

Language is different: grammar acquisition in a language game model

December 13, 2001

Formal models of language acquisition

The need for Universal Grammar

- Representation bias (search space)
- Procedural bias (search procedure)
- Collective dynamics (“learners learn from learners”)

Assumptions

- Infants acquire language
- There is a (possibly finite) “grammar universe” $(x_0 \dots x_n)$
- There are errors in learning (Q_{ij})
- Communicative success is dependent on similarity $(F_{ij} \propto a_{ij})$
- There is a “poverty of stimulus” (b)

The logical problem of language acquisition

“The basic result of the field [of learnability theory] is 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)

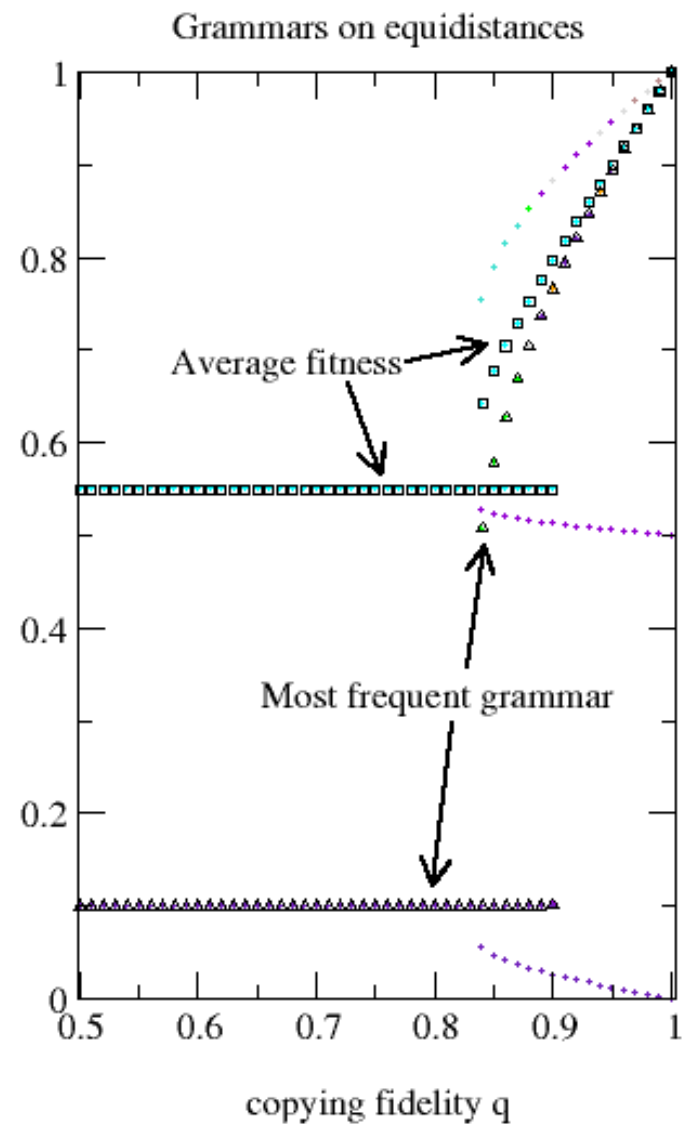
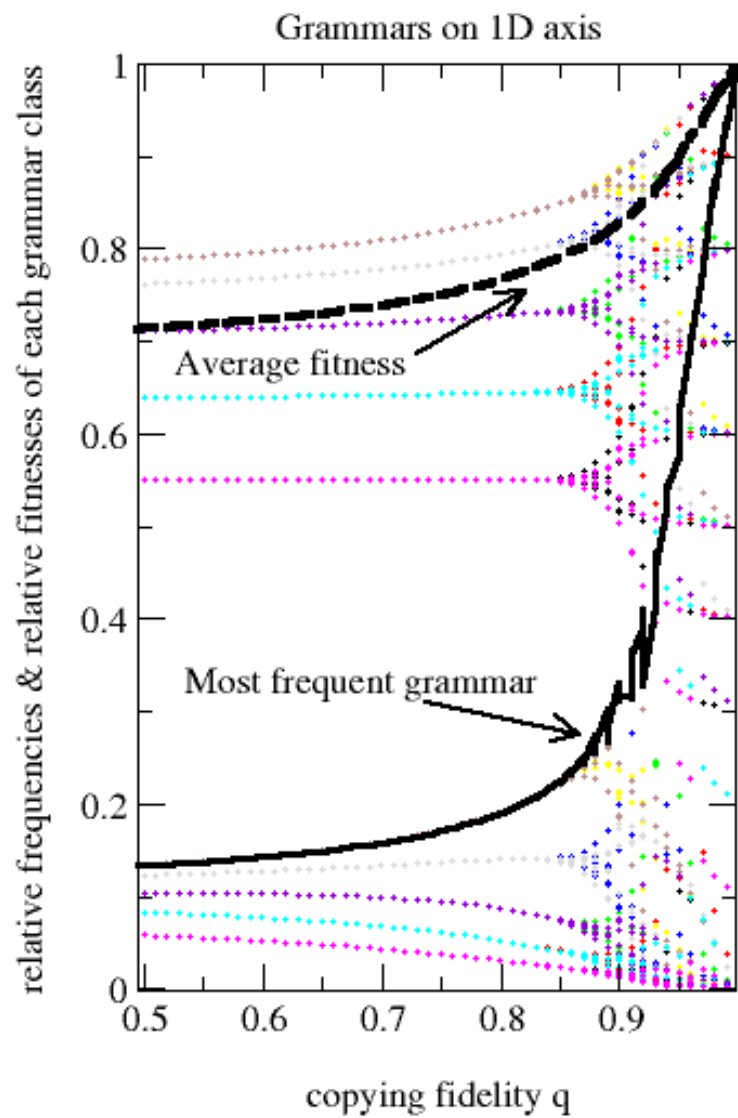
- Formalism: rewriting grammars (Chomsky)
- No negative evidence
- No semantics or context
- Criterion: “identification in the limit” (Gold, 1967), for every possible grammar

The coherence threshold

There is a minimum learning accuracy q_1 necessary to maintain grammatical coherence in the population. Because the best possible learner (the “batch learner”) needs a minimum amount b_c of sample sentences to reach q_0 , and b_c is proportional to the number of possible grammars n , the Universal Grammar can only be of small size (Nowak et al, 2001).

$$\dot{x}_i = \sum_{j=0}^N x_j f_j Q_{ji} - \phi x_i \quad (1)$$

- Successful grammars are more likely to be transmitted (fitness, role model) ($f_i = \sum_j (x_j F_{ij})$)
- Every grammar is equally likely, is equally expressive and is equally similar to every other grammar



Representation of the linguistic abilities

- Context-free grammars of the form: $A \mapsto t$, $A \mapsto BC$ or $A \mapsto Bt$
- Almost in Chomsky Normal Form, thus no restrictions on expressiveness
- For determining the language L of a certain grammar G we use simple depth-first exhaustive search of the derivation tree.
- In the communication between two agents, the speaker chooses a random element s of its language, and the hearer checks if s is an element of its own language. If so, the interaction is a success, otherwise it is a failure.

The learning algorithm

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; for every substring z of the right-hand sides of all rules, calculate the compression effect $v(z)$ of substituting z with a nonterminal A ; replace all occurrences of the substring $z' = \operatorname{argmax}_z v(z)$ with A if $v(z') > 0$, and add a rule $A \mapsto z'$ to the grammar. The compression effect is measured as the difference between the number of symbols in the grammar before and after the substitution.

Generalization: equate two nonterminals, such that the grammars becomes smaller and the language larger; for every combination of two nonterminals A and B ($A, B \neq S$), calculate the compression effect v of equating A and B . 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.

- off-line learning
- avoiding overgeneralization
- enforcing sufficient expressiveness

```

teach(i, j)
  repeat T times
    teacher i generates random string s from  $L_i$ 
    student j calls incorporate(s)
  repeat until  $G_j$  doesnot change anymore
    student j calls compress()
  repeat until  $G_j$  doesnot change anymore
    student j calls generalize()
  repeat N times, or until generalization is rejected
    student j generates random string  $s'$  from  $L_j$ 
    if ( $s' \notin L_i$ ) reject generalization
if ( $size(L) < M$ )
  repeat  $M - size(L)$  times
    generate random string s of maximum size  $l_0$ 
    student i calls incorporate(s)

```

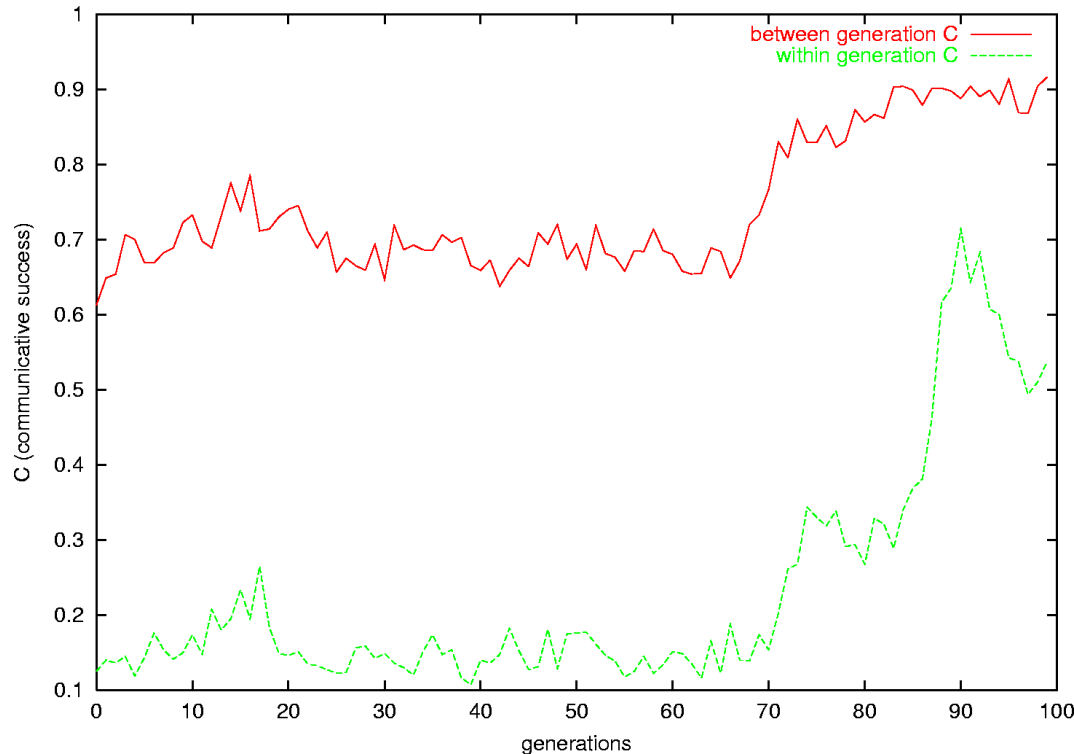
Transmission

Iterated learning : agents in a *chain* learn from the previous agent and teach to the next;

Fitness proportional selection : the fitness of an agent is determined by its success in communicating with the agents of its own generation. The expected number of offspring is proportional to this fitness.

- *Mutualistic*: the fitness is calculated by counting the number of times an agent is successful as a hearer and as a speaker
- *Parasitic*: the fitness is calculated by counting the number of times an agent is successful only as a hearer.

Results



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

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.

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 / representation bias) 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 “logical problem of language acquisition” can be solved without binary and a priori restrictions of the search space.