Conditions of cognitive plausibility of computational models of category induction

Daniel Devatman Hromada

Laboratoire Cognition Humaine et Artificielle (ChART)

Universite Paris 8

hromi@wizzion.com

Abstract. We present two axiomatic and three conjectural conditions which a model inducing natural language categories should dispose of, if ever it aims to be considered as "cognitively plausible". 1st axiomatic condition is that the model should involve a bootstrapping component. 2nd axiomatic condition is that it should be data-driven. 1st conjectural condition demands that the model integrates the surface features – related to prosody, phonology and morphology – somewhat more intensively than is the case in existing Markov-inspired models. 2nd conjectural condition demands that asides integrating symbolic and connectionist aspects, the model under question should exploit the global geometric and topologic properties of vector-spaces upon which it operates. At last we shall argue that model should facilitate qualitative evaluation, for example in form of a POS-i oriented Turing Test. In order to support our claims, we shall present a POS-induction model based on trivial k-way clustering of vectors representing suffixal and co-occurrence information present in parts of Multext-East corpus. Even in very initial stages of its development, the model succeeds to outperform some more complex probabilistic POS-induction models for lesser computational cost.

Keywords: categorization, part-of-speech induction, surface features, vector spaces, categorization-oriented Turing Test, partitioning of grammatical feature space, K-means clustering, cognitive plausibility

1. Introduction

The notion of "cognitive plausibility" and "part-of-speech induction" shall be defined in subsection 1.1. Subsection 1.2 shall clarify the position of syntactic category induction within the field of Natural Language Processing (NLP). The last subsection (1.3) shall offer a brief overview of the history of the problem, arguing that the current paradigm is probabilistic and English-centered one.

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

1.1 Cognitive plausibility

This article enumerates some basic conditions which should be fulfilled, we believe, by engineers aiming to transform their computational models into "cognitively plausible" artificial agents. **We label as "cognitively plausible" a model which tends to** address some basic function of human cognitive system not only by simulating, in a sort of "black-box apparatus", the mapping of inputs (stimuli, corpus data etc.) upon outputs (results), but also tends *to* **faithfully represent the way how the respective function/skill is accomplished by a human mind** and its material substrate – the brain.

In other terms, we believe that a cognitively plausible model should not only aim to attain the most quantitatively accurate results, but also to do so by processing the information similarly to the way mind does it.

The aim of this article is to elucidate the notion of "cognitive plausibility" (CP) by relating it to one particular problem, that of construction of grammatical categories present in natural languages. More concretely, we shall try to illustrate our point on the problem of construction of part-of-speech (POS) classes. We precise that the term POS-induction (POS-i) designates the process which endows the human or an artificial agent with the competence to attribute the POS-labels (like "verb", "noun", "adjective") to any token observable in agent's linguistic environment. For the simplicity of the argument, only parts of textual corpora like Multext-East (Erjavec, 2012) shall be considered as such "linguistic environment" of the computational agent introduced below.

1.2 Part-of-Speech induction in Natural Language Processing and Language Acquisition studies

POS-i is often considered to be "one of the most popular tasks in research on unsupervised NLP" (Christodoulopoulos et al., 2010). The problem of construction of grammatical categories is closely related to problem of "grammar induction" and language acquisition. Since "syntactic category information is part of the basic knowledge about language that children must learn before they can acquire more complicated structures" (Schütze, 1993), it is hard to imagine any computational model of grammar induction - aiming to discover the set of rules of the grammar of the language under study- without it being able to construct, in the first place, the equivalence classes upon which the rules-to-discover shall be applied (Elman, 1989; Solan et al., 2005).

Acquisition of formal grammatical categories, be it parts-of-speech or others, is thoroughly studied in psycholinguistic literature – for introductory overview c.f. Levy et al.,(1988). Such studies often aim to address the question "**whether grammatical categories are innate, or induced through interaction with environment by means of imitation and analogy?".** The result of this never-ceasing Nature&Nurture debate is vast amount of both empiric and theoretic knowledge which could be ideally useful for any tentative to bring together disparate disciplines of artificial intelligence and developmental psychology.

1.3 POS-i paradigm(s)

While already latent in worthy POS-i models, like that of (Elman, 1989) existed before, or were published more or less in parallel (Schütze, 1993), the paradigm currently dominating the POS-i domain was fully born with article published by Brown et al. in 1992. Without going into detail, we precise that the model was successful because of its ability to apply both Markovian

probabilistic concepts and those coming from information theory (Shannon & Weaver, 1949) upon the information contained in the co-occurrences of the words in the sequences, thus becoming the flagship of what we label hereby as "co-occurrence distribution" or "contextual distribution" (CD) paradigm. In decades to follow, the CD paradigm have clearly dominated the POS-i field. Be it hidden Markov Models tweaked with variational Bayes (Johnson, 2007) , Gibbs sampling (Goldwater & Griffiths, 2007), morphological features (Berg-Kirkpatrick, Bouchard-Côté, DeNero, & Klein, 2010; Clark, 2003) or graph-oriented methods (Biemann, 2006) – all such approaches and many others consider contextual co-occurrence to be the primary source of POS-irelevant information.

But as comparative study of (Christodoulopoulos et al., 2010) indicates when demonstrating that models integrating morphological features tend to better than those who do not, it seems plausible that the uncontested primary role of CD in POS should be revised. While it is evident that the CD indeed must furnish relevant information if ever
distributional hypothesis is valid (Harris, 1954) and it is axiomatic that distributional hypothesis applies in case of any agent creating categories consistently with Hebb's law (Hebb, 1964) we shall argue in subsection 3.1 that pertinent POS-I clues can be extracted not only from word's "external" contextual properties but also from word's very "internal" Μορφε.

2. Axiomatic conditions of Cognitive Plausibility

This section deals with what we believe are necessary (i.e. *sine qua non*) conditions of cognitive plausibility of a computational model . Subsection 2.1 deals with the "bootstrapping" condition stating that categories which are being built are based on categories which have already been built. Emergence of bootstrapping effect shall be illustrated on a trivial multi-iterative re-clustering of clusters pre-clustered according to CD features. Subsection 2.2 discusses the assumption that in order to be cognitively plausible, the model should be data and/or oracle-driven.

2.1 Bootstrapping the bootstrapping

From biochemistry to social sciences it is a well known fact that *structuring structures are the structures structured*. Computational Linguistics and NLP in particular is not an exception. The most general definition of the term bootstrapping (B) – i.e. that B is a self-sustaining multi-iterative process whereby outputs of the previous iteration modify the very execution of the next iteration – could be indeed apply upon so many computational "recurrent", "self-feeding" (Riloff & Jones, 1999), "auto-organizing" (Nowak et al., 1999) approaches that have been already applied in so many NLP studies, that to state about a NLP algorithm X that "X bootstraps" may sometimes seem to be plain tautology.

In certain sense almost any POS-i model based on CD paradigm are, *ex vi termini,* bootstrapping ones because even in the most simplistic models, the information about the membership of the target word W_T in the candidate class C is inferred from the probabilities of membership of W_L (W_T's left context) and W_R (W_T's right context) to their respective candidate POS classes. Given the fact that the W_T plays the role of right context for W_L and the role of left context for W_R, whole problem is circular and as such often calls for a bootstrapping solution.

Solan et al. (2005) refer to a crucial 4th component of their automatic distillation of structure (ADIOS) algorithm as "generalized bootstrapping". Differently from the "geometric approach" which shall be presented in our experiment below, ADIOS implements graph-like structures in

order to attain its aim of construction of equivalence classes useful in subsequent grammar induction. But in its very essence, the approach of Solan et al., i.e. that one should substitute the vertices "subsumed" by a "subsuming" non-terminal class-denoting vertex is analogical, *mutatis mutandi,* to the approach presented in the following paragraphs.

1.1.1 1st experiment: Bootstrapping k-way POS clustering seeded by token co-occurrence features

Experiment was performed with data contained in English (en), Czech (cs) and Slovak (sk), corpora contained in 4th version of Multext-East corpus (Erjavec, 2012).

Table 1 . Overall statistics of analyzed corpora

Corpus	Word Types	Tokens	$Tags_{POS}$	$Feat_{cooc}$
Cs	19283	100368	13	70426
En	10511	134832	12	36774
Sk	20588	103452	13	74912

Table 1. presents summary statistics concerning the quantities of distinct word tokens, word types (i.e. tokens without context) and the most coarse-grained "gold standard" POS-tags is presented along with total number of distinct co-occurrence features which is equivalent to the number of columns (dimensions) in the resulting co-occurrence matrix.

Every word W_T type was characterized by a (row) vector of values [W_{1L}, W_{2L} ...W_{NL}, W_{1R}, W_{2R} ... W_{NR}], W_{1L} referring to cases when the word W_1 occurred to the left of W_T, W_{2L} to cases when W_{2L} was to the left, W_{3R} to cases when W_3 was to the right from the target word. What results is a simple co-occurrence matrix with N rows and maximum of $Feat_{COOC}$==2*N columns. Given that in the experiment we were actually looking two words to the left and two words to the right from W_T, the maximum possible number of columns was $Feat_{COOC}$ =4*N. But since not all word couples do occur asides each other, the final number $Feat_{COOC}$ was always below the theoretical limit.

The matrix has been clustered in C={2 … 50} clusters by the fast & frugal repeated bisection k-way clustering algorithm as implemented in the clustering tool CLUTO (Karypis, 2002). Columns were scaled according to IDF principle and the clustering was done according to cosine metrics. Once finished, comparison with "gold standard" yielded V-measure (Rosenberg & Hirschberg, 2007) values which are also illustrated as NO curves on Figure 1.

We have implemented the bootstrapping component in a following manner: **After each clustering, the information about the proposed cluster is added as a new feature to target's word vector description**. Thus, if matrix with 20 columns entered the first iteration which clustered the vectors into 5 clusters, the matrix entering the second iteration shall have 20+5 columns. If second iteration yields 6 clusters, a matrix with 25+6 columns will become the input for the third iteration etc. Figure 1 shows that in case of all 3 studied corpora, the bootstrapping BO method always attains higher scores than the static NO approach.[1]

1 Note that the V-measure of NO-bootstrap curves seem to be relatively stable in regards to increase of number of clusters. Contrary to many-to-one accuracy (purity) which increases with number of clusters, V-measure thus seems to be better evaluation measure for cases when solutions containing different numbers of clusters have to be compared.

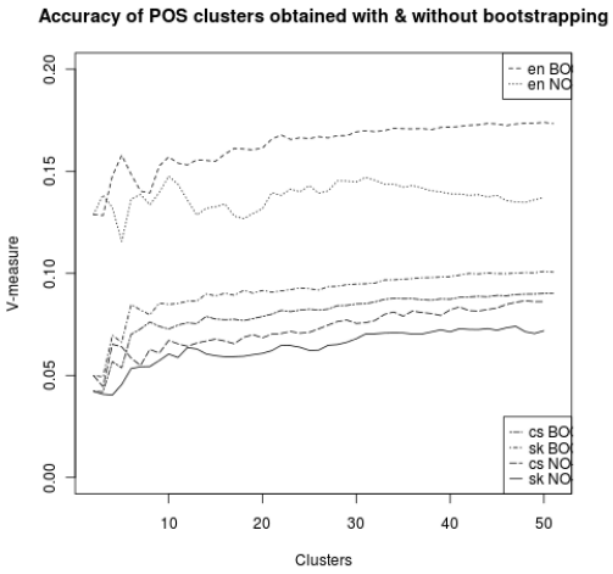

Fig. 1. Bootstrapping of contextual co-occurrence statistics

2.2 Data and oracle-driven learning

Computational models unable to analyze what they have previously synthesized and synthesize what they have previously analyzed could be hardly labeled as "cognitively plausible". But even the presence of such "dialectic" component cannot be the guarantee of absolute success, if ever the model's initial *prima materia* – the data with which the whole bootstrapping is initiated – are not adapted to model's prewired "innate" state.

It is unfortunately often the case in computational linguistics that whenever the model does not attain the expected performance, huge amount of effort is invested into tuning the model by diverse *ad hoc* modifications. After hours of exhaustive search, both intellectual as well as automatic, diverse parameters, meta-parameters and hyper-parameters are finally discovered which allow the model to attain somewhat superior performances when confronted, for example, with Wall Street Journal (WSJ) corpus But human categorization faculties – POS-i included – do not develop in such a way. While it seems plausible that same sort of "tuning of parameters" indeed takes place during initial period of language acquisition, it seems to be so efficient because the data itself is well adapted to ever-evolving state of baby's neuro-linguistic structures. Said more concretely, parents do not recite to its children the WSJ or Eulex corpora in order to adjust the synaptic weights in the brains of their children, they rather modify all their narrative intentions by pragmatic, prosodic, phonological as well as semantic Babytalk (Ferguson, 1964) cognitive filters. In doing so – by pre-processing the stimuli before it even attains perceptual buffers of child agent's ears – parents affirm themselves in the role of computational oracle (Turing, 1939).

Since it was already demonstrated by Clark (Clark, 2010) with sufficient analytical clarity that the "supervision" coming from external oracle machines can significantly reduce the complexity of the grammar induction and POS-i problems, we found it worthwhile to state that "fully unsupervised approaches are very rare because the **engineer's decision to confront the algorithm with corpus X and not Y, and to do so in the moment T_1 and not T_2, is already an act of supervision**".

By saying so we do not want to underestimate the importance of using the same corpora for mutual comparison of scientific results. We simple want to indicate that, because it determines everything which follows, the question of corpus choice should not be neglected. More concretely, cognitively plausible models of POS-i should be firstly tuned and "raised" with corpora like CHILDes (MacWhinney, 2000) and only later should be their scope of validity extended by means of confrontation with corpora of adult and expert utterances.

3. Conjectural conditions of model's Cognitive Plausibility

Subsection 3.1 discuss the role of non-distributional "surface" features for POS-induction. Discussion is followed by results of an experiment suggesting that features like suffix can indeed offer quite strong clues for the creation of syntactic categories. Subsection 3.2 introduces a conjectural condition for model's CP by proposing to base it principally on geometric grounds. It is followed by subsection 3.3 arguing that CP model should facilitate evaluation by means of qualitative inspection. In general, these sections deal with CP's conjectural conditions, meaning that while they may seem less self-evident that the axiomatic ones, we nonetheless consider them as valid.

3.1 Integration of surface features

Natural languages are very redundant communication channels (de Saussure., 1922; Shannon & Weaver, 1949). Three facets of the word – its morpho-phonological signifiant, its invisible signifiée and its its syntactic function – are not independent from one another and more often than not do they significantly overlap (Jackendoff, 2003; Lakoff, 1990). Thus it is not surprising that especially in morphologically rich languages, token's very syntactic function is encoded by morphemes present in the surface, i.e. objectively perceivable form, of the token itself. And results obtained by Clark (Clark, 2003) or (Berg-Kirkpatrick et al., 2010) indeed point in this direction – it may be no coincidence that approaches which exploit morphological features turned out, in (Christodoulopoulos et al., 2010) comparative study, to perform better than models which do not use such features.

1.1.2 2nd experiment : Assessing the impact of sufixal features on part-of-speech categorisation

We used the same three Multext-East corpora as in the first experiment. Ultimate character trigram was extracted from every word type and considered to be a feature. Word types are subsequently clustered in C clusters according these $\mathbf{Feat_{SUFFIX}}$ orthogonal dimensions. The comparison with Mutext-East gold standard *subsequently* yields V-measures (V), entropies (H) and purities (P) presented in Table 2.

Table 2. Performance of model's inducing C categories solely according to suffixal features

	C=10	C=30	C=50
Cs	V=0.178	V=0.24	**V=0.26**
534	H=0.487	H=0.392	H=0.34
	P=0.582	P=0.642	P=0.69
En	**V=0.248**	V=0.215	V=0.2
286	H=0.428	H=0.4	H=0.39
	P=0.639	P=0.652	P=0.66
Sk	V=0.17	V=0.272	**V=0.274**
523	H=0.5	H=0.373	H=0.339
	P=0.504	P=0.685	P=0.714

Amount below the corpus name in the above table denotes the length of the $\mathbf{Feat_{SUFFIX}}$ vector, i.e. the number of distinct suffixal trigrams observed in their respective corpora.

$Feat_{SUFFIX}$-driven model attains lesser V-measures as had obtained (Christodoulopoulos et al., 2010) when evaluating models of (Clark, 2003) or (Berg-Kirkpatrick et al., 2010) within their 2013 comparative study. The very same study however also indicates that even the simplistic $FEAT_{SUFFIX}$-driven model can be worth of certain interest since it seems to be quite fast – in comparison to models harnessing the power of more than dozen computational cores to attain comparable or even better V-measures than $FEAT_{SUFFIX}$-driven method , we are glad to state that in order to attain results presented above, our dual-core Pentium needed in average T_{EN}=1.8, T_{SK}=3.2, T_{CS}=3.6 seconds per simulation.

3.2 Knowledge is geometric

After the Turing machine symbol-operating paradigm started to put more importance upon ever-still more & more fine-grained modular to probabilistic and connectionist models. But in recent years, a "geometric" paradigm starts to gain momentum in diverse fields of cognitive sciences including computational linguistics and NLP. In experiments described above such paradigm was harnessed in a sense that *instead of modulating weights along different dimensions, geometers often modulate the number of dimensions itself*. It could be possibly reproached to such a geometric approach that associating every plausible feature with a new dimension can induce some serious matrix-sparsity problems and|or that such an approach would be, sooner or later, confronted with insurmountable computational and memory limits. It is true that methods by means of which some older approaches deal with the problem of huge co-occurrency matrices can be very costly, as is the case, for example, in singular value decomposition within LSA (Landauer & Dumais, 1997). But since very elegant, simple and concise representations of sparse matrices can be very easily generated (Karypis, 2002) and since lemma of Johnson-Lindenstrauss (W. B. Johnson & Lindenstrauss, 1984) indicates that sparse high-dimensional matrices can be easily projected into low-dimensional as is often done in random-indexing (Sahlgren, 2005), it seems to be plausible to state that construction of vector spaces which are 1) dense but 2) transformable for low computational cost 3) encode huge amount of features attributed to huge amount of objects is not so problematic as it used to be in time when HMM-mastered POS-i paradigm was born.

Series of articles by Sahlgren (2002; 2005), Cohen (2010), Widdows (2004) and their colleagues offer valuable initiation into advantages of random-projection based semantic models. For more general discussion of "geometrization of thought" in diverse fields of cognitive sciences, see (Gärdenfors, 2004). Within all such geometric models, categories can be considered as local subspaces of a global space derived from the data.

3.3 Mix of quantitative and qualitative evaluation

Performance of early grammatical category induction models was evaluated manually by introspection into induced equivalence classes and articles published in the period of "golden age" of POS-i often used to enumerate members of at least one particularly pleasing class or presenting their dendograms. Such an approach was later critiqued by Clark (2003) as "inadequate" and attention of POS-I community turned towards more quantitative measures like perplexity, conditional entropy, cross-validation (Gao & Johnson, 2008), one-to-one (Haghighi & Klein, 2006) or many-to-1 accuracy (purity); variation of information (Meila, 2003) , substituable F-score (Frank et al., 2009) etc.

For the purposes of this article we had decided to present our simulations principally in terns of V-measure. Given its elegance, stability in regards to growing number of clusters but also certain "strictness" (note that even the best performing models present in comparative study (Christodoulopoulos et al., 2010) rarely surpass the V>0.6 limit), we consider the V-measure to be very valuable quantitative measure of performance of clustering POS-i algorithms.

But we also believe that the "old school" many-to-1 purity measure can be of certain interest, especially for those aiming to create a "semi-supervised bridge" between POS-induction and POS-tagging models; or by those aiming not to evaluate the performance of the model by rather to gain insights of correct annotations of analyzed corpora. In other terms, asides to "global" statistic measures informing the researcher about the overall performance of the model, more "local" measures can still offer interesting and useful information about individual induced classes themselves. Values presented in Table 3 represent the number C of clusters into which the corpus has to be partitioned in order to obtain at least **Φ** absolutely pure (i.e. Purity=1) classes.

Table 3. Distillation of absolutely pure categories

	SFFX	CD	CD+BO	SFFX+CD+BO
Φ=1	72	168	107	69
Φ=2	92	194	142	71
Φ=3	105	196	180	80
Φ=4	126	248	189	90
Φ=5	131	281	194	96
Φ=10	160	377	256	116

For example, in order to obtain an absolutely pure cluster on the basis of contextual distribution (CD) features, one would have to partition the English part of Multext-East corpus into 168 clusters among which shall emerge following noun-only cluster:

authority, character, frontispiece, judgements, levels, listlessness, popularity, sharpness, stead, successors, translucency, virtuosity

Interesting insights can also be attained by inspection of some exact points of the clustering procedure. Let's inspect, as an example, the case when one clusters the English corpus into 7 clusters **according to features both internal to the word – i.e. suffixes – and external – i.e. co-occurrence with other words co-occurrence.** Such an inspection indicates that the model somehow succeeds to distinguish verbs from nouns. As is shown on Table 4, whose columns represent the "gold standard" tags and rows denote the artificially induced clusters, our naïve

computational model tends to put nouns in clusters 4 and 6 while putting verbs into clusters 2, 3 and 5.

Table 3 . Origins of Noun-Verb distinction

	N	V	M	D	R	A	S	C	I	P	X	G
0	10	3	0	0	413	30	0	0	0	0	1	0
1	**568**	67	0	0	1	0	1	2	0	1	0	0
2	97	**668**	0	0	1	137	3	2	0	0	0	0
3	13	**1011**	1	0	275	0	2	0	0	0	0	0
4	**1173**	67	4	0	6	133	0	0	0	4	3	0
5	608	**958**	72	67	252	321	99	72	7	106	3	12
6	**1977**	97	22	0	42	1091	3	0	3	0	2	0

The objective of our ongoing work is to align as much as possible such "seeding" states like that presented on Table 4. with data consistent with psycholinguistic knowledge about diverse stages of language acquisition process.

At last but not least, we believe that the **temporal aspects of model's performance**, i.e. the answer to the question "How long does the model need to run in order to furnish reasonable results?" **should be always seriously considered**. One way how to evaluate such temporal aspects of categorization could be a simplistic Turing-Test (TT) like POS-i oriented scenario where the evaluator asks the model (or an agent) to attribute the POS-label to word posed by evaluator, *or at least to return a set of members of the same category*. In such a real-life scenario, an absolute perfection of possible future answer could be possibly traded off for less perfect (yet still locally optimal) answer given in reasonable time.

But because with this TT_{POS} proposal we already depart from the domain of unsupervised induction towards semi-supervised "learning with oracle" or fully supervised POS-tagger, we conclude that we consider the condition "cognitively plausible model of part of speech induction should be evaluated by both quantitative and qualitative means" to be the weakest among all proposals concerning the development of an agent inducing the categories of natural language in a "cognitively plausible" way.

4. Conclusion

Model should be labeled as "cognitively plausible" model of certain human faculty if and only if it not only accurately emulates the input (problem) → output (solution) mapping executed by the faculty, but also emulates the basic "essential" characteristics associated to such mapping operation in case of human cognitive systems, i.e. emulates not only WHAT but also HOW the problem → solution mapping is done.

In relation to the problem of how part-of-speech induction is effectuated by human agents, two characteristic conditions have been defined as axiomatic (necessary). First postulates that POS-i should involve a "bootstrapping" multi-iterative process able to subsume terminals sharing common features under a new non-terminal and to subsequently exploit the information related to occurrence of the new non-terminal to extend the (vectorial) definition terminals represented in the memory. Ideally the process should converge to partitions "optimally" corresponding to the gold standard. First experiment has shown for three distinct corpora that even a very simple model based on clustering of the most trivial co-occurrence information can attain higher accuracies if such a bootstrapping component is involved. The second necessary condition of

POS-i's CP is that it should be data or oracle-driven. It should perform better when first confronted with simple corpora like CHILDes (MacWhinney, 2000) and only latter with more complex ones than if it would be first confronted with complex corpora.

Another condition of POS-i's CP proposed that morphological and surface features should not be neglected and instead of playing a secondary "performance increasing role", they should possibly "seed" whole bootstrapping process which shall follow. This condition is considered to be conjectural (i.e. "weaker") just because it points to somewhat orthogonal direction than does a traditionally acclaimed distributional hypothesis (Harris, 1954). It may be the case, however, that especially native speakers of some morphologically rich languages shall consider the "syntax-is-also-IN-the-word" paradigm not only as conjectural but also axiomatic.

Another "weak" condition of cognitive plausibility postulates that many phenomena related to mental representations and thinking, POS-i included, can be *not only described but also explained and represented* in geometric and topologic terms. Ideally, the geometric paradigm (Gärdenfors, 2004) should not be contradictory but rather complenetary to symbolic and connectionist paradigms. The last and weakest condition of CP proposed that computational models of part-of-speech induction should be not only easily quantitatively analyzed but should be also transparent for researcher's or supervisor's qualitative analyses. They should facilitate and not complicate posing of all sorts of "Why?" questions and the results should be easily interpretable. A sort of categorization-faculty Turing Test was proposed which could be potentially embedded into the linguistic component of the hierarchy of Turing Tests which we propose elsewhere (Hromada, 2012).

It may be the case that the list of conditions of cognitive plausibility presented in this article is not sufficient one and should be extended with other terms like "modularity", "self-referentiality" or notions coming from complex systems and evolutionary computing. Regarding the problem of elucidation of how could a machine induce, from the environment-representing corpus, the categories in a way analogical to that of a child learning by imitating its parents, we consider even the list of 2 strong precepts and 3 weak precepts hereby presented as quite useful and possibly necessary.

Bibliography

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., & Klein, D. (2010). Painless unsupervised learning with features. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (p. 582–590).

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. *Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (p. 7–12).

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based ngram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two Decades of Unsupervised POS induction: How far have we come? *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (p. 575–584).

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics- Volume 1* (p. 59–66).

Clark, A. (2010). Towards general algorithms for grammatical inference. *Algorithmic Learning Theory* (p. 11–30).

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256.

Elman, J. L. (1989). *Representation and structure in connectionist models*. DTIC Document.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, *46*(1), 131–142.

Ferguson, C. A. (1964). Baby talk in six languages. *American anthropologist*, *66*(6_PART2), 103–114.

Frank, S., Goldwater, S., & Keller, F. (2009). Evaluating models of syntactic category acquisition without using a gold standard
Proc. 31st Annual Conf. of the Cognitive Science Society (p. 2576–2581).

Gao, J., & Johnson, M. (2008). A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (p. 344–352).

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *ANNUAL MEETINGASSOCIATION FOR COMPUTATIONAL LINGUISTICS* (Vol. 45, p. 744).

Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (p. 320–327).

Harris, Z. S. (1954). Distributional structure. *Word*.

Hebb, D. O. (1964). *The Organization of Behavior: A Neuropsychlogical Theory*. John Wiley & Sons.

Hromada, D. D. (2012). Taxonomy of Turing Test Scenarios. *Proceedings of AISB/IACAP 2012 Symposium.* Birmingham, United Kingdom.

Jackendoff, R. (2003). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (p. 296–305).

Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, *26*(189-206), 1.

Karypis, G. (2002). *CLUTO-a clustering toolkit*. DTIC Document.

Lakoff, G. (1990). *Women, fire, and dangerous things*. Univ. of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211–240.

Levy, Y., Schlesinger, I. M., Braine, M.D.S. (1988). *Categories and Processes in Language Acquisition*. Lawrence Erlbaum.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Transcription, format and programs* (Vol. 1). Lawrence Erlbaum.

Meilua, M. (2003). Comparing clusterings by the variation of information. *Learning theory and kernel machines* (p. 173–187). Springer.

Nowak, M. A., Plotkin, J. B., & Krakauer, D. C. (1999). The evolutionary language game. *Journal of Theoretical Biology*, *200*(2), 147– 162.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. *Proceedings of the National Conference on Artificial Intelligence* (p. 474–479).

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (Vol. 410, p. 420).

Sahlgren, M. (2005). An introduction to random indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE* (Vol. 5).

Sahlgren, M., & Karlgren, J. (2002). Vector-based semantic analysis using random indexing for cross-lingual query expansion. *Evaluation of Cross-Language Information Retrieval Systems* (p. 169–176).

De Saussure, F., Bally, C., Séchehaye, A., Riedlinger, A., Calvet, L. J., & De Mauro, T. (1922). *Cours de linguistique générale*. Payot, Paris.

Schütze, H. (1993). Part-of-speech induction from scratch. *Proceedings of the 31st annual meeting on Association for Computational Linguistics* (p. 251–258).

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information. *Urbana: University of Illinois Press*, *97*.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*(33), 11629.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, *2*(1), 161–228. *Language and Speech*, *40*(1), 47–62.

Vlachos, A., Korhonen, A., & Ghahramani, Z. (2009). Unsupervised and constrained Dirichlet process mixture models for verb clustering. *Proceedings of the workshop on geometrical models of natural language semantics* (p.74–82).

Widdows, D., & Kanerva, P. (2004). *Geometry and meaning*. CSLI publications Stanford.