceptions and narratives of particular events. Twitter also provides a practical way to examine collective experience related to a topical event, to study behaviors and attitudes where social desirability bias may occur in official surveys, and to collect large amounts of data with a limited budget (Himelboim et al., 2013). Twitter conversations also illustrate the views of the reading public and show dominant viewpoints, which emerge quickly and are difficult to change (Xiong et Liu, 2014).

We collected every tweet published between 2011 and 2015 that contained any reference to the financial crisis in the National University of Colombia with the hashtag #CrisisUNAL. We chose this period to track Twitter conversations around this topic, from the time it became a Twitter trend in 2011 through 2015 (the last year in which we collected data). Our collection formed a corpus of over 45,000 tweets. In the next section we describe how we used topic modeling.

4. Method

Topic modeling is a machine-learning method used to discover hidden thematic structures in large collections of documents. In this work we used LDA, a widely used method in topic modeling (Jelodar et al., 2017; Fligstein

et al., 2014), which assumes that there is a set of topics to be found in a collection of documents. The intuition behind LDA is that documents exhibit multiple topics. A topic is formally defined as a distribution of words over a fixed vocabulary (Blei, 2012). For LDA, topics must be specified before any data are generated. For each document in the collection, this method generates the words in a two-stage process. During the first stage, it randomly chooses a distribution over topics (step 1). In the second stage, for each word in the document, it randomly chooses a topic from the distribution over topics in step 1 (step 2a), and a word from the corresponding distribution over the vocabulary (step 2b). At the end, each document exhibits topics in different proportions (step 1) and each word in each document is drawn from one of the topics (step 2b), where the selected topic is chosen from the per-document distribution over topics (step 2a) (Blei, 2012). To run the LDA model, we followed five steps. First, we transformed the tweets into a corpus and exported this corpus to a document-term matrix; the terms were stemmed and the stop words, punctuation, numbers and terms shorter than three letters were removed. Second, we used the mean term frequency-inverse document frequency (tf-idf) to select the vocabulary. We only included terms with a tf-idf value of at least 0.1, which is a bit less than the median, to make sure that the most frequent terms were omitted. Third, we defined the number of topics k by estimating the log-likelihood of the model for each topic number, from 1 to 300 topics; we selected $k = 12$ as having the highest log-likelihood value (LL = -198000). Fourth, we run the LDA model for $k = 12$ topics. Fifth, we labeled the k = 12 topics previously identified by choosing the top N terms, ranked according to the probability of that topic. For this we used the R package topicmodels.

5. Results

Table 1 displays the 12-topic solution and lists the 10 highest-ranking terms for each topic. We call attention to four sets of topics: six topics concerned with social protest (dark shading), three topics on educational reform (medium shading), two topics calling for investment (light shading), and one topic emphasizing the role of the National University of Colombia in the Colombian peace process (no shading). To more easily interpret the topics, after reviewing the list of terms we examined those tweets that exhibited each topic with the highest probability.

5.1 Protest topics

Protest topics are the focus of the Twitter conversations on the financial crisis in the National University of Colombia. Topic 1 covers the protests of the education workers. The most highly ranked terms were *sintraunal* (the labor

union covering all workers at public universities), *protest, strike, campus, riot, gas, blocked*, and *wall.* The tweets in which this topic was strongly represented locate protests in national and international contexts with terms like *nation* and *clacso (Latin American Council of Social Sciences*), indicating that the protests were a matter of concern in Colombia and in Latin America. Topic 3 also refers to the protests of the education workers. Some of the top words are *sintraunal*, *gases, wall*, and *block.* This topic frequently exhibits tweets that show negative aspects of protests, such as *confrontation*, *death*, and *bombs*.

Table 1: 12-topic solution

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
sintraunal	agricultural	sintraunal	agrarian	solidarity	no to the
protest	strike	gases	protest	no to the	reform
strike	graffiti	wall	movement	reform	universities
campus	hate	block	mobilization	justice	listen
riot	block	undefined	participation	march	sciences
gas	bombs	bombs	people	respect	confrontation
blocked	terrorists	hood	bombs	charge	media
wall	crash	criticism	poor	help	classrooms
nation	delinquents	death	assembly	block	abandoned
clacso	guevara	confrontation	disturbance	upedagogica	mobilization
				studying	block
Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
defend	no to the	Stamp	intimidation	peace	revolutionary
university	reform	demand	blocked	process	victory
improvement	propose	support	abandoned	mobilization	popular
campus	threat	public	public	research	campus
crisis	oblivion	university	eviction	studying	strike
infrastructure	save	strike	strike	participation	eviction
cement	closed	resources	che	talks	denounce
hospital	blocked	deserve	graffiti	intellectuals	deserve
architecture	abnormality	financial	protest	solidarity	abandonment
sociology	upedagogica	pride	worker	civil	took
	uncertainty				

Topics 2 and 4 refer to the agricultural sector protests. While Topic 4 is related to the *mobilization* of *people* to take part in these *protests,* Topic 2 emphasizes the participation of *terrorists* and *delinquents* in *agricultural strikes.* In this context, social protest is associated with the Argentine Marxist revolutionary Ernesto Che Guevara. Che is also mentioned in Topic 10, which deals with the protests of the working class and the intimidation of protesters. The most highly ranked terms in this topic are *intimidation*, *blocked*, *abandoned*, *public*, *eviction*, *worker*, *strike*, *che*, *graffiti*, and *protest.* Finally, Topic 12 covers the revolutionary cause of social protest and includes the words *revolutionary*, *victory*, *popular*, *campus*, and *strike*.

5.2 Anti-reform topics

Five topics deal with the reforms of higher education proposed by the government. According to the terms included in Topic 5, public universities reject this reform and called for justice and respect; terms in this topic include *solidarity*, *no to the reform*, *justice*, *march*, and *respect*; tweets representing this topic show strong solidarity among public universities, specially from the Universidad Pedagógica (*upedagogica*). Topic 8 is also related to the rejection of the planned educational reform to save public education; this includes terms like *no to the reform*, *propose*, *threat*, *oblivion*, and *save*; Universidad Pedagógica (*upedagogica*) is mentioned as well. In the same way, Topic 6 indicates that public universities reject the reform of higher education, mobilize to denounce the government's abandonment, and demand to be listened to; some of the words in this topic are: *no to the reform*, *universities*, *listen*, *sciences*, *confrontation*, *media*, *classrooms*, *abandoned*, *mobilization*, and *block.*

5.3 Investment topics

Topics 7 and 9 cover demands for investment to face the crisis. Topic 7 calls for infrastructure investment. Many tweets in which this topic is prominent focus on the *infrastructure crisis* of the *campus* buildings, in particular the *sociology* and *architecture* buildings and the university's *hospital*. The top terms in this topic include *defend*, *university*, *improvement*, *campus*, *crisis*, *infrastructure*, *cement*, *hospital*, *architecture*, and *sociology.* Topic 9 plays a similar role in investment demands focusing on the pro-National University of Colombia *stamp*, created to acquire *financial resources* to improve the university facilities. Some tweets containing this topic highlight the role of the University as a national *pride*. The top ranked terms include *stamp*, *demand*, *support*, *public*, *university*, *resources*, *financial*, *strike*, *deserve*, and *pride.*

5.4 Peace topic

Topic 12 represents the integration of the crisis in the National University of Colombia into a broader frame of national concern associated with the Colombian peace process. The top-ranked terms are *peace*, *process*, *mobilization*, *research*, *studying*, *participation*, *talks*, *intellectuals*, *solidarity*, and *civil.* Tweets in which this topic was strongly represented are related to the role of the university as facilitator in peace talks among the government, rebel groups involved in the Colombia's internal armed conflict (which began in the mid-1960s and is currently in negotiation, in a process known as the Colombian peace process), intellectuals, and representatives of civil society.

6. Conclusions

Producing an interpretable way to study Twitter conversations efficiently and effectively is only the beginning. The solution of this issue presents meaningful categories to address the analytic question that motivated the study: *how was the financial crisis in the National University of Colombia framed on Twitter?* The 12-topic solution showed that it was framed through four categories: protest, anti-reform, investment, and peace.

Each topic constitutes a frame, in that it includes terms calling attention to particular ways in which the crisis under study may arouse controversy: protest frames emphasize public displays, demonstrations and the civil disobedience of the working class; anti-reform frames refer to the rejection of the reform of higher education by public universities; investment frames focus on investment demands to face the crisis; and the peace frame draws attention to the role the National University of Colombia played in acting as a facilitator in the Colombian peace process. Each of these frames represents a discursive environment for the financial crisis, which broadcasts not just the structural characteristics of the crisis (investment demands and education reform), but also symbolic representations of ongoing social events (workers protests and peace process), which can be seen as claims about ongoing social processes and demands of reparation.

These results provide substantive insight into Twitter conversations about the financial crisis in the National University of Colombia. Using LDA to discover topics allowed us to locate two narratives: one focused on the structural characteristics of the crisis and the other concerned with symbolic representations of ongoing social events surrounding that crisis. For cultural sociologists, this is only the beginning of the analysis. A topic model allows a starting point to be found, which in this case is the structure of Twitter data. Used properly, with appropriate validation, topic models are valuable complements to other interpretive approaches, offering new ways to extract topics and make sense of online data.

References

Alexander, J. (2004). Toward a theory of cultural trauma. In Alexander, J., Eyerman, R., Giesen, B., Smelser, N. and Sztompka, P. *Cultural trauma and collective identity*. Univ of California Press.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6): 570–606.

El Espectador (2011). Universidades con ánimo de lucro, apuesta del gobierno. March 10.

Eyerman, R., Alexander, J. C., and Breese, E. B. (2011). *Narrating trauma: on the impact of collective suffering*. Routledge.

Fligstein, N., Brundage, J. S., and Schultz, M. (2014). Why the Federal Reserve failed to see the financial crisis of 2008: The role of "Macroeconomics" as a sense making and cultural frame. *IRLE Working Paper* No. 111–14.

Griffiths, T., and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, pp. 5228–5235.

Himelboim, I., McCreery, S., and Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2): 154–174.

Jelodar, H., Wang, Y., Yuan, C., and Feng, X. (2017). Latent Dirichlet allocation (LDA) and Topic modeling: Models, applications, a survey. *arXiv preprint arXiv:1711.04305*.

Kovanović, V., Joksimović, S., Gašević, D., Siemens, G., and Hatala, M. (2015). What public media reveals about MOOCs: A systematic analysis of news reports. *British Journal of Educational Technology*, 46(3): 510–527.

Ministry of National Education (2010). Proposal for the education reform in Colombia. April 12.

National University of Colombia (2014). *Estadísticas e indicadores de la Universidad Nacional de Colombia*. 19. ISSN 2357-5646.

Newman, D., and Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the Association for Information Science and Technology*, 57(6): 753–767.

Sanandres, E., and Otálora, J. (2015). Application of topic modeling for Trauma Studies: The case of Chevron in Ecuador. *Investigación & Desarrollo*, 23(2): 228–255.

Semana (2011). Reforma a la Ley 30: por qué sí, por qué no. April 1.

Yang, T., Torget, A., and Mihalcea, R. (2011). Topic modeling on historical newspapers. In K. Zervanou & P. Lendvai (Eds.), *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 96–104.

What volunteers do? A textual analysis of voluntary activities in the Italian context

Francesco Santelli, Giancarlo Ragozini, Marco Musella
University of Naples Federico II
francescosantelli@unina.it marcomusella@unina.it

Abstract
The complex phenomena of volunteering was mainly analyzed in economic literature with respect to its "economic value added", i.e the capability of this kind of activities to increase the level of productivity of some specific gods or services. In this paper, the point of view switches and voluntary organizations are analyzed as place of job market innovation, where new jobs arise and where people acquire new skills. Thus, volunteering can be thought as "social innovation" factor. In order to analyze the contents of voluntary works we use data coming from Istat survey "Multiscopo, Aspetti della vita quotidiana" (Multi-purposes survey, daily life aspects), for the year 2013. In our textual analysis, we use information included in the open answers given by people about the description of the tasks performed individually as volunteer. After stemming, lemmatization, and cleaning, data have been analyzed by means of Community Detection based on Semantic Network Analysis in order to discover patterns of jobs and through Correspondence Analysis on Generalized Aggregated Lexical Tables (CA-GALT) in order to discover profiles of volunteers. In particular, we look for differences given by gender, age, educational level, region of residence and type of voluntary association.
Keywords: Text Mining, Volunteers, Lexical Correspondence Analysis, Semantic Network Analysis

1. Introduction

Volunteer work differs from the traditional forms of work for several features. Nevertheless, most of the authors approaching the volunteering phenomenon are interested mainly in the economic value that this sector is able to add to the labour market (Ironmonger, 2000; Salamon et al., 2011) considering it like a special case of job in the economic theory framework. From this point of view, volunteering is assumed to be a peculiar sector of the production with a considerable number of divergent rules and dynamics compared to the standard work patterns, but still able to provide goods and services to the community like all the other sectors. It will lead, of course, to increase the overall economic value of the society.

In this work, the focus will be instead from a different perspective: volunteering will be considered as a laboratory of social innovation embedded in the labour market. The main concept behind it is that volunteering is based on different guidelines and different principles (Zamagni, 2005); therefore, it could develop new professional profiles and modify pre-existent ones. Text Mining approach will be performed on open-end questions given by volunteers, assuming that their self-concepts is a consistent proxy of volunteering world. The empirical statistical analysis will make use of two tools chosen for their capability to profile both groups of words and cluster of volunteers. The latter, in the Italian context, will be analyzed in parallel with the traditional categories applicable to the classic labor theory. It will be shown that most of the determinants of the segmentation of the professions (Colombo, 2003), such as gender, age or geographic area of origin, can be adopted as well in this framework.

2. Data and statistical approach

Data are taken from the Istat Survey of 2013 "Multiscopo, Aspetti della vita quotidiana" (Multi-purposes survey, daily life aspects) (Istat, 2013). It is a large annual sample survey that covers the resident population in private households, by interviewing a sample of about 20000 households and about 50000 people with P.A.P.I. technique. The main dimensions questionnaires concern education, work, family and social life, spare time, political and social participation, health, life style and access to the services.

From the whole sample, we selected about 5000 persons that declared to be involved in volunteering and that answered to open-end questions about their voluntary activities and if they carried out it within an organization or by themselves. The main core of the statistical text mining procedure will be focused on these brief descriptions of their own volunteering jobs. We analyzed the descriptions along with the socio-demographic variables available: gender, age, geographic macro-area and educational level. Given the definition of volunteering (Istat, 2013; Wilson, 2000), several descriptions were erased from the database as they do not belong to voluntary activities (e.g., people donating blood to AVIS organization, or people that provides help to family members).

Therefore, after this preliminary procedure in order to delete inappropriate or missing answers, the valid number of volunteers are 4254 from the original 5000. Before going through the analysis, we perform a preliminary transformation of the original lexical data by removing punctuation and stop-words, and by stemming the words, i.e. deleting all the derivational and inflectional suffixes (Lovins, 1968; Willet, 2006). Therefore, all the words that evolved from the same root will be considered to be the same after the

stemming. For this task we use the Porter Stemming Algorithm using software R implemented in the package tm (Meyer et al., 2008).
After the preliminary analysis, in order to discover groups of activities that can be described as jobs we apply a Semantic Network Analysis (van Atteveldt, 2008; Drieger, 2013), and in order to profile of voluntary jobs with respect to socio-demographic dimensions we use Correspondence Analysis on Generalized Aggregated Lexical Tables (CA-GALT) (Kostov et al., 2015). The former is an extension of Social Network Analysis that treats text as graph structure: each word is defined as a node, and the ties between words are undirected links weighted by the count of co-occurrences (how many times do these words appear together in the same answers). Groups of terms corresponding to semantic clusters can be found through community detection algorithms (Fortunato, 2010). We use the Fast Greedy method that is suited to deal with undirected and weighted edges (Clauset et al., 2004). On the other hand, the CA-GALT method allows us to jointly analyze in a multiple correspondence framework both the lexical table and socio-demographic profiles, combining the document-term matrix and the matrix containing the individual characteristics.

3. Main findings of the analysis

After the preliminary transformations, the overall corpus shows a high degree of heterogeneity with 1649 different words, and a high level of sparsity, close to 100% due to the large number of documents and their shortness. The term frequency distribution has a median equal to 2, and a $p_{0.75}$ percentile equal to 4. Given the sparsity, we focus the analysis on the most frequent words that profile and describe voluntary activities, taking into account only words that are above the $p_{0.90}$ percentile (frequency equal to 11), and ending up in a vocabulary consisting of 175 words. The most used of them are *organizz* (to organize, or organization) that appears 296 times, *assistent* (assistent) with 225 occurrencies, *attiv* (activity) that occurs 215 times, then *assoc* (association), *aiut* (to help) and *volontar* (volunteer and derived words). Those terms can be considered pretty generic, and could be related to several aspects inside the volunteers' community, without showing additional informative power to profile volunteers. They are followed by terms describing specific field of intervention: sport, *fond* (fund), event, *bambin* (child/children), *anzian* (senior/old). Further, some of them are expressing just one semantic meaning, and can be considered bi-grams (Collins, 1996): *croce rossa* (red cross), *croce verde* (green cross), *croce bianca* (white cross), *protezione civile* (civil protection/defense), *vigili fuoco* (firefighters), *capo scout* (scoutmaster). We merge them in the following.
Applying the Semantic Network and the community detection algorithm to

these data, we found 7 groups/communities. In Fig. 1 we plot the semantic network along with the communities, in which words are colored according to the community. It is possible to identify a set of "jobs" related to the typical charity organizations, mainly in a religious context: the care of old people and hospitalized people -*ospedal, malat, assistenz, ascolt, accud, cur, sostegn-* (orange), the education and animation od disadvantaged children, mainly in religious organizations -*insegn, parrocc, scuol, orator, cateches, anim-* (purple), the food and cloth drive and its distribution to the poor -*cibo, vestiar, caritas, raccolt, aliment, mens, pover-* (green). Another large group is related to the executives and officers of organizations and to the cultural events organizers -*organizz, event, cultural, membr, consigl, dirigent, reunion-* (blue). Related to this large group we found the musicians (black) characterized by *suon, band, musical*. Finally, the last important area of the network is associated to the organized volunteers on the territory -*vigilefuoc, protezionecivil, territor, croceross, soccors, ambul-* (red). The coaches are mixed with this group -*squadr, allen, calc, pallavol-* (brown). All these activities are mainly done in nonreligious organizations and are not directly related to charity aims.

Analyzing categories and lexical CA in (fig:2) is possible to profile individuals according to their demographic status. In this context is not performed a real clustering procedure, but as in classical Correspondence Analysis the two spaces, units and variables, are linked taking into account that words close to a specific categories are more likely to occur for people belonging to the given category. It is clear that there is a gender gap: men are related to sport activities, they play music in band, they are driver (mainly ambulance) and they are involved in administration tasks. Women are more involved in providing services to individuals (taking care of children and old people), also carrying out food and cloth drive for the poor. Geographic differences come up as well: volunteers from North-Est and North-Ovest describe their activities as *manutenzion, dirigent, addett, consigl*, showing a higher organization level. South and Islands are more related to a female style of volunteering, with a predisposition for religious organization and mainly aimed to assistance. Educational level and age have an impact: lowest level of education, crossed with age information, profile a group of old and less educated volunteers involved in religious volunteering. Highest educated people carry out mainly administrative tasks. The central group of age (35-64) shows, on the other hand, an average profile close to the origin of axis, as well as people from Center Italy.

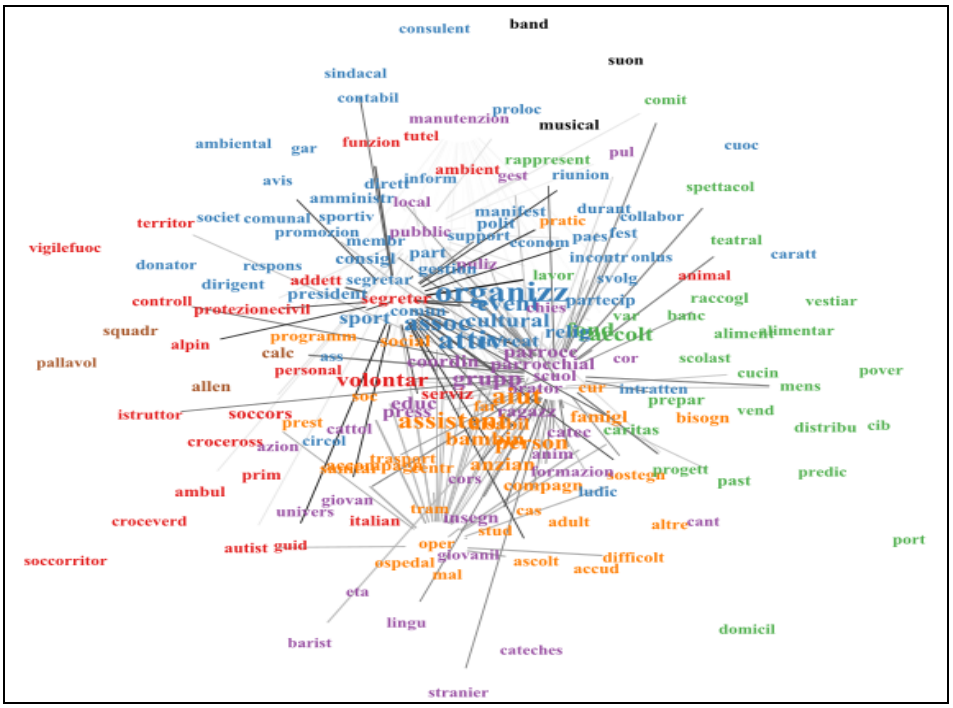
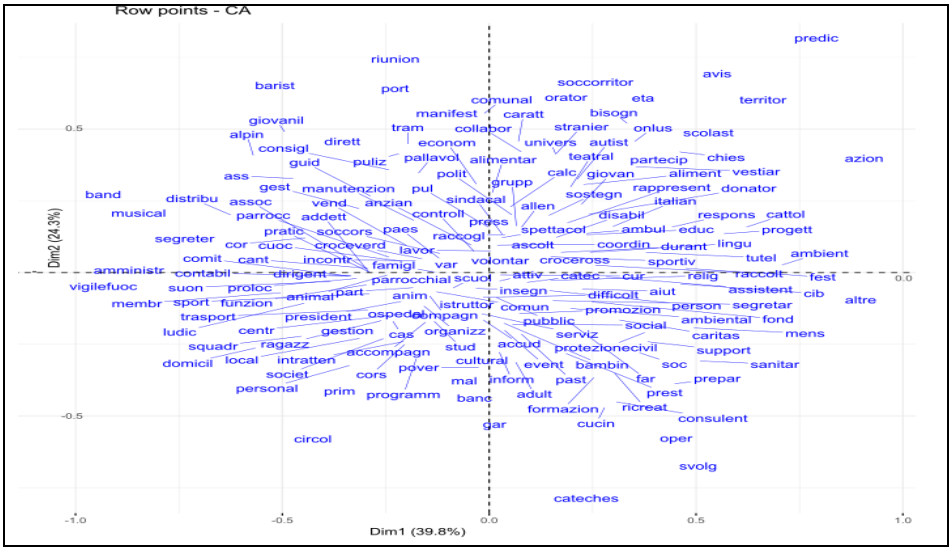


Figure 1: Semantic network: different colors for different communities identified by FastGreedy algorithm. Size of words and width of the edges are proportional to the weights

4. Discussion and conclusion

As introduced in the first section, the aim of this work is to present a general perspective about volunteering work in Italy under the assumption that is possible to study it in an analogue way in which labour market is studied in classic economic literature. Some authors already gave example how it follows also the rule of supply and demand under given condition (i.e. Wolff et. al, 1993) and also volunteering companies make use of marketing strategies similarly to business companies (Dolnicar et Randle, 2007). The two different statistical tools presented in previous section give to the empirical analysis different hints, and are somehow complementary. Communities in Semantic Network of (fig:1) are based on the connection level between words, without taking into account other previously known characteristics of individuals. Communities thus discovered are groups of words that define several activities and so clusters of jobs in some specific fields. In the second analysis, both spaces build in Ca-Galt, individuals and categories, stress out how segmentation is clearly present in volunteering as in labour market, and words used (and so activities done) change for gender, education, age and macro-area, in an equivalent way as for standard jobs. It

gives so an overview about the relationships between words (as description of activities) and categories (socio-demographic variables). Summing up, both analysis highlight how volunteering is complex and heterogeneous; it shows that people involved are in some cases highly skilled, often using some of the competencies trained in their life. Generally, they are able to describe their activities in a thorough way, explaining openly the aim of their voluntary jobs. The Text Mining analysis presented in this work could lead to figure out some needs of the population that are not adequately satisfied, given the assumption that volunteers spend their time and use their skills to give something to individuals that strongly ask for demands, in a framework similar to supply and demand mechanism. Furthermore, to have a more exhaustive overview for future policies to undertake, next step could be likely to go on the other side; another survey should be done asking people why do they *ask help* to volunteers. It will lead to better understand the real needs of individuals that are not fully satisfied of what they get in terms of assistance, especially from official institutions welfare.



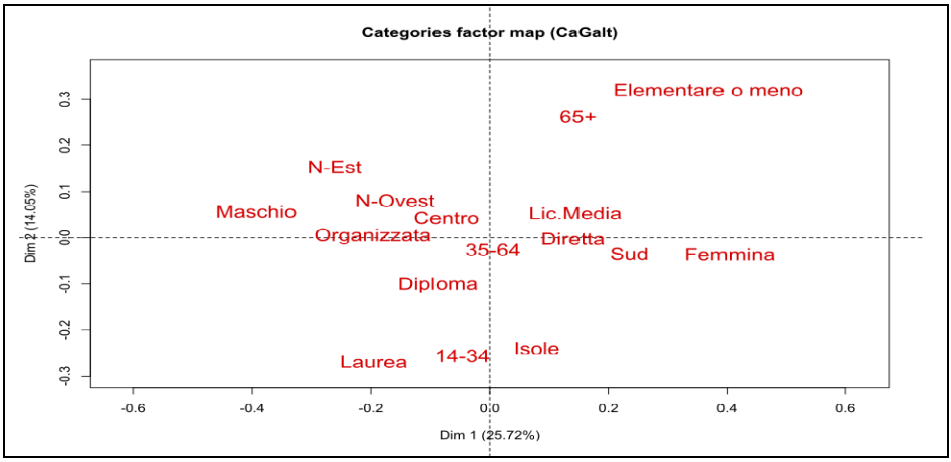


Figure 3: Ca-Galt for both terms (blue) and categories (red). Overlapping both factor maps is possible to profile cluster of individuals.

References

Amati, F., Musella, M. and Santoro, M. (2015). *Per una teoria economica del volontariato.* (Vol. 1). G. Giappichelli Editore, Torino

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6), 066111.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In Proceedings of the *34th annual meeting on Association for Computational Linguistics*, 184-191, Association for Computational Linguistics

Colombo, A. (2003). *Razza, genere, classe. Le tre dimensioni del lavoro domestico in Italia*. Polis, 17(2), 317--344,

Dolnicar, S. and Randle, M. (2007). The international volunteering market: Market segments and competitive relations. International Journal of Nonprofit and Voluntary Sector Marketing, 12(4), 350-370.

Drieger, P. (2013) Semantic network analysis as a method for visual text analytics, *Procedia-social and behavioral sciences*, 79, 4 – 17

Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.

Indagine Istat Multiscopo sulle famiglie: aspetti della vita quotidiana, (2013), Retrieved from http://www.istat.it/it/archivio/91926

Ironmonger, D. (2000). Measuring volunteering in economic terms. Volunteers and Volunteering, *The Federation Press*, Sydney, 56--72

Kostov, B., Bécue Bertaut, M. and Husson, F. (2015). Correspondence analysis on generalised aggregated lexical tables (CA-GALT) in the FactoMineR package, *R Journal*, 7(1), 109 -- 117,

Lovins, J. (1968). Development of a stemming algorithm. Mech. Translat. *Comp. Linguistics*, 11(1-2), 22--31, (1968)

Meyer, D., Hornik, K., and Feinerer, I. (2008). Text mining infrastructure in R. Journal of statistical software, 25(5), 1-54.

Salamon, L., Sokolowski and S., Haddock, M. (2011). Measuring the economic value of volunteer work globally: Concepts, estimates, and a roadmap to the future, *Annals of Public and Cooperative Economics*, 82(3), 217--252, (2011)

van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content*, BookSurge Publishers, Charleston SC

Willett, P. (2006). *The Porter stemming algorithm: then and now.* Program, Vol. 40 Issue: 3, 219--223, doi: https://doi.org/10.1108/00330330610681295

Wilson. J. (2000). Volunteering, *Annual review of sociology*, 26(1), 215—240

Wolff, N., Weisbrod, B. A., and Bird, E. J. (1993). The supply of volunteer labor: The case of hospitals. Nonprofit Management and Leadership, 4(1), 23-45.

Zamagni, S. (2005). Gratuità e agire economico: il senso del volontariato. *In Working Paper presented at Aiccon meeting*, Bologna

A longitudinal textual analysis of abstract presented at Italian Association for Vocational guidance and Career Counseling' Conferences from 2002 to 2017

S. Santilli[1]., S. Sbalchiero[2], L. Nota[3], S. Soresi[4]
[1] University of Padova – sara.santilli@unipd.it
[2] University of Padova – stefano.sbalchiero@unipd.it
[3] University of Padova – laura.nota@unipd.it
[4] University of Padova – salvatore.soresi@unipd.it

Abstract
This new century is characterized by phenomena such as globalization, internationalization, and rapid technological advances, that influence people life and the ways in which they seek and do their jobs. Changing the shape of organizations changes the shape of careers. To better account for the complexities of work due to the least socio economic crisis, the Life Design paradigm, a new paradigm for career theory in the 21st century (Savickas et al., 2009) has been recently developed an it represent the third wave of career theory and practice. The first wave emerged as the psychology of *occupations* in the first half of the 20th century to match people to jobs. The second wave comprised the psychology of *careers* ascending at mid-20th century to manage worker and other life roles across the lifespan.
The main aims of the present study was illustrate the changes in theory, technique e measure emerged in the Italian vocational guidance and career counseling psychology by the analysis of the abstract presented at Italian Association for Vocational guidance and Career Counseling' Conferences. The corpus was composed of 1,250 abstracts that have been collected from 2002 to 2017. In order to compare and contrast the main semantic areas over time, a topic analysis by means of Reinert's method (1983) was conducted (IRaMuTeQ and R software) to detect the clusters of words that characterized the different orientations over time. The results show that career counseling theories and technique evolved during the time to better assist workers in adapting to fluid societies and flexible organization and to better help clients design their lives in 21st century.

Keywords: longitudinal textual analysis, career counseling, vocational psychology

1. Introduction

In Western countries the economic recession that characterized the years

2008–2009 lead to a dramatic loss of jobs throughout the Union's private sector. Furthermore fast moving global economy and phenomena such as globalization, internationalization, and rapid technological advances, influence people's lives and the ways in which they seek and do their jobs. The world of work is in general much less clearly defined or predictable, and employees face greater challenges in coping with work transitions (Savickas et al., 2009). Therefore, life in a 21st-century requires new models and methods to deal with the new issues such as uncertainty, inequalities, poverty, immigration precariousness in the labor market, and with the worrying consequences also on individual and relational wellbeing. For these reasons existing traditional career guidance assumptions have been swept away, together with other certainties, by the sudden changes that have taken place in the world of work and in the economic field. To better account for the complexities of work, the Life Design paradigm, a new paradigm for career theory and intervention in the 21st century (Savickas et al., 2009) has been developed. The psychology of life design advances a contextualize epistemology emphasizing human diversity, uniqueness, and purposiveness in work and career to make a life of personal meaning and social consequence. Rather than matching self to occupation, it reflects a third wave of career theory and practice. The first wave emerged as the psychology of *occupations* in the first half of the 20th century to match people to jobs. The second wave comprised the psychology of *careers* ascending at mid-20th century to manage worker and other life roles across the lifespan. The third wave arose as the psychology of *life design* to make meaning through work and relationships.

The main aims of the present study was illustrate the longitudinal changes that emerge in the Italian context regarding the models and the theoretical paradigms that drive vocational guidance and career counseling by the analysis of the abstract presented at the Italian Association for Vocational guidance and Career Counseling 'Conferences. Specifically, we analyzed differences between the abstract presented before the economic recession (from 2002 to 2008) and during/after the economic recession (form 2009 to 2017) in the topics related to research, theories, and practice. The corpus was composed of 1,250 abstracts that have been collected from 2002 to 2017.

2. Corpus and method

All the abstracts have been collected by the Italian Association for Vocational guidance and Career Counseling - SIO. SIO represents at the national and international level a focal center in which the main scholars and practioners converge, gather, share and compare the theories and practices in terms of vocational guidance and career counseling. The Abstracts from the first

SIO's Conference (2002) to the latest one (2017) were collected. No abstract were collected during the year 2003, 2007, 2016, and 2014, becouse SIO has not organize national conferences. The corpus is composed of 1,250 abstracts. The corpus was pre-processed by means of IRaMuTeQ and R software (Ratinaud 2009; Sbalchiero e Santilli, 2017). The corpus was normalized replacing uppercase with lowercase letters, and punctuation, numbers and stop words have been removed because are not significant to analyse the content of abstract. The pre-processing steps were useful to reduce the redundancy and to provide homogeneity among forms. The lexicometric measures (Tab.1) indicate that it is plausible to apply statistical analysis of textual data to the corpus (Lebart et al., 1998). The corpus is composed of 20,932 word-type and 462,034 word-tokens.

Tab. 1. Lexicometric Characteristics of the corpus

Number of texts	1250
(V) Word-type	20932
(N) Word-tokens	462034
(V1) Hapax	8902
(V/N)*100 = Type/Token Ratio	4,53
(V1/V)*100 = Percentage of hapax	42,53

Using the Reinert method (Reinert, 1983), we extracted a series of 'lexical worlds'. The texts was divided into elementary content units of similar length, then, the algorithm provides reports on 'words x units' matrix. The classification of units consents to identify and extract only parts of texts relating to the same topic, so for each cluster the list of the most significant words calculated using the chi square measurement, are identified (Reinert, 1993; Sbalchiero and Tuzzi 2016; Sbalchiero e Santili, 2017).

3. Results

The analysis conducted by means of Reinert's method detected five different lexical worlds, as the dendrogram shows (Fig. 1). The methods identify the lexical worlds quite well because 98,42% of the abstracts have been classified and the words in the same sematic area are semantically associated, i.e. they refer to the same issue.

Specifically, the first class of the present corpus refers to career counselor's professional knowledge, skills, resources and training. The second class refers to the principals variables and constructs related to vocational guidance and career counseling, such as self-efficacy, personality, coping, intelligence, emotions, satisfaction, optimism. The third class include the statistic measure and instruments used in vocational guidance to assess

people career self and personality. The fourth class refers to context variables, to the supports and barriers for inclusion, rights of people with vulnerabilities (people with disabilities, psychological sidelines, etc.). The fifth class includes the guidance services, projects, career guidance activities that are provided by local centers (university, region, province).

As already mentioned, differences between the abstract presented before the economic recession (pre-crisis: from 2002 to 2008) and during/after the economic recession (post crisis: form 2009 to 2017) were analysed. These two period in vocational guidance history are specific because the stable employment and secure organizations of the pre-crisis have in post crisis given way to a new social arrangement of flexible work and fluid organization, causing people tremendous distress, making difficult to comprehend career with theories that emphasize stability than mobility. Furthermore, it seemed interesting to analyze whether differences could be found in the theories and techniques presented in the abstract in pre e post crisis. To differentiate between papers presented pre- and post-crisis, a specific procedure was used based on the Chi2 association of semantic classes (Ratinaud, 2014) over the two period of time (Fig. 2).

The classes related to the pre-crisis are three and five characterised by statistic measure and instruments used in vocational guidance to assess people and guidance services, projects, and career guidance activities. The post-crisis period is characterized by the class four, that refers to context variables, to the support and barriers for inclusion, rights of people with vulnerabilities.

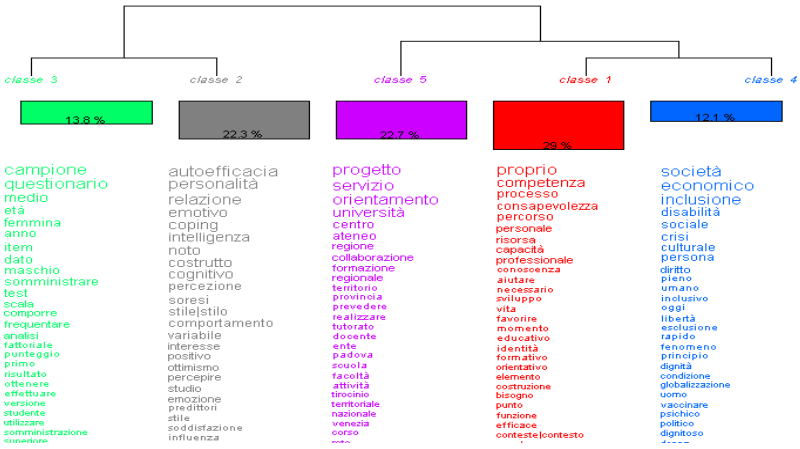


Fig. 1: Cluster Dendrogram and list of most relevant words for each lexical world (in descending order according to the Chi² value of each class).

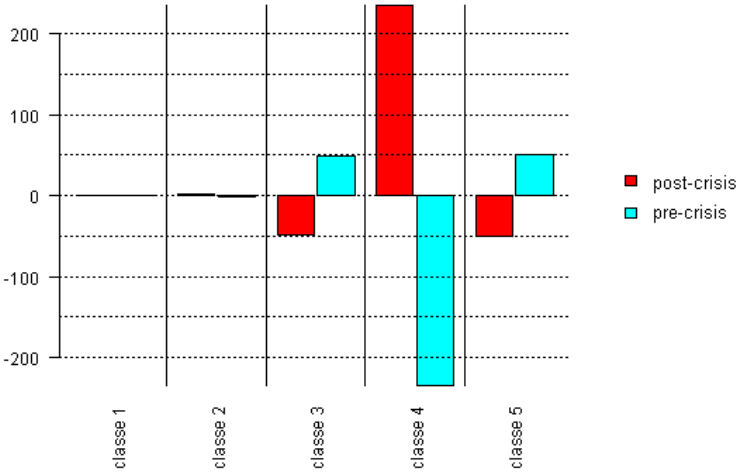


Fig. 2: Comparison among pre-crisis and post-crisis papers

These results highlighted that the topics presented in the abstract related to pre-crisis are more oriented towards "people" focusing on the assessment and measure with a statistical background. In the post-crisis period, the attention of counsellors is more oriented toward the "environment" in which people live and the relation between people and their context, so the uniqueness and the vulnerability of people are considers in relation to social and work inclusion. Finally, in order to compare and contrast the main semantic areas over time, the classes were analysed using the Chi2 association of semantic classes and their distribution over years (Fig. 3).

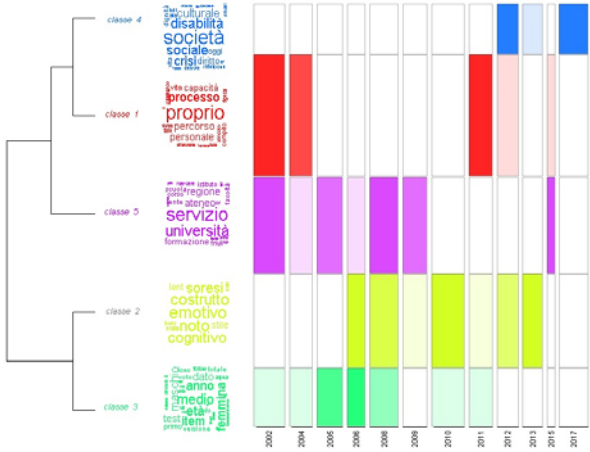


Fig. 3: Comparison among classes and their distributions over years

In addition to the classes already analyzed in the pre and post crisis periods, the comparison among classes and their distributions over years, highlights also class 1 and class 2, which can be considered as evergreen in the vocational guidance and career counseling field because they are present throughout almost the entire period considered. The class 1 refers to career counselor's professional knowledge, skills, and competences. The class 2 refers to variables and constructs related to vocational guidance and career counseling such self-efficacy, coping, life satisfaction, and positive attitudes.

4. Conclusions and discussion

The aim of the present study were to highlight the changes in theory, technique e measure emerged in the Italian vocational guidance and career counseling psychology by the analysis of the abstract presented at Italian Association for Vocational guidance and Career Counseling' Conferences.

The results show five different lexical worlds classes, related to career counselor's professional knowledge, variables and constructs of vocational guidance and career counseling, measure and instruments to assess people career self and personality, context variables to support inclusion of people with vulnerabilities, and career guidance services and center.

Differences between the abstract presented before the economic recession (pre-crisis: from 2002 to 2008) and during/after the economic recession (post crisis: form 2009 to 2017) were also analysed. The results shows that career counseling theories and technique evolved during the time to better assist workers in adapting to fluid societies and flexible organization and to better help clients design their lives in 21st century. In fact, while in the abstracts relented to the pre-crisis period, emphasis is given to all those guidance activities that consider particularly important to allow the person to collect information about their characteristics and needs before advancing decision-making hypotheses (measure and instrument for the assessment), in the abstracts related to the post-crisis period attention is paid to the "contexts" where people live. Career guidance practices that are limited to the analysis of "attitudes" and "interests" are considered obsolete, while current policies, challenges, socio-economic conditions, the way in which vulnerability is conceptualised are inputs from the environment which act at various levels and on which scholars should pay attention (Shogren, Luckasson, & Schalock, 2014).The evolution of social sciences that revolve around orientation is undoubtedly a very complex phenomenon. Career scholars and practioners should support people's needs taking into account the organizational and environmental context in which they develop and take shape. Currently the career guidance theory and model are numerous and not always denominated and defined in the same way by the various authors

and scholars. For these reasons is important to analyze and understand the different model developed over the time in order to activate a continuous comparison in the field of career counselor's competences that produces precise trajectory regard the constructs to develop in the people by program and activity provided by career services. In fact, noteworthy is the result that highlights how the classes that refer to vocational guidance and career counseling are presented throughout the entire period considered.

Nevertheless, these are just some results and other analyzes will be useful for examining the peculiarities that these specific classes assume during the years considered, in order to identify the specific skills and constructs that characterized different historical periods. It could also be important to compare the results that emerged in the Italian context with those of other European and North American contexts, to generalize the results obtained.

References

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Kluwer Academic Publishers: Dordrecht.

Ratinaud, P. (2014). Visualisation chronologique des analyses ALCESTE: application à Twitter avec l'exemple du hashtag #mariagepourtous. *Actes des 12es Journées internationales d'Analyse statistique des Données Textuelles*. Paris Sorbonne Nouvelle–Inalco.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.

Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» a` travers l'analyses tatistique d'un corpus de re´cits de cauchemars. *Langage & Société*, 66, 5–39.

Shogren, K. A., Luckasson, R. & Schalock, R. L. (2014). The fefinition of "context" and its application in the field of intellectual disability. Journal of Policy and Practice in Intellectual Disabilities, 11(2), 109-116.

Savickas, M. L., Nota, L., Rossier, J., Dauwalder, J. P., Duarte, M. E., Guichard, J., ... & Van Vianen, A. E. (2009). Life designing: A paradigm for career construction in the 21st century. Journal of Vocational Behavior, 75, 239-250.

Sbalchiero, S. & Santilli, S. Some introductory methodological notes. In L. Nota & S. Soresi (Eds.), For A manifesto in favor of Inclusion. Florence: Hogrefe Editore

Sbalchiero, S., & Tuzzi, A. (2016). Scientists' spirituality in scientists' words. Assessing and enriching the results of a qualitative analysis of in-depth interviews by means of quantitative approaches. Quality & Quantity, 50(3), 1333-1348.

A la poursuite d'Elena Ferrante

Jacques Savoy
Université de Neuchâtel (Suisse) – Jacques.Savoy@unine.ch

Abstract
The objective of an authorship attribution model is to determine, as accurately as possible, the true author of a document, literary excerpt, threatening email, legal testimony, etc. Recently a tetralogy called *My Brilliant Friend* has been published under the pen-name Elena Ferrante, first in Italian and then translated into several languages. Various names have been suggested as possible true author (e.g., Milone, Parrella, Prisco, etc.). Based on a corpus of 150 contemporary Italian novels written by 40 authors, two computer-based authorship attribution methods have been employed to answer the question "Who is the secret hand behind Elena Ferrante?" To achieve this objective, the nearest neighbor (*k*-NN) approach was applied on the 100 to 2,000 most frequent tokens using the Delta model. As a conclusion, we found that Domenico Starnone is the true author behind Elena Ferrante's pseudonym. As a second approach and using the entire vocabulary, Labbé's model confirms this finding.

Résumé
L'objectif d'un modèle d'attribution d'auteur consiste à identifier, de la manière la plus fiable possible, le véritable auteur d'un document, extrait d'une œuvre, d'un courriel menaçant ou d'un testament. Récemment, la tétralogie débutant avec *L'amica geniale* (*Une Amie Prodigieuse*) a été publié sous le nom de plume d'Elena Ferrante, d'abord en italien puis traduite dans plusieurs langues. Plusieurs noms ont été proposés comme le possible véritable écrivain (par exemple, Milone, Parrella, Prisco, etc.). En s'appuyant sur un corpus composé de 150 romans contemporains italiens écrit par 40 auteurs, deux méthodes d'attribution d'auteur ont été utilisés pour déterminer qui se cache derrière le pseudonyme Elena Ferrante. Dans ce but, la technique du plus proche voisin a été appliquée sur la base des 100 à 2 000 vocables les plus fréquents avec le modèle Delta. Comme conclusion, on aboutit au nom de Domenico Starnone comme la véritable identité de Elena Ferrante. Comme deuxième approche basée sur l'ensemble du vocabulaire, le modèle de Labbé confirme cette conclusion.
Keywords : Authorship attribution, corpus linguistics.
Mots-clés : Attribution d'auteur, linguistique de corpus.

1. Introduction

Avec la parution de *L'amica geniale* (2011) débute une tétralogie sur la vie à Naples depuis les années 50. Cette série de romans rencontre un étonnant succès, en particulier aux États-Unis. Toutefois, l'auteur indiquée, Elena Ferrante, représente un pseudonyme dont la véritable identité n'a pas été révélée. Des érudits et journalistes ont proposé plusieurs noms en tenant compte de possibles similarités stylistiques ou en affirmant que l'auteur doit connaître le Naples d'après-guerre, voire être une femme (par exemple, Erri De Luca, Francesco Piccolo, Michele Prisco, Fabrizia Ramondino, …). Sur la base des royalties versés, le journaliste C. Gatti (Gatti, 2016) affirme que la plume de Ferrante est tenue par Anita Raja (femme de l'écrivain Domenico Starnone). Aucune étude scientifique approfondie n'a abordé cette question, mais une première ébauche indique que le véritable auteur serait Domenico Starnone (Tuzzi *et al*., 2018). L'identification du véritable auteur de ces romans nous rappelle les investigations sur les relations Gary-Ajar en France dans les années 1970. Dans le monde anglo-saxon, la parution de *The Cuckoo's Calling* (2013) sous la signature de R. Galbraith correspond à une affaire similaire puisque le véritable auteur était J. K. Rowling (Juola, 2016). La découverte d'un poème inédit soulève également la question de son véritable auteur (Thisted & Efron, 1987), (Craig & Kinney, 2009). Pour lever le voile sur l'identité exacte de Ferrante, notre étude dispose d'un corpus de 150 romans italiens contemporains. De plus, on s'appuiera sur deux méthodes d'attribution d'auteur (Juola, 2006) reconnues et ayant fait l'objet de plusieurs études. En effet, afin d'admettre une preuve devant un tribunal celle-ci doit posséder plusieurs caractéristiques (Chaski, 2013) comme, par exemple, correspondant aux meilleures pratiques dans le domaine, avoir été testée et pouvant être vérifiée et répliquée. Enfin, nous faisons l'hypothèse que le véritable auteur derrière la signature Ferrante est bien l'un des 39 écrivains italiens présents dans notre corpus (attribution dans un ensemble fermé).

2. Travaux reliés

Afin de déterminer l'identité d'un écrivain, trois paradigmes principaux ont été proposées (Juola, 2006), (Stamatatos, 2009). D'abord, on s'est appuyé sur des mesures stylométriques admises comme invariantes pour chaque auteur, à l'exemple de la longueur moyenne des phrases, la taille du vocabulaire par rapport à la taille du document (TTR) (Rexha *et al*., 2016). Face à des textes de tailles variables, ces mesures s'avèrent d'être instables (Baayen, 2008).

Deuxièmement, les choix lexicaux permettent de différencier les auteurs, tant dans la sélection des mots que dans leur fréquence d'occurrences ; « Le style c'est l'homme » disait Buffon en 1753). Dans ce but, Mosteller & Wallace (1964) proposent de sélectionner semi-automatiquement les vocables les plus

pertinents. Burrows (2002) choisit les mots les plus fréquents et, en particulier, les mots fonctionnels (déterminants, prépositions, conjonctions, pronoms et verbes auxiliaires). Ces derniers possèdent l'avantage d'être plus fortement reliés au style de l'auteur qu'à la sémantique. Cette liste comprendra entre 50 à 1 000 vocables les plus fréquents (Hoover, 2007), voire l'ensemble du vocabulaire (Labbé, 2007). D'autres auteurs proposent de définir a priori une telle liste (Hughes *et al*., 2012). Sur cette base, chaque texte est représenté par les fréquences relatives d'occurrence des vocables sélectionnés. Ensuite, une mesure de distance (ou de similarité) permet d'estimer la proximité de deux textes. L'attribution s'établit habituellement selon la règle du plus proche voisin. Troisièmement, en recourant à des modèles d'apprentissage automatique (Stamatatos, 2009) les attributs les plus pertinents (mots, bigrammes de mots ou de lettres, partie du discours, émoticons, etc.) peuvent être sélectionnés. Ensuite un classifieur est entraîné pour générer les profils des auteurs retenus (Naïve Bayes, régression logistique, SVM, apprentissage en profondeur (Kocher & Savoy, 2017), etc.). Enfin, le texte d'attribution douteuse est représenté et le nom du profil le plus similaire est retourné comme réponse.

3. Le corpus de romans italiens contemporains

Grâce aux efforts de A. Tuzzi et M. Cortelazzo (Université de Padoue), le corpus PIC (Padova Italian Corpus) a été créé en 2017. Cette collection contient 150 romans italiens couvrant la période de 1987 à 2016. Comme l'indique le tableau 1, ce corpus contient des œuvres de 40 auteurs (dont Elena Ferrante avec sept textes). Lors de sa création, les auteurs originaires de Naples et de sa région ont été favorisés (10 noms indiqués en italique dans le tableau 1), de même que les femmes (12, pour 27 hommes).

Ce corpus contient 9 609 234 formes, avec une moyenne de 64 061 mots par œuvre (un seul écrit comprend moins de 10 000 formes). La longueur moyenne des romans signés par Ferrante s'élève à 88 933 mots. Enfin, un contrôle éditorial a été appliqué afin d'éliminer les éléments non-textuels (titre courant, numérotation des pages, etc.) ainsi qu'une inspection de l'orthographe. Ce corpus renferme donc des écrits de la même époque et langue, du même genre littéraire et dont la qualité a été vérifiée. Le 7 septembre 2017, un *workshop* regroupant sept équipes de chercheurs s'est tenu à l'Université de Padoue durant lequel le nom de Domenico Starnone a été identifié unanimement comme l'auteur derrière les œuvres de Elena Ferrante. Pour atteindre cette conclusion, notre approche s'appuie sur les techniques suivantes.

Tableau 1 : Nom des écrivains inclus dans le corpus avec le nombre de romans

Nom	H/F	Nombre	Nom	H/F	Nombre	Nom	H/F	Nombre
Affinati	H	2	Giordano	H	3	*Prisco*	H	2
Ammaniti	H	4	Lagiola	H	3	Raimo	H	2
Baiani	H	3	Maraini	F	5	*Ramondino*	F	2
Balzano	H	2	Mazzantini	F	4	*Rea*	H	3
Baricco	H	4	Mazzucco	F	5	Scarpa	H	4
Benni	H	3	*Milone*	F	2	Sereni	F	6
Brizzi	H	3	*Montesano*	H	2	*Starnone*	H	10
Carofiglio	H	9	Morazzoni	F	2	Tamaro	F	5
Covacich	H	2	Murgia	F	5	Valerio	F	3
De Luca	H	4	Nesi	H	3	Vasta	H	2
De Silva	H	5	Nori	H	3	Veronesi	H	4
Faletti	H	5	*Parrella*	F	2	Vinci	F	2
Ferrante	?	7	*Piccolo*	H	7			
Fois	H	3	Pincio	H	3			

4. Identifier l'auteur derrière la signature Elena Ferrante

Notre étude débute par l'application du modèle Delta (Burrows, 2002) dans lequel la sélection des attributs stylistiques correspond aux k vocables les plus fréquents. Toutefois, aucune limite précise pour le paramètre k n'est indiquée et des travaux précédents (Savoy, 2015) soulignent que des valeurs entre 200 et 500 tendent à apporter les meilleures performances. Cette limite fixée, la méthode Delta estime un Z score pour chaque vocable t_i basé sur la fréquence relative (dénotée rtf_{ij} pour le terme t_i et dans le document D_j) comme indiqué par l'équation 1 (avec $mean_i$ indique la fréquence moyenne du vocable et s_i son écart-type).

$$\text{Z score}(t_{ij}) = (rfr_{ij} - mean_i) / s_i$$

Pour chaque auteur, on concatène tous ses écrits pour générer son profil A_j. Enfin, on calcule la distance entre la représentation du texte à attribuer (dénotée Q) et les profils des auteurs A_j (voir équation 2). Ensuite, les différents auteurs peuvent être triés avec la plus faible distance signalant l'auteur le plus probable. Le tableau 2 redonne les trois premiers auteurs avec des valeurs pour k = 200, 300 et 500. Dans la dernière colonne (*Stopword*), les vocables choisis correspondent uniquement aux mots fonctionnels de l'italien (k = 307).

$$\Delta(Q, A_j) = {}^{1}/_{k} \cdot \sum_{i=1}^{k} \left| \text{Z score}(t_{iq}) - \text{Z score}(t_{ij}) \right|$$

Le tableau 2 nous renseigne sur l'attribution du roman *L'amica geniale* (2011). En considérant les six autres ouvrages, le même nom apparaît au premier rang. De même, si le nombre de vocables s'élève à 50, 100, 150, 250, 400, 1 000, 1 500 ou 2 000, nous retrouverons toujours Starnone en première place et ceci pour toutes les œuvres de Ferrante.

Une analyse plus fine des distances du tableau 2 indique que la différence (en pourcentage) entre les distances du premier et deuxième rang présente des valeurs nettement supérieures à celles entre le deuxième et troisième rang. Ainsi, si $k = 200$, la différence entre 0,524 et 0,686 s'élève à 30,9 % tandis que celle entre 0,686 et 0,700 n'est que de 2,0 %. Le premier nom proposé se détache clairement des autres.

Dans une deuxième série d'expériences, nous avons regroupé tous les romans attribués à Elena Ferrante pour en former qu'un seul texte (ou profil). En variant le nombre de vocables de 50, 100, 150, 200, 250, 300, 400, 500, 1 000, 1 500 à 2 000, Starnone se retrouve toujours au premier rang des auteurs ayant la plus forte similarité avec le profil d'Elena Ferrante.

Tableau 2 : Listes triées des auteurs les plus probables pour L'amica geniale *(méthode Delta)*

	$k = 200$		$k = 300$		$k = 500$		*Stopword*	
Rang	Distance	Auteur	Distance	Auteur	Distance	Auteur	Distance	Auteur
1	0,524	Starnone	0,515	Starnone	0,505	Starnone	0,421	Starnone
2	0,686	Veronesi	0,684	Brizzi	0,686	Veronesi	0,640	Milone
3	0,700	Balzano	0,719	Veronesi	0,710	Brizzi	0,660	Veronesi

Comme second modèle d'attribution d'auteur, l'approche de Labbé (2007) suggère de recourir à l'ensemble du vocabulaire. Dans ce cas, la distance entre deux textes A et B (indiquée par D(A,B) dans l'équation 3) dépend des fréquences absolues des vocables dans les deux textes (dénotées par tf_{iA}, respectivement tf_{iB}, avec $i = 1, 2, \ldots, k$). La variable n_A (ou n_B) signale la longueur de l'écrit A (en nombre de formes). Comme les deux textes ne possèdent pas des tailles identiques, les fréquences du plus long (B dans l'équation 3) seront multipliées par le rapport des tailles (voir partie droite de l'équation 3). Enfin, les valeurs D(A,B) seront comprises entre 0 (aucun mot en commun) et 1 (mêmes mots avec des effectifs identiques).

$$\mathrm{D}(A,B) = \sum_{i=1}^{k} \left| tf_{iA} - \widehat{tf}_{iB} \right| \Big/ (2 \cdot n_A) \quad \text{avec} \quad \widehat{tf}_{iB} = tf_{iB} \cdot n_A / n_B$$

En appliquant cette méthode, une distance est calculée entre chaque roman et la distance permet de trier les couples d'écrits, de la plus faible distance à la plus grande. Le corpus PIC génère (150 x 149) / 2 = 11 175 couples. Un extrait est repris dans le tableau 3.

Dans ce tableau, la première place correspond aux deux œuvres les plus similaires, deux romans écrits par Ferrante dans notre cas, soit *Storia di chi fugge e di chi resta* (Id : 51, (2013)) et *Storia della bambina perduta* (Id : 52, (2014)). Les deux autres romans de la tétralogie suivent, du deuxième au quatrième

rang, soit avec *Storia del nuovo cognome* (Id : 50, 2012) et *L'amica geniale* (Id : 49, (2011)). En cinquième position, on rencontre deux écrits de Faletti, soit *Niente di vero tranne gli occhi* (Id : 42, 2004) et *Io sono Dio* (Id : 44, 2009), puis deux romans de Veronesi (Id : 145, *Caos calmo* (2009) et Id : 147, *Terre rare* (2014)). Avec des distances faibles, les appariements s'opèrent entre des œuvres rédigés par le même auteur et dans un intervalle de temps assez court.

Tableau 3 : Liste triée des romans les plus similaires (méthode Labbé)

Rang	Distance	Id.	Auteur 1	Id.	Auteur 2
1	0.140	51	Ferrante	52	Ferrante
2	0.148	50	Ferrante	51	Ferrante
3	0.155	49	Ferrante	50	Ferrante
4	0.157	50	Ferrante	52	Ferrante
5	0.165	42	Faletti	44	Faletti
6	0.166	145	Veronesi	147	Veronesi
…	…	…	…	…	…
43	0.228	47	Ferrante	127	Starnone
…	…	…	…	…	…
63	0,241	108	Raimo	147	Veronesi

Lorsque la distance augmente, la probabilité de rencontrer le même auteur pour les deux ouvrages reliés diminue. Le premier lien apparament incorrect se situe au 43e rang avec un écrit de Ferrante (Id : 47, *I giorni dell abbandono* (2002) apparié avec un de Starnone (Id : 127, *Eccesso di zelo* (1993)). Un appariement entre ses deux auteurs apparaît également au rang 44, 53, et 54, avant que l'on découvre un autre type d'erreur en position 63 reliant un roman rédigé par Raimo (Id : 108, *Il peso della grazia* (2012)) et un autre de Veronesi (Id : 147, *Terre rare* (2014)). Puis, on découvre à nouveau un appariement entre Ferrante et Starnone aux rangs 65, 69, 71, 72, 73, 74, soit un total de dix couples entre ces deux auteurs et seulement un seul avec des autres écrivains. Sachant que Ferrante correspond à un pseudonyme, la forte similarité de style avec celui Starnone fait de ce dernier un choix de premier ordre.

5. Analyse

Les choix lexicaux ne sont pas le fruit du hasard et chaque auteur a ses préférences qui sont détectables par les mesures stylistiques. Le rapprochement entre Ferrante et Starnone s'explique également en analysant quelques exemples. Dans notre corpus, les sept romans de Ferrante correspondent à 6,5 % de la taille tandis que 6,4 % est constitué par les dix œuvres de Starnone. Si les fréquences d'occurrences de certains mots s'écartent de ces proportions et dans la même direction pour les deux auteurs, nous pouvons rapprocher leur style.

Le nom *padre* apparaît 9 815 fois dans le corpus PIC. Dans les œuvres de Ferrante, on en dénombre 833 (8,5 % du total) et 1 170 chez Starnone (11,9 %). Ce mot est clairement employé plus fréquemment par ces "deux" auteurs. De manière similaire, le mot *madre* possède une fréquence de 8 246 dans le corpus pour 1 104 occurrences (13,4 %) sous la plume de Ferrante et 762 (9,2 %) avec Starnone. D'autres vocables fonctionnels possèdent des distributions similaires. Ainsi le mot *persino* (même) apparaît 1 351 fois dans la collection PIC et on en compte 266 (19,7 %) chez Ferrante et 205 (15,2 %) chez Starnone. On notera également que ce terme peut également s'écrire *perfino* (avec une fréquence d'occurrences de 20 avec Ferrante, 18 chez Starnone). Pour Ferrante et Starnone, on voit une préférence pour une forme, tandis que d'autres auteurs recourent uniquement à l'une des orthographes (Baricco : uniquement *perfino*, Tamaro : seulement *persino*). Enfin certains écrivains ignorent les deux mots (Covacich, Parrella) ou l'utilisent très rarement (De Luca ou Balzano). Comme exemples complémentaires, certains mots ne sont employés que par Ferrante et Starnone comme *risatella* (gloussement, 16 occurrences chez Ferrante, 4 avec Starnone) ou *contraddittoriamente* (contradictoirement, Ferrante : 6; Starnone : 9).

Pour un écrivain italien, le lexique peut inclure des formes provenant du dialecte comme celui de Naples avec le terme *strunz* (*stronzo* en italien). Ce terme apparaît 85 dans le corpus, avec 63 occurrences dans les romans de Starnone et 18 chez Ferrante (et deux fois chez De Silva et Raimo).

Certains *n*-grammes de mots s'avèrent plus fréquents chez Ferrante et Starnone comme *no essere che* (ne pas être ça) qui apparaît 23 fois (100 %) dans le corpus mais 6 (26,1 %) sous la plume de Ferrante et 7 (30,4 %) sous celle de Starnone. Ensemble ces deux auteurs apportent plus de 56 % des occurrences de cette séquence.

6. Conclusion

Cette étude s'appuie sur deux méthodes d'attribution d'auteur reconnues d'une part, et, d'autre part sur un corpus de 150 romans contemporains rédigés par 40 auteurs. Comme attributs stylistiques, nous avons retenu les 100, 150, 200, 250, 300, 400, 500, 1 000, 1 500 et 2 000 mots les plus fréquents pour la méthode Delta (Burrows, 2002). Avec ces différentes valeurs, le premier nom retourné comme le probable auteur s'avère toujours Domenico Starnone et ceci pour les sept romans parus sous le nom Ferrante. En s'appuyant sur l'ensemble du vocabulaire et la méthode de Labbé (2007), la même conclusion est obtenue.

En analysant quelques choix lexicaux, on découvre des relations étroites entre Starnone et Ferrante. Par exemple, le mot *persino* est sur-employé dans les romans des deux auteurs, et la second forme *perfino* n'apparaît que plus

rarement. Chez d'autres écrivains, on rencontre habituellement une préférence pour l'un des deux termes ou l'absence de leur usage. Enfin, suite à l'atelier qui s'est tenu à Padoue le 7 septembre 2017 aboutissant à désigner Domenico Starnone comme l'écrivain derrière la signature Ferrante, celui-ci a démenti en être le véritable auteur (Fontana, 2017).

Remerciements
Cette recherche a été possible grâce à A. Tuzzi et M. Cortelazzo qui nous ont transmis le corpus PIC.

Références

Baayen, H.R. (2008). *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267-287.

Chaski, C. (2013). Best practices and admissibility of forensic author identification. *Journal of Law and Policy*, 21(2):333-376.

Craig, H., & Kinney, A.F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

Fontana, E. (2017). Lo scrittore Domenico Starnone: "Io non sono Elena Ferrante". Il Giornale, 9 sept.

Gatti, C. 2016. La véritable identité d'Elena Ferrante révélée. BublioObs, 2 octobre 2016.

Hoover, D.L. (2007). Corpus stylistics, and the styles of *Henry James*. *Style*, 41(2):160-189.

Hughes, J.M., Foti, N.J., Krakauer, D.C., & Rockmore, D.N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the PNAS,* 109(20), pp. 7682-7686.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information*, 1(3):233-334.

Juola P. (2016). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(1), i100-i113.

Kocher, M., & Savoy, J. (2017). Distributed language representation for authorship attribution. Digital Scholarship in the Humanities, 2017, to appear.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1):33-80.

Mosteller, F., & Wallace, D.L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading.

Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2016). Towards a more fine grained analysis of scientific authorship. *Proceedings ECIR 2016*, pp. 26–31.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, 30(2):246-261.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):433-214.

Tuzzi, A., & Cortelazzo, M. (2018). What is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer. *Digital Scholarship in the Humanities*, to appear.

Regroupement d'auteurs dans la littérature du XIXe siècle

Jacques Savoy
Université de Neuchâtel (Suisse) – Jacques.Savoy@unine.ch

Abstract
This paper presents the author clustering problem in which a set of n texts written by several distinct authors must be regrouped into k clusters, each of them corresponding to a single author. The proposed model can use different distance measures and feature sets (composed of the most frequent word types). The evaluation is based on a French corpus composed of 200 excerpts of novels written during the 19th century. By varying different parameter settings, the evaluation indicates a better performance achieved with words instead of n-grams of letters. The Cosine distance achieves lower performance levels compared to the Tanimoto (L_1) or Matusita (L_2) functions. The text size plays an important role in the effectiveness of the solution, showing that size of 10,000 tokens produces significantly better results than text size of 5,000 to 500 tokens. A more detailed analysis provides reasons explaining stylistic aspects of some authors.

Résumé
Cette communication présente le problème du regroupement d'auteurs dans lequel un ensemble de n textes écrits doit être regroupé dans k grappes distinctes, une pour chaque auteur. Le modèle proposé permet l'emploi de différentes mesures de distance et divers ensembles d'attributs (vocables les plus fréquents). L'évaluation s'appuie sur un corpus composé de 200 extraits de romans français du XIXe siècle. En variant différents paramètres, notre étude indique que les vocables s'avèrent meilleur que les n-grammes de lettres. La fonction cosinus génère un taux de réussite plus faible que le fonction Tanimoto (L_1) ou Matusita (L_2). La taille des textes joue un rôle important dans la qualité de réponse et une longueur de 10 000 mots permet une performance significativement supérieure à des valeurs variant de 5 000 à 500 mots. Une analyse apporte quelques explications sur le style de différents auteurs.
Keywords : Automatic classification, unsupervised machine learning, authorship attribution.
Mots-clés : Classification automatique, apprentissage non-supervisé, attribution d'auteur.

1. Introduction

Le problème d'attribution d'auteur (Juola, 2006) rencontre un intérêt grandissant avec la multiplication des canaux électroniques. La présence de messages anonymes ou pseudo-anonymes soulève de nombreux défis en criminalité (Olsson, 2008), (Chaski, 2013) à l'exemple des chats calomnieux ou des courriels menaçants. Pourtant des questions plus classiques méritent notre attention comme, par exemple, déterminer la véritable identité de la romancière Elena Ferrante (Gatti, 2016) ou sur les relations de Shakespeare et de ses co-auteurs (Michell, 1996), (Craig & Kinney, 2009).

Dans ce cadre, notre communication présente les problèmes liés à la question du regroupement d'auteurs avec une application en littérature française du XIXe siècle. Ce problème se résume ainsi. Disposant d'un ensemble de n extraits de romans, on doit regrouper en k classes disjointes, chacune contenant tous les écrits du même auteur. Ce problème a été posé lors de la campagne d'évaluation CLEF-PAN 2016 et 2017 (Stamatatos *et al*., 2016) mais les collections tests n'ont pas été rendues publiques. Ce problème présente une difficulté majeure par l'absence de données d'entrainement.

2. Travaux reliés

Afin d'identifier l'auteur d'un écrit, trois familles d'approches ont été proposées (Juola, 2006). En premier lieu, des mesures stylométriques supposées invariantes ont été évoquées comme la longueur moyenne des phrases, la taille du vocabulaire par rapport à la longueur du document (rapport TTR) (Rexha *et al*., 2016). Toutes ces mesures possèdent l'inconvénient d'être instables face à des textes de tailles différentes (Baayen, 2008). Une deuxième famille d'approches se fonde sur le vocabulaire. Mosteller & Wallace (1964) proposent de sélectionner de manière semi-automatique les vocables les plus pertinents. Burrows (2002) sélectionne les mots les plus fréquents et, en particulier, les mots fonctionnels (déterminants, prépositions, conjonctions, pronoms et verbes auxiliaires). Ces derniers possèdent l'avantage d'être plus fortement reliés au style de l'auteur qu'à la sémantique. Cette liste comprendra entre 50 à 1 000 vocables les plus fréquents (Hoover, 2007). D'autres auteurs proposent de définir a priori une telle liste (Hughes *et al*., 2012). Ainsi, chaque texte peut être représenté par les fréquences d'occurrence de ces vocables. Ensuite, une mesure de distance (ou de similarité) permet d'estimer la proximité de deux textes. L'attribution s'établit habituellement selon la règle du plus proche voisin.

Troisièmement, des modèles d'apprentissage automatique (Stamatatos, 2009) permettent de sélectionner les attributs (mots, bigrammes de mots ou de lettres, POS, émoticons, etc.) possédant le meilleur pouvoir discriminant. Ensuite un classifier est entraîné sur un ensemble d'apprentissage (SVM,

régression logistique, etc.). Cependant, dans le cadre du regroupement d'auteurs, aucune donnée d'entraînement n'est disponible rendant caduc de telles approches. Dès lors, pour résoudre ce problème, des approches proposent de déterminer en premier lieu le nombre *k* d'auteurs sur l'ensemble *n* d'écrits (Stamatatos *et al*., 2016). Cette valeur fixée, on applique un algorithme de classification *k-means* afin d'identifier les différents groupes de textes. Par itération, le nombre *k* d'auteurs peut être affiné. Comme second paradigme, la distance entre chaque écrit est calculée, puis on applique un algorithme de classification hiérarchique (Lebart *et al*., 1998) pour former les grappes de documents. Dans cette étude, nous suivrons cette seconde stratégie de résolution, choix qui nous a permis d'obtenir le deuxième rang lors de la dernière campagne d'évaluation PAN-CLEF 2016.

3. Corpus de test et méthodologie d'évaluation

L'évaluation empirique tient une place importante en attribution d'auteur. Comme les corpus des campagnes PAN-CLEF 2016 et 2017 n'ont pas été rendus publics, nos évaluations seront basées sur une collection extraite de la littérature française du XIXe siècle. Ce corpus nommé St-Jean[1] contient 200 extraits de romans écrit par 30 auteurs (entre 1801 (Châteaubriant, *Attala*) et 1901 (Régnier, *Les Rencontres de Monsieur de Bréot*)). Ce nombre d'écrivains et de textes étant élevé, la tâche demeure ardue. Chaque auteur est représenté par au moins trois extraits (avec un maximum de treize pour Balzac) provenant d'un à six romans et aucun écrivain ne représente plus de 5 % du corpus. Chaque extrait contient en moyenne 10 073 formes (min : 10 026 ; max : 10 230 ; standard déviation : 25). Au total, ce corpus contient 2 014 641 formes pour 51 661 vocables extraits de 67 romans. Disposant de *n* textes, notre approche produira une liste ordonnée de liens entre textes avec une indication de la distance entre eux. Un exemple est présenté dans le tableau 1. Avec ce corpus, la solution se compose de 30 groupes requérant la présence de 670 liens intra-auteurs. Comme mesure d'évaluation, nous reprenons la précision moyenne (AP) (la moyenne des précisions obtenues pour chaque lien pertinent), mesure usitée lors des campagnes PAN-CLEF 2016 et 2017. Ainsi, une valeur unique de performance reflète la qualité de chaque modèle de classification. Comme seconde mesure, la valeur HP (haute précision) indique le nombre de liens correctement établis depuis le début jusqu'à la présence du premier lien erroné. Dans notre tableau 1, la valeur HP = 168 signalant que les 168 premiers liens sont justes.

[1] Ce corpus a été créé par D. Labbé et est disponible (www.unine.ch/clc/home/corpus.html) soit sous la forme de textes, soit lemmatisé. Les encodages UTF-8 et Windows sont disponibles.

Tableau 1 : Exemple d'un extrait d'une liste ordonnée selon la distance (Tanimoto)

Rang	Distance	Texte 1	Texte 2
1	0,239	51 Flaubert	62 Flaubert
2	0,242	3 Flaubert	20 Flaubert
3	0,248	29 Sand	115 Sand
4	0,248	122 Staël	140 Staël
5	0,253	125 Fromentin	159 Fromentin
6	0,255	37 Flaubert	62 Flaubert
7	0,256	132 Régnier	162 Régnier
...	…	…	…
169	0,324	42 Maupassant	51 Flaubert

4. Sélection des attributs et mesure de distance

Afin de regrouper les documents selon leur auteur, nous devons les représenter en fonction de leur style et non en fonction des thèmes qu'ils abordent. Comme mentionné précédemment, plusieurs études ont démontré que les vocables les plus fréquents constituent des attributs pertinents pour détecter le style d'un auteur. Dans le cadre de l'attribution d'auteur, le thème pourrait perturber des affectations correctes lorsque, par exemple, deux auteurs abordent des sujets similaires. Pour cerner les aspects stylistiques, une étude récente a démontré que tenir compte des 200 à 300 mots les plus fréquents (Savoy, 2015) apporte de bonnes performances comparées à d'autres fonctions de sélection (rapport des cotes, gain d'information, chi-carré, etc.). Sur la base du corpus St-Jean, les mots les plus fréquents de notre corpus sont : de (4,11 % des occurrences), et (2,44 %), la (2,36 %), le (1,94 %), et à (1,9 %). Comme alternative, plusieurs études proposent de recourir aux fréquences des lettres et des bigrammes de lettres et, plus généralement, des *n*-grammes afin de distinguer les différents styles (Kjell, 1994), (Juola, 2006). On remarquera toutefois que les composantes stylistiques et thématiques seront toutes les deux présentes dans la génération de tels *n*-grammes. Dans cette étude, la distinction entre majuscules et minuscules est ignorée et les signes de ponctuation sont éliminés. Par contre, on tiendra compte du fait qu'une lettre débute ou termine un mot. Le nombre maximal d'attributs s'élève à (27 x 27) + 27 = 756. Pour la langue française, on retrouve 594 (ou 78,6 %) combinaisons possibles dans notre corpus. Les lettres françaises les plus fréquentes sont : e (15.6 % des lettres), s (8,3 %), a (8,3 %), i (7,5 %), et t (7,2 %). En indiquant par _ l'espace, les bigrammes de lettres les plus usuels sont : e_ (5,1 % des bigrammes), s_ (3,5 %), t_ (2,7 %), _d (2,4 %), et _l (1,8 %).

Dès que chaque document est représenté par *m* de mots (ou de *n*-grammes de lettres), on peut calculer sa distance avec les autres entités du corpus. Le choix de cette fonction de distance (ou de similarité) peut s'opérer selon des critères théoriques (par exemple, symétrie, inégalité triangulaire) ou

empiriques (efficacité). Basée sur le profilage d'auteur, une étude récente (Kocher & Savoy, 2017) indique qu'aucune mesure de distance s'avère toujours la meilleure. Par contre un groupe restreint permet d'obtenir de bonnes performances comme la distance de Manhattan ou de Tanimoto basée sur la norme L_1, ou celle de Matusita (norme L_2). Nous avons repris ces mesures en y ajoutant la distance du cosinus. Ces quatre mesures respectent la symétrique et respectent l'inégalité triangulaire (Kocher & Savoy, 2017).
Dans la définition de ces mesures de distance, les lettres majuscules indiquent les vecteurs représentants les documents. Les minuscules (a_i, b_i) correspondent aux fréquences relatives des termes sélectionnés.

$$dist_{Manhattan}(A,B) = \sum_{i=1}^{m} |a_i - b_i|$$

$$dist_{Tanimoto}(A,B) = \sum_{i=1}^{m} \left(|a_i - b_i| \Big/ Max(a_i, b_i) \right)$$

$$dist_{Matusita}(A,B) = \sqrt{\sum_{i=1}^{m} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2}$$

$$Sim_{Cosine}(A,B) = \sum_{i=1}^{m} a_i b_i \Big/ \sqrt{\sum_{i=1}^{m} a_i^2} \sqrt{\sum_{i=1}^{m} b_i^2}$$

$$dist_{Cosinus}(A,B) = cos^{-1}(Sim_{Cosine}(A,B)) \Big/ \pi$$

5. Évaluation

Notre première évaluation concerne l'efficience des différentes mesures de distance ainsi que la performance du nombre de vocables les plus fréquents retenus comme attributs. Le tableau 2 indique les valeurs de précision moyenne (AP) et de haute précision (HP) en représentant les textes par les 100 à 1 000 vocables les plus fréquents, ou tout le vocabulaire. La dernière ligne et colonne nous renseigne sur la moyenne des APs.

Tableau 2 : Précision moyenne (AP) et haute précision (HP) selon diverses mesures de distance avec des représentations construites entre 100 vocables et tout le vocabulaire

	Manhattan		Tanimoto		Matusita		Cosinus		Moyenne
Attributs	AP	HP	AP	HP	AP	HP	AP	HP	AP
100	0,674	185	0,695	192	0,655	181	0,626	152	0,663
200	0,692	186	0,708	193	0,687	222	0,628	145	0,679
300	0,705	**190**	0,720	**196**	0,727	**244**	**0,629**	**148**	0,695
500	0,720	186	0,735	189	**0,750**	212	0,627	149	0,708
1 000	**0,730**	183	**0,743**	186	0,745	204	0,617	142	**0,709**
Tout	0,713	166	0,672	168	0,568	135	0,599	142	0,691
Moyenne	0,706	183	0,712	187	0,689	200	0,621	146	0,681

Ces résultats indiquent que les différences de précision moyenne restent faibles entre les mesures de Manhattan, Tanimoto et Matusita. Toutes les trois s'avèrent supérieures au cosinus. En considérant la haute précision (HP), Matusita tend à apporter une meilleure efficacité. Reste à déterminer a priori cette valeur maximale, sans connaître les attributions correctes. Enfin, une représentation par 300 à 500 voire 1 000 vocables les plus fréquents fournit les meilleurs taux de succès. En remplaçant les vocables par des *n*-grammes de lettres (performances indiquées dans le tableau 3), les valeurs de performance s'avèrent inférieures aux vocables. La variation des taux de succès entre une combinaison uni- et bigrammes de lettres (deuxième ligne du tableau 3) ou des séquences plus longues s'avère peu élevée. Par contre les temps de traitement s'accroissent rapidement (8,2 minutes pour les uni- et bigrammes à plus de 4 heures pour les 5-grammes comparé à 3 minutes avec les 500 mots les plus fréquents). Enfin, la fonction cosinus retourne les performances les moins bonnes. Nos premières évaluations se fondaient sur l'ensemble du texte disponible, soit environ 10 000 mots. Si l'on réduit cette taille à 5 000 voire à 500, les taux de réussite obtenus sont indiqués dans le tableau 4. La première ligne est reprise du tableau 2 puis les tailles décroissent comme le signale la première colonne. La réduction moyenne des performances est reprise dans la dernière colonne. Ainsi, en réduisant les textes à 5 000 mots, la baisse moyenne s'élève à 25,8 %. Si l'on doit œuvrer avec des longueurs de 1 000 à 500 mots, les taux de réussite s'avèrent faibles générant une réduction de 80 à 90 %. Est-il vraiment raisonnable d'effectuer des attributions d'auteur avec de telles tailles ?

Tableau 3 : AP et HP selon diverses mesures de distance avec des n-grammes de lettres

	Manhattan		Tanimoto		Matusita		Cosinus		Moyenne
n-grams	AP	HP	AP	HP	AP	HP	AP	HP	AP
uni & bi	0,559	139	0,559	139	0,503	128	**0,538**	94	0,540
3-gram	0,527	108	0,527	108	0,471	130	0,476	108	0,500
4-gram	0,570	153	0,570	153	0,507	147	0,481	**112**	**0,532**
5-gram	0,587	177	0,587	177	0,541	181	0,543	73	0,565
6-gram	**0,588**	**200**	**0,588**	**200**	**0,557**	**188**	0,415	36	0,588
Moyenne	0,566	155	0,566	155	0,506	147	0,510	97	0,545

Tableau 4 : AP et HP selon diverses mesures de distance avec des textes de tailles différentes (représentation sur la base de 300 vocables)

	Manhattan		*Tanimoto*		*Matusita*		*Cosinus*		*Moyenne*	*Différence*
Taille	*AP*	*HP*	*AP*	*HP*	*AP*	*HP*	*AP*	*HP*		
10 000	*0,705*	*190*	*0,720*	*196*	*0,727*	*244*	*0,629*	*148*	*0,695*	
5 000	*0,526*	*55*	*0,545*	*58*	*0,526*	*85*	*0,466*	*74*	*0,516*	*-25,8%*
2 500	*0,326*	*31*	*0,342*	*39*	*0,306*	*35*	*0,284*	*11*	*0,315*	*-54,8%*
1 000	*0,152*	*4*	*0,152*	*2*	*0,116*	*1*	*0,141*	*3*	*0,140*	*-79,8%*
500	*0,093*	*2*	*0,089*	*2*	*0,079*	*3*	*0,086*	*2*	*0,087*	*-87,5%*

En analysant la liste triée obtenue avec la fonction Matusita et en représentant les textes par les 300 vocables les plus fréquents, les distances les plus faibles se retrouvent entre des extraits de la même œuvre. La distance la plus faible se trouve avec le roman *Les Rencontres de Mr de Bréot* (1901) de Régnier, puis on trouve *Bouvard et Pécuchet* (1881) de G. Flaubert, *Delphine* de Mme de Staël (1803), *Mme Bovary* (1857) de G. Flaubert et *La Petite Fadette* (1832) de G. Sand. Si l'on analyse les appariements les plus difficiles entre deux œuvres du même auteur, les romans *Graziella* (1852) et *Geneviève* (1863) de A. de Lamartine constitue le lien le plus distant. Ensuite, on rencontre *La double Maîtresse* (1900) de H. de Régnier, *Aurélia* (1855) et *Les Illuminés* (1852) de G. de Nerval et *Le père Goriot* (1833) et *La Maison Nucingen* (1838) de H. de Balzac. Ces auteurs peuvent adopter des styles assez dissemblables, rendant une attribution plus ardue. Parmi les œuvres dont le style est perçu comme proche par la machine mais qui sont écrites par deux auteurs distincts, on trouve en tête *Bel-Ami* (Maupassant, 1885) et *Mme Bovary* (Flaubert, 1857), puis *Volupté* (Sainte-Beuve, 1834) et *Dominique* (Fromentin, 1862), *Notre Cœur* (Maupassant, 1890) et *Mme Bovary* (Flaubert, 1857), et enfin *L'Assommoir* (Zola, 1879) et *Mme Bovary* (Flaubert, 1857).

6. Conclusion

Parmi les fonctions de distance, notre étude indique que le cosinus n'apporte pas de bons résultats. Par contre, les différences de performance entre les fonctions Manhattan, Tanimoto ou Matusita demeurent faibles. Afin de cerner une partie importante du style des auteurs, le recours à une représentation sur la base de vocables s'avère plus efficiente que le recours aux *n*-grammes de lettres (pour *n* variant de 1 à 6). Représenter le style avec les 300 à 500 vocables les plus fréquents s'avère pertinent.

Lorsque l'on compare la précision moyenne (AP) et la haute précision (HP), le choix des paramètres optimaux diffère quelque peu d'une mesure de performance à l'autre. Notons que l'AP ne punit pas sévèrement les erreurs d'affectation, erreurs qui entraînent immédiatement une baisse de la valeur HP. Enfin, la taille des textes joue un rôle essentiel dans une attribution d'auteur et des valeurs inférieures à 1 000 mots ne permettent que des affectations souvent douteuses. Parmi les auteurs retenus, le style du roman *Mme Bovary* se rapproche de celui de Maupassant (*Bel-Ami*) ou de Zola (*L'Assommoir*).

Remerciements

L'auteur remercie D. Labbé pour avoir mis à sa disposition le corpus St-Jean.

Références

Baayen, H.R. (2008). *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267-287.

Chaski, C. (2013). Best practices and admissibility of forensic author identification. *Journal of Law and Policy*, 21(2):333-376.

Craig, H., & Kinney, A.F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

Gatti, C. 2016. La véritable identité d'Elena Ferrante révélée. BublioObs, 2 octobre 2016.

Hoover, D.L. (2007). Corpus stylistics, and the styles of *Henry James*. *Style*, 41(2):160-189.

Hughes, J.M., Foti, N.J., Krakauer, D.C., & Rockmore, D.N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the PNAS,* 109(20), pp. 7682-7686.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information*, 1(3):233-334.

Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifier. *Literary and Linguistics Computing*, 9(2):119-124.

Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. *Information Processing & Management*, 53(5):1103-1119.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1):33-80.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data.* Dordrecht, Kluwer.

Michell, J. (1996). *Who Wrote Shakespeare?* Thames and Hudson: New York (NY).

Mosteller, F., & Wallace, D.L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading.

Muller, C. (1992). *Principes et méthodes de statistique lexicale*. Honoré Champion, Paris.

Olsson, J. (2008). *Forensic Linguistics.* Continuum, London.

Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2016). Towards a more fine grained analysis of scientific authorship. *Proceedings ECIR 2016*, pp. 26–31.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, 30(2):246-261.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):433-214.

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2016). Clustering by authorship within and across documents. *Working Papers, CLEF-2016.*

What's Old and New? Discovering Topics in the American Journal of Sociology[1]

Stefano Sbalchiero, Arjuna Tuzzi
University of Padova – stefano.sbalchiero@unipd.it; arjuna.tuzzi@unipd.it

Abstract
Nowadays the field of text mining techniques seems to be very active in dealing with the increasing mass of available digital texts and several algorithms have been proposed to analyze and synthesize the vast amount of data that today represents a challenging source of information overload. Topic modeling is a collection of algorithms which are useful for discovering themes, i.e. topics, in unstructured text. The Latent Dirichlet Allocation (LDA) by Blei (et al., 2003) was one of the first topic modeling algorithms and since then the field seems to be active and many variants and other algorithms have been suggested. The present study considers a topic as an indicator of the relevance of a research area in a specific time-span and its temporal evolution pattern as a way to identify the paradigm changes in terms of theories, ideas, forgotten topics, evergreen subjects and new emerging research interests. The study aims to contribute to a substantive reflection in Sociology by exploring the temporal evolution of topics in the abstracts of articles published by the American Journal of Sociology in the last century (1921-2016). Within the classical LDA perspective, the study also focus on topics with a significant increasing or decreasing trend (Griffiths et Steyvers, 2004). The results show different shifts that involved relevant reflections on various issues, from the early debate on the "institutionalization" process of Sociology as a scientific discipline to recent developments of sociological topics that clearly indicate how sociologists have reacted to new social problem.

Keywords: Chronological corpus, History of Sociology, Academic Journals, Text Mining, Latent Dirichelet Allocation

[1] This study was supported by the University of Padova, fund CPDA145940 (2014) "Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature" (P.I. Arjuna Tuzzi, University of Padova).

1. Introduction: topic modeling

As evidenced by the literature on topic modelling (Blei et al., 2003; Ponweiser, 2012; Grimmer et Stewart, 2013; Griffiths et Steyvers, 2004), text mining approaches can mitigate the problem of analysing huge collections of textual data when they increase in number and size and complicate all information processing. From a methodological point of view, since the topics emerge directly from data, text mining approaches can tone down some problems about the role of analysts in coding and interpreting the content hidden in corpora, e.g. research bias or errors that notoriously affect most approaches in comparative and quanti-qualitative researche (Strauss et Corbin, 1990; Corbetta, 2003). A popular approache to extract information by summarizing the main contents embedded in relevant collection of texts in digital form is known as topic modeling (Blei et Lafferty, 2009), which is essentially a collection of algorithms that are exploited to discover themes, i.e. topics, in unstructured and complex texts. The Latent Dirichlet Allocation (LDA) is one of the first topic modeling algorithms, namely a "generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" (Blei et al., 2003, p. 996). LDA is a technique that facilitates the automatic discovery of themes in a collection of documents. Since a text document can deal with different topics and the words that occur in that document reflect a set of possible topics, in "statistical natural language processing, one common way of modeling the contributions of different topics to a document is to treat each topic as a probability distribution over words, viewing a document as a probabilistic mixture of these topics" (Griffiths et Steyvers, 2004, p. 5228). Actually we cannot directly observe topics but only documents and words, as topics are part of the latent and hidden text structure. The model infers the latent topic structure given by observed words and documents: this is the LDA's generative processes which recreate (generate) the documents of the corpus by assigning the probability of topics (the relevance) to documents and the probability of words to topics. The result is a probabilistic distribution of topics over documents that is characterized and described by a cluster of co-occurring words (Blei et al., 2003). This list o words enable the researcher to interpret the meaning of all the generated topics. For the purposes of the present study, the temporal variable is crucial to analyse the direction and evolution of topics, and particularly to the extent that they have a direct relationship with the most significant shifts in the development of Sociology as a discipline over time. For these reasons, we propose a LDA-based topic detection procedure as this "method discovers a set of topics expressed by documents, providing quantitative measures that can be used to identify the

content of those documents, track changes in content over time" (Griffiths et Steyvers, 2004, p. 5228). An additional estimation procedure exploits a metavariable (year) to explore the topics trends: LDA offers the opportunity to estimate the slope of a linear model that represents the distribution of topics by year. The model permits to identify "hot and cold topics" (Griffiths et Steyvers, 2004), i.e. topics with significant increasing (hot) and decreasing (cold) trends through time.

2. Corpus and data

The American Journal of Sociology (AJS), established in 1895 as the first U.S. scholarly journal in its field, can be considered one of the world's preeminent journals and a leading voice for research in social sciences. The journal fosters pathbreaking work from all areas of sociology, with an emphasis on theory building and innovative methods. AJS is a multi-disciplinary journal that strives to speak to a "general sociological reader" and is open to sociologically informed contributions from anthropologists, statisticians, economists, educators, historians, and political scientists. Manuscripts are subjected to a double-blind review process and published articles are considered representative of the best current theoretical and methodological debates. Our corpus includes all the abstracts of the papers published by AJS that have been retrieved from popular archives (Scopus and Web of science) and the journal webpages. We decided to work on the abstracts since they provide concise information about the main contents of all articles. With regard to selection criteria, they were based on the following consideration: when abstracts did not provide any information about the content or did not refer to relevant scientific contributes (e.g. editorials, master heads, errata, acknowledgements, rejoinders, notes, announcements, corrections, list of consultants, obituary, etc.) we decided to disregard them in further analyses.

The corpus is composed of 3,992 abstracts, collected for a period of almost a century (mean: 41 per year), from the Volume No. 27, Issue No. 1 (1921) to the latest, No. 121, Issue No. 4 (2016). The collected texts had relevant contents for the purpose of the present analysis based on the following consideration and hypothesis: If we consider a topic as an indicator of the relevance of a research area in a specific time-span, then the temporal evolution pattern of subject matters can portray main paradigm changes in terms of theories, ideas, forgotten topics, evergreen subjects and new emerging research interests in Sociology. The corpus has been pre-processed by means of TaLTaC2 software package. After the tokenization (the identification of words given character sequences chopping it up into pieces), the corpus has been normalized replacing uppercase with lowercase letters. An automatic search procedure identified relevant multi-words (MWs), i.e.

informative sequences of words (Pavone, 2010) repeated at least five times in the corpus (849 MWs in total). This procedure retrieved most interesting MWs in the abstract (e.g. *united states*, fr. 395; *social structure*, fr. 115; *social science*, fr. 101; *labor market*, fr. 89; *social change*, fr. 78) and contributed to increase the amount of information conveyed by sequences of words[2]. Then, the corpus has been processed by means of R software packages[3]: punctuation marks and numbers have been removed, as well as some grammatical words (articles, conjunction, prepositions, pronouns). The corpus is composed of 24,418 word-types and 512,410 word-tokens (tab. 1), and the measures show that there is a sufficient level of redundancy to proceed with statistical analyses of textual data (Lebart et al., 1998; Trevisani et Tuzzi, 2015; Bolasco, 2013).

Table 1. Basic lexical measures of the corpus of AJS abstracts

(V) WORD-TYPES	24,418
(N) WORD-TOKENS	512,410
(V/N)*100 = TYPE/TOKEN RATIO	4.76
(V1/V)*100 = PERCENTAGE OF HAPAX	47.08

3. Topic detection

As the LDA algorithm "fits" the terms in the document into a number of topics that must be specified *apriori*, this represents an important and sensitive decision that affects results and findings: few topics will produce broad subjects and mixed-up contents, while too many topics will produce minimal subjects and results too detailed to be readable and interpretable. To set the number of topics in a data driven manner we have the opportunity to calculate different metrics (Arun et al., 2010) and estimate the optimal number of topics (Griffiths et Steyvers, 2004) by means of the maximum log-likelihood of LDA for a number of topics ranging from 2 to 50 (Fig. 1).

[2] If MWs did not appear at least 5 times in the corpus, that is about once every 20 years, it was not considered important; however, the MWs that appeared with a frequency equal to or greater than 10 are 417.

[3] The analysis were implemented by R pakages: Tm, Lda, Topic model.

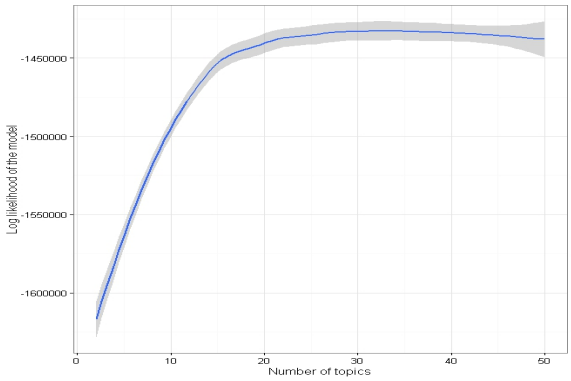


Fig. 1: Fitting the model: log-likelihood calculated for increasing number of topics

The best number of topics is the one with the highest value of log-likelihood that is around 30 and can be established as the optimal number of topics.
Figure 2 shows the general trend of all the 30 topics as depicted by the fitted model A clue of how these topics change over time is shown by 30 panels with a topic trend line each, that lists the number of topics with positive or negative trends. All of the topics are ordered by slope: decreasing topics appear in the first panels (top left), and increasing ones in the last panels (bottom right)., Since the main aim of this study is to detect the temporal evolution of old, new and emerging topics in Sociology, we can resort to a limited number of topics that show prototipycal temporal patterns(Ponweiser, 2012; Griffiths et Steyvers, 2004).

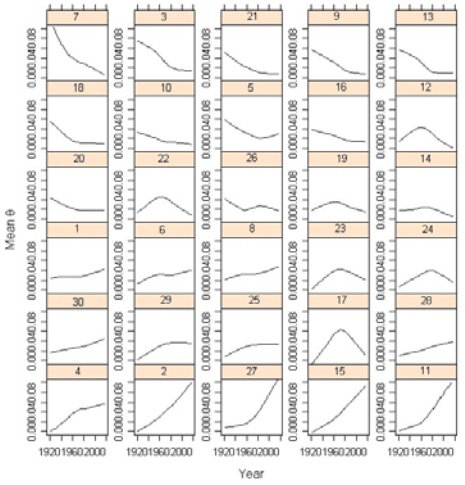


Fig. 2: Temporal patterns of the 30 topics in Sociology sorted by slope of linear models

Consistent with the idea that topics show different trends and embrace theoretical, conceptual, and methodological shifts, the analysis of time-dependent phenomena identifies three specific temporal patterns of topics: topics whose trajectory has grown in time and it is increasing over time (28, 4, 2, 27, 15, 11); topics whose trajectory decreased (7, 3, 21, 9, 13, 18); and topics whose peak-like journey (meteor) was high only in a specific interval of time (14, 17, 28, 15) or shows more irregular temporal trajectories.

4. What's old and new in Sociology?

To focus on major increasing or decreasing topics from 1921 to 2016, we explored the contents of five coldest and hottest topics. Figure 3 provides the top term for these topics.

The groups of coldest topics correspond on one hand to the methodological development of sociological perspectives, and on the other hand to some specific objects of research. These topics were very popular in about 20s and 50s. First of all, the debate on the "institutionalization" process of Sociology as a scientific discipline characterized the early debate (topic 7). The main need was to create a strong scientific and knowledge base from the development of ideas advanced by the "founding fathers", e.g. Durkheim. At the same time, the debate on the "measurement" of social phenomena arose. The issue of migration between cities and farms (topic 3) by economic and social groups gives the net law of rural-urban social selection. The emerging of a scientific social reflections about health and illness (topic 21) by using empirical data to evaluate how social life affects morbidity and mortality rate, and vice versa, increased in efforts for better educated public and to improve health legislation. The development of psychological sociology (topic 9) and the general progress of psychological interpretations of social processes and institutions have decreased over time; researches in this tradition have been criticized because they mainly exemplified the biological background of social interpretations, also supplied by the impulse from the Darwinian doctrine. Class culture, conflict and leisure (topic 13) were popular issues in the 30s and 50s: the industrialization had raised many questions, from the class conflict to the growth of leisure hours of after work hours, providing new insights for social thought. The group of hottest topics (Fig. 4) is related to articles that have a focus of interest in a wide range of empirical case studies that underline most significant changes that have occurred since the mid-1960s.

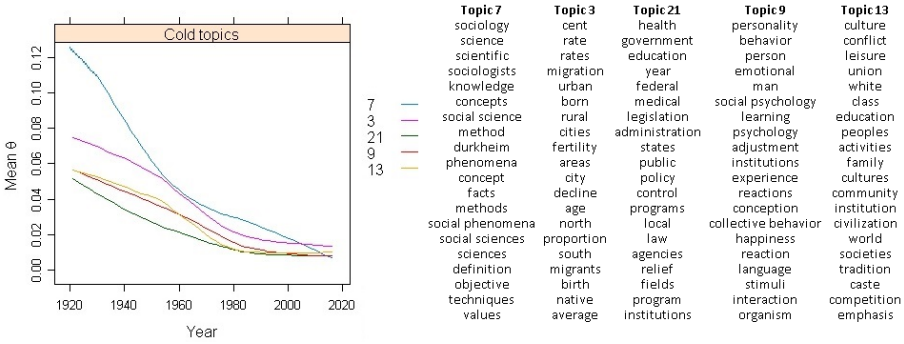


Fig. 3: Decreasing topics: the five coldest (significant neg., p level 0.005)

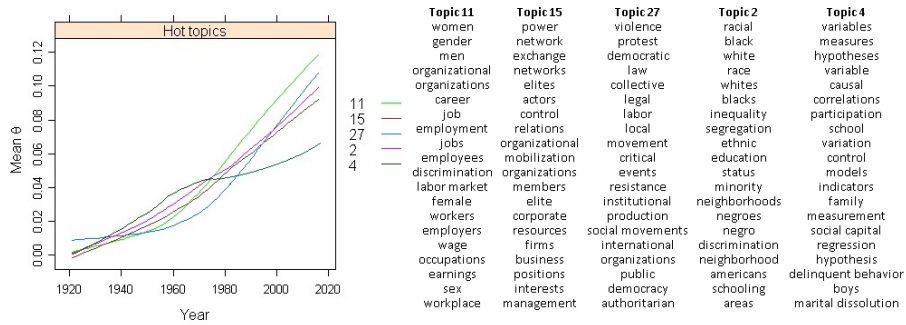


Fig. 4: Increasing topic attention: the five hottest (significant pos., p level 0.005)

In those years, gender revolution (topic 11), ethnic discrimination (topic 2), mobilization, power and élite (topic 15), protests and social movements (topic 27), and the "measurement" of social phenomena in a post-positivist fashion, especially until the 70s (topic 4), offered to sociologists the opportunity to deal with a social effervescence of a particular historical moment. These hot topics indicates the 'birth' and recent developments of some sociological topics that clearly indicate how Sociology (as a discipline) and sociologists have reacted to new social problems.

In conclusion, through the topic detection analysis of the abstracts of articles, different shifts that involved reflections on various issues have been identified. During the twentieth century, Sociology expanded its scope and influence, and motivated much research studies as well as a diversification of the field. Other studies have offered a remarkable theoretical contribution to the historical 'shape' of Sociology as a discipline (Kalekin-Fishman et Denis, 2012), even in a critical perspective (Turner, 1998), either emphasizing the content of the various domains of sociology (Scott et Desfor Edles, 2011; Blau, 2004), or specifically within the intellectual ground of American Sociology since the mid-nineteenth century (Calhoun, 2007). Even if they show an

interesting round of paradigmatic reflection in Sociology, there has been a lack of research studies on the history of Sociology through empirical data and evidence to fast-moving sociological topics over time. To the extent that the history of Sociology is a continuous approach to the Sociology of the present, a new way of reading the history of a discipline is rely on topic detection of articles published in mainstream journals which mirror the sociological scientific debate of a specific historical moment. We analysed these trends exploiting topics as emerged from a text corpus and highlighted two distinct directions of topics, characterized by different theoretical and methodological implications that coexist within the same period considered: the hot-increasing and cold-decreasing topics. Results show how Sociology has become one of the main social science to provide fresh thinking about a whole range of topics affecting the public sphere and, as a consequence, the discipline developed shifting priorities in universities and social research agenda towards specialization and fostered the birth of a wide range of sub-disciplines over time. This is just the tip of the iceberg: further analyses will shed light on many more aspects that need a deeper reflection.

References

Arun R., Suresh V., Veni Madhavan C. E. and Narasimha Murthy M. N., (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.), *Advances in knowledge discovery and data mining*, Springer Berlin Heidelberg, pp. 391-402.

Blau J. R. (2004). *The Blackwell Companion to Sociology*, Malden, MA: Blackwell.

Blei D. M., Ng A. and Jordan M. I., (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Blei D. M and Lafferty J.D., (2009). Topic Models. In A. Srivastava, M. Sahami (eds.), *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Press.

Bolasco, S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining.* Carocci, Rome.

Calhoun G. (2007). *Sociology in America: A History*. Chicago: University of Chicago Press

Corbetta P. (2003). *Social Research: Theory, Methods and Techniques*, SAGE Publications Ltd., London.

Griffiths T. and Steyvers M., (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 101(Supplement 1):5228-5235.

Grimmer G. and Stewart B. M., (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, in *Political*

Analysis, 21 (3): 267-297.

Kalekin-Fishman D. and Denis A. (2012). *The Shape of Sociology for the 21st Century: Tradition and Renewal*, London, SAGE.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring textual data*. Kluwer Academic Publishers: Dordrecht

Pavone, P. (2010). Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus. In S. Bolasco, I. Chiari and L. Giuliano (Eds*.) Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles, 9-11 June 2010*, Sapienza University of Rome, pp. 131-140. LED.

Ponweiser M., (2012). *Latent Dirichlet Allocation in R*, Vienna University of Business and Economics.

Scott A. and Desfor Edles A. (2011). *Sociological Theory in the Contemporary Era: Text and Readings*, Thousand Oaks, Pine Forge Press.

Strauss, A.L. and Corbin, J. (1990). *Basics for Qualitative Research: Grounded Theory Procedures and Techniques*, Newbury Park, Sage.

Trevisani, M. and Tuzzi, A. (2015). A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal. *Quality & Quantity*, 49(3): 1287-1304.

Trevisani, M. and Tuzzi, A. (in press). Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based System.*

Turner S. (1998). Who's Afraid of the History of Sociology? *Swiss Journal of Sociology*, 24: 3-10.

Comparison of Neural Models for Gender Profiling

Nils Schaetti, Jacques Savoy
Université de Neuchâtel - Rue Emile-Argand 11 - CH2000 Neuchâtel - Switzerland

Abstract
This paper describes and evaluates two neural models for gender profiling on the PAN@CLEF 2017 tweet collection. The first model is a character-based Convolutional Neural Network (CNN) and the second an Echo State Network-based (ESN) recurrent neural network with various features. We applied these models to the gender profiling task of the PAN17 challenge and have demonstrated that it can be applied to gender profiling. As features, we propose using pre-trained word vectors, part-of-speech (POS) and function words (FW) for the ESN model, and character 2-grams matrix with punctuation marks, smilies, beginning and ending 2-grams for the deep learning model. We finally compared these strategies to a baseline and found that an ESN model based on Glove pre-trained word vectors achieves the highest success rate and outperforms the baseline and the character-based CNN model.

Keywords: Author Profiling, Gender Profiling, Deep-Learning, Convolutional Neural Network, Reservoir Computing, Echo State Network, Natural Language Processing

1. Introduction

At the age of big data, a large number of applications are based on an exponential amount of various data such as pictures, videos, articles, links and blogs shared directly from computers, websites, smartphones and sensors. Social networks and blogs are the new platforms of communication based on fast interactions, generating a large varieties of content with their own characteristics. These contents are difficult to compare with traditional texts, such as novels and articles.

This issue raises new questions : Can we determine if the author of a textual content is a man or a woman ? Can we identify the author's place of origin, his age group or his (or part of) psychological profile ? Answering these questions can help solve current issues of the social network era, such as fake news, plagiarism and identity theft. Author profiling is, therefore, a particular and pertinent subject interest.

In addition, author profiling is central to applications involving marketing, security and forensics. For example, forensic linguistics and police

investigation forces would like to know specific defining characteristics, such as the gender, the age group and the socio-cultural background of an author of harassing messages. When we apply this to marketing, companies and resellers could make use of these profile characteristics while targeting their consumers' preferences, based on the analysis of individual consumer social network posts and online product consulting. In order, to extract this information, the classic statistical methods are employed as they have proven to be effective for text classification.

Deep learning has gained increasing popularity just over the last decade, becoming a "breakthrough" technology in image recognition and computer vision. Yet, it faces difficulties in natural language processing (NLP) tasks. But recurrent neural networks (RNN), as well as long short-term memory (LSTM) obtained better results in such tasks. In this view, we therefore decided to test such an approach on the gender profiling tasks with two neural models, one based on Convolutional Neural Networks (CNN) and 2-grams of characters, and the other on the Reservoir Computing Paradigm. Finally, we compare them to a baseline composed of both a random and a naive Bayes classifier

This paper is organized as follows. Section 2 introduces the data set used to train and test both of the models and the methodology used for evaluation. Section 3 describes and evaluates our deep-learning model. Section 4 introduces the proposed echo state network-based reservoir computing model. Section 6 compares the results with the baseline. In the last section, we draw conclusions on our findings and possible future improvements.

2. Methodology

To compare our two models on the gender profiling task, we needed a common ground composed of the same dataset and evaluation measures. To create this common ground, the PAN CLEF evaluation campaign was launched [1] and allowed multiple research groups to propose and compare profiling algorithms with the same methodology.

For the PAN CLEF 2017 evaluation campaign, four test collections of tweets were generated written in several languages including English. Based on these collections, the challenge was to classify Twitter profiles per language variety (e.g., UK vs. US English) and gender. We were then able to use this common ground for our two models and compare their capacities on the gender profiling task.

The dataset was collected on Twitter and is composed of tweets from different authors with 100 per author. For each author, a label indicates the correct gender (male, female). The collection included 3,600 authors, residing in the United-States, Great Britain, Ireland, New Zealand, Australia and

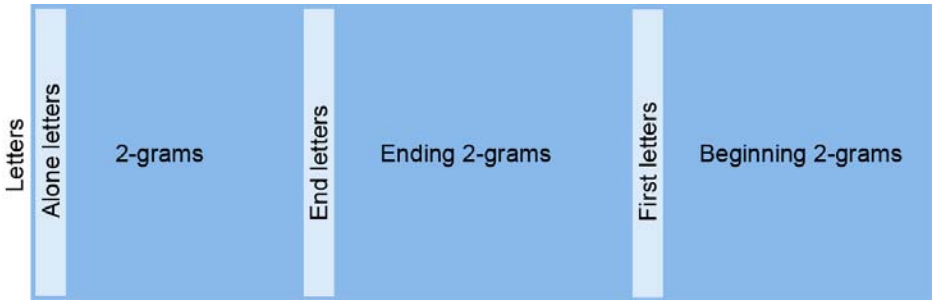
Canada, 600 per country, and 1,800 for each group, for a total of 360'000 tweets. The table velow resumes dataset properties.

Authors	Tweets	Genders
3600	360k	(male) 1800 ; (female) 1800

The overall performance of a model is based on the accuracy on the gender profiling task. The accuracy is the number of correctly classified author gender divided by the number of authors. Based on data depicted in the table above, a random baseline will produce an accuracy rate of 0.5 (or 50%).

3. Character N-grams Matrix-based Convolutional Neural Networks

A Convolutional Neural Network (or CNN) is a variety of feed-forward artificial neural networks inspired by the visual cortex [2]. In our first model, we applied a CNN to a character bigram representation matrix for an author in a collection. The first shows the structure of the representation matrix.



For each letter, one can find one row. In the first position, the relative frequency of this letter is provided. Then, from left to right, the matrix is composed of the relative frequencies of each character bigram (e.g., at row "t" and column "h", the relative frequency of the bigram "th" is given). The hird part is optional and composed of relative frequencies of ending character bigrams, and finally, the last part is the same optional matrix representing the starting character bigrams of each word. This matrix representing an author is the input for the CNN

The first two layers are composed of 20 and 10 kernels respectively, with a size of 5 × 5. These layers are followed by a drop-out layer. The last two are linear layers based on ReLU. The outputs are finally obtained by a Softmax function and give the author's predicted class. The predicted class is therefore the class with the highest corresponding output from this function. The training set is composed of 90% of the dataset and the remaining 10% is use to estimate the performance. This procedure is repeated 10 times with

non-overlapping test sets to obtain the 10-fold cross validation estimator.

Matrix / Alphabet	English	+ Punctuation	+ Punctuation & Smilies
Bigrams	75.26%	76.16%	76.51%
+ starting bigrams	76.02%∗	77.63%∗†	77.50%∗
+ ending bigrams	75.94%	77.22%†	77.25%
+ starting & ending bigrams	76.12%	77.83%†	78.33%∗†

4. Echo State Network-based Reservoir Computing models

4.1. Echo State Networks

An Echo State Network was introduced in [3] and corresponds to the first equation. In this model, the highly non-linear dimensional vector x_t, denoting the activation vector at time t, is defined by

$$x_{t+1} = (1 - a) * x_t + a * f(W_{in} * u_{t+1} + W * x_t + W)$$

where $x_t \in R\ N_x$ with N_x the number of neurons in the reservoir. The scalar *a* represents the leaky rate allowing to adapt the network's dynamic to the the task to be learned. The input signal is represented by the vector u_t with dimension N_u, multiplied by the weight matrix in $W \in R^{Nx \times Nu}$. In addition, the matrix $W \in R^{Nx \times Nx}$ stores the internal weights. Finally, W_{biais} is the bias, and usually the initial vector is fixed to $x_0 = 0$, corresponding to a null state.
The network's output $\hat{y}$ is defined by $\hat{y}_t = g(W * x_t)$ and the learning phase consists of finding the values of the matrix $W_{out} \in R^{Ny \times Nx}$, e.g., by applying the ridge regression method. This matrix is defined by

$$W_{out} = Y * X^T(X * X^T + \lambda * I)^{-1}$$

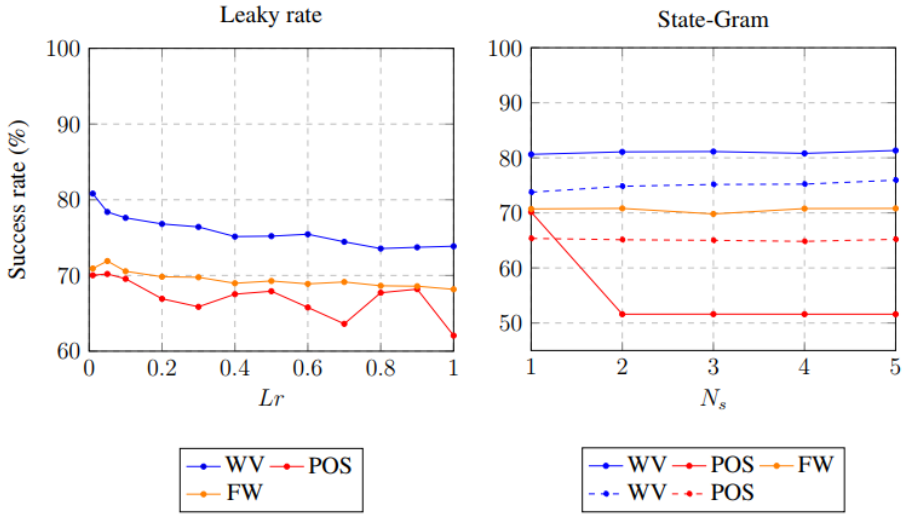
Where $Y \in R^{Ny \times T}$ is the matrix containing each target output $\hat{y}$ for t = 1, 2, . . ., T where T denotes the training size, and N_y the number of outputs (categories). Similarly, the matrix $X \in R^{Nx \times T}$ stores the reservoir states x_t obtained during the training phase. Finally, the parameter λ is a regularization factor.

4.2. From texts to temporal signals

In order to apply ESN for text classification, we must first transform input texts as a temporal signal. In this study, we have evaluated three signal converter methods. First, each word sequence in a text (e.g. "to the citizens

of") can be viewed as a word vector (WV) (e.g., vec(to), vec(the), vec(citizens), vec(of), each vector extracted from word embeddings pre-trained with Glove), Part-Of-Speech (POS) vector (size : number of POS tags), and as a function word (FW) (size : number of FW).

As output, the ESN generated the vector $y_{t,g}$ with g∈{male, female} denoting the probability that the tokens in the ESN's memory at time t as been written by a man or a woman. We then end up with an output temporal signal of gender probabilities (over t = 1, 2, . . ., T), and the final predicted class of a document is the one with the highest average across time.



4.3. State-Gram

In addition, the output layer can take account of more than one state to estimate the class probabilities. A state-gram value of 2 means that the training is performed, not only on a single x_t , but on $x_{t-1} \cup x_t$. Such a model was effective for handwritten digits recognition [4].

5. Results

In the second table, one can see the results of the deep-learning CNN model with different vocabulary and starting and ending bigrams. The statistical tests indicate that the starting bigrams can significantly improve the performance with respect to the base model (first row). The combination of starting and ending bigrams (last row) shows a significant improvement only for the vocabulary composed of punctuation marks and smilies. The best result (78.33%) is achieved by a CNN model with punctuation and smilies, with starting and ending character bigrams.

The left plot in the second figure shows the three features (WV, POS, FW) with a leak rate between 0.01 and 1.0. Using the same three feature sets, the right-side plot indicates the accuracy rate obtained by the state-gram model with value between 1 and 5. With a solid line, the best leak-rate parameter value is used, and with the dotted curves, a leak-rate value of 1 was used. Overall, Figure 2 indicates that the pre-trained word vector (WV) is the best feature set with a maximum value of 80.81% with a leak rate of 0.01. As the best accuracy rates is obtained with a leak rate between 0.01 and 0.05 (left plot in Figure 2), we can conclude that the author profiling task has a very slow temporal dynamics. The right-side plot signals that no significant improvement is achieved by increasing the value of the stage-gram parameter for the best leak-rate parameter value. Moreover, a high value of N_s decreases the performance for POS feature. The performance slightly increases for a leak-rate parameter value of 1, but these results show that the leak-rate parameter is a better lever to increase the accuracy rates.
The following table compares the accuracy rates that can be achieved by a random classifier, the naive Bayes model together with the CNN and ESN models (with $N_x = 1{,}000$).

Classifier	10-CV success rate
Random baseline	50.0 %
Naive Bayes classifier baseline	75.5 %
CNN 2-grams + starting-grams + ending-grams	78.3 %
ESN on Glove with $N_x = 1000$	80.6 %

6. Conclusion

This paper presents a comparison of two neural models composed of a character-based CNN model and an echo-state network (ESN) model with POS, function words (FW) or pre-trained word vectors (WV) as possible feature sets. Based on the CLEF-PAN 2017 dataset, the best CNN model achieves a success rate of 78.3% with a feature set composed of the vocabulary, the punctuation marks, and smilies. The best ESN model obtains a success rate of 80.6% with 1,000 neurons and a leak-rate of 0.01. Based on our experiment setting, this model achieves the best performance. In comparison, the naive Bayes classifier obtains a success rate of 75.5% and the average and best performance for the gender profiling task in PAN 2017 was respectively 75.88% and 82.5%.
Our results indicate that the two models can significantly improve the accuracy rate on the gender profiling task. Moreover, they demonstrated that

a simple model, thanks to its simple linear regression algorithm, such as the echo state network can achieve a higher success rate than a more complex model such as a character-based CNNs. This higher result level can be explain by the recurrent architecture of the ESN model, allowing it to take into account word order. In the future, we want to explore more features for the ESN and word vectors pre-trained for Twitter applications to achieve hopefully a better performance. We will also apply classical and deep ESN architectures to other natural language processing tasks such as authorship identification and author diarization.

References

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at pan 2016 : cross-genre evaluations. Working Notes Papers of the CLEF, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11) :2278–2324, 1998.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany : German National Research Center for Information Technology GMD Technical Report, 148(34) :13, 2001.

Nils Schaetti, Michel Salomon, and Raphaël Couturier. Echo state networks-based reservoir computing for mnist handwritten digits recognition. In Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC), 2016 IEEE Intl Conference on, pages 484–491. IEEE, 2016.

Segments répétés appliqués à l'extraction de connaissances trilingues

Lionel SHEN
Université Sorbonne Nouvelle - Paris 3 – lionel.shen@sorbonne-nouvelle.fr

Abstract
In a context of globalized societies, multilingualism is becoming an economic and social phenomenon. Translation constitutes a crucial element for communication. A good translation guarantees the quality of the transmission of all information. However, face to the challenge of multilingual information monitoring, can we simply use translation? With the advent of the digital age and the integration of all new technologies, corporate governance is undergoing a complete metamorphosis. One of the priorities remains the efficient exploitation of accumulated big data. The objectives of this paper hope to highlight the specificity and efficiency of the Repeated Segments tool through information discovering of trilingual thematic corpora (French, English and Chinese).

Résumé
Dans un contexte de sociétés mondialisées, on peut parler de multilinguisme ou encore de plurilinguisme. Aujourd'hui, la frénésie autour du phénomène des mégadonnées et le multilinguisme sont en train de métamorphoser tous les services et les comportements de notre époque. La traduction devient alors un élément capital pour la communication entre les peuples. Une bonne traduction garantit la qualité de la transmission de toutes les informations. Cependant, devant la gageure que constitue le projet de réaliser une veille multilingue, peut-on utiliser simplement la traduction ? Cet article s'articule autour d'explorations de corpus thématiques trilingues appliquées à l'extraction de connaissances et tente de mettre en lumière la spécificité et l'efficacité des cooccurrences en trois langues, français, anglais et chinois.

Keywords: segments répétés, textométrie, veille multilingue, multilinguisme, fouille d'informations, text-mining, cooccurrences, poly-cooccurrences

1. Introduction

Le monde, qui utilise des centaines de langages depuis des millénaires, a formalisé les mots et les grammaires pour transcrire, enseigner et transmettre sur des supports, les savoirs, les faits et les pensées. Des hiéroglyphes aux

idéogrammes, en passant par les alphabets, ces représentations diffusent ainsi l'image du monde à travers les époques, les évolutions, les moeurs et les courants de pensée. Cela représente aujourd'hui des centaines de milliards de mots dans des corpus différents, avec des occurrences variables. Il n'est pas possible à un être humain d'aborder par lui-même la masse des publications archivées ou en circulation. Seul l'usage de l'informatique peut, à présent, dans le cadre de la mondialisation, permettre un balayage massif des séquences des corpus nécessaire à l'étude des occurrences et des usages des mots, au moins dans les langues essentielles diffusant le savoir, l'information et la communication entre les humains. L'utilité de ces recherches est étendue, allant des besoins sociaux, humains, scientifiques aux guerres économiques, en passant par les médias et les enjeux stratégiques des politiques. C'est la capacité à détecter, enregistrer, analyser et comprendre dans les meilleurs délais, qui va permettre aux différentes forces de pouvoirs d'anticiper les décisions et d'agir efficacement. Cette force de veille, implantée de manière continue et basée sur des outils performants, élaborés et mis en œuvre par des chercheurs, des informaticiens, des stratèges, des économistes, sous l'autorité des décideurs... va donc construire les forces de demain, parfois à l'échelle de la planète. Dans un contexte de sociétés mondialisées, on peut parler de multilinguisme ou encore de plurilinguisme. Aujourd'hui, la frénésie autour du phénomène *Big-data* et le multilinguisme sont en train de métamorphoser tous les services et les comportements de notre époque. La traduction devient alors un élément capital pour la communication entre les peuples. Une bonne traduction garantit la qualité de la transmission de toutes les informations. Cependant, devant la gageure que constitue le projet de réaliser une veille multilingue, peut-on utiliser simplement la traduction ?

Cet article s'articule autour d'explorations de corpus thématiques trilingues appliquées à l'extraction de connaissances et tente de mettre en lumière la spécificité et l'efficacité de l'outil Segments répétés en trois langues.

2. Corpus

Pour constituer ce travail, deux types de corpus sont mobilisés : un corpus comparable (nommé ENRG) et un corpus parallèle (nommé CLRG), composés de données textuelles extraites des discours de presse, ainsi que ceux des ONG. La construction de ces deux corpus s'effectue autour de trois thèmes d'actualité ayant pour objet, l'environnement, l'énergie et le changement climatique.

La construction de ces deux corpus s'opère à partir d'articles de journaux issus de nos trois sphères de communication, à savoir, le Monde pour la France (4 817 articles), le NYT pour les États-Unis (3 993 articles) et 1200

médias pour la Chine (14 509 articles) comme le présente les deux figures (figure 1 et figure 2) ci-dessous.
Les données textuelles extraites du corpus comparable proviennent des discours de la presse, tandis que celles du corpus parallèle sont issues de ceux des ONG.

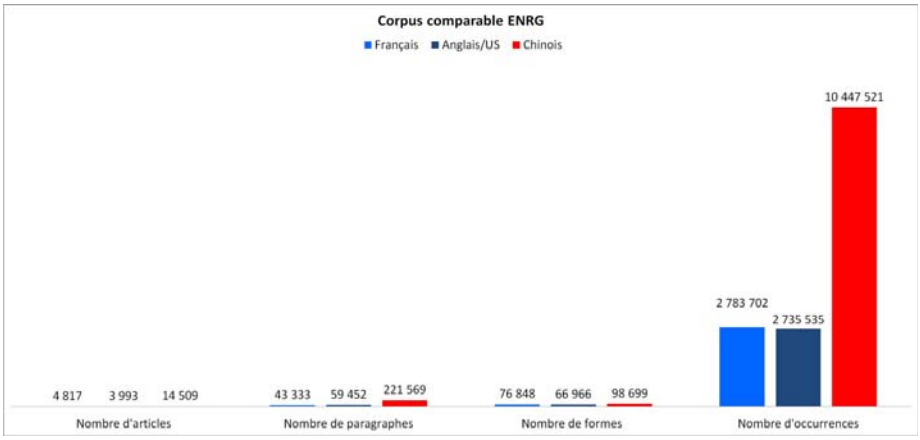


Figure 1 : volumétrie du corpus comparable ENRG

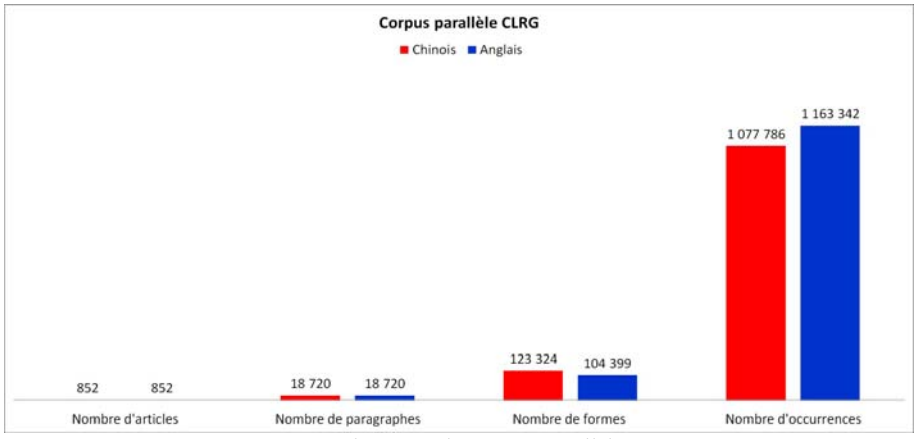


Figure 2 : volumétrie du corpus parallèle CLRG

Quant à l'aspect temporel des données du corpus comparable, il diffère selon les sources et couvre des périodes plus ou moins étendues : de 1999 à 2012 pour le Monde, de 2005 à 2012 pour le NYT, de 2008 à 2013 pour les médias chinois. Concernant le corpus parallèle, les articles datent de 2006 à 2014. La figure 3, ci-dessous montre les différentes périodes couvertes par les médias retenus.

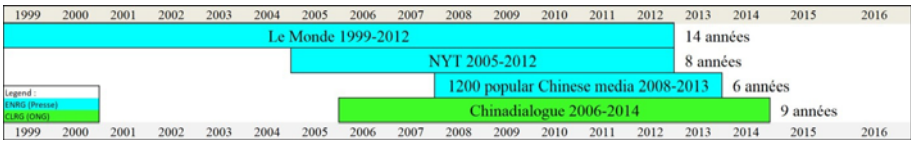


Figure 3 : périodes couvertes par les corpus ENRG et CLRG

Les dépouillements sont réalisés à l'aide des outils de la textométrie, notamment grâce aux analyses factorielles des correspondances (AFC), aux spécificités du modèle hypergéométrique, aux segments répétés, aux réseaux cooccurrentiels et poly-cooccurrentiels ou encore à la carte des sections. Les caractéristiques locales et globales, les convergences, les divergences et les particularités de ces différents corpus ont été mises successivement en évidence. Après avoir présenté rapidement les deux corpus utilisés, nous allons nous polariser sur l'outil Segments répétés appliqué d'abord au corpus parallèle puis ensuite au corpus comparable. Nous nous intéresserons, plus particulièrement dans cet article à la spécificité des segments répétés appliqués à l'extraction de connaissances multilingues. Comme le souligne André Salem, « *L'outil prend toute sa valeur lorsque l'unité linguistique traitée n'est pas le mot, mais le segment répété (suite de mots d'une longueur 2, 3, 4, 5)* » (Salem 1987).
Nous rappelons que «*Un segment répété est une suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus*».
Nous émettons l'hypothèse suivante : l'outil Segments répétés serait plus performant en chinois, qu'en anglais et qu'en français.

Corpus parallèle : segments répétés anglais-chinois
Nous examinons maintenant les segments les plus répétés obtenus à partir des deux volets (anglais-chinois) du corpus parallèle CLRG.

Tableau 1 : segments les plus répétés du corpus parallèle CLRG

longeur	segment répété	équivalent français	fréquence	longeur	segment répété	équivalent français	fréquence
2	of the	du	8159	2	中国 的	Chine	1252
2	in the	dans le	5877	2	温室 气体	gaz à effet de serre	1016
2	to the	au	2790	2	中的	milieu de	897
2	and the	et le	2558	2	的 是	est	888
2	climate change	changement climatique	2468	2	气候变化 的	Changement climatique	830
2	on the	sur	2143	2	的 问题	problème de	828
2	the world	le monde	1853	2	可 再生能源	énergies renouvelables	803
2	to be	être	1839	2	新 的	nouveau/nouvelle	758
2	for the	pour le	1720	2	的 影响	effets / influence	752
2	is a	est un	1586	2	我们 的	notre	733
2	at the	au	1560	2	上 的	sur	693
2	that the	que le	1530	2	可 持续	durable	684
2	from the	du / à partir de	1413	2	更 多	plus	659
2	will be	sera	1365	2	的 国家	pays	630

Le tableau 1 ci-dessus illustre les 14 segments les plus répétés de CLRG. Nous constatons que la fréquence de segments répétés du volet anglais est

beaucoup plus élevée que celle du chinois.
Par exemple, la fréquence du segment *climate change* est de 2 468 dans le volet anglais, tandis que dans le volet chinois, la fréquence est de 830.
La signification des segments répétés du volet anglais relève peu d'informations intéressantes. Les mots-outils ou les mots syntaxiques sont les plus répétés, un seul thème relatif à notre recherche est présent, *climate change*. En revanche, les segments répétés en chinois nous révèlent les véritables thèmes de notre recherche, *gaz à effet de serre*, *changement climatique*, *énergies renouvelables*, *nouveau/nouvelle*.
Nous pouvons dire que deux types de répétitions se manifestent : d'une part de mots grammaticaux pour l'anglais, et d'autre part, de mots de contenu pour le chinois. Rappelons que la forte répétition de mots grammaticaux est la cause du grand nombre d'occurrences en anglais. Plus l'emploi des mots grammaticaux est intensif, plus le nombre d'occurrences est important. Ce phénomène dissymétrique des segments répétés dans les deux volets est absolument normal, car la structure syntaxique des deux langues est complètement différente. Le fait d'avoir des traductions de l'un à l'autre ne prouve nullement l'emploi symétrique des segments qui se répètent de la même manière dans les deux langues. Cependant, un prétraitement de l'anglais pour éliminer les mots outils donnerait plus de sens à l'étude des segments répétés (Shen, 2016). Les remarques formulées par André Salem viennent étayer notre hypothèse, renforcée également par celles de Damon Mayaffre. « *L'analyse des voisinages récurrents permet d'utiliser les segments répétés pour documenter les analyses statistiques faites à partir des formes simples. On trouvera enfin une analyse typologique effectuée à partir des segments répétés.* » (Salem, 1986). «*Moins encore que la fréquence d'un mot, la récurrence de segments ne peut être naïvement attribuée au hasard : soit elle pointe une contrainte syntaxique, soit elle indique une détermination ou option sémantique. Dit rapidement, le mot est une unité graphique, le plus souvent ambiguë, sans sens explicite, pas même doté de signification. Le segment, lui, devient une unité linguistique porteuse de sens*» (Mayaffre, 2007).
Ces résultats de l'étude bilingue (anglais-chinois) des segments répétés parallèles ainsi que leurs analyses montrent que, pour une même information énoncée et décrite en deux langues, la répétition événementielle et thématique est plus saillante en chinois en raison de la faible pratique des anaphores (Shen, 2016). De plus le contenu est plus diversifié, puisque nous retrouvons nos principaux thèmes de recherche.
Nous abordons l'étude des segments répétés dans le corpus comparable ENRG, composé de trois sous-corpus : sous-corpus français ENRG-FR, sous-corpus américain ENRG-US, sous-corpus chinois ENRG-CN.

Corpus comparable : segments répétés trilingues (français, anglais/US, chinois)

Tableau 2 : segments les plus répétés du corpus comparable ENRG

ENRG-FR			ENRG-US				ENRG-CN			
longeur	segment répété	fréquence	longeur	segment répété	équivalent français	fréquence	longeur	segment répété	équivalent français	fréquence
2	de la	23101	2	of the	du	19529	2	减 排	réduire les émissions	12554
2	de l	19787	2	in the	dans le	13169	2	低 碳	faible teneur en carbone	10730
2	à l	7426	2	to the	au	6361	2	新 能源	nouvelle(s) énergie(s)	8753
2	à la	7244	2	on the	sur le	5339	2	气候 变化	changement climatique	8181
2	dans le	4975	2	and the	et le	4943	2	风电	l'énergie éolienne	7658
2	et de	4189	2	for the	pour le	4932	2	不 是	pas	6658
2	dans les	3972	2	that the	que le	4537	2	这 一	celui-ci	6409
2	sur le	3961	2	at the	au	3794	2	光 伏	photovoltaïque	6081
2	dans la	3720	2	from the	du	3326	2	大 的	grand	5995
2	que les	3600	2	to be	être	3277	2	的 是	ce que	5911
2	a été	3328	2	by the	par le	3185	2	的 一	l'un des	5628
2	sur les	3205	2	New York	New York	3137	2	也 是	aussi	5338
2	sur la	3092	2	in a	dans un	3095	2	就 是	est justement	5322
2	y a	2887	2	of a	d'un	2898	2	这 是	ceci est	5124
2	dans l	2870	2	he said	il a dit	2881	2	不 能	ne peut pas	5014
2	que le	2763	2	with the	avec le/la	2684	2	环境 保护	protection de l'environnement	4866

Le tableau 2 ci-dessus présente les 16 segments les plus répétés d'ENRG. Comme pour le corpus parallèle, notre premier constat est une répétition thématique particulièrement saillante pour le sous-corpus chinois (ENRG-CN). Par exemple, la fréquence du segment réduire les émissions est de 12 554, placé comme le segment le plus répété dans ENRG-CN, formes absentes dans le haut du tableau des deux autres sous-corpus. Cependant, ces formes existent, mais sont classées bien plus bas dans les résultats des segments répétés. Les autres sous-thèmes représentés par les séquences répétées comme *faible teneur en carbone*, *énergie éolienne*, *photovoltaïque*, etc., directement liés aux énergies et au changement climatique sont également mis en valeur dans le tableau 2. Pour les sous-corpus français et américain, seuls des mots grammaticaux ou mots-outils apparaissent dans les segments les plus répétés. Ce phénomène est dû essentiellement au mécanisme des anaphores ou au mécanisme déictique qui n'est pas le même en français et en anglais américain (Shen, 2016). Toutefois, nous remarquons qu'en chinois, ce sont des termes clés qui se répètent, tandis qu'en anglais et en français, il s'agit souvent d'entités nommées (noms propres, toponymes, etc.).

3. Conclusion

Dans le processus d'extraction de connaissances trilingues, nous pouvons conclure que les segments répétés mettent en lumière très efficacement les caractéristiques les plus saillantes en chinois que dans les deux autres langues occidentales. Deux types de répétitions se manifestent : d'une part des mots grammaticaux pour le français et l'anglais, et d'autre part, des mots de contenu pour le chinois.

De plus, nous soulignons que les cooccurrences ou poly-cooccurrences permettent également d'extraire des connaissances du corpus grâce à la

coprésence de formes éloignées. Selon Mayaffre, «*L'étude des segments répétés offre une alternative à la lemmatisation. Elle permet de désambiguïser les termes de manière formelle et surtout de manière endogène, en corpus et non en référence (arbitraire) au dictionnaire ou à la langue*» (Mayaffre, 2007).

A juste titre, en raison de la forte présence des mots-outils, les cooccurrences ou poly-cooccurrences par rapport aux segments répétés permettent de récupérer les séquences répétées non contigües au travers des phrases ou des paragraphes.

A partir des résultats des segments répétés des deux corpus, nous pouvons affirmer que l'outil Segments répétés présente l'avantage d'extraire rapidement des informations clés en chinois, alors qu'en français et en anglais, le mécanisme des cooccurrences et poly-cooccurrences met en valeur des informations non détectables par des moyens traditionnels (par exemple, les concordances).

Aussi, l'outil Segments répétés constitue un atout fondamental pour la fouille d'informations multilingues.

Bibliographie

Bonnafous S. and Tournier M. (1995). Analyse de discours, lexicométrie, communication et politique. In : Langages, 29e année, n°117, Paris, Larousse, pp. 67-81.

Habert B., Nazarenko, A., Salem A. (1997). Les linguistiques de corpus. Paris, Armand Colin/Masson, 254 p.

Habert B. and Zweigenbaum P. (2002). Problèmes épistémologiques : Régler les règles. TAL. Paris, Association pour le traitement automatique des langues, vol. 43, no3, pp. 83-105.

Lafon P. (1981). Analyse lexicométrique et recherche des cooccurrences. In: Mots, n°3, octobre 1981. Butor-Rousseau, Péguy, Presse du Zaïre, "la nouvelle droite", vocabulaires, communiste et socialiste, cooccurrences? pp. 95-148.

Lafon P. and Salem A. (1983). L'inventaire des segments répétés d'un texte. In: Mots, n°6, mars 1983. L'oeuvre de Robert-Léon Wagner. Vocabulaire et idéologie. Analyses automatiques. pp. 161-177.

Lamalle C. and Salem A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels » in A. Morin et P. Sébillot (éds), JADT 2002. Saint-Malo : IRISA-INRIA, vol. 1, 403-411.

Lebart L. and Salem A. (1995). Statistique textuelle

Longrée D., Luong X., Mellet S. (2004). « Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé » in G. Purnelle, C. Fairon, A. Dister (éds), JADT04. Louvain : Presses universitaires de Louvain, vol. 2, 743-752.

Longrée D., Luong X., Mellet S. (2006). « Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées » in J.-M. Viprey (textes réunis par), JADT' 06. Besançon : Presses universitaires de Franche-Comté, vol. 2, 643-654.

Mayaffre D. (2004). Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République. Paris : Champion.Mayaffre, Damon (2004). Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République. Paris : Champion.

Mayaffre D. (2007). L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan : Retour sur les travaux actuels de topographie/topologie textuelle. Lexicometrica, Andrée Salem, Serge Fleury, 2007, pp.1-12.

Rastier F. (2001). Arts et sciences du texte. Paris : Puf.

Salem A. (1986). Segments répétés et analyse statistique des données textuelles. In: Histoire & Mesure, 1986 volume 1 - n°2. Varia. pp. 5-28.

Salem A. (1987). Pratique des segments répétés. Essai de statistique textuelle, 1987

Shen L. (2016). Méthodes de veille textométrique multilingue appliquées à des corpus de l'environnement et de l'énergie : « Restitution, prévision et anticipation d'événements par poly-résonances croisées », Thèse : Sciences du langage, Université Sorbonne Nouvelle – Paris 3, octobre 2016, 474 p.

Viprey J. (2005-a). « Philologie numérique et herméneutique intégrative », in Adam J.-M. et Heidmann U. (éds.), Sciences du texte et analyse de discours. Genève : Slatkine, 51-68.

Viprey J. (2005-b). « Corpus et sémantique discursive : éléments de méthode pour la lecture d corpus », in A. Condamines (dir.), Sémantique et corpus. Paris : Lavoisier, pp. 245-276.

Viprey J. (2006). « Structure non-séquentielle des textes », Langages, 163, 71-85.

Misurare, Monitorare e Governare le città con i Big Data

Sandro Stancampiano
Istat – stancamp@istat.it

Abstract
Several new data sources are investigated in the production process of official statistics. This paper describes the results of the analysis of online reviews about four points of interest in Rome, Italy. The reviews, collected from the web using web scraping and data wrangling techniques, was written by tourists and visitors during the 2017. The general aim of this research is to extract useful information to help civil servants and citizens in decision-making processes. Within the activities related to this study were automatically collected and stored in a Data Base 9227 documents (each document is a review) used to build the corpora. The paper intends to classify the reviews and qualify the sentiment of the texts using tools and techniques of text mining.

Abstract
Numerose nuove fonti di dati vengono analizzate nel processo di produzione delle statistiche ufficiali. Questo documento descrive i risultati dell'analisi delle recensioni online su quattro punti di interesse della città di Roma, in Italia. Le recensioni, raccolte con tecniche di *web scraping* e *data wrangling*, sono state scritte da turisti e visitatori nel corso del 2017. Lo scopo generale di questa ricerca è di estrarre informazioni a supporto dei processi decisionali sia dei dipendenti pubblici sia dei cittadini. Tra le attività correlate a questo studio sono stati raccolti e archiviati automaticamente in una base di dati 9227 commenti utilizzati per creare un corpora analizzato utilizzando strumenti e tecniche di *text mining*. Il documento intende classificare le recensioni e qualificare il sentimento dei testi.

Keywords: big data, Internet as data source, text mining, cluster analysis, web scraping.

1. Introduzione

Questo progetto si propone di indagare soluzioni relative all'uso dei Big Data per produrre statistiche ufficiali a supporto della pubblica amministrazione. L'Istat ha incluso questo tema, condiviso a livello europeo, nel Piano

triennale della ricerca tematica e metodologica[1]. L'Istat sta considerando la possibilità di utilizzare i Big Data nel processo di produzione dei dati, in modo da attenuare il *trade-off* tra tempestività e accuratezza (Alleva, 2016).

2. Background della ricerca

Questo lavoro si focalizza sul tema della gestione dei beni culturali, indagando mediante tecniche esplorative multivariate (Bolasco, 2014) fonti dati non convenzionali. Si vogliono mostrare le enormi potenzialità dei dati presenti sul web per produrre statistiche al fine di ottimizzare i processi decisionali. Il risultato della ricerca potrà essere di ausilio agli amministratori nella gestione dei servizi dedicati ai fruitori dei beni culturali presenti sul territorio. L'esperimento, che si concretizza in un progetto pilota replicabile ed estendibile su ampia scala, utilizza l'analisi testuale (*text mining*) per estrarre informazioni da dati scaricati dal web mediante tecniche di *web scraping*. Si vogliono scoprire regolarità nei testi esaminati utilizzando la *cluster analysis* (analisi dei gruppi). Questa tecnica, applicata attraverso il software IRaMuTeQ, consente di definire la distanza tra gli oggetti che si vogliono classificare (Ceron et al., 2013).

3. Obiettivo e ipotesi di ricerca

Tra i molti siti web utilizzati dagli utenti per produrre contenuti, è stato scelto Tripadvisor. Gli utenti registrati utilizzano il sito per scrivere le loro recensioni sui luoghi in cui si sono recati condividendo le loro esperienze (Iezzi e Mastrangelo, 2012). Sono state scelte quattro tra le più celebri attrazioni della città di Roma frequentate quotidianamente da numerosi turisti (Colosseo, Pantheon, Fontana di Trevi e Piazza Navona). Il Colosseo con oltre sei milioni di visitatori ha determinato, anche per il 2016, l'incremento degli incassi garantiti dai musei italiani[2] e la supremazia della regione Lazio in questa graduatoria. Molti visitatori lasciano valutazioni relative ai luoghi aggiungendo considerazioni sullo stato di conservazione dei beni, sui servizi e i disservizi che hanno notato. Si ritiene che analizzando questi commenti, sia possibile dedurre preziose informazioni.

L'analisi ha permesso di ottenere una classificazione gerarchica delle recensioni basata sui termini caratterizzati da un utilizzo superiore alla media con riferimento alla variabile monumento.

[1] https://www.istat.it/it/files/2011/07/Piano-strategico-2017-2019.pdf (pp.27-28)

[2] http://www.beniculturali.it/mibac/export/MiBAC/sitoMiBAC/Contenuti/MibacUnif/Comunicati/visualizza_asset.html_892096923.html

4. Corpus e metodo

I commenti sono stati raccolti in una base dati mediante l'applicativo Diogene[3]: progettato con il paradigma OOA/D e realizzato con metodologia *agile* (Larman, 2005). Utilizzando lo stesso software è stato creato il corpus delle recensioni.Le 9227 recensioni raccolte, pubblicate dal 1 gennaio al 31 dicembre 2017, sono così suddivise: Colosseo 3483 (37.8%), Piazza Navona 1020 (11%), Fontana di Trevi 2829 (30.6%) e Pantheon 1895 (20.5%).

Si è proceduto in prima istanza con l'analisi lessicale ricavando informazioni utili alla successiva analisi testuale volta a localizzare unità di testo di rilevo per gli obiettivi del presente studio (Bolasco, 2013). L'analisi ha permesso di individuare gruppi di parole omogenei al loro interno ed eterogenei tra loro riguardo ai "concetti" espressi nelle recensioni. Il corpus analizzato è composto da 9227 testi, 1788819 occorrenze, 11891 forme, 366 hapax di cui il 3.08% relativi alle forme e lo 0.02% relativi alle occorrenze e media 193.87.

La ricchezza lessicale del corpus è molto bassa[4] (V/N*100 = 0.66%), difatti a fronte di un testo ampio si riscontra un vocabolario ridotto. Osservando le 30 forme attive con la frequenza assoluta maggiore, notiamo come il linguaggio utilizzato privilegi i sostantivi e gli aggettivi rispetto ai verbi. Gli aggettivi esprimono positività (*bello*, *bellissima*, *grande*) e i sostantivi sono legati alla fruizione dei beni oggetto di studio (*monumento*, *piazza*, *visita*, *luogo*, *consiglio*, *interno*) così come i verbi (*visitare*, *fare*, *vedere*, *dire*, *entrare*, *trovare*).

5. Gli scriventi e le recensioni

I dati relativi ai giorni della settimana in cui è stata scritta la recensione, evidenziano la tendenza degli utenti a mettere nero su bianco i dettagli delle loro esperienze nei giorni centrali della settimana, con una predilezione per i mercoledì (vedi Figura 5.1).

Le persone durante i fine settimana si dedicano alle visite dei beni culturali e preferiscono descrivere quanto visto e vissuto martedì, mercoledì e giovedì.

Nel periodo oggetto di studio le recensioni relative alle quattro piazze sono state in media 741 al mese con un minimo di 572 a giugno e un massimo di 1129 a gennaio.

Dalla Figura 5.2 risulta che i primi mesi dell'anno, da gennaio ad aprile, sono quelli in cui si concentra il maggior numero di recensioni (oltre il 42% del totale).

[3] Diogene è un software sviluppato in java per effettuare processi di *data wrangling.*

[4] Il calcolo è stato effettuato applicando la formula RL=V/N dove V = ampiezza del vocabolario e N = numero totale di parole nel testo.

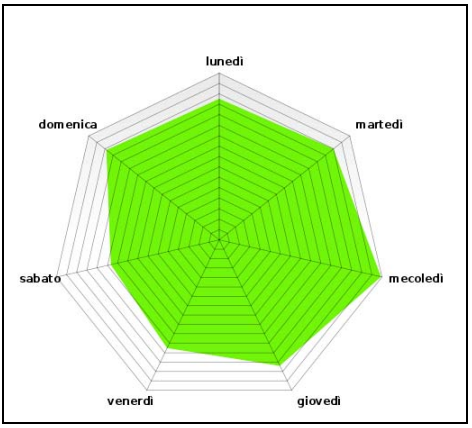


Figura 5.1: Numero di recensioni per giorno della settimana (gennaio – dicembre 2017)

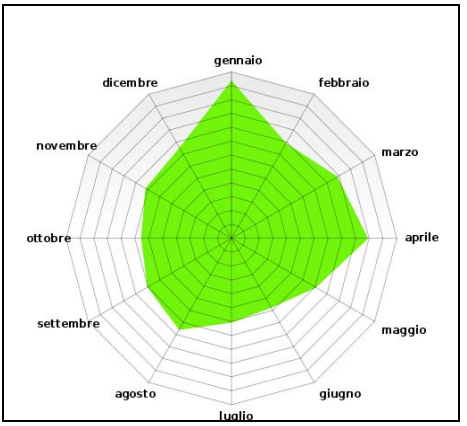


Figura 5.2: Numero di recensioni per mese (gennaio – dicembre 2017)

6. Cluster Analysis

La *cluster analysis* ci consente di raggruppare le unità statistiche massimizzando coesione e omogeneità delle parole incluse in ciascun gruppo e minimizzando al tempo stesso il legame logico tra quelle assegnate a gruppi/classi differenti.

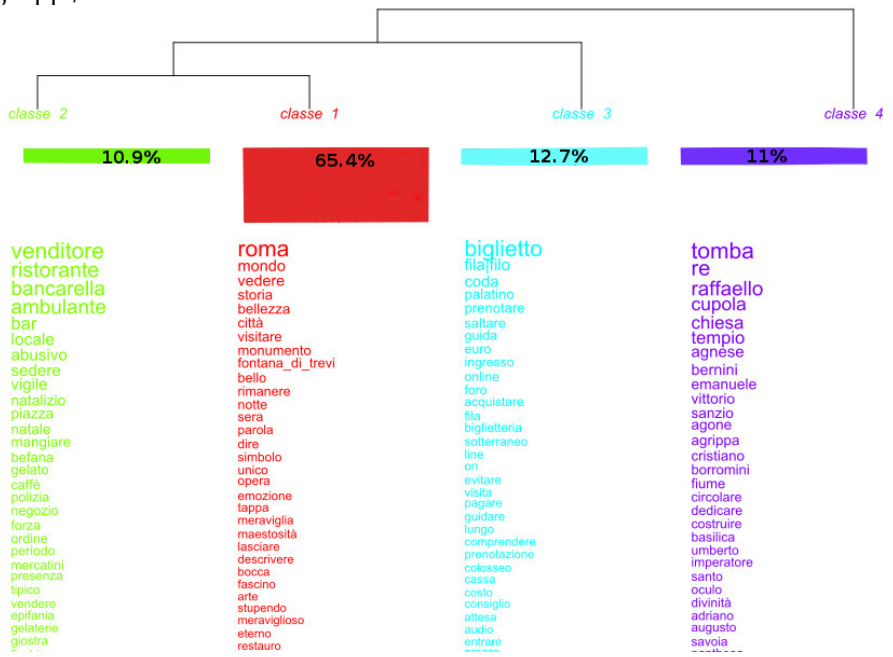


Figura 6.1: Dendrogramma delle classi secondo similarità

Il dendrogramma (Figura 6.1) mostra la divisione del corpus in 4 classi. Le parole contenute in ciascuna classe permettono di individuare le tipologie di argomenti trattati nel corpus, applicando la metodologia Alceste proposta da Max Reinert e implementata nel software IRaMuTeQ[5].

In Figura 6.2 osserviamo le parole appartenenti ai quattro gruppi e come si dispongono sul piano fattoriale. Questa visualizzazione chiarisce meglio il significato delle classi individuate.

Il gruppo di parole in rosso (65.4%), che si concentrano intorno all'origine, è composto dai termini più utilizzati: trasversali a tutto il corpus e di conseguenza a tutti e quattro i beni esaminati. Si tratta di parole tema come *roma, simbolo, monumento, città, storia,* dei verbi *visitare, vedere, tornare, dire* e di sostantivi e aggettivi come *bello, emozione, luce, bellezza* che esprimono positività e azioni legate alla visita.

La classe 2, in verde (10.9%), rappresenta i commenti pubblicati da persone che sono attente a quello che accade nei luoghi e considerano prioritaria la sicurezza, la legalità e la qualità dei servizi che trovano.

Si distinguono parole come *venditore, abusivo, presenza, peccato, fastidioso, ordine, municipale, polizia, fischietto.* Ci sono inoltre parecchi riferimenti alle attività commerciali (*bar, bancarella, locale, ristorante, gelateria, trattoria*) con particolare riguardo a cosa si può mangiare (*aperitivo*, *pizza, granita, gelato, vino*) e alle modalità di fruizione (*tavolino, tavolo, panchina*). Questo gruppo di parole evidenzia considerazioni che non sono strettamente correlate alla visita culturale ma piuttosto a tutto quello che ruota intorno a una escursione turistica.

La classe 3, in celeste (12.7%), rappresenta tematiche connesse ad aspetti economici e pratici che in alcuni casi possono causare disagio durante la visita. Emergono parole come *acquistare, prenotare, saltare, fila, coda, interminabile, biglietto, pagare, guida, audioguida, gratis, costo, euro, ticket.*

Gli argomenti sottesi sono relativi al costo del biglietto, all'attesa per l'ingresso e alla modalità della visita con connotazione sia positiva sia negativa a seconda della situazione particolare descritta dall'utente.

La classe 4, in viola (11%), rappresenta coloro che descrivono e raccontano l'esperienza dal punto di vista culturale citando eventi, luoghi e personaggi storici. Le parole più utilizzate sono *tomba, re, raffaello, sanzio, chiesa, colonna, fiume, barocco, agone, agnese, borromini, savoia, papa, pagano, cristiano.* Si tratta di riferimenti a luoghi di culto e opere (Sant'Agnese in Agone, la fontana dei Quattro Fiumi, le tombe dei re custodite nel Pantheon, ecc.), agli artisti che le

[5] IRaMuTeQ è un software realizzato per effettuare analisi multidimensionali di testi che fornisce una interfaccia grafica a R, altro software di elaborazione dati particolarmente efficiente per l'analisi di grandi dataset.

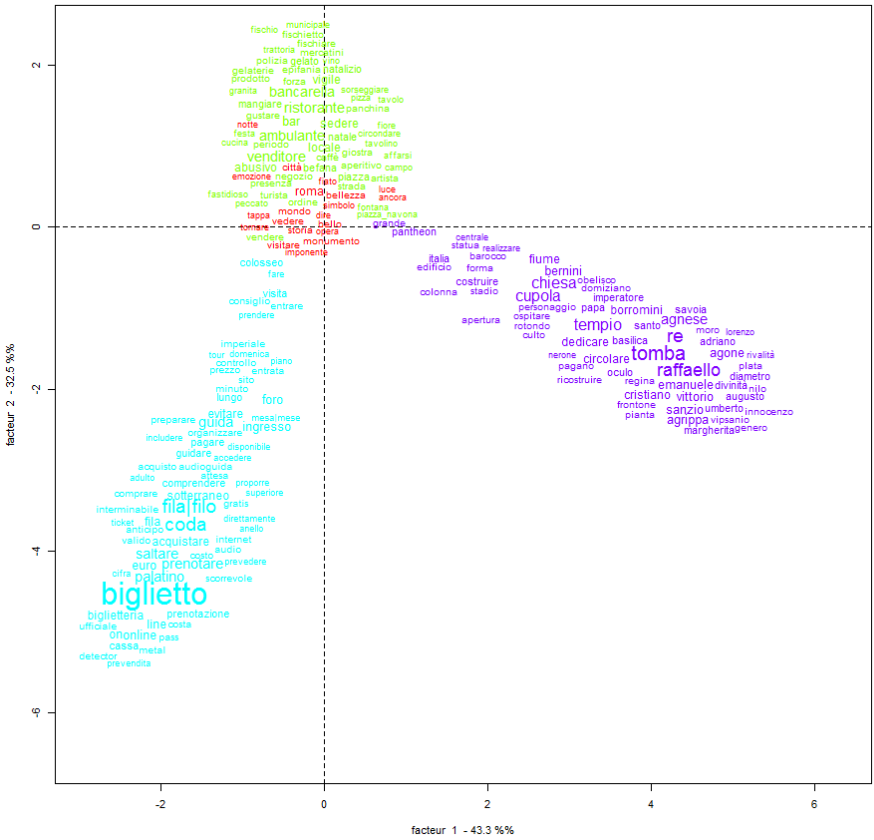


Figura 6.2: La disposizione delle parole sul piano fattoriale

hanno realizzate (Raffaello Sanzio e Borromini su tutti), alla storia e al contesto sociale e culturale di pertinenza dei siti visitati.

La disposizione dei termini sul piano fattoriale, a prescindere dai gruppi, evidenzia il continuum della visita, che inizia con la *prenotazione,* la *biglietteria* e il successivo *acquisto* seguito dalla *fila* per *entrare* e dalla constatazione della *bellezza* del *monumento* per poi *visitare* e immergersi negli aspetti artistici e nella storia del luogo in cui ci si trova.

7. Conclusioni e sviluppi futuri

Le tematiche palesate sono di sicuro interesse per gli amministratori pubblici, che possono ascoltare direttamente dalla voce dei cittadini quali sono i principali problemi dal punto di vista degli utenti. Sulla base di questo genere di analisi il decisore può valutare se e come intervenire per migliorare

la gestione dei luoghi e dei beni culturali.
Il flusso informativo parte dal cittadino che alla fine del processo può ottenere dei benefici tangibili grazie ai dati che lui stesso ha immesso in rete.
Il processo descritto in questo lavoro mostra un uso classico di Big Data: dati prodotti con una finalità specifica vengono utilizzati successivamente per raggiungere altri obiettivi apportando un innegabile valore aggiunto (Rudder, 2015).
Le tecniche di *text mining* applicate hanno permesso di valorizzare informazioni che altrimenti sarebbero rimaste inutilizzate.
Ulteriori e più approfondite analisi potranno essere condotte con la stessa metodologia e i medesimi software adoperati in questo lavoro. Si potrà continuare il monitoraggio, incrementando il corpus per condurre un'analisi longitudinale su questi stessi monumenti o studiare altre città e altri beni culturali al fine di migliorare le politiche di gestione e ottimizzare i processi decisionali.

References

Alleva G. (2016). Più forza ai dati: un valore per il Paese. *Relazione di apertura della 12° conferenza nazionale di statistica.*

Bolasco S. (2014). *Analisi Multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Carocci editore.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Carocci editore.

Ceron A., Curini L., Iacus S. M. (2014). *Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la Rete.* Springer Italia.

Iezzi Domenica F., and Mastrangelo M. (2012). Il passaparola digitale nei forum di viaggio: mappe esplorative per l'analisi dei contenuti. *Rivista Italiana di Economia, Demografia e Statistica,* 66 (3-4), pp. 143-150.

Larman C. (2005). *Applicare UML e i Pattern. Analisi e progettazione orientata agli oggetti.* Luca Cabibbo (a cura di), Pearson Education Italia.

Rudder C. (2015). *Dataclisma. Chi siamo quando pensiamo che nessuno ci stia guardando.* Mondadori.

Exploration textométrique d'un corpus de motifs juridiques dans le droit international des transports

Fadila Taleb[1], Maryvonne Holzem[2]
[1]Université Rouen Normandie – fadila.taleb@etu.univ-rouen.fr
[2]Université Rouen Normandie– maryvonne.holzem@univ-rouen.fr

Abstract
Within the framework of a research whose objective consists of responding to a need formulated by the IDIT, which helps to interpret the jurisprudential texts contained in its database, we are looking to highlight the interpretive paths considered as modal scenarios. We propose here a preliminary textometric analysis in order to define the linguistic profile of the corpus and to detect certain repeated segments that may represent a relevant constraint to complete and enrich the interpretive paths identified in the case law.

Résumé
Dans le cadre d'une recherche dont l'objectif consiste à répondre à un besoin formulé par l'IDIT[1], celui d'aider à l'interprétation des textes jurisprudentiels contenus dans sa base de données, nous cherchons à mettre au jour des parcours interprétatifs envisagés comme des scénarios modaux. Nous proposons ici une analyse textométrique préalable afin de cerner le profil linguistique du corpus et de détecter certains *segments répétés* pouvant représenter une contrainte pertinente pour compléter et enrichir les parcours interprétatifs identifiés dans les textes jurisprudentiels.
Keywords: textométrie, parcours interprétatif, scénario modal, segments répétés, motifs juridiques, droit des transports.

1. Introduction

1.1. Contexte

Dans le cadre d'un projet pluridisciplinaire « PlaIR »[2], des chercheurs informaticiens, linguistes, juristes posent la question de l'aide à l'interprétation[3] du fond jurisprudentiel de la base de données de l'IDIT. Du point de vue linguistique, notre tâche préalable à une implémentation consiste en l'étude de décisions de justice dans le but de comprendre leur

[1] Institut du Droit International des Transports.

[2] *Plateforme d'Indexation Régionale*

[3] Notre objectif est celui d'une aide instrumentée centrée sur l'agir de l'utilisateur *cf*. travaux du *groupe v* (Holzem et Labiche, 2017).

structure, le mécanisme argumentatif mis en œuvre et les mouvements de transformations textuelles susceptibles de déclencher des parcours interprétatifs pouvant aider à la lecture de ces décisions. Notre recherche s'écarte des modèles prédictifs, *justice prédictive ou legaltech*, qui, sous l'influence des *big data et du Machine Learning*, produisent des résultats de contentieux sur des bases algorithmiques. De ce point de vue, nous partageons les craintes de bon nombre de juristes de voir ces *legaltech* « *devenir eux mêmes une nouvelle forme de justice* » (Garapon, 2017). Il s'agit d'une pratique textuelle (et intertextuelle) comprise comme régime de transformation et d'interprétation. Dans cette perspective, notre recherche se place donc du côté de la *jurilinguistique* et son objectif est d'essayer de comprendre dans une approche linguistique et à travers l'étude du matériel textuel les décisions de justice et les stratégies argumentatives mises en œuvre pour ainsi aider à leur interprétation.

1.2. Questionnement et hypothèse

Pour aider à l'interprétation nous cherchons à cerner les stratégies argumentatives mises en œuvre par le juge, notamment dans sa manière d'intégrer et de prendre en charge les discours des autres (celui des parties du procès, celui des experts, celui du législateur etc.). Notre hypothèse est fondée sur des recherches antérieures (Holzem 2014 et Taleb 2014[4]) qui ont montré l'intérêt de la prise en compte des modalités linguistiques suivant le modèle développé dans (Gosselin 2010) pour la constitution d'un parcours interprétatif (Rastier 2001) envisagé ici comme scénario modal susceptible d'aider à l'interprétation. Mais avant de procéder à une telle analyse textuelle menée directement sur des textes pleins, nous avons eu besoin de cerner dans sa globalité et ses spécificités le profil linguistique de notre corpus d'étude. Pour cela, nous avons eu recours à une analyse textométrique approfondie, menée avec le logiciel TXM. Au fil de nos investigations textométriques, nous nous sommes rendu compte de l'importance de certaines fonctions offertes par ces outils pour la détection, par exemple, de *segments répétés*[5], qui peuvent représenter une contrainte pertinente pour compléter les parcours interprétatifs identifiés grâce à l'étude des modalités. L'objectif de cet article est de présenter, dans ses grandes lignes, en raison de la place, l'analyse textométrique menée sur notre corpus.

[4] Un mémoire de master 2 recherche en science du langage soutenu en juin 2014 : « *Étude du scénario modal et du syllogisme juridique pour la compréhension du processus de production du texte. Cas des textes du droit.* »

[5] *Suite de formes graphiques identiques attestées plusieurs fois dans le texte.*

2. Corpus et méthodologie

2.1. Description globale

Nous avons, à la suite de Rastier (2011), retenu le critère du genre comme critère définitoire du corpus de référence. Il regroupe des textes (décisions de justice) relevant du discours judiciaire[6] et appartenant au genre jurisprudentiel[7]. En reprenant la typologie du corpus proposée par B. Pincemin (1999) et reprise par (Rastier 2011), nous avons distingué quatre niveaux de corpus : (i) un corpus *existant/latent* (*archives* pour Rastier) qui correspond dans notre recherche à la base de données de *l'IDIT ;* (ii) un corpus de *référence* qui renvoie à l'ensemble des documents numérisés dans le fond jurisprudentiel de l'IDIT ; (iii) un corpus *d'étude* qui contient un nombre délimité de ces décisions sélectionnées pour les besoins de notre recherche et enfin (iv) un corpus *distingué (corpus d'élection ou sous-corpus* pour Rastier*)* correspondant à des passages précis des textes étudiés nommés « *les motifs* ». Ces derniers constituent le cœur du jugement, le juge exposant « *(…) les raisons de faits et de droit qui justifient la décision (…).*» (Cohen et Pasquino, 2013). Notre intérêt pour cette zone textuelle est doublement motivé. Premièrement notre objectif consiste à repérer les moments clés de transformations du jugement pour cerner les stratégies argumentatives mises en œuvre et partant aider à leur interprétation. Deuxièmement la motivation est une composante commune[8] à toutes les décisions de toutes les juridictions. Elle doit faire face à une double exigence : logique et persuasive. L'une est due à la forme syllogistique du raisonnement juridique imposée et l'autre à la nécessité de persuader l'auditoire de la décision[9] de sorte à éviter les recours et faire accepter la solution juridique apportée comme étant la seule possible.

---

[6] Il renvoie aux discours produits par (ou au sein) des juridictions. Il est à distinguer du discours *juridique* qui désigne, entre autre, les domaines du droit ou ses sources (lois, réglementation etc.). L'un concerne la création du droit, l'autre rend compte de son aspect applicatif.

[7]Le terme de jurisprudence renvoie ici à l' « *ensemble des décisions rendues par les tribunaux d'un pays, pendant une certaine période dans une certaine manière.* » (*Dictionnaire du vocabulaire juridique 2017*,éd. LexisNexis) P.322)

[8] Ce qui n'est pas le cas pour les autres composantes. Ainsi, *l'exposé du litige* ne figure pas dans les arrêts de la cour de cassation, car celle-ci étant une juridiction d'ordre suprême, son rôle est de veiller à la bonne application des normes juridiques, elle considère l'appréciation des faits par les juges de fond comme étant souveraine.

[9] Composée certes des parties du litige directement concernées par la décision, mais aussi les autres juges des autres juridictions et un public encore plus large, le destinataire universel.

2.2. Caractéristiques quantitatives

Le volume textuel du corpus d'étude est de 878848 occurrences dont 22456 formes. Le sous-corpus des motifs représente à lui seul près de la moitié des occurrences du corpus d'étude. Il contient 393092 occurrences pour 14742 formes. La dysmétrie de la distribution des formes dans les différentes zones délimitées montre l'importance et le rôle des motifs dans les décisions de justice, ils sont leur raison d'être, et tout juge est dans l'obligation de motiver son jugement**.**

2.3. Encodage et prétraitement

Notre corpus présente l'avantage d'être accessible en ligne. Cependant, l'ensemble des textes au format PDF n'est pas homogène : certains documents proviennent d'un format image (non *océrisé*[10]). Le format PDF n'étant pas pris en charge par TXM, nous avons tout d'abord procédé à une conversion (avec la technique d'océrisation pour les fichiers annotés et numérisés) au format TXT, puis dans un second temps à un codage XML en s'inspirant des recommandations de la TEI[11] pour l'encodage des données textuelles. Ce dernier nous permet une navigation plus fine dans le corpus grâce à des métadonnées péritextuelles, comme celles relatives au type de la juridiction *: tribunal de commerce (TC)*, *cour d'appel (CA)*, *cour de cassation (CC)*, à la date et au lieu, et des métadonnées intratextuelles, telles que celles relatives à des parties spécifiques dans les textes. Nous avons relevé quatre parties principales : *faits*, *moyens*, *motifs*, *conclusions*. *Les motifs et les conclusions* sont présents dans toutes décisions étudiées. *Les faits* sont absents des CC, et *les moyens* ne sont pas toujours indiqués comme tels dans les arrêts CA, ils sont souvent rappelés dans la zone des faits sous forme de discours indirect.
La figure suivante représente les différentes phases de préparation du corpus avant son traitement textométrique :

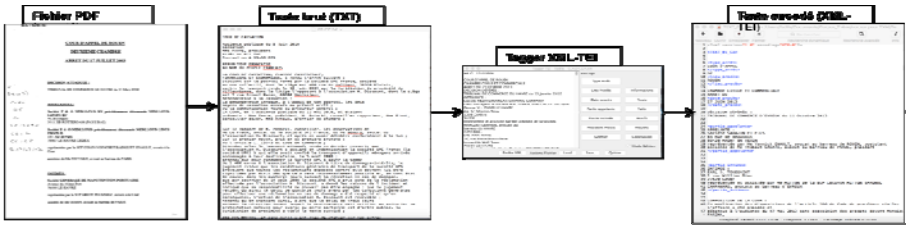


*Figure **1** Les étapes de préparation du corpus*

[10] OCR (*Optical Character Recognition*) Reconnaissance Optique de Caractères, étape nécessaire pour déchiffrer les formes et les traduire ici en lettres.

[11] *Text Encoding Initiative* : recommandations standard pour l'encodage des documents numériques.

Pour le passage du format TXT au format XML-TEI nous avons créé les balises spécifiques au genre du corpus étudié : <TypeJuridiction>, <DateProcès>, <LieuProcès> *etc*. Nous avons eu recours à un encodage semi-automatique au moyen d'un *tagger* conçu spécialement pour notre étude par Eric Trupin, MCF en informatique au laboratoire LITIS[12]. Cette étape indispensable de préparation du corpus pour le traitement textométrique a été à la fois chronophage et délicate : traitement des annotations manuscrites et nettoyage de documents plus anciens.

3. Exploration textométrique du corpus distingué : *la zone des motifs*

3.1. *Etude occurrentielle : les spécificités lexicales*

Une première étude contrastive au moyen d'un traitement textométrique phare, *le calcul de spécificités*[13], permet d'avoir une vue globale sur les caractéristiques lexicales du corpus distingué « *les motifs* ». Le tableau ci-dessus dresse la liste des 20 premières formes les plus spécifiques à cette zone. Il est trié par ordre décroissant sur l'indice de spécificité de celle-ci :

Unités	Fréquence T 828700	MOTIFS t=391346	score	CorpusR t=437354	score
Attendu	801	797	250,8	4	-250,8
que	16573	9373	129,9	7200	-129,9
Considérant	236	226	59,3	10	-59,3
Mais	163	163	53,1	0	-53,1
ne	7197	4031	50,3	3166	-50,3
est	4843	2778	45,1	2065	-45,1
attendu	181	170	41,3	11	-41,3
pas	4136	2372	38,5	1764	-38,5
moyen	396	308	34,9	88	-34,9
donc	802	530	26,4	272	-26,4
sera	516	356	22,9	160	-22,9
résulte	378	267	19,6	111	-19,6
Que	532	353	18,3	179	-18,3
succombe	59	58	17,4	1	-17,4
équité	51	51	16,6	0	-16,6
convient	240	177	16,2	63	-16,2
marchandises	1549	892	15,7	657	-15,7
suivant	156	123	15,3	33	-15,3
condamnera	55	53	14,6	2	-14,6
inéquitable	62	58	14,2	4	-14,2

Figure 2 : spécificité lexicales de la zone des motifs

Nous portons ici attention à un usage excessif d'occurrences caractéristiques du discours judiciaire et constitutif de la zone des motifs : *Attendu*, *que*, *Considérant*[14], *attendu*, de même pour les connecteurs : *Mais* et *donc*.

[12] Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, Université Rouen Normandie

[13] Le calcul de spécificités implémenté dans TXM repose sur la *loi hypergéométrique* développée par Lafon (1984). Le seuil de pertinence d'une distribution est fixé à 2 : +2 l'indice de spécificité est positivement significatif, -2 il est négativement significatif. L'indice se situant entre les deux est banal.

[14] Dans notre corpus la forme *Considérant* n'apparaît que dans les CA. Son absence dans les CC serait donc significative.

L'ensemble de ces marqueurs jouent un rôle spécifique ici, celui de ponctuer l'argumentation du juge en assurant sa progression syllogistique. L'usage excessif du futur, représenté avec les verbes être (*sera : 22,9*) et condamner *(condamnera : 14,6)* n'est pas surprenant, car avant de prononcer le verdict final dans un acte exclusivement directif (énoncé réservé à la zone des dispositifs), les juges avancent au préalable dans la zone des motifs les résultats (comme le montre d'ailleurs le suremploi du verbe *résulte (19,6))* de leurs argumentations : « *Le jugement entrepris* ***sera*** *confirmé en ses autres dispositions qui ne sont pas critiquées*».[15]; « *Le tribunal* ***condamnera*** *Monsieur le capitaine du […] ;».*[16] L'emploi significatif d'autres mots, comme *équité*, *marchandises*, *inéquitable* renvoie à la thématique des textes étudiés : *le droit des transports*. L'emploi significatif des adverbes de négation *: ne (+50,3), pas (+38,5)* révèle une caractéristique particulière de l'argumentation juridique car, fidèle au principe spinoziste *Determinatio negatio es,* la négation manifeste une valeur réplicative et résultative (*i.e.* portée référentielle en réponse à ce qui a été énoncé précédemment et qui n'a plus lieu d'être) préparatoire à la transformation juridique de l'énoncé.

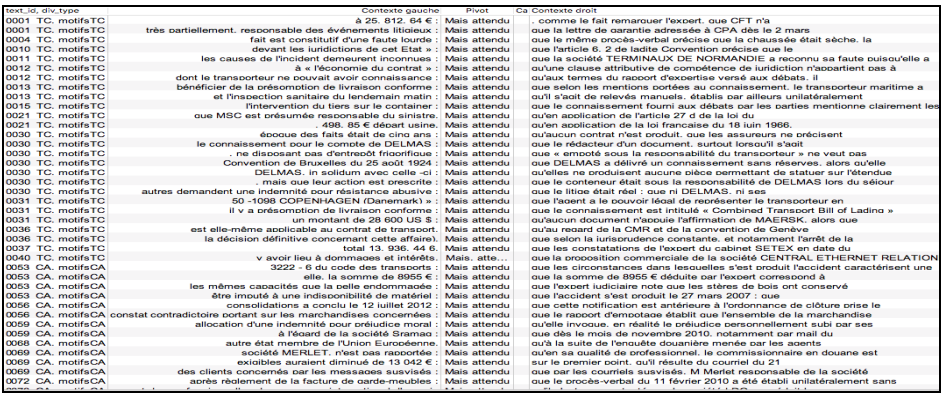
3.2. Etude contextuelle

Au-delà des investigations menées sur des unités lexicales minimales, les outils que propose la communauté ADT problématisent la notion de contexte selon des paliers différents pour privilégier un retour au texte. Nous allons ici donner l'exemple de la contextualisation des « *attendu* » dans la zone des motifs dont le suremploi a été relevé dans le tableau ci-dessus. Suite à une étude cooccurrentielle autour du mot-pôle « attendu », nous avons repéré une très forte attractivité avec le connecteur « *Mais* » (l'indice de spécificité[17]= +95).

[15] CA Rouen, 03/10/2013

[16] TC de Rouen, 15/12/2003

[17] Le calcul des cooccurrences qui repère les affinités et répulsions lexicales selon un indicateur de probabilité de rencontre repose sur le même modèle que celui du calcul des spécificités (Lafon, 1984).



text_id, div_type	Contexte gauche	Pivot	Ca	Contexte droit
0001 TC. motifsTC	à 25. 812. 64 € :	Mais attendu		. comme le fait remarquer l'expert. que CFT n'a
0001 TC. motifsTC	très partiellement. responsable des événements litigieux :	Mais attendu		que la lettre de garantie adressée à CPA dès le 2 mars
0004 TC. motifsTC	fait est constitutif d'une faute lourde :	Mais attendu		que le même procès-verbal précise que la chaussée était sèche. la
0010 TC. motifsTC	devant les juridictions de cet Etat » :	Mais attendu		que l'article 6. 2 de ladite Convention précise que le
0011 TC. motifsTC	les causes de l'incident demeurent inconnues :	Mais attendu		que la société TERMINAUX DE NORMANDIE a reconnu sa faute puisqu'elle a
0012 TC. motifsTC	à « l'économie du contrat » :	Mais attendu		qu'une clause attributive de compétence de juridiction n'appartient pas à
0012 TC. motifsTC	dont le transporteur ne pouvait avoir connaissance :	Mais attendu		qu'aux termes du rapport d'expertise versé aux débats. il
0013 TC. motifsTC	bénéficier de la présomption de livraison conforme :	Mais attendu		que selon les mentions portées au connaissement. le transporteur maritime a
0013 TC. motifsTC	et l'inspection sanitaire du lendemain matin :	Mais attendu		qu'il s'agit de relevés manuels. établis par ailleurs unilatéralement
0015 TC. motifsTC	l'intervention du tiers sur le container :	Mais attendu		que le connaissement fourni aux débats par les parties mentionne clairement les
0021 TC. motifsTC	que MSC est présumée responsable du sinistre.	Mais attendu		qu'en application de l'article 27 d de la loi du
0021 TC. motifsTC	. 498. 85 € départ usine.	Mais attendu		qu'en application de la loi française du 18 juin 1966.
0030 TC. motifsTC	époque des faits était de cinq ans :	Mais attendu		qu'aucun contrat n'est produit. que les assureurs ne précisent
0030 TC. motifsTC	le connaissement pour le compte de DELMAS :	Mais attendu		que le rédacteur d'un document. surtout lorsqu'il s'agit
0030 TC. motifsTC	. ne disposant pas d'entrepôt frigorifique :	Mais attendu		que « emporté sous la responsabilité du transporteur » ne veut pas
0030 TC. motifsTC	Convention de Bruxelles du 25 août 1924 :	Mais attendu		que DELMAS a délivré un connaissement sans réserves. alors qu'elle
0030 TC. motifsTC	DELMAS. in solidum avec celle -ci :	Mais attendu		qu'elles ne produisent aucune pièce permettant de statuer sur l'étendue
0030 TC. motifsTC	. mais que leur action est prescrite :	Mais attendu		que le conteneur était sous la responsabilité de DELMAS lors du séjour
0030 TC. motifsTC	autres demandent une indemnité pour résistance abusive :	Mais attendu		que le litige était réel : que ni DELMAS. ni ses
0031 TC. motifsTC	50 -1098 COPENHAGEN (Danemark) » :	Mais attendu		que l'agent a le pouvoir légal de représenter le transporteur en
0031 TC. motifsTC	il v a présomption de livraison conforme :	Mais attendu		que le connaissement est intitulé « Combined Transport Bill of Lading »
0031 TC. motifsTC	un montant de 28 600 US $:	Mais attendu		qu'aucun document n'appuie l'affirmation de MAERSK. alors que
0036 TC. motifsTC	est elle-même applicable au contrat de transport.	Mais attendu		qu'au regard de la CMR et de la convention de Genève
0036 TC. motifsTC	la décision définitive concernant cette affaire).	Mais attendu		que selon la jurisprudence constante. et notamment l'arrêt de la
0037 TC. motifsTC	total 13. 936. 44 6.	Mais attendu		que les constatations de l'expert du cabinet SETEX en date du
0040 TC. motifsTC	v avoir lieu à dommages et intérêts.	Mais. atte...		que la proposition commerciale de la société CENTRAL ETHERNET RELATION
0053 CA. motifsCA	3222 - 6 du code des transports :	Mais attendu		que les circonstances dans lesquelles s'est produit l'accident caractérisent une
0053 CA. motifsCA	elle. la somme de 8955 € :	Mais attendu		que la somme de 8955 € déduite par l'expert correspond à
0053 CA. motifsCA	les mêmes capacités que la pelle endommagée :	Mais attendu		que l'expert judiciaire note que les stères de bois ont conservé
0053 CA. motifsCA	être imputé à une indisponibilité de matériel :	Mais attendu		que l'accident s'est produit le 27 mars 2007 : que
0056 CA. motifsCA	consolidations a conclu le 12 juillet 2012 :	Mais attendu		que cette notification est antérieure à l'ordonnance de clôture prise le
0056 CA. motifsCA	constat contradictoire portant sur les marchandises concernées :	Mais attendu		que le rapport d'empotage établit que l'ensemble de la marchandise
0059 CA. motifsCA	allocation d'une indemnité pour préjudice moral :	Mais attendu		qu'elle invoque. en réalité le préjudice personnellement subi par ses
0059 CA. motifsCA	à l'égard de la société Sramao :	Mais attendu		que dès le mois de novembre 2010. notamment par mail du
0068 CA. motifsCA	autre état membre de l'Union Européenne.	Mais attendu		qu'à la suite de l'enquête douanière menée par les agents
0069 CA. motifsCA	société MERLET. n'est pas rapportée :	Mais attendu		qu'en sa qualité de professionnel. le commissionnaire en douane est
0069 CA. motifsCA	exiables auraient diminué de 13 042 € :	Mais attendu		sur le premier point. qu'il résulte du courriel du 21
0069 CA. motifsCA	des clients concernés par les messages susvisés :	Mais attendu		que par les courriels susvisés. M Merlet responsable de la société
0072 CA. motifsCA	après règlement de la facture de garde-meubles :	Mais attendu		que le procès-verbal du 11 février 2010 a été établi unilatéralement sans

Figure 3: Concordancier "Mais attendu" dan la zone des motifs

Nous avons remarqué une systématicité dans l'usage des « *Mais attendu* » qui vient clore un enchaînement de propositions subordonnées introduites par des « *Attendu que* », repris parfois par la conjonction *que*. L'étude approfondie des contextes de ce « *Mais attendu* » révèle une incidence particulière de celui-ci sur ses contextes droits :

« ***Attendu que*** *les marchandises ont été totalement perdues du fait de leur décongélation.* ***Attendu que*** *la première évaluation des marchandises a été établie à 18. 498, 85 € départ usine,* ***Mais attendu qu****'en application de la loi française du 18 juin 1966, le montant de la marchandise s'évalue en valeur CIF (coût + assurance + fret).* ***Attendu qu'en*** *l'espèce la valeur CIF des marchandises se monte à 21. 163, 96 €, c'est bien ce montant que le tribunal retiendra en préjudice principal. ».*

Dans l'extrait ci-dessus *Mais attendu* introduit non seulement un mécanisme de renforcement argumentatif[18], mais il joue également le rôle de déclencheur de transformation modale entre deux modalités de type axiologiques[19]. Dans l'exemple cité ici, le *Mais attendu* accompagné d'une référence juridique *« application de la loi française […]* assure cette transformation entre une norme liée au domaine du transport (*marchandises totalement perdues du fait de leur décongélation* : modalité axiologique négative) et les modes d'édiction d'une norme juridique cette fois. La marchandise dépréciée se trouve alors revalorisée (axiologique positif du point de vue juridique) par le changement des co-occurrents à droite (*valeur CIF (coût + assurance + fret)).*

[18] Voire les travaux pionniers de A. Ducrot (1984) sur les valeurs argumentatives de *Mais*.

[19] *Les modalités axiologiques sont propres aux jugements de valeur de nature morale, idéologique et/ou légale*. (Gosselin, 2010).

4. Conclusion

À travers cette contribution nous avons voulu montrer l'intérêt que représente une étude textométrique pour l'appréhension de son corpus d'étude. Si notre objectif principal, celui de mettre au jour des parcours interprétatifs nommés scénarios modaux (Taleb 2015), est difficilement envisageable en se limitant à une stricte étude textométrique (car elle repose sur l'étude modale propre à chaque texte). L'approche textométrique s'est avérée néanmoins pertinente pour décrire et cerner le profil linguistique du corpus. Son principe différentiel essentiel du point de vue sémantique, nous a incitées à adopter cette démarche d'analyse contrastive indispensable. L'analyse contextuelle à plusieurs paliers nous a permis le repérage de *constructions lexicales répétitives,* comme l'exemple des *« Mais attendu »* exposé ici, qui se révèlent être des moments clés du jugement et donc parcours interprétatifs corrélatifs à une transformation modale.

Références

Cohen M. et Pasquino P. (2013). *La motivation des décisions de justice, entre épistémologie sociale et théorie du droit. Le cas des Cours souveraines et des cours constitutionnelles*. CNRS, New York University, University of Connecticut.

Ducrot A. (1982). *Le dire et le dit*. Les Éditions de minuit, Paris.

Garapon A. (2017). Les enjeux de la justice prédictive. *La semaine juridique LexisNexis*, N°12: 47-52.

Gosselin L. (2010). *Les modalités en français. La validation des représentations*. Amsterdam-New-York : Rodopi B.V.

Holzem M. (2014). Le Parcours interprétatif sous l'angle d'une transformation d'états modaux, dans Numes Correia C. et Coutinho M. A. (eds), *Estudos Linguisticos : Linguistic studies* , n° 10, p. 283-295.

Holzem M. Labiche J (2017) *Dessillement numérique : énaction, interprétation, connaissances*. Bruxelles, Bern, Berlin : PIE Peter Lang.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Pincemin B. (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thése de Doctorat en Linguistique, Universit. Paris IV Sorbonne, chapitre VII.

Rastier F. (2001). *Art et science du texte.* Puf. Rastier 2011

Rastier F. (2011). La mesure et le grain. Paris, Éditions Honoré Champion.

Taleb F. (2015). Les modalités linguistiques pour aider à l'interprétation de textes juridiques. *Actes Interface TAL IHM (ITI'2015), 22ème Congrès TALn*, Caen.

The Framing of the Migrant: Re-imagining a Fractured Methodology in the Context of the British Media.

James M. Teasdale
Sapienza University of Rome - teasdale.1650019@studenti.uniroma1.it

Abstract 1
This study analyses the portrayal of migrants and migration in the British press over two periods, using frame analysis as a foundation methodology, while attempting to improve upon the methodology used in similar studies. The study holds the 'frame' to be the key organising feature in the portrayal of migrants and these frames can be located through a cluster analysis of textual data. The first aim of the work is to ascertain how far location and time affect the deployment of one frame or another, what these frames consist of and, therefore, provide a detailed analysis of how migration is portrayed in the British press: a focus sorely lacking in previous frame analysis studies to date. The study demonstrates that six frames can be identified over two periods; four being thematic and two being episodic. The 'negative' and 'positive' migrant frames were present in the first period, as the 'local' focus provided an ideal ground for the former's deployment as the subject was located closer to home and was depicted as a threat. While the second period saw the dominance of the 'positive' migrant frame with the death of Alan Kurdi and the corresponding conceptual shift to the 'global' removing the subject from the immediate border and placing them in a wider context. This was coupled with the overlap of the domestic responsibility frame with the 'positive' migrant frame as the two became intimately linked in the second period, while the European responsibility frame also arose. This demonstrated that the hegemony of one frame can be challenged but only if the corresponding situation is 'drastic' enough to allow.

Abstract 2
Questo studio analizza la raffigurazione dei migranti e della migrazione nella stampa britannica durante il corso di due periodi di tempo, utilizzando la teoria del frame analysis come metodologia di base e cercando di migliorare il procedimento di analisi utilizzato in studi analoghi. La ricerca pone il "frame" come principio organizzatore di base nella rappresentazione dei migranti. Questi frames possono essere rintracciati attraverso l'analisi clustering di dati testuali. Il primo scopo dello studio è quello di accertare

quanto posizione e tempistiche possano influenzare l'impiego di un frame rispetto ad un altro, in che cosa consistano questi frames e dunque fornire un'analisi dettagliata di come il processo migratorio venga descritto nella stampa britannica. Si tratta di un focus fortemente mancante negli studi basati sulla teoria del frame sino ad oggi. L'osservazione dimostra che, durante i sopra citati due periodi di tempo, sono sei i frame che possono essere identificati: si tratta di quattro di tipo tematico e due di tipo episodico. I frame "negativo" e "positivo" riguardo i migranti si possono rintracciare nel primo periodo, dal momento che il focus "locale" ha fornito un terreno ideale per l'impiego degli stessi. I soggetti erano infatti situati in prossimità del territorio ed erano dunque raffigurati come una minaccia. Al contrario, il secondo periodo di tempo vede il prevalere del frame "positivo" riguardo ai migranti, innescato dalla morte di Alan Kurdi e dal corrispondente slittamento concettuale che ha portato alla rimozione "globale" del soggetto dai confini immediatamente prossimi per ricollocarlo in un contesto più ampio. Questo si è appaiato al sovrapporsi del frame della responsabilità nazionale con il frame "positivo" riguardo ai migranti. Si può notare come i due frames siano diventati profondamente interconnessi durante il secondo periodo, proprio mentre si registrava l'insorgere del frame della responsabilità europea. Ciò dimostra come l'egemonia di un singolo frame possa essere sfidata, ma solo nel caso in cui la situazione corrispondente sia "drastica" al punto da permetterlo.
Keywords: migration, frame analysis, cluster analysis, British media, text mining

1. Introduction

1.1 Frame analysis and the migration crisis

Over the last two decades frame analysis has become an increasingly popular tool for analysing the portrayal of a subject in the media, due to its ability to demonstrate the latent and manifest meaning of the news and the recurring themes and elements that exist in common between individual texts (Zhongdang and Kosicki, 1993). According to Entman, 'framing essentially involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.' (Entman 1993). A reality is presented to the audience, a reality that can be considered a package of information of which the constituent parts together form the frame being deployed (Gamson et al. 1983). One frame is distinguishable from another precisely because this collective package is the sum of its parts. These parts are defined as framing

devices and reasoning devices, which can be discovered alongside one another thereby indicating the presence of one frame or another. These framing devices can consist of metaphors, visual images, lexical choices, stereotypes, idioms etc. (Tankard et al. 2001) which in turn support reasoning devices within the same frame which define the problem, assign responsibility, pass judgement and present possible solutions (Entman 1993).

As a relatively new approach, and apart from the shared inheritance from cognitive psychology (Bartlett 1932), anthropology (Bateson 1972) and the seminal work of Erving Goffman (Goffman 1974), frame analysis remains a fluid approach with a lack of empirical and methodological consistency across studies. Some authors have even contended if the school in of itself can even be considered a paradigm due to this diversity (D'angelo 2002:871; Entman 1993:51). This paper is not concerned with this contention, but does strive to arrive at a methodology which incorporates various elements of previous techniques in order to arrive at a complimentary approach which in turn minimises the criticism normally fired at more extreme approaches deployed in the past due to their perceived rigidity and shortcomings.

To date very little frame analysis has been directed towards migration, especially in the British context. Despite the migration crisis showing no signs of abating, the response of Europe has generally been categorised by two approaches; (i) strengthening internal and external borders to restrict movement throughout Europe (ii) disrupting attempted crossings by means of the Mediterranean. Britain is particularly interesting within this context, not only as a state which has consistently tried to curb entry at an official level, but also because of the media's and public's keen obsession with migration which was ultimately exemplified in the Brexit referendum. The media can be considered as central to this response. Whether one considers it to be the embodiment of public opinion or of elite opinion, it is nonetheless an incarnation of a country's position and can be seen as acting as an arbiter of said country's opinion. The current migration crisis is as complex as it is pressing, and the 'reality' presented by the media should not be seen as natural, ready to be recorded and transmitted from one human being to another, but rather as something that is constructed and then transmitted according to constructivist theory (Goffman 1974). The media is therefore able to set the agenda and frame the debate on the migration crisis, in turn affecting the reality in the mind of the population and government.

This paper has two aims in mind. The first is to develop a methodology which combines previous qualitative and quantitative approaches in order to improve validity and reliability while the second is to use said methodology to ascertain how migration is portrayed by the British media and how far this portrayal is affected by factors such as time and geographical focus.

2. Methodology

The study's methodology was constructed with historical criticisms directed at frame analysis in mind; either that the process is too qualitative and therefore lacks reliability, or that it is conducted too quantitively, and therefore lacks reliability. The first step was to collect the data, which was obtained manually from four daily British newspapers' online archives (the Daily Express, the Guardian, the Telegraph and the Daily Mail), and included all newspaper articles which included 'migration', 'migrant', 'refugee' etc. in the title, or whose content largely dealt with such topics. The two periods of investigation are 28th to 31st July 2015 and 2nd to 6th September 2015, these dates were chosen in order to ascertain whether frames could be consistently identified across two periods, even in the short term, but also to investigate whether dominant frames can be challenged if events are deemed drastic enough (the tragic death of Alan Kurdi became the dominant news story in the second period, whereas the first was primarily concerned with the Calais crisis). In total 505 were gathered, 160 for the first period and 345 for the second.

The quantitative aspect of the study consists of a computer assisted approach, by using cluster analysis to process the data and indicate the presence of 'frames'. Because, as mentioned above, framing is considered to be the grouping and salience of certain elements to the neglect of others, one can consider the cluster generated by a computer to precisely be a direct indication of the presence of one frame or another, as words are the primary form framing elements assume. The software used was the R program in conjunction with the Iramuteq interface. The clustering method used is that of Reinert (Reinert 1983), whose conception of clusters as a 'cognitive-perceptive framework' lends itself perfectly to frame analysis, concerned as it is with discerning different representations of a perceived reality. The second, more qualitative step of the study, was to conduct a deep read of all the texts, where the researcher intuitively coded texts and created a frame matrix which allowed an awareness of the context of the text as well as those framing and reasoning devices which seemed re-occurring and therefore significant. Combined, this allowed the reliability of the initial cluster analysis generated by the computer to be complemented by the in depth familiarity of the researcher, which provided a validity to the interpretation of results.

3. Results

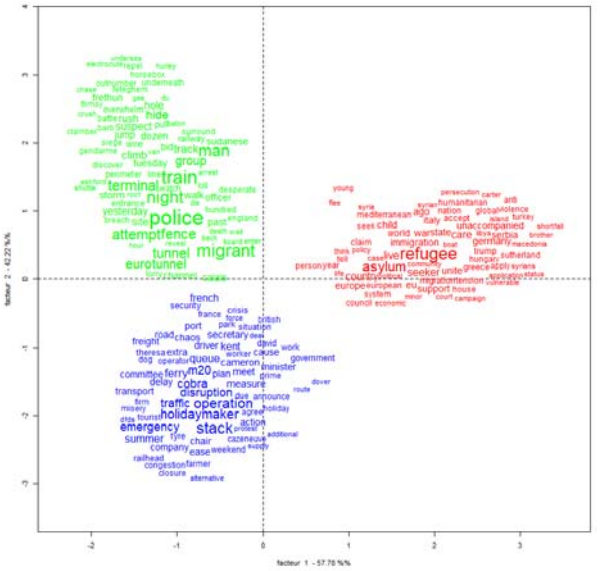


Figure 1. Cluster analysis for first period

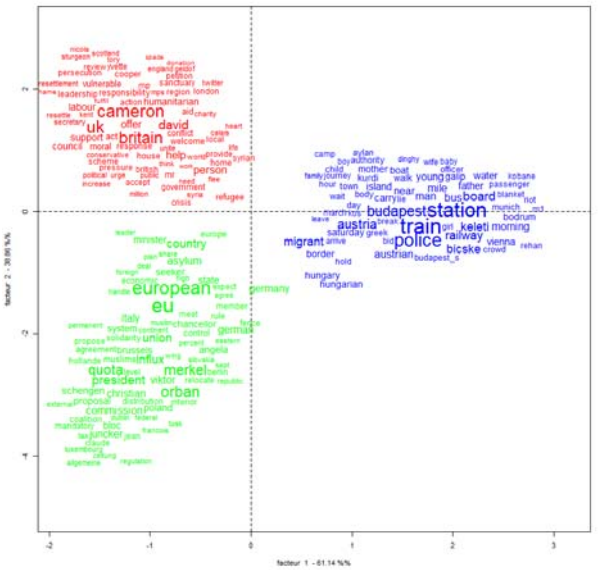


Figure 2. Cluster analysis for the second period

The two cluster analyses seem to identify three distinct clusters, yet those identified in the second period varying dramatically in respect to the first.

The first period under investigation generated three clusters, which have been labled *The Refugee Cluster* (Red), *The Migrant Cluster* (Green) and the *Calais Crisis Cluster* (Blue). However, the second period produced three different clusters: *Migration as a Domestic Issue Cluster* (Red), *Migration as a European Issue Cluster* (Green) and the *Migrant Crisis Cluster* (Blue). At first glance these results seem to refute the basis of framing theory; that frames are not produced by the journalist, but are deployed from the cultural repertoire they cognitively hold in common with the rest of society (Goffman 1974). This is because, if framing theory is correct, then in the space of one month it would be impossible for frames to mutate completely, and one would expect the clusters identified in the first period to be identical to those found in the second. However, if one makes a distinction between issue-specific and generic frames and episodic and thematic frames (de Vreese 2005) the two cluster groups are far more similar than first meets the eye.

For instance, the first period produced two frames which are predominantly concerned with the figure of the migrant and two differing portrayals of the migrant; the migrant as a helpless victim and the migrant as an opportunistic individual. These are both clusters which one can consider thematic frames as the clusters do not refer to one story but rather represent a thematic perspective. The third frame, however, can be categorised as being an issue specific frame, concerned as it is only with the Calais crisis, the 'Jungle' camp and the stories of migrants attempting to enter the channel tunnel. The second period, similarly, consists of two thematic frames (that which considers migration as an issue for the British government and that which considers it to belong to the realm of European governance) and one episodic frame (those stories relating specifically to the death of Alan Kurdi and those migrants attempting to move through Hungary and Austria in the early days of September 2015). If the two episodic frames are laid aside, one is left with four remaining; the *'negative' migrant frame*, *the 'positive' migrant frame*, the *domestic responsibility frame* and the *European responsibility frame*. What is interesting to note in the second period, is that 'positive' migrant frame from the first period does not disappear, but overlaps with and bolsters/is bolstered by the the arising domestic responsibility frame. For example, many of the key terms of the 'positive' migrant frame (vulnerable, refugee, conflict, persecution, support, receive, community etc.) are emblematic of those found in the so-called domestic responsibility frame (vulnerable, refugee, sanctuary, hazardous, save, help etc.) This means that rather than 'disappearing', the frame which represents migrants as individuals in need has been combined with arising domestic responsibility frame.

However, this does not account for the disappearance of the 'negative' migrant frame. The reason for this lack of presence, and likewise the merging

of the 'positive' migrant frame and the domestic responsibility frame in the second period, is due to the shock events linked to the tragic death of Alan Kurdi on September 2nd 2015. The event seems to have made the deployment of the 'negative' migrant frame untenable in the second period, while at the same time the 'positive' migrant frame persists as the period proved more fertile for this perspective. This is one reason why the two frames overlapped in the second period; the outrage and shock at the death of a toddler ultimately led to the locating of the solution to the 'positive' migrant frame in the domestic responsibility frame. Interestingly, this overlap did not occur with the European responsibility frame, which may be due to British political actors (the majority of those interviewed across the articles) actively positioning themselves as ready to help migrants in order to show themselves in a positive light.

Another interesting finding is how location affected or at least was linked to the change in hegemony between the 'positive' and 'negative' migrant frames. In the first period, the obsession with the Calais crisis (demonstrated by the presence of the corresponding episodic frame) seemingly provided conceptual ground in which the 'negative' migrant frame could flourish, whereas in the second period, dominated as it was by news of the death of Alan Kurdi (and the presence of a more international episodic frame) ensured the continued presence of the 'positive' migrant frame. One reason for this could be that as the migrant is located nearer to the British boarder, the 'negative' migrant frame (characterised by terms such as arrest, siege, repel, overwhelm) was more easily deployed due to the greater unease of foreign migrants entering the country, whereas when the focus was positioned more globally this unease was overcome by the moral shock of Alan Kurdi's death, lessening the unease and therefore the appropriateness of the previous frame.

Despite demonstrating some continuity of frames across the two periods, that geographical focus affects the deployment of one frame or another and that shock events can seemingly shift the frames in play to a great extent, the study is not without shortcomings. Firstly, the two time periods, and the limitation of four days to each, has greatly reduced the data available. This in turn makes it impossible to understand how far and how robust the identified frames are across an extended period of time and whether other frames come into play depending on the specific moment or the dominating news story. One solution could be to extend the time frame, but this might in turn lead to a drop in validity and insight due to the limitations of the researcher to deal with the data to the same extent as a computer. The second issue, as has already been mentioned, is determining precisely the characteristics of one frame in relation to another. One possible solution

would be to predetermine those terms which are identified as framing elements or reasoning devices as variables in the cluster analysis, which would in turn limit the identification of episodic frames in favour of thematic frames and over a longer period more clearly define the continuation, and the fluctuation in presence, of identified frames. The drawback of this, however, is that arguably the subjectivity of the researcher enters at too early a stage and harms the validity of the methodology. A third point is that, although the cluster analysis did capture many of the framing devices (as they are commonly exhibited as words), it was unable to capture all (for instance accompanying images) and was largely unable to identify the presence of reasoning devices (as the unit of analysis needs to be bigger than single word choice).

References

Bartlett, F. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* University of Chicago press.

D'Angelo, P. (2002). News Framing as a Multiparadigmatic Research Program: A Response to Entman. *Journal of Communication,* 52(4): 870-888.

Entman, R.M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication,* 43(4): 51-58.

Gamson, William A. and Kathryn E. Lash. (1983). The Political Culture of Social Welfare Policy. In S.E. Spiro and E. Yuchtman-Yaar, *Evaluating the Welfare State: Social and Political Perspectives*. Academic Press.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harper and Row.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2): 187-198.

De Vreese, C.H. (2005). News Framing: Theory and Typology. *Information Design Journal and Document Design,* 13(1): 51-62.

Zhongdang, P. and Kosicki G.M.. (1993). Framing Analysis: An approach to news discourse, *Political Communication,* 10(1): 55-75.

Tankard, J.W. and Severin W.J. (2001). *Communication Theories: Origins, Methods and Uses in the Mass Media, 5th Edition*. Pearson.

Results from two complementary textual analysis software (Iramuteq and Tropes) to analyze social representation of contaminated brownfields

Marjorie Tendero[1], Cécile Bazart[2]
[1] University of Rouen – CREAM and Agrocampus Ouest - marjorie.tendero@agrocampus-ouest.fr
[2]University of Montpellier, Montpellier – CEE-M - cecile.bazart@umontpellier.fr

Abstract
The aim of this paper is to demonstrate the complementarity of two types of textual analysis software, Iramuteq and Tropes, to analyze a corpus of data extracted from an open-ended question from a national cross-sectional survey. Descendant hierarchical classification made with Iramuteq lead to more homogeneous and less groups of discourse than the references fields made with Tropes. References fields allow to reveal how the corpus' thematic are articulated made with Iramuteq.

Résumé
Cette communication présente l'apport complémentaire de deux logiciels d'analyse de contenu, Iramuteq et Tropes, pour analyser les représentations sociales à partir de réponses données à une question ouverte dans un questionnaire d'enquête. Il montre que les classifications hiérarchiques descendantes opérées à l'aide du logiciel Iramuteq peuvent être approfondies de façon complémentaire à l'aide des classifications sémantiques par univers de références et l'outil scénario du logiciel Tropes. Les classes de discours sont moins nombreuses et plus homogènes que les univers de références mis en évidence par logiciel Tropes. Ces derniers montrent l'articulation des thématiques du corpus.
Keywords: Brownfield; Classifications; Iramuteq; textual data analysis; Tropes.

1. Introduction

L'analyse de contenu regroupe les techniques permettant une analyse systématique et objective des communications écrites et orales. Il s'agit d'une approche multidisciplinaire croisant des méthodes quantitatives et qualitatives, et dont les domaines d'application sont très nombreux : sciences de la communication, sociologie, psychologie, informatique, et économie par exemple. Ces techniques étudient la structure d'un texte, ou d'un discours,

ainsi que sa logique afin de mettre en évidence le contexte dans lequel il est produit, et sa signification réelle à partir de données objectives. Ces méthodes permettent de traiter les réponses à des questions ouvertes en soutenant l'interprétation du phénomène étudié sur des critères quantitatifs et objectifs (Garnier and Guérin-Pace 2010). Pour analyser les réponses données à des questions ouvertes, un des avantages de ces méthodes et d'éviter les biais liés à la codification thématique *a posteriori*. Toutefois, cette méthode fait l'objet de critiques. Ces dernières sont relatives aux étapes à mettre en place pour préparer le corpus, pour effectuer les analyses, et interpréter les résultats. Ainsi, lors de la phase de préparation du corpus, une lemmatisation peut être effectuée. Or, celle-ci regroupe parfois des formes dont l'emploi, dans un contexte donné, mène à des contresens (Lemaire 2008). C'est le cas lorsqu'une forme au pluriel est lemmatisée au singulier. De plus, les dictionnaires des expressions utilisés par les logiciels peuvent ne pas rendre compte des marqueurs de modalités comme la négation (Fallery and Rodhain 2007). Par ailleurs, des différences interprétées en termes d'analyse de contenu peuvent en réalité provenir de différences sociales dans la façon dont un individu s'exprime à l'oral ou à l'écrit. Les problèmes d'homonymies, de polysémies, de synonymies peuvent donc amener à construire des classes lexicales différentes alors qu'elles relèvent de modes d'expression hétérogènes sur la forme mais en réalité très similaire sur le fond ; ce qui est le cas des opinions exprimées par des périphrases, des paraphrases ou des ellipses. Une attention particulière doit donc être portée au traitement des ambiguïtés afin d'éviter toute erreur d'interprétation. Pour cette raison, il est intéressant de combiner deux approches complémentaires, et donc différents logiciels, d'analyse de contenu ; ce qui permet d'assurer la validité des résultats (Vander Putten and Nolen 2010; Lejeune 2017). C'est par exemple ce qui a été fait sur un corpus d'entretien pour comparer les logiciels Nvivo et Wordmapper (Peyrat-Guillard 2006). Dans cette communication nous soulignons l'apport complémentaire des logiciels Iramuteq et Tropes pour l'analyse des représentations sociales associées aux friches polluées à partir des réponses données à une question ouverte dans le cadre d'une enquête administrée au niveau national auprès de 803 individus résidant sur une commune impactée par ce type de foncier. Nous présentons dans la section qui suit la méthodologie adoptée, les données récoltées et les analyses effectuées. Dans une troisième section, nous présentons les résultats obtenus à l'aide du logiciel Iramuteq ; puis ceux obtenus à partir du logiciel Tropes dans une quatrième section. Nous discutons des apports complémentaires de ces deux logiciels pour l'étude des représentations sociales à partir de l'analyse des réponses données à une question ouverte dans une dernière section.

2. Méthodologie

Nous avons élaboré un questionnaire afin d'étudier la perception individuelle vis-à-vis du risque de pollution du sol, et les représentations, et perceptions relatives aux friches urbaines et à leur reconversion. Le questionnaire a été administré aux riverains résidant sur les communes impactées par une friche polluée[1]. Au total, 803 réponses complètes ont été collectées sur 503 communes impactées par la présence d'une friche polluée.

Pour analyser les représentations sociales, associées aux friches polluées, nous avons utilisé la question ouverte suivante : « à quoi associez-vous l'expression de friches urbaines ? ». Nous avons procédé à une analyse de données textuelles car cette technique d'analyse des données se prête particulièrement bien à l'étude des représentations, individuelles ou sociales, en rendant compte de la dynamique représentationnelle et cognitive d'un phénomène (Abric 2003; Beaudouin and Lahlou 1993; Kalampalikis 2005; Negura 2006).

Toutes les questions étaient obligatoires. Cependant, tous les participants n'ont pas réussi à y répondre : certaines réponses n'étaient qu'une suite de caractères permettant de passer à la question suivante. De plus, cette question ouverte se situait dans la seconde partie du questionnaire. Ce dernier était relativement long ; il en a résulté une perte d'attrition. Nous avons donc écarté ces réponses de notre analyse. Au total, 539 réponses ont pu être conservées ; soit 67,12 % des réponses collectées.

Les données ont été formatées pour pouvoir être analysées à partir du logiciel IRaMuteQ (Interface de R pour les analyses multidimensionnelles de textes et de questionnaires) version 0.7 alpha 2 dans un premier temps. C'est un logiciel libre développé par Pierre Ratinaud au sein du LERASS (Laboratoire d'Études et de Recherche Appliquées en Sciences Sociales) distribué sous les termes de la licence GNU GPL (v2) (Baril and Garnier 2015; Ratinaud and Déjean 2009). Le tableau 1 ci-dessous montre un extrait des réponses analysées.

Tableau 1 : Extrait du corpus analysé

0001	percept_eleve	affecte_non	prevent_non	gestion_non
	danger_oui	exist_non	gestfri_non	intention_oui
	confiance_non	sexe_h age_4059 reg_centre		
Abandonnée, sale, nuisible				
0002	percept_eleve	affecte_non	prevent_oui	gestion_non
	danger_oui	exist_oui gestfri_non	intention_oui	confiance_oui

[1] Ces communes ont été identifiées à partir d'une extraction de la base de données BASOL sur les sites et sols pollués (ou potentiellement pollués) appelant une action des pouvoirs publics, à titre préventif ou curatif.

	sexe_f	age_1924	reg_als		
Zones non_habité					
0003	percept_eleve	affecte_non	prevent_non	gestion_non	
	danger_non	exist_non	gestfri_non	intention_non	
	confiance_non	sexe_f	age_4059	reg_als	
Un jardin en ville, laissé à l'abandon.					
0004	percept_moyen	affecte_non	prevent_non	gestion_non	danger_non
	exist_non	gestfri_non	intention_oui	confiance_non	
sexe_f	age_4059	reg_rha			
zone abandonnée, zone polluée ville					

Le corpus de texte analysé a les caractéristiques décrites dans le tableau ci-dessous.

Tableau 2 : Statistiques descriptives associées au corpus analysé

	Corpus « friche »
Nombre de réponses	539
Nombre de mots (occurrences)	2 177
Nombre moyen de mots utilisés	4,04
Nombre de formes actives (total)	1 537
Nombre de formes supplémentaires (total)	640
Nombre d'hapax	275
Nombre de formes	482
Nombre de formes actives (différentes)	402
Nombre de formes supplémentaires (différentes)	80

Nous comparons les analyses suivantes : statistiques descriptives et classification hiérarchique descendante effectuée à l'aide du logiciel Iramuteq et univers de références et scénario à l'aide du logiciel Tropes. Il s'agit d'un logiciel d'analyse sémantique de textes créé en 1994 par Pierre Molette et Agnès Landré à partir des travaux de Rodolphe Ghiglione sur l'analyse propositionnelle de discours (Molette, Landré, and Ghiglione 2013).

3. Résultats de l'analyse avec Iramuteq

3.1. Statistiques descriptives

Le tableau ci-dessous décrit les termes les plus fréquemment employés par les individus (effectif ≥ 20) lorsqu'ils évoquent les friches polluées. Ces dernières sont des « terrains » (99 occurrences), des « zones » (36) laissées à « l'abandon » (106). Il s'agit de terrains sur lesquels étaient implantées d'anciennes « usines » (29) aujourd'hui « désaffectées » (17).

Tableau 3 : Termes les plus fréquemment employés (statistiques descriptives à partir du logiciel Iramuteq)

Formes actives	Effectif	Type	Forme active	Effectif	Type
Abandon	106	Nom	Usine	29	Nom
Terrain	99	Nom	Pollution	28	Nom
Laisser	63	Verbe	Ancien	28	Adjectif
Abandonner	49	Verbe	Espace	25	Nom
Ville	46	Nom	Bâtiment	25	Nom
Zone	36	Nom	Sol	20	Nom
Terrain vague	34	Nom			

3.2. Classification hiérarchique descendante

65.49 % des réponses données sont classifiées au sein de quatre catégories. Le tableau 4 ci-après indique la significativité des termes associés à chaque classe. La première classe regroupe les termes faisant référence aux anciennes activités industrielles. La deuxième classe renvoie aux problème de la gestion de déchets en milieu urbain en évoquant les « décharges », les « saletés », et la « pollution ». La troisième classe correspond aux termes caractérisant ce type d'espace. La quatrième classe, quant à elle, fait référence aux espaces de nature auxquels les friches correspondent, en particulier dans le cas de parcelles agricoles laissées en jachère.

4. Résultats complémentaires apportés par Tropes

Nous avons formaté le corpus pour l'analyser avec le logiciel Tropes. L'analyse des univers de références nous permet de mettre en évidence les principaux thèmes utilisés dans le texte en regroupant les termes dans des classes d'équivalent sémantiques. Le tableau 4 ci-après présente les résultats obtenus par les univers de références à l'aide du logiciel Tropes.

Les classifications sont données par ordre décroissant et indiquent le nombre de termes qui s'y rapportent. Ces classifications ne permettent pas toujours de couvrir l'ensemble des termes utilisés dans le corpus : seuls les substantifs les plus significatifs du texte y apparaissent. Il est toutefois possible de paramétrer ces classifications à partir du mode scénario du logiciel ; la figure 1 en montre un extrait.

5. Discussion et conclusion

Le tableau 6 précise les avantages et contraintes respectifs liés à l'utilisation de ces deux logiciels pour analyser les représentations sociales des friches polluées. En particulier, la classification sémantique par univers de références

et l'outil scénario font apparaître des classes plus nombreuses et moins homogènes que dans le cas de la classification hiérarchique descendante effectuée sous Iramuteq.

Tableau 4 Résultats de la classification hiérarchique descendante à partir du logiciel Iramuteq

Classe 1 (39,7 %)			Classe 2 (15 %)			Classe 3 (33,7 %)			Classe 4 (11,6 %)		
Anciennes activités industrielles			Problèmes de gestion des déchets en milieu urbain			Zone abandonnée et inutilisée			Espace agricole en jachère		
Forme active	χ^2	*p*	Forme active	χ^2	*p*	Forme active	χ^2	*p*	Forme active	χ^2	*p*
Abandonner	58,95	< 0,0001	Pollution	151,38	< 0,0001	Terrain	107,94	< 0,0001	Espace	114.47	< 0,0001
Usine	42,73	< 0,0001	Sol	59,79	< 0,0001	Abandon	84,27	< 0,0001	Nature	62.29	< 0,0001
Ancien	29,1	< 0,0001	Ville	32,66	< 0,0001	Laisser	82,57	< 0,0001	Vert	46.45	< 0,0001
Bâtiment	28,82	< 0,0001	Repos	17,13	< 0,0001	Friche	16,1	< 0,0001	Libre	41.31	< 0,0001
Industriel	22,13	< 0,0001	Désert	17,13	< 0,0001	Milieu_urbain	12,58	0,00038	Non_exploité	38.60	< 0,0001
Polluer	17,24	< 0,0001	Saleté	11,41	0,00073	Sauvage	10,6	0,00113	Champ	30.79	< 0,0001
Désaffecté	15,66	< 0,0001	Décharge	8,25	0,00408	Aller	5,95	0,01471	Non_entretenu	24.89	< 0,0001
Site	15,66	< 0,0001	Terre	7,85	0,00507	Non_utilisé	2 ,97	NS (0,08500)	Rntretenir	5.81	0,01596
Immeuble	14,05	0,00017	Culture	7,85	0,00507				Non_cultivé	2.71	NS (0,09962)
Zone	13,72	0,00021	Non_cultivé	4,06	0,04402						
Industrie	10,9	0,00096									
Lieu	10,87	0,0023									
Non_construit	9,29	0,00513									
Endroit	7,83	0,00547									
Vieux	7,72	0,00547									

Tableau 5 : Principaux univers de références associés au corpus

Univers de références 1			**Univers de références 2**		
Référence	**Eff.**	**Exemple de termes associés**	**Référence**	**Eff.**	**Exemple de termes associés**
Ville	74	Ville, taudis, zone urbaine	Ville	73	Ville, taudis, milieu urbain, zone urbaine
Lieu	59	Zone	Lieu	59	Site, zone, lieu
Habitat	55	Bâtiments, immeubles, logement, appartements	Industrie	50	Industrie, zone industrielle, usines

Industrie	50	Zone industrielle, industrie, usine	Immeuble	36	Bâtiments, immeuble
Écologie	39	Pollution, dépotoir	Pollution	33	Polluant, pollution, dépotoir
Plantes	33	Végétation, herbe, ronce	Déchet	22	Déchet, détritus
Déchet	22	Déchet, détritus	Agriculture	21	Jachère, cultures
Agriculture	22	Jachère, cultures	Terre	20	Sols, terre
Terre	20	Terre			

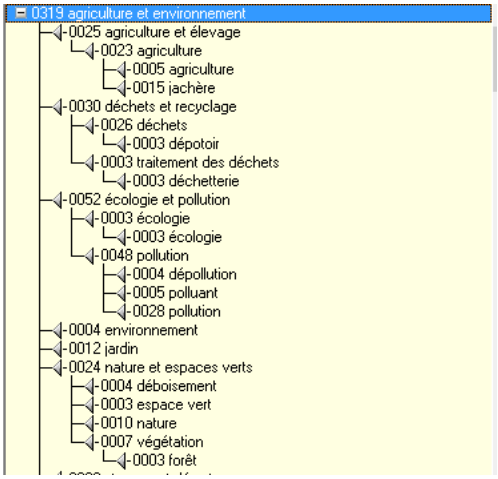


Figure 1 : Extrait des scénarios sous Tropes (ordre croissant)

Cet outil permet d'approfondir et de valider l'interprétation effectuée à partir de la classification hiérarchique descendante à l'aide du logiciel Iramuteq. Ces deux logiciels apparaissent donc comme complémentaires. Ces complémentarités restent toutefois à vérifier à l'aide d'autres type de corpus (entretiens par exemple). Enfin, pour étudier les représentations sociales de friches polluées auprès de populations impactées par ce type de site, il serait intéressant d'identifier le lexique émotionnel et affectif utilisée à l'aide d'EMOTAIX par exemple (Piolat and Bannour 2009). En effet, cela permettrait de mieux identifier la dimension affective dans les intentions comportementales à l'égard de ce type de site.

Tableau 6 : comparaison des fonctionnalités d'Iramuteq et de Tropes pour l'analyse des représentations sociales

Logiciels / Procédures	Iramuteq	Tropes
Découpage du texte	Segments de texte	Propositions canoniques
Style du texte		✓
Mise en scène		✓
Épisodes et rafales		✓
Classifications	Classification hiérarchique descendante	Univers de références
Scénario		✓
Statistiques descriptives	✓	✓
Analyse de similitude	✓	Indirectement par mots avec des graphes en aire ou étoilé
Analyse de spécificité et analyse factorielle des correspondances	✓	
Analyse prototypique	✓	
Principaux atout pour l'étude des représentation sociales	Richesse des analyses et des résultats	Formatage des corpus moins contraignant
Principaux inconvénients pour l'étude des représentations sociales	Formatage des corpus longs	Lemmatisation et classification automatisées aboutissent à des résultats peu lisible

References

Abric, Jean-Claude. 2003. *Méthodes D'étude Des Représentations Sociales*. ERES.

Baril, Élodie, and Bénédicte Garnier. 2015. 'Utilisation d'un outil de statistiques textuelles : IRaMuteQ 0.7 alpha 2. Interface de R pour les analyses multidimensionnelles de textes et de questionnaires'. Institut National d'Études Démographique.

Beaudouin, V, and S Lahlou. 1993. 'L'analyse Lexicale : Outil D'exploration Des Représentations'. *Cahier de Recherche* C (48): 25–92.

Fallery, Bernard, and Florence Rodhain. 2007. 'Quatre approches pour l'analyse de données textuelles :lexicale, linguistique, cognitive, thématique'. In *XVIème Conférence de l'Association Internationale de Management Stratégique*. Montréal, Canada.

Garnier, Bénédicte, and France Guérin-Pace. 2010. *Appliquer les méthodes de la statistique textuelle*. Les collections du CEPED (Centre Population et

Développement). Paris: CEPED.
Kalampalikis, Nikos. 2005. 'L'apport de la méthode Alceste dans l'analyse des représentations sociales'. In *Méthodes d'étude des représentations sociales*, edited by Jean-Claude Abric, 147–63. Hors collection. ERES.
Lejeune, Christophe. 2017. 'Analyser Les Contenus, Les Discours, Ou Les Vécus ? À Chaque Méthode Ses Logiciels !' In *Les Méthodes Qualitatives En Psychologie et Sciences Humaines de La Santé*, Dunod, 203–24. Psycho Sup.
Lemaire, Benoît. 2008. 'Limites de La Lemmatisation Pour L'extraction de Significations'. In *9ème Journées Internationales d'Analyse Statistique Des Données Textuelles*, 725–32. Lyon, France.
Molette, Pierre, Agnès Landré, and Rodolphe Ghiglione. 2013. *Tropes. Version 8.4. Manuel de référence*. http://tropes.fr/doc.htm.
Negura, Lilian. 2006. 'L'analyse de Contenu Dans L'étude Des Représentations Sociales'. *SociologieS* Théories et recherches (October).
Peyrat-Guillard, Dominique. 2006. 'Alceste et WordMapper : L'apport Complémentaire de Deux Logiciels Pour Analyser Un Même Corpus D'entretien'. In *Journées d'Analyse Statistique Des Données Textuelles*, 725–36. Besançon, France.
Piolat, Annie, and Rachid Bannour. 2009. 'EMOTAIX : Un Scénario de Tropes Pour L'identification Automatisée Du Lexique Émotionnel et Affectif'. *L'Année Psychologique* 109 (04): 655. https://doi.org/10.4074/S0003503309004047.
Ratinaud, Pierre, and Sébastien Déjean. 2009. 'IRaMuTeQ: Implémentation de La Méthode ALCESTE D'analyse de Texte Dans Un Logiciel Libre'. *Modélisation Appliquée Aux Sciences Humaines et Sociales MASHS*, 8–9.
Vander Putten, Jim, and Amanda L Nolen. 2010. 'Comparing Results from Constant Comparative and Computer Software Methods: A Reflection About Qualitative Data Analysis'. *Journal of Ethnographic and Qualitative Research* 5: 99–112.

Remerciements
Nous remercions Jean-Marc Rousselle pour avoir administré en ligne ce questionnaire sous Limesurvey. Cette enquête a bénéficié du soutien financier du SRUM 2015, de l'université de Montpellier, du CEE-M (LAMETA), de l'ADEME, de la Région Pays-de-la-Loire, et du CREAM (Université de Rouen).

Multilingual Sentiment Analysis

Matteo Testi[1], Andrea Mercuri[1,2], Francesco Pugliese[1,3]
[1]Deep Learning Italia – m.testi@deeplearningitalia.com
[2]Tozzi Institute – a.mercuri@deeplearningitalia.com
[3]Italian National Institute of Statistics – francesco.pugliese@istat.it

Abstract
In recent years, Sentiment Analysis (SA) has attracted significant attention in different areas of Research and Business. This is because "sentiments" can influence opinions of product vendors, politicians and the public opinion. The sentiments of users are generally categorised into three classes: negative, positive or neutral. Lately, more and more Deep Learning (DL) models have been employed to SA thanks to their automatic high-dimensional feature extraction capability. However, DL supervised models are greedy of data and the shortage of sentiment's data sets in specific languages (other than English) is a big issue. In order to address this multilingual issue of training sets we propose a very deep Recurrent Convolutional Neural Network model (RCNN) which achieves "state-of-art" accuracy in sentiment classification. Extracting keywords from the final max-pooling layer we are able to create a corpus of domain-specific keywords. By exploiting these "discriminative" extracted words we scrape a long sequence of sentences (in two different languages) in order to feed a Neural Machine Translation model. A sequence-to-sequence model with attention and beam-search has been implemented to translate one language sentences (i.e. English) into another language sentences (i.e. Italian). As example, we train our RCNN on an English twitter sentiment training-set and extract keywords to generate the machine translation model. During the test stage, we translate our test sentences (i.e tweets) into another language for which we have poor training set (i.e. Italian). Results highlight a significant accuracy gain of this technique with regard to a model exclusively trained on a poor training set expressed in a language different from English.
Keywords: sentiment, analysis, multilingual, deep, learning, recurrent, convolutional, neural, machine, translation

1. Introduction

In recent years, Sentiment Analysis (SA) has attracted significant attention in different areas of Research and Business. This is mainly due to the fact that "sentiments" (which are exhibited on the web by users) can affect opinions of product vendors, politicians and readers in general, namely the public

opinion. According to one of the most accredited definitions: Sentiment Analysis is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes (Qurat Tul Ain et al, 2017; Liu, 2012). This user point of view may usually be expressed under the unstructured form of an opinion, review, news, disapproval, etc. The rising demand of SA comes from the need of summarising a general direction of user opinions from social media (Haenlein et Kaplan, 2010). In fact, the aggregate data from Sentiment Analysis can represent a valuable information in order to orient decisions in politics, digital marketing or finance. Therefore, SA arises as a multidisciplinary field joining computational linguistics, information retrieval, semantics, natural language processing and artificial intelligence in general (Aydogan et Akcayol, 2016). Ultimately, SA can be seen as the process of automatically categorise utterances into three different classes: negative, positive or neutral. Generally these sequences of text or sentences come from social networks, opinion web-sites, e-commerce feedbacks, etc. Twitter is one of the most useful microblogging platforms for Sentiment Analysis and Opinion Mining since it offers very good API to download tweets and it is very popular amongst different categories of people (Pak et Paroubek, 2010). Traditionally, SA is a text classification problem and relies on two kinds of approaches: a) *"lexicon-based"* which is usually applied to problems without a training set. This technique generally makes use of a fixed number of keywords to orient the classification process by means of decision trees such as k-Nearest Neighbours (k-NN) or Hidden Markov Model (HMM); b) *"machine learning-based"* where extracted features typically consist of Parts of Speech (POS) tags, n-grams, bi-grams, uni-grams and bag-of-words. Classification can be performed by Naïve Bayes or Support Vector Machines (SVMs) (Singh et al., 2016). Traditional lexicon-based approaches are not effective anymore in combination with the modern textual Big Data corpuses, especially as far as sentiment concerns. On the other hand, Machine learning approach can be supervised and unsupervised (less common) and it is a methodology able of automation over enormous corpus of data, this is a critical requirement for a reliable Sentiment Analysis. Deep Learning is a branch of Machine Learning proposed by G.E. Hinton in 2006 and adopts Deep Neural Network for text classification (Hinton et Salakhutdinov, 2006). Deep Learning enhance traditional neural networks introducing more than thousands of neurons, millions of connections, new regularisation techniques (dropout, data augmentation, batch normalisation), new pre-processing (skip-gram, word embeddings, etc) and different new models both supervised and unsupervised: Convolutional Neural Networks (CNN)

(Krizhevsky et al., 2012), Deep Belief Networks (DBN) (Hinton et al., 2006). and many more. Lately, more and more Deep Learning (DL) models have been employed to SA thanks to their automatic high-dimensional feature extraction capability (Vateekul and Koomsubha, 2016). For instance, in Financial Sentiment Analysis (FinTech), Deep Learning has contributed to investigate how to harness different media and financial resources in order to improve the accuracy of stock price forecasting (Day et Lee, 2016). The experimental results show how news sentiment categorisation, by means of Deep Neural Networks, has different effects to investors and their investments. However, SA is a challenging field due to the lack of supervised data and to the nature inherently subjective of sentiments. In this work we tackle one of the biggest problems for modern machine learning-based Sentiment Analysis: the shortage of data sets in specific less common languages (Italian, German, etc.). In order to address the classification of sentiments we examined some of *"state-of-art"* text classifiers: many deep learning models have been employed in Sentiment Analysis previously, such as those invented by Stanford University: Recursive Neural Networks (RNNs) (Socher et al., 2011b) and Recursive Neural Tensor Networks (RNTNs) (Socher et al., 2013). Furthermore, Stanford released the Sentiment Treebank that is the first corpus with fully labeled parse trees to train RNTSs. RNTNs reach an accuracy ranging from 80% up to 85.4% on a Sentiment Treebank's test set. Although Recursive Models are very efficient in terms of constructing sentences' sentiment representations, their performance heavily depends on the performance of the textual tree construction. Constructing such a textual tree exhibits a time complexity of at least O(n^2), where n is the length of the text. For this reason, we decided to make use of a Recurrent Convolutional Neural Network model (RCNN) (Lai et al., 2015) achieving a rather competitive accuracy in sentiment classification with regard to Recursive Models. RCNNs exploit a recurrent structure to capture contextual information as much as possible when learning word representations, which may introduce considerably less noise compared to traditional window-based neural networks. Moreover, the benefit of exhibiting a time complexity of O(n) is a big added-value of RCNNs. To provide the support to a Multilingual Sentiment Analysis, a Neural Machine Translation (NMT) model has been employed in order to translate one language sentence (i.e. English) into another language sentence (i.e. Italian). Basically, a NMT model is a Neural Network structured in an encoder-decoder pattern which turned out as a competitive alternative to the traditional Statistical Machine Translation (SMT). The encoder consists of two independent recurrent networks: *"forward"* which reads the the sentence in the natural order and *"backward"* which reads the sentence in reverse order. Instead, the decoder is

an RNN capable to compose the sentence to be translated. This sequence-to-sequence model can be trained on a training set made of pairs of sentences: the first is expressed into the source language and the second into the target language (Cho et al., 2014).

2. Materials and Methods

The novelty of our Recurrent Convolutional Neural Network, with respect to the original paper, is that we introduced two new recurrent models called Long Short Term Memories (LSTM) instead of simple RNNs. These two LSTM bi-directionally scan the text. The topology of the RCNN (see Fig. 1) is intentionally designed to capture the context of each word (see the original paper for further details). The RCNN has been trained on a corpus of 1.6 million tweets composed from various Semeval training-sets (Strapparava et Mihalcea, 2007) and divided into positives (800k) and negatives (800k). To input textual sequences into the neural network we insert a pre-trained embedding layer on top (Mikolov et al, 2013). The embedding layer, which has been pre-trained on an English Wikipedia Corpus, transforms indexed words into numerical vector. Embedding vectors are characterised by a semantical relationship amongst them according a chosen metrics, a cosine distance in this case. Size of embedding vecotrs is 300.

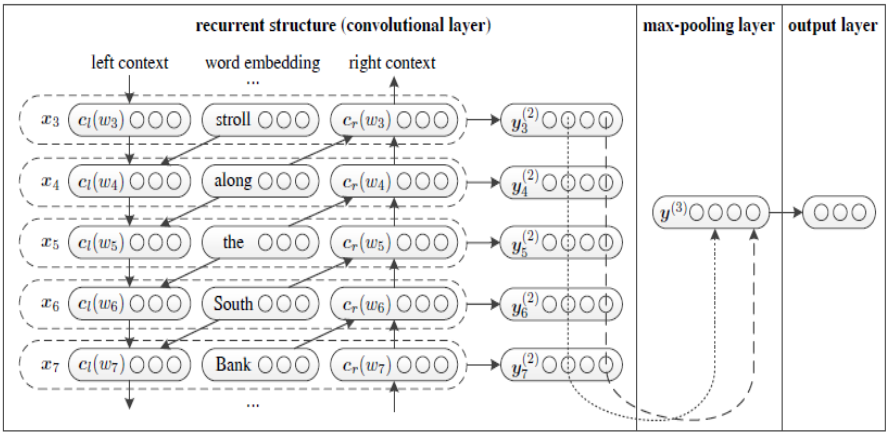


Figure 1. The structure of the RCNN scanning the sentence "A sunset stroll along the South Bank affords an array of stunning vantage points" (Lai et al., 2015).

During the training stage, the RCNN achieves 84% of accuracy on a validation set (selected at the 20% of the original dataset). On a test set of 380 tweets (provided by Semeval), the model returns around 82% of accuracy on positive tweets and 78% of accuracy on negatives, with an approximative

80% overall on a mixed tweets set. We followed recomended settings within the original paper for the hyper-parameters selection.
Finally, we have modified the RCNN in order to extract the most significant keywords that are specific for the model to drive the sentiment classification. Basically, the third layer, that is the max-pooling layer, relies on an element-wise *"max"* function as follows :

$$y^{(3)} = \max_{i=1}^{n} y_i^{(2)}$$

The most "*discriminative"* words for the sentiment classification are those most frequently selected in the max-pooling layer. Hence, we extracted the indices of words corresponding to the max values of activation identified within the third layer. During the training we determined 3.2 millions of keywords, namely 2 for each tweet, the most important and the second in order of signinifancy. Many of the resulting keywords come duplicated or altered for multiple reasons : they might belong to a common slang or undergo typing errors. Then, we removed doubles and we matched the rest with the embedding corpus cointaining 2.5 mln words of the English Language. This process turned out with 85,000 correct english keywords. By exploiting these keywords as seed, we scraped a long sequence of sentences in English from a website of Contextual Translations such as *"Reverso Context"* (context.reverso.net) and its Italian translation in many different form of expression. This stage led to a training set of 800,000 pairs of sentences English-Italian and a Validation set of 50,000 pairs. A multi-level sequence-to-sequence model with attention and beam-search has been implemented to be trained on the training set of pairs (see Fig. 2) (Bahdanau et al., 2014; Luong et Manning, 2016).

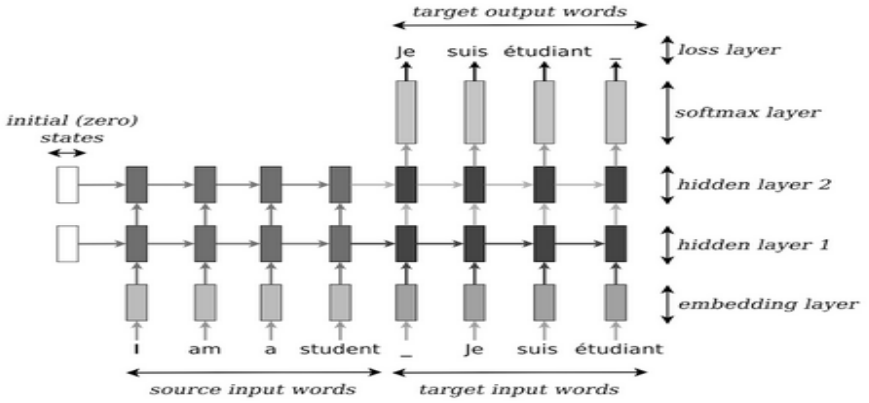


Figure 2. Multiple levels encoder-decoder (Luong et Manning, 2016).

"Attention-based" models enable the decoder to "focus" specifically on some words rather than others, selectively orienting towards a more efficient combination of words within the destination language sentences (Chorowski, et al., 2015). *"Beam search"* is a greedy algorithm maximising the probability of the ouput words (Britz et al., 2017). The NMT model was trained with an embedding matrix randomly initialised and trained within the same process. Embedding vectors size was 512. Both encoder and decoder are made of two LSTM cells with an hidden state size equal to 512. Training algorithm was the Stochastic Gradient Descent (SGD) with 32 sized batches; initial learning rate of 1 and a decay factor of 0.5 starting from the 5-th epoch, plus early stopping to reduce the overfitting. Beam search amplitude has been set to 5. In Fig.3 they are reported some resulting translations from Italian to English, on a test example.

```
adottare un vocabolario condiviso è un suggerimento perfetto su come scrivere frasi comprensibili
adopt a shared vocabulary is a perfect suggestion on how to write understandable sentences

un altro suggerimento su come scrivere frasi semplici: evita le negazioni inutili
another suggestion about how to write simple sentences : avoid unnecessary <unk>

quasi 90 persone sono morte per una tempesta tropicale nelle filippine
nearly 90 people died for a tropical storm in the philippines
```

Figure 3. Some translations from Italian to English by means of the neural model trained by us.

In the same time, we have trained the RCNN model on the most popular Italian Sentiment Polarity Training set of tweets called SentiPolc 2016 (Barbieri et al., 2016). which is made of 7,000 annotated tweets and 300 test tweets. In this case (Italian language) our model reaches 45% of validation set accuracy and 43% on test set. For the embedding layer we have adpoted a pre-trained language model on an Italian Wikipedia Embedding Corpus.

3. Results

We have tested the English RCNN model on the same italian SENTIPOLC 2016 test-set translated into English by our neural machine translation model. Results highlight a boost of performance : **78%** of accuracy on the test set versus the **43%** of the Italian trained RCNN model proving our strategy of stacking NMT and RCNN models is successful.

4. Conclusion

Despite of the imperfections of the Neural Machine Translation producing translations with some errors, the RCNN is tolerant to minimal errors and is

able to hold the accuracy to high levels on a test set. This is because RCNN was previously trained on a solid and huge English corpus of tweets. This entire process of keywords extraction, specifically to the task of sentiment classification from the training set, is a fully novel approach to tackle the problem of the lack of Sentiment training sets in other languages. Keywords allow generating a domain-specific training set for the Neural Machine Translation. Arguably, we believe this way of stacking NMT and RCNN lead to a cutting-edge Multilingual Sentiment Classifier that can benefit other fields of Text Classification in future. Future directions might be towards a closer integration of NMT and Text Classifier and a reduction of translation errors.

References

Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat and A. Rehman (2017). *Sentiment Analysis Using Deep Learning Techniques: A Review.* International Journal of Advanced Computer Science and Applications (ijacsa).

Haenlein, M., and Kaplan, A. M. (2010). *An empirical analysis of attitudinal and behavioral reactions toward the abandonment of unprofitable customer relationships.* J. Relatsh. Mark.

Aydogan, E. and Akcayol, M. A. (2016). *A comprehensive survey for sentiment analysis tasks using machine learning techniques.* Int. Symp. Innov.

Liu, B. (2012). *Sentiment analysis and opinion mining (synthesis lectures on human language technologies).* Morgan & Claypool Publishers.

Pak, A., and Paroubek, P. (2010, May). *Twitter as a corpus for sentiment analysis and opinion mining.* In LREc (Vol. 10, No. 2010).

Singh, J., Singh, G., and Singh, R. (2016) *A review of sentiment analysis techniques for opinionated web text.* CSI Trans. ICT.

Hinton, G. E., and Salakhutdinov, R. R. (2006). *Reducing the dimensionality of data with neural networks*. science, 313(5786), 504-507.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* In Advances in neural information processing systems (pp. 1097-1105).

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). *A fast learning algorithm for deep belief nets. Neural computation.* 18(7), 1527-1554.

Vateekul, P., and Koomsubha, T. (2016, July). *A study of sentiment analysis using deep learning techniques on Thai Twitter data.* In Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on (pp. 1-6). IEEE.

Day. M., and Lee C. (2016) *Deep Learning for Financial Sentiment Analysis on Finance News Providers*. no. 1, pp. 11271134.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. 2011b. *Semi-supervised recursive autoencoders for predicting sentiment distributions*. In EMNLP, 151–161.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP, 1631–1642.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). *Recurrent Convolutional Neural Networks for Text Classification.* In AAAI.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint arXiv:1406.1078.

Strapparava, C., and Mihalcea, R. (2007, June). *Semeval-2007 task 14: Affective text*. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 70-74). Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473.

Luong, M. T., and Manning, C. D. (2016). *Achieving open vocabulary neural machine translation with hybrid word-character models*. arXiv preprint arXiv:1604.00788.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). *Attention-based models for speech recognition.* In Advances in Neural Information Processing Systems (pp. 577-585).

Britz, D., Goldie, A., Luong, T., and Le, Q. (2017). *Massive exploration of neural machine translation architectures*. arXiv preprint arXiv:1703.03906.

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016, December). *Overview of the EVALITA 2016 SENTiment POLarity Classification Task.* In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016).

A linguistic analysis of the image of immigrants' gender in Spanish newspapers

Juan Martínez Torvisco
Universidad de la Laguna – jtorvisc@ull.edu.es

Abstract 1 (in English)
The phenomenon of immigration has been studied from diverse perspectives is important to understand that immigration is a fact associated with times of crisis. The reason for the avalanche of immigrants to the Canary Islands (Spain) is because it is the gateway to Europe, and therefore, immigrants want to enter from this point. This research arises from the need to linguistically determine the treatment of the phenomenon of immigration in the Spanish press as a result of the arrival of thousands of foreign citizens to the coast of the Canary Islands in 2006 and in 2015. It attempts to analyse four Spanish newspapers using Iramuteq qualitative analysis software, two from the Canary Islands (El Día and Canarias 7) and two Spanish national newspapers (El País and ABC). Also, we wanted to know how it is the informative treatment of gender. Our hypothesis is that the word male (immigrant) appear more than woman and on the contrary woman (refugee) has a higher frequency than male. Results are presented on a dendogram figures.

Abstract 2 (in Spanish)
El fenómeno de la inmigración se ha estudiado desde diversas perspectivas, y es un hecho asociado a tiempos de crisis. El motivo de la avalancha de inmigrantes en las Islas Canarias (España) se debe a que es la puerta de entrada a Europa y, por lo tanto, los inmigrantes quieren entrar desde esta parte de Europa, buscando una major vida. Esta investigación surge de la necesidad de determinar lingüísticamente el tratamiento del fenómeno de la inmigración en la prensa española como resultado de la llegada de miles de ciudadanos extranjeros a la costa de las Islas Canarias en 2006 y 2015. Se analizan cuatro periódicos españoles utilizando el software Iramuteq de análisis cualitativo, dos de ámbito regional de Canarias (El Día y Canarias 7) y dos periódicos de ámbito nacional (El País y ABC). También queríamos saber cómo aparece el género en las noticias de estos diarios. Nuestra hipótesis es que los inmigrantes son mayoritariamente hombres por tanto debe aparece más que la mujer y al contrario, la palabra mujer (refugiada) tiene una frecuencia mayor que la del hombre. Los resultados se presentan

dos figuras de dendograma con el Análisis Jerárquico Descendiente (DHC) y reflejan que la mujer aparece en 2015 pero no está presente en las noticias de los diarios en 2006 y a la inversa ocurre con el hombre.Keywords: a set of keywords describing the content of the paper.

1. Introduction

The media have become a powerful tool to make visible conflicts, or show realities that sometimes remain hidden from the world. Such a fact seems unquestionable. One of the most-recent cases are the so-called "immigration crisis" or the "refugees' crisis," it began before the dates analyzed in the current research, however, achieve an uncertain projection until these citizens reached the coasts of Europe, in the case of the Canary Archipelago. The concept "immigrant" as Shier, Engstrom & Graham (2011) suggest that they define an "immigrant" is a person arriving (immigrating) who has come to live in a country from some other country with the purpose to settle there.

The journalistic enterprises face the challenge of attracting new audiences, being aware of the transformation of the sector and the emergence of a new ecosystem. These companies require narrative treatments contrasting from those already known, since these information units synthesize the content and preponderance of the published news; these elements are deciding to capture the attention of the readers (Jarvis, 2014).

Through the selection of the headlines, it is possible to highlight the role of new professionals in the newsrooms that are responsible for defining what kind of news be published. As Ramonet (1998) makes evident, the variety of sources guarantee objectivity. However, information is a social good that concerns and understand the whole society. This society must establish moral norms that govern the responsibility of the media (Fraerman, 1998).

The phenomenon of immigration has been analyzed from diverse perspectives is important to understand that immigration is a fact associated with times of crisis. But the gender issues are not treated deeply. Thus, one important aim is to know whether journalists take account that fact.

The Canary Islands (Spain) is a point of gateway to Europe and this is the reason for the avalanche of immigrants, males and females. The evidence suggests immigrant's networks wanting to enter by this point to reach European land. Most migration researchers understand these networks as consisting of a set of "strong ties" based on kinship, friendship, or a shared community of origin that connects migrants and non-migrants (Massey et al. 1998). Migration network approach is that a multidirectional flow of information and resources forms the basis of every migratory process (Dekker & Engbersen, 2014).

The migration phenomenon in Europe has had two phases of maximum

activity in the years 2006 and 2015 where, despite being displaced people from the place of origin to another destination, including a change of residence. In the first case, the citizens who enter Europe through the Canary Islands are the so-called undocumented immigrants. These people left their countries as a free choice and for a "personal interest," in line with the definition of International Conference on Migration (IOM). In the second case, refugees have carried out the displacement (also present in 2006, but in a very small percentage) to save their lives or preserve their freedom, as United Nations High Commissioner for Refugees (UNHCR) states.

The data analyzed in this paper focuses on international migration and the movement across national borders, consequently this work takes care of the time-span analysis that separates two massive arrivals and the evolution that originates in the field of communication in that period. The search terms "immigrant," in 2006 and "refugee" in 2015 and also the words "man" and "woman" were used as keywords to search the headlines and full news of database and locate information about immigration, and refugees (MUGAK, 2016). The study analyses the year 2006 matching with 2015 and aims to probe the narrative production generated by two Spanish newspapers (ABC and El País) and two Spanish regional newspapers (Canarias 7 and El Día), in relation to the immigration phenomenon that took place in the Canary Islands in those years.

2. Method

In the present study carried out in the years 2006 and 2015, statistical methods are mainly concerned with the non-linguistic information from a text; e.g. term frequencies, inverse frequency and the position of a keyword in a text. For data analysis, for the study we apply Iramuteq software (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires; Ratinaud, 2009; Ratinaud & Marchand, 2012, 2015).

In our study, for the data processing, apply the Descending Hierarchical Classification (DHC) by Reinert method (1983, 1986, 1990) defined by lexical classes, where each of them represents a subject matter, and they can be described according to the vocabulary that defines them.

From the most frequent words given in the text segments, lexical analysis was performed. This analysis overcomes the dichotomy between quantitative and qualitative research, as it allows employing statistical calculations on qualitative data, the texts. The vocabulary related to "immigration, immigrant/s, refugee/s, man and woman etc." are identified and quantified in the frequency and, in some cases in relation to its position within the text.

3. Results

Below, the author illustrate the data of the text corpus of the years 2006 and 2015 period of study. The corpus used in this analysis is ad-hoc constructed. It contains 4.703 newspaper headlines and news published throughout 2006 and 2015 in Spanish. We used four newspapers two of nationwide (El Pais and ABC) and two of regional scope (Canarias 7 and El DIA), of which 169 news corresponds to El Pais, 291 news for ABC, whereas Canarias 7 published 512. The information of three newspapers was obtained through MUGAK (Centre of Studies and Documentation on Immigration, Racism and Xenophobia, Basque Country, Spain, 2016) database, in case of the newspaper El DIA, 3.731 news; the information was taken directly from the newspaper database.

Table 1 - Statistical data from the text corpus of study

	Corpus 2006	**Corpus 2015**	**Subcorpus 2006**	**Subcorpus 2015 (text in web editions)**
Occurrences	426.135	30.531	147.468	6.148
Forms	11.993	4.792	9.747	1.487
Hapax	5.093	2.440	4.525	827
Texts	7	11	7	4

In addition, the characteristics of each text, the number of occurrences detected in the online version of the newspapers is broad and reflects 20% of the occurrences of the entire corpus, observed the lexicometry while the remaining 60% belongs to the activity developed in the profiles enabled in the social networks of each newspaper. It can be observed the following cloud of words by collecting in generic terms the forms that characterize the selected texts.

As it can observe some of the words, with bulkier characters and therefore most relevant, are related to the area of study that concerns us: period 2006 the word immigrant is the most used in the newspapers analyzed, followed by Canarias, *patera* and *cayuco* (two types of small boats) as a form to arrive to the Canary Islands. However, in 2015 appears the term refugee (refugiado), *immigrant* (immigrant), *welcome* (bienvenida), *governmen*t (gobierno), rescue (rescate) or the *Canary Islands* (Canarias).

In addition, some forms of refugeing, offering, asking or rescue appear, as Crespo (2008) points out, a certain ideological position that undoubtedly helps to construct a certain image about the migratory phenomenon and its

consequences for the receiving countries. The graphs generated by the Iramuteq software of this corpus of text can be inferred that some specific forms give positive or negative value. Depending on the verbs used for this purpose and the profile of the migrant to which reference is made, in our case display the data of the two analyzed periods. These appear related to the terminology of the topic that occupies us and previously used in the construction of the press holders.

3.1. Data from Descending Hierarchical Classification Analysis 2006

Iramuteq 0.7 alpha 2 software (Ratinaud, 2014) provides multivariate analysis through DHC and calculates descriptive results of clusters according to its main vocabulary (Camargo & Justo, 2013). Likewise, its location in the dendrogram, the resulting forms' clusters reflect the different work scenarios beside how some social realities cross: class 1 (social, immigrant aid), class 2 (immigrants and their local rescue), Class 3 (social and family), class 4 (institutional). As well, a concept that appears common to two conglomerates in "immigrant" and "immigration" as can be seen in the figure below. (Fig.1). The word "male" appears 184 times, X2 =521,9.

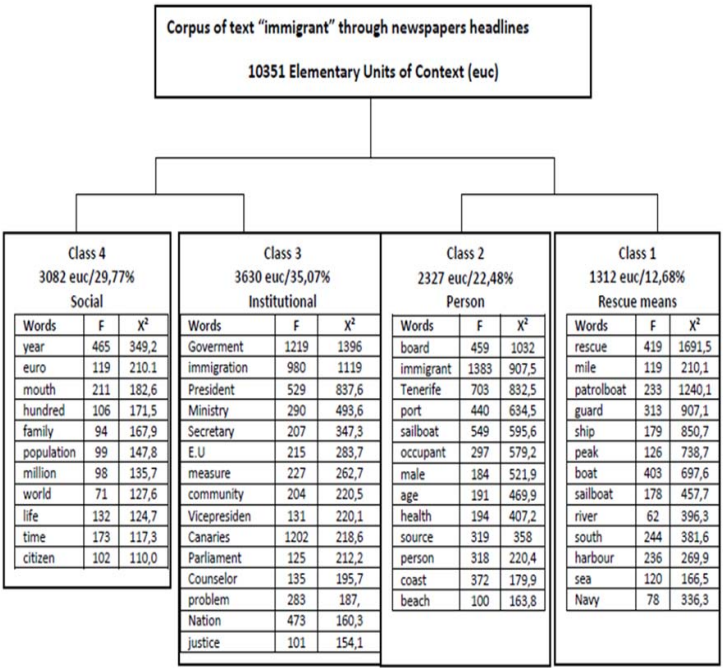


Figure 1 - DHC Dendrogram 2006

3.2 Data from 2015 DHC

The data shown in the graphs below (Fig. 2) of this text offer an estimated viewing on the figure of the "refugee" and the "immigrant" and their evolution in the context of the knowledge acquired by the media as the phenomenon is going forward. In such a way, we find two words, "refugee" and "immigrant", that appear in the journalistic headlines.

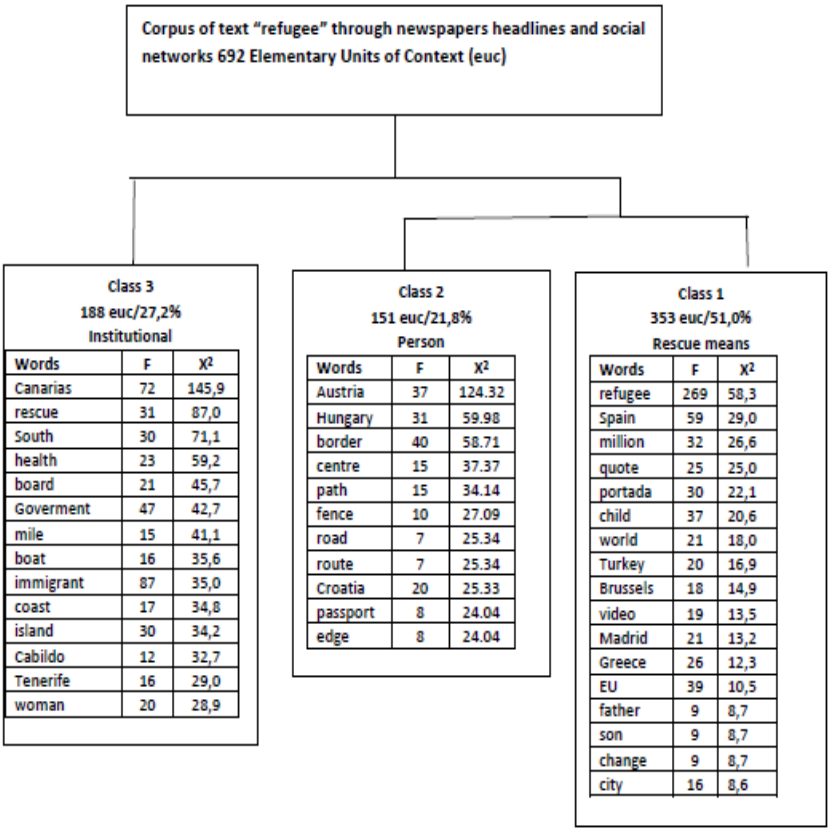


Figure 2 - DHC Dendrogram 2015

The result of the above dendrogram reflects the different work scenarios and how some social realities are mixed: class 4 (local), class 2 (institutional), class 3 (social) and class 1 (European). The word "woman" appears 20 times with X2=28,9. It is worth mentioning the founding of the term "to receive", an element that is similar to the rest of the verbs that accompany it in the constellation of words in which it is lodged (to propose, to find, to celebrate or to dispose among many others). However, it becomes more relevant due

to its preponderance and strategic situation in an environment in which it appears with vocabulary with which it keeps linguistic similarities.

4. Conclusion

This object of study that evolves in parallel to the population movement, as well as certain informative personalization through the introduction of adjectives that indicate narrative subjectivity. Our findings suggest a vast of knowledge that covers countless issues related immigrants and refugees and woman and man. It can be said that the word "man" does not appear during the 2006 and it does "male", however in 2015 appears "woman" instead "female and it does not "male" like in 2006. The mechanization of publishing systems marks a clear dividing line between some texts and others and the shortage of human and technical resources used for this activity, causes local media to be less interventionist in drafting their texts than national ones. Finally, it should be notice for the future researches the role of journalists and the usage they do of the gender topic as a way to know how the immigration phenomenon man/woman behaves.

References

Crespo, E (2008). El léxico de la inmigración: atenuación y ofensa verbal en la prensa alicantina. En M. Martínez (Ed.) *Inmigración, discurso y medios de comunicación* (pp.45-62). Alicante: Instituto Alicantino de Cultura Juan Gil Albert, Diputación Provincial de Alicante.

Dekker, R & Engbersen, G. (2014). How social media transform migrant networks and facilitate migration. *Global Networks 14*, 4, 401–418.

Jarvis, J. (2014). *Geeks Bearing Gifts*. CUNY Journalism Press, New York. Spanish *El fin de los medios de comunicación de masas. ¿Cómo serán las noticias del futuro?* Barcelona: Ediciones Gestión 2000.

Massey, D. S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino and J. E. Taylor (1998) *Worlds in motion: understanding international migration at the end of the millennium*, New York: Oxford University Press.

Mugak (2016) *Centre of Studies and Documentation on Immigration, Racism and Xenophobia,* Basque Country, Spain. Available in www.mugak.eu

Ramonet, I (2011). *La tiranía de la comunicación*. Madrid: Debate.

Ratinaud, P. (2009). IRAMUTEQ: *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* [Computer software] Retrieved 5th march 2013 in http://www.iramuteq.org.

Ratinaud, P. (2014). Visualisation chronologique des analyses ALCESTE : application à Twitter avec l'exemple du hashtag #mariagepourtous. In Actes des 12eme *Journées internationales d'Analyse statistique des Données Textuelles.* JADT 2014 (p. 553- 565). Paris, France. Disponible

Ratinaud, P. & Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IraMuTeQ. Em: Actes des 11eme Journées Internationales d'Analyse statistique des Données Textuelles. JADT 2012. Liège.

Ratinaud, P., & Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, 108, 57- 77

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8, 2, 187- 198.

Reinert, M. (1986). Un logiciel d'analyse lexicale: ALCESTE. *Les cashiers de l'Analyse des Données,* 4, 471-484.

Reinert, M. (1990). ALCESTE. Une méthologie d'analyse des données textuales et une application: Aurelia de G. de Neval. *Bulletin de méthologie sociologique*, 28, 24-54

Shier ML, Engstrom S & Graham JR (2011) International migration and social work: A review of the literature, *Journal of Immigrant and Refugee Studies*, 9, 1, pp. 38-56. http://dx.doi.org/10.1080/15562948.2011.547825.

Lo strano caso delle frequenze zero nei testi legislativi euroistituzionali

Francesco Urzì
combinazioni.lessicali@gmail.com

Abstract
In this paper we intend to verify the actual impact of the so-called *universals of translation* – i.e. those linguistic features which typically occur in translated rather than original texts - on the legislative texts produced by the European Union. To this aim, a number of text segments have been heuristically selected in order to ascertain if their statistical absence, or quasi-absence, from European legislation should be traced back to the effects of the abovementioned universals and to identify possible EU-internal factors that might explain such conspicuous statistical absences.
Keywords: universals of translation. European Union, Eur-lex, euroitaliano, terminology.

1. Introduzione

Negli ultimi tempi si sono moltiplicati gli studi su corpora comparabili volti a verificare l'effettiva incidenza dei cosiddetti *universali della traduzione*, ossia dei tratti linguistici comuni ai testi tradotti e non riconducibili a un'influenza sistemica della lingua sorgente (Baker 1993 e 1996 e Laviosa 2002). Per l'italiano disponiamo delle analisi di Garzone 2005 e di Ondelli-Viale 2010. Ondelli-Viale, che si avvalgono esclusivamente di un corpus di estrazione giornalistica, rilevano ad esempio la minore ricchezza lessicale e la frequenza lievemente maggiore del *Vocabolario di base* nelle traduzioni, per effetto dell'universale traduttivo della *semplificazione.*

Meno numerosi sono gli studi sui tratti specifici dell'*euroitaliano*, ossia di quella varietà della nostra lingua rappresentata dall'italiano delle traduzioni dell'UE. In tale ambito Cortelazzo 2013 ha operato un confronto quantitativo di due corpora di una certa ampiezza costituiti rispettivamente da direttive europee e leggi italiane di recepimento, utilizzando tra l'altro misure lessicometriche (ad es. *type/token ratio* e *hapax*) e prendendo anche in considerazione i "segmenti ricorrenti" (che secondo l'autore confermano per il corpus UE scelte lessicali "leggermente più povere e omogenee di quelle nazionali").

Con il presente contributo ci proponiamo di stabilire sulla scorta di segmenti scelti euristicamente, casi eclatanti di frequenze zero o prossime allo zero sul

dominio di secondo livello europa.eu, e più specificamente su Eur-lex, che ne costituisce un sottoinsieme. Lo scopo di tale esercizio è di verificare

- se l'irrilevanza statistica di determinate lessie in questi corpora, praticamente costituiti solo da testi tradotti - ricordiamo la pluricitata affermazione di Umberto Eco secondo cui "la lingua dell'Europa è la traduzione" - non forniscano una prova incontrovertibile degli effetti degli *universali traduttivi*, in particolare quelli della *semplificazione* e della *normalizzazione* (o *conservatorismo linguistico*);
- se non sia pure ravvisabile un processo di "autoinibizione" da parte dei traduttori UE all'utilizzo di tali lessie. Non opererebbero in altre parole solo le tendenze generali ascrivibili al processo traduttivo in sé (gli *universali della traduzione* appunto), ma anche e soprattutto la specifica cultura traduttiva euroistituzionale e lo specifico contesto tecnico-operativo che contraddistingue i servizi di traduzione delle Istituzioni europee.

Essendo tale analisi di tipo eminentemente qualitativo, l'utilizzo di un corpus "rumoroso" come Google non inficia la rilevanza dei risultati quantitativi, che tendono unicamente a individuare solo grandi scarti di frequenza, per cui è vero in questo caso che "more data is better data".

2. La *cultura traduttiva* delle Istituzioni europee

2.1 Confusione fra 'termine'e 'parola'

Un tratto soggiacente della cultura di categoria dei traduttori euroistituzionali è la non percezione della differenza teorica fondamentale fra 'termine' e 'parola'. E' diversa infatti nel termine e nella parola la natura del riferimento,

"che nel termine è specializzata all'interno di una particolare disciplina, mentre nella parola è generale in una varietà di argomenti (Cfr. Scarpa 2008: 52, che cita Sager 1994: 43).

Cabré (1999, 33-34), sulle orme di Wüster (1981), menziona due specificità della terminologia. La prima è che

"words in dictionaries are described with respect to their use in context; they are considered as elements of discourse. For terminology, on the other hand, terms are of interest on their own account";

la seconda che

"lexicology and terminology present their inventories of words or terms (…) in different ways because they start from different viewpoints: terminology starts with the concept and lexicology, with the word".

Cabré (*ibidem*, 36) nota inoltre che

"whereas a terminological inventory usually contains only nouns, in a general language dictionary all grammatical categories are represented".

2.2 Referenzialità intertestuale

La natura "ciclica" degli atti legislativi dell'Unione - che molto spesso modificano e aggiornano testi legislativi precedenti – che fa sì che le soluzioni traduttive già consacrate dall'ufficialità finiscano per essere trasferite di peso sui nuovi atti, con un fenomeno che si potrebbe definire di *common law* linguistica, in cui il precedente esercita forza vincolante sul giudizio linguistico autonomo del traduttore. E' in questa fase che il traduttore UE spesso assegna status di 'termini' a sintagmi che pur non rispondendo teoricamente a tale definizione (v. 2.1) hanno comunque acquisito il crisma dell'ufficialità per essere stati "validati" in testi legislativi precedentemente pubblicati o anche solo verificati sul piano qualitativo e ritenuti idonei a a essere immessi nel successivo iter legislativo. E' così che determinate soluzioni traduttive tendono a perpetuarsi all'interno delle "filiera testuale" della materia trattata. Al riguardo va citato anche l'effetto di condizionamento subito dai traduttori più giovani, i quali trovano arduo sostenere scelte linguistiche innovative in contrasto con la "tradizione" dei testi dell'*acquis communautaire* e, soprattutto, tendono a non discostarsi dall'approccio traduttivo dei colleghi più anziani.

3. Il contesto tecnico-operativo dei servizi di traduzione delle Istituzioni europee

3.1 House Rules

I servizi di traduzione delle Istituzioni europee hanno a disposizione un "Manuale di convenzioni redazionali" (OPOCE 2011), nella cui pagina di benvenuto si legge che "la sua applicazione [del Manuale] è **obbligatoria** [grassetto originale] per chiunque intervenga nella preparazione di ogni documento (su carta o elettronico) nelle istituzioni, organi o servizi dell'Unione europea". Non viene fatta nel Manuale alcuna distinzione fra le varie tipologie di testi e le differenti funzioni comunicative che competono a ciascuna di esse. Inoltre molte regole di redazione sono presentate sotto forma di prescrizione assoluta Ad esempio, si prescrive "direttiva" (atto legislativo) con la minuscola (il che non sorprende visto il numero di volte in cui il termine viene utilizzato nei testi UE), nonostante la regola secondo cui (Lesina 2009) "nei casi in cui un nome generalmente usato in senso comune viene utilizzato in senso proprio, con un significato restrittivo o particolare (…) l'iniziale maiuscola *può* [corsivo mio] essere utile per ragioni di chiarezza, al fine di segnalare al lettore la particolare accezione del nome". Conoscendo la scarsa frequentazione degli italiani (anche di buona cultura) con la terminologia degli atti legislativi comunitari, sorprende che il Manuale di convenzioni redazionali prescriva che "direttiva", anche quando non seguita dagli estremi completi dell'atto legislativo (ad es. direttiva

2049/39/CE), debba essere *sempre* scritta con la minuscola (dunque *anche* nei testi a carattere divulgativo destinati alle pagine web).

3.2 Effetto standardizzante delle tecnologie CAT e MT

Attualmente i traduttori delle Istituzioni europee beneficiano di una memoria di traduzione comune a tutti i servizi denominata "Euramis" e che provvede alla pretraduzione dei testi sia quando la traduzione è curata dai servizi interni sia quando è esternalizzata ad agenzie di traduzione. Da qualche anno è entrata in servizio anche la traduzione automatica che, su richiesta del traduttore, integra l'output della traduzione assistita. Poiché ad alimentare la memoria Euramis sono esclusivamente segmenti di testo "validati" (ossia già sottoposti al processo interno di controllo di qualità e dunque ritenuti idonei al successivo dibattito politico o alla pubblicazione) i traduttori preferiscono non discostarsi da soluzioni ritenute "sicure" (e la cui adozione, va pure sottolineato, si traduce in un notevole risparmio di tempo).

4. Esempi paradigmatici di "grandi assenti"

Ad esemplificazione di quanto sopra passiamo di seguito in rassegna una serie di sintagmi, che presentano casi clamorosi di frequenze zero o prossime allo zero. Nelle relative tabelle il numero di occorrenze preceduto da asterisco indica dei "falsi positivi". L'asterisco fra parentesi segnala che sono dei falsi positivi almeno una parte delle occorrenze. Le forme prese in considerazione sono una forma aggettivale gerundiva (*costruendi*), alcuni sintagmi nominali con aggettivo relazionale (*indagini poliziesche, attività manutentive, servizi consulenziali)*, un composto aggettivale determinativo formato da due aggettivi relazionali (*politico-programmatico*) e due costrutti, rispettivamente con fattorizzazione (dati *quali- quantitativi*) e zeugma preposizionale (*valutare e tener conto [di]*). Laddove utile sono state proposte, a titolo comparativo, le statistiche relative alla forme più in uso nel corpus legislativo europeo.

4.1 Gerundivo

Token	Google	Europa.eu	Eur-lex
Costruendi	11.800	*2	*1

I due unici esempi di europa.eu - 'i costruendi locali' e 'sepolcri esistenti e costruendi', entrambi provenienti dalla banca elettronica TED[1], sono riferiti ad aree territoriali italiane. In questo caso sembra aver operato il

[1] TED - *Tenders Electronic Daily*, ossia il supplemento alla Gazzetta ufficiale dell'Unione europea dedicato agli appalti pubblici europei

conservatorismo linguistico, che ha indotto ad evitare una forma non registrata dai dizionari[2] e probabilmente ritenuta dai traduttori troppo ardita.

4.2 Aggettivi relazionali semplici e composti

Un analogo comportamento linguistico convenzionale e semplificatorio da parte dei traduttori si osserva nel caso degli aggettivi relazionali. Non tutti i suffissi che formano aggettivi relazionali sono infatti suffissi "dedicati", ossia deputati a codificare esclusivamente il rapporto di relazione; alcuni formano anche aggettivi qualificativi. Tale è ad esempio il suffisso *-ivo*[3] come in *attività produttive* vs. *prefisso produttivo*. Spesso basta questa ambivalenza semantica a dissuadere il traduttore dall'utilizzare tali aggettivi in funzione relazionale e a indurlo a fargli preferire soluzioni alternative (ad es. con l'impiego della preposizione 'di' o con locuzioni preposizionali del tipo 'relativo/riguardo a/in materia di'. Nel caso di 'indagini poliziesche', potrebbe forse aver agito anche il proposito di evitare una indesiderata connotazione.

Token	Google	Europa.eu	Eur-lex
Indagini di polizia	164.000	793	85
Indagini poliziesche	14.700	(*)2	0

Da notare che una delle 2 occorrenze di 'indagini poliziesche' in europa.eu è un comunicato stampa, dunque scritto con ogni probabilità da un giornalista e non da un traduttore.

Token	Google	Europa.eu	Eur-lex
Attività di manutenzione	1.230.000	6.730	354
Attività manutentive	89.400	(*)139	*1

Da osservare che l'unico risultato di Eur-lex per 'attività manutentive' lo si ritrova in un testo italiano, che riportiamo (grassetto mio)

"Regolamento del sottosegretario di Stato per l'Edilizia abitativa, la Pianificazione territoriale e l'Ambiente recante definizione di nuove

[2] Tale forma non registrata ad esempio nel Sabatini Coletti 2008 che però riporta 'istituendo' e 'costituendo', mentre il Grande dizionario Garzanti riporta solo 'costituendo'.

[3] Suffisso usato prevalentemente per la formazione di aggettivi qualificativi (Wandruska 2004: 391)

prescrizioni relative alla prevenzione di perdite accidentali di fluidi frigorigeni nell'ambito dell'utilizzo di o dell'esecuzione di **attività manutentive** su impianti di refrigerazione e, in relazione alle stesse, recante modifica del regolamento prescrizioni impermeabilità impianti di refrigerazione 1997"
Dei 139 risultati in europa.eu 114 provengono dalla banca TED e, come conferma un controllo a campione eseguito da chi scrive, si riferiscono ad avvisi di appalto riguardanti il territorio italiano.

Token	Google	Europa.eu	Eur-lex
Servizi di consulenza	6.870.000	29.300	16
Servizi consulenziali	96.600	(*)21	0

Anche in questo caso, dei 21 risultati di europa.eu 3 provengono da TED, altri (anche se non tutti) da regioni italiane.
Per quanto riguarda gli aggettivi relazionali composti, del tipo: *libero professionale* (relativo alla libera professione) oppure *marittimo-portuale* (relativo ai porti marittimi), si è scelto come caso eclatante di assenza il composto 'politico-programmatico'. L'assenza è tanto più significativa in quanto non mancano certo nell'Unione europea i documenti funzionalmente analoghi al Documento politico-programmatico italiano, ma è solo a quest'ultimo documento che fanno riferimento le pochissime occorrenze di questo termine riscontrate su europa.eu e Eur-lex. Ancor più che nel caso degli aggettivi relazionali semplici, l'assenza si spiega con il senso di incertezza semantica che le formazioni aggettivali costituite da due aggettivi relazionali possono ingenerare, visto che spesso la loro disambiguazione (stabilire cioè se si tratta di composto coordinativo o determinativo) può avvenire solo in relazione a un dato cotesto.

Token	Google	Europa.eu	Eur-lex
Politico-programmatico	34.900	8	*1

Delle 8 occorrenze di europa.eu, almeno 2 provengono da documenti redatti da curatori italiani. L'unica occorrenza in Eurlex (dove la versione inglese è *policy and planning platform*), fa pensare a un brano di testo originariamente redatto in italiano e a una lettura coordinativa, anziché determinativa, del composto in sede di traduzione.

4.3 Fattorizzazioni e costruzioni zeugmatiche

Questi due costrutti, i cui meccanismi sono di difficile reperimento nelle grammatiche, sono ampiamente utilizzati nel linguaggio giuridico e amministrativo italiano per evidenti ragioni di economia linguistica. Si è scelta a tal fine la sequenza 'dati qualitativi e quantitativi', che è un'espressione che ricorre sovente in testi che riportano dati statistici e che viene pertanto utilizzata in una pluralità di settori. Per lo zeugma grammaticale si sono ricercate le occorrenze della sequenza 'valutare e tener conto'[4], che è risultata non ben accetta dai traduttori in quanto probabilmente troppo "audace". Oltretutto costrutti di questo tipo vengono sovente attribuiti a un'influenza della lingua inglese[5], motivo questo di ulteriori spinte puristiche da parte dei traduttori.

Token	Google	Europa.eu	Eur-lex
Dati qualitativi e quantitativi	23.100	370	1
Dati quali-quantitativi	10.400	*9	0

I 9 risultati europa.eu si riferiscono *tutti* a progetti italiano nati in ambito regionale

Token	Google	Europa.eu	Eur-lex
Valutare e tener conto	1930	(*)5	0

Dei 5 esempi in europa.eu 2 si devono all'eurodeputata Pasqualina Napolitano (doc. A6-0502/2008) mentre 3 sono di provenienza esterna all'UE.

[4] Come nel seguente esempio (grassetto mio):

Art. 5. (Coordinamento per la sicurezza e salute ex decreto legislativo n. 81 del 2008)

1. Ai sensi dell'articolo 90, comma 1-bis, del decreto legislativo n. 81 del 2008, il Tecnico incaricato è obbligato a considerare, **valutare e tener conto**, al momento delle scelte tecniche per la fase progettuale oggetto dell'incarico, **dei** principi e **delle** misure generali di tutela di cui all'articolo 15 del citato decreto legislativo n. 81 del 2008. (http://bandieconcorsi.comune.trieste.it/contenuti/allegati/schema_contratto_incarico.pdf).

[5] Fanfani 2010

Riferimenti bibliografici

Baker M. (1993), "Corpus Linguistics and Translation Studies – Implications and Applications", in: M. Baker/G. Francis/Tognini Bonelli (a cura di), *Text and Technology: In Honour of John Sinclair*, Amsterdam-Philadelphia: Benjamins, 233-250.

Baker M. (1996), "Corpus-based Translation Studies: The challenges that Lie Ahead", in: H. Somers (a cura di), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, Amsterdam-Philadelphia: Benjamins, 175-186.

Cabré, M. T. (1999), *Terminology – Theory, methods and applications*, Amsterdam-Philadelphia: John Benjamins.

Cortelazzo M. A (2013), "Leggi italiane e direttive europee a confronto", in: Stefano Ondelli (a cura di), "Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto", Trieste, EUT Edizioni Università di Trieste, 2013, pp. 57-66.

Fanfani M. (2010) *Anglicismi*, in Simone R., Berruto G. D'Achille P. (a cura di) "Enciclopedia dell'italiano". Istituto della Enciclopedia italiana, Roma

Garzone G. (2005), "Osservazioni sull'assetto del testo italiano tradotto dall'inglese", in: A. Cardinaletti/G. Garzone (a cura di), *L'italiano delle traduzioni*, Milano: Franco Angeli, 35-58.

Grande Dizionario Garzanti di italiano (2017), De Agostini Scuola s.p.a. – Garzanti linguistica (versione elettronica)

Laviosa S. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications*, Amsterdam-New York: Rodopi.

Laviosa S. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications,* Amsterdam-New York: Rodopi.

Lesina R. (2009), *Il Nuovo Manuale di stile*¸Bologna: Zanichelli

Manuale interistituzionale di convenzioni redazionali, Ufficio delle pubblicazioni dell'Unione europea (OPOCE), 2011, ISBN 978-92-78-40704-9

Ondelli S. e Viale M. (2010), *L'assetto dell'italiano delle traduzioni in un corpus giornalistico. Aspetti qualitativi e quantitativi*. In Rivista internazionale di tecnica della traduzione, n.12/2010, pp. 1-62. ISSN 1722-5906.

Sabatini F e Coletti V. (2008), *Il Sabatini Coletti. Dizionario della lingua italiana,* Milano, Rizzoli-Larousse.

Sager J. (1994), *Language Engineering and Translation Consequences of Automation*, Amsterdam-Philadelphia: John Benjamins.

Scarpa F. (2008), *La traduzione* specializzata, seconda edizione, Milano: Hoepli.

Urzì F. (2016), "Il paradosso degli aggettivi di relazione composti derivati da sintagmi N+A. Una risorsa non utilizzata in traduzione", in: R. Bombi/V. Orioles (a cura di), *Lingue in contatto-Contact Linguistics*, Roma: Bulzoni,

163-178.

Wandruszka U. (2004), "Aggettivi di relazione", In M.Grossmann/F. Rainer (a cura di), *La formazione delle parole in italiano*, Tübingen, Niemeyer, 382-394.

Wüster E. (1976), "La théorie générale de la terminologie - un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets", in H. Dupuis (a cura di), *Essai de définition de la terminologie. Actes du colloque international de terminologie (Québec, Manoir du lac Delage, 5-8 octobre 1975)*, Québec, Régie de la langue française, pp. 49-57.

Wüster E. (1981), "L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique e les sciences des choses", in Rondeau, Guy/Felber, Helmut (a cura di), *Textes choisis de terminologie – I Fondements théorique de la terminologie*, Québec, GIRSTERM, 55-114.

Les traductions françaises de *The Origin of Species* : pistes lexicométriques

Sylvie Vandaele
Université de Montréal – sylvie.vandaele@umontreal.ca

Abstract
In order to develop a sound methodology that would guide the analysis of the translations of important writings, we used Hyperbase to perform a lexicometic analysis of specificities on two corpora based on the various English and translated editions of Charles Darwin's *The Origin of Species*. We show that the translated corpus is characterized by a notable lexical dispersion. compared to the source corpus. By combining the use of Hyperbase with Logiterm. a text alignment software. we were able to target and analyse contexts of interest. This approach allows for the rapid identification of contexts that are significant both statistically and in terms of the analysis of the translation strategies themselves.

Résumé
Afin de mettre au point une méthode raisonnée d'analyse des traductions d'œuvres conséquentes, nous avons soumis les versions originales de *The Origin of Species,* de Charles Darwin ainsi que leurs traductions en français à une analyse lexicométrique des spécificités à l'aide du logiciel Hyperbase. Nous montrons que le corpus de traductions se caractérise par une dispersion lexicale notable, contrairement au corpus anglais source. Les spécificités ont permis, à l'aide du logiciel d'alignement bilingue Logiterm, de cibler l'analyse de contextes bilingues montrant les différences de choix de traduction, Cette approche permet de repérer rapidement des contextes significatifs tant sur le plan statistique que sur le plan de l'analyse des stratégies de traduction.

Keywords: *The Origin of Species*; retranslation; translation choices; specificities; Hyperbase; Logiterm,

1, Introduction

La retraduction, fréquente en littérature (voir Monti et Schnyder, 2011), est rare en science, *The Origin of Species* [désormais OS], l'œuvre célèbre de Charles Darwin, fait exception : six éditions de langue anglaise (de 1859 à 1872), six traductions en français dont deux modernes (voir Vandaele et

Gendron-Pontbriand [2014] pour les détails). Cependant, l'ampleur de l'œuvre rend l'analyse des traductions difficile. Nous proposons une méthode consistant à isoler les spécificités lexicales des originaux et des traductions, puis à repérer les contextes bilingues alignés correspondants, soumis ensuite à une analyse qualitative. Nous accédons ainsi rapidement aux éléments saillants de l'évolution de l'œuvre et de ses traductions.

2. Corpus et méthodologie

Les deux corpus[1] sont constitués par les chapitres intégraux des six éditions originales anglaises de l'OS (1859-1872) et les six traductions en français, à l'exclusion du paratexte et des notes de bas de page. Les césures en fin de ligne ont été éliminées, les numéros de page, placés entre deux phrases, les appels de notes, enlevés. Nous avons eu recours au logiciel Hyperbase v. 10[2] réalisé par Étienne Brunet (Brunet 2011). L'annotation syntaxique et la lemmatisation ont été réalisées au préalable avec Cordial v. 14 (Synapse) pour le français, et à la volée, pour l'anglais, avec la version de TreeTagger incluse dans Hyperbase. L'alignement des versions originales et traduites a été réalisé avec Logiterm v, 5.7.1. (Terminotix).

3. Les versions originales anglaises de l'OS

Le corpus anglais compte un peu plus d'un million d'occurrences, Darwin a procédé à des ajouts, mais aussi à des retraits.[3] La 6e édition (1872[4]) est 28 % plus longue que la 1re (1859), soit 48 000 occurrences de plus. L'analyse de la richesse du vocabulaire montre la proximité lexicale des six éditions originales : on compte 8559 lemmes pour tout le corpus, 6082, pour la 1re édition et 7431, pour la 6e (tableau 1).

Les lemmes communs forment la majorité du corpus : pour les textes 2 à 2, leur nombre varie de 5597 à 6600, tandis que le nombre des lemmes privatifs fluctue de 136 à 1795. L'examen des formes donne des résultats du même ordre. L'accroissement chronologique des lemmes montre un léger appauvrissement pour la 2e et la 3e édition, mais un enrichissement notable

[1] Les textes anglais viennent du site Darwin Online (John van Wyhe, dir. 2002- . *The Complete Work of Charles Darwin Online* - http://darwin-online.org.uk/). Les textes français ont été obtenus par Gallica ou Google livres, ou ont été numérisés par nous.

[2] Téléchargeable à <http://ancilla.unice.fr/>.

[3] Voir le variorum en ligne (van Wyhe, 2002-; < http://darwin-online.org.uk/Variorum/1859/1859-1-dns.html>).

[4] Celle de 1876, dite 6b, est quasiment identique à celle de 1872. C'est l'édition de 1872 qui a été traduite par Edmond Barbier (1876), raison pour laquelle nous l'avons choisie sans notre analyse.

du vocabulaire dans la 6e édition (tableau 1), essentiellement redevable à un grand nombre d'hapax, souvent des noms d'espèces.[5] Ce résultat reflète le fait que Darwin apporte de plus en plus de données à l'appui de sa théorie.

Tableau 1 – Corpus des éditions originales de l'OS

Année de publication et édition	Code	Nombre d'occurrences[6]	Richesse du vocabulaire	Accroissement chronologique
			Effectif des lemmes N (écarts réduits)	Écarts réduits (calculés sur les lemmes
1859, 1re éd,	OS01	170 634	6082 (2,67)	4,5
1860, 2e éd,	OS02	171 665	6210 (4,21)	-6,5
1861, 3e éd,	OS03	181 974	6019 (0,34)	-4,9
1866, 4e éd,	OS04	200 608	6914 (9,59)	1,8
1869, 5e éd,	OS05	199 963	7072 (11,67)	0,3
1872, 6e éd,	OS06	218 870	7431 (14,06)	16,5
Total		1 143 714	8559	

L'analyse arborée (selon Luong, 1994; cité dans Brunet 2011) met en évidence la faible distance séparant les textes, ce qui est attendu (figure 1), mais permet de situer les différentes éditions entre elles : qu'il s'agisse des fréquences (1A) ou des présences (1B)[7], on note une grande proximité entre les 1re et 2e éditions, ce qui est corroboré dans les préfaces. La 5e et la 6e sont proches, cette dernière se distinguant par les nombreux hapax. La 3e et la 4e sont intermédiaires. Nombre de lemmes privatifs passent sous la barre des 5 %, les spécificités sont peu nombreuses, ce qui est attendu, mais révélateur. Les spécificités positives ne repèrent aucun mot plein pour les quatre premières éditions, mais font apparaître le pronom *I* et le déterminant *my*.
C'est à la 5e édition que l'on note l'apparition de deux spécificités de mots pleins statistiquement significatives : *survival* et *fittest*, avec un écart réduit de 4,6 et de 4, respectivement, pour les formes, ou *survival* (substantif, 4,6) et *fit* (adjectif, 4) pour les lemmes. Dans la 6e édition, apparaissent *Mr* (7,1), *through* (6,1) *cambrian* (5,8) *orchids* (4,3), *developed* (4,9) *et development* (4,2), *lower* (4,2),

[5] Le nombre d'hapax augmente considérablement dans la 6e édition : respectivement, 45, 40, 61, 133, 134, et 622 occurrences (lemmes) de la 1re à la 6e édition (écart réduit de 33,5 pour la 6e édition).

[6] Les valeurs reportées dans les tableaux sont fournies par Hyperbase. Il y a de légères différences avec des valeurs publiées antérieurement, dues à la préparation des textes et aux logiciels utilisés pour le décompte.

[7] Respectivement selon Labbé et Jaccard, cités dans Brunet 2011.

beneficial (4,1) et *spontaneous* (4,1). L'analyse des lemmes fait, en plus des précédents, remonter *survival* (substantif, 4,6), *spine* (substantif, 5,3), *increased* (adjectif, 4,2), *movement* (substantif, 4,1), *fit* (adjectif, 4,1), *beneficial* (adjectif, 4,1) et *spontaneous* (adjectif, 4,1).

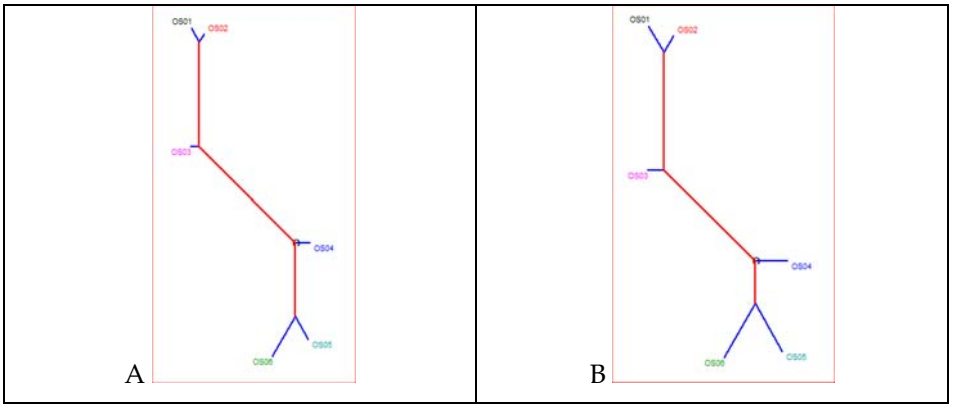


Figure 1 – Analyse arborée sur les lemmes : A – sur les fréquences; B – sur les présences

Le regroupement des spécificités en catégories reflétant le contenu sémantique (établi à partir des contextes) est instructif : concepts théoriques (*fittest, fit, survival, through* [expression de la causation]), données et citations (*cambrian, orchids, spine, Mr*), vision dynamique du vivant de Darwin (*develop, development, increased, movement, spontaneous*), jugements de valeur (*beneficial*, *lower* [certaines occurrences]). Ainsi , les spécificités, même rares, se démarquent par leur saillance : elles captent l'introduction du fameux concept de Spencer (1864), *survival of the fittest* et permettent de présumer une affirmation de la pensée de Darwin – à savoir sa vision profondément dynamique de la nature. Enfin, les spécificités négatives signalent que les fréquences relatives du déterminant possessif *my* et du pronom *I* diminuent avec le temps, ce qui traduit l'ajout de passages non argumentatifs contenant des données, et ce qui corrobore l'augmentation des hapax, constitués par majoritairement par des noms d'espèces.

4. Analyse du corpus français

Le corpus français comprend un peu plus de deux millions d'occurrences (tableau 2) : trois traductions d'époque (Clémence Royer [1862, 3e éd.], Jean-Jacques Moulinié [1873, 5e éd.], Edmond Barbier [1876, 6e éd.]); celle de Daniel Becquemont (2008), qui part de la traduction de Barbier et la modifie pour remonter à la 1re édition; deux modernes, par Augustin Berra (2009, 6e éd.) et Thierry Hoquet (2013, 1re éd.) (voir Vandaele et Gendron-Pontbriand [2014] pour les références bibliographiques). Les textes comptent de 181 785 à

248 863 occurrences, soit un écart de 67 078 occurrences. Les différences de coefficients de foisonnement[8] révèlent déjà que les traducteurs ont travaillé avec des stratégies de traduction distinctes. L'homogénéité lexicale diminue par rapport aux originaux. La contribution de chacun des textes à la richesse lexicale est beaucoup plus importante en français qu'en anglais : les lemmes partagés dans les textes pris deux à deux se situent entre 4498 (13Ho et 62Ro) et 5649 (73Mo et 76Ba) pour un total de 11712 lemmes (soit 3153 lemmes de plus que dans le corpus anglais). Chacun des textes français contribue pour un pourcentage moindre au vocabulaire commun (figure 2A). Les effectifs des lemmes privatifs sont plus importants (de 772 à 3000) et fluctuent d'un traducteur à l'autre (figure 2B). Sont mises en évidence les différences entre Becquemont (08Bq) et Hoquet (13Ho) pour la 1re édition, et entre Barbier (76Ba) et Berra (09Be) pour la 6e édition, mais aussi la proximité (attendue) entre Barbier et Becquemont.

*Tableau 2 – Traductions françaises de l'OS – * d'après la traduction de Barbier de la 6e édition,*

Année de publication	Édition originale anglaise	Traduit par	Code	Nombre d'occurrences	Coefficient de foisonnement	Richesse du vocabulaire
						Effectif des lemmes N (écart réduit)
1862	1861 (3e)	C. Royer	62Ro	207 633	14 %	6357 (-6,7)
1873	1869 (5e)	J.-J,.Moulinié	73Mo	211 691	6 %	7036 (0,8)
1876	1872 (6ea)	E. Barbier	76Ba	241 170	10 %	6971 (-3,8)
2008	1859 (1e)*	D. Becquemont	08Bq	186 440	9 %	6260 (-4,8)
2009	1876 (6eb)	A. Berra	09Be	248 863	14 %	7804 (5,0)
2013	1859 (1e)	T. Hoquet	13Ho	181 785	7 %	6579 (-0,2)
Total				**1 277 582**		**11 712**

Les distances lexicales intertextuelles (figure 3) confirment la proximité de Becquemont et de Barbier, mais révèlent deux faits inattendus : 1) Royer (62Ro) se situe sur la même branche que Berra et Hoquet; 2) Moulinié (73Mo) se place entre Becquemont et Barbier lorsque l'on passe des fréquences aux présences.

[8] Le coefficient de foisonnement est l'accroissement du nombre d'occurrences observé lorsque l'on traduit de l'anglais au français. Il est généralement admis, en traduction dite « pragmatique » (par opposition à la traduction littéraire) que le taux de foisonnement se situe généralement entre 10 % et 15 %, une des causes étant que le français recourt à plus de mots grammaticaux que l'anglais. Une forte concision peut diminuer ce taux.

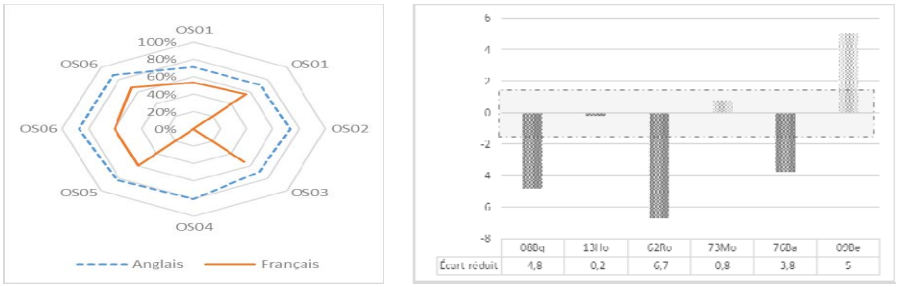


Figure 2 – A – Contributions respectives de chacun des textes aux parties communes des corpus anglais et français (lemmes)[9] – B – Richesse lexicale (lemmes), Le pointillé indique le seuil de 5 %,

Diverses hypothèses explicatives doivent être explorées, mais il n'est en tout cas plus permis de douter que les manières de traduire sont décisives au point de brouiller, sur le plan lexical, la chronologie des versions originales, et que cette approche permet de mettre ces particularités en évidence.

Figure 3 – Analyse arborée (méthode Luong) sur les lemmes
A – calculée sur les fréquences (Labbé); B - calculée sur les présences (Jacquard)

Nous nous sommes ensuite concentrée sur les spécificités positives des lemmes des mots pleins et, parmi elles, avons sélectionné les unités dont la signification paraissait la plus caractéristique du propos central de l'OS : ainsi, sélection, préservation, pouvoir… ont été retenus, mais pas aujourd'hui, grandement,

[9] Le schéma a été obtenu à partir des effectifs des lemmes pour chacun des textes, ramenés en pourcentage du nombre total de lemmes par corpus (représentation « radar » fournie par Excel v.16). Les effectifs des lemmes des textes traduits ont été disposés en regard des textes anglais (ceux de OS1 et OS6 ont donc été dupliqués); de plus, la forme assymétrique du tracé pour le français rend compte de l'absence de traduction d'OS2 et d'OS4. À cause de ces particularités, l'aire délimitée par les traits n'est pas représentative des valeurs totales pour chacun des corpus, mais le schéma reste visuellement parlant.

inclure… Nous nous sommes ensuite concentrée sur les spécificités positives des lemmes des mots pleins et, parmi elles, avons sélectionné les unités dont la signification paraissait la plus caractéristique du propos central de l'OS : ainsi, sélection, préservation, pouvoir… ont été retenus, mais pas aujourd'hui, grandement, inclure…

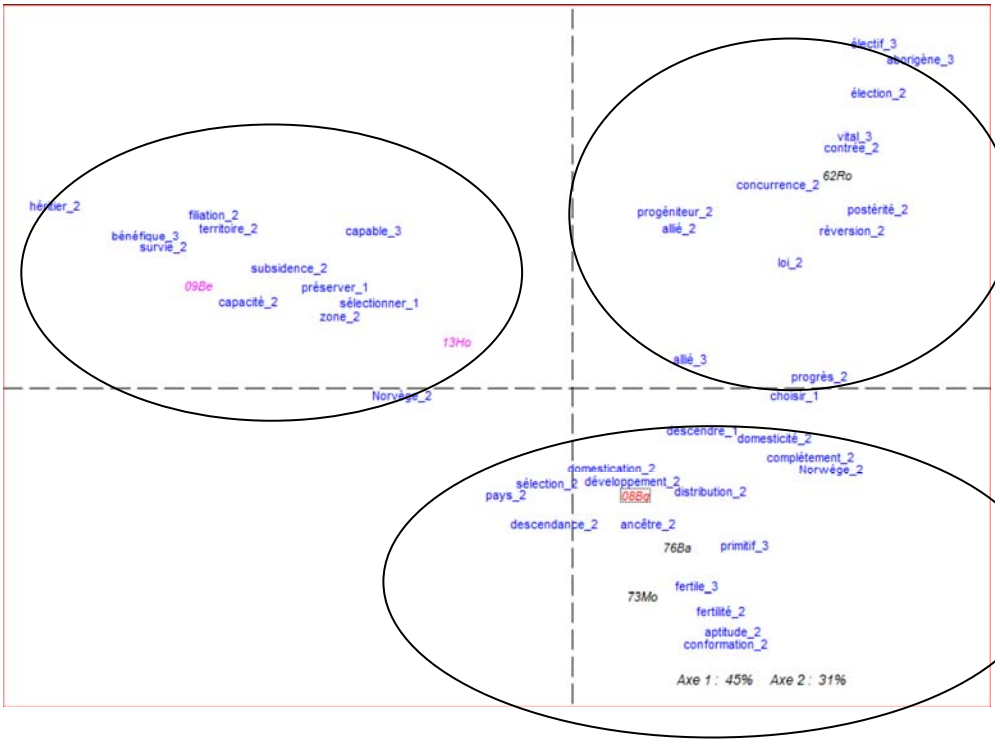


Figure 4 – Analyse factorielle de correspondances : sélection de lemmes parmi les spécificités

La quarantaine de lemmes ainsi obtenus a permis de générer un graphe (figure 4) représentant le résultat d'une analyse de correspondances (menée selon le programme de Lebart, inclus dans Hyperbase, sur les données pondérées). Le graphe montre que les modernes (Berra, Hoquet) s'opposent aux anciens (Barbier, Moulinié) ou quasi-ancien (Becquemont), Royer se situant à part. La consultation des contextes ciblés par cette méthode dans les corpus alignés par Logiterm permet d'analyser qualitativement les choix de traduction. L'exemple le plus frappant est le choix de <u>élection</u> et de <u>électif</u> par Royer, qui s'oppose au choix de <u>sélection</u> par les autres traducteurs (tab. 3).

Tableau 3 – Traductions alignées d'une phrase commune à toutes les éditions anglaises (Introduction)

Darwin	*and we shall then see how **Natural Selection** almost inevitably causes much Extinction of the less improved forms of life…*
62Ro	*Nous verrons comment cette **élection naturelle** cause presque inévitablement de fréquentes extinctions d'espèces parmi les formes de vie moins parfaites…*
73Mo	*Nous y verrons comment la **sélection naturelle** détermine presque inévitablement l'extinction des formes moins perfectionnées…*
76Ba	*Nous verrons alors que la **sélection naturelle** cause, presque inévitablement, une extinction considérable des formes moins bien organisées…*
08Bq	*Nous verrons alors que la **sélection naturelle** cause presque inévitablement une extinction considérable des formes moins bien organisées*
09Be	*nous verrons alors de quelle façon la **sélection naturelle** cause presque inévitablement une forte extinction des formes de vie moins améliorées…*
13Ho	*Et nous verrons comment la **Sélection Naturelle** cause presque inévitablement une grande Extinction des formes de vie moins améliorées…*

5. Conclusion

Le ciblage de contextes, repérés au moyen d'une analyse lexicométrique préalable, dans des corpus alignés conséquents est une stratégie de choix. Elle permet d'arriver assez vite à des observations statistiquement significatives et de pointer d'emblée sur des éléments majeurs sans hypothèse préalable. Comme le souligne Brunet (2002), l'intérêt de travailler sur des traductions est que certains paramètres sont fixés. L'inconvénient actuel de l'entreprise tient à la faible ergonomie du processus, c'est-à-dire aux nombres de clics liés au passage d'un logiciel à l'autre. Restent les nombreuses modifications sous le seuil de 5 %, qui peuvent recéler, malgré l'absence de signification statistique, des éléments cruciaux en matière de choix de traduction. D'autres stratégies de filtrage sont alors nécessaires pour leur étude.

Remerciements

Nous remercions vivement Étienne Brunet, Damon Mayaffre et Laurent Vanni pour leurs conseils sur l'utilisation d'Hyperbase. Il va de soi que les éventuelles erreurs sont nôtres. Merci aussi à Marie-Joëlle Stratford-Desjardins, étudiante auxiliaire de recherche, pour son aide à la préparation du corpus. La présente recherche a bénéficié d'une subvention de recherche du Conseil de recherche en sciences humaines du Canada (2015-2018).

Références

Brunet É. (2002). Un texte sacré peut-il changer ? Variations sur l'Evangile. In Cook J., dir. *Bible and Computer*, Leiden / Boston : Brill, pp. 79-98.

Brunet É. (2011). *Hyperbase – Manuel de référence*. *Hyperbase pour Windows, version 8.0 et 9.0.*

Luong X. (1994). L'analyse arborée des données textuelles : mode d'emploi. *Travaux du cercle linguistique de Nice,* 16 : 27-42.

Monti E. et Schnyder, P., dir. (2011). *Autour de la retraduction : Perspectives littéraires européennes*. Coll. Universités, Paris : Orizons,

Spencer H. (1864). *The Principles of biology*. Vol. 1, New York: Appleton.

Vandaele S. et Gendron-Pontbriand E.-M. (2014). *Des « vilaines infidèles » aux grands classiques : traduction et retraduction de l'œuvre de Charles Darwin.* In: Pinilla J. et Lépinette B., dir, Traducción y difusión de la ciencia y de la técnica en España en los siglos XVIII y XIX,Valence : Universitat de València, pp. 249-276.

Circuits courts en agriculture : utilisation de la textométrie dans le traitement d'une enquête sur 2 marchés

Pierre Wavresky[1], Matthieu Duboys de Labarre[2],
Jean-Loup Lecoeur[3]
[1]Umr Cesaer Inra-Agrosup Dijon – pierre.wavresky@inra.fr
[2]Umr Cesaer Inra-Agrosup Dijon – matthieu.duboys-de-labarre@inra.fr
[3]Umr Cesaer Inra-Agrosup Dijon – yajintei@hotmail.fr

Abstract
Semi-structured interviews about short food supply chains have been done with producers and consumers on two different markets. Our work gives an insight to the themes common to producers and consumers that are not attributable to the interviews guides. It also underlines the advantages of a textometric approach and the precautions necessary to interpret such a corpus.

Résumé
Des entretiens semi-directifs sur le thème des circuits courts alimentaires ont été menés sur deux marchés, auprès de producteurs et des consommateurs. Notre travail s'intéresse notamment aux thématiques communes aux producteurs et consommateurs et qui ne soient pas imputables aux grilles d'entretiens. Il souligne par ailleurs les apports d'une approche textométrique, ainsi que les précautions d'interprétation sur un tel corpus.
Keywords: short food supply chain, semi-structured interviews, textometry

1. Introduction et méthodologie

Les circuits courts alimentaires interviennent de plus en plus dans le débat social. Ils sont devenus l'emblème d'une opposition au « modèle conventionnel ». Ils s'inscrivent également dans des enjeux de politique publique (définition légale en 2009 avec le plan Barnier[1]), et scientifique. Ils comprennent des formes innovantes comme les AMAP, mais aussi des formes plus anciennes comme les marchés ou la vente à la ferme.
La sociologie a abordé les circuits courts sous des angles variés : la consommation engagée (Dubuisson-Quellier, 2009), la sociologie de

[1] Circuit de commercialisation comprenant au plus un intermédiaire entre le producteur et le consommateur.

l'innovation (Chiffoleau et Prévost, 2012), d'autres ont approché la question en décalant le point de vue vers le développement local (Traversac, 2010) ou au travers de la notion de proximité (Mundler et Rouchier, 2016). Les travaux de sociologie insistent sur l'intérêt économique des circuits courts, mais aussi sur leur capacité à recréer du lien social (Prigent-Simonin et Hérault-Fournier, 2014). De nombreux dispositifs s'appuyant sur les circuits courts de commercialisation se caractérisent par un rapport direct entre consommateurs et producteurs. Ce lien a été l'objet de différentes analyses et interprétations dans la littérature. Il est perçu comme un déplacement de l'espace de référence des agriculteurs vers celui des consommateurs (Dufour et Lanciano, 2012). Il a aussi été analysé comme le lieu de rencontre autour d'attentes plurielles (Chiffoleau et Prévost, 2012). Plus généralement, il s'ancrerait dans des logiques communes de *re-localisation* des pratiques agricoles et alimentaires (Duboys de Labarre, 2005). C'est ce lien que nous allons analyser au travers d'un dispositif textométrique. Nous mettrons en lumière les intérêts et les éventuelles limites interprétatives liés au type de corpus (faible nombre d'entretiens semi-directifs). Cela nous éclairera également sur les thématiques abordées et leur spécificité. Dans le cadre du projet européen H2020 « Strength2food » [2] , pour la France, nous avons interrogé 23 personnes[3] (12 vendeurs-producteurs et 11 consommateurs) sur deux marchés (en milieu rural et en milieu urbain) par entretien semi-directifs. Nos deux sous-populations relèvent d'initiatives différentes dans leur structuration et leur ancienneté[4]. Dans les deux cas, les parties-prenantes restent attachées à la consommation/production bio et sont assez engagés. Ce corpus n'est donc pas représentatif (ni des consommateurs ni des producteurs) et nous considérons ce travail comme exploratoire.

Le corpus est analysé grâce au logiciel de textométrie Iramuteq[5], les thèmes communs ou spécifiques des producteurs et consommateurs seront recherchés essentiellement par classification descendante hiérarchique (Reinert, 1983) et par analyse de spécificité. Parmi les variables caractérisant les textes, a été incluse une variable à 4 modalités : consommateur-rural,

[2] https://www.strength2food.eu/. Ce projet a été financé par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de la convention de subvention n° 678024

[3] Ces entretiens, structurées autour de 6 thèmes, sont semi-directifs et visent à favoriser l'expression des acteurs. Ils sont retranscrits mot à mot et incluent des annotations de l'intervieweur.

[4] Celle en milieu urbain est un marché de plein vent traditionnel, celle en milieu rural est un marché de producteurs innovant.

[5] http://www.iramuteq.org/ (Pierre Ratinaud)

consommateur-urbain, producteur-rural, producteur-urbain[6]. Comme la longueur des interviews est très variable (de 102 à 560 segments de texte) et le nombre d'interviewés assez faible (23), les statistiques relatives à cette variable peuvent être essentiellement imputables à une interview, il est donc d'autant plus nécessaire de revenir à l'interview. De plus il peut arriver que le lien, en termes de Khi², entre une des quatre catégories (ou une interview) et une thématique (classe de la classification) soit faible. Or quelques segments de textes énoncés par cette catégorie sous-représentée sont parfois très liés à cette thématique, et dire que le lien est faible serait erroné. D'où l'analyse, aidée par une représentation graphique, des segments de textes les plus caractéristiques d'une classe, pour chaque catégorie étudiée.
Deux annotations de l'intervieweur, caractérisant la parole de l'interviewé, ont été conservées au sein du corpus, et seront donc analysées comme les autres mots : « rire » (codé « _rire ») et « blanc », signifiant un délai avant la réponse ou en son sein (codé « _blanc). Le but étant de voir si des hésitations (« _blanc ») sont cooccurrentes d'autres lemmes.

2. Analyse statistique du corpus réponse

Les 5 lemmes les plus courants sont : *aller, voir, bio, gens, marché*. Ce qui ressemble à un programme : *aller* au *marché*, donc favoriser un mode de circuit court, pour acheter ou vendre des produits *bio* et pour *voir* des *gens*, donc avec un aspect relationnel important. Il est probable que les lemmes *bio, aller* et *marché* soient liés au contexte d'enquête (nature des enquêtés pour *bio* et nature des dispositifs pour *aller* et *marché*). Enfin, le caractère assez homogène de l'importance quantitative de ces 5 lemmes peut être interprété comme le reflet d'un horizon commun partagé par nos informateurs et ce en dépit de de leur groupe d'appartenance (producteur ou consommateur) ou du dispositif étudié.

2.1. Classification descendante hiérarchique : 12 types de discours

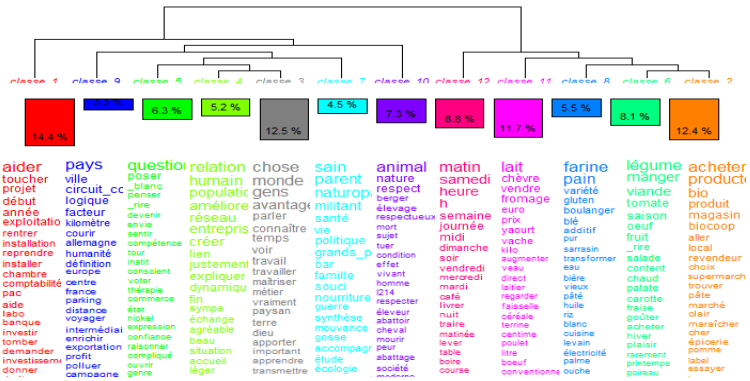
Une classification descendante hiérarchique[7] (Reinert 1983) a permis de dégager 12 types de discours. Nous nous focaliserons sur 2 ensembles de classes[8], selon qu'elles sont plutôt spécifiques ou peu spécifiques d'une catégorie (producteur ou consommateur).

---

[6] Producteur-urbain signifiant producteur vendant sur le marché de la ville moyenne, en opposition avec producteur-rural qui vend sur le marché du village.

[7] 5264 segments de texte sur les 6231, soit 84%, ont été retenus par la classification.

[8] Nous écartons la classe 3 (12,5%) car elle est peu interprétable (lemmes polysémiques : *chose, gens, monde*...).

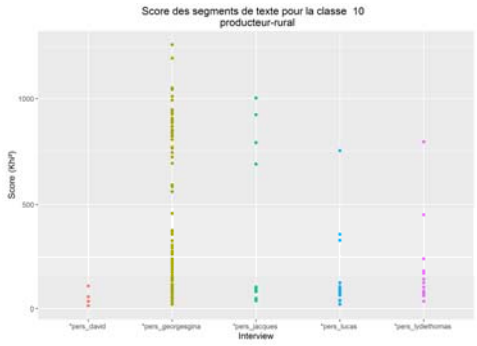


Graphique 1 : les 12 classes de discours

Le premier ensemble regroupe les classes 1, 2, 6, 9 et 11 qui sont caractéristiques d'un sous-groupe. Les classes 1 et 11 concernent surtout les producteurs, par contre les classes 2, 6 et 9 émanent principalement de consommateurs. Dans la classe 1 (14.4%) il est question des *aides*, de *projet*, d'*installation*, de *reprise (*d'*exploitation)*, d'*investissement*. Il y a des critiques sur la *PAC (*notamment sur le fait que ce soit compliqué), mais pas seulement : « *Bah comme on a de la surface un peu ouais ça commence c'est super compliqué la* ***PAC*** *je sais pas si tu veux qu'on en parle _rire même nous on a du mal* » (Lydie, productrice rurale). La classe 11 (11,7%) est orientée autour des produits laitiers (*lait, chèvre, fromage, yaourt, vache, faisselle, litre, cabri*…), avec un aspect monétaire (*euro, prix*). Dans la classe 6 (8.1%) c'est de nourriture dont il est question, notamment le fait de manger des fruits et légumes de saison (*manger, tomate, fraise, saison*, pas en *hiver*). C'est un discours de consommateurs, surtout urbains. Melissa et Jennifer parlent surtout des courses qu'elles font, où elles les font (sur le marché de la ville moyenne essentiellement, où elles ont été interrogées). Toutefois l'autre thème (manger des fruits de saison) est celui qui est le plus typique de cette classe.
Dans la classe 9 (3.3%) il est question de ville (vivre en *ville*/à la *campagne*) et de distance, aussi bien en termes de proximité que de nombre d'intermédiaires (*distance, kilomètre, circuit_court, intermédiaire*). C'est plutôt une classe de consommateurs. Enfin dans la classe 2 (12.4%) les 4 premiers lemmes forment une phrase : *acheter produit bio producteur*. *Revendeur* et *local* sont présents aussi. Il est donc question du comportement d'achat, mais pas des produits qu'on achète, comme dans la classe 6, plutôt de certaines de leurs propriétés (*bio*) et de la qualité du vendeur (*producteur*). Les classes 1, 2 et 6 renvoient directement à des thèmes abordés dans les guides d'entretiens respectifs des groupes et la classe 11 à une catégorie de produit agricole

spécifique qui était surreprésentée dans l'échantillon des producteurs transformateurs (5 informateurs sur 12). Ces classes parlent des pratiques liées aux groupes (professionnelles, d'achat et de consommation alimentaire) et permettent de les caractériser. Nous noterons que les classes 1, 2 et 6 renvoient à la notion de maîtrise ou de contrôle. Pour la classe 1 parce que les aides PAC sont parfois perçues comme extérieures et complexes. Pour les classes 2 et 6 au contraire parce qu'elles traduisent l'idée que le consommateur maîtrise sa pratique (choix de se fournir directement auprès d'un producteur et en aliments bio, locaux et de saison).

Le second ensemble regroupe les classes 4, 5, 7, 8, 10 et 12. Elles sont peu spécifiques d'une catégorie. La Classe 10 (7.3%) est celle du respect des animaux et plus généralement du respect du vivant. On peut remarquer que le lemme *_rire* y est particulièrement rare : dans cette classe, le respect des animaux est abordé comme une question sérieuse. « *C'est un* ***animal*** *pour l'****élevage*** *donc je le mange s'il a été* ***élevé*** *dans le* ***respect*** *des* ***lois*** *de la* ***nature*** *et de l'****univers*** *s'il a été* ***élevé*** *d'une* ***manière respectueuse*** *par* ***rapport*** *à l'environnement* » (Théophile, producteur urbain) [Les mots en gras sont spécifiques de la classe]. Il n'y a pas de différence marquée rural/urbain ou producteur/consommateur.



Graphique 2 : Score des segments de texte (classe 10)

Mais si on considère le nombre de segments de texte caractéristiques (graphique 2), on voit que Jacques n'en parle pas beaucoup mais il en a énoncé certains très caractéristiques. Autrement dit, il parle peu mais intensément du bien-être animal : « *Et nous nos* ***animaux*** *on est en bio on fait attention au bien-être* ***animal*** *on fait le* ***choix*** *de garder tous les petits pour pas qu'ils partent dans des* ***élevages industriels intensifs*** *et la suite logique*» (Jacques, producteur rural [score=925][9]).

La classe 7 (4,5%) renvoie à deux univers de sens différents autour du lemme *vie* : d'une part la notion de trajectoire de vie en relation avec la parentèle (*famille, parent* [d'origine *agricole*], *grand_parent, enfant*), et d'autre part à une forme de souci de soi (*mode* de *vie sain*, *santé* reliée à *nourriture* et *alimentation*). « *En* ***amont*** *dans un* ***mode*** *de* ***vie*** *qui devrait te permettre d'avoir une* ***vie*** *plus*

[9] La somme des Khi² (mesurant le lien entre chaque lemme et la classe) donne le *score du segment de texte.*

harmonieuse plus ***saine*** *plus en* ***meilleure santé physique*** *psychique mentale* ***sociale*** *parce_que tu crées du* ***lien*** *aussi enfin y a une… ça va dans une même* ***mouvance*** » (Claire, consommatrice rurale).

La classe 8 (5.5%) concerne les céréales (*farine, pain, gluten, variété, vieux, boulanger*), notamment les *vieilles variétés*.

La classe 5 (6.3%) est celle du doute (on se *pose* des *questions*, il y a des *_blanc* : ces 3 lemmes sont entre 8 et 9 fois plus nombreux qu'attendu). « Se poser des questions » et *penser* évoque aussi une prise de conscience de problèmes. Mais c'est également « poser des questions » aux vendeurs sur leur production.

La classe 4 (5.2%) est celle des relations et de leur importance. « *Eh ben les* ***relations humaines*** *on côtoie une diversité de* ***population*** *quoi des* ***gens*** *et en fait on se parle c'est* ***agréable*** *_rire* » (Christine, consommatrice rurale).

Enfin la classe 12 (8.8%) est celle du temps (temps passé [*heure*], horaire précis [*h*]). Les jours de la semaine sont cités, les moments de la journée aussi, avec *matinée, nuit, café, boire*… Les 2 individus les plus impliqués dans cette classe sont François et Thérèse (éleveurs urbains). Il n'y a pas de spécificité forte d'une des 4 catégories car s'il y a surreprésentation de certains producteurs dans cette classe, d'autres parlent très peu de cet aspect (David et Théophile). Or les deux producteurs qui sont principalement impliqués dans cette classe se sont installés dans un cadre familial (ils ont repris l'exploitation de leurs parents). Alors que ceux qui en parlent le moins sont des hors cadres familiaux. La littérature (Dufour et Lanciano, 2012) souligne que les contraintes temporelles sont plus importantes dans le cadre d'une production en circuits courts. Cette dernière serait vécue différemment en fonction de la trajectoire des agriculteurs (cadres ou hors cadres familiaux).

Le caractère commun de ces classes nous permet de proposer quelques pistes de réflexions concernant les liens qui se nouent entre producteurs et consommateurs. La classe 5 (celle du doute) renvoie partiellement à une forme de réflexivité partagée par ces deux groupes. Le respect des animaux et de la nature (classe 10)[10] et l'aspiration à un mode de vie, un souci de soi (classe 7) dessinent un lien entre préoccupations personnelles et engagements globaux (respect des animaux et cause environnementale) (Pleyers, 2011). Enfin, la classe 4 souligne l'horizon commun que constitue l'importance du lien social attaché aux circuits courts.

[10] Cette classe commune émerge dans le discours alors qu'elle n'est pas un thème des deux guides d'entretiens.

2.2. Pronoms personnels et spécificités

L'analyse des spécificités des 4 catégories d'interviewés, toutes classes confondues, a mis notamment en évidence un emploi très différencié des pronoms personnels. Les consommateurs ruraux citent souvent deux des producteurs par leur prénom. Le lemme *discuter* est également présent. Donc ils parlent de gens avec lesquels ils sont en lien fort.

Les consommateurs urbains citent beaucoup *je* et *j,* ainsi que *vous* : « *Oui et puis […] si* ***vous*** *voulez vos salades au bout de 3 ou 4 jours en grande_surface elles ont pas été vendues elles ont quand même pas la même tête que celles que* ***j****'achète qui ont été cueillies la veille hein* » (Mélissa, consommatrice urbaine). Il est donc question de ce que l'interviewé fait (*je, j*) et de ce qu'il ne fait pas (*vous*). Donc de son comportement d'achat : ce qu'il achète, du lieu où il achète ou pas (*marché, supermarché, …*), de la façon dont c'est produit ou vendu (*bio, label, équitable, local, transport*). Il y a également le lemme *rencontre* : le lien est présent, mais de façon plus conceptuelle, moins proche que dans le groupe des consommateurs ruraux.

Chez les producteurs ruraux les pronoms *tu* et *nous* sont très employés. Le *nous* peut renvoyer à un couple de producteurs (Georges et Gina) ou à une communauté à laquelle on appartient : (les producteurs diversifiés, les producteurs du marché du village rural) : « ***Nous*** *ce qui fait la caractéristique du secteur c'est que c'est des exploitations qui sont tournées vers beaucoup d'espèces on n'a pas de spécialisation enfin pas de très très grosse spécialisation* » (David, producteur rural). Il nous semble que cette spécificité dans l'utilisation des pronoms peut-être rattachée à la nature différente des dispositifs (et non à leur caractère rural ou urbain). Dans un cas, le marché de plein vent traditionnel, nous avons affaire à une structure de taille importante qui préexiste aux acteurs. S'il est bien un lieu de rencontre, il est plus fortement marqué par une dimension individuelle tant pour les producteurs que pour les consommateurs (d'où la présence du *je*). Dans l'autre, le petit marché de producteurs engagés, nous avons affaire à un projet de taille plus réduite construit par une partie des acteurs. Les relations interpersonnelles, l'identification à un ou des collectifs mais également la dimension participative y sont donc plus marquées.

3. Conclusion et perspectives

De nombreux thèmes sont apparus fortement dans le discours des interviewés : l'importance des relations, l'importance d'acheter au producteur des produits bio, de manger des produits de saison, d'utiliser des variétés de blé ancienne, de respecter l'environnement et les animaux. D'autre part, l'emploi de pronoms personnels différents et l'usage ou non de prénoms, révèlent une proximité avec les producteurs locaux (discours des

consommateurs ruraux), l'appartenance à un groupe (discours des producteurs ruraux), une norme dans le comportement d'achat (discours des consommateurs urbains). Il est important de ne pas tenir compte uniquement de la spécificité globale d'une catégorie (ou d'un interviewé) pour juger de sa plus ou moins grande implication dans une thématique (cas de Jacques). De ce fait, les thèmes révélés par la classification ne sont pas toujours très spécifiques d'une catégorie. Malgré un corpus restreint et spécifique, la textométrie permet de mettre au jour des éléments factuels identifiés dans la littérature et d'esquisser des liens analytiques avec des approches théoriques plus générales. Ces résultats nous amèneront à poursuivre ce travail, dans le cadre du projet Strenght2Food, en y intégrant une comparaison internationale (avec tout ou partie du corpus des 6 pays partenaires sur cette thématique).

Références

Chiffoleau Y., Prévost B. (2012). Les circuits courts, des innovations sociales pour une alimentation durable dans les territoires, *Norois*, 224.

Duboys de Labarre M. (2005). Le mangeur contemporain, une sociologie de l'alimentation. *Thèse de sociologie, soutenue à Bordeaux, 426p.*

Dubuisson-Quellier S. (2009). La consommation engagée. *Paris, Presses de la Fondation nationale des sciences politiques (Contester)*.

Dufour A., Lanciano E. (2012). Les circuits courts de commercialisation: un retour de l'acteur paysan ? *Revue Française de Socio-Économie (n° 9*), pp. 153-169.

Mundler P., Rouchier J. (2016). Alimentation et proximités: Jeux d'acteurs et territoires. *Educagri.*

Pleyers G. (dir.) (2011) La consommation critique, mouvements pour une alimentation responsable et solidaire. *Desclée de Brouwer*.

Prigent-Simonin A-H., Hérault-Fournier C. (2014). Au plus près de l'assiette. *Editions Quæ*.

Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, VIII(2) :187-198.

Traversac J.B. (2010). Circuits courts : contribution au développement régional. *Educagri.*

On the phraseology of spoken French: initial salience, prominence and lexicogrammatical recurrence in a prosodic-syntactic treebank *Rhapsodie*

Maria Zimina, Nicolas Ballier
Université Paris Diderot
mzimina@eila.univ-paris-diderot.fr; nicolas.ballier@univ-paris-diderot.fr

Abstract
This paper focuses on specific quantitative characteristics of spoken language phraseology in the *Rhapsodie* speech database (ANR *Rhapsodie* 07 Corp-030-01). A recent study (Zimina & Ballier, 2017) has shown that prosodic segmentation into IPE: Intonational PEriods (segments of speech with distinctive pitch and rhythm contours) available within the *Rhapsodie* database offers new insights for the observation of the functions of formulaic expressions in speech. Recurrent lexicogrammatical patterns at the beginning of Intonational PEriods (IPE) are strongly related to spoken formulaic language. These variations of initial salience depend upon several factors (interactional needs, social context, genres, etc.). Further experiments have shown that initially salient patterns also have specific prosodic characteristics in terms of prominence (prosodic stress) across major speech genres of the *Rhapsodie* dataset (oratory, narrative, description, argumentation, procedural) and corresponding speaking tasks. These specific prosodic characteristics are likely to reflect communicative needs of speakers and listeners (interactions, uptakes, speaking turns, etc.).

Keywords: phraseology, prosodic constituents, prominence, salience, textometrics

1. Introduction

Our research examines the notions of phraseology and formulaic language in speech production on the basis of prosodic transcriptions indicating specific events in speech: boundary tones, pitch accents, disfluent segments, etc. (Yoo et Delais-Roussarie, 2009). We believe that such speech events coded in spoken corpora are relevant for identifying the prosodic characteristics of formulaic language.

Corpus-based studies of phraseology often exploit recurrent patterns detected using repeated segments, co-occurrences and pattern-matching

techniques to explore formulaic strings of written texts (Granger, 2005; Sitri et Tutin, 2016). This approach seems equally applicable to oral discourse. Following this approach, our initial objects of study are predictable and productive sequences of signs called *lexicogrammatical patterns* (lexical signs, grammatical constructions). Made of permanent **'pivotal'** signs and a more productive '*paradigm*', these patterns may be discontinuous and may or may not be syntactic constituents (Gledhill, 2011; Gledhill et al., 2017). For example:

§ et donc euh **c'est pour** *ça* **qu**'aujourd'hui *je suis* en italien en XXX …
§ c'est-à-dire § ouais § un mois **c'est pour** *ça* **que** *ça s'appelle* radio Timsit …
§ mais bien sûr donc <**c'est pour** *ça* bien sûr bien sûr **que** *je parlais* oui XXX …
§ **c'est pour** *cela* **que** *je tenais* à vous rencontrer la veille de notre fête …

We then explore the ways in which prosodic features may correlate with extended lexical patterns, as well as the extent to which prosody corresponds to patterns which have a particular register or discourse function. These lexicogrammatical patterns combined with prosodic features extracted from speech databases are possible methodological tools for identifying phraseological characteristics of oral discourse.

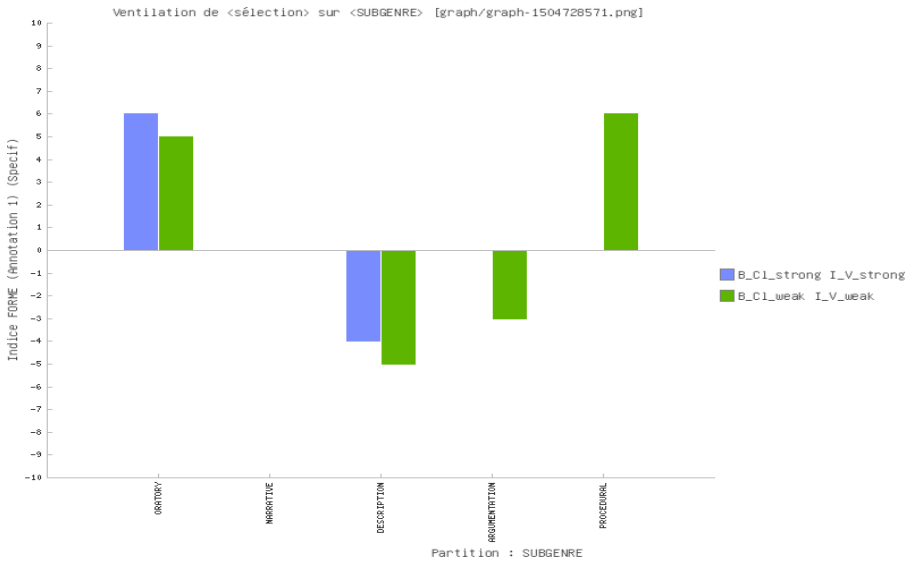


Figure 1: Characteristic elements of initial prominence across speech genres: Clit + Verb in initial position of Intonational PEriods (IPE)

2. The *Rhapsodie* speech database

Large spoken corpora are rarely distributed with a fine-grained prosodic annotation. Fortunately, for French, a reference corpus, the *Rhapsodie* speech database (ANR Rhapsodie 07 Corp-030-01), is freely available online (http://www.projet-rhapsodie.fr/). This syntactic and prosodic treebank is composed of 57 short samples of spoken French (approximately 5 minutes long), orthographically and phonetically transcribed (approximately 33,000 words).

2.1. Database structure

The *Rhapsodie* data file is available online in tabular form. The corpus data is structured as follows: all *Rhapsodie* texts are first identified by codes. Each text is further divided into separate units and segmented into tokens. The remaining columns display more than 60 linguistic annotations of the tokens, including microsyntax (rection, dependency, constituency), macrosyntax and prosody. This data set was transformed from the spreadsheet format into a *Trameur* base file using regular expressions (Fleury, 2013). The data structure is composed of two parts: (1) a *Thread*, which is a list of items with position identifiers; (2) a *Frame*, which is a list of corpus partitions defined on the *Thread*. Each partition has a name and a list of named constituents identified through their first and last token positions on the *Thread*. Thus, each annotated token from the *Rhapsodie* corpus becomes an item identified by its position on the *Thread* (Fleury et Zimina, 2014).

2.2. Prosodic annotation

The corpus covers several discourse types and speaking styles: oratory, narrative, description, argumentation, procedural; interactive, semi-interactive and non-interactive; public and private, planned, spontaneous and semi-spontaneous, etc. (Lacheret et al., 2017). The transcriptions and the annotations are aligned on the speech signal (Lacheret et al., 2014). A combination of manual and automated annotations allowed a segmentation of speech into prosodic periods (Lacheret et Victorri, 2002), which relies on the initial characterization of two types of speech events retained from the manual annotation: prosodic prominence and disfluencies. Organized around rhythmic and melodic components, the hierarchy of prosodic constituents includes: Intonational PEriods (IPE); Intonational PAckages (IPA): sub-constituents internal to periods; Rhythmic Groups (RG): sub-constituents internal to intonational packages; Metrical Feet (MF): sub-constituents inside rhythmic groups; Syllables, with Prominence levels, including: 0 (non-prominent), W (weak) and S (strong).

3. Quantitative analysis of the prosodic dimension of phraseology

As the link between the "marked status" as a +phrase/expression/formulaic expression etc. and prosodic constituents is still to be revealed, some of our research questions are of an exploratory nature and more than 60 layers of morpho-syntactic, syntactic, macro-syntactic and prosodic annotation in *Rhapsodie* necessarily open new perspectives for the exploration of the prosodic dimension of phraseology.

3.1. Preliminary research

Previous research on spoken phraseological units (Lin, 2013) did not take into account the prosodic hierarchy, in other words, the various sizes of the prosodic constituents (Nespor et Vogel, 2007). We first replicated this methodology, which consisted in describing the prosodic characterisations of phraseological units attested in speech-to-text transcription. For example, in the categorization of the stress hierarchy, we can distinguish between stressed (*strong*) and less prominent (*weak*) syllables.

Preliminary analyses of repeated segments from the *Rhapsodie* corpus, such as *jeune fille* (F=20) or *je veux dire* (F=21) led us to observe the non-congruency of the recurrence of prosodic features, such as prominence, and traditional phraseological units recovered from transcribed speech data:

… une|*dis-weak* **jeune**|***strong*** **fille**|***weak*** euh|*weak* habillée|*weak* tout|*strong* en|*strong* …
… c|*tail'*|*tail* est|*tail* une|*tail* **jeune**|***tail*** **fille**|***tail*** # pauvre|*strong* et|*strong* affamée|*strong* …
… dans|*strong* la|*strong* rue|*strong* avec|*weak* une|*weak* **jeune**|***weak*** **fille**|***filled-dis*** #
§ vous|*weak* voyez|*weak* ce|*weak* que|*strong* **je**|***strong*** **veux**|***strong*** **dire**|***strong*** #
§ euh|*tail* vous|*weak* voyez|*weak* ce|*weak* que|*weak* **je**|***weak*** **veux**|***weak*** **dire**|***weak*** §

Our first analyses of these examples made us simultaneously consider multiple prosodic properties of these collocations, as well as the necessity of taking into account the hierarchy of prosodic units corresponding to these speech contexts.

3.2. Prosodic constituents: a genre-based analysis of lexicogrammatical recurrence and initial salience

A recent study (Zimina et Ballier, 2017) has shown that prosodic segmentation into IPE: Intonational PEriods (segments of speech with

distinctive pitch and rhythm contours) available within the *Rhapsodie* database offers new insights for the observation of the functions of formulaic expressions in speech. Lexicogrammatical patterns at the beginning of IPE are strongly related to spoken formulaic language. Recurrent prominences observable after speech breaks can be revealed by textometric analysis of **repeated Part-Of-Speech (POS) segments** (Salem, 1987) at the beginning of the IPEs. This method can be used to isolate a set of pivotal elements associated with what is commonly perceived as a strong prosodic boundary. Computation of **characteristic elements** (Lebart et al., 1998), applied to **repeated segments** (Salem, 1987), describes these regularities with respect to genre and speaking styles, categorized in *Rhapsodie* as 'subgenres' (Lacheret et al., 2014).

Discovered variations of initial salience depend upon several factors (interactional needs, social context, genres, etc.). They reflect specific communicative needs of the speakers. For example, **Cl + V** is a positive characteristic element at the beginning of the IPE in the speech contexts of oratory genre (specificity index: +10). The following examples reveal some lexicogrammatical realizations of this productive pattern in *Rhapsodie* (categories corresponding to **Cl + V** appear in bold):

il faut les faire grandir
§ **ce sera** un coup franc
je souhaite que l'Europe

These pivotal elements reflect the structure of regular rhetorical units with a predictable/definable discourse function (performative utterances).

3.4. Prosodic salience and prominence: systemic combination

To fine-tune our study of regular prosodic features of phraseology, we have added another layer of analysis, namely the prominence of final syllables. For these purposes, we have combined three annotation levels: (1) the positional properties of units within the IPE structure with *BILOU* tags (*Beginning*, *Inside*, *Last*, *Unit*-length and *Outside*); (2) POS tags; (3) final stress prominence: *strong*, *weak*, *pause_* and % (inaudible or non-transcribed due to overlap). The base file has been automatically re-annotated with *Le Trameur* to add this new combined annotation layer to the *Rhapsodie* data. Repeated segments have then revealed that at the beginning (B) of IPEs, the final prominence of the pivot **Cl + V** is either *weak weak* (F=114) or *strong strong* (F=95). On Figure 1, the results of characteristic elements analysis in different speech genres show that these weak realisations of final syllables are positive characteristic elements (specificity index: +06) of procedural utterances, such

as instructions in travel planning, while both strong (+06) and weak (+05) realisations of **Cl + V** are characteristic elements of initial prosodic salience in oratory genre. A finer-grained analysis according to speaking task, presented on Figure 2, shows that this prosodic richness can be attributed to the subgenre of political discourse, a well-known subtle and complex discourse phenomenon (Dorna, 1995; Mayaffre, 2002). For political speech, pragmatic strategies influence the choice of a *weak weak* (+06) or of a *strong strong* (+03) sequence to realize specific discourse functions, for example:

... reculer la pauvreté # § **ce**|***strong*** **sera**|***strong*** tout le sens du combat de la France… (*focus*) … ces valeurs # § en les faisant vivre # **nous**|***weak*** **serons**|***weak*** plus forts pour aborder les temps qui viennent… (*fonction performative*)

The strong prominence of **Cl + V** also corresponds to emphatic realisations at the beginning of IPE in sermons (+03), as evidenced in Figure 2. Similarly, advertising favours recurrent overuse of stressed syllables in this position (+04).

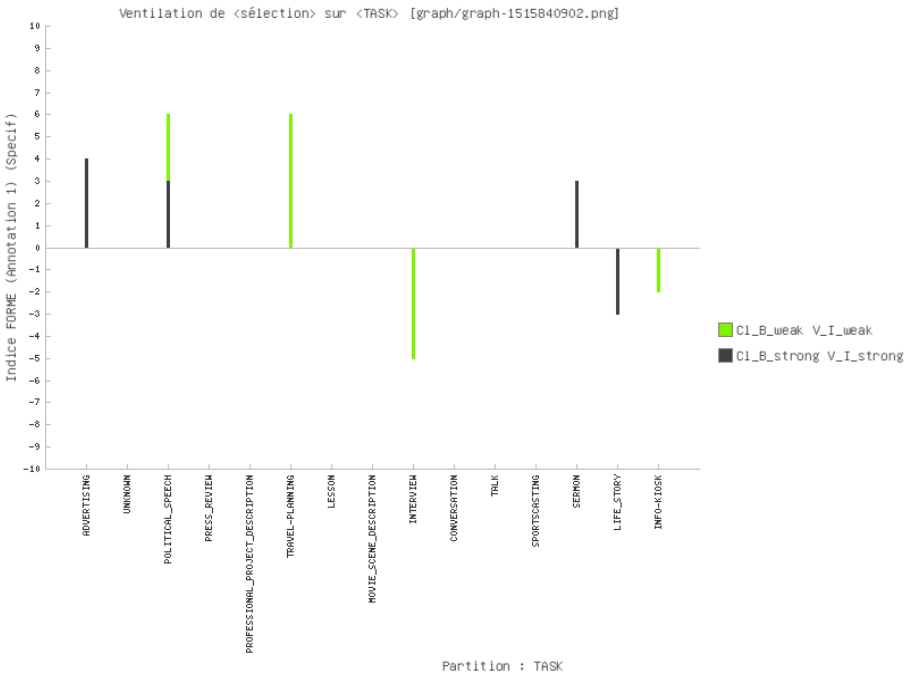


Figure 2: Characteristic elements of initial prominence in different speaking tasks: Clit + Verb in initial position of Intonational PEriods (IPE)

4. Conclusions

The prosodic hierarchy (Nespor et Vogel, 2007) acknowledges several layers

of granularity, from the prosodic utterance to the phoneme. For our investigation, the IPE was the best initial candidate for the proper level of granularity in the prosodic hierarchy. There were structural reasons for this, such as the fact that speaking turns were likely to be signalled in initial position of IPE. The textometric analysis of this prosodic constituent of the prosodic hierarchy has shown revealing features such as the limited distribution of POS categories in the initial position, as well as the role of prosodic prominence (stressed syllables) and its relevance for the distinction of speech genres. It appears that the recurrent patterns reflected by such sequences as *"je salue"*, *"elle souhaite"*, *"il faut"*, *"on continue"* are not unlike the stable lexicogrammatical patterns that can be observed in written data. In all likelihood, the initial characteristic distributions with specific prosodic characteristics correspond to communicative needs (interactions, uptakes, speaking turns, etc.).

Because of the complexity of the *Rhapsodie* speech database, we regard these explorations as preliminary: we have only based our analysis on few layers of annotations. Besides, other layers of granularity within the prosodic hierarchy might also be relevant: Intonational PAckage (IPA), Rhythmic Group (RG), etc. We also surmise that other variables (channel, planning type, event structure: monologal vs. dialogal tasks) are likely to reveal related features.

5. Future research

In future work, various lines of investigation can be pursued, such as the examination of the various layers of the prosodic hierarchy. Looking for collocational structures may lead us to question the recurrence of patterns within prosodic units, in other words, the embedding of prosodic constituents or the complexity of the boundaries of the constituents across the layers of the prosodic hierarchy. Other correlates might be considered such as duration and prosodic contours (the whole corpus also includes the sound files). It may well be the case that annotators were influenced by the genre of the recordings, and auditory analysis across genres based on our characteristic elements analysis may nuance the levels of prominence assigned by the annotators. If the identification of these collocational prosodic patterns is robust, it should also remain transparent and decisive for subjects when resynthesized. The acoustic signal can be modified so as to erase lexical contents, only keeping the melody (humming). Resorting to humming should enable us to test the relevance of prosodic sequences which should robustly remain identifiable as characteristic signals of collocations in perception tests.

References

Dorna, A. (1995). Les effets langagiers du discours politique. *Hermès, La Revue* 1995/2 16, 131–146.

Fleury, S. (2013). *Le Trameur. Propositions de description et d'implémentation des objets textométriques*. Sorbonne nouvelle – Paris 3, http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf

Fleury, S., Zimina, M. (2014). Trameur: A Framework for Annotated Text Corpora Exploration. In: *Proceedings of 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland, pp.57–61, http://www.aclweb.org/anthology/C14-2013.pdf

Gledhill, C. (2011). The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts. *ASp (Anglais de Spécialité)* 59, 05–23.

Gledhill C., Patin S., Zimina M. (2017). Identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *CORPUS* 17, 113–143.

Granger, S. (2005). Pushing back the limits of phraseology. How far can we go? In: Cosme, C., Gouverneur, C., Meunier, F., Paquot, M. (eds.): *Proceedings of PHRASEOLOGY 2005. An Interdisciplinary Conference*, Université Catholique de Louvain, Louvain-la-Neuve, pp. 165–168.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J-P., Obin, N., Pietrandrea, P., Tchobanov, A. (2014). *Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Lacheret, A., Kahane, S., Pietrandrea, P. (eds.) (2017). *Rhapsodie: a prosodic and syntactic treebank of spoken French*, John Benjamins, Amsterdam-Philadelphia.

Lacheret, A., Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum* 24 (1-2), 55–73.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht, Boston.

Lin, Ph. M.S. (2013). The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics* 18(4), 561–588.

Mayaffre, D. (2002). Discours politique, genres et individuation socio-linguistique. In: Morin, A., Sébilot, P. (eds.). *Actes des JADT 2002*, Saint-Malo, France, IRISA-INRIA, pp.517–529.

Nespor, M., Vogel, I. (2007). *Prosodic Phonology*. Berlin. Mouton De Gruyter.

RHAPSODIE Homepage, http://www.projet-rhapsodie.fr

Salem, A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Klincksieck, Paris.

Sitri, F., Tutin, A. (dir.) (2016). Phraséologie et genres de discours. Patrons, motifs, routines. *LIDIL* 53.

Yoo, H-Y, Delais-Roussarie, E. (eds.) (2009). *Actes de la conférence Interface Discours & Prosodie (IDP 2009)*, Paris, France, http://makino.linguist.jussieu.fr/idp09/actes_fr.html

Zimina, M., Ballier, N. (2017). Intonational PEriods (IPE) and Formulaic Language: A Genre-based Analysis of a French Speech Database. *Proceedings of Europhras 2017 Conference: Computational and Corpus-based Phraseology: Recent Advances and Interdisciplinary Approaches*, London, http://www.tradulex.com/varia/Europhras2017-II.pdf

Abstracts

What kind of contributions does research provides? Mapping issue based statements in research abstracts

Filippo Chiarello[1], Gualtiero Fantoni[1],
Andrea Bonaccorsi[1], Silvia Fareri[2]
[1]School of Enginner, University of Pisa – filippo.chiarello@destec.unipi.it
[2]Marco Biagi Foundation, University of Modena and Reggio Emilia

Abstract
Sentiment analysis is the study of the polarity (positive/negative) of documents (Pang and Lee 2008). Lexicon based techniques are one of the most used methods to compute the polarity of a document. Lexicons are collections of words, each annotated with its own positive or negative orientation. The overall sentiment of a document is therefore computed upon the prior polarity of the contained words.
Given that one of the most used approach to build sentiment lexicons is terms extraction from polarized documents (e.g. restaurant or movie reviews, social network post), we may hesitate to apply it to styles of text dramatically different from what they were validated on. Lexicons need to be redesigned every time we want to shift from one context to another.
Another well-known problem of lexicon approaches is that writers make use of valence shifters. These are words that affects the polarized words: a negation flips the sign of a polarized word (e.g., "it is not good"), an amplifier increases the impact of a polarized word (e.g., "I totally enjoy that."), or a de-amplifier reduces the impact of a polarized word (e.g., "it is almost perfect"). So if valence shifters occur frequently a simple dictionary lookup may not be measuring the sentiment appropriately.
In this paper, we try to address both these problems with a specific focus on extracting negative sentences from abstracts of scientific articles. We thus propose a negative sentence extractor for appropriately dealing with the research paper domain. We will start with the collection of a set of abstracts belonging to the same field of knowledge. These documents are then pre-processed using state of the art natural language processing tools (sentence splitter, tokenizer, lemmatization). Then for each sentence we will compute a negative sentiment polarity and take in to consideration only the sentences having a negative polarity score below a thresh-hold level. Finally we will apply topic modelling algorithm on the negative sentences, with the aim of give a graphical synthesis of the main problems of a research field.
For the polarity computation, we will rely on a lexicon developed by the authors that extracts advantages and disadvantages of inventions from

patents (Chiarello, 2017) and a novel dictionary lookup approach that tries to incorporate weighting for valence shifters (Rinker 2018). Following a bottom up-approach we will redesign these lexicon for an optimal application on paper documents.

References

Pang B. and Lee L. (2008). Opinion Mining and Sentiment Analysis, *Foundation and Trends in Information Retrieval,* vol.(2)

Chiarello, F., Fantoni, G., Bonaccorsi, A., (2017). Product description in terms of advantages and drawbacks. Exploiting patent information in novel ways. ICED

Rinker, T. W. (2018). sentimentr: Calculate Text Polarity Sentiment version 2.3.2. http://github.com/trinker/sentimentr

Technical sentiment analysis: predicting the success of new products using social media

Filippo Chiarello[1], Giacomo Ossola[1], Gualtiero Fantoni[1],
Andrea Bonaccorsi[1], Andrea Cimino[2], Felice Dell'Orletta[2]
[1]School of Engineer, University of Pisa – filippo.chiarello@destec.unipi.it
[2]Institute for Computational Linguistics of the Italian National Research Council

Abstract

Nowadays, social media has become an inseparable part of modern life, providing a vast record of mankind's everyday thoughts, feelings, and actions. For this reason, there has been an increasing interest in research of exploiting social media as information source of knowledge although extracting a valuable signal is not a trivial task. In fact, social media data is noisy and must be filtered before proceeding with the analysis. In this domain, sentiment analysis, which aims to determine the sentiment content of a text unit, is considered one of the best data mining method. It relies on different approaches (Collomb, 2013): machine learning, lexicon-based, statistical and rule-based.

In this work, we try to understand if sentiment analysis is really the best available method to analyze consumer's perception of products. In particular, we compare state of the art sentiment analysis based on machine learning methods with a lexicon approach based on a dictionary of advantages and drawbacks related to products, important aspects evaluated by consumers during the buying decision process. The lexicon has been developed by researchers of the authors (Chiarello, 2017) to extract advantages and drawbacks of inventions from patents.

Our work started with the selection of an event able to polarize Twitter users' attention and a set products to analyze. In particular, we chose a premiere trade-show for the video game industry, and two video game consoles disclosed during the event. We collected tweets about products published before, during and after the trade-show. Since social media data is noisy (for example it may contains spam and advertising), before proceeding with the analyses, we filtered our dataset. In particular, after removing too short and non-English tweets, we manually classified a randomly extracted subset of posts to train the automatic classifier which provide us the cleansed dataset. Finally, we built product-related clusters of tweets.

Once obtained the final dataset, we conducted a sentiment analysis of the posts using state of the art machine learning techniques. We classified each tweet as positive, negative or neutral. Then, we applied our lexicon

identifying advantages tweets and drawbacks tweets. In order to compare the outputs of the two analyses, we considered advantages tweets as positive, drawbacks as negative, and tweets without words from our lexicon as neutral. We found consistent and interesting differences between the two methodologies. In particular we found that when a product has a certain technological complexity and fuels a more technical social media discourse, sentiment analysis seems to be less performing while advantages and drawbacks analysis is abler to produce technical-functional judgements about the products.

References

Collomb C. , Costea ,C. and Brunie L. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation.

Chiarello, F., Fantoni, G., Bonaccorsi, A., (2017). Product description in terms of advantages and drawbacks. Exploiting patent information in novel ways. ICED

Citizens and neighbourhood life: mapping population sentiment in Italian cities

Fiorenza Deriu, Domenica Fioredistella Iezzi[2]
[1]Sapienza University of Rome – email@provider.be
[2]Tor Vergata University – email@provider.be

Abstract 1
In recent years, citizens are increasingly taking an active part in neighbourhood life by writing comments on social media. Citynews is an online platform, that is active in 45 Italian provinces, and offers to readers facts and news about Italy. It records every month 85 million visits and counts over 565 thousand registered users. It has the characteristic of being a citizen journalism: an increasing part of these contents are in fact write to the platform by readers. We scraped the comments of users a sample of 6,418 posts published on the Citynews site from January 2017 to January 2018, on 9 metropolitan cities (Bari, Bologna, Florence, Genoa, Milan, Naples, Rome, Turin and Venice). We applied two analysis procedure: 1. Exploratory analysis of the contents of citynews posts with the aim of classifying the opinions of active citizens; 2) Sentiment analysis of the sampled cities, with assignment of scores by postcode. In the first phase, we mapped the contents of the reviews and identify the areas in which citizens report major events (positive or negative) occurred in the zone. In the second phase, we built a sentiment index, normalized by postcode and designed the maps of citizenship mood.

Abstract 2
Negli ultimi anni, i cittadini partecipano sempre più attivamente alla vita di quartiere scrivendo commenti sui social media. Citynews è una piattaforma online, attiva in 45 province italiane, che offre ai lettori notizie e commenti sull'Italia. Registra ogni mese 85 milioni di visite e conta oltre 565 mila utenti registrati. Ha la caratteristica di essere un *citizen journalism*: una parte crescente di questi contenuti è infatti scritta sulla piattaforma dai lettori. Abbiamo raccolto i commenti degli utenti su un campione di 6.418 post pubblicati sul sito Citynews da gennaio 2017 a gennaio 2018, su 9 città metropolitane (Bari, Bologna, Firenze, Genova, Milano, Napoli, Roma, Torino e Venezia). Abbiamo applicato due procedure di analisi: 1. Analisi esplorativa dei contenuti dei post delle città con l'obiettivo di classificare le opinioni dei cittadini attivi; 2) Analisi del sentimento delle città campionate, con assegnazione dei punteggi per codice postale. Nella prima fase dell'analisi, abbiamo mappato i contenuti delle recensioni e identificato le aree in cui i cittadini segnalano i principali eventi (positivi o negativi) accaduti nella zona. Nella seconda fase, abbiamo costruito un indice di sentiment normalizzato, e disegnato le mappe dell'umore della cittadinanza.

Keywords: active citizenship, map, sentiment analysis.

Vax network: profiling influential nodes with social network analysis on twitter

Francesca Di Carlo, Rosy Innarella, Brizio Leonardo Tommasi
Tor Vergata University – francesca.dicarlo90@gmail.com;
rosy.innarella@gmail.com; brizio.tommasi@gmail.com

Abstract

We live in a society increasingly conditioned by opinions of third parties. If our objective is to spread a new concept or idea effectively into the society and the public opinion, we must consider two basic factors: the presence of influential individuals that accelerate the spread of the process and the susceptibility of people to be "infected" by the new idea. In order to keep under control the spread of new behaviors, it is very important to collect as much as information about the profiles of the people belonging to the two categories (Sawyer, 2011). This paper analyses a social network based on two clusters of people. The first cluster identifies people with opinion "pro-vax" the second one identified by "anti-vax" opinions emerged in the last three years. This analysis consists of a description of a data collection of public tweets available online; every collection is based on about one month of tweets per year. Tweets are extracted with specific "key-words" in the context of vaccine and anti-vaccine factors. In particular the data collection is based on a "key-words search model" consisting of a combination of context words, for example: (i) "*core cluster*" with: vaccination, vaccine, vaxxer, antivaccine, antivaxxer, anti-vaccine, anti-vaxxer; (ii) "*effects cluster*" with vaccine dependent variable: Autism, MMR**,** Pharma and vaccine; (iii) "*community cluster*" with vaccine dependent variable: Cdc, Aaps, Fda, Hrc, Wakefield and vaccine. We used twitter fetcher included in the semantic and social network analysis software Condor (Gloor, 2009); in this way we could fetch 120 days of tweets distributed over three years: apr-2015, nov-2016, jun-2017. The total volume of collected tweets is about 300,000. All the collected tweets were written in English. In this elaboration we were able to distinguish behavior emerged in tweets in each cluster analysis by carrying out a preliminary sentiment analysis of the collected tweets. This network of Twitter, named "*vax network*", consists of about 800,000 links with an average degree of 3 relations. The analysis of "*vax network*" is based on the mainly centrality measures (e.g. degree, betweenneess, closeness, clustering coefficient, eigenvector) for the identification of relevant node in terms of potential influence of "*vax networks"*. Instead of the sentiment analysis is based on main methods (e.g. word frequencies, activity, emotionality, sentiment, complexity) of natural language used in the tweets of "*vax network*", for profiling the relevant node in terms of its main characteristic correlated with own centrality in the network.

Consequently we can identify the profiled nodes that have prevalent behavior for influencing or clustering the trend of the *"vax network"*. In the last three years we have identified a specific trend of the sentiment analysis, calculated on the twitter discussions about vaccine network. The sentiment analysis of the tweets has a negative meaning, regardless of using positive or negative words in the vaccine context. In general the 80/20 law emerges between "pro" and "anti" nodes. The sentiment together with emotionality has a growing positive trend, due to an awareness communication on vaccine utility. **Keywords:** network analysis, sentiment analysis, text mining, twitter, profiling, influence, vaccine, complex network.

Alteryx

Davide Donna
Managing Partner, The Information Lab

Alteryx is an analytics platform that allows users to build complex workflows for an end to end data management and analytics without scripting. This is possible through an intuitive interface where users drag and drop on the canvas the different tools that he needs to connect, blend, transform, analyze and export the data.
Alteryx is a Data Science platform that includes three macro areas: ETL with capability to connect to most of the databases and services, clean, prepare and blend the data; Advance Analytics, integrating several predictive and statistical models scripted in R and Pyton; Geospatial analytics.
In the textual data analytics contest Alteryx finds several applications, presented during the speech:
- web scraping: download html pages and parsing data through regular expressions (Regex)
- big data management
- Connection to external services for advance analytics. I.e. with the direct interaction with Azure Cognitive Services Alteryx can implement key phrases extraction, sentiment analysis and language dectection
- Connection to external services for data input: i.e. with the connectors to Twitter and other social media the information can be directly imported into the analytical platform
- Integration of R and Python libraries: the models are written into Alteryx and integrated and run within the entire flow
- Cleaning, blending and transformation of the information downloaded from different sources
- Export the results in the favorite format. I.e. Visualanalytcs or reporting tool like Tableau for a clear and smart representation.

Complexity of US President Speeches

Valerio Ficcadenti[1], Roy Cerqueti[2], Marcel Ausloos[3,4]
[1]Marche Polytechnic University - v.ficcadenti@univpm.it
[2]University of Macerata - roy.cerqueti@unimc.it
[3]University of Leicester - ma683@le.ac.uk
[4]GRAPES - Angleur, Liege, Belgium - marcel.ausloos@ulg.ac.be

Abstract
This work is devoted to the exploration of the rhetoric dynamics of a large collection of US Presidents' speeches. In particular, speeches are viewed as complex systems and are first analysed through rank-size laws, being the words of each speech ranked in terms of their frequencies. The building of the dataset itself represents a relevant step of the study. In this respect, by using a web scraping routine on the Miller Center website, a large span of 978 speeches have been downloaded. After a pre-processing phase the set is reduced to 951; for each one, the words' frequencies are stored. A best fit procedure with Zipf-Mandelbrot laws is performed over the 951 talks individually. Thanks to these estimations, it is possible to reach interesting conclusions on how 45 United States Presidents, from April 30, 1789 till February 28, 2017 delivered political messages. Our analysis shows some remarkable regularities, not only inside a given speech, but also between different speeches. Results are discussed under a political and linguistics points of view.

Keywords: US Presidents' speeches, speeches' framework, US Presidents' rhetoric, rank-size analysis Zipf-Mandelbrot law

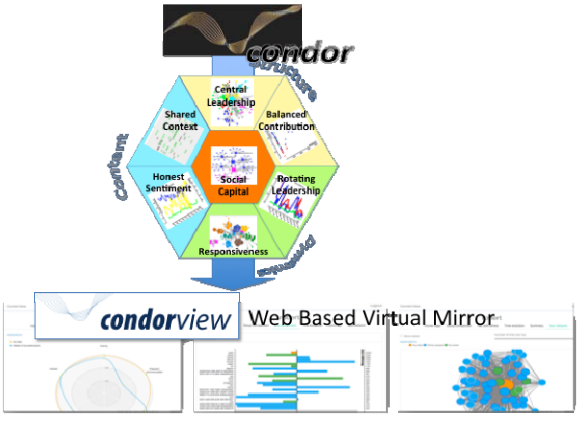
Measuring the Dynamics of Social Networks with Condor

Peter A. Gloor
MIT Center for Collective Intelligence, Cambridge MA – pgloor@mit.edu

Condor is an easy-to-use dynamic semantic social network analysis tool, which automatically imports many types of communication data such as e-mail, Twitter, Blogs, Wikipedia, and Facebook, converting it into social networks for longitudinal network, content and sentiment analysis using machine learning and deep learning. It has been developed over the last 15 year at the MIT Center for Collective Intelligence together with other universities (Colgone, Rome Tor Vergata, etc.).

Condor consists of four parts:

1. A series of fetchers to directly load data from e-mail, for example from Gmail or Exchange, from Twitter, from Google, Wikipedia, and Facebook.
2. Interactive preprocessing functions to modify and reduce the graph, filter by content and by structure, annotate by geocode, merge multiple e-mail addresses, and create modified graphs.
3. Visualization functions, to show the static network, a dynamic movie of the network over time, geographical word maps, and different views for structure, content, and sentiment of actors.
4. Export functions to export time series of all variables for later longitudinal analysis in statistics packages such as R or SPSS (or Excel).



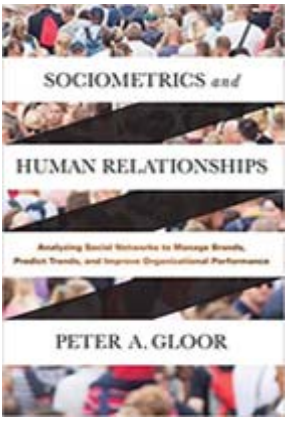
Condor can….

…Analyze structure, dynamics, and content of E-Mail, constructing the communication matrix between different influencers.

…Coolhunt the Internet using Twitter, Blogs, and Facebook for trends and trendsetter to analyze interactions among influencers in different online information spheres (Twitter, Blogs, Forums, Wikipedia,…)

…Measure collaboration and communication efficiency on seven proprietary metrics (honest signal of collaboration) by structure, time, and contents: central leadership, rotating leadership, balanced contribution, responsiveness, honest sentiment, shared context, social capital.

…Find digital tribes on social media using deep learning, and mapping tribal affiliations to brands, products, and people.



Requirements

Condor runs on Windows, Mac, and Linux, and needs Java (MySQL optional) installed.
Download Condor and the Condor manual
from http://guardian.galaxyadvisors.com/guardian

Condor is described in Gloor, P. Sociometrics and Human Relationships: Analyzing Social Networks to Manage Brands, Predict Trends, and Improve Organizational Performance , Emerald Publishing, London 2017

Contact: Peter A. Gloor, MIT Center for Collective Intelligence and Galaxyadvisors AG
www.galaxyadvisors.com, www.transparencyengine.org, pgloor@mit.edu

"BIG DATA" Words Trend Analysis using the multidimensional analysis of texts

Iolanda Maggio[1], Domenica Fioredistella Iezzi[2], Matteo Fatighenti[3]
[1]Rhea Group – iolanda.maggio@esa.int
[2]Tor Vergata University – stella.iezzi@uniroma2.it
[3]CapGemini – matteo.fatighenti@gmail.com

Abstract
Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information. Big Data includes so many specialized terms that it's hard to know where to begin. An evolution in languages was experienced in the last few years and this paper shows how the relevant terms are changing. This analysis started as a complement of a thesis activity for the Master Programme in Data Science of University of Tor Vergata. To perform the text analysis the IRaMuTeQ software has been used that will be described in the next paragraphs. The exercise aims at analyse both contents with a tool/software for multidimensional analysis of texts to give evidence of BIG DATA words trends. This approach provides the users with different text analyses, either simple ones, such as the basic lexicography related to lemmatization and word frequency; or more complex ones such as descending hierarchical classification, post- hoc correspondence factor analysis and similarity analysis. The vocabulary distribution is presented in a comprehensive and clear way with graphical representations derived from the lexicographic analysis. These analyses can be performed using texts referring to a certain thematic (text corpus) grouped in one text archive; or data from spreadsheets (matrices with individuals in a row and words in a column), like the dataset derived from free evocation tests.

Keywords: big data, text mining, Zipf diagram, Clustering, Dendograms, Corpus.

Itinerari turistici, network analysis e text mining

Mario Mastrangelo
Tor Vergata University – mario.mastrangelo@uniroma2.it

Abstract

Appare ormai evidente come gli strumenti propri del web 2.0 abbiano modificato in maniera sostanziale il mondo in cui viviamo, e che le ricadute di tali strumenti siano di grande entità soprattutto in alcuni settori economici. Tra questi, sicuramente il turismo. Le opportunità fornite dai cosiddetti user generated Contents -e dall'avvento dei Big Data- hanno trasformato sensibimente questo settore, sia dal lato della domanda, in termini di scelta delle destinazioni, di pianificazione del viaggio, di soddisfazione in merito ai servizi e ai fornitori, sia dal lato dell'offerta, in termini di analisi dei flussi, di marketing, e di customer satisfaction. L'ipotesi, sempre più suffragata dai fatti, è pertanto quella che per una quota crescente dei turisti contemporanei il parere espresso da propri pari tramite i social media, sia di tipo specifico, come Tripadvisor, che di tipo generico, come ad esempio Facebook e Twitter, sia ormai da porre sullo stesso piano e in certi casi possa risultare addirittura più incisivo di gran parte dei messaggi veicolati dai canali informativi tradizionali. In questa direzione, il presente contributo intende illustrare una applicazione pratica di alcune tecniche di analisi dei dati testuali (analisi delle corrispondenze lessicali, text clustering, sentiment analysis, etc) che, impiegate in serie o in parallelo, possono contribuire a fornire un valido supporto per l'individuazione di itinerari turistici sempre più individuali e "sociali" in un determinato contesto territoriale.

In particolare, in questa simulazione sono stati presi in considerazione 38 siti di Roma, tutti molto noti, piuttosto diversificati - da monumenti, a palazzi istituzionali, a sedi museali – e tutti vicini tra loro, così da poter creare piccoli itinerari raggiungibili a piedi. Questi siti divengono i nodi di un grafo orientato, i cui archi costituiscono le distanze tra ciascun sito, caratterizzati da una serie di attributi, alcuni dei quali di natura "tradizionale" come la classificazione iniziale in base a parametri storico-artistico-funzionali (periodo romano, periodo barocco, periodo rinascimentale, periodo contemporaneo, siti museali e siti "misti", ovvero quelli che non rientrano con esattezza in nessuna delle precedenti categorie) e altri derivanti dai social media, nel caso specifico mediante opportune operazioni di text mining su

un Corpus assemblato a tal fine da un insieme di recensioni in lingua inglese dei 38 siti su Tripadvisor.
La rappresentazione dei siti sotto forma di grafo, insieme ai due insiemi di attributi associati ai nodi, permette di eseguire con facilità alcune operazioni a partire dai desiderata del turista; ad esempio suddividere i siti in base a come i turisti parlano di loro, e individuare insiemi di percorsi tra i siti caratterizzati da un sentiment più elevato, e dunque dal fatto che altri turisti, in posizione simmetrica, hanno espresso con le loro parole un alto gradimento. In altre parole, dati un punto di partenza e un punto di arrivo, a percorsi più o meno tradizionali, definiti in modo "oggettivo" (ad esempio, tour barocco, tour religioso, e, in prospettiva, tour gastronomico, etc) si potrebbero aggiungere percorsi definiti in modo soggettivo sulla base delle valutazioni espresse dagli altri turisti (ad esempio, tour delle tappe più ammirate, più emozionanti, più commentate, etc). In tal modo si unirrebbero le due anime attuali dell'informazione turistica, quella tradizionale indispensabile ma più statica e quella basata sui contenuti generati da altri utenti, effimera e soggettiva ma più dinamica; dal punto di vista scientifico, questo approccio permette di applicare tecniche ben consolidate ma raramente utilizzate in maniera congiunta, allo scopo di estrapolare, interpretare ed impiegare in maniera efficace ed originale la grande mole dei contenuti generati dai vari utenti in un settore particolarmente fecondo in questo senso.

Text Mining per l'analisi qualitativa e quantitativa dei dati amministrativi utilizzati dalla Pubblica Amministrazione

Maria Francesca Romano[1], Guido Rey[2],
Antonella Baldassarini[2] Pasquale Pavone[4]
[1]Istituto di Economia, Scuola Superiore Sant'Anna, [2]Istituto di Management, Scuola Superiore Sant'Anna, [3]ISTAT, [4]Centro di Analisi delle Politiche Pubbliche – Università di Modena e Reggio Emilia

Abstract
Nella Pubblica Amministrazione la maggior parte delle informazioni (sia qualitative che quantitative) è contenuta in testi in linguaggio naturale e spesso i testi "nascondono" al loro interno anche molti dati numerici. Articoli di giornale e interventi in dibattiti televisivi richiamano spesso l'attenzione dell'opinione pubblica sulla mancata o incompleta informatizzazione della giustizia. Un esempio positivo sono le sentenze emesse dalla Corte di Cassazione già visualizzabili ed interrogabili sul sito www.italgiure.it. Si tratta di un totale di oltre 475.000 documenti a partire dall'anno 2012. E' già stata effettuata l'estrazione dal sito Italgiure di un sottoinsieme di sentenze di merito (circa 4.700), e le analisi testuali condotte con il software TalTac2.0 mostrano (Romano 2017) come sia possibile giungere a buoni risultati nella verifica della :

- completezza rispetto alla individuazione delle parti offese, degli imputati, delle eventuali parti civili, e di enti, /aziende pubbliche e private: I nominativi sono tutti in chiaro ed esistono con pochissime eccezioni, coperte da omissis.
- attendibilità / certezza rispetto ad eventi criminosi e successiva elaborabilità con strumenti informatici
- identificazione dei soggetti (Enti Pubblici) e del ruolo di imputati e di enti / aziende private;
- dell'ammontare economico degli atti criminosi collegati ad attività economiche (corruzione, aste pubbliche, gare di appalto, ecc.);
- luoghi in cui si sono svolti gli atti criminosi.
- presenza di appartenenti ad organizzazioni criminali e loro insediamenti

Il paper discute i problemi metodologici derivanti dall'uso di testi in linguaggio naturale di carattere giuridico, propone una metodologia di estrazione delle informazioni rilevanti ed i vantaggi derivanti da una

integrazione tra dati e informazioni estratti dalle sentenze di merito e archivi di natura amministrativa.
Sarà discussa la possibilità di integrare più basi di dati, mettendo in relazione i dati estratti dalle sentenze con i dati presenti in altre basi di dati anche di natura statistica; saranno presentati i risultati di un esercizio su di un insieme di sentenze della Corte di Cassazione da collegare con archivi statistici come, ad esempio, il registro delle imprese attive tracciando possibili scenari conoscitivi per un'analisi economica di eventi criminosi.
Componente centrale dell'analisi è il confronto con altre basi di dati, statistiche e testuali, al fine di acquisire elementi qualitativi e quantitativi su particolare fenomeni criminosi, nonché convalidare la correttezza dei dati estratti dalle sentenze non sempre esenti da errori.

Riferimenti bibliografici

Romano M.F. (2017), Dalle parole ai numeri: estrarre dati dalle sentenze della magistratura, in Rey G.M. (a cura di), La mafia come impresa. Analisi del sistema economico criminale e delle politiche di contrasto, FrancoAngeli, pp.121-154.

Taglio cesareo e Vbac in Italia al tempo dei Big Data: una proposta di ulteriore contributo informativo.

Alessandro Cesare Rosa
Dipartimento di Epidemiologia del S.S.R. del Lazio – a.rosa@deplazio.it

Abstract 1
The project starts with the curiosity to investigate, for a topic of epidemiological (but also popular) interest such as the caesarean section in Italy, which is the level of information and awareness acquired by a sample of citizens who frequents and writes an opinion on the web, comparing freely. Conceptually linked to the caesarean section, the theme of VBAC, acronym of "Vaginal Birth After Cesarean", will be analyzed in parallel with the same methodology. The objective is to highlight connections, if they exist, between aspects related to the correct practice of the caesarean section that the woman and / or the single health care provider should adopt (ministerial guidelines created for the public) with the opinion expressed by a sample of web-users. To do so, it was decided to scrape, from the web, the textual content of the opinions present in some websites and online forums of frequentation, precisely in order to extract, where possible, semantic macro-areas comparable to the concepts of departure

Abstract 2
Il progetto nasce dalla curiosità di indagare, per una tematica di interesse epidemiologico (ma anche divulgativo) quale il taglio cesareo in Italia, quale sia il livello di informazione e consapevolezza acquisita da un campione di cittadini che frequenta e scrive la propria opinione sul web, confrontandosi liberamente. Concettualmente legato al taglio cesareo, anche il tema del "Parto Vaginale dopo Cesareo",acronimo inglese VBAC, verrà analizzato in parallelo con la medesima metodologia. Ci si pone l'obiettivo di evidenziare eventuali connessioni tra gli aspetti legati alla corretta pratica del taglio cesareo che la donna e/o il singolo operatore sanitario dovrebbero adottare,delineati nelle linee guida ministeriali rivolte al pubblico, con la percezione in merito espressa da un campione di web-users. Per fare ciò, si è deciso di attingere, dal web, al contenuto testuale delle opinioni presenti in alcuni siti e forum online di stampo e frequentazione volutamente generalista, proprio al fine di elicitare,dove possibile,macro-aree semantiche confrontabili con i concetti di partenza.
Keywords: Exploratory Textual Data Analysis, Text Mining, Discourse analysis.

Finito di stampare in proprio
nel mese di giugno 2018
UniversItalia di Onorati s. r. l.
Via di Passolombardo 421, 00133 Roma Tel: 06/2026342
email: editoria@universitaliasrl. it – www. universitaliaeditrice.it