Evolutionary Consequences of Language Learning

Partha Niyogi
Robert C. Berwick
Center for Biological and Computational Learning
Massachusetts Institute of Technology, Room E25-201
45 Carleton St.
Cambridge, MA 02142
Replies to: berwick@ai.mit.edu
Note: figures follow paper; sent in postscript form.

April 12, 1996

Abstract

Linguists' intuitions about language change can be captured by a dynamical systems model derived from the dynamics of language acquisition. Rather than having to posit a *separate* model for diachronic change, as is typically done in the diachronic literature by drawing on assumptions from population biology (cf. Kroch, 1989), this new model dispenses with the need for these independent assumptions by showing how the behavior of *individual* language learners leads to emergent, global *population* characteristics of linguistic communities over several generations. As the simplest case, we formalize the example of two grammars (languages) differing by exactly one binary parameter, and show that even this situation leads directly to a nonlinear (quadratic) dynamical system. We study this one parameter model in a variety of situations for different kinds of acquisition algorithms, maturational times and show how different learning theories can have very different evolutionary consequences. We are thus able to precisely formulate an evolutionary criterion for the adequacy of linguistic and learning theories and by applying the computational model to the historical loss of Verb Second from Old French to modern French we show that otherwise adequate grammatical theories can fail our new evolutionary criterion.

1 Introduction: Language Ontogeny & the Paradox of Language Change

Much research on language learning has focused on how children — individuals — acquire the grammar of their caretakers from "impoverished" data presented to them during childhood. Cast formally, the logical problem of language acquisition requires a learner to converge to its correct target grammar — the language of its caretakers, and presumably a member of the class of possible natural language grammars. Posed this way, language acquisition mirrors the familiar case of biological ontogenesis — the development of a mature individual biological faculty from its initial state.

Language scientists have also long been occupied with describing phonological, syntactic, and semantic change, often appealing to a relation between language change and evolution, but rarely going beyond analogy. For instance, Lightfoot (1991, chapter 7, pp. 163–65ff.) talks about language change in this way: "Some general properties of language change are shared by other dynamic systems in the natural world."

The overall goal of this paper is to move from this analogy to formal modeling. Just as in the biological sciences, we can logically move from the analysis of *individual* biological development to *population* development — that is, from description of language change at the individual level — language acquisition — to description of language change at the ensemble population level — a distribution of final attained states over time. In the usual biological setting, this amounts to the sufficient logical requirements for a model of evolution, leaving aside natural selection, as noted by Lewontin (1978, 184):

> A sufficient mechanism for evolution by natural selection is contained in three propositions:
>
> 1. There is variation in... behavioral traits among members of a species (the principle of variation);
> 2. The variation is in part heritable... in particular, offspring resemble their parents (the principle of heredity);
> 3. Different variants leave different numbers of offspring either in immediate or in remote generations (the principle of differential fitness)

Clearly, the first two conditions are met in our case, where differing grammars (languages) and language acquisition respectively serve as the principles of variation and heredity.[1]

Since all the requirements for an evolutionary model are satisfied, we have in place all the elements to formally model diachronic language change — change in the ensemble properties of language populations — using the formal armamentarium of evolutionary biology, and, furthermore, *deriving* population changes over time from individual ontogenesis, just as in the biological case. In brief, this is the aim of the current paper: to put the study of language change on the same firm formal foundation as evolutionary population biology, deriving a model of ensemble language change from a model of individual language change, just as in the biological case.

Indeed, from at least one perspective, linguistics has a substantial *advantage* over traditional biological studies of genetic change (evolution): in the case of ordinary biological evolutionary models, the mapping from an individual's gene frequencies (their *genotype*) to a developed organism, or ontogenesis, is essentially completely unknown, yet is required for a full model of evolution. However, in the analogous case of language, we *do* have a model of language ontogenesis — namely, the models of language acquisition that have been a focus of language research for many years.[2] In this strong sense, then, not only can we draw on evolutionary biology to model language change, we can possibly advance beyond what is possible in biological evolutionary modeling.

To begin, we note that Lewontin's principle of variation requires that individuals differ in their final attained "phenotypes", or grammars. Cast in our terms, this comes to the following paradox. The language acquisition problem, if solved perfectly, would lead to language stasis: If generation after generation children successfully attained the grammar of their parents, then languages would never change with time. Yet languages do change.

We can resolve this paradox by introducing an explicit, formal evolu-

[1] We leave aside the principle of differential fitness for now, though it might be easily accommodated in the mathematical modeling that follows. example, the notion of "selection" may be readily and exactly incorporated in any number of ways, viz., as so-called cultural effects. Similarly, so-called "least effort" effects, if they prove relevant, may be so incorporated.

[2] This fact may be surprising to some readers. However, again as Lewontin (1978) observes, biological evolution involves a mapping from genotype space to phenotype space (the organism's "external form" on which selection actually acts). In no case except the most trivial sense is this mapping known; certainly not even for the simplest organism in full.

tionary model for language change, grounded on language acquisition as the source of slight variation that can arise from generation to generation. We introduce a computational dynamical systems model for this purpose, to the best of our knowledge, the first such model employed to describe diachronic language processes, and investigate its consequences. Specifically, we show that a computational population language change model emerges as a natural consequence of individual language learnability — as expected from general evolutionary considerations. Our computational model establishes the following:

1. *Learnability* is a well-known criterion for the adequacy of grammatical theories. Our model provides an *evolutionary* criterion: By comparing the trajectories of dynamical linguistic systems to historically observed trajectories, one can determine the adequacy of linguistic theories or learning algorithms.

2. We derive explicit dynamical systems corresponding to parameterized linguistic theories (e.g. the Head First/Final parameter in HPSG or GB grammars) and memoryless language learning algorithms (e.g. gradient ascent in parameter space).

3. In the simplest possible case of a 2-language (grammar) system differing by exactly 1 binary parameter, the system reduces to a quadratic map with the possibility of the usual chaotic properties (dependent on initial conditions). That such complexity can arise even in the simplest case suggests that formally modeling language change may be quite mathematically rich.

4. We illustrate the use of dynamical systems as a research tool by considering the loss of Verb Second position in Old French as compared to Modern French. We demonstrate by computer modeling that one grammatical parameterization advanced in the linguistics literature does not seem to permit this historical change, while another does.

5. We can more accurately model the time course of language change. In particular, in contrast to Kroch (1990) and others, who mimic population biology models by imposing an S-shaped logistic change by *assumption*, we explain the time course of language change, and show that it need not be S-shaped. Rather, language-change envelopes are *derivable* from more fundamental properties of dynamical systems: sometimes they are S-shaped, but they can also be nonmonotonic.

2 The Acquisition-Based Model of Language Change

We show how a combination of a grammatical theory and a learning paradigm leads directly to a formal dynamical systems model of language change.

First, informally, consider a linguistically homogeneous adult population speaking a particular language. Individual children exposed to example sentences attempt to attain their caretaker target grammar. After a finite number of examples, some are successful, but others may misconverge. The next generation will therefore no longer be linguistically homogeneous. The third generation of children will hear sentences produced by the second—a different distribution—and they, in turn, will attain a different set of grammars. Over generations, the misconvergences of individual learners will propagate leading to the evolution of the linguistic composition of the population as a dynamical system. In the remainder of this paper we formalize this intuition, showing the evolution of language types over successive generations within a single community. We return to the details later, but let us first formalize our intuitions.

2.1 Grammatical theory, Learning Algorithm, Sentence Distributions

Let us formally specify the following objects that will play a key role in determining the nature of our dynamical system for language change.

1. Denote by $\mathcal{G}$, a family of possible (target) grammars. Each grammar $g \in \mathcal{G}$ defines a language $L(g) \subset \Sigma^*$ over some alphabet Σ in the usual way.

2. Denote by P, the distribution with which sentences of Σ^* are presented to the individual learner (child). More specifically, let P_i be the distribution with which sentences of the ith grammar $g_i \in \mathcal{G}$ are presented if there is a speaker of g_i in the adult population. Thus, if the adult population is linguistically homogeneous (with grammar g_1) then $P = P_1$. If the adult population speaks 50 percent $L(g_1)$ and 50 percent $L(g_2)$ then $P = \frac{1}{2}P_1 + \frac{1}{2}P_2$.

3. Denote by $\mathcal{A}$ the learning algorithm that children use to hypothesize a grammar on the basis of input data. If d_n is a presentation sequence of n randomly drawn examples, then learnability requires the learner to converge to the target grammar in the limit (for every target grammar g_t in the class), i.e.,

$$Prob[\mathcal{A}(d_n) = g_t] \longrightarrow_{n \to \infty} 1$$

Learnability serves as an important criterion for the adequacy of linguistic theories. Thus linguists attempt to characterize the class of possible human languages by $\mathcal{G}$ in such a way that the class is learnable. Developmental psychologists attempt to characterize the learning algorithm by means of which children actually choose grammars in this class on exposure to primary linguistic data. By combining the results of each research enterprise, we attempt to derive the evolutionary consequences of particular theories of language and associated theories of learning.

2.2 Dynamical System Model

We now define the resultant dynamical system by providing its two necessary components:

A State Space ($\mathcal{S}$): a set of system states. Here, the state space is the space of possible linguistic compositions of the population. Formally, a state is described by a distribution P_{pop} on $\mathcal{G}$. The distribution P_{pop} describes the proportion of the population speaking each of the languages corresponding to the grammars in $\mathcal{G}$. The state space $\mathcal{S}$ is therefore the space of all possible probability distributions on $\mathcal{G}$. Note that the state space depends only upon the grammatical theory and nothing else.

An Update Rule: how the system states change from one time step to the next. Typically, this involves specifying a function, f, that maps $s_t \in S$ to s_{t+1}. In our case the update rule can be derived directly from the learning algorithm $\mathcal{A}$ in conjunction with the sentence distributions P_i's. Learning is the key that changes the distribution of languages spoken from one generation to the next.

Let us outline the procedure for obtaining the update rule. Given the state at generation t, i.e., $P_{pop,t}$, we see that any $\omega \in \Sigma^*$ is presented to the learner with probability

$$P(\omega) = \sum_{h_i \in \mathcal{G}} \in P_i(\omega) P_{pop,t}(h_i)$$

where $P_{pop,t}(h_i)$ is the proportion of the adult population who have internalized the grammar h_i and P_i is distribution with which such speakers produce sentences.

The learning algorithm $\mathcal{A}$ uses the linguistic data (n examples, indicated by d_n) and conjectures hypotheses ($\mathcal{A}(d_n) \in \mathcal{G}$). One can, in principle, compute the probability with which the learner will develop an arbitrary hypothesis, h_j, after n examples:

$$Prob[\mathcal{A}(d_n) = h_j] = p_n(h_j) \quad (1)$$

Imagine that after n examples, maturation occurs, i.e., the child retains for the rest of its life the hypothesis it has after n examples. Then, with probability $p_n(h_j)$, an arbitrary child will have internalized grammar h_j. Thus, in the next generation, a proportion $p_n(h_j)$ of the population will have grammar h_j, i.e., the linguistic composition of the next generation is given by $P_{pop,t+1}(h_j) = p_n(h_j)$ for every $h_j \in \mathcal{G}$. In this fashion, we have an update rule, $P_{pop,t} \stackrel{\mathcal{A}}{\longrightarrow} P_{pop,t+1}$.

Maturation is a psychologically plausible theory that captures the notion that there is a finite learning phase after which humans do not attempt to change their grammars any further. In other words, humans are not forever entertaining the possibility of changing their current grammatical hypotheses with the availability of more data, but after a period of time, they "mature" and retain their mature hypothesis for the rest of their adult lives. There might be some debate about when exactly this maturation occurs but for our purposes we assume that there is some value n that characterizes this. From a mathematical perspective, we could take the limit of eq. 1 as n tends to infinity to derive the dynamical system in the absence of any maturational theory. Such a limit is however not guaranteed to exist and the maturational theory therefore aids us in making sure that the update rule always exists.

Generality of the approach. Note that such a dynamical system exists for every choice of $\mathcal{A}$, $\mathcal{G}$, and P_i (relative to the constraints mentioned earlier). In short then, we have outlined the procedure for the following transformation

$$(\mathcal{G}, \mathcal{A}, \{P_i\}) \longrightarrow \mathcal{D}(\text{ dynamical system})$$

Importantly, this formulation does *not* assume any particular linguistic theory, learning algorithm, or distribution over sentences. One can now investigate within this framework, the evolutionary implications of a variety of learning theories and grammatical theories and compare the evolutionary predictions against historical data.

3 Language Change in Parametric Systems

We now instantiate our abstract system by modeling some specific cases. Suppose we have a "parameterized" grammatical theory, such as HPSG or GB (Chomsky, '81), with n boolean-valued parameters and a space $\mathcal{G}$ with 2^n different languages (in this case, equivalently, grammars). Further take the assumptions of Gibson and Wexler (1994), regarding sentence distributions and learning: P_i is uniform on unembedded sentences generated by g_i and $\mathcal{A}$ is a local, online, error-driven learning algorithm called the TLA (Triggering Learning Algorithm). For concreteness, we provide a formal description of the TLA.

TLA (Triggering Learning Algorithm

- [Initialize] Step 1. Start at some random point in the (finite) space of possible parameter settings, specifying a single hypothesized grammar with its resulting extension as a language;
- [Process input sentence] Step 2. Receive a positive example sentence s_i at time t_i (examples drawn from the language of a single target grammar, $L(G_t)$), from a uniform distribution on the degree-0 sentences of the language (we relax this distributional constraint later on);
- [Learnability on error detection] Step 3. If the current grammar parses (generates) s_i, then go to Step 2; otherwise, continue.
- [Single-step hill climbing] Step 4. Select a single parameter uniformly at random, to flip from its current setting, and change it (0 mapped to 1, 1 to 0) *iff that change allows the current sentence to be analyzed*;

Clearly, the TLA is a memoryless learning algorithm that updates its grammatical hypothesis after every example sentence in an attempt to attain the target grammar (parameter settings). There are variants of the TLA that we will consider later in this paper.

To derive the relevant update rule for our dynamical system of language change we need to be able to quantify eq. 1. To that end, we are helped by the following results (the first straightforward; see Niyogi, 1994):

Claim 1 *Any memoryless incremental learning algorithm that attempts to set the values of the parameters on the basis of example sentences, can be modeled exactly by a Markov Chain. For an n-parameter system, this Markov chain has 2^n states with each state corresponding to a particular*

grammar. The transition probabilities depend upon the distribution P with which sentences are provided to the learner, and the manner in which the learning algorithm $\mathcal{A}$ updates its hypothesis.

Ofcourse, it is hardly surprising that a memoryless learning algorithm can be modeled by a first order Markov chain. The usefulness of the Markov analysis however, is that it now allows us to characterize the probability with which the learner will attain each of the possible parameter settings. Specifically,

Lemma 1 *The probability that the memoryless learner internalizes hypothesis h_i after m examples (solution to equation 1) is given by:*

$$Prob[\ Learner's\ hypothesis\ = h_i \in \mathcal{G}\ after\ m\ examples]$$

$$= \{\tfrac{1}{2^n}(1,\ldots,1)'T^m\}[i]$$

Here, T is the transition matrix of the Markov chain characterizing the hypothesis changes made by the memoryless learner, $(1,\ldots,1)$ is a 2^n- dimensional row vector with all ones, and the learner starts with an initial hypothesis chosen uniformly at random.

The above lemma characterizes eq. 1 that we can now use to derive the update rule. Thus, we obtain our required dynamical system for parameter-based theories and memoryless acquisition algorithms. The evolution can be characterized in the following manner.

1. Let $\mathbf{\Pi_1}$ be the initial population mix. Assume P_i's as above. Compute the distribution of primary linguistic data to the children (P) accordingly from $\mathbf{\Pi_1}$, and P_i's.

2. Compute T (the transition matrix of the learner) according to the claim.

3. Use the lemma to obtain the update rule, to get the population mix $\mathbf{\Pi_2}$.

4. Repeat for the next generation.

Let us now apply this mathematical model to the simplest possible case: two grammars (languages) differing by exactly one binary parameter. We

shall see that even here the mathematics becomes nontrivial. Following the detailed analysis of the one parameter case, pursuing a more realistic setting, we turn to a 3-parameter, 8 grammar model, concluding with an application to a real case of diachronic syntax change, the loss of verb second in the change from Old French to Modern French.

4 One Parameter Models of Language Change

Consider the following simple scenario.
$\mathcal{G}$: Imagine that due to UG constraints there are only two possible grammars (parameterized by one boolean valued parameter) associated with two languages in the world, L_1 and L_2.
$\mathcal{P}$: Suppose that speakers who have internalized grammar g_1 produce sentences with a probability distribution P_1 (on the sentences of L_1). Similarly, assume that speakers who have internalized grammar g_2 produce sentences with P_2 (on sentences of L_2).

One can now define

$$a = P_1[L_1 \cap L_2]; 1 - a = P_2[L_1 \setminus L_2]$$

and similarly

$$b = P_2[L_1 \cap L_2]; 1 - b = P_2[L_2 \setminus L_1]$$

$\mathcal{A}$: Assume that the learner uses the TLA to set parameters.
$\mathcal{N}$: Let the learner have just two example sentences before maturation occurs, i.e., after two example sentences, the grammatical hypothesis the learner has will be retained for the rest of its life.

Given this framework, it is possible, as discussed, to characterize the behavior of the individual learner by a Markov chain (see Niyogi and Berwick, 1994) with two states, one corresponding to each grammar. If sentences were provided according to distribution P_1, the transition matrix would be T_1 and if sentences were provided according to P_2 the transition matrix would be T_2 as shown below:

$$T_1 = \begin{bmatrix} 1 & 0 \\ 1-a & a \end{bmatrix}$$

$$T_2 = \begin{bmatrix} b & 1-b \\ 0 & 1 \end{bmatrix}$$

The TLA learner's hypothesis would change from g_1 (corresponding to language L_1) to g_2 (correspondingly L_2) from example to example according to transition probabilities shown in the matrices above. Thus, if the learner happens to pick L_1 as its random initial hypothesis, and the target grammar happens to be L_2, then, with probability b the learner will retain its hypothesis after one example, and with probability $1-b$, it will change it.

What happens when there is no single unique target grammar? It is possible to show that if sentences are drawn with probability p from L_1 and probability $1-p$ from L_2, then the transition matrix characterizing the learner's hypotheses is provided by:

$$T = pT_1 + (1-p)T_2$$

This would allow us to characterize the evolving linguistic composition of the population over time.

4.1 The Linguistic Population

At any given point in time, the population consists only of speakers of L_1 and L_2. Consequently, the linguistic composition can be represented by a single variable, p: this will denote the fraction of the population speaking L_1. Clearly $1-p$ will speak L_2.

It is possible to show that the linguistic composition will evolve as:

Theorem 1 *The linguistic composition in the $(n+1)$th generation (p_{n+1}) is related to the linguistic composition of the nth generation (p_n) in the following way:*

$$p_{n+1} = Ap_n^2 + Bp_n + C$$

where $A = \frac{1}{2}((1-b)^2 - (1-a)^2)$; $B = b(1-b) + (1-a)$ *and* $C = \frac{b^2}{2}$.

A few points are in order:

1. When $a = b$, the system has exponential growth. When $a \neq b$ the dynamical system is a quadratic map (which can be reduced by a transformation of variables to the logistic, and shares the same dynamical properties). See fig. 1

2. The scenario $a \neq b$ is much more likely to occur in practice— consequently, we are more likely to see logistic growth rather than exponential ones. Crucially, the logistic form has now been *derived* rather than *assumed* as in previous work (e.g., Kroch, 1990).

3. We get a class of dynamical systems. The quadratic nature of our map comes from the fact that $N = 2$. If we choose other values for N we would get cubic and higher order maps. We show the explicit derivation of some of these. There are already an infinite number of maps in the simple one parameter case. For larger parametric systems and more complicated learning algorithms, the mathematical situation is significantly more complex.

4. Logistic maps are known to be chaotic. In our system, it is possible to show that:

Theorem 2 *Due to the fact that $a, b \leq 1$, the dynamical system never enters the chaotic regime.*

This naturally raises the question—is this true for all grammatical dynamical systems, specifically the linguistically "natural" cases? Or are there ones where chaos will manifest itself? (It would obviously be quite interesting if all the natural grammatical system spaces were nonchaotic.) Further research on this subject is planned.

4.2 Other Choices of N:

Let us now consider the case where the maturation time, N, is equal to 3. Everything else is just as before, i.e., there are two languages, and the learning algorithm is the TLA. How does the population evolve?

As before, the state of the population at any point in (generational) time can be characterized by a single variable p taking values in $[0, 1]$. Thus, p represents the proportion of the population speaking language L_1 (correspondingly, having internalized grammar, g_1).

It is possible to prove:

Theorem 3 *The evolution of p is given by the cubic map*

$$p_{n+1} = Ap_n^3 + Bp_n^2 + Cp_n + D$$

where $A = (a-b)^2(2-a-b); B = (a-b)(2-2a+4b-ab-3b^2); C = 2(1+b)(1-a)+b^2(2-3b+a); D = b^3$

Interestingly enough, if $a \neq b$, we get a cubic map. If $a = b$, however, the system degenerates to a first order map again. Note, however, that this first order map is different from that obtained when $N = 2$ and $a = b$. The cubic nature of this map arises clearly due to our choice of $N = 3$.

As the value of N increases, we will get higher order maps. However, the coefficients of the map (indicated by A, B, C, D in the above cubic case) depend in nontrivial ways upon the parameters (a and b). Consequently, the coefficients are not independent of each other. Furthermore they are not able to take on all possible values since $0 < a, b < 1$.

Now consider the case where $N = \infty$. In other words, the child has infinite amount of time to mature and attain its linguistic hypothesis. It is possible to prove

Theorem 4 *The proportion of L_1 speakers evolves according to the update rule*

$$p_{n+1} = \frac{p_n(1-a)}{(1-b) + p_n(b-a)}$$

A number of observations are worthwhile. First, notice that if $a = b$, we find that $p_{n+1} = p_n$ for all n. In other words, the population *never* changes its linguistic characteristics from generation to generation. If, $a < b$ it can be shown that p_n tends to 1 as n tends to infinity from all initial conditions. This makes sense for $a < b$ implies that language L_1 is easier to learn than L_2 because there are less ambiguous (in the sense of being analyzable by both grammars) sentences in L_1 and consequently over generations we find that all speakers tend to acquire L_1. The reverse is true when $a > b$ when it is possible to show that p_n tends to 0 as n tends to infinity.

In summary, if the maturation time is $N = \infty$, we find that populations either remain stable all the time with no change at all ($a = b$) or one language type is completely eliminated from existence over time.

Thus we see that the number of examples the child is given in order for it to form its mature, adult, hypothesis might significantly affect the dynamics of the evolutionary systems that result. One could, therefore, in principle, concretely quantify the evolutionary effect of different maturational theories and use this to judge the adequacy of such theories for human language acquisition.

4.3 Other Choices of $\mathcal{A}$:

Let us now return to the situation where $N = 2$. As we discussed, we get a variety of quadratic maps when this happens. How this map will depend upon the parameters a, b depend upon the details of the learning algorithm. The learning algorithm is just a mapping from the data to grammatical hypotheses. This mapping might be deterministic or it might be stochastic

corresponding to randomized learning algorithms. We examine below a few simple variations in the learning algorithms and the corresponding language change models that correspond to these.

Consider the following variations in algorithms:

Algorithm 1 ($\mathcal{A}_1$) : Start with a default parameter setting of L_1. Now follow the TLA. In other words:

1. Let Initial Hypothesis be g_1.
2. Receive new example sentence.
3. Flip hypothesis if and only if the new example sentence is not analyzable by the current hypothesis.
4. Go to 2.

The above algorithm is an online, memoryless one that differs from the TLA only in that it chooses g_1 as a default initial hypothesis rather than randomly choosing one. Its behavior can easily be analyzed by the Markov chain method, eq. 1 can be characterized and the language change model derived. In contrast, consider the following two algorithms that operate in batch mode. In other words, the learner makes a hypothesis after analyzing *both* the examples it receives. Of course, one could have batch algorithms for arbitrary N in which case the learner would make its hypothesis after *all* N examples have been received. We have considered only the case of $N = 2$ for simplicity. The first algorithm ($\mathcal{A}_2$) is one that is biased to prefer g_2 as a hypothesis. The second algorithm ($\mathcal{A}_3$) is a stochastic version of the first.

Algorithm 2 ($\mathcal{A}_2$) :

1. Collect two example sentences.
2. Choose g_2 unless both examples are not analyzable by g_2. Otherwise, choose g_1.

Algorithm 3 ($\mathcal{A}_3$) :

1. Run algorithm $\mathcal{A}_2$.
2. Flip the output of $\mathcal{A}_2$ with probability η and output the result as the grammatical hypothesis.

It is possible to show:

Theorem 5 *The linguistic population evolves according to the following rules in each of the three cases:*

(1) $p_{n+1} = (a-b)(1-b)p_n^2 + (1-b)(2b+1-a)p_n + b^2$

(2) $p_{n+1} = (1-a)^2 p_n^2$

(3) $p_{n+1} = \eta + (1-2\eta)p_n^2$

With this analysis in hand, we can make some summary observations. The form of the update rule, though quadratic, differs in the three cases. In case 1, we get a logistic update only if $a \neq b$. In case 2, it is easy to show that the the population always moves to a fixed stable point of a completely homogeneous L_2 speaking community, i.e., p_n tends to 0 in all cases except where $a = 0$. This makes sense as the learning algorithm is biased towards g_2 and it becomes difficult for L_1 speakers to survive for very long. Interestingly, case 1 and case 2 correspond to two situations where the learning algorithm is biased but with drastically different evolutionary consequences. While one language type is entirely eliminated in case 2, both language types are usually always present in the evolutionary trajectories corresponding to case 1. In this fashion, one could concretely determine the evolutionary implications of different kinds of learning algorithms. Case 3 corresponds to a randomized learner and here we get a quadratic map if $\eta > 0$. As a point of interest, we note that when $\eta = 1$, we actually get period doubling behavior and in the limit, the populations are always homogeneous except that generations alternately speak L_1 and L_2. Of course, $\eta = 1$ corresponds to a learner that returns a hypothesis that is the exact opposite of $\mathcal{A}_2$. Needless to say, this latter algorithm is not a very good learner.

5 Example 2: A Three Parameter System

Turning next from the simplest situation to a more realistic setting, let us consider a specific example to illustrate the derivation of the previous section: the 3-parameter syntactic subsystem described in Gibson and Wexler (1994). Specifically, posit 3 Boolean parameters, Specifier first/final; Head first/final; Verb second allowed or not, leading to 8 possible grammars/languages (English and French, SVO−Verb second; Bengali and Hindi, SOV−Verb second; German and Dutch, SOV+Verb second; and so forth). The learning algorithm is the TLA. For the moment, take P_i to be a uniform distribution on unembedded sentences in the language. The key results we obtain by computer simulation of the resulting dynamical systems are as follows:

1. **All +Verb second populations remain stable over time.** Non-verb second populations tend to *gain* Verb second over time (e.g., English-type languages change to a more German type) contrary to historically observed phenomena (loss of Verb second in both French and English) and linguistic intuition (Lightfoot, 1991). This evolutionary behavior suggests that either the grammatical theory or the learning algorithm are incorrect, or both.

2. **Rates of change can vary from gradual S-shaped curves to more sudden changes** (fig. 5).

3. **Diachronic envelopes are often logistic, but not always.** Note that in some alternative models of language change, the logistic shape has sometimes been *assumed* as a starting point, see, e.g., Kroch (1990). However, Kroch concedes that "unlike in the population biology case, no mechanism of change has been proposed from which the logistic form can be deduced." On the contrary, we propose that language learning (or mislearning due to misconvergence) could be the engine driving language change. The nature of evolutionary behavior *need not* be logistic. Rather, it arises from more fundamental assumptions about the grammatical theory, acquisition algorithm, and sentence distributions. Sometimes the trajectories are S-shaped (often associated with logistic growths): sometimes not as in fig. 5.

4. **In many cases a homogeneous population splits into stable linguistic groups.**

5. **Varying maturation time affects evolutionary trajectories.** See fig. 5.

6. **Different initial population mixes lead to phase space plots with possible fixed points.** In the previous simulations, we always initialized the dynamical system with a homogeneous poulation. Instead of starting with homogeneous populations, one could, of course, consider any nonhomogeneous initial condition, e.g. a mixture of English and German speakers. Each such initial condition results in a grammatical trajectory. One typically characterizes dynamical systems by their phase-space plots. These contain all the trajectories corresponding to different initial conditions, exhibited in fig. 5.

It remains to precisely characterize the fixed points in such settings and for different parameterizations. Note, though, that the possibility of fixed points for different initial language mixtures gives rise to the following (perhaps important) possiblility, which we again leave for future work:

7. **The existence of fixed points for language mixtures offers a potentially novel alternative account of the "universality" of**

creole languages. Bickerton (1981, 1990) has argued that since geographically distinct creole languages — languages relatively rapidly and diachronically evolved from different contact languages — all seem to share the same distinctive syntactic/morphological properties despite their disparate geographic origins, it must be the case that all creoles mirror a common "universal grammar". In other words, the explanation for the commonality of creole structure lies in an apparent reversion to an underlying universal grammar. However, the dynamical systems model suggests another *logically possible* explanation: common creole characteristics are a by-product of the common fixed point for the initial language mixes.[3]

6 The Case of Modern French

We next briefly consider a different parametric system (studied by Clark and Roberts, 1993) as a test of our model's ability to impose a diachronic criterion on grammatical theories. The historical context is the evolution of Modern French from Old French, in particular, the loss of "Verb second," the appearance of a verbal element in exactly the second position of a sentence. *Loss of Verb-Second* (from Clark and Roberts, 1993)

Mod.	*Puis entendirent-ils un coup de tonerre.
	then they heard a clap of thunder.
Old	Lors oirent ils venir un escoiz de tonoire.
	then they heard come a clap of thunder

Recall that simulations in the previous section indicated an (historically incorrect) tendency to gain Verb second over time. We now consider Clark and Roberts' (1993) alternative 5-parameter grammatical theory. These parameters include: (1) Null subjects or not; (2) Verb second; and three other binary parameters having to do with case theory that we need not detail here, yielding 32 possible languages (grammars). It has been generally argued that in the middle French period, word forms like Adv(erb) V(erb) S(ubject) decreased in frequency, while others like Adv S V increased; eventually bringing about a loss of Verb second. We can now test this hypothesis with the model, varying initial conditions about population mixtures, for-

[3]Of course, this possibility does not exclude Bickerton's account of creole's universality — from another point of view, it strengthens his conclusion, but it does so for another reason. However, in our view the fixed point possibility does weaken the argument of universality from observed common convergence.

eign speakers, etc.

Starting from just Old French, our model shows that, even without foreign intrusion, eventually speakers of Old French die out altogether, and within 20 generations, 15 percent of the speakers have lost Verb second completely. However, note that this is not sufficient to attain Modern French, and the change is too slow. In order to more closely duplicate the historically observed trajectory, we consider an initial condition consisting more like that actually found: a mix of Old French and data from Modern French (reproducing the intrusion of foreign speakers and reproducing data similar to that obtained from the Middle French period, see Clark and Roberts, 1993 for justification).

Given this new initial condition, fig. 6 shows the proportion of speakers losing Verb second after *one* generation as a function of the proportion of sentences from the "foreign" Modern French source. Surprisingly small proportions of Modern French cause a disproportionate number of speakers to lose Verb second, corresponding closely to the historically observed rapid change.

7 Conclusions

Learning theory attempts to account for how individual children solve the problem of language acquisition. By considering a *population* of such *individual* "child" learners, we arrive at a model of *emergent*, global, population language behavior. Consequently, whenever a linguist proposes a new grammatical or learning theory, they are also implicitly proposing a particular theory of language change, one whose consequences need to be examined. In particular, we saw the gain of Verb second in the 3-parameter case did not match historically observed patterns, but the 5-parameter system did. In this way the dynamical systems model supports the 5-parameter linguistic system to explain some changes in French. We have also greatly sharpened the informal notions of the time course of linguistic change and grammatical stability, indeed, showing that the rich results of population biology theory can be directly drawn on to study language change. Such evolutionary systems are, we believe, useful for testing grammatical theories and explicitly modeling historical language change.

While the computational study of language acquisition has become well established, the computational study of language change has not been as far advanced. Our aim here is to take a step in this direction and arrive at a

research program for the computational study of language change. Such a research program requires that one fix (a) the relevant components of the grammatical theory that capture the variations across languages; (b) a computational account of language acquisition; and (c) the relevant historical data that is to be captured by the evolutionary theory.

In deriving the evolutionary consequences in this paper, several simplifying assumptions were made. First, it was assumed that all children in a community receive example sentences drawn from the same linguistic distribution. This ignores (geographic and cultural) neighborhood effects, but in a way that may be easily remedied in the future. For example, although the total adult population might be half Spanish speaking and half English speaking, the speakers might live in neighborhoods that are entirely Spanish and entirely English speaking. The children living in these neighborhoods are exposed to different distributions of primary linguistic data from the Spanish and English sources. A second simplifying assumption made was that of non-overlapping generations. In other words, the entire population was divided into adults and children. Children received primary linguistic data only from adults and not from other children. Such a clean generational division of sources is not strictly true in practice. Again, this more complex mathematical case of age structured populations could covered by existing results from population biology. Finally, we have not entertained the possibility of children acquiring more than one grammar simultaneously and the consequences of that for language change. Finally, we assumed that the analog of "selection" in biological models was the identity mapping; this could be altered in obvious ways to accommodate models of "least effort" principles; cultural and sociological change, and the like.

In short, as with all mathematical modeling, especially initial steps like the one presented here, we have made certain simplifying assumptions in order to highlight the basic properties of the dynamical logic: the move from individuals to population thinking in language. In the best case, these simplifying assumptions themselves can, and will be, systematically altered as our principled approach to studying language change advances.

Acknowledgements

Thanks to Morris Halle and David Lightfoot for useful discussions. This research was performed at the Center for Biological and Computational Learning at MIT. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC–9217041.

References

Bickerton, D. (1981). *Roots of Language*. Ann Arbor: Karoma Press.

Bickerton, D. (1990). *Language and Species*. Chicago: University of Chicago Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht, Netherlands: Foris Publications.

R. Clark and I. Roberts. (1993) A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.

E. Gibson and K. Wexler. (1994) Triggers. Linguistic Inquiry, *25(4)*, 407-454.

Anthony S. Kroch. (1990) Reflexes of grammar in patterns of language change. *Language Variation and Change*, pages 199–243.

Lewontin, Richard. (1978) Adaptation and Evolutionary Theory. From *Studies in the History and Philosophy of Science*, 9:3, 181–206.

D. Lightfoot. (1991) *How to Set Parameters*. MIT Press, Cambridge, MA.

P. Niyogi. (1994) *The Informational Complexity of Learning From Examples*. Ph.D. thesis. Massachusetts Institute of Technology, Cambridge, MA.

P. Niyogi and R. C. Berwick. (1994) A Markov Model for Finite Parameter Spaces. *Proc. of the 32nd ACL Conference*, Las Cruces, New Mexico.

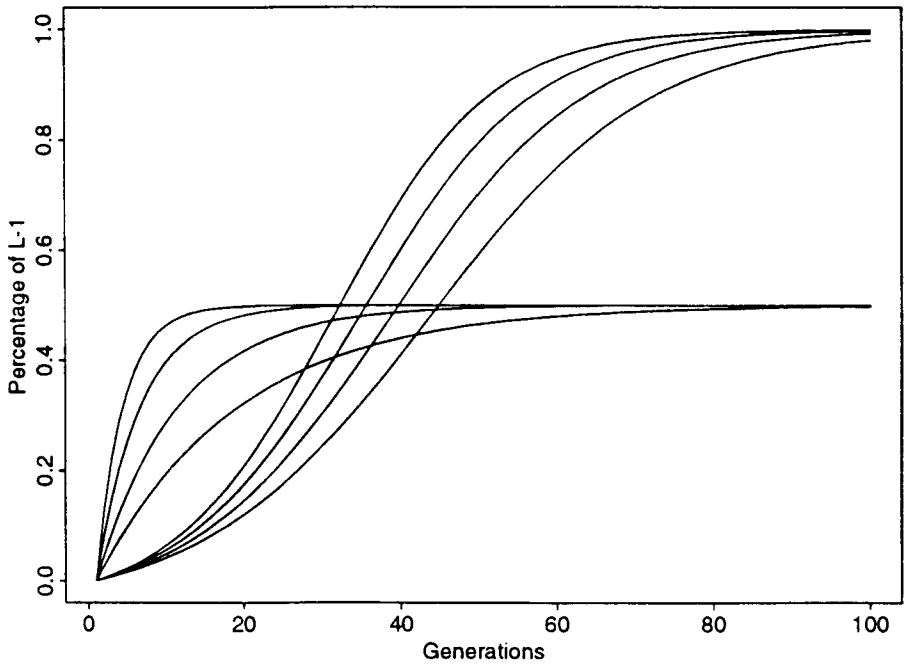

Figure 1: Evolution of linguistic populations whose speakers differ only in the $V2$ parameter setting. This reduces to a one parameter model as discussed. Note the exponential growth when $a = b$. The different exponential curves are obtained by varying the value $a = b$. When a is not equal to b, the system has a qualitatively different (logistic) growth. By varying the values of a and b we get the different logistic curves.

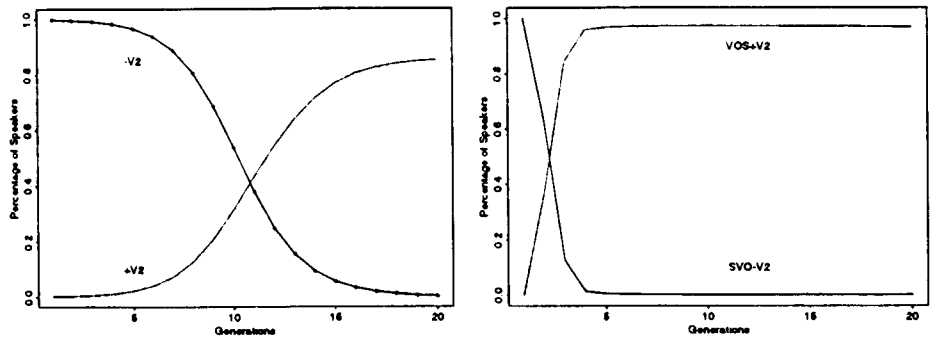

Figure 2: **Left:** Percentage of the population speaking languages of the basic forms V(erb) O(bject) S(ubject) with and without Verb second. The evolution has been shown upto 20 generations, as the proportions do not vary significantly thereafter. **Right:** Percentage of the population speaking languages S V O −Verb second (English) and V O S (+Verb second) as it evolves over the number of generations. Notice the sudden shift over a space of 3-4 generations.

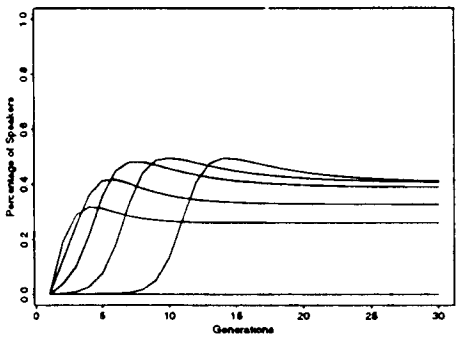

Figure 3: Time evolution of linguistic composition for the situations where the learning algorithm used is gradient ascent. Only the percentage of people speaking V(erb) O(bject) S(ubject) (+Verb second) is shown. The initial population is homogeneous and speaks V O S (−V2). The maturational time (number of sentences the child hears before internalizing a grammar) is varied through 8, 16, 32, 64, 128, 256, giving rise to six curves. The curve with the highest initial rate of change corresponds to the situation where only 8 examples were allowed to the learner to develop its mature hypothesis. The initial rate of change decreases as the maturation time N increases.

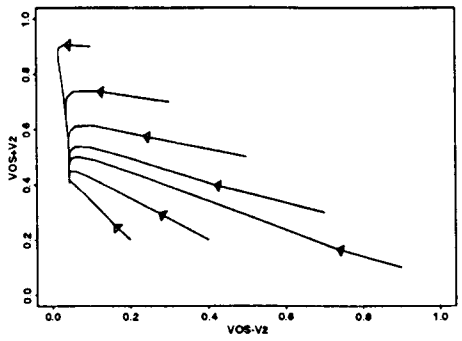

Figure 4: Subspace of a Phase-space plot. The plot shows the number of speakers of V(erb) O(bject) S(ubject) ($-$Verb second and $+$Verb second) as t varies. The learning algorithm was single step, gradient ascent. The different curves correspond to grammatical trajectories for different initial conditions.

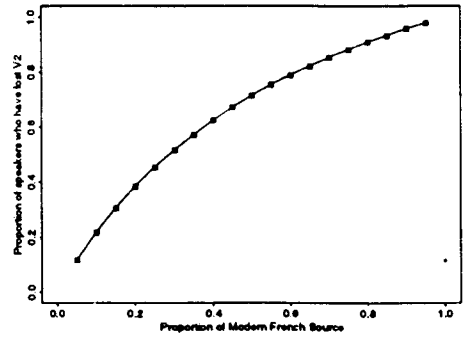

Figure 5: Tendency to lose Verb second as a result of new word orders introduced by Modern French sources in the dynamical systems model.