

Language Adaptation helps Language Acquisition

Willem Zuidema

Artificial Intelligence Laboratory

VUB, Brussels, Belgium

and

Language Evolution and Computation Unit

University of Edinburgh, Scotland

Selforganization & Linguistics

In formal linguistics there exists a tendency to assume detailed, innate specifications to explain observed linguistic phenomena.

In artificial life / complex systems there is a desire to explain complex patterns from simple interactions.

Two lines of attack

1. Show that interesting language-like patterns can emerge without detailed innate specifications.
2. Show that arguments that “prove” the need for such a specification are wrong.



Social transmission: Learners learn from learners



- meaning repertoire (Steels & Kaplan, 1999)
- sound repertoire (de Boer, 1997)
- Saussurean sign (Oliphant & Batali, 1996)
- shared lexicons (Steels, 1996)
- phonemic code (Oudeyer, 2001)
- color words (Belpaeme, 2002)
- compositional morphology (Batali, 1997)
- compositional word order (Kirby, 1999)
- recursive syntax (Kirby, 2000; Batali, 2002)
- word order universals (Kirby, 1998)

Horizontal Transmission: the negotiation model

John Batali (2002) The negotiation and acquisition of recursive grammars as a result of competition among exemplars, In: Linguistic evolution through language acquisition (Ted Briscoe, ed.), Cambridge University Press

- Exemplars of form (strings) – meaning (predicate logic) associations
- Learning: Cost based on success
- Production & Interpretation: Cost based on number of modifications
- Population – every agents communicates with and learns from every other agent

Results: more general (i.e. compositional) exemplars are used more often and become less costly

Vertical Transmission: the iterated learning model

Simon Kirby (2000) Syntax without natural selection, In: The Evolutionary Emergence of Language (Knight, C., Hurford, J. & Studdert-Kennedy, M., eds.). Cambridge University Press.

- Context-free grammars, predicate logic
- Induction algorithm: minimal description length
- Chain – every next agent learns from the previous
- Bottle-neck: limited number of training examples

Results: more general (i.e. compositional) rules have a higher probability of surviving the bottle-neck

Language Adaptation

1. Languages are transmitted culturally and are subject to change
2. Languages will change more if they are difficult to learn or less adequate
3. Over time, languages that are easy to learn/adequate are more likely to occur (Deacon, 1996; Kirby 2000)

- Language adaptation helps language acquisition

.. The challenge is to show in computational model the relevance of this for “the argument from the poverty of the stimulus”.

Universal Grammar as an innate specification

“Chomsky’s hypothesis is that many aspects of the formal structure of language are encoded in the genome.” (Wexler, 1999, “Innateness of Language”, MITECS)

- “Hypothesis testing”, “querying the UG” in 1st language acquisition
- “Loosing access to the UG” in 2nd language acquisition
- “language organ”
- “language genes”

The Poverty of the Stimulus (I)

“Every aspect of language that can not reasonably be assumed to be present in the primary linguistic data, must be part of the innate machinery of the human brain” (Cook, 1993, Linguistics and 2nd language acquisition)

The Poverty of the Stimulus (II)

“The basic results of the field [of learnability theory] include the formal, mathematical demonstration that without serious constraints on the nature of human grammar, no possible learning algorithm can in fact learn the class of human grammars.” (Wexler, 1999, “Innateness of Language”, MITECS)

What this talk is about ...

- Criticising arguments for this extreme view of UG
- Arguing that “every *possible* language is equally likely” is NOT the proper assumption for natural languages.

... and not about

- denying humans have an innate bias for language (nurture rather than nature, empiricism vs. nativism);
- denying that this bias is language-specific and has been selected for in the evolution of language;
- nor about presenting a “realistic” model for language acquisition and evolution.

The model

Search space (Grammar Universe)

Context-free grammars of the form: $A \mapsto t$, $A \mapsto BC$ or $A \mapsto Bt$. Convenient, and no restrictions on expressiveness.

Parsing: depth-first search (Norvig's optimization), with maximum depth d

Derivation: random string from parsable language

Interaction: speaker derives random string s , and hearer checks if s is a string from its own language. If so, the interaction is a success, otherwise it is a failure.

Search procedure (Grammar Induction)

Incorporation: *extend the language, such that it includes the encountered string; if string s is not already part of the language, add a rule $S \mapsto s$ to the grammar.*

Compression: *substitute frequent and long substrings with a nonterminal, such that the grammar becomes smaller and the language remains unchanged;*

1. for every valid substring z of the right-hand sides of all rules, calculate compression effect $v(z)$ of substituting z with a nonterminal A ;
2. replace all valid occurrences of the substring $z' = \operatorname{argmax}_z v(z)$ with A if $v(z') > 0$;
3. and add a rule $A \mapsto z'$ to the grammar.

Generalization: *equate two nonterminals, such that the grammar becomes smaller and the language larger;*

1. for every combination of two nonterminals A and B ($B \neq 'S'$), calculate the compression effect v of equating A and B .
 2. Equate the combination $(A', B') = \operatorname{argmax}_{A, B} v(A, B)$ if $v(A', B') > 0$; i.e. replace all occurrences of B with A .
- The compression effect is measured as the difference between the number of symbols before and after replacing and deleting redundant rules;
 - Compression and generalization are each repeated until nothing changes anymore;
 - Off-line learning (all examples all received at once);
 - Sufficient expressiveness enforced.

Example

Training data:

abcabcabcd, abcabcabcabcd, abcd, abcabcd

After incorporation:

S->abcabcabcd, S->abcabcabcabcd, S->abcd, S->abcabcd,

After compression:

S->Tabcd, S->Tabcabcd, S->Ud, S->Td, T->UU, U->abc,

After generalization:

S->Uabcd, S->Uabcabcd, S->Ud, U->UU, U->abc,

Poverty of Stimulus (I): Nurture nor nature

$S \rightarrow Uabcd$, $S \rightarrow Uabcabcd$, $S \rightarrow Ud$, $U \rightarrow UU$, $U \rightarrow abc$

Structural ambiguity:

$abcd$: $((abc)(d))$

$abcabcd$: $((abc)(abc))d$

$abcabcd$: $((abc)(abcd))$

This property is neither in the data nor in the algorithm, it results from the interaction between the two.

Poverty of Stimulus (II): Identification in the limit

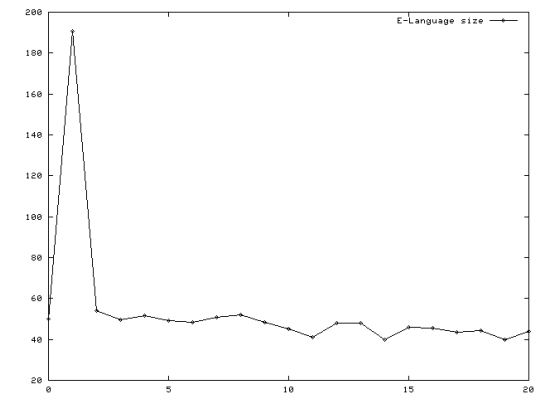
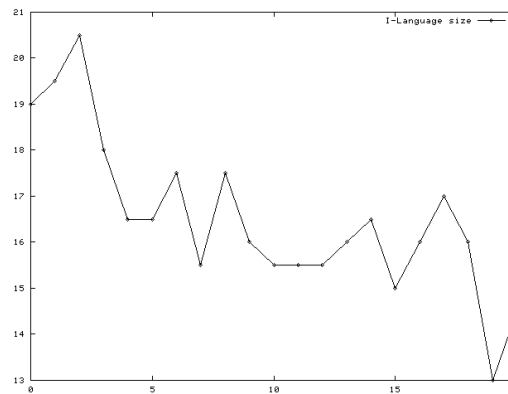
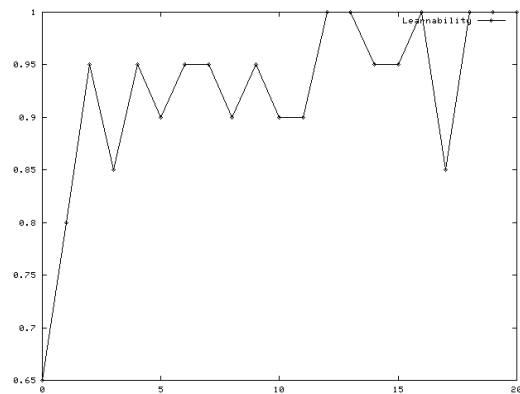
“the results [...] show that only the most trivial class of languages considered is learnable (in the sense of identification in the limit) from text [...].” (Gold, 1967, “Language identification in the limit”)



$\dots \mapsto \text{sentences} \mapsto \left\{ \begin{array}{l} \text{incorporation} \\ \text{compression} \\ \text{generalization} \end{array} \right\} \mapsto \text{sentences} \mapsto \dots$

Iterated learning

Transmission: individuals in a *chain* learn from the previous individual and teach to the next;



(a) learnability (b) I-Language size (c) E-Language size Parameters: $V_t = \{a, b, c, d\}$, $V_{nt} = \{S, X, Y, Z, A, B, C\}$, $T=30$, $E=20$, $l_0=3$. Shown are the average values of 2 simulations.

Conclusions

- In this model, language adapts to the bias of the learning algorithm. The algorithm therefore needs less training samples than Nowak et al. predict as a lower bound.
- Results that “prove” the need for Universal Grammar (i.e. restrictions of the search space) are based on the assumption that any target grammar from that space is equally likely. Here we show that in iterated learning that assumption is not reasonable.

- Limitations of the learning procedure make the learning in future generations easier. The collective dynamics give “emergent” restrictions; the “poverty of stimulus” does not make binary and a priori restrictions of the search space necessary.

The poverty of stimulus is now no longer a problem; instead, the ancestors' poverty is the solution for the child's.

- What is Universal Grammar?
 1. Representable languages
 2. Learnable languages
 3. Stable outcomes of iterated learning

The Poverty of the Stimulus (III)

(Nowak, Komarova & Niyogi, 2001, Science)

1. There is a minimum learning accuracy q_1 necessary to maintain grammatical coherence in the population.
2. The best possible learner (the “batch learner”) needs a minimum amount b_c of sample sentences to reach q_1

“If all the b sentences happen to be consistent with more than one grammar (say, with r grammars), then the [batch] learner can pick any of the r grammars with probability $1/r$.”

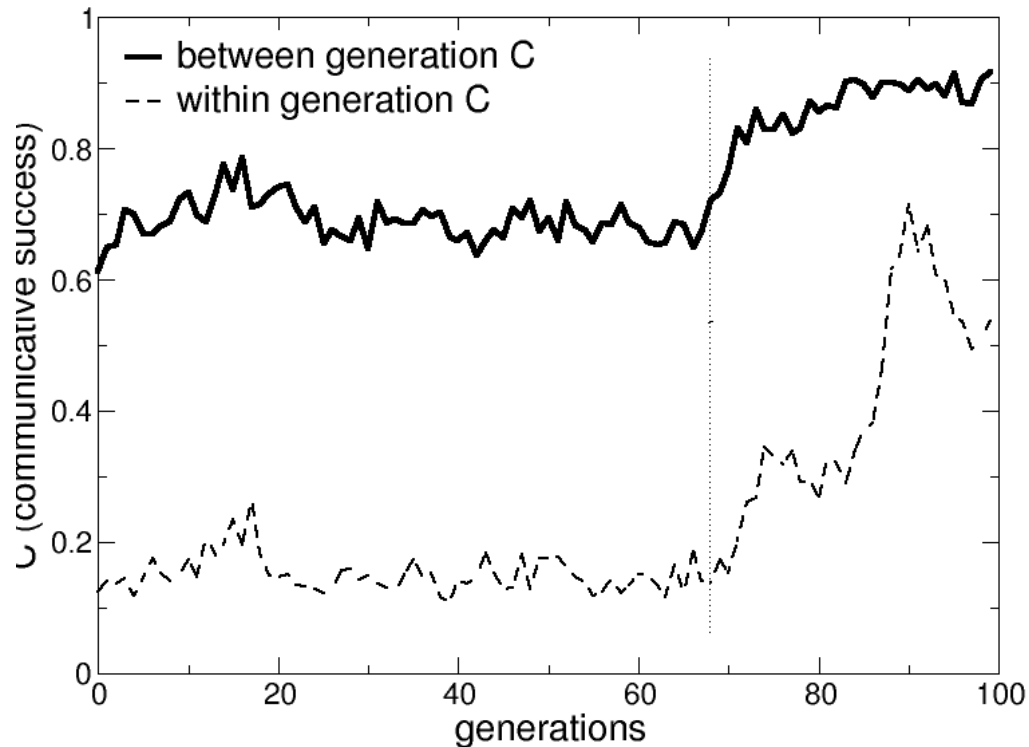
3. b_c is proportional to the number of possible grammars n
 - The Universal Grammar can only be of small size.

Fitness proportional selection

“We assume that the learning mechanism employed by humans lies somewhere between these two extremes [i.e. the batch and memory-less learners].”

Transmission: the fitness of an individual is determined by its success in communicating with the individuals of its own generation. The expected number of offspring is proportional to this fitness.

Results



Parameters: hearer benefit condition, $V_t = \{0, 1, 2, 3\}$,
 $V_{nt} = \{S, a, b, c, d, e, f\}$, $P=20$, $T=100$, $M=100$, $d=8$,
 $l_0=12$

There are regions of grammar space where the dynamics are apparently under the “coherence threshold”, while there are other regions where the dynamics are above this threshold. The parameters, including the number of sample sentences T , are still the same, but the language has adapted itself to the **bias** of the learning algorithm.

Future directions

- Semantically constrained language evolution:
 - learnability
 - language universals
- Baldwin effect
- Interaction evolution – learning
 - Parametrization of the learning algorithm (currently only: NT_COST)
 - Baldwin effect (part of the rules considered innate)
- Origins of compositionality
 - Apparent compositionality — productive compositionality
 - Needs a model that does not presupposes it (Zuidema & Westermann — Jordan Pollack)