Evolutionary modelisation of ontogeny of linguistic structures

Rigorous Thesis Examination

Daniel D. Hromada[1,2]

[1]Slovak University of Technology
Faculty of Electronic Engeneering and Informatics
Department of Robotics and Cybernetics

[2]Université Paris 8
École Doctorale Cognition, Langage, Interaction
Laboratoire Cognition Humaine et Artificielle

4-11-2014

doc. Ing. Ivan Sekaj, PhD.
prof. Ing. Vladimír Kvasničkа, DrSc.

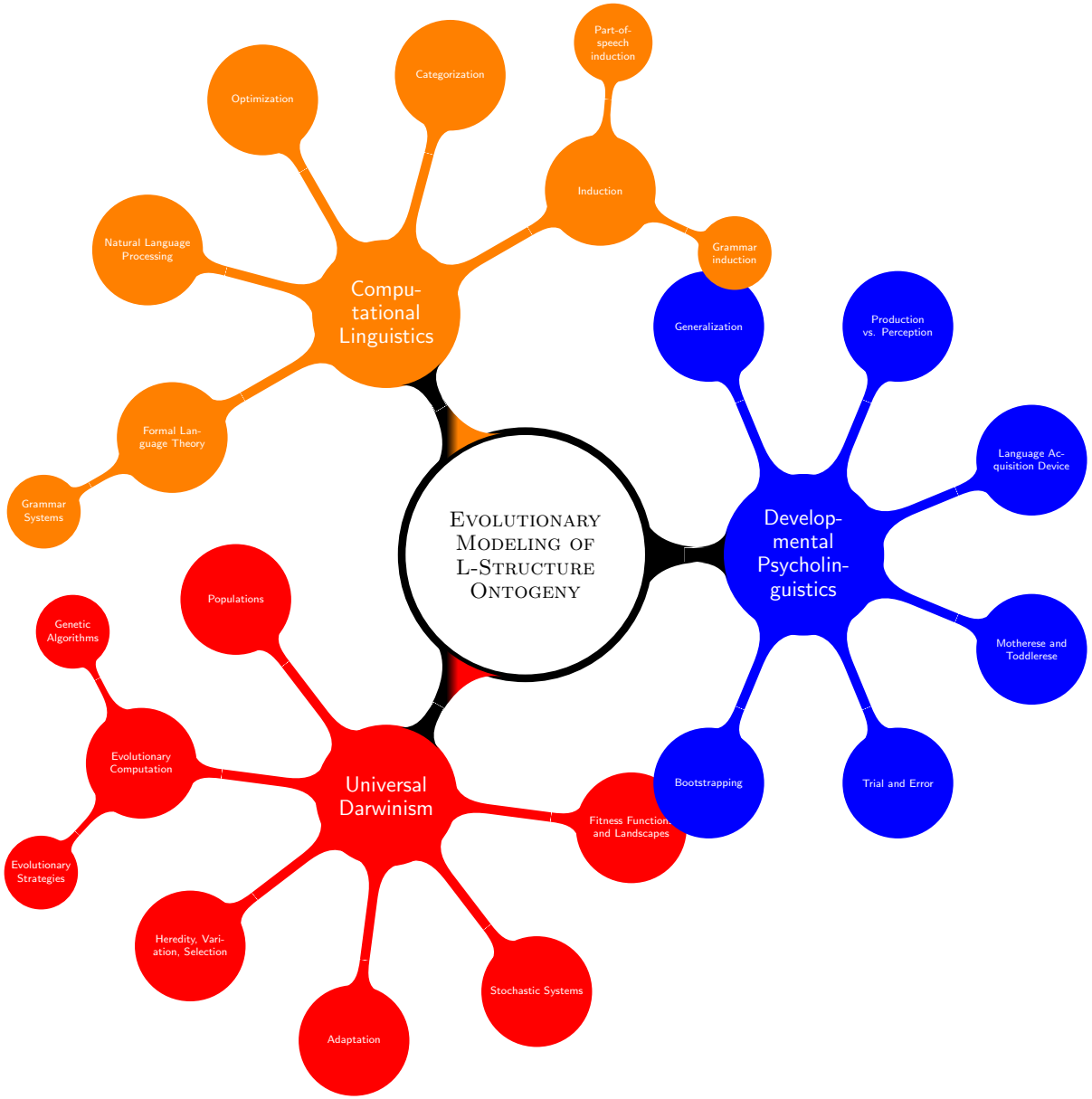
Administrative Position

Development

- 2010: enrolled for PhD. at Ecole Doctorale Cognition, Langage, Interaction of University Paris8
- 2011: attribution of double PhD. scholarship by french government, inscription at STU as external doctorant
- 2012: inter-university convention signed by FEI STU's Dean and President of Paris8, start of scholarship
- 2013: summer semester in Paris, presentation of prof. Tijus in Bratislava
- 2014: summer semester in Paris, end of scholarship

Thesis is to be written and defended in English language.

An interdisciplinary enterprise



Universal Darwinism

Definition

A general theoretical framework aiming to explain the emergence and optimization of diverse complex phenomena in terms of interaction of three basic processes:

1. information variation
2. information selection
3. information replication

UD-consistent disciplines

- biology (Darwin 1859, Mendel 1866) and genetics (Morgan 1916, Watson & Crick, 1953)
- sociobiology (Hamilton 1974, Wilson 1978) and evolutionary psychology (Cosmides & Tooby 1997)
- memetics (Dawkins 1976, Blackmore 2000)
- evolutionary epistemology
- neural darwinism
- evolutionary computation, artificial life, ...
- evolutionary linguistics

Evolutionary Epistemology

An ambiguous definition

Evolutionary Epistemology aims to explain source, existence, nature, scope and diversity of forms of knowledge in evolutionary terms.

Two possible intepretations:

1. biological evolution of cognitive and mental faculties in animals and humans
2. knowledge *per se* evolves by selection and variation

The second interpretation can be further analyzed:

1. knowledge can emerge by variation&selection of ideas shared by a group of mutually interacting individuals (Popper 1972)
2. knowledge can emerge by variation&selection of cognitive representations within one individual

Genetic Theories of Learning and Creativity

"Genetic" not in contemporary (i.e. DNA-related) sense but as related to "origins" (genesis) and "heredity" (genus).

Piaget's Genetic Epistemology

1. aims to explain how human cognitive systems (CS) develop from birth onwards
2. CS pass through series of stages, every stage involves equilibration of cognitive schemas
3. schemas change through process of assimilation and acommodation

Campbell-Simonton's Theory of Creativity

1. scientific discovery and creativity can be explained in terms of *blind variation and selective retention* (Campbell 1970)
2. *"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."* (Simonton 1990)

Neural Darwinism

Edelman (1987) postulated that complex adaptations in the brain arise through some process similar to natural selection.

Another variant of ND is theory of Changeux and Dehane (1989): "the production and storage of mental representations, including their chaining into meaningful propositions and the development of reasoning, can also be intepreted, by analogy, in variation-selection (Darwinian) terms within psychological time-scales."

Fernando et al. (2012) propose two "toy models .. of a means by which a higher-order unit of neuronal evolution *above* the synaptic level may be able to replicate."

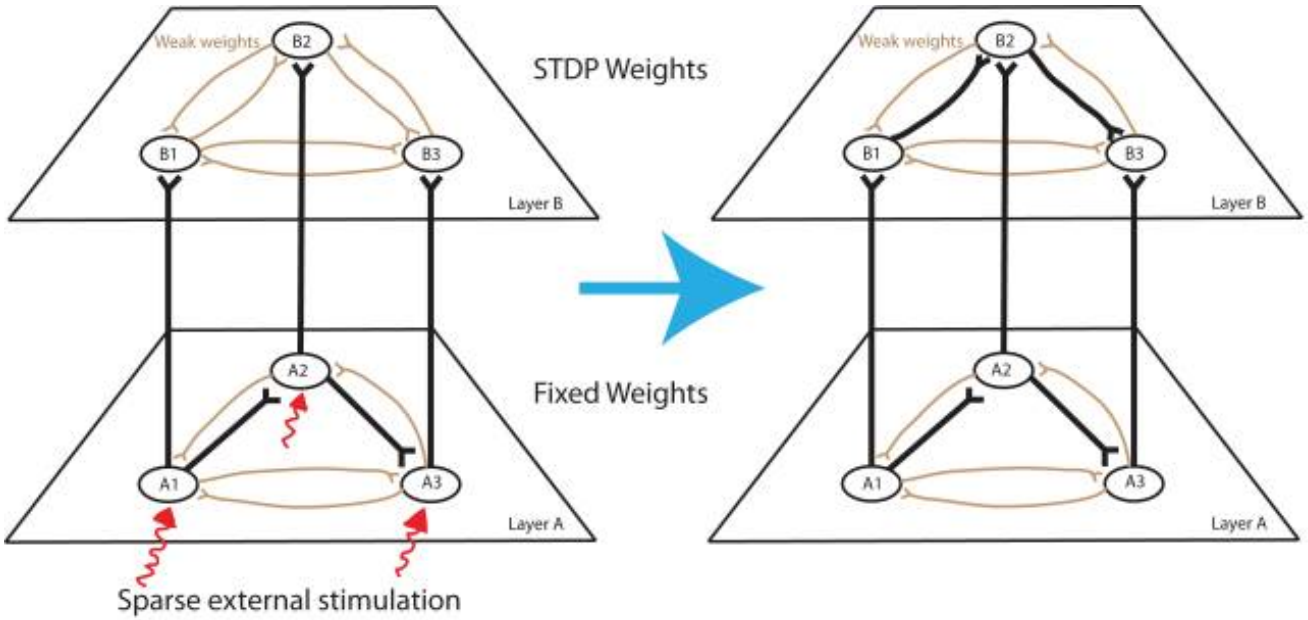


Figure: reproduced from (Fernando et al., 2012).

Evolutionary Computation

Definition

"Evolutionary computation uses computational models of evolutionary processes as key elements in the design and implementation of computer-based problem solving systems" (Spears et al., 1993)

- genetic algorithms (c.f. next slide)
- evolutionary programming (stronger genotype-phenotype distinction, FSAs, little recombination)
- evolutionary strategies (involves more recombination, self-adaptation, other nature-inspired approaches)
- genetic programming (does not search for solutions but for programs)

Grammatical evolution

Variant of Genetic Programming which uses evolutionary search to discover specific sequences of application of rules of production which generate program code which yields wished solutions.

- swarm intelligence (Kennedy & Eberhart, 2001)
- artificial life (no exogenous fitness function: Tierra, AVIDA, etc.)

Genetic algorithms

Canonic GA (Holland, 1975)

Encoding: binary vector
Initial population: randomly generated
Selection: fitness proportionate ($p_i = f_i / \sum_{j=1}^{N} f_j$)
Crossover: one-point
Mutation: bit-flip with probability p (0.001)

```
rand init
evaluate
select
repeat
        crossover
        mutation
        evaluate
        select
until stop
```

Schema theorem

A schema is a subset of strings with similarities at certain positions. Schema theorem states that short, low-order (i.e. with few fixed positions) schemata with above-average fitness increase exponentially in successive generations:

$$E(m(H, t+1)) \geq \frac{m(H, t) f(H)}{a_t} [1 - p$$

m(H,t) is the number of strings belonging to schema H at generation t, f(H) is the observed average fitness of schema H and a_t is the observed average fitness at generation t. P is the probability that crossover or mutation will disrupt H.

Convergence to global optimum

Rudolph (1994) has proven that CGAs are certain to converge to global optimum only if they "keep track of the best solution found over time" (i.e. involve a form of *elitism*).

Evolutionary Language Game (Nowak et al., 1999)

Let's have a population of N agents. Each agent is described by an $r * c$ 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. From this associative matrix A, one can derive

- the active "speaker" matrix S by normalizing A's rows: $s_{ij} = \frac{a_{ij}}{\sum_{n=1}^{r} a_{in}}$
- the "hearer" passive matrix H by normalization of A's columns: $h_{ij} = \frac{a_{ij}}{\sum_{n=1}^{c} a_{nj}}$

Subsequently, we can imagine two individuals A and A', the first one having the language L (H,S), the other having the language L' (H', S'). The payoff related to communication of two individuals is calculated as follows:

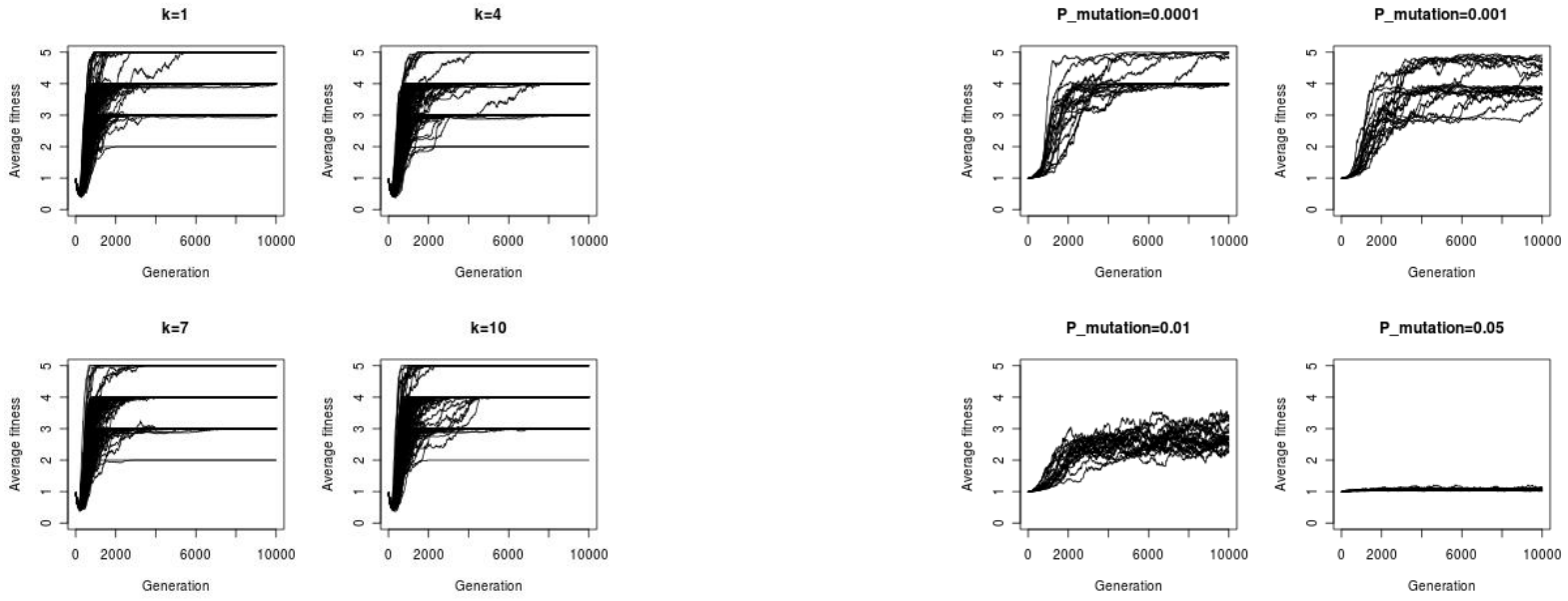
$$F(A, A') = \sum_{i=1}^{r} \sum_{j=1}^{c} s_{ij} h'_{ji} = Tr(SH')$$

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

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

By implementing EC, these fitness values can subsequently direct evolution of the population toward states where individual matrices are more optimally "aligned". In ELG, 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.

Evolutionary Language Game #2



- parent-child information transfer modelled by matrix sampling procedure
- parameter *k* specify the quantity of repetition during the matrix sampling
- all experiments with N=100 having memones of size 5x5 (i.e. their associative matrices could encode max 5 "sounds" and 5 "meanings")
- convergence to globally optimal state is assured only if MS involves small but nonzero amount of noise!
- beautiful model how "language" is sure to arise *ex nihilo* in communities wherein information transfer between individuals exists
- Nowak et al. (1999) and Kvasničkа & Pospíchal (2007) use it to iluminate emergence of language in phylogeny of homo sapiens sapiens species, but couldn't be an analogical approach used to model transfer from mother to child in ontogeny?

Evolutionary Linguistics

Definition

Scientific study of both the origins and development of language as well as the cultural evolution of languages.

- Schleicher's (1853) language tree (Stammbaumtheorie) theory
- lack of fossil records, difficult to empirically verify, banned by Societe linguistique de Paris in 1866
- revived at the end of 20th century (c.f. Pinker & Bloom, 2011)
- quantitative comparative linguistics, phylogenetic trees...
- focuses on phylogeny and not ontogeny

Why EL should focus on ontogeny

"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, 1975)"

Formal Language Theory

Alphabet A is a finite, nonempty set of symbols.
A word or a string over an alphabet A is a finite sequence of symbols of from A.
A* is the set of all words over A*.

A language L over A is a subset of A*.

A grammar G is a quadruple=(N,T,P,S) where N is the nonterminal alphabet, T is the terminal alphabet, $S \in N$ is the axiom and P is the set of rewriting (production, substitution) rules, written as $x \to y$. Grammars are called

- REGULAR when the form of all rules in P is $X \to \alpha, X \to \alpha B, \alpha \in T, A, B \in N$
- CONTEXT-FREE when all its rules have form $X \to x$ where $X \in N$, $x \in A_G^*$
- CONTEXT-SENSITIVE when P contains only rules of the form $x_1 X x_2 \to x_1 w x_2$ x_1, x_2, w being strings over $A_G, X \in N$

Language L GENERATED from grammar G is a set of all sequences of terminals which can be derived from axiom S by recursive application of rules in P.

Language L can be PARSED by grammar G if, for all sequences $s \in L$, there exist at least one sequence of application of production rules which, when applied in an inverse fashion (i.e. substitute left side of production rule for the right side), shall end at axiom S.

Grammar Systems

Introduction

A Grammar System is a set of grammars working together, according to a specified protocol, to generate a language.

- a syntactic theory of multi-agent, distributed and parallel systems
- multiple independent grammars share their productions in "string environment" (analogic to AI "blackboard" approaches)
- environment can change on its own (so called "eco-grammar" systems) or not (language colonies)

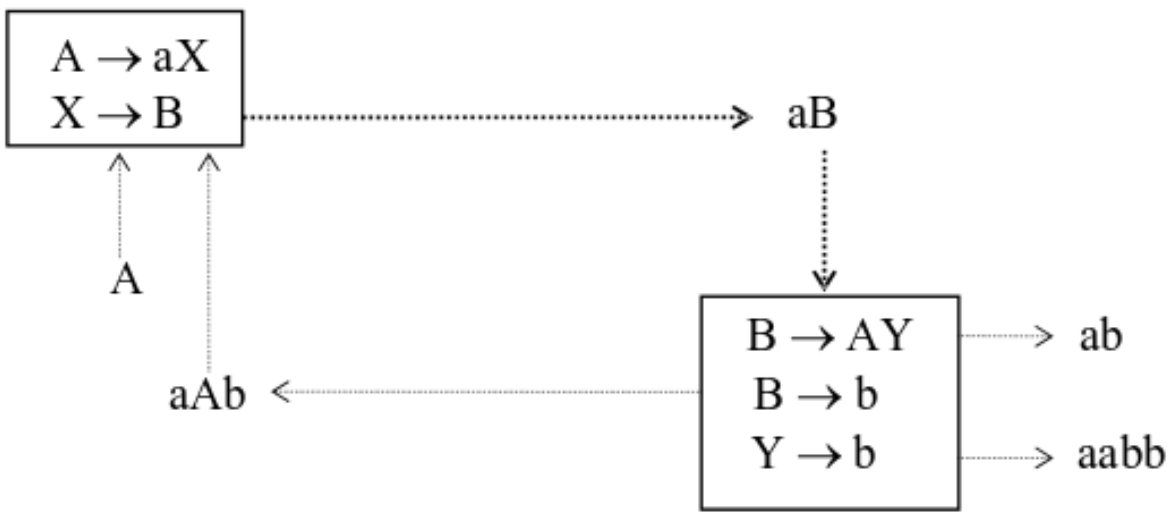


Figure: Reproduced from Kelemen's (2004) article "Miracles, colonies, and emergence".

Natural Language Processing

- uses computers to process human languages
- implements AI, data-mining, information retrieval and machine learning methods (both supervised and unsupervised)
- first and ultimate NLP challenge posed by Turing (1950)
- other problems: anaphora resolution, automatic summarization, discourse analysis, machine translation, morpohological segmentation, named entity recognition, natural language understanding, POS-induction and tagging, parsing, question answering, sentiment analysis, speech recognition, word sense disambiguation etc...

- in NLP, statistics often plays more important role than FLT
- in NLP, methods based on aNN, Naive Bayes, SVM-ba methods are predominant, EC is much less used

POS-induction and Grammar induction

Part-of-speech induction

The goal is to group tokens, present in the pure-text corpus C, into clusters grouping members of diverse parts-of-speech (nouns, verbs, adjectives, etc.).

Grammar induction

The goal is to infer, from pure-text corpus C, a grammar G which could have generated the corpus C.

POS-i and GI problems are strongly intertwined. Clusters discovered by POS-i can be denoted by non-terminal symbols.

C: John loves Mary. Mary hates John. Mary sleeps. John weeps.

ideal grammar:
N$\rightarrow$John||Mary
V$\rightarrow$love||hate||sleep||weep
S$\rightarrow$NVs||NVsN

least general grammar: S$\rightarrow$C
most general grammar: S$\rightarrow$A*

Learning of semantic categories

How can machines work with semantic categories?

Semantic categories (i.e. concepts) can be characterized

- by extensive (listing the instances) or ostentative (pointing the finger) definition
- in terms of sufficient and necessary features
- as (convex) subspaces of N-dimensional semantic feature space (Gardenfors 2004)
- as prototypes (points) within such spaces

Principle(s) behind construction of semantic spaces

In neurosciences: "neurons that fire together, wire together" (Hebb 1964)
In linguistics: "a word is characterized by the company it keeps" (Harris 1954) In philosophy: "the meaning of a word is its use in the language (Wittgenstein 1953)

Conjecture

Development of vocabulary in human children is a variant of multi-class classification problem and as such can be simulated by an algorithm creating and partitioning semantic feature vector spaces.

Developmental Psycholinguistics

Developmental Psycholinguistics (DP) is a scientific discipline studying changes occuring in human faculty of understanding and production of natural languages. As such, it is closely related to developmental psychology (a sub-field of psychology) and developmental linguistics (a sub-field of linguistics).

Language Development (DEF)

Language development (LD) - or ontogeny of natural language L in human individual H - is a constructivist process gradually transforming L into evermore optimized communication channel facilitating the exchange of information between H and her social surroundings.

- language is social and pragmatic (allows children to manipulate objective world)
- comprehension precedes production: C-representations offer preliminary targets for P-productions
- physiological predispositions of language are innate but useless without triggering epigenetic stimuli
- children are not "ideal learners" (in Gold's theorem sense)
- brains simultaneously encode multiple language registers and grammars

Motherese

- parents modify their language in order to make themselves understood
- higher pitch (267 Hz in comparison to 198Hz), slower tempo, greater rhytmicity, longer pauses between utterances
- "much of the speech addressed to babies consists of short, routine, repetitive utterances produced with great consistency and frequency in the same contexts, day after day" (Clark 2003)"
- repetitions three times more frequent in speech to two-years-old than in speech to ten-years-old

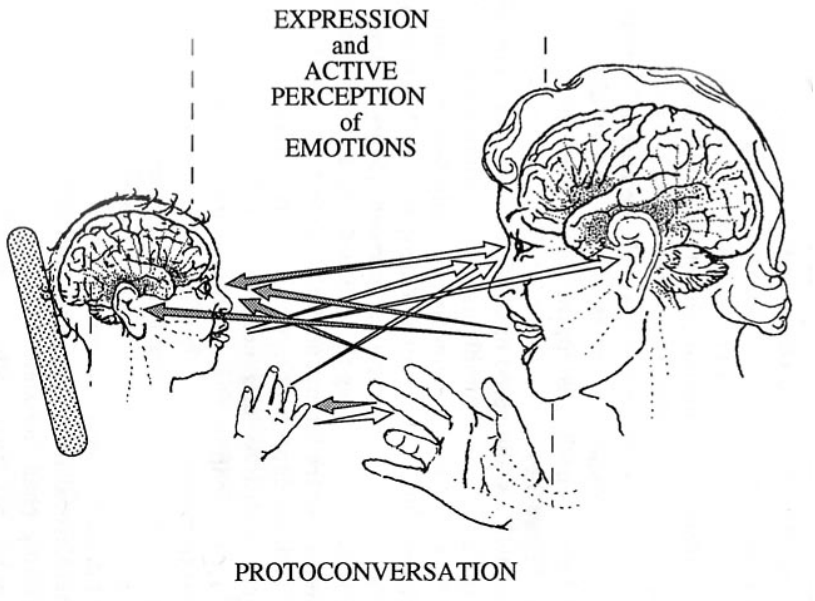


Figure: Reproduced from Tverarthen (1993).

Toddlerese

- babys expressive faculties start with 1bit communication channel (need soothing/don't need soothing)
- gets more subtle and fine-grained with time: more information transmitted with less signal
- babbling starts cca at 8 months of age, first as repetition of same syllables (mamamama, babababa), later syllables shall start to vary within the sequence (babadadabebe)
- around 1 year: consistent vocalizations in specific contexts (protowords)
- children tend to be quite accurate in their first productions but later versions of the same words appear to be further from adult targets
- "continuous exploration, experimentation, practice and intense involvement with linguistic structure" (Labov, 1978)
- LD reveales, upon closer inspection, a constantly changing series of small experiments where child progressively scrutinizes and tries out different options (Clark 2003)
- a lot of variability in children's word forms (Ferguson and Farewell 1975)
- first grammar form around "pivot" words, e.g.: "mama auch, tato auch, nana auch, baba auch" ($S \rightarrow Nauch$; $N \rightarrow mama|tato|nana|baba$)
- toddlerese: 10 - 30 months

Quantitative laws of language acquisition

Piotrowski law

In both linguistic phylogeny as well as ontogeny (e.g. sentence length, vocabulary size) does development follow the logistic equation: $\frac{c}{1+ae^{-bt}}$. Note that in ecology, the same equation is considered to yield *the law of population growth* (Lotka, 1920).

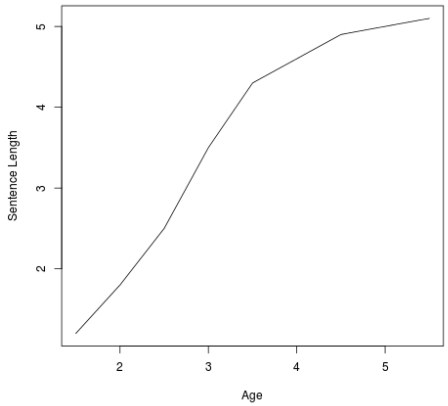
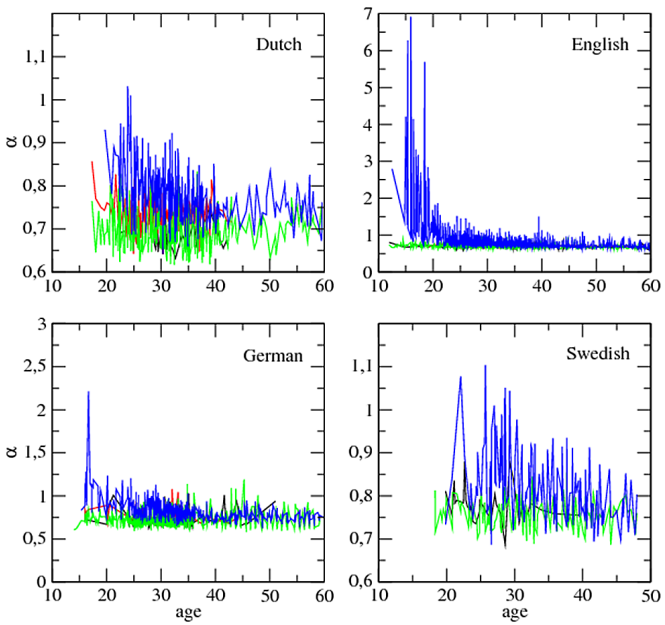


Figure: reproduced from (Baixeries et al., 2013).



Zipf's law

Zipf (1949) showed that if the most frequent word in a text is assigned rank 1, the second most frequent word is assigned rank 2 etc. than frequency f(r) of a word of rank r obeys $f \approx r^{-\alpha}$ (i.e. follows the power-law distribution). Recent (Baixeries et al., 2013) analyses of CHILDES corpus indicate that the exponent α depends on age and is much higher and decreases faster in small children.

Common aspects of both LD and evolution

Axiomatic

1. Convergence: different trajectories, same result
2. Variation: children PLAY, children forget
3. Non-monotonicity: locally "correct" behaviours are lost

Hypothetic

1. Adaptation: gradual convergence of L_T towards L_M and possibly G_T towards G_M
2. Replication: both vertical (repetition) and horizontal (non-local storage)
3. Parallel coexistence of schemas
4. Selection: correct behaviours are rewarded

Subject and Method

Subject

My own daughter. 0-30months (0-2;6)

Method

Phenomenological method based principially on amazed observations.
Long-term journal.
Little or no experimental (artificial) interactions beyond natural and normal scenarios.

Cognitive Crossover Cases

Case 1 - Banan

Banan was called "baja" in (1;6) and "anan" in (1;10). At (1;11) a following interaction took place:
F: banan ; C: anan
F: banan ; C: anan
F: baja ; C: bajan
F: bajan ; C: banan

Case 2 - Olol

Very intensive "Krtko & Orol" period between 1;10-1;11. Word "OLOL" used with high frequency on a regular basis. During one pre-sleep monologue, subject said "KOLOL" when enumerating the names of her friends from creche, one among them being named Nikol.

Bilingual crossovers

oči+augen=oge
opica+afe=api
voda+wasser=vava
etc...

Quantitative observations

Corpus

CHILDES - Child Language Data Exchange System (MacWhinney and Snow 1984)

1. more than 130 corpora of transcripted child verbal interactions
2. more than 20 languages

Variation operators whose impact shall be analyzed

1. Substitutions - *papija* $\rightarrow$ *babija* $\rightarrow$ *mamija*
2. Reduplications - *hau* $-$ *hau*
3. Omissions - *vlak* $\rightarrow$ *ak*

Method

Matching with Perl-compatible regular expression (Hromada 2011). Reduplication, for example, can be easily detected with regexp (\d{2,})\1.
Note that strings can evolve by substituting substrings for other strings and the substitution rule itself is also a string.

Grammar Induction

- inducing not one monolithic grammar but populations of individual grammars
- fitness function promotes individuals which
  1) match patterns present in environment
  2) generate utterances which shall be "accepted" by environment
  3) minimize number of utterances which shall not be accepted by environmennt
- individual grammar is encoded as an ordered sequence of production rules

Corpus

#Mutter#
#Vater#

$Grammar_1$

$Vat \rightarrow A$
$\#A \rightarrow B$
$er\# \rightarrow A$
$\#Mu \rightarrow A$
$tA \rightarrow A$
axioms: AA BA

$Grammar_2$

$er\# \rightarrow B$
$\#Va \rightarrow A$
$\#Mu \rightarrow A$
$tB \rightarrow B$
$er\# \rightarrow B$
axioms: AB

Concept Construction

Attaching meanings to words interpreted as supervised learning of multiclass classifier. In most recent experiments I crossover four ideas in order to create it:

- RANDOM PROJECTION - exploiting lemma Johnson-Lindenstrauss to project problem into D-dimensional space
- BINARIZATION - transposition of problem from real-valued spaces to binary (Hamming) spaces (Hromada 2014)
- THEORY OF PROTOTYPES - every category C can be characterized by a prototype P_C which is as close as possible to members of C and as far as possible members of other category (Rosch 1973)
- EVOLUTIONARY COMPUTATION - thus searches for such a set of K prototypes $P_1, ..., P_N$ which maximizes the "prototype fitness function":

$$F(I) = \sum_{i}^{N}(\sum_{C}^{K} H(P_B, i)_{P_C=P_B if L_i \neq C} - H(P_G, i)_{P_C=P_G if L_i=C})$$

where $H(P_G, i)$ is the Hamming distance between the binary vector denoting the prototype P_G and the document i contained in training document set of cardinality N.
Every individual is a binary vector obtained by concatenation of vectors of all K prototypes $|I| = D * K$.

Concept Construction - Preliminary Results

Trained on training part and evaluated on testing part of 20newsgroups corpus (K=20)
LSB parameters D=128, S=3, I=2
CGA (N=100, P_M=0.001, one-point crossover) with 1/8 elitism

Figure: Evolutionary induction of semantic prototypes - training

Figure: Evaluation of induced prototypes against the testing set

Algorithm seems to perform better than "deep learning" Semantic Hashing method of Salakhutdinov & Hinton (2009).
An Evolutionary Computation Algorithms are capable of generalization and can be thus considered a case of Machine Learning.

Thesis

1. At some level of abstraction, ontogeny of syntactic and semantic categories is a process consistent with tenets of Universal Darwinism.
2. Representations in human mind are subjects of variation, selection and replication.
3. In young children this process is still not completely internalized (Vygotsky 1934) and is thus visible to external observer.
4. Evolutionary Computation is a means how this process can be successfully simulated *in silico*.

Merci

Thank You.