Slovak University of Technology in Bratislava
Faculty of Electrical Engineering and Information Technology
Department of Control and Cybernetics

&&

University Paris 8
Ecole Doctoralle Cognition, Langage, Interaction
Cognition Humaine et Artificielle

Thesis for rigorous examination

Daniel D. Hromada

Topic of doctoral Thesis :	Evolutionary models of ontogeny of linguistic categories and rules
Advisors :	doc. Ing. Ivan Sekaj, PhD. (ivan.sekaj @ stuba.sk) prof. Charles Tijus (tijus @ univ-paris8.fr)
Form of study :	external and under double supervision
Study began :	september 2010 at University Paris 8 september 2011 at Slovak University of Technology
Study program :	Cybernetics (Slovak University of Technology) Psychology (University Paris 8)
Study discipline :	9.2.7 Cybernetics (Slovak University of Technology) Cognitive Psychology (University Paris 8)

Abstract

Language development is a process by means of which a human baby constructs an adequate competence to encode & decode meanings in language of her parents. Computationally it can be described as a trinity of mutually interconnected problems : clustering of all tokens which baby heard into 1) semantic and 2) grammatical categories ; and 3) discovery of grammatical rules allowing to combine the members of diverse equivalence classes into syntactically correct and meaningful phrases. A theoretical, « psycholinguistic » claim of our Thesis is that similary to those theories which explain emergence of cultural or creative **thinking as the result of evolutionary process occuring within an individual mind**, the emergence of linguistic representations and faculties within a human individuum can be also considered as a case where basic tenets of Universal Darwinism apply. The practical, « cybernetic » aim of the Thesis is to create a computational models of concept learning, part-of-speech induction and grammar induction having comparable performance to existing models but based principially on evolutionary algorithms. It shall be argued that the « fitness function » , which determines the « survival rate » of « candidate grammars » emerging and disappearing in baby's mind should be based upon the idea that the most fit is such a grammar G which « minimizes the distance » between the utterances successfully parsed from linguistic environment E by the application of grammar G and the utterances potentially generated by the grammar G.

Keywords : evolutionary computing, language acquisition, genetic epistemology, part-of-speech induction, grammar induction, optimal clustering, machine learning, concept construction, grammar systems, motherese, toddlerese

List of most important abbreviations

CC – Concept Construction
EA – Evolutionary Algorithms | EP – Evolutionary Programming
ES – Evolutionary Strategies
ET – Evolutionary Theory
GA – Genetic Algorithms
GE - Genetic Epistemology
GI – Grammar Induction | Grammar Inference
LA - Language Acquisition | LD – Language Development
NLP – Natural Language Processing
POS-i – Part-of-Speech Induction ; POS-t – Part-of-Speech Tagging
UD – Universal Darwinism

Convention

« *Italics* » – citation
(x | y | z) - disjuncted token – i.e. read token « (neural|quantum) darwinism » as « neural darwinism ; quantum darwinism »

Table of Contents

0. Introduction

A general form of Evolution Theory (ET) postulates that entities evolve and adapt to their environment by a process of accumulation of information. Such a generalized theory – often referred to as « Universal Darwinism » - can be and often is applied in diverse scientific disciplines as diverse as biology, linguistics or even anthropology and psychology. Since principal concepts and tenets of ET can be easily formalised into stochastic « evolutionary » algorithms, ET can yield not only a theoretical framework but also a computational experimental methodology for any scienctific discipline whose basic concepts and principles can be « reduced » into a ET-consistent form.

The aim of my doctoral Thesis is to empirically – i.e. by means of computational experiments - demonstrate that certain phaenomena observed by « developmental linguists » and « psycholinguists » can be explained in terms of principles of Universal Darwinism and as such can be modelled by « computational linguists » and « Natural Language Processing (NLP) engineers» who shall found their computational models upon methods offered by Evolutionary Computing paradigm. More concretely, I shall try to indicate that « evolutionary » optimization can be used to yield solutions to at least three problems of language development:

1) induction of semantic categories, i.e. construction of « concepts »
2) the problem of induction of part-of-speech grammatical categories of words natural languages, i.e. the problem of how equivalence classes like « nouns », « verbs », « adjectives » etc. are constructed by the language-acquiring agent
3) the problem of grammar induction, i.e. the problem of how an agent can acquire a grammar from the corpus or its environment

It shall be indicated that the term «language-acquring agent » could be interpreted both as an organic agent (e.g. a human baby) trying to learn the language of its environment (e.g. its parents) as well as a computational agent (e.g. a Turing Machine) inducing the structural properties of the language which generated the corpora with which the agent has been confronted. In other terms, it shall be indicated that ET is generalizable in such an extent, that its correct implementation may allow two systems based upon Darwinian principles « replicate, mutate, select » *to converge to same optimal or quasi-optimal categories* regardless the fact that the substrate by means of which they computate is organic or not.

The first chapter will more closely present the above-mentioned basic principles of the universal ET doctrine and enumerate certain scientific disciplines for which the ET furnishes a useful theoretical framework. Asides biology where the role of ET is evident, a discipline of «evolutionary psychology » shall be mentioned principially in

order to avert the reader that our aims are not limited to those posited by evolutionary psychology. The « memetic theory », on the contrary, shall more precisely elucidate our ultimate aim since it already introduces a novel level of representation, « a meme », supposed to be « the basic unit of immitation » and as such offers an interesting starting point for any Darwin-consistent theory of evolution of non-organic (e.g. cultural) structures and artefacts. It is, however, the constructionist « genetic epistemology » (GE) of Jean Piaget which shall resonate even more strongly with our aims – since what GE ultimately postulates is that the human psyche – with all its linguistic, moral, object-manipulating faculties – pass through the sequence of « stages » . For it is our belief that such Piagetian « stages » can be explained, in computational terms, as «quasi-optimal attractors» within a very complicated « search space » of agent's internal representations and that a sort of evolutionary process occurs not only on a social-memetic level between the agents imitating each other, but, in the first place, within the agents (him|her|it)self. This Thesis is our tentative to base this « learning=evolution » belief on solid ground of complexity theory.

The second chapter will address the topic of language development (LA). The topic is so vaste and deep that only most fundamental subproblems (i.e. vocabulary development, acquisition of part-of-speech categories and acquisition of gramamrs) shall be briefly described and some basic notions like « variation set » or « motherese» will be introduced. We shall try to evit the dispute between diverse linguistic doctrines and schools (e.g. nativists, cognitivists, comparativists) ; focus shall be put upon points of consensus supported by empiric evidence.

While the goal of first chapter is to furnish the theoretical framework and the goal of the second is to specify the problem, it is the third chapter which deals with the concrete computational tentatives to unify the two. Major part of the chapter shall deal with the question of evaluation of diverse inductive models. Some most successful computational models of part-of-speech induction (POS-i) and grammar induction (GI) shall be mentioned in order to pave the way for the evolutionary ones. As shall be indicated by this section, the tentatives to apply evolutionary algorithms (EAs) to solve the POS-i and GI problem are, regardless the good results reported in the litterature, very rare. In specific subsections of the chapter, we shall mention certain models, both psycholinguistic and computational, which justify our claim that the process of ontogeny of linguistic faculty can be not only interpreted but also modelled as a process of evolutionary optimization of cognitive structures.

1. Universal Darwinism

Universal Darwinism (UD) is a scientific paradigm regrouping diverse scientific theories extending the Darwinian theory of evolution and natural selection (Darwin 1859) beyond the domain of biology. It can be understood as a generalized theoretical framework aiming to explain the emergence of many complex phenomena in terms of interaction of three basic processes: 1) variation 2) selection 3) retention. According to UD paradigm, interaction of these three components yields a « *universal algorithm valid not only in biology, but in all domains of knowledge where we can extract informational entities – replicators, which are able to reproduce themselves with variations and which are subjects to selection* » (Kvasnička and Pospíchal 1999). This generic algorithm is nothing else than traditional Evolutionary Theory (ET) which, when when considered as substrate-neutral, can be applied to such a vaste number of scientific fields that it has been compared to a kind of « universal acid » which « *eats through just about every traditional concept, and leaves in its wake a revolutionized world-view, with most of the old landmarks still recognizable, but transformed in fundamental ways*» (Dennett 1996) .

As of 2013, the existing scientific disciplines which could be labeled as UD-consistent include: biology ; evolutionary (art | psychology | music | linguistics | ethics |economics | anthropology | epistemology | computation); sociobiology ; memetics ; (quantum | neural | psycho) darwinism ; artificial life and many others. In regards to the overall aim of our Thesis, some of the most relevant instances of UD are described in following sub-sections.

1.1. Biological evolution

Evolutionary Theory was born when young Charles Darwin realised that the « *gradation and diversity of structure » (Darwin 1906),* which he had encountered among mockingbirds of Galapagos islands, could be explained by natural tendency of species to « *adapt to changing world* ». Parallely to Darwin's work which was gradually clarifying the terms of variability and its close relation to environment-originated selective pressures, Gregor Mendel was assessing statistical distributions of colours of flowers of his garden peas in Brno in order to finally converge to fundamental principles of heredity . But it was only in 1953 when the double-helix structure of the material substrate of heredity of biological species – the DNA molecule – was described in (Watson & Crick, 1953) paper.

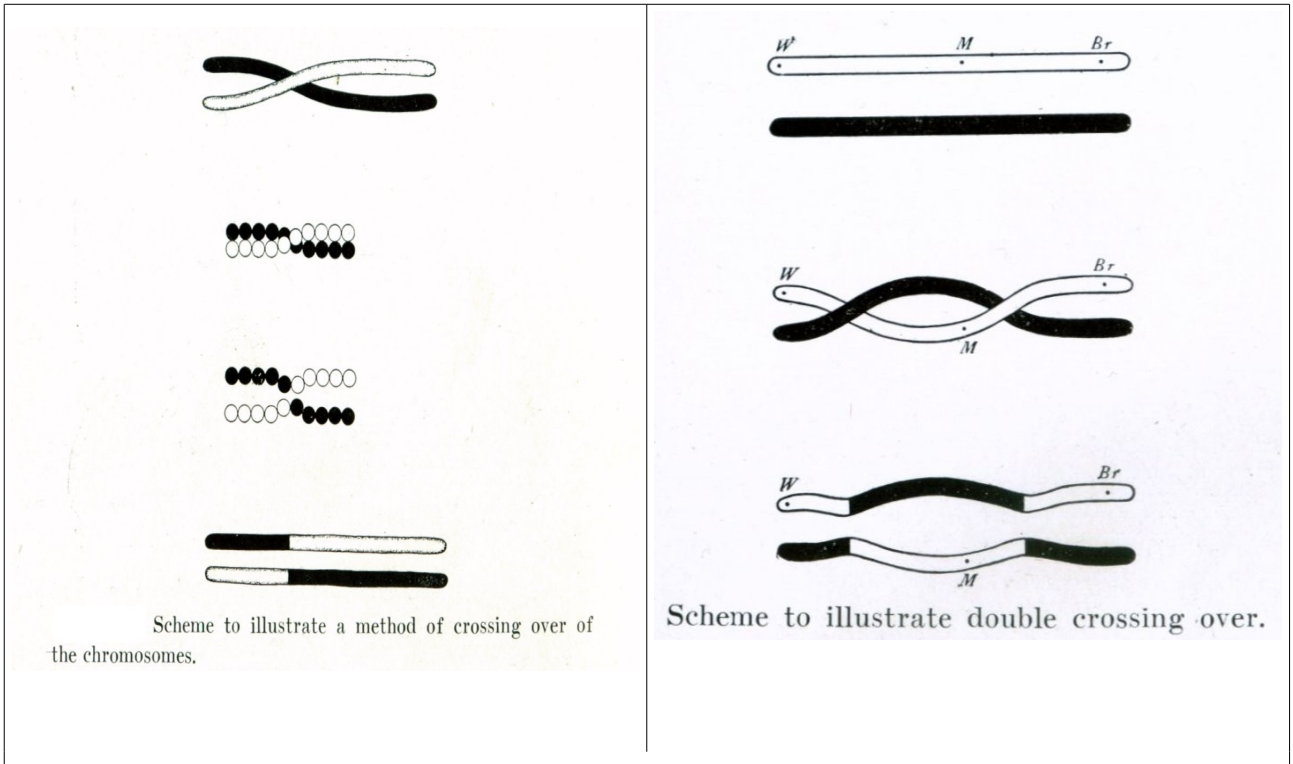
In simple terms : In the DNA molecule, information is encoded as a sequence of nucleotides. Every nucleotide can contain one of four nucleobases, it thus ideally carry 2 bits of information. Continuous sequence of three nucleotids gives a « triplet » which, when interpreted by a intracellular « ribosome » machinery, can be « translated » into an amino-acid. Sequences of amino-acides yield proteins which interact one with another in biochemical cascades. The result is a living organism with its particular phenotype aiming to reproduce its genetic code.

If, in the given time T there are two organisms A and B whose genetic code differs in such an extent that their phenotype differs, and if ever the phenotype of organism A

augments probability of A's survival and reproduction in the external world W, while the B's phenotype diminishes such probability , we say that the A is better adapted to world W than B, or more formally that fitness(A) > fitness(B). Evolutionary Theory postulates that in case that there is a lack of resources in world W, descendants of the organism B shall be gradually, after multiple generations, substituted by descendants of a more fit organism « A ». This is so because during every act of reproduction, the material reason for having a more fit phenotype - the DNA molecule – is transferred from parent to offspring and the whole process is cumulative across generations.

It can, however, happen, that the world W changes. Or a random (stochastic) event – a gamma ray, the presence of a free radical - can occur which would tamper A's genetic code. Such an event – called « mutation » - shall result, in majority of cases, in decrease of A's fitness. Rarely, however, can mutations also increase it.

Another event which can transform the genetic sequence is called « crossover ». It can be formalised as an operator which substitutes one part of genetic code of the organism A with corresponding sequence of organism B, and vice versa, the part of B with the corresponding part of A. It is indeed especially the crossover operation, first described by in the article of T.H. Morgan (Morgan 1916), which is responsible for « mixing of properties » in case of a child organism issued from two parent organisms. In more concrete terms : the genetic code of such « diploid » organisms is always stored in X pairs of chromosomes. Each chromosome in the pair is issued from either father or mother organism which, during the process of meiosis, divide their normally diploid cells into haploid gamete cellls (i.e. sperms in case of father and eggs in case of mother). It is especially during the first meiotic phase that



Scheme to illustrate a method of crossing over of the chromosomes.

Scheme to illustrate double crossing over.

Figure 1 : Two types of crossover operation. Figures reproced from (Morgan, 1916)

crossover occurs, the content of DNA sequence of two grand-parents being mixed and mapped during crossover operation into the chromosome contained in the gamete which, if lucky, shell fuse with the gamete of another parent in the act of fecondation.

Resulting « zygote » is again diploid, contains mix of fragments of genetic code originally present in the cells of all four grand-parents of the nascent organism. Zygote subsequently exponentially divides into growing number of cells which differentiate from each other according to instructions contained in the genetic code which are triggered by biochemical signals coming from cell's both internal and external environment. If the genetic code shall endow the organism with such properties that will allow it to survive in its environment until its own reproduction, approximately half of the genetic information contained in its DNA shall be transfered to the offspring organism. If not, the information as such shall disappear from the population due to its incompatibility with the environment.

1.2. Evolutionary Psychology

It was already Darwin who posited that ET shall have profound impact upon psychology :

« In the distant future I see open fields for far more important researches. Psychology will be based on a new foundation that of the necessary acquirement of each mental power and capacity by gradation. » (Darwin 1859)

While two possible intepretations of this Darwin's idea exist, Evolutionary Psychology (Ev.Psych.) focuses only on the first one. It aims to explain diverse faculties of human soul & mind in terms of selective pressures which moulded the modular architecture of human brain during millions of years of its phylogenetic history. Its central premises state : *« The brain's adaptive mechanisms were shaped by natural and sexual selection. Different neural mechanisms are specialized for solving problems in humanity's evolutionary past » (Cosmides and Tooby 1997).*

In more concrete terms, Evolutionary Psychology explains quite successfully phaenomena as diverse as emergence of cooperation and altruistic behaviour (Hamilton 1963); male promiscuity and parental investment (Trivers 1972) or even the obesity of current anglo-saxxon population (Barrett 2007). All this and much more is explained as a result of adaptation of homo sapiens sapiens (and all its biological ancestors) to dynamism of its ever-changing ecological and social niche.

Thus, in the long run, Ev.Psych. tends to explain and integrate all innate faculties of human mind in the evolutionary framework. The problem with Ev.Psych., however, is that in its grandious aim to « *assemble out of the disjointed, fragmentary, and mutually contradictory human disciplines a single, logically integrated research framework for the psychological, social, and behavioral sciences* » (Cosmides and Tooby 1997) it can sometimes happen that Ev.Psych. posits as innate, and thus explainable in terms of biological natural selection, cognitive faculties which are not innate but acquired. Thus it may be more often than rarely the case that whenever it comes to the famous nature vs. nurture (Galton 1875) controversy, evolutionary

psychologists tend to defend the nativist cause even there, where it means to commit a epistemological fallacy to do so[1].

And what makes things even worse for the discipline of Evolutionary Psychology as is currently performed is, that the forementioned Darwin's precognition has, asides the nativist & biological one, also another intepretation. Id est, when Darwin spoke about mental powers and capacities acquired by gradation, one cannot exclude that he was speaking not only about gradation in phylogeny, but also ontogeny.

1.3. Memetics

Theory of memes or memetics is, in certain sense, a counter-reaction to Evolutionary Psychology's aims to explain human mental and cognitive faculties in terms of innate propensities. Similiarly to Ev.Psych., memetics is also issued from the discipline of sociobiology which was supposed to be « *The extension of population biology and evolutionary theory to social organization* » (Wilson 1978). But differently to both Ev.Psych. and sociobiology, memetics does not aim to explain diverse (cultur|psychologic|soci)al phenomena solely in terms of evolution operating upon biochemical genes, but also in terms of evolution being realised on the plane of more abstract information-carrying replicators called « memes » (Dawkins 2006).

The basic definition of the classical memetic theory is: « *Meme is a replicator which replicates from brain to brain by means of imitation*» (Blackmore 2000). These replicators are somehow represented in the host brain as some kind of « cognitive structure » and if ever externalised by the host organism – no matter whether in form a word, song, behavioral schema or an artefact – they can get copied into other host organism endowed with the device to integrate such structures[2]. Similary to genes which often network themselves into mutually supporting auto-catalytic networks (Kauffman 1996), memes can also form more complex memetic complexes, « memplexes », in order to augment the probability of their survival in time. Memes can thus do informational crossovers with one another (syncretic religions, new recepts from old ingredients or DJ mixes can be nice examples of such memetic crossover) or they can simply mutate, either because of the noise present during the imitation (replication) process, or due to other entropy-related decay-like factors related to the ways how active memes are ultimately stored in brains or other information processing devices.

Memetic theory postulates that the cumulative evolutionary process applied upon such information-carrying stuctures shall ultimately lead to emergence of such complex phaenomena as culture, religion or language.

1.4. Evolutionary Epistemology

Epistemology is a philosophical discipline concerned with the source, nature, scope ,

1 If ever we accept the notion of falsifiability as an important criterion of accpetation or rejection of the scientific hypothesis (Popper et al. 1972), many hypotheses issued from EP would have to be rejected because, since being based in the distant past which is almost impossible to access, they are less falsifiable than hypotheses explaining the same phaenomena in terms of empiric data observable in the present.

2 In neurobiological terms, the faculty to imitate and hence to integrate memes from external environment is often associated to so-called « mirror neurons » (Rizzolatti and Craighero 2004).

existence and divesity of forms of knowledge. Evolutionary epistemology (EE) is a paradigm which aims to explain these by applying the evolutionary framework. But under one EE label, at least two distinct topics are, in fact, addressed :

1) EE_1 which aims to explain the biological evolution of cognitive and mental faculties in humans and animals
2) EE_2 postulates that knowledge itself evolves by selection and variation

EE_1 can be thus considered as sub-discipline of Ev.Psych. and as such, is subject to Ev.Psych.-directed criticism presented on previous page. EE_2, however, is closer to memetics since it postulates the existence of a second replicator, i.e. of an information-carrying structure which is not materially encoded by a DNA molecule.

The distinction between EE_1 and EE_2 can also be characterised in terms of « phylogeny » and « ontogeny ». Given the definition of phylogeny as the « processus which shapes the form of species » and contrasting it to ontogeny defined as « processus shaping the form of individual », we find it important to reiterate that while EE_1 is more concerned with knowledge as a result of phylogenetic moulding of DNA, EE_2 points more in the direction of « ontogeny». In fact, EE_2 paves the way for at least two other sub-interpretations :

$EE_{2\text{-}1}$ Knowledge can emerge by variation&selection of ideas shared by a group of mutually interacting individuals (Popper 1972)

$EE_{2\text{-}2}$ Knowledge can emerge by variation&selection of cognitive structures within one individuum

It is worth noting that while a so-called recapitulation theory stating that « *ontogeny recapitulates phylogeny* » (Haeckel 1879) is considered to be discredited by many biologists and embryologists ; it is still held as valid by many reseachers in human and cognitive sciences observing a « *strong parallelism between cognitive development of a child and … stages suggested in the archeological record* » (Foster 2002)100 years after one of Darwin's companion has noted : « *Education is a repetition of civilization in little* » (Spencer 1894).

1.5. Individual Creativity

In fact, the evolutionary epistemology was born with the tentative of D.T. Campbell to explain both creative thinking and scientific discovery in terms of « *blind variation and selective retention* » of thoughts (Campbell 1960). Departing from introspective works of mathematician Henri Poincare who stated « *To create consists precisely in not making useless combinations and in making those which are useful and which are only a small minority. Invention is discernment, choice...Among chosen combinations the most fertile will often be those formed of elements drawn from domains which are far apart...What is the cause that,among the thousand products of our unconscious activity, some are called to pass the threshold, while others remain below?»* (Poincaré 1908), Campbell suggests that what we call creative thought can be described as a Darwinian process whereby the **previously acquired knowledge blindly varies in unconscious mind** of the creative thinker and that only some such structures are

subsequently selectively retained. As (Simonton 1999) puts it: « *How do human beings create variations? One perfectly good Darwinian explanation would be that the variations themselves arise from a cognitive variation-selection process that occurs within the individual brain.* »

1.6. Genetic Epistemology

« *The fundamental hypothesis of genetic epistemology is that* ***there is a parallelism between the progress made in ... organization of knowledge and the corresponding formative psychological processes****. Well, now, if that is our hypothesis, what will be our field of study? Of course the most fruitful, most obvious field of study would be reconstituting human history: the history of human thinking in prehistoric man. Unfortunately, we are not very well informed about the psychology of Neanderthal man or about the psychology of Homo siniensis of Teilhard de Chardin. Since this field of biogenesis is not available to us, we shall do as biologists do and turn to ontogenesis. Nothing could be more accessible to study than the ontogenesis of these notions.* ***There are children all around us****.* » (Piaget 1974)

Strictly speaking, Piaget's developmental theory of knowledge, which he himself called Genetic Epistemology (GE) may seem to be utterly non-Darwinian. In fact it is not even concerned with biochemical genes : Piagetian uses the term « genetic » to refer to a more general notion of « heredity » defined as structure's tendency to guard its identity through time.

The basic structural primitives of Piagetian theory are behavioral « schemas » which can be defined as « *a basic set of experiences and knowledge that has been gained through personal experiences that define how things should be and act in the person's environment. As the child interacts with their world and acquires more experiences these schemes are modified to make sense, or used to make sense of the new experience* » (Bee and Boyd 2003).

There are two ways how such schemes can be modified. Either they « assimilate » data from external environment. Or, if ever such assimilation is not possible because it is simply not possible that child's cognitive system matches the perceived external datum with the internal pre-existing category, the process of « accomodation » takes place which transforms the internal category to match the external datum.

Ultimately, the set of schemes gets so out-dated or so altered by past modifications that they are not useful anymore. Whenever such «equilibriation » occur, old set of schemas is rejected, the child tends to « *start fresh with a more up-to-date model* » (Bee and Boyd 2003), thus attaining new substage or stage of its development. In the Piagetian system – which is based on very precise yet exhaustive observations of dozens of children including his own – the order of stages is fixed and it is very difficult, or even fully impossible, for evolving psyche to attain *pre-operational* stage 2 or *concrete operational* stage 3 if it had not even mastered all that is to master during the *sensorimotor* stage 1 .

Given the fact that the GE paradigm involves :

- heredity – schemes are structures which tend to keep their identity in time
- variation – schemes are altered by the environment-driven assimilation or

accomodation [3]

- selective pressures – only those schemas which are most well adapted to environment and/or form most functionally fit complexes with other schemas shall pass through the equilibriation milestone

it can be briefly stated that Piaget's GE could be aligned with ET and UD. And what more, it may be the case that notion of Piagetian stages is consisted with the notion of attractor or locally optimal states whose emergence is, according to complex system theory (Kauffman 1996; Flake 1999), inevitable in a system as complex as child's psyche definitely is.

1.7. Evolutionary computation

We have already mentioned (c.f. 1.1.) that evolution, as defined within UD, can be thought of as a universal, generic algorithm. Not only can « evolutionary theory » serve us to explain diverse phenomena around us, it can be also exploited for finding solutions to diverse problems. Thus it is of no surprise that many researchers in informatics realized that not only can be the evolutionary process encoded as an informatic algorithm, but that such algorithms could be useful as a heuristic which could potentially lead to a discovery of useful (quasi)-optimal solutions to wide range of diverse problems. First explorations in the domain were done by Rechenberg's « evolutionary strategies » (Rechenberg 1973) and Holland's « genetic algorithms » (Holland 1975) which, along with « evolutionary programming » (Fogel et al. 1966) and « genetic programming » form the « evolutionary computation » subdiscipline of computer science.

All four approaches differ from classical optimization methods in following aspects :

1. *using a population of potential solutions in their search*
2. *using explicit « fitness » instead of function derivatives*
3. *using probabilistic, rather than deterministic, transition rules »*

(Kennedy et al. 2001)

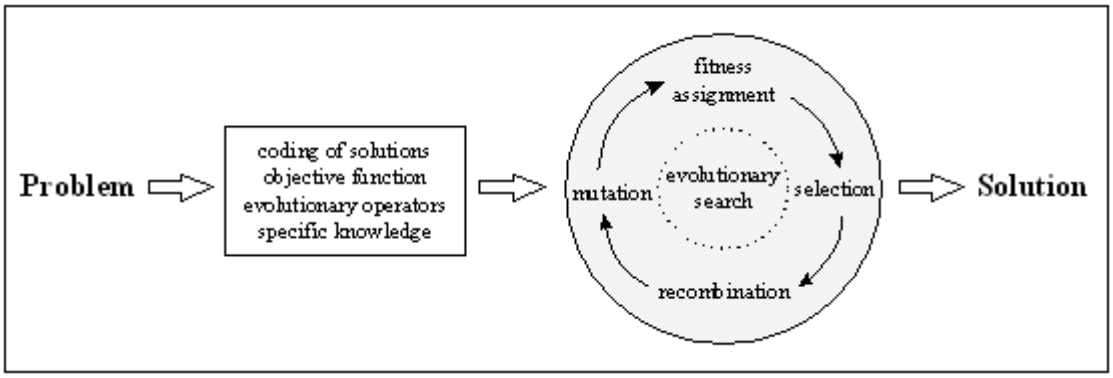


Figure 2: Basic genetic algorithm schema. Reproduced from (Pohlheim 1996)

3 Note that in terms of theory of evolutionary computation, one can relate the Piagetian notion of assimilation to an operator of local variation which attracts the cognitive system to locally optimal agreement with its environment, while accomodation suggests an interpretation in term of more global variation operators (like cross-over), which could potentially allow the cognitive system to reach a state of global equilibrium in regards to environment.

1.7.1. Genetic algorithms & fitness landscapes

Basic principle of « genetic algorithms » is illustrated on Figure 2. The core component of every genetic algorithm is the objective « fitness function » able to attribute a cardinal value or ordinal rank to any individum in the population of potential solutions. In other terms, the fitness function yields the criterium according to which one candidate individum is evaluated as « more fit » a solution, in regards to the problem under study, than other potential solutions present in the population. Population is the set of individual solutions. Every individual solution is encoded as a vector of values (also called « chromosome » or « genome ») which can vary in time. Designer choice related to the way how the problem solutions are encoded in chromosomal vectors, e.g. the type (Boolean ? Integer ? Float ? Set?) of different elements of the vector is also a crucial one and can often determine whether the algorithm shall succeed or fail.

In every generation – i.e. in every iteration of the algorithmic cycle represented by the circle on Figure 2. - all N individuals in the population are evaluated by the fitness function. Every individual thus obtains the « fitness » value, which subsequently governs the « selection » procedure choosing a subset of individuals from the current generation as those, whose genetic information shall reproduce into next generations.

In our Thesis we plan to exploit especially the « fitness proportionate selection » as the selection operator. This operator, also called « roulette wheel operator » transforms the fitness f_i of individual i into the probability p_i of its survival by means of a formula :

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$$

where N is the number of individuals in the population.

Once the « most fit » candidates are selected by the selection operator, they are subsequently mutually recombined by means of « crossover » operators and/or modified by means of « mutation » operators. Many different types of selection, mutation and crossover operators exist, for their overview c.f. (Sekaj 2005). For the purpose of this work let's just note that the probabilities of occurrence of mutation or crossover have to be fairly low, otherwise no fitness-increasing information could be transferred among generations and whole system will tend to present non-converging chaotic behaviour (Nowak et al. 1999).

Another useful strategy, which guarantees that maximal fitness shall either increase or at least stay constant, is called elitism. In order to implement the strategy, one simply guards one (or more) individual(s) with highest fitness unchanged for next generation, thus protecting « the best ones » from variations which would, most probably, decrease rather than increase the fitness[4].

Yet another widely used approach reinforces the selection pressure by removal of the

4 Note that in nature, elitism is often but not always the case. For it can happen that, due to stochastic factors, the most fit individuals die before they succeed to reproduce themselves.

weakest individuals. Both elitist « survival of the fittest » and the contrary « removal of the weakest » are often combined.

The selection of the most fit individuals from the old generation, their subsequent replication and/or recombination and diversification yields a new generation. Because individuals with lower fitness have been either completely or at least partially discarded by the selection process, one can expect that the overall fitness of new generation shall be higher than the fitness of the old generation. With little bit of luck, one can also hope that the most fit individuals of the new generation shall be little bit more fitter than the most fit individuals discovered in the new generation – this can happen if ever a « benign » mutation have occured, i.e. a modification which had moved the individual from the lower point on the « fitness landscape » to somewhat higher state.

The notion of fitness landscape, first introduced in (Wright 1932) is a metaphor useful for understanding&explanation of diverse evolutionary phenomena. The landscape is depicted as a mountain range with peaks of varying height. The height at any point on the landscape corresponds to its fitness value; i.e. the higher the point, the greater the fitness of an individual represented by the given point of the landscape. In such a representation, the evolution of the organism to more and more « fit » forms can be depicted as a movement up-hill, towards the most closest peak (i.e. local optimum) or towards the highest peak of the whole landscape (i.e. global optimum). Figure 3 illustrates a fitness landscape of a very simple organism with only one gene (whose potential values are encoded by illustration's X axis).

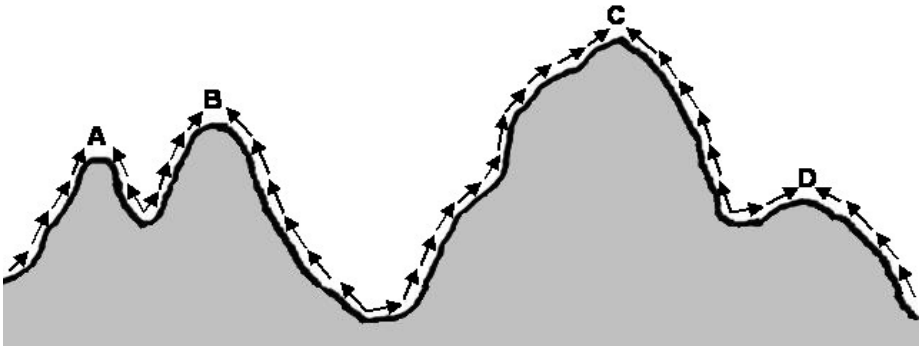


Figure 3: Possible fitness landscape for a problem with only one variable. Horizontal axis represents gene's value, vertical axis represents fitness.

Every arrow on the figure represents one possible individual. Its length represents the variation which can be brought in by the mutation operator. The fact that individuals always tend to move « upwards » indicates that selection pressures are involved. It has to be added that without the implementation of the crossover operator, the globally optimal state (encoded by point C) could not be attained for individuals who haven't originated at the slopes of C. Only some sort of crossover operator could ensure that individuals who attained the local optima (encoded by peaks A, B, D) could be mutually recombined (for example B with D) in a way that shall allow them to leave the locally stable states and approach the globally optimal C.

The fact that genetic algorithms, thanks to « crossover » operators, can combine two

individuals from diverse sectors of the fitness landscape, allow them to find solutions to problems where heuristics based on « gradient descent » should fail.

1.7.2. Evolutionary programming & evolutionary strategies

Evolutionary programming (EP) and evolutionary strategies (ES) are methods whose overall essence is very similar to GAs. There are, however, some subtle differences among the approaches.

In EP, mutation is the principal and often the only variation operator. While recombination is rarely used, « operators are freely adapted to fit the problem at hand » (Kennedy et al. 2001). EP algorithms often double the size of population by mixing children with parents and then halving the population by selection. Tournament selection operator is often used. Another difference is that while GAs were developped in order to optimize the numeric parameters of mathematical function under study – and variation thus directly modifies the genotype – in EP, one mutates the genotype but evaluates the fitness according to phenotype. EP is thus often used for construction & optimization of such structures like « finite state automata » (Fogel et al. 1966). A self-adaptation approach (Bentley 1999) allowing for mutation of the parameters of the evolution itself – e.g. the mutation rate – is also frequently used.

Such an approach of « evolving the evolution » is also used in ES which where discovered - in parallel but independetly with Holland's GAs – by (Rechenberg 1973). The biggest difference between EP and ES is thus fact that ES often recombines its individuals before mutating them. Popular and well-performing strategy thus seems to be :

1. Initialize the population
2. Perform recombination using P parents to form C children[5]
3. Perform mutation on all children
4. Evaluate children population and select P members from it.
5. If the termination criterion is not met, go to step 2 ; terminate otherwise.

Given the fact that in our Thesis, we shall often: 1) encode the problem of linguistic category induction by non-numeric chromosomes 2) evaluate the fitness of individuals by means of additional « phenotypic algorithms » we consider the works of Fogel & Rechenberg to be of particular importance for our study.

1.7.3. Genetic programming

Contrary to GAs, E.Prog and E.Strat which operate upon the chromosomes (vectors) of fixed length of numeric/boolean/character values, do individuals evolved by means of Genetic Programming (GP) encode programs of arbitrary length and complexity. In other terms, one may state that while above-mentioned EC methods look for the most optimal solution of a given problem, GP tends to produce a hierarchical tree

5 Frequently used C/P ratio is 7

structure encoding a sequence of instructions (i.e. a program) able to yield optimal solutions to a whole range of problems. Simply said : GP is simply a way how computer programs can automatically « discover » new and useful programs.

The most important thing to do in order to prepare a GP framework is to specify how shall be the resulting individuals (programs) encoded. Original choice of the founder of the discipline, John Koza, was to encode all individuals as trees of LISP S-expressions composed of sub-trees, which are, themselves, also LISP S-expressions. Within such arborescent S-expressions, the terminal (i.e. leave nodes where the branches end) nodes represent program's variables and constants while the non-terminal nodes (i.e. internal tree points) represent diverse functions contained in the function set (e.g. arithmetic functions like +, -, *, / ; mathematic functions like log, cos ; boolean functions like AND, OR, NOT ; conditional operators if/else etc.)

Figure 4 illustrates how, during the initial run of the algorithm, an individual – calculating, for example, the square root of X+5 – could be possibly randomly generated by implementing a following procedure :

1) « Root » of the program tree is randomly chosen from the function set, it is the function sqrt.
2) The function sqrt has only one argument (arity(sqrt)=1), therefore it will take only one input from the randomly determined functor + (addition)

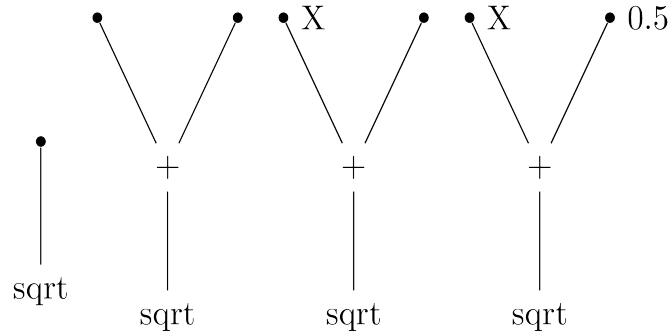


Figure 4: Sequence of steps constructing the program sqrt(x+5)

3) Functor + takes two inputs (arity(+)=2), therefore the tree bifurcates into two lines in this node. It randomly choses, as the first argument, the constant 5 ; and the variable X as the second argument.

Note that in step 3, both arguments were chosen from the terminal set. If they would have been chosen from the function set, the tree could bifurcate further. In order to prevent such growth of trees ad infinitum, a limiting « maximal tree depth » parameter is more than often implemented in GP scenarios.

Once such a program has been generated, one can evaluate its fitness by confronting it with diverse input arguments and comparing its output with a golden standard. Such a random-program generation & evaluation is repeated for all N initial candidate programs, subsequently the most individuals are selected and varied. While GP's selection techniques can sometimes closely ressemble selection techniques as used in GAs, variation operators are often of essentially different nature. This is so, because GP's not individual genomes or their linear sequences can be mutated or crossed-over, but rather complex and hierarchical networks of expressions. In a case of cross-over, for example, one switches whole sub-tree encoded within one

individual, for a sub-tree encoded within another one.

GP-based solutions cannot be expected to function correctly if they do not satisfy the theoretical properties of closure and sufficiency. In order to fulfill the closure condition, each function from the non-terminal set must be able to successfully operate both on output of any function in the non-terminal set and on any value obtainable by a member of the terminal set. Even behaviour of some simple operators thus has to be a priori adjusted (e.g. return 1 in case of division by zero) in order to assure correct functioning of the resulting program.

On the other hand, sufficiency property demands that the set of functors and terminals is sufficiently exhaustive. Otherwise the solution could not be found. One can not, for example, hope to discover equation for generating the Mandelbrot set if the initial set of terminals does not contain the notion of imaginary number, nor does the function set contain any other explicit or implicit reference to the notion of complex plane. Thus, while the closure constraint delimits the upper bound beyond which the discovery of the solution is not feasible, the sufficiency constraint delimits the lower bound of the minimal set of « initial components » which have to be defined a priori, so that discovery of the adequate program should be at least theoretically possible.

Other theoretical notions as well as diverse subtleties (special operators, methods how to distribute the initial population in the search space, fitness function proposals, domains of application, etc.) of practical implementation, are to be found in possibly the most important GP-concerning monography (Koza 1992).

1.7.4. Grammatical evolution

Grammatical Evolution (Gr.Ev) is a variant of GP in a sense that it also use evolutionary computing in order to automatically generate computer programs. The most important difference between Gr.Ev and GP is that while GP operates directly upon phenotypic trees representing program's code itself (for example in form of LISP expressions), Gr.Ev uses the evolutionary machinery for the purpose of generating grammars, which would subsequently generate the program code.

In Formal Language Theory, grammar is represented by the tuple {N, T, P, S} where N denotes the set of non-terminals, T the set of terminals, S is a symbol which is member of N and P denotes the set of production rules that substitute elements of N by elements of N, T or their combinations[6]. Consider a grammar exhaustive enough to encode programs able to perform arbitrary number of operations of addition or subtraction of two variables:

N = {expr, op, var}
T = { +, -, x, y}
S = expr
P = {
<op> → + | -

6 This is the case for so-called context-free and context-sensitive grammars.

<var> → x | y
<expr> → <var> | <expr> <op> <expr>

Such a grammar contains three non-terminals, non-terminal <op> which could be subtituted for either terminal + or terminal - ; non-terminal <var> which could be subtituted for either terminal x or terminal y ; and non-terminal <expr> which could be substituted for either a non-terminal <var>, or a sequence of non-terminals <expr> <op> <expr>. The fact that in this last production, the non-terminal <expr> is present both on left and right side of the substitution rule gives this grammar a possibility to recursively generate infinite number of expressions like :

x+x
x+y
y+x
y+y
x-x
x-y
y-y
y-x
x+x
x+x+x
x+x-x
x+x+y
x+x-y
x-x+y-y
x-x-y+y+x
y+y+x+x+y-x
etc.

Thus, even a very simple grammar with only four terminal symbols and three non-terminal symbols to each of which are associated only two production rules can theoretically produce an infinite number of distinct individual programs able to perform basic arithmetic operations with two variables.

Generation of a given resulting expression is determined by the order of application of specific production rules, starting with non-terminal symbol S. Such a sequence of application of production rules is called derivation. For example, in order to derive the individual « x+x », one has to apply production rules in following order :

S = <expr>
<expr> ::= <expr> <op> <expr>
<expr> ::= <var>
<var> ::= x
<op> :: = +
<expr> :: = <var>
<var> :: = x

while the individual « y-x » would be generated, if ever the starting symbol S should be expanded by a following sequence of production rules :

S = <expr>
<expr> ::= <expr> <op> <expr>
<expr> ::= <var>
<var> ::= y
<op> :: = -
<expr> :: = <var>
<var> :: = x

In Grammatical Evolution, it is this « order of application of production rules» which is encoded in the individual chromosome. In other terms, individual chromosomes encode when and where distinct production rules shall be applied. Figure 5 more closely illustrates, and puts into analogy with biological systems, the sequence of transformations which every binary chromosome undergoes during the process of unfolding into fully functional program :

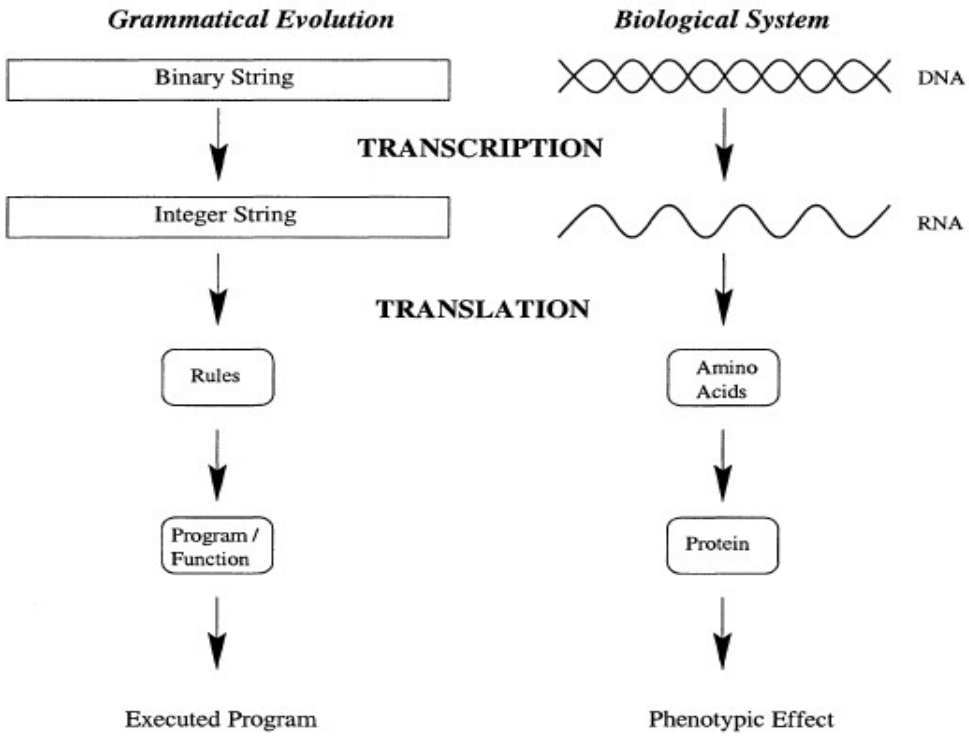


Figure 5: Sequence of transformations from genotype until phenotype in both Gr.Ev and Biological systems. Figure reproduced from (O'Neill & Ryan 2003).

It can be easily infered from the above-displayed schema that the approach of Gr.Ev is quite intricate and involves multiple steps of information processing. Whole process starts with binary chromosome subsequently split into 8-bit codons which yield an integer specifying which production rule to use in a given moment of program's generation. On many different layers does the « generation » process, as implemented in Gr.Ev, introduce and implement very original ideas like:

1. « Degenerate genetic code » - similary to « nature's choice » to encode one amino-acid by means of many different triplets, can one encode application of a unique production rule by more than one codon.

2. «Wrapping » - under certain conditions can be whole genome « traversed » more than once during the process of phenotypic expression. Specific codon can be thus used more than once during the compilation of single individual.
   etc.

Rationale for usage of such « biologically inspired tricks » is more closely presented in the work of the founders of Grammatical Evolution field (O'Neill & Ryan 2003) . They claim that the focus on genotype-phenotype distinction, especially in combination with implementation of « degenerate code » and « wrapping » notions, could result in compression of representation (& subsequent reduction of size of program search-space) and account for phenomenas like « neutral mutation », well-observed in biological systems, whereby a mutation occures in the genotype but does not have any effect upon the resulting phenotype. Another important advantage mentioned by O'Neill and Ryan is that Gr.Ev approach makes it very easy to generate programs in any arbitrary language. This is due to the versatility and generality of notion of « grammar ».

When compared with traditional GP technique, Gr.Ev was outperformed in a scenario when one had to find solutions to problem of symbolic regression. But in more case complex scenarios like « symbolic integration », « Santa Fe ant trial » or in scenario where one had to discover a most precise « caching algorithm », Gr.Ev significantly outperformed GP. Seminal work of (O'Neill and Ryan 2003) presents also some other interesting examples of practical application of Gr.Ev, for example in the domain of financial market prediction.

We note that while in many points (« grammar », « evolution ») does the work of O'Neilly and Ryan significantly overlap with ours, their aims significantly differ from those that shall be presented in our Thesis. More concretely, while Gr.Ev tends to offer a very general toolbox to generate useful computer programs in arbitrary programming language and used for solving arbitrary problems, our Thesis shall deploy the evolutionary computation machinery to shed some light upon diverse facets of one sole problem : that of « Natural Language Development».

Other important difference between the approach of Gr.Ev and the one we shall present in our thesis is that while in Gr.Ev, grammars are considered to be « generative devices », i.e. tools used for generation of programs, in our Thesis we shall use them as both « generative » and « parsing » devices. Another, even more fundamental difference is due to the fact that while « *At the heart of GE lies the fact that genes are only used to determine which rule is applied when, not what the rules are.* » (O'Neill and Ryan 2003), the evolutionary model of language-induction proposed in our Thesis shall aim to determine not only the order of application of the rules, but also the content of the rules themselves.

1.7.5. Tierra

Another example of how can one materialise evolutionary principles within an *in silico* framework is offered by Tierra, an artificial life simulation environment

programmed between 1990-2001 by Thomas S. Ray and his colleagues. Since Ray is an ecologist, his objective was not to develop an EC-like model in order to find or optimalize solutions of a given problem, rather he aimed to create a system where artificially entities could spontaneously evolve, co-evolve and potentially create whole artificial ecosystems.

An artificial entity in Tierra's framework (Ray 1992) is a program composed of sequence of instructions, chosen from instruction set containing 32 quite traditional assembler instructions somewhat tuned by the author so that their usage would facilitate « replication » of the code. Every artificial entity runs in its own « virtual CPU » but its code stays encoded in the « soup », i.e. piece of RAM which is potentially read-accessible to all other entities as well. Rare «cosmic ray » mutations flip the bits of « soup » from time to time, more variation is ensured by bit-flipping during the procedure whereby the entity replicates (i.e. copies) its code from the « mother cell » section of the soup to the « daughter cell » section.

Selection is in certain sense emulated by a so-called Reaper process which tends to stop the execution of programs which are either too old or contain too much flawed instructions. Other than that, there is nothing which ressemble the traditional notion of exogenously defined « fitness function ». For within Tierra, the survival (or death) of diverse species of programs is a direct consequence of species ability (or inability) to obtain access to limited ressources (CPU & memory).

Thus, after one seeds the initially empty soup with a manually constructed individual, containing 80-instructions allowing the individual to copy his code into the daughter cell of the memory, after the memory has been filled and the battle for ressources has started and once the mutation have generated sufficiently enough of variation, one can observe the emergence of dozens of new forms of replicable programs. Some of them being parasites, some of them being able to create algorithmic counter-mesures against parasites, one can literally observe an emergence of artificial yet living ecological system. It is therefore little surprising that Tierra could automatically evolve, among others, an individual containing just 22 instructions, capable of replication. That is, a replicator almost 4 times shorter than the replicator manually programmed by the conceptor of the system and injected into initial « soup ».

Currently the most famous descendant of Tierra is an AVIDA system (Ofria and Wilke 2004). Contrary to Tierra, however, is every AVIDA's individual encapsulated within its own virtual CPU and memory space. Tierra's Darwinian metaphore[7] of computer programs evolving by means of fighting for limited ressources is thus not so strictly followed.

7 http://life.ou.edu/pubs/tierra/node3.html

2. Language development

Language development (LD) is a constructionist process which endows humans with the capacity for transfering of information to, and obtaining of information from, other humans by means of verbal communication. Term « language development » shall be used preferably to « language acquisition » in order to mark the fact that the child not only passively « acquires » the language from environmental input but rather gradually builds it, in interaction with its environment. Sometimes the term « language learning » shall be used as well to denote the same process.

In our Thesis we shall focus only on modeling of development of « first language » , i.e. we shall aim to present a computational and evolutionary model of the process by means of which a human baby learns the language of its closest social environment.

Child's closest social environment are her parents, most notably her mother. Hundreds of studies were conducted to study the nature of « motherese », a special simplified language between mothers and their children (M. Harris 2013). While many studies point in divergent directions, they more or less agree that « *Maternal speech has certain characteristics that distinguish it from speech to other adults. These characteristics are in essence simplicity, brevity and redundancy.* » What's more, it seems to be a well-established fact that there exists a reciprocal link between the complexity of motherese and complexity of child's production. In other terms, mother's adjust their language according to the stage of child's linguistic development.

Other studies also indicate an existence of causal link between the quantity and simplicity of motherese utterances on one hand and child's linguistic development. More concretely, studies like that of (Furrow et al. 1979) indicate that child's confrontation with frequent and simple utterances facilitates their linguistic development while more complex style can slow their development down. Other studies, like that of Ellis & Wells (1980) precise that « *children who showed the earliest and most rapid language development received significantly more acknowledgments, corrections, prohibitions and instructions from their parents* ».

This causal link between mother's linguistic productions and child's developping linguistic competence shall play an important role when we shall discuss the « fitness function problem ». More concretely, we shall try to integrate into our computational models an idea that the fitness function evaluating the performance of child's internal categorization mechanism and/or candidate grammar shall be external to the child. The fitness function shall be given by mother's behaviour.

2.1. Ontogeny of semantic categories (concepts)

Natural language furnishes a communication channel for exchange of meanings. Meaning (also called « signifié » in traditional linguistics) is intentional, it refers to some external entity (also called « referent ») . Within the language L, meaning M can be denoted by a token (also called « signifiant ») and it is by exchange of physical (phonic, in case of spoken language, graphemic in case of written language etc.) manifestations of these tokens that producer (speaker|writer) and reciever

(hearer|reader) communicate.

Traditionally meaning of the word, i.e. its « semantics », was often considered as something almost «sacred » and impossible to formalize by mathematical means. Maximum which could be done, and had been done since Aristotle until middle of 20th century, was to define concept in terms of lists of « necessary and sufficient features ». Two types of features were considered to be both necessary and sufficient for definition of majority of concepts : first specifying concept's genus (or superordinated concept) and second specifying the particular property (differentia) which distinguished the concept from other members of the same genus. Thus, for example, « dog » could be defined as domesticated (differentia) canine (genus). Important property of such system of concepts was, that it allowed no ambigous or fuzzy border cases : the logical « law of excluded middle » guaranteed that all entities which were not both canines and domesticated at the same time (e.g. a chihuahua which passed all her life in wilderness) could not be called a dog.

The change of paradigm came slowly with works of late Wittgenstein[8] but especially with empirical studies of Eleanore Rosch (Rosch 1999) who realized that not only are concepts often defined by bundles of features which are neither necessary not sufficient, but that the degree with which a feature can be associated with a concept often varies. Subsequently, Rosch has proposed a « prototype theory » of semantic categories whose basic postulate is, that some members of the category (or some instances of the concept) can be more « central » in relation to the category (resp. concept) than others. Prototypical theory as well as other both theoretic and empirical advances, in combination with development of information-processing technologies, have paved the way to operationalization of semantics which allows us to transform meanings of words into mathematically commesurable entities.

In computational semantics, meaning of a token X observable within language corpus C is often characterized as a vector of relations which X holds with other tokens observable within the corpus. The set of such vectors associated to all tokens observable in C yields a « semantic space » which is a vector space within which one can effectuate diverse numeric and|or geometric operations. In short, concepts can be operationalized as geometric entities (Gärdenfors 2004).

« In the most simple case can be the vector which denotes concept X calculated as a linear combination of vectors of concepts in context of which X occurs » (Hromada 2013a). This is an algebraic form of famous « distributional hypothesis » stating that « a word is characterized by the company it keeps » (Z. S. Harris 1954) which can be considered to be the central dogma of statistical semantics. Distributional hypothesis is in certain a variation to an old « associationist » explanation of functioning of mind, which stated that the essence of mind is somehow related to mind's ability to create relations, i.e. associations, between successive mental states.

Both mind's faculty to create associations -considered by philosophers like Hume and Locke to be primary faculty of mind - as well as distributional hypothesis that

8 « *For a large class of cases of the employment of the word 'meaning'—though not for all—this way can be explained in this way: the meaning of a word is its use in the language.* » (Wittgenstein 2009)

meaning of symbol X can be defined in terms of meanings of symbols with which X co-occurs, can be, we believe, neurologically explained in terms of postulate first stated by Hebb, the neurologist :

« The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other. » (Hebb 1964)

One can assume that 1) if not only on single neurons but, mutatis mutandi, also whole neural circuits are governed by Hebb's rule, and 2) if distinct words W_x and W_y are somehow processed and represented by distinct neural circuits N_x and N_y THEN it shall follow that whenever a hearer shall hear (or speaker shall speak) the two-word phrase W_xW_y, the ensemble of material (synaptic?) relations between N_x and N_y shall get reinforced. In more geometrical terms, on a more « mental » level, such a « rapprochement » of N_x and N_y would be characterized by convergence of the geometrical representations of both circuits to their common geometrical centroid. Thus, after processing the phrase W_xW_y, the vectorial representations of both N_x and N_y will be closer to each other than before hearing (or generating) the phrase.

In our Thesis we shall presuppose that an associationist principle, similar to the one described above, is indeed at work whenever a human mind constructs a concept. We use term « concept » synonymously to the term « semantic class » : we define both concept and semantic classes as either subspaces of « semantic vector space », or as centroid points of such subspaces.

Theoretically, there are multiple (and possibly infinitely) many ways how a cognitive system can internally represent an external environment E (or, in case of a computational linguistic agent, a corpus C) as « semantic space » S of dimensionality D. It is important to notice that the overall partitioning of cognitive system's vector space determines how the system classifies the world. If system's ability to correctly classify the world determines the reproductive fitness of an organism within which the cognitive system is embedded, one can state that the topology of internally represented semantic space can quite directly influence organism's fitness.

Consider, for example, reproduction fitness of a member of prey species which sometimes mis-classifies a predator species for a sexual mate, and compare it to the fitness of such a an individual among prey species whose semantic space is optimized so that the probability of such mis-classification is practically reduced to zero.

A question whether such « semantic space optimization » occurs during the phylogeny of human species or whether it occurs principially during early years of child's developpement (i.e. ontogeny) is a variant of « nature vs. nurture » (Galton 1875) debate between « nativists » who bet on the «innateness » of certain faculties of human psyche (c.f. discussion above Evolutionary Psychology above); and empiricist who believe that practically all knowledge we dispose of and use in everyday life is acquired from environment. Being aware of results of studies suggesting that children of very small age dispose of knowledge concerning basic relations among physical objects, or even social and moral skills (Haidt 2012) we consider as unwise the tentative to label nativist position as a priori invalid. On the

other hand, being aware of the force with which processes like socialisation, acculturation and learning mould the psyche of an adult individual, we shall definitely consider as true the statement «topology of semantic space represented within the cognitive system of human individual can be optimized by supervised assimilation of knowledge encoded in surrounding environment».

Notwithstanding the answer to nature & nurture question in regards to human faculty of categorization, the part of our Thesis devoted to «evolutionary models of concept construction » shall simply suggest that something like optimization of semantic spaces by means of evolutionary computing is, indeed, possible.

2.2. Ontogeny of formal categories (parts-of-speech)

Words of language can be also partitioned into classes independently from their semantic content. For example, while there is practically no manifestly evident semantic feature between words like « apple » and « process », they can be both considered as belonging to the same category of « nouns ». Principal reason for this being the fact that within a sentence like, for example, «This apple makes me happy» one can freely substitute « apple » for « process » and still obtain a grammaticaly correct sentence.

Sometimes the formal categories and semantic categories partially overlap. Such is the case, for example, in many indo-european languages where one often finds « feminine » nouns marked with markers of one formal group and « masculine » nouns marked with markers of other group. Even more extreme case of such « overlap » of semantic and formal categorization processes was observed among Diyarbal aborigines of Australia who use the same determiner « balan » (in certain sense analogic to German article « die ») in front of all nouns referring to « woman, fire and dangerous things» (Lakoff 1990). In modern linguistic tradition, however, are semantic and formal categories considered to be independent from each other.

There exist multiple dimensions along which linguistic tokens can be categorized into formal classes. Most importantly, the appartenance of word W to class C can be principially infered from : 1) its position in regards to other words 2) its morphology (i.e. its internal composition with all prefixes, word root, suffixes etc.)[9]. It is also important to realize that the same token can belong to many different categories in the same time and that the relations between categories themselves could be either inclusive, for « nested » categories, or « orthogonal ». Thus, for nested categories, appartenance of , for example, german token « die Schönheit » to «gender» subcategory «feminine» immediately implies that it also belongs to part-of-speech «noun ». On the other hand the sole fact that it is « feminine » does not inform us whether it could be attributed to « nominative » or « accusative » subsubcategories of grammatic subcategory « case ». Thus, subcategories of « case » and « gender », while being both « nested » within the part-of-speech category « nouns » are orthogonal to each other[10].

9 C.f. (Hromada 2014a) for a comparative study assessing the impact of morphology and word-order features upon POS-induction in Bulgarian, Czech, Estonian, Farsi, English, Hungarian, Polish, Romanian, Russian and Slovak.

10 The theoretical importance of existence of this distinction in regards to current formal grammar models of natural

On the most abstract level, linguistic tokens can be categorized into two principal 0-level formal categories of «functional» and « lexical » items. The set of grammatical items is closed, and it contains such parts-of-speech as determiners, conjunctions, pronouns, prepositions. On the other hand, classes of « lexical items » are opened and include meaning-carrying parts-of-speech like nouns, verbs, adverbs, adjectives etc. Study by (Shi et al. 1999)offers evidence that even newborn children (1-3 days old !) react differently to lexical and functional words and are thus «*able to categorically discriminate these sets of words based on a constellation of perceptual cues that distinguish them*».

Once children are able to distinguish functional words from lexical ones, the process of ontogeny of formal categories can proceed towards development of part-of-speech categories. While it would be definitely mistaken to state that all languages of the world can be partitioned into & mapped upon part-of-speech languages known from English or other indo-european languages (i.e. noun, adjectives, pronouns, verbs, adverbs, preposition, conjunction, interjections), linguists generally agree that some kind of « noun»-ressembling and «verb»-ressembling categories are to be observed in all systems of human verbal communication.

It is undoubtably the case that between the birth and cca 2-years of age, prototype for such part-of-speech clusters are being formed within the child's cognitive system. This has to be so, around age of 2, children usually start to apply specific rules to specific items (i.e. start to conjugate the verbs or declinate the nouns). Subsequently, the learning of much more subtle distinctions, related to nature of grammatical categories like genus, casus, numerus for nouns or modus, tempus, etc. for verbs can take place. For diverse case studies concerning the acquisition of formal categories, c.f. (Y. E. Levy, Schlesinger, & Braine, 1988).

Acquisition of both semantic and formal linguistic categories is facilitated by so-called « variation sets » (VS). One observes a linguistic variation set whenever the identific word/cluster of words occur in identical or slightly variated form within multiple consequent utterances. Not only nursery rhymes and lullabies are filled with such « alternations in maternal self-repetitions » (Hoff-Ginsberg 1986) VS are also highly frequent in standard « motherese ». In Turkish, for example, VS seem to make up approximately 20% of child-directed speech (Küntay and Slobin 1996) and very similar proportions are also reported for English language (Brodsky et al. 2007).

Note that the notion of « variation set » can be intepreted in terms of evolutionary theory, given that:

- maternal self-repetition can be intepreted as a form of « replication in time», whereby every single utterance is considered to be an independent individual
- alteration of form between subsequent utterances can be interpreted as a result of a variation operator influencing mother's production of new sentences

In context of our tentatives to explain language development in terms of evolutionary theory and suggest its validity by means of evolutionary computation model, we find

languages shall be further extended in fulll version of the Thesis.

this insight « *the image that best characterizes the young language leaner is that of a multilevel analyzer who is working with* ***several types of analysis simulatenously****, with different degrees of success, as learning progresses* » (Levy 1988).

It may be stated that the reason why categorization processes develop in the first place is congitive system's the tendency to optimize its functions and structures. As Maratsos (1998) put it: « *Once the speaker hears just one grammatical use of a new word which suffices to identify its membership in a category, he can refer to the whole system of rules involving this category* » (Maratsos 1988).

Thus, both semantic as well as formal categories can reduce cost of processing and storing of information by and within the cognitive system.

2.3. Ontogeny of grammars (grammar induction)

Partitioning of words into grammatical categories can be useful only if it is accompanied by development of grammatical rules which combine members of diverse categories in order to produce meaningful sentences. We reiterate that strictly formally, grammar is defined as the tuple {N, T, P, S} where N denotes the set of non-terminals, T the set of terminals, S is a symbol which is member of N and P denotes the set of production rules that substitute elements of N by elements of N, T or their combinations.

Within such formal framework, the problem of partitioning of words into diverse grammatical categories can thought to be as equivalent to problem of discovery of production rules which 1) associate members of T (words) to members of N (labels of distinct categories) 2) combine elemets of N in order to produce new elements of N. In fact, the problem of construction of formal categories and discovery of grammatical rules are mutually intertwined, some researchers go even so far as to state : « ***Category symbols****, whether in phrase structure rules or in the lexicon,* ***are logically equivalent to the rules written on them****, and as such are completely system-dependent : They are shorthand descriptions of the rule system as a whole. By anyone's theory, young children's linguistic system does not possess all the features of the endstate system. In other words, their language cannot be describe by the same grammar as the adult system*» (Ninio 1988).

In litterature, development of language is often described as a process composed of three « stages » which can be subsequently subdivised in a followin manner :

«Pregrammatical :

> *a.* ***Rote-learning*** *– item-based acquisition is manifested in the use of formally unanalyzed units or chunks ;*
>
> *b.* ***Initial modifications*** *– formal alternations apply to a small number of highly familiar, good exemplars ;*

Structure-bound

> *c.* ***Interim schemata*** *– transitional or bridge strategies take the form of productive, but nonnormative rules ;*

d. ***Grammaticization*** *– structure-bound rules are those of the endstate grammar ;*

Discourse-oriented :

e. ***Convention and variety*** *– grammatical rules are deployed with appropriate, discourse-sensitive lexical restrictions, stylistic alternations, usage conventions, register distinctions etc.*

» (Berman 1988).

In our Thesis, we shall put aside the intricacies of the third, « Discourse-oriented » stage and shall focus on « Pregrammatical » and « Structure-bound » stages. More concretely, we shall aim to explain acquisition of words and word chunks in phase a. as the result of the **« crossover » between structures present in the environment and structures represented within the cognitive system**; while the gradual emergence of categories and associated production rules which can be observable during phases b. c. d. shall be explained not only in terms of informatic crossover of structures present in environment and represented in cognitive system but also as the result of purely internal replication, variation and decay, proper to the cognitive system, and resulting in complexity-increasing « battle for resources » among structures represented within it.

We are convinced that introduction of such «cognitive-system internally variying operators » like « entropy-induced decay » (associated to the phenomenon of « forgetting ») and « structural merging » (associated to the phenomenon of « dreaming ») we can, for example, offer a very simple&natural yet effective solution to a so-called « overgeneralization »[11] problem. When it comes to overgeneralization of grammatical rules, they are often observable in phases c. & d. (i.e. between 2-4 years of age) whenever the child applies the production rule beyond the scope of its validity. The most famous example of overregularization in English is that practically all children apply the rule $V_{past} \rightarrow V_{Present}$+ed on all verbs. Thus, especially during MLU stage 4 and 5[12], they generate past participles like « throwed » or « braked » which are not correct. What is fascinating about the problem of overregularization is not only that all children shall start to employ irregular forms of past participles so that errors are not reproduced anymore ; but especially the fact that often, children used the correct « irregular form » even before (i.e. in one-word phases a. and b.). Only later did they converge to incorrect overregularization : « *Initially, children's uses of -ed past tense are all accurate. They may say melted or dropped, but not, as they later do, runned and breaked* » (Maratsos 1988) .

We see an important analogy between observations of such sequence of correct/incorrect/correct behaviour, and general behaviour of evolutionary systems which also often « reject » locally optimal solutions and descend into fitness

11 According to the domain (formal, semantic) the problem is also sometimes named as « overextension », « overregularization » or the problem of « overinclusive grammar ».

12 MLU means «Mean Length of Utterance » and is a measure traditionally used in developmental psycholinguistics for assessing of child's linguistic performance at given age. In period when child produces one-word utterances like « mama » , « tato », MLU is considered to be 1 ; later when child starts to say two-word utterances like « mama nene », MLU increases towards 2 etc.

landscape valleys in order to subsequently climb towards more optimal states. Thus, we believe that the term « conflict » present in the following principle can be also interpreted in evolutionary sense :

« *Whenever a newly acquired specific rule (i.e. a rule that mentions a specific lexical item, like throw, make, allow, report) is in* ***conflict*** *with previously learned general rule (i.e. a rule that would apply to that lexical item but also to many others of the same class), the specific rule eventually takes precedence* » (Braine 1971).

McWhinney uses a similar term « competition » to label its Competition Model of linguistic competence. « *The competition model assumes that lexical elements and components to which they are connected can vary in their degree of activation. Activation is passed along connections between nodes. During processing, items are in competition with one another. In auditory processing …, in allomorphic processing …, in the processing of role relations, in polysemy …, the item that wins out in a given competition is the one with the greatest activation* » (MacWhinney 1987).

If one could interpret the last phrase of the above citation as « the component which has the greatest activation has the greatest fitness and thus the highest probability of being replicated within the cognitve system », one could consider MacWhinney's connectionist model as an evolutionary one, and thus pointing in our direction. But since that is not the case, and since it seems that MacWhinney's model does not, at least not explicitly, involve any processes of replication, nor sources of random variation nor does it explicitely work with «populations of grammars», we are obliged to look for another theoretical framework which could more easily integrate such notions.

It may be the case that a so-called theory of « Grammar Systems » (Csuhaj-Varjú 1994) and « Language Colonies » (Kelemen and Kelemenová 1992) could furnish such a framework for our tentative to explain ontogeny of grammar in human individuum as an evolutionary process. Both will be introduced in part 4 of this text.

3. Computational Models of Text Processing

Majority of models and algorithms presented in this chapter are results of intellectual work of computational linguists working in domain of « Natural Language Processing » (NLP). In NLP, one processes data encoding natural (human) languages with computational methods which often involve machine learning, data mining, information retrieval, statistical inference or artificial intelligence (AI) algorithms. Among principal objectives of NLP can one include : 1) to allow machines to « understand » and|or work with meanings 2) to develop an autonomous artificial agent (Hromada, 2012) able to pass the Turing Test (Turing 2008); and 3) to elucidate, by means of computational simulations, possible ways how human cognitive system treats natural language.

Computational aim of our Thesis overlaps especially with NLP's third objective. Such an aim bring with itself many complex problems not easy to tackle and thus, in order to reduce their amount and complexity we shall reduce the notion of « Natural Language » to the notion of « text ». It is true that in doing so, we shall completely ignore the phonetic, phonologic and prosodic aspect of language which has been, during practically all human history, a principal way how human speakers encoded their messages in order to transfer them to other human hearers. It is only during few centuries that the communication by means of text became prominent and only within last decades it became dominant, mainly because of increasing role of computers in our lives. This is at least partially so because computers are essentially machine built for processing of sequences of discrete symbols and that's what a text is – a sequence of discrete symbols. Contrary to flux of spoken language, which is also a sequence, but composed of units whose boundaries are often unclear and whose features overlap.

3.1. Concept construction

We define the « concept construction » (CC) problem as an open-class variant of « classification » or « categorization » problem. In classical, « closed-class » categorization problem, the objective is to assign a label which denotes the membership to a category C_1 to a set of objects disposing of particular combination of properties (also called « features » in AI community) ; and assign to categories C_2, C_3 etc. other objects disposing of different features. Problem of « binary classification » where only two categories are involved, is well studied and dozens of diverse algorithms exists which allow to train, in machine learning scenario, such classification models (« classifiers ») which will subsequently quite successfully classify such objects of the « testing set » which absent in the « training set ».

In NLP one often solves classification problem by means of so-called « Support Vector Machines » . During the traininig of SVM, algorithm tries to discover a hyperplane « that has the largest distance to the nearest training data point of any class » (Vapnik et al. 1997). SVMs belong to group of « linear classifiers » which all base their classification decisions on linear combinations of characteristics (features) of objects-to-be-classified. Other machine learning algorithms as diverse as Linear

Discrimant Analysis, Naive Bayes classifiers, logistic regression or perceptron also belong to group of linear classifiers.

« Multiple class » variants of these algorithms also exist, allowing for classification of objects into more than 2 categories. In case of all these algorithms, however, all the classes-to-be-looked-for are known in advance ; datapoints in the training set are labeled with labels belonging to a finite set and after the training, during the testing phase, one's objective is simply to attribute the correct label to a new object. While the object itself was most probably not present in the training set and is not « new », the finite set of all class/category labels-to-be-attributed are well known from the very beginning of training. In this sense all algorithms mentioned above address the closed-class variant of classification problem.

On the contrary, in open-class variant of classification problem one can be potentially asked, in the testing phase, to attribute to an object, which was not present during training phase, a label which was also not present in the turing phase. In other terms, in open-class variant of classification problem one does not know in advance neither the number nor even the nature of categories which are to be constructed.

3.1.1. Non-evolutionary model of CC

One possible way how one can address problem of Concept Construction – which we consider to be the instance of an « open-class classification problem » as defined above – is described as follows:

1. During the (train|learn)ing phase, use the training corpus to create a D-dimensional semantic vector space, i.e. attribute the vectors of length D to all members of the set of entities (word fragments, words, documents, phrases, patterns) E which includes all observables within the training corpus
2. During the testing phase :

   2.1 characterize the object (text) O by a vector $\vec{o}$ calculated as a linear combination of vectors of features which are observable in O and whose vectors were learned during the training phase

   2.2 characterize labels-to-be-attributed L_1, L_2, ... by vectors $\vec{l}_1, \vec{l}_2$ ...

   2.3 associate the object O with the closest label. In case we use cosine metric, we minimize angle between $\vec{o}$ and label vectors, i.e. $arg\,max \cos(\vec{o}, \vec{l}_x)$

Note that in order to make this approach functional, two important conditions have to be fulfilled. Primo, vectors associated to entities observables within the training corpus must be commesurable, i.e. have to be of same dimensionality and be members of the same vector space. Secundo, the set of all entities E observed during learning has to be sufficiently exhaustive, so that potentially any novel label or object which shall appear during the testing phase could be at least partially characterized in terms of members observables during the training phase.

The first condition of « entity commesurability » is not fulfilled by many vector space models which often yield multiple spaces for entities of different « types ». In such

models, « word » entities are often encoded as rows of the matrix and « context » or « documents » entities, i.e. entities within which the words entities occur, are encoded as column of the same matrix, or are encoded in a completely different matrix. On the contrary, algorithms like Random Indexing (RI) or Reflective Random Indexing (RRI) construct semantic vector spaces from initial textual corpora in a way that everything they encounter – be it syllables, words or whole documents – is ultimately represented as rows of the same matrix.

RI and RRI have also other advantages which are more closely described elsewhere (Sahlgren 2005; Cohen et al. 2010; Hromada 2013b). For the purpose of this article let's just underline the fact that both RI and RRI can be quite computationally efficient since they are able to « project » semantic relations hidden in the text upon a vector space with restrained dimensionality. Theoretically, this is permitted due to a so-called lemma Johnson-Lindenstrauss stating that « *a small set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved* » (Johnson and Lindenstrauss 1984)

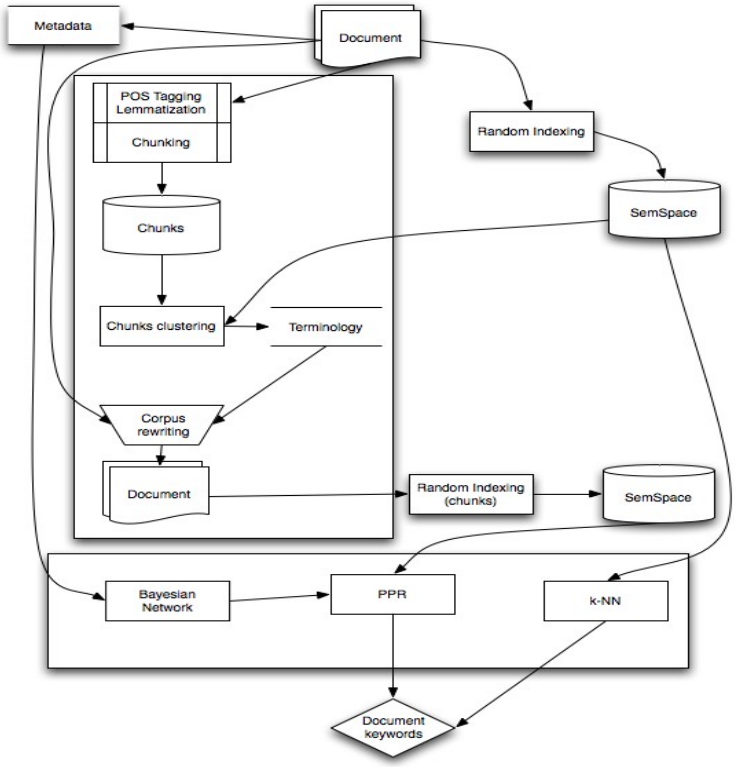


Figure 6: Description of DEFT2012 system for automatic attribution of keywords to scientific articles. Figure reproduced from poster

In 2012, a hybrid system with RRI semantic component at its very core, was deployed in a francophone datamining competition DEFT2012 (El Ghali et al. 2012). The goal of the competition was to create such an automatic NLP system which would be able to attribute to scientific articles the same keywords as were attributed by their authors. In other terms, the goal was to artificially simulate the cognitive

activity of « attributing a conceptual label » to a scientific article. The tricky thing about the problem was that it was not a standard « closed class » classification problem, but indeed an « open class » problem since there were many keywords labels which have not been present in the training set, yet were to be associated in the testing scenario. Figure 6 illustrates relations among diverse components of this hybrid system.

As may be easily seen, whole « artillery » of diverse NLP tools like POS-taggers, lemmatizers and chunkers was deployed in order to yield sufficiently exhaustive set of features from which two distinct semantic spaces were composed by means of RRI. Resulting semantic spaces were subsequently post-optimized by combining probabilistic Bayesian Networks and production rules.

In the first simpler task of the competition DEFT2012, the system has attained F-Score of 94.8%. The task was simpler because a list of candidate labels was furnished within training corpus and subsequently another list of candidate keywords was furnished with the testing corpus. The system has attained F-score of 58.7% in a second, more difficult task where no such lists were given. In both tasks it outperformed the systems deployed by other 9 participants of the competition.

3.1.2. An evolutionary model of CC

Task 4 of 2014 edition of the datamining competition Defi en Fouille Textuelle (DEFT) was understood as an instance of classification problem with opened number of classes. More concretely, the challenge was to create an artificial system which would be able attribute a specific member of the set of all class labels to scientific articles of the testing corpus. The training corpus of 208 scientific articles presented in diverse sessions of diverse editions of an annual TALN/RECITAL conference was furnished to facilitate the training of the model.

To solve this problem, we have proposed an algorithm consisting of two nested components, as represented on Figure 7. The inner component, which we call Reflective Space Indexing (RSI) is responsable for construction of the vector space. Its input is a genotype, the list of D features which trigger the whole reflective process, its output -a phenotype - is a D-dimensional vector space consisting of vectors for all features, objects (documents) and classes. The inner component is « reflective » in a sense that it multi-iteratively not only characterizes objects in terms of their associated features, but also features in terms of associated objects. RSI's principal parameter is the number of dimensions of the resulting space (D). Input of RSI is a vector of length D whose D elements denote D « triggering features », the initial conditions to which the algorithm is sensible in the initial iteration. After the algorithm has received such an input, it subsequently characterizes every object O (document) by a vector of values which represent the frequency of triggering feature in object O. Initially, every document is thus characterized as a sort of bag-of-triggering-features vector. Subsequently, vectors of all features – i.e. not only triggering ones – are calculated as a sum of vectors of documents within which they occur and a new iteration can start. In it, initial document vectors are discarded and

new document vectors are obtained as a sum of vectors of features which are observable in the document. Whole process can be iterated multiple times until the system converges to stationary state, but it is often the second and third iteration which yields most interesting results. Note also that what applies for features and objects applies, *mutatis mutandi*, also for class labels.

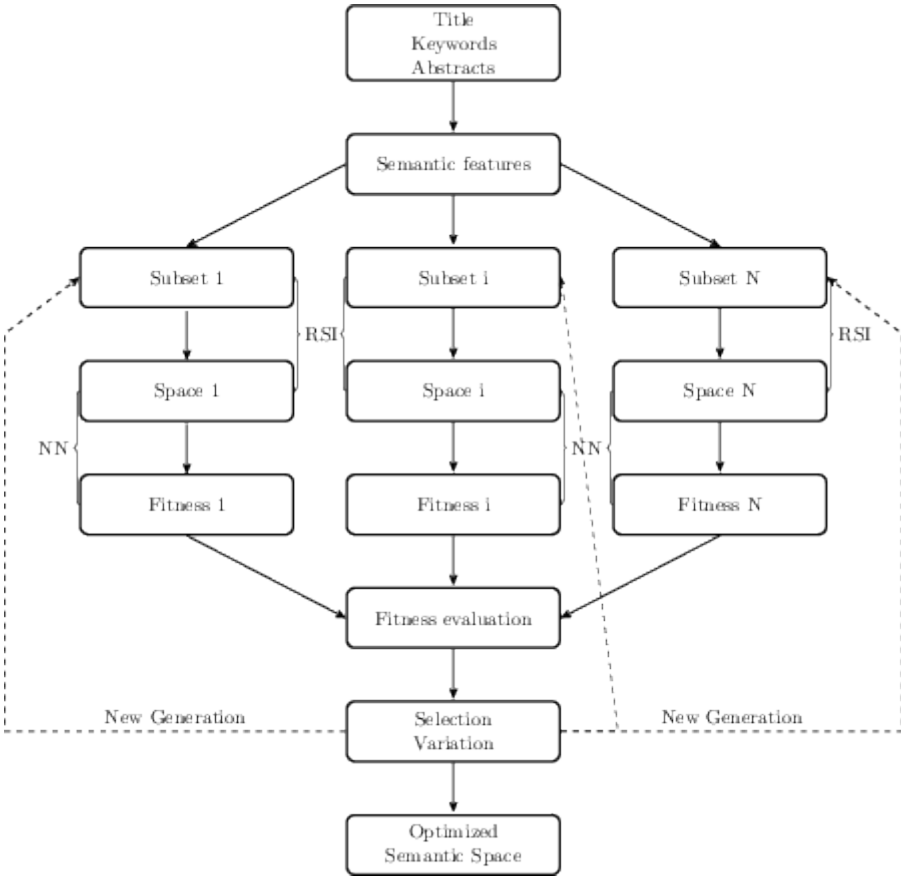


Figure 7: Diagram of DEFT2014 model, embedding the construction of semantic spaces within an evolutionary framework.

For purposes of DEFT 2014, every individual RSI run consisted of 2 iterations and yielded 200-dimensional space.

The envelopping outer component is a trivial evolutionary algorithm whose task was to find the most « fit » combination of features to perform the classification task. In every « generation », evolutionary component injects multiple individual lists of triggering features (i.e. « genomes ») into the inner component and subsequently evaluates the fitness function of resulting vector spaces. It subsequently mutates, selects and crosses-over genotypes which had yielded the vector spaces wherein the classification was most precise.

The evolutionary component of the system was conceived as a sort of feature selection mechanism. The objective of the optimization was to find such a genotype – i.e. such a list of « triggering features » – which would subsequently lead to discovery of a vector space whose topology would facilitate construction of a most classification-friendly vector space.

As is common in evolutionary computing domain, whole process was started by creation of a random population of individuals. Each individual was fully described by a genome composed of 200 genes. Initially, every gene is assigned a value randomly chosen from the pool of 5849 feature types observable in the training corpus. In DEFT2014's Task 4 there were thus 5849^{200} possible individual genotypes one could potentially generate and we consider it important to underline that classificatory performance of phenotypes, i.e. vector spaces generated by RSI from genotypes, can also substantially vary.

What's more, our observations indicate that by submitting the genotype to evolutionary pressures -i.e. by discarding the least « fit » genomes and promoting, varying and replicating the most fit ones - one also augments the classificatory performance of the resulting phenotypical vector space. In other terms, search for a vector space1 which is optimal in regards to subsequent partitioning or clustering can be accelerated by means of evolutionary computation.

During the training, evaluation of fitness of every individual in every generation proceeded in a following manner :

- pass the genotype as an input to RSI (D=200, I=2)
- within the resulting vector space, calculate cosines between all document and class vectors
- attribute N documents with highest score to every class label (N was furnished for both testing and training corpus)
- calculate the precision in regards to training corpus golden standard. Precision is considered to be equivalent to individual's fitness

Size of population was 50 individuals. In every generation, after the fitness of all individuals has been evaluated, 40% of new individuals were generated from the old ones by means of a one-point crossover operator whereby the probability of the individual to be chosen as a parent was proportional to individual's fitness. For the rest of the new population, it was generated from the old one by combination of fitness proportionate selection and mutation occuring with 0.01 probability. Mutation was implemented as a replacement of a value in a genome by another value, randomly chosen in the pool of 5849 feature types. Advanced techniques like parallel evolutionary algorithms or parameter auto-adaptation were not used in the study.

While algorithm succeeded to optimize the vector space generated to training corpus with precision of 87%. However, the resulting model over-fit the training corpus and failed to be fully transferable on testing corpus. Possibly due to implementation error – c.f. (Hromada 2014b) for closer discussion- the model has thus achieved only 27 % precision when confronted testing data. While being definitely more performant than a random baseline, our approach was the least performant among 5 participants of DEFT2014.

Notwithstanding the failure of our model in DEFT2014, we consider as an important our observation that « by evolutionary selection of chromosome of features which initially « trigger » the reflective process one can, indeed, optimize the topology and hence the classification performance of the resulting vector space » (Hromada 2014b).

3.2. Part-of-speech induction and part-of-speech tagging

The term Part-of-speech-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 linguistic token observable in agent’s linguistic environment. POS-i can be understood as a « partitioning problem » since one’s objective is to partition the initial set of all tokens occuring in corpus C (which represent agent’s linguistic environment E) into N subsets (partitions, clusters) whose members would correspond to grammatical categories as defined by the gold standard. Because one does not use any information about « ideal » gold standard grammatical categories during the training phase and uses it only for final evaluation of the performance of the model, POS-i is considered to be an « unsupervised » machine learning problem.

POS-i’s « supervised » counterpart is the problem of POS-tagging. In POS-tagging, one trains the system by serving it, during the training phase, sequence of couples (word W, tag T) where tag T is the label denoting the grammatical category into which the word W belongs. POS-tagging is thus simpler than POS-i where no information about ideal labels is furnished during the learning. Training of POS-tagging systems is of particular importance especially for languages where many word forms can potentially belong to many part-of-speech categories (in English, for example, can almost any noun play also role of the verb; token like « still » can be intepreted as substantive, verb, adjective and even adverb, its POS-category being determined by its context). On the contrary, in morphologically rich languages where such a « homonymy of forms » is present in lesser degrees and relations between word types and classes are less ambigous, one can often simply train the POS-tagging system by simply memorizing an exhaustive list of (W, T) couples.

3.2.1. Non-evolutionary models of POS-i

The paradigm currently dominating the POS-i domain was fully born with article published by Brown et al. in 1992. Brown and his colleagues have applied the information theoretic notion of « mutual information » :

$$M(w_1 w_2) = \log \frac{Pr(w_1 w_2)}{Pr(w_1) Pr(w_2)}$$

upon all bigrams (i.e. sequences of two words) composed of tokens w_1, w_2 and had subsequently devised a merging algorithm able to group words into classes in a way that the mutual information within a class would be maximized.

In two decades since publication of study of Brown et al., their approach has inspired hundreds of studies : 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 co-occurence of words with n-gram sequences to be the primary source of relevant information for subsequent creation of part-of-speech clusters. In all these models, one aims to discover the ideal parameters of Markovian statistical models, often employing a so-called Expectation-Maximization (EM) algorithm to discover the optimal partitioning. Unfortunately, EM is unable to quit locally optimal states once they were discovered. Notwithstanding this disadvantage, comparative study of (Christodoulopoulos et al. 2010) suggests that probabilistic models of part-of-speech induction can be indeed very performant.

POS-i induction can be also realized by means of k-means clustering algorithm, or one of its variants. K-means algorithm (Karypis 2002) partitions N observations, described as vectors in D-dimensional space, into K clusters by attributing every observation into the cluster with the nearest centroid (i.e. mean). If one considers these centroids to denote prototypes of the categories in center of which they are located, then one can consider the k-means algorithm to be consistent with « prototype theory of categorization », as proposed by Rosch. Table 1 illustrates simple K-mean partitioning of tokens present in English version of Orwell's 1984.

Table 1. K-means clustering of tokens according both suffixal and co-occurence informations. Table partially reproduced from (Hromada 2014c)

	Noun	**Verb**
0	10	3
1	**568**	67
2	97	**668**
3	13	**1011**
4	**1173**	67
5	608	**958**
6	**1977**	97

In this example case we have clustered all tokens observable in the corpus into 7 clusters according to features both internal to the token – i.e. suffixes – and external – i.e. co-occurrence with other tokens**.** Note that even such a simple model where no machine learning or optimization were performed, K-means algorithm somehow succeeds to distinguish verbs from nouns. As is shown in the Table 1, whose columns represent the "gold standard" tags and rows denote the artificially induced clusters, even such a naïve computational model has assigned 83.6% of nouns to clusters 1, 4 and 6 while assigning 91.8% of verbs into clusters 2, 3 and 5.

3.2.2. Evolutionary models of POS-i & POS-t

Usage of evolutionary computing in NLP is - in comparison to other methods like neural networks, Hidden Markov Models, Conditional Random Fields or SVMs –

still very rare. This is also the case to NLP's sub-problem of part-of-speech tagging and thus we are aware of only one tentative to use genetic algorithms to train a part-of-speech tagger :

In his (Araujo 2002) proposal, Araujo describes a system of POS-t involving crossover and mutation operators. What is particularly interesting about Araujo's system is that **separate evolution process is run for every separate sentence of the test corpus**. Training corpus, on the other hand, serves mainly as a source of statistical information concerning co-occurrences of diverse words and tags in diverse word & tag contexts. This information concerning the « global » statistic properties of the training corpus is later exploited in computation of fitness.

Let's take, for example, the phrase « Ring the bell ». Since words like « ring » and « bell » are in English sometimes used as verbs, and sometimes used as nouns, such a sentence can be tagged at least in 4 different ways :

N D[13] N

V D V

N D V

V D N

Such sequences of tags yiels individual members of Araujo's initial population of chromosomes. In languages like English where almost every word can be attributed to more than one POS category & the number of possible tag sequences therefore increases with length of the phrase-to-be-tagged, one will be most probably obliged to randomly choose such initial individuals. Fitness of every individual possibly tagging the sentence of n words is subsequently calculated as a sum of accuracies of tags (genes) on position i :

$$\sum_{i=0}^{n} f(g_i)$$

Accuracy g_i of an individual gene is calculated as :

$$f(g_i) = \log\left(\frac{context_i}{all_i}\right)$$

whereby values of $context_i$ and all_i are extracted from the training table which was constructed during the training phase and represent the overall frequency of occurrence of word wi within specific ($context_i$) and all (all_i) contexts.

Once fitness is evaluated, fitness-proportional crossing-over (50%) and mutation (5%) is realized. Notwithstanding the fact that Araunjo doesn't seem to have used any other selection mechanism, in less than 100 generations, populations seemed to converge into sequence of tags which were more than 95% correct in regards to gold standard. This is a result comparable to other POS-tagging systems but with lesser computational cost. It is also worth noting that Araujo's experiments indicate that working solely with contextual window W_L, W, W_R , i.e. just looking one word to the

13 We denote, by a non-terminal symbol D, the category of « determiners » into which belongs also article « the ».

left and one word to the right, seems to yield, in case of POS-tagging of English language higher scores than extracting data from larger contextual spans.

When it comes to the « unsupervised » variant of the POS-t problem, id est the problem of Part-of-speech induction, up to this date there have been -as far as we know - no tentatives to address the POS-i problem by means of evolutionary computing. For this reason, and for the reason that we see strong analogies between problems of CC and POS-i, our Thesis shall aim to solve this problem with a model similar to the one which we have presented in part 3.1.2 of this work.

3.3. Grammar induction

Input of Grammar Induction (GI) process is a corpus of sentences written in language L, its output is, ideally a grammar (i.e. a tuplet G={S,N,T,P} as defined in above chapters) or at least a model able to generate language sentences of L, including such sentences that were not present in the initial training corpus.

The nature of resulting grammar is closely associated to the content of the initial corpus as well as to the nature of the inductive (learning) process. According to their « expressive power », all grammars can be located somewhere on a « specificity – generality » spectrum. On one extreme of the spectrum lies the grammar having following production rules :

$$1 \rightarrow 2^*$$

$$2 \rightarrow a \mid b \mid c \ldots Z$$

whereby * means « repeat as many times as You Want ». This very compact grammar can potentially generate any text of any size and as such is very general. But exactly because it can accept any alphabetic sequence and thus does not have any « discriminatory power » whatsoever, is such a grammar completely useless as an explication of system of any natural language.

On the other extreme lies a completely specific grammar which has just one rule :

$$1 \rightarrow \text{<corpus>}$$

This grammar contains exactly what corpus C contains and is thus not compact at all (it is even two symbols longer than C). Such a grammar is not able to encode anything else than the sequence which was literally present in the training corpus and is therefore also useless for any scenario were novel sentences are to be generated (or accepted).

The objective of GI process is to discover, departing solely from corpus C (which is written in language L), a grammar which is neither too specific, nor too general. If it is too general, it shall « overgeneralize », i.e. shall be able to generate (or accept) sentences which aren't be considered as grammaticaly correct by common speaker of L. If it is too specific, it shan't be able to represent all sentences contained in C or, if it shall, it shan't be able to generate (or accept) any sentence which is considered to be sentence of L but was not present in the initial training corpus C.

3.3.1. Non-evolutionary models of grammar induction

One of the first serious computational models of GI is a « Syntagmatic – Paradigmatic » (SNPR) model presented in (Wolff 1988). Its core algorithm is presented in Table 2.

TABLE 2 Outline of Processing in the SNPR Model (reproduced from Wolff, 1988)

1. Read in a sample of language.
2. Set up a data structure of elements (grammatical rules) containing, at this stage, only the primitive elements of the system.
3. WHILE there are not enough elements formed, do the following sequence of operations repeatedly:
   BEGIN
   - 3.1 Using the current structure of elements, parse the language sample, recording the frequencies of all pairs of contiguous elements and the frequencies of individual elements. During the parsing, monitor the use of PAR elements to gather data for later us in rebuilding of elements.
   - 3.2 When the sample has been parsed, rebuild any elements that require it.
   - 3.3 Search amongst the current set of elements for shared contexts and fold the data structures in the way explained in the text.
   - 3.4 Generalize the grammatical rules.
   - 3.5 The most frequent pair of contiguous elements recorded under 3.1 is formed into a single new SYN element and added to the data structure. All frequency information is then discarded.

   END

We consider the SNPR model to be of particular importance because of its aim to explain the process of Grammar Induction as a sort of cognitive optimization : « *The central idea in the theory is that language acquisition and other areas of cognitive development are, in large part, processes of building cognitive structures which are in some sense optimal for the several functions they have to perform* » (Wolff 1988). Wolff also associates his « cognitive optimization hypothesis » with a «law of cumulative complexity » postulated in a study (Brown 1973) which is considered to be tha big classics of language development litterature : «*if one structure contains everything that another structure contains and more then it will be acquired later than that other structure* » (Wolff 1988).

Grammar resulting from such a contact between language sample and SNPR inducing mechanism is displayed on figure 7.

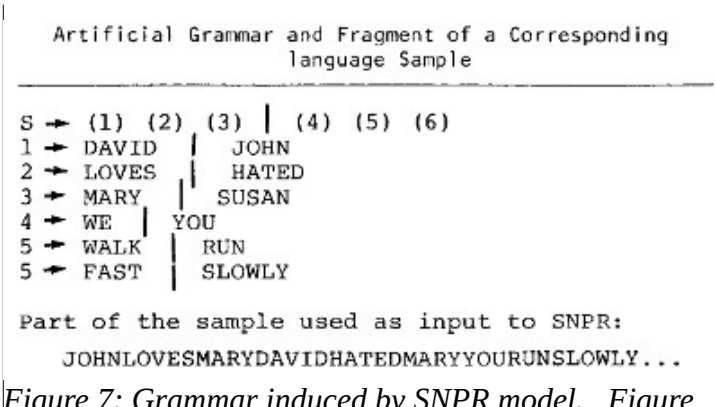


Figure 7: Grammar induced by SNPR model. Figure reproduced from (Wolff, 1988)

In Wolff's theory optimalization is further understood as compression. Within the SNPR model is such compression realized in part 3.5 of his algorithm, where the most frequent pair of contiguous elements

(either terminals or non-terminals) is substituted for a new non-terminal symbol. For this reason, the size of grammar able to generate the initial language sample ideally decreases with every cycle of model's « while » loop until the process converges to state where there is no redundancy to « compress ».

Wolff proposes that Grammar Induction is a process which should maximize the coding capacity (CC) of the resulting grammar while minimizing its size[14]. He defines the ratio between grammar's CC/MDL to denote grammar's efficiency and it may be the case that within a more evolutionary framework where one would work with populations of grammars, a very similarly defined notion of efficiency could be used as the core component of the fitness function. Unfortunately, Wolff's 1988 SNPR model is not evolutionary since it does not involve any stochastic factors nor notion of multiple candidate solutions. Wolff's SNPR is simply confronted with the language sample, deterministically compresses redundancies in a way that can sometimes ressembles human grammar (and sometimes not), gets subsequently stuck in local optimum and there's no way how to get out of it.

Another famous model of GI is that of (Elman 1993). Contrary to Wolff's algorithm which is principially « symbolic », is Elman's model « connectionist » one. More concretely, Elman had succeeded to train a simple recurrent neural network which was «*trained to take one word at a time and predict what the next word would be. Because the predictions depend on the grammatical structure (which may involve multiple embeddings), the prediction task forces the network to develop internal representations which encode the relevant grammatical information.* » (Elman 1993).

The most important finding of Elman's study seems to be the evidence for a so-called « less is more hypothesis » (Newport 1990) which Elman himselfs labels with terms « importance of starting small » : « *Put simply, the network was unable to learn the complex grammar when trained from the outset with the full "adult" language. However, when the training data were selected such that simple sentences were presented first, the network succeeded not only in mastering these, but then going on to master the complex sentences as well.* » (Elman 1993). Something similar occured also when he tuned the capacity of « internal memory » of his networks rather than the corpus itself. Elman observed: « *If the learning mechanism itself was allowed to undergo "maturational changes" (in this case, increasing its memory capacity) during learning, then outcome was just as good as if the environment itself had been gradually complicated.* »

Thus, not only results of Elman's computational model point in the same direction as many developmental and psycholinguistic studies of « motherese » (c.f. citations from Harris in part 2 of this work) ; they also show the importance of gradual physiological changes for ultimate mastering of maternal language. He goes even so far to state that prolonged infancy of human children can possibly go hand in hand with the fact that only humans develop language in an extent we do : «*In isolation,*

14 In current research, it is more common to speak about grammar's Minimal Description Length (MDL).

we see that both learning and prolonged development have characteristics which appear to be undesirable. Working together, they result in a combination which is highly adaptive» (Elman 1993).

Notwithstanding these interesting results which are not to be underestimated, we see two disadvantages of Elman's approach. Primo, as is often the case for connectionist neural networks, his resulting model is somewhat difficult to interpret : given the training constraints mentioned above, the network seems to predict quite well the next word in the phrase, but it is not evident why it does what it does. Elman himself dedicates major part of his article to descriptions of his tentatives to understand how his « blackbox » functions. Secundo, Elman confronted his model only with artificial corpora, i.e. corpora generated from manually created grammars. Thus, his model accounts only for a limited subset of properties of one language (English) and as such is still quite far from full-fledged solution to problem natural language's GI.

Last model we present in this brief overview, called « Automatic Distillation of Structure » (ADIOS) seem to be in lesser extent touched by this second disadvantage since as its authors state : « *In grammar induction from large-scale raw corpora, our method achieves precision and recall performance unrivaled by any other unsupervised algorithm. It exhibits good performance in grammaticality judgment tests (including standard tests routinely taken by students of English as a second language) and replicates the behavior of human subjects in certain psycholinguistic tests of artificial language acquisition. Finally, the very same algorithmic approach also is proving effective in other settings where knowledge discovery from sequential data is called for, such as bioinformatics.* » (Solan et al. 2005).

ADIOS is a graph-based model. It considers the sentences to be a path in the directed pseudograph (i.e. loops and multiple edges are allowed), each sentence being delimited by special « begin » and « end » vertices. Every lexical entry (i.e. a word type) is also a vertex of the graph, thus if more than two sentences share the same word X, they cross themselves in the vertex V_X ; if they contain the same subsequence XY, their paths share the common subpath (edge) $V_X V_Y$ etc.

Authors of ADIOS describe their algorithm as follows : « *The algorithm generates candidate patterns by traversing in each iteration a different search path (initially coinciding with one of the original corpus sentences), seeking subpaths that are shared by a significant number of partially aligned paths. The significant patterns (P) are selected according to a context-sensitive probabilistic criterion defined in terms of local flow quantities in the graph...Generalizing the search path, the algorithm looks for an optional equivalence class (E) of units that are interchangeable in the given context [i.e., are in complementary distribution]. At the end of each iteration, the most significant pattern is added to the lexicon as a new unit, the subpaths it subsumes are merged into a new vertex, and the graph is rewired accordingly... The search for patterns and equivalence classes and their incorporation into the graph are repeated until no new significant patterns are found.* » (Solan et al. 2005).

In other terms, ADIOS starts with a so-called Motif Extraction (MEX) procedure which looks for bundles of graph's subpaths which obey certain conditions. Once such « patterns » are found, they are subsequently « substituted » for non-terminal symbols and a graph is « rewired » to incorporate such newly constructed non-terminals. Such a « pattern distillation » procedure of generalization bootstraps itself until no further rewiring is possible. Output of the whole process is a rule grammar combining patterns (P) and their equivalence classes (E) into rules, able to generate even phrases which weren't present in the initial corpus. Example of how ADIOS progressively discovers more and more abstract combinatorial patterns is presented on Figure 8.

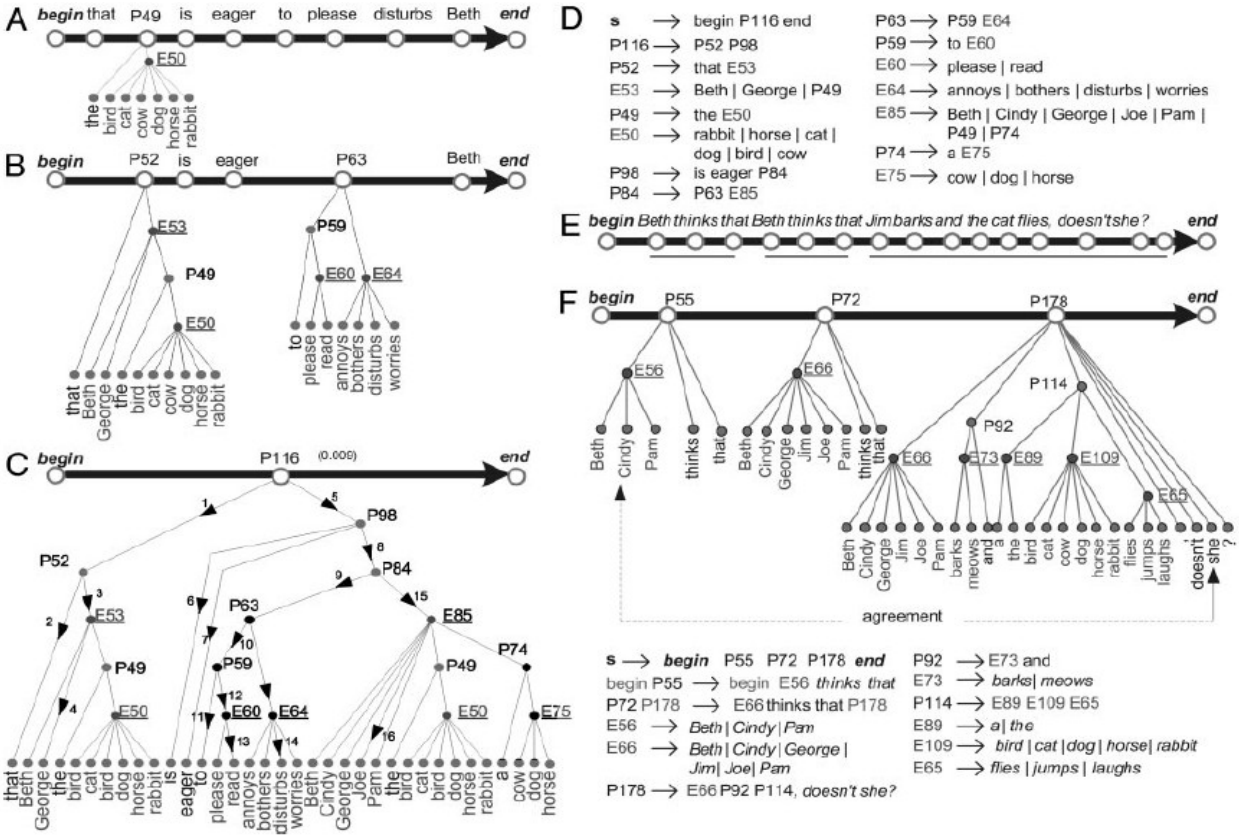


Figure 8: Equivalence classes and production rules induced from English language samples by ADIOS algorithm. Figure reproduced from (Solan et al. 2005)

ADIOS is undoubtably one of the most performant GI systems which currently exist. It combines both statistic, probabilistic and graph-theory notions with notion of rule-based grammar and as such is also of great theoretical interest. On the other hand, ADIOS does not involve any source of stochasticity, seems to be purely deterministic and as such incapable to deal with highly probable convergence towards locally optimal grammars. In confrontation with some partial corpora this may possibly not cause any problems but, we predict, without any stochastic variation whatsoever, ADIOS could not account for more than few « advanced » & real-life properties of natural languages and as such shall possibly share the destiny of SNPR model.

3.3.2. Evolutionary models of grammar induction

Multiple authors have proposed to solve the GI problem with different variants of evolutionary computinng - in following paragraphs we shall describe five different approaches:

1) Tomita's (1982) hill-climbing induction of finite state automata
2) Dupont's (1994) GIG method for inference of regular languages
3) Evolution of stochastic Context-Free Grammars as presented by Keller & Lutz (Keller and Lutz 1997)
4) Evolutionary method of (Aycinena et al. 2003) inducing grammars from POS tags of nine different English language corpora
5) Genetic algorithm of Smith & Witten (Smith and Witten 1995) for inducing a LISP s-expression grammar from a simple corpus of English sentences

Tomita's 1982 paper can be considered to be one of the first empiric studies of grammatical inference. The study focused on inference of grammars of 14 different regular languages – which are often called « Tomita languages » in subsequent litterature – by means of deteministic finite state automata. Tomita had first encoded any possible finite state machine with n states in a following manner :

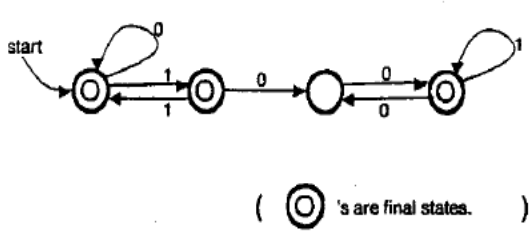


Figure 9: Finite state automaton matching all strings over (1 + 0) without an odd number of consecutive 0's after an odd number of consecutive 1's. Figure reproduced from (Tomita 1982)*

$$((A_1, B_1, F_1) (A_2 , B_2 , F_2) . . . . (A_n , B_n , F_n))$$

whereby every block « *(A_i, B_i, F_i) corresponds to the state i, and A_i and B_i indicate the destination states of the 0-arrow and the 1-arrow from the state i, respectively. If A or B is zero, then there is no 0-arrow or 1-arrow from the state i, respectively. F_i indicates whether state i is one of the final states or not. If F_i is equal to 1, the state i is one of the final states. The initial state is always state 1* » (Tomita, 1982).

Thus, for example, the string ((1 2 1) (3 1 1) (4 0 0) (3 4 1)) encodes the finite state automaton illustrated on figure 9.

Such encoding allowed Tomita to subsequently apply his hill-climbing approach. Hill-climbing can be considered to be a precursor to more extended genetic programming, since it employs both random mutations to explore surounding search-space and sort of selection algorithm which always prefers to use, in following iteration of the algorithm, such individual solutions for which the value of evaluation function E increases. Tomita's definition of E is very simple:

$$\mathbf{E = r - w}$$

« where r is the number of strings in the right-list accepted by the machine, and w is the number of strings in the wrong-list accepted by the machine». Right-list is a positive sample corpus while wrong-list is the negative sample. Thus, if a random mutation transforms an individual X_n into individual X_{n+1} so that $E(X_{n+1}) > E(X_n)$, - i.e. if an automaton is discovered which matches more positive sequences, or less negative sequences, or both - it will be X_{n+1} which will be mutated in the next cycle of the algorithm.

Tomita's approach cannot be considered to be fully evolutionary because he haven't used populations nor did he employed any kind of cross-over operator. For this reason, Tomita's regular grammar-infering algorithm did sometimes got stuck in local maxima from which there was no way out. Notwithstanding this small imperfection – of which Tomita himself was well aware – his work served, and still serves, the role of an important hallmark on the path to full-fledged GI.

Dupont (1994), for example, has also focused his study on induction of 15 different regular Tomita languages. In his formally very sound work, he defines the problem of inference of regular languages as a problem of finding of optimal partition of a state space of a finite « maximal canonical automaton » (MCA) able to accept the sentences from positive sample. Fitness function takes into account also the system's tendency to reject the sentences contained in the negative sample. By using a so-called « left-to-right canonical group encoding », Dupont succeeds to represent diverse individuals automata in a very concise way which allows him to subsequently evolve them by means of structural mutation («*the structural mutation consists of a random selection of a state in some block of a given partition followed by the random assignment of this state to a block* », e.g. MUTATE({{1,3,5},{2},{4}}) → {{1,5},{2,3},{4}}) and structural crossover («*the structural crossover consists of the union in both parent partitions of a randomly selected block* », for example CROSS({**{1,4}**,{2,3,5}},{**{1,3}**,{2},{4},{5}}) → {**{1,3,4}**,{2,5}},**{1,3,4}**,{2},5}).

Because « *the search space size dramatically increases with the size of the positive sample, making the correct identification more difficult when we have a larger positive information on the language* », Dupont has also proposed an incremental procedure allowing to start the search process from smaller yet pertinent region of the search space. Procedure goes as follows : « *first sort the positive sample I_+ in lexicographical order. Consequently,* ***the shortest strings are first taken into account****. Starting with the first sentence of I_+, we construct the associated MCA(I_+) and we search for the optimal partition of its state set under the control of the whole negative sample I_. Let A_1 denote the derived automaton with respect to this optimal partition. Let s_{next} denote the next string in I_+. If s_{next} is already accepted by A_1, we skip it.* » (Dupont 1994). Otherwise, the aumaton A_1 is be extended so that it can cover also s_{next}. The search under the control of whole negative sample is then restarted and whole process is repeated until all sentences from positive sample have been considered.

With population size of 100 individuals, maximum number of 2000 evaluations, crossover rate 0.2, mutation rate/bit 0.01 and semi incremental procedure implemented, Dupont's approach have attained, in average, classification rate of 94.4%. For five among fifteen Tomita's languages, grammars were constructed which attained 100% accuracy (i.e. accepted all sentences from positive sample and rejected all strings from negatives sample). Results have also indicated that if ever the semi-incremental procedure is applied, the sample size has positive influence upon the accuracy of infered grammars – bigger sample yields more accurate grammars.

While Tomita's results indicate and Dupont's results further confirm the belief that induction of grammars by means of evolutionary computing is a plausible thing to do, they do so only in regards to most similar type of grammars – the regular ones. Grammars of natural languages, however, are definitely not regular languages and models of GI of more expressive « context free » (CFG) or « context sensitive » grammars are needed.

Keller and Lutz employed a genetic algorithm to evolve parameters of stochastic context-free grammars (SCFG) of 6 different languages. SCFGs are similar to traditional CFGs[15], but extended with probability distribution, so that there is a probability value in the range [0,1] associated to every production rule of the grammar. These values are called SCFG's parameters and these are the values which the algorithm of Keller & Lutz aims to optimize by means of GAs. Their approach involves following steps :

> «
>
> 1. *Construct a covering grammar that generates the corpus as a (proper) subset.*
> 2. *Set up a population of individuals encoding parameter settings for the rules of the covering grammar.*
> 3. *Repeatedly apply genetic operations (cross-over, mutation) to selected individuals in the population until an optimal set of parameters is found.*
>
> *» (Keller and Lutz 1997)*

Their fitness function F(G) is based on idea of Minimal Description Length (MDL). More formally, Keller & Lutz aimed to maximize:

$$F(G)=\frac{K_c}{L(C|G)+L(G)}$$

by minimizing the denominator which is defined as a sum of number of bits needed to encode the grammar G (L(G)) plus the number of bits needed to encode corpus G, given the grammar G (L(C|G)). Numerator K_c is just a corpus dependent normalization factor assuring that the value of fitness shall be in range [0,1]. When

15 « In formal language theory, a context-free grammar (CFG) is a grammar inn which every production rule is of the form V → w, where V is a single non-terminal symbol, and w is a string of terminals annd/or non-terminals. The term « context-free » expresses the fact that non-terminals can be rewritten without regard to the context in which they occur » (Choubey and Kharat 2009)

confronted with positive samples of cca 16000 strings (typically of length 6 or 8) of 6 different context-free languages :

1. *EQ : language of all strings consisting of equal numbers of as and bs*
2. *language* $a^n b^n (n \geq 1)$
3. *BRA1 : language of balanced brackets*
4. *BRA2 : balanced brackets with two sorts of bracketing symbols*
5. *PAL1 : palindromes over {a,b}*
6. *PAL2 : palindromes over {a,b,c}*

their algorithms have converged, in majority of cases, to such combinations of parameters of their SCFGs which had allowed them to accept more than 95% of strings presented in the positive sample. Such results indicate that genetic algorithms can be used as a means for unsupervised inference of parameters of stochastic context-free grammars. Note that Keller & Lutz confronted, during both testing and training, their algorithm only with positive sample. While doing so for training is justifiable - since the objective of their study was to study whether grammars can be infered solely from positive evidence – not doing so during testing phase makes uncertain the extent to which their infered grammars overgeneralize.

Another huge disadvantage in regards to aims of our Thesis is the simple fact that their approach also seems to be very costly (« *number of parses that must be considered increases exponentially with the number of non-terminals* »). And since they confronted their algorithms only with corpora composed of sentences of artificial and not natural languages, we shall not try to imitate their approach of « tuning SCFG parameters » in our Thesis.

By being context-free and not simply regular, the grammars studied by Keller & Lutz or (Choubey and Kharat 2009) could be considered to be more similar to grammars of natural languages. Nonetheless, languages composed of palindromes and sequences of balanced brackets are still far way off from natural languages and the question « in what extent are results concerning GI of artificial languages applicable to GI of natural languages ? » is far from being answered. Rather than trying to answer it, we proceed now to discussion of two approaches where evolutionary GIs have been applied upon natural language sentences :

The first method, proposed in (Aycinena et al. 2003) has focused on induction of CFG grammars from nine different part-of-speech tagged natural language corpora. Sentences contained in these corpora, composed thus of sequences of part-of-speech tags (c.f. Section 3.2) were used as positive examples, while randomly generated sequences of POS-tags have yielded negative examples.

Initial population was composed of linear encodings of randomly generated context-free grammars, for example the string SABABCBCDCAE would represent this CFG :

S → AB
A → BC
B → CD
C → AE

During the evaluation of individual grammar G, one would first try to parse both positive and negative corpora with the grammar G and subsequently calculate the final fitness by applying the following formula :

$$F(\alpha)=\gamma^{max(0,|\alpha|-¿P|)}C(\alpha)-\delta I(\alpha)$$

« *where P is the set of preterminals, C(α) is the number of parsed sentences from the corpus, I(α) is the number of sentences parsed from the randomly generated corpus,* δ *is the penalty associated with parsing each sentence in the randomly generated corpus, and* γ *is the discount factor used for discouraging long grammars » (Aycinena et al. 2003)*

In their study, Aycinena had placed randomly generated population of 100 individual grammars on a two-dimensional 10 x 10 torus grid. Subsequently, they had applied a following select-breed-replace strategy :
«
1. Select and individual randomly from the grid
2. Breed that individual with its most fit neighbor to produce two children
3. Replace the weakest parent by the fittest child » (Aycinena et al. 2003)

In their framework, «*cross-over is accomplished by selecting a random production in each parent. Then a random point in these productions is selected and cross-over is performed, swapping the remainder of the strings after the cross-over points*». Every symbol of a resulting string can be subsequently mutated (mutation rate=0.01). «*A mutation is simply the swapping of a non-terminal or pre-terminal with another non-terminal or pre-terminal* » (Aycinena et al. 2003)

Figure 10 shows the number of generations each run was able to complete, the grammar G that last evolved, the percentage of positive examples parsed by G, the percentage of negative examples parsed by G and G's fitness.

While results displayed above may seem encouraging authors, have noticed that in majority of cases, their approach « *gives a grammar that is very capable of detecting whether a sentence is valid in English, but it has not learned much English structure* ». In other terms, Aycinena et al. have succeeded to breed grammars which have certain discriminatory power but are practically useless as models of English language. They go even so far as to state, in the ultimate paragraph of their work that « *It is still possible that English grammar is too complex to be learned from a corpus of words* » and that other external clues are necessary for successful GI of English.

The big disadvantage of above-mentioned algorithm was also the fact that its input were sequences of already attributed POS-tags and not sequences of words themselves. Thus, even if the approach would discover some interesting grammars, a reproach could be made and justified that in fact it only re-discovered the rules of the tagging system which was used in the first place. From perspective of our Thesis, another disadvantage of Aycinena et al.'s approach is related to the fact that their approach is anything but model of grammar development in human child. For it is evident (c.f. Section 2) that children learn the grammar of their language in an incremental fashion – they are not confronted with whole corpus from the very beginning. Nor does the corpus stay identic after each iteration of the learning process. On the contrary : as child grows, its linguistic environment - the corpus – also grows. Both in length and complexity.

Corpus	Number of generations completed	grammar	% positive examples parsed	% negative examples parsed	fitness
aliceinwonderland	200000	0V00R000J0VN0RJ0PN0PJ00 N00V0VJ0P00NN0N00TN0J00 T00TJ	92.50%	8.40%	657.364
brown1_a	48500	0N442N0TN0J420V0P03RN0N 040N0NV0R27T00T400T2V00 0P40V00V00R00J0V400N0JN	94.10%	6.10%	1667.85
brown1_b	200000	0JJ0NJ0R00N00T000N0J00V0 0VN0P00TN0TJ0PN0RJ0PJ0V J	94.70%	6.70%	1227.17
brown1_c	15500	0447R47TT1VJ22P40J5P72JN 40T71T4V42O140V1044PT03P 70600N0R77JN6JT4P11T6575 6N24JP70N07J4P500R4NN2V V52T4PN34N61504J0NV77J0 N44056N24JN0R622N5574RP 0NT4T004P0NP3V01154RO	80.50%	4.70%	302.996
brown1_d	45000	3V000J5TV00V0R230N0NV0T N1T33NN00T3TV00N0T33N00 J300P03020R20P00R30T0P30 NP	88.20%	5.60%	583.596
brown1_e	122000	0PJ0P00V00N00TN0T00TJ0VJ 0JJ0RJ0J00R00PN00N0NJ	93.90%	5.90%	1762.67
children	200000	0PJ0V00VN0JJ0NJ00N0N00TJ 0VJ0J00R00RJ0P00TN0T0	91.80%	5.70%	677.211
tomsawyer	200000	0PJ0VN0NN00V0T00TN0TJ00 N00J0NJ0P00PN0VJ0V00R00 J00N00RJ	92.70%	8.60%	2292.89
wizardofoz	200000	O00OPJOVOOROOPOORJOVN OTN0OOO0JOTOOPNONVOVJ OTJONOONNOJN	89.50%	9.20%	920.852

Figure 10: Grammars evolved from nine different POS-tagged corpora. Figure reproduced from (Aycinena et al., 2003).

An interesting evolutionary approach of GI which both tries to create own

non-terminal categories and also takes such « incrementality » into account is presented in the work of (Smith and Witten 1995). In their scenario, candidate grammars are evolved after presentation of every new sentence. Grammars have form of LISP s-expressions whereby AND represets a concatenation of two symbols (i.e. a syntagmatic node) and OR represents a disjunction (i.e. a paradigmatic node). Whole process is started as follows : « *The GA proceeds from the creation of a random population of diverse grammars based on the first sample string. The vocabulary of the expression is added to an initially empty lexicon of terminal symbols, and these are combined with randomly chosen operators in a construction of a candidate grammar...If the candidate grammar can parse the first string, it is parsed into the initial population* ». Figure 11 displays two sample grammars for the sentence « the dog saw a cat ».

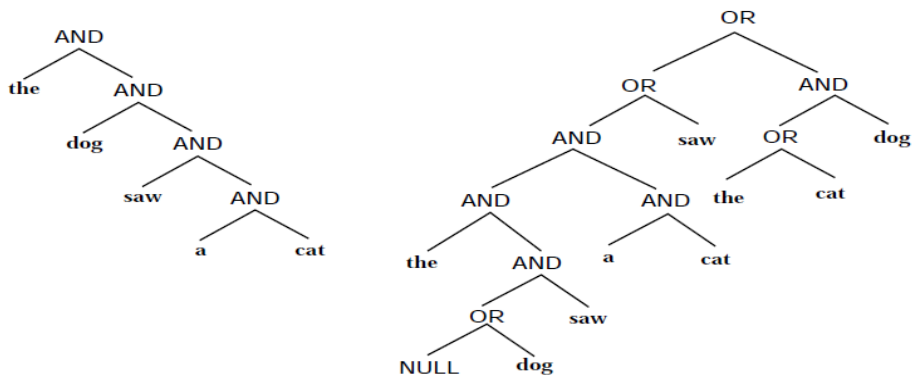


Figure 11: Two simple grammars covering the sentence "the dog saw a cat".
Figure reproduced from (Smith & Witten, 1995)

S-expression sequences representing individual grammars are subsequently mutated. Couple of parent grammars can also switch their nodes – probability of being chosen for such cross-over is inversely proportional to grammar's size : shorter grammars are prefered. Cross-over is non-destructive, parents thus also persist. The events of reproductions are grouped in cycles, at the end of each cycle, population of candidate grammars is confronted with new sentence from sample of positive evidence.

In their article (Smith and Witten 1995)show, how after presentation of sentences : «the dog saw a cat », « a dog saw a cat », « the dog bit a cat », « the cat saw a cat », « the dog saw a mouse » and « a cat chased the mouse » their system naturally converged to a grammar which had quite correctly subsumed determiners like « a », « the » under one group of OR nodes, verbs like « chased », « saw », « bit » under another, and nouns like « dog », « cat », « mouse » under yet another. The grammar which they finally obtain is not ideal but, as they argue, it could get better if confronted with new sentences. «*It is an adaptive process whereby the model is graudally conditioned by the training set. Recurring patterns help to reinforce partial inferences, but intermediate states of the model may include incorrect generalizations that can only be eradicated by continued evolution. This is not unlike the developing grammar of a child which includes mistakes and overgeneralisations that are slowly*

eliminated as their weaknenesses are made apparent by increasing positive evidence ». (Smith and Witten 1995)

While strongly agreeing with above citation, we nonetheless cannot ignore certain drawbacks of Smith & Witten's approach. Most importantly, by using LISP's s-expressions as a way of representing their grammars, they ultimately have to end up with highly bifurcated binary trees (since arity of AND|OR operators is 2). Thus, one can easily subordinate two non-terminals to one terminal (e.g. OR(cat,dog)), but in case of three subordinated terminals, one is obliged to use complex expression involving three non-terminals (e.g. OR(OR(cat,dog),OR(mouse,NULL)). Therefore, in such an s-expression based representation, is any class having more than two members neccessarily represented by a longer sequence → is more prone to mutation → is highly « handicapped » in regards to much shorter expressions subordinating just two nodes.

Another drawback of Smith & Witten's work which cannot be ignored is related to the fact that while they used English language sentences to train their system, the sentences were very simple and the relevance of their findings to GI of « natural » English is more than disputable. In fact, they seem to achieve, with quite complex evolutionary machinery, even less than Wolff's deterministic SNPR model have achieved almost a decade before. Notwithstanding these two drawbacks we nonetheless consider as particularly inspiring their approach aiming to solve the problem of GI of natural languages by uniting, in one framework, the notions adaptability, evolvability and statistical sensitivity to recurring patterns.

We summarize : all five above-mentioned approaches indicate that evolutionary computing can potentially yield useful solutions to the problem of Grammar Induction of both artificial (regular, context-free) and natural language grammars. The length of the candidate grammar is frequently used as an input argument of the fitness function. Note also that both solutions of Dupont and Smith & Witten also use a sort of « incremental » procedure whereby individual solutions gradually adapt to every new sentence. Especially Dupont's findings are reminiscent of what was already told about « importance of starting small » when discussing works of Elman & Harris.

On the other hand, none of the above mentioned models was confronted with corpus of child-directed (i.e. « motherese ») or child-originated utterances. The objective of our Thesis shall be to fill this gap.

3.4. Evolutionary Language Game

Evolutionary Language Game (ELG) first proposed in (Nowak et al. 1999) is a stunningly simple yet mathematically feasible stochastic model addressing the question « How could a coordinated system of meanings&sounds evolve in a group of mutually interacting agents ?».

In most simple terms, the model can be described as follows: Let's have a population of N agents. Each agent is described by an n x m associative matrix A. A's entry a_{ij} specifies how often an individual, in a role of a student, observed one or more other individuals (teachers) referring to object i by producing signal j. Thus, from this matrix A, one can derive the active « speaker » matrix P by normalizing rows :

$$p_{ij} = a_{ij} \Big/ \left(\sum_{l=1}^{m} a_{il} \right),$$

while the « hearer » passive matrix Q by normalization of A's columns:

$$q_{ji} = a_{ij} \Big/ \left(\sum_{l=1}^{n} a_{lj} \right).$$

The entries p_{ij} of the matrix P denote the probability that for an agent-speaker, object i is associated with sound j. The entries q_{ji} of the matrix Q denote the probability that for an agent-hearer, a sound j is associated with the object i.

Subsequently, we can imagine two individuals A and A', the first one having the language L (P, Q), the other having the language L' (P', Q'). The payoff related to communication of such two individuals is, within Nowak's model, calculated as follows:

$$F(A, A') = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} q'_{ji} = Tr(PQ')$$

And the fitness of the individual A in regards to all other members of the population can be obtained as follows :

$$f(A) = \frac{1}{|P| - 1} \sum_{\substack{A' \in P \\ (A' \neq A)}} F(A, A')$$

After the fitness values are obtained for all population members, one can easily apply traditional evolutionary computing methods in order to direct the population toward more optimal states, i.e. states where individual matrices are mutually « aligned ». In Nowak's framework this alignment represents the situation when hearer and speaker mutually understand each other, i.e. speaker has encoded meaning M by sound S and hearer had subsequently decoded sound S as meaning M.

ELG beautifully illustrates how such an alignment of sound-meaning matrices – a mutually shared communication protocol - can emerge practically ex nihilo given that there is some « mutual learning » procedure mechanism involved, which allows to transfer information from one individual to individual another. This is attained by creating a blank « student » matrix and then filling its elements, by means of stochastic « matrix sampling » procedure, in a way so that the resulting student matrix will partially correspond to| be aligned with matrices of pre-existing « teacher » (or teachers).

Further discussion and experiments with ELG are described (Kvasničkа and Pospíchal) and (Hromada 2012). All these studies point in the same direction and

suggest that not only emergence of mutually shared communication protocol practically *ex nihilo* is possible whenever there exists a means of transfer of information among individuals but also that presence of certain low amount of noise during the learning process is the only way how to make certain that the system will converge to « communicatively optimal » state.

The role of ELG model within the context of our Thesis is quite opened. For while it is the case that ELG sheds some light upon the question of emergence of language within a community of symbolicaly interacting agents, it does not, principially address the problem of language learning by a concrete individual. Thus, ELG is rather a model of macroscopic phylogeny than microscopic ontogeny - it addresses the problem of how small communities of homo habilis could, hundred years ago, gradually converge to system of signs within which, for example, « baubau » could mean a banana and « wauwau » mean a lion. But it does not address a problem of how today's human baby learns the complex language of her mother.

On the other hand, it is not completely *hors propos* to imagine a slight variation of Nowak's model wherein one population of matrices would be fixed (representing the linugistic competence of a teacher or mother organism) while the second population of matrices would represent the linguistic competence of a « child ». Given that the fitness function would somehow succeed to represent the degree of alignment between such « mother » and « child », we postulate that something like child's language competence could spontaneously emerge.

4. Remark concerning the Theory of Grammar Systems

A branch of Formal Language Theory which could be of particular use for pursposes of our Thesis is devoted to study of Grammar Systems (GS). A GS is a « *set of grammars working together, according to a specified protocol, to generate a language*» (Jiménez-López 2000). Thus, contrary to classical Formal Language Theory within which one grammar generate ones language, in GS several grammars work together in order to generate one language. Grammar Systems can be therefore considered as a sort of multi-agent variants of traditional « monolithic » formal grammar theory.

The very nature of multi-agent systems often implies cooperation, communication distribution, modularity parallelism, or even emergence of complexity. For example, Figure 12 illustrates a very simple bimodular « language colony » variant of a GS.

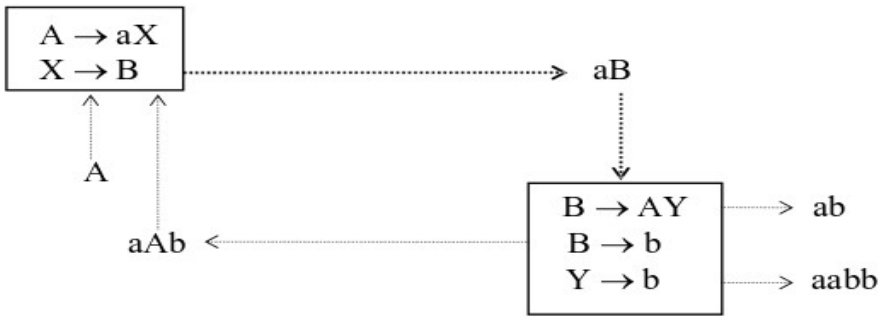


Figure 12: Language colony of two finite grammars cooperating to generate an infinite language. Figure reproduced from (Kelemen 2004)

By allowing the finite grammar components to communicate through a common symbolic environment[16], one ultimately generates a language which is infinite ! (Kelemen 2004) applies the term « miracle » to such behaviour, which is very common in the world of GS.

Since the Theory of Grammar Systems is formally very well developped - most notably thanks to life-long work of Erzsébet Csuhaj-Varju and substantial contributions by George Paun and Jozef Kelemen– it is impossible for us to introduce, within the limited scope of this text, the formalism of GS Theory in closer detail. This will be done in the final version of our Thesis, if ever we decide to poursuit our research in direction. If that will be the case, we will often refer to the doctoral Thesis of (Jiménez-López 2000) which contains many persuasive arguments for application of GS upon the study of natural human languages. On the other hand, the Thesis of Jimenez-Lopez is limited by the fact that it mostly proposes to use the Grammar System Theory as a framework explaining the final, i.e. « adult » linguistic component, and not as a framework which could elucidate the very process of language development and language acquistion[17].

The only tentative to use Grammar System apparatus for grammatical inference is that of (Sosík and Štỳbnar 1997). Contrary to other authors of GS who focus principially on the productive (i.e. generative) aspects of GS, Sosik & Štýbnar have focused on GS's language-accepting properties. In a hybrid connectionist-symbolic architecture, they have used a « neural pushdown automaton » to infer a language colony able to cover some simple artificial context-free grammars able to cover balanced parenthesis or palindrom languages.

As far as we know, no tentative is reported in the litterature to solve the problem of grammar induction of natural languages by means of evolutionary optimization of Grammar Systems.

5. Thesis

The Thesis hereby introduced is done under double supervision of dpt. Cybernetics at Slovak University of Technology (STU) and « cognitive psychology » laboratory affiliated to University Paris 8 (P8). Ideally, both « engineering » approach – common to STU – as well as more cognition-oriented « experimental » approach of P8, should be equally reflected in the final Thesis. In order to do so, the Thesis shall, in fact, introduce multiple «theses » among which some shall be addressing more « theoretical » psychology and linguistics related phenomena and problems.

But due to its affiliation to STU, the text shall also introduce more concrete,

16 A common symbolic environment which is shared by different modules plays the central role in practically all variants of Grammar Systems. It is reminiscent of the role which « short term memory » or «working memory » plays in cognitive psychology.

17 In terms of Grammar System Theory, it seems to be more appropriate to speak about «language emergence »

pragmatic and operational theses aiming to offer a computationally and formally sound affirmative answer to the question : « Can a language development be modelled as an evolutionary process ? »

5.1. Theoretical Thesis

At first, a child has to learn

- how to segment the world into groups of discrete objects and processes
- how to segment phonetic flux into sequences of discrete linguistic tokens

The subsequent problem of language development can be analyzed as a trinity of sub-problems:

1) vocabulary development (learning of mappings between objects and tokens)
2) induction of grammatical categories
3) induction of grammatical rules

These tasks are deeply and strongly intertwined. Without ability to segment world into objects there are no stable referents to which linguistic tokens could refer. Without ability to perceive recurrent tokens, there are no conventional symbols with which a child could denote specific objects. Without vocabulary development (which relates to induction of semantic classes which we have called « concept construction » in the text above), there is no need for grammatical rules nor categories. Without grammatical categories, grammatical rules are just a senseless tautological formal game and there is no way to distinguish useful grammars from useless ones. Without useful grammars, vocabulary development shall halt at some locally optimal level of a « pidgin » language.

Left on their own, these problems pose us in front of us a variant of a chicken & egg problem which seems almost impossible to tackle. Baby's brain, however, resolves these problems with such such an ellegance that one is tempted to say that they even do not exist.

Aim of the Thesis which shall follow is to demonstrate that if one interprets the above mentioned set of problems interpreted in terms of

- parent-child communication (imitation)
- partitioning of vector spaces (categorization)
- gradual accomodation and assimilation of knowledge (generalization)

one could subsequently state that the key theoretical Thesis we aim to defend is

T_t process of language development is an auto-organizing and potentially evolutionary process

Note the word « potentially », because in order to be labeled as « evolutionary », following conjectures have to be validated :

C_1) Not only imitation but also repetition are forms of replication : Information

replicates not only between the brains but also in the brain.

C_2) Fitness of a linguistic structure is related to its ability to represent certain recurrent aspect of agent's environment : If cognitive structure matches some aspect of environment, it gets activated. By being activated, it augments it probability of being (at least partially) replicated.

C_3) Problems of both generalization and overgeneralization are to be solved by variation|decay operators endogenously transforming the information represented in the memory of a language-inducing system.

Acceptation of above-mentioned conjectures lead us to model of language development based not on tuning of parameters of single monolithic grammar, but rather based on a population of « microgrammars », a « language colony » (Kelemen and Kelemenová 1992) of mutually communicating, co-operating, decaying and replicating sequences of production rules unceasingly trying to match the language of linguistic environment.

We postulate that if such an environment has certain properties of « motherese », a linguistic competence : an ability to generate utterances in still more & more complex « toddlerese », shall spontaneously emerge.

Thus, three notions will be of utmost importance in the Thesis which we hereby introduce : « motherese », « microgrammar » and « matching ». The corpus of « motherese », more concretely the CHILDes corpus (MacWhinney 2000) , will be considered to be sufficiently adequate image of initial stages of child's linguistic environment. Development of child's linguistic competence will be explained in terms of gradual evolution of individual « microgrammars », i.e. chromosomes whose genomes can be understood as individual production rules. At last but not least, the notion of « matching » shall furnish us the first principle which could potentially allow us to to explain the mystery of language acquisition as an evolutionary process:

P_1 «If (internal) rule R or substitional schema S succeeds to match some aspect of (external) environment, then it shall be replicated into another microgrammar»

5.2. Operational Thesis

The operational Thesis (T_O) is stated as follows :

T_O « There exists an evolutionary algorithm A which, when confronted with the corpus of motherese language (L_M) as an input, can produce the toddlerese grammar (G_T) able to generate the L_M-ressembling toddlerese language L_T »

The term « evolutionary » means that the algorithm A shall involve incremental replication, mutation and selection of information-representing structures. More concretely, these information-representing structures, i.e. g*enomes, shall be ordered sequences of genes, whereby each gene shall contain an individual substitution rule.* Thus, every individual genome shall represent a « microgrammar » aiming to transform linguistic token (i.e. sequence of terminals) currently observable in the environment, into sequence of non-terminals. Whenever such « successful parse »

shall occur, the principle P_1 shall apply and useful genes shall be reproduced into other individual micro-grammars. This could potentially cause the micro-grammars to gradually adapt their structures to those of environment.

On the other hand, in order to prevent excessive adaptation, a variation operator shall be also integrated in the algorithm A, aiming to vaguely modeling a well-known phenomenon of « forgetting ».

5.3. The organization of the Thesis

The Thesis shall be composed of five parts each of which is composed of multiple major chapters. Every chapter consists of introduction and conclusion preceding resp. following more specific subchapters which can fractally branch into sub-chapters , sub-sub-chapters etc. All such parts, chapters, sub-chapters etc. can be considered to be « non-terminal » nodes of structure presented by this text.

The first part, labeled Theses, is a stem of whole text. It will introduce multiple theses at varying degrees of generality which shall be all - in one way or another - more directly addressed in subsequent sections. In order to weave the basic conceptual fabric, some definitions of terms like « evolution » and « language learning » shall be also offered along the path delimited in Section 1. All variants of the thesis shall be briefly related to other cognitive sciences.

The second branch, labeled « Theoretical position » is composed of chapters dedicated to Universal Darwinism, Developmental Psycholinguistics and Natural Language Processing. In these chapters, the theses presented in the first chapter shall be more deeply interpreted and contextualized in terms of respective disciplines.

The third branch, labeled «Observations» will describe multiple longtitudinal observations of one concrete human child. In certain cases, the generalizability of such individual observations shall be verified or falsified by means of text-mining the CHILDes corpora. Subsequent interpretations in terms of the evolutionary theoretical framework shall follow.

The penultimate branch, called «Simulations» shall present multiple computational models addressing four problems related to language acquisition process : 1) The problem of segmentation 2) The problem of induction of grammatical categories 3) The problem of induction of grammatical rules 4) The problem of concept induction. Specific chapter will be dedicated to every problem in which existing solutions shall be described. Special focus shall be put on evolutionary solutions, if they exist. To every of four above-mentioned problems we shall try to offer our own unique evolutionary solution and subsequently we shall discuss its performance. PERL source codes of diverse versions of the algorithm A shall be also attached in order to allow reproducibility of our results by other scientists.

The conclusive branch labeled « Synthesis » shall primarily discuss results obtained in parts « Observations » and « Simulations ». If the results turn out to be consistent with theory, the work shall end with a tentative to integerate theses T_t and T_t in one unified framework. If unsuccessful, potential reasons of the failure shall be analysed.

6. Bibliography

Araujo, Lourdes. 2002. Part-of-speech tagging with evolutionary algorithms. In *Computational Linguistics and Intelligent Text Processing*, 230–239. Heidelberg, Germany: Springer.

Aycinena, Margaret, Mykel J. Kochenderfer, and David Carl Mulford. 2003. An evolutionary approach to natural language grammar induction. *Final Paper Stanford CS224N June*.

Barrett, Deirdre. 2007. *Waistland: A (R) evolutionary View of Our Weight and Fitness Crisis*. New York, NY : WW Norton & Company.

Bee, Helen L., and Denise Roberts Boyd. 2003. *The developing child*. Boston, MA : Allyn & Bacon.

Bentley, Peter. 1999. *Evolutionary design by computers*. San Francisco, CA : Morgan Kaufmann.

Berman, Ruth A. 1988. Word class distinctions in developing grammars. *Categories and processes in language acquisition*: 45–72.

Blackmore, Susan. 2000. *The meme machine*. Oxford, England : Oxford University Press.

Braine, Martin DS. 1971. On two types of models of the internalization of grammars. *The ontogenesis of grammar*: 153–186.

Brodsky, Peter, H. R. Waterfall, and Shimon Edelman. 2007. Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the 29th Cognitive Science Society Conference, ed. DS McNamara & JG Trafton*, 833–38.

Brown, Roger. 1973. *A first language: The early stages.* Cambridge, MA : Harvard University Press.

Campbell, Donald T. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review* 67: 380.

Choubey, Nitin S., and Madan U. Kharat. 2009. Grammar Induction and Genetic Algorithms-An Overview. *Pacific Journal of Science and Technology* 10: 884–888.

Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman. 2010. Two Decades of Unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 575-584).

Cohen, Trevor, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics* 43: 240–256.

Cosmides, Leda, and John Tooby. 1997. Evolutionary psychology: A primer. *Retrieved from http://www.cep.ucsb.edu/primer.html*.

Csuhaj-Varjú, Erzsébet. 1994. *Grammar systems: a grammatical approach to distribution and cooperation*. Yverdon, Switzerland : Gordon and Breach Science Publishers.

Darwin, Charles. 1859. On the Origin of Species. London, England : John Murray.

Darwin, Charles. 1906. *The voyage of the Beagle*. 104. JM Dent & sons.

Dawkins, Richard. 2006. *The selfish gene*. Oxford, England : Oxford university press.
Dennett, Daniel C. 1996. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. 39. New York, NY : Simon & Schuster.
Dupont, Pierre. 1994. Regular grammatical inference from positive and negative samples by genetic search: the GIG method. In *Grammatical Inference and Applications*, 236–245. Heidelberg, Germany: Springer.
El Ghali, Adil, Daniel Hromada, and Kaoutar El Ghali. 2012. Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. *JEP-TALN-RECITAL 2012*: 77.
Elman, Jeffrey L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48: 71–99.
Flake, G. W. 1999. *The computational beauty of nature*. Cambridge, MA : MIT press.
Fogel, Lawrence J., Alvin J. Owens, and Michael J. Walsh. 1966. Artificial intelligence through simulated evolution. New York, NY : John Wiley & Sons.
Foster, Mary LeCron. 2002. Symbolism: the foundation of culture. *Companion encyclopedia of anthropology :* 366. Canada : Routledge.
Furrow, David, Katherine Nelson, and Helen Benedict. 1979. Mothers' speech to children and syntactic development: Some simple relationships. *Journal of child language* 6: 423–442.
Galton, Francis. 1875. *English men of science: Their nature and nurture*. .
Gärdenfors, Peter. 2004. *Conceptual spaces: The geometry of thought*. MIT press.
Haeckel, Ernst Heinrich Philipp August. 1879. *The evolution of man*. Vol. 1. [sn].
Haidt, Jonathan. 2012. *The righteous mind: Why good people are divided by politics and religion*. Random House LLC.
Hamilton, William D. 1963. The evolution of altruistic behavior. *The American Naturalist* 97: 354–356.
Harris, Margaret. 2013. *Language experience and early language development: From input to uptake*. Psychology Press.
Harris, Zellig S. 1954. Distributional structure. *Word*.
Hebb, Donald Olding. 1964. *The Organization of Behavior: A Neuropsychlogical Theory*. John Wiley & Sons.
Hoff-Ginsberg, Erika. 1986. Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology* 22: 155.
Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, MI : University of Michigan Press.
Hromada, Daniel Devatman. 2012. Variations upon the theme of Evolutionary Language Game. Unpublished manuscript. Slovak University of Technology.
Hromada, Daniel Devatman. 2013a. Geometrizácia ontológií - prípadová štúdia SNOMED. Unpublished manuscript. Slovak University of Technology.
Hromada, Daniel Devatman. 2013b. Random Projection and Geometrization of String Distance Metrics. In *Proceedings of the Student Research Workshop associated with RANLP*, 79–85. Hissar : Bulgaria.
Hromada, Daniel Devatman. 2014a. Comparative study concerning the role of surface morphological features in the induction of part-of-speech categories. In

Proceedings of TSD2014 conference. Heidelberg, Germany : Springer.

Hromada, Daniel Devatman. 2014b. Introductory experiments with evolutionary optimization of reflective semantic vector spaces. In *TALN-RECITAL-DEFT 2014*. Marseille.

Hromada, Daniel Devatman. 2014c. Conditions for cognitive plausibility of computational models of category induction. In *Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Heidelberg, Germany: Springer.

Jiménez-López, MD. 2000. Grammar systems: a formal-language-theoretic framework for linguistics and cultural evolution. PhD Dissertation. Tarragona, Spain : Rovira i Virgili University, .

Johnson, William B., and Joram Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* 26: 1.

Karypis, George. 2002. *CLUTO-a clustering toolkit*. DTIC Document.

Kauffman, Stuart. 1996. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford, England : Oxford University Press.

Kelemen, Jozef. 2004. Miracles, colonies, and emergence. In *Formal Languages and Applications*, 323–333. Heidelberg, Germany: Springer.

Kelemen, Jozef, and Alica Kelemenová. 1992. A grammar-theoretic treatment of multiagent systems. *Cybernetics and System* 23: 621–633.

Keller, Bill, and Rudi Lutz. 1997. Evolving stochastic context-free grammars from examples using a minimum description length principle. In *1997 Workshop on Automata Induction Grammatical Inference and Language Acquisition*.

Kennedy, James F., James Kennedy, and Russel C. Eberhart. 2001. *Swarm intelligence*. San Francisco, CA : Morgan Kaufmann.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1). Cambridge, MA : MIT press.

Küntay, Aylin, and Dan I. Slobin. 1996. Listening to a Turkish mother: Some puzzles for acquisition. *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*: 265–286.

Kvasnička, Vladimír, and Jirí Pospíchal. Evolúcia jazyka a univerzální darwinizmus. In *Myseľ, inteligencia a život.* Bratislava : Slovenská Technická Univerzita.

Lakoff, G. 1990. *Women, fire, and dangerous things*. Chicago, IL : University of Chicago Press.

Levy, Yonata. 1988. The nature of early language: Evidence from the development of Hebrew morphology. *Categories and processes in language acquisition*: 73–98. Lawrence Erlbaum Associates.

MacWhinney, Brian. 1987. The competition model. *Mechanisms of language acquisition*: 249–308.

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

Maratsos, Michael. 1988. The acquisition of formal word classes. *Categories and processes in language acquisition*: 31–44. Lawrence Erlbaum Associates.

Morgan, Thomas Hunt. 1916. *A Critique of the Theory of Evolution*. Princeton University Press.

Newport, Elissa L. 1990. Maturational constraints on language learning. *Cognitive science* 14: 11–28.

Ninio, Anat. 1988. On formal grammatical categories in early child language. *Categories and processes in language acquisition.* Lawrence Erlbaum Associates.

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

Ofria, Charles, and Claus O Wilke. 2004. Avida: A software platform for research in computational evolutionary biology. *Artificial life* 10: 191–229.

O'Neill, Michael, and Conor Ryan. 2003. *Grammatical evolution: evolutionary automatic programming in an arbitrary language*. Genetic Programming Series, Vol. 4. Heidelberg, Germany: Springer.

Piaget, Jean. 1974. Introduction à l'épistémologie génétique. *Paris, PUF*.

Pohlheim, Hartmut. 1996. GEATbx: Genetic and evolutionary algorithm toolbox for use with MATLAB documentation. Retrieved from *http://www. geatbx. com/docu/algindex. html*.

Poincaré, Henri. 1908. *L'invention mathématique*.

Popper, Karl Raimund, Karl Raimund Popper, and Karl Raimund Popper. 1972. *Objective knowledge: An evolutionary approach*. Oxford, England : Clarendon Press.

Ray, Thomas S. 1992. Evolution, ecology and optimization of digital organisms. *Santa Fe*.

Rechenberg, Ingo. 1973. Evolutionsstrategie–Optimierung technisher Systeme nach Prinzipien der biologischen Evolution. Stuttgart, Germany : Fromman-Holzboog.

Rizzolatti, Giacomo, and Laila Craighero. 2004. The Mirror-Neuron System. *Annual Review of Neuroscience* 27: 169–192.

Rosch, Eleanor. 1999. Principles of categorization. *Concepts: core readings*: 189–206. Cambridge, MA : MIT press.

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

Sekaj, I. 2005. *Evolučné výpočty a ich využitie v praxi*. Iris.

Shi, Rushen, Janet F Werker, and James L Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition* 72: B11–B21. doi:10.1016/S0010-0277(99)00047-5.

Simonton, Dean Keith. 1999. Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry* 10: 309–328.

Smith, Tony C., and Ian H. Witten. 1995. A genetic algorithm for the induction of natural language grammars. In *Proc. of IJCAI-95 Workshop on New Approaches to Learning for Natural Language Processing*, 17–24.

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

Sosík, Petr, and Leoš Štỳbnar. 1997. Grammatical inference of colonies. In *New*

Trends in Formal Languages, 236–246. Heidelberg, Germany: Springer.
Spencer, Herbert. 1894. *Education: Intellectual, moral, and physical*. CW Bardeen.
Tomita, Masaru. 1982. Dynamic construction of finite-state automata from examples using hill-climbing. In *Proceedings of the fourth annual cognitive science conference*, 105–108.
Trivers, R.L. (1972). Parental investment and sexual selection. In B. Campbell (Ed.), Sexual selection and the descent of man, 1871-1971 (pp. 136–179). Chicago, IL: Aldine.
Turing, A. M. 2008. Computing machinery and intelligence. *Parsing the Turing Test*: 23–65.
Vapnik, V., S. E Golowich, and A. Smola. 1997. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems 9*.
Wilson, Edward O. 1978. What is sociobiology? *Society* 15: 10–14.
Wittgenstein, L. 2009. *Philosophical investigations*. Wiley-Blackwell.
Wolff, J. Gerard. 1988. Learning syntax and meanings through optimization and distributional analysis. *Categories and processes in language acquisition* 1.
Wright, Sewall. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics*, 1:356–366.