The Evolutionary Dynamics of Natural Language
Dimacs Workshop on Computation and Language, 1998

Charles D. Yang, Robert C. Berwick
MIT Artificial Intelligence Laboratory
charles, berwick@ai.mit.edu
Partha Niyogi
Lucent Technology Inc.
niyogi@research.bell-labs.com

1 Introduction

Essential to Darwinian evolution are the concepts of popultional and variational thinking [15, 13]. The variant properties among individuals in a population allow the forces of evolution such as natural selection to operate. Evolutionary changes are therefore changes in the distribution of variant individuals in the population. This paper reports two projects that draw insights from these perspectives to investigate the evolutionary dynamics of natural language.

The first project [19] studies the problem of child language acquisition. We propose that the state of the learner can be viewed a population of variant hypotheses/grammars with a probabilistic distribution, and that language acquisition is a selectional process where the fitness of a hypothesis is be defined as its compatibility with the linguistic data. We prove some convergence results and show that the model predicts certain psychological findings that are otherwise hard to obtain.

2 A Selectional Model of Language Acquisition

2.1 The Problem

One of the central problems in modern linguistics and cognitive science is the problem of language acquisition: how does a human child come to acquire the target language in her linguistic environment with such ease, yet without favorable learning conditions such as (effective) correction or negative examples?

The *empiricist* approach to this problem has recently (re)gained popularity in computational linguistics and machine learning. The child is viewed as an inductive learner that derives linguistic regularities from the statistical distribution of patterns in the input data. This approach, however, cannot be correct.

Consider the following finding from child psycholinguistic research: different aspects of grammatical knowledge are learned at different rates. In French, inflected verbs precede negation and adverbs, e.g. "Jean n'*aime* pas Marie" (literally, "John likes not Mary", an ungrammatical sentence in English). The rule for this kind of verb placement is acquired before the child's first utterance at about the 20th month, as evidenced by the extreme rarity of incorrect verb

placement in child speech [16]. On the other hand, the rule that requires English to use a sentential subject is acquired much later, as English children continue to produce subjectless sentences until the 30th month [2]. Now if we look at the actual input to children (the CHILDES corpus compiled by [14]), we find that virtually all parental English sentences have a subject, while only 7-8% of parental French sentences contain an inflected verb followed by negation/adverb. The purely inductive learning approach, e.g. one that builds statistical models based on pattern distribution in the input data, predicts that the use of sentential subjects in English should be learned much earlier than the placement of the verb in French – exactly the opposite of the actual findings in child language.

Another leading approach to language acquisition, largely in the tradition of generative linguistics, capitalizes on the fact that although child language is patently different from adult language, it is different in highly *restrictive* ways. Given the input to the child, there are logically possible, simple generalizations (inductive rules) to describe the data that are never attested in child language. For example, forming a question in English involves inversion of the auxiliary verb and the subject:

Is the man t tall?

where "is" has been fronted from the position t, the position where it assumes in a declarative sentence. A possible inductive rule to describe the above sentence is this: front the *first* auxiliary verb in the sentence. This rule, though logically possible and simple to use, is never attested in child language [4, 8]; that is, children are never seen to produce sentences like:

$\star$ Is the cat that the dog t chasing is scared?

where the first auxiliary is fronted (here, "is"), instead of the auxiliary following the subject of the sentence (here, the second auxiliary verb in the sentence).

Findings like these lead linguists to postulate that the human language capacity is constrained in some *a priori* space – Chomsky's Universal Grammar (UG). Previous studies [3, 18, 1, 11] of language learnability in the UG framework are *transformational*, borrowing a term from evolution [13], in that the learner moves from one hypothesis/grammar to another as input sentences are processed. Since at any time the state of the learner is identified with a particular hypothesis/grammar, it is hard to explain (a) the inconsistent patterns in child language, which cannot be described by any single grammar [2], and (b) the smoothness of language development, whereby the child gradually converges to the target grammar, rather than the abrupt jumps that would be expected from binary changes in hypotheses/grammars [17] .

2.2 A Population of Grammars

We propose that the state of the learner is a *population* of grammars that are made possible by the biological endowment of the human language faculty. Each grammar G is associated with a *weight* p_G. For the target grammar T, we say

that learning *converges* if $\lim_{t\to\infty} p_T) = 1$.

The learning algorithm is as follows:
For an input sentence s, the child

1. with the probability $p(G)$, selects a grammar G
2. - if $s \in G$ then $p'(G) = \frac{p(G)+v}{1+v}$
   - if $s \notin G$ then $p'(G) = \frac{p(G)-w}{1-w}$

The algorithm is an on-line Hebbian one. It is on-line to reflect the rather limited computational capacity of the child language learner–sophisticated data processing and a large memory to store previously seen input examples are deemed psychologically implausible. The Hebbian type of associative learning is motivated by biological evidence on the development of specific neural substrates that are guided by specific input stimulus from the environment [12, 7, 10].

Input sentences can be grouped into two classes with respect to the target grammar T:

1. $\nexists G \in \overline{T}, \text{such that} s \in G$.
2. $\exists G \in \overline{T}, \text{such that} s \in G$.

Sentences belonging to type 1 above are refered to as *unambiguous triggers*, that is sentences that can only be analyzed by the target grammar. For example, suppose that English is the target grammar, which, as mentioned earlier, requires a sentential subject. Languages like Italian, Spanish, and Chinese, on the other hand, have the option of dropping the subject:

(lui)	ha	parlato	**Italian**
he	has	spoken	**English**

Therefore, sentences with subjects are not necessarily useful to distinguish English from Italian. However, there exists a certain type of English sentences that is informative:

There is a man in the room.

The sentential subject "there" does not carry any referential meaning in the statement, unlike thematic subjects that denote the agent or the participant of an action. Thus the presence of "there" is for purely structural reasons, to satisfy the requirement that the pre-verbal subject position in English must be filled. Italian lacks this requirement, and thus lacks this sentence type.

Suppose at time t, the weight of the target grammar T is $p(t)$. Let's look at the expected value of $p(t+1)$ at time $t+1$, after an input sentence has been presented to the learner:

1. $s \in T$ and $\nexists G \in \overline{T}, s \in G$
   (a) with probability $p(t)$, T is chosen:
   $p(t+1) = \frac{p(t)+v}{1+v}$

(b) with probability q_i, $G_i \in \overline{T}$ is chosen:
$p(t+1) = \frac{p(t)}{1-w}$

2. $s \in T$ and $\exists G \in \overline{T}, s \in G$

(a) with probability $p(t)$, T is chosen:
$p(t+1) = \frac{p(t)+v}{1+v}$

(b) with probability q_i, $G_i \in \overline{T}$ is chosen:

- with probability α_i,[1] $s \in G_i$:
$p(t+1) = \frac{p(t)+v}{1+v}$
- with probability $(1-\alpha_i)$, $s \notin G_i$:
$p(t+1) = \frac{p(t)}{1-w}$

It is easy to show that $E[p(t+1)] \geq p(t)$. Viewed as a stochastic process, $p(t)$ is a sub-martingale [9], which converges to a unique point. Some special cases of $p(t)$ will be studied in details.

The similarity between the proposed model and biological evolution is clear. A population consists of grammars with different properties, which are translated into different "fitness" values in a particular linguistic environment. Over time, the target grammar, which by definition has the highest fitness value, will gradually win out. It is important to note that the target's rise to dominance is through a step-wise Markovian process, as depicted in the learning algorithm, thus preserving psychological plausibility by reducing the computational load of the learner.

2.3 Predictions of the Model

Suppose that natural language grammars vary in a parametric space, as cross-linguistic examination suggests [5, 6]. We can then study the dynamical behaviors of grammar classes that are grouped along a parametric dimension. Specializing the general result on $p(t)$, the weight of the target grammar class is governed by a recursive function:

$$p(t+1) = p(t) + cp(t)q(t), \qquad \text{where} \qquad q(t) = 1 - p(t) \tag{1}$$

where c is a constant determined by the frequency (u) of unambiguous triggers in the input data and by the learning rates v and w. The convergence to a single target grammar can then be viewed as the intersection of parametric grammar classes, each of which is converging to the target value of the respective parameter. This process thus corresponds to the selection of independent loci in biological evolution.

We examine the predictions of the proposed model in two well-studied cases of child language acquisition: the acquisition of English obligatory subject and

[1] This variable indicates how compatible a non-target grammar G_i is with the input data. Thus, by definition, $0 \leq \alpha_i < 1$.

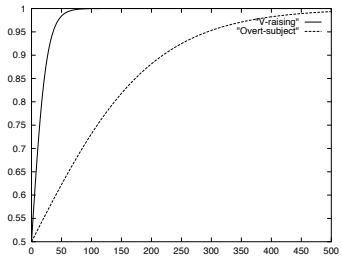

Figure 1: The X-axis represents the effective sample size, the number of sentences that change the distribution in the grammar population; the Y-axis represents the frequency of correct use of the target parameter values, as predicted by the learner modeled in 1 (assuming the null hypothesis that the child accesses the same population of grammars (and thus their distribution) to analyze as well as produce sentences).

the acquisition of French inflected verb placement. (Section 2.1 pointed out the difficulties they pose for previous studies of language learnability.) Note that other things being equal, the rate of learning is determined by the frequency of unambiguous triggers in the input data. For the relevant cases, the patterns of such triggers are shown in as follows:

1. French verb placement

   Jean n' *aime* pas Marie.

   Jean *embrasse* souvent Marie.

2. English obligatory subject

   There are cookies in the jar.

   Is *there* a toy train on the floor?

Using naturalist parental speech to children recorded in an on-line corpus [14], we estimated the frequencies of sentences of types 1 and 2 are 8% and 1%, respectively. Figure 1 shows the predicted learning curves for the acquisitions of these two aspects of grammar. The sharp contrast reported in Section 2.1 between the developmental time courses of these aspects of grammar is predicted.

We will also present evidence from other aspects of actual language development in support of the model. This project, if successful, will be a first step towards a "population genetics" theory of language acquisition which bridges the gap between discrete grammars and continuous language development. It will bring formal rigor to the study of natural language and cognitive systems and will situate it in a broader biological framework.

References

[1] Berwick, Robert C. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

[2] Brown, Roger (1973). *A first language*. Cambridge, MA: Harvard University Press.

[3] Chomsky, Noam (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

[4] Chomsky, Noam (1975). *Reflections on language*. New York: Pantheon.

[5] Chomsky, Noam (1981). *Lectures on government and binding*. Dordrecht: Foris.

[6] Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

[7] Jean-Pierre Changeux (1983). *L'Homme Neuronal*. Paris: Fayard.

[8] Steven Crain (1991). Language acquisition without experience. With peer commentaries. *Behavioral and Brain Sciences*.

[9] Joseph Doob (1953). *Stochastic Processes*. New York: Wiley.

[10] Edelman, Gerald (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.

[11] Gibson, Edward and Kenneth Wexler (1994). Triggers. *Linguistic Inquiry* 25: 355-407.

[12] Hubel, David H. and Torsten N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160: 106-54.

[13] Lewontin, Richard C. (1983). The organism as the subject and object of evolution. *Scientia* 118: 65-82.

[14] MacWhinney, Brian and Catherine Snow (1985). The Child Language Date Exchange System. *Journal of Child Language* 12, 271-296.

[15] Mayr, Ernst (1982). *The growth of biological thoughts*. Cambridge, MA: Harvard University Press.

[16] Pierce, Amy (1992). *Language acquisition and syntactic theory: a comparative analysis of French and English child grammar*. Boston: Kluwer.

[17] Pinker, Steven (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

[18] Wexler, Kenneth and Peter Culicover (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

[19] Yang, Charles D. (forthcoming). The variational dynamics of natural language. Ph.D. Dissertation. Department of Electrical Engineering and Computer Science, MIT.