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 2, 1997

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 has sometimes been done by drawing on assumptions from population biology (cf. Cavalli-Sforza and Feldman, 1973; 1981; Kroch, 1990), this new model dispenses with 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 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 and maturational times, showing how different learning theories can have very different evolutionary consequences. This allows us to formulate an evolutionary criterion for the adequacy of grammatical and learning theories. An application of the computational model to the historical loss of Verb Second from Old French to Modern French is described showing the how otherwise adequate grammatical theories might fail the 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 the 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 to one side for now 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.[3]

[1] We leave aside the principle of differential fitness for now, though it might be easily accommodated in the mathematical modeling that follows. For 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 in phonology, 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 latter being the organism's "external form" on which selection actually acts). In no case except the most trivial is this mapping known; certainly not even for the simplest organism in full.

[3] Indeed, in what as far as we know is the earliest and only fully mathematical treatment of a framework similar to this one, Cavalli-Sforza and Feldman (1973, 1981) explicitly ground a general model for cultural change on the Mendelian model for genetic inheritance: if we assume that L is the 'language' of one parent (or group) and l that of the other, corresponding to the usual genetic contribution of the two sexes, then in the case of one binary property ('trait'), Cavalli-Sforza and Feldman note that there are, as is conventional, four possible inherited 'genotypes': LL, Ll, lL and ll, passed on with the corresponding probabilities b_0, b_1, b_2, and b_3 that each genotype produces a surface trait

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.[4]

We can resolve this paradox by introducing an explicit, formal evolutionary 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 system model for this purpose 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 initial results:

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 over a diverse range. Note that learnability plays an essential role because it fixes the parameters of the resulting dynamical systems that model language change.

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). This analytical work complements and extends those who have used explicit computer *simulations* of diachronic systems, e.g., Hare and Elman's (1995) model for morphological change, based on a neural-network acquisition algorithm for learning and representing past tense morphemes. Unlike Hare and Elman's system, we distinguish between

of type L (so, e.g., in the *biological* case of type LL which corresponds to 'homozygous' or monolingual situation, b_0 is typically assumed to be 1, while b_3 is 0, since the homozygous genetic type ll cannot produce a surface trait of L). As we shall see, in our model the b's correspond to the convergence properties of the child's (or adult's) language acquisition algorithm. We point out some of the similarities and differences between our model and that of Cavalli-Sforza below, though a full account is beyond the scope of the current paper.

[4] Clearly, this condition holds literally only if the parental population is held fixed, e.g., no outside intervention, sampling effects caused by finite population size, and the like. We show how to introduce such effects later.

individual and population ensemble effects (they follow what amounts to a single indvidual over successive generations while we follow an ensemble of individuals) and derive more general, analytical results that can be applied to a wider variety of grammatical systems and learning algorithms, including, but not limited to, neural network representations or even 'simulated annealing'. It is also mathematically straightforward to model the situation in which a contact population (e.g., Scandanavian) speaking another language influences a base population (e.g., English) in the sense of 'second language acquisition' whereby the base population attempts to acquire the language of the contact population, possibly arriving at a bilingual state that has further diachronic effects. All this can be effortlessly modeled in the framework given here.[5]

3. In the simplest possible case of a 2-language (grammar) system differing by exactly 1 binary parameter, the model reduces to a quadratic map with the possibility of the usual chaotic properties. That such complexity can arise even in the simplest case suggests that formally modeling language change may be quite mathematically rich. However, we show that because the mappings rest on probabilities whose values must necessarily lie between 0 and 1, true chaotic behavior is not ever mathematically possible.

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.

[5] We take this 'bilingual' account to be roughly the import of Kroch's work on diachronic language change, in which speakers acquire and produce some distributional *mix* of two languages, though Kroch does not supply an explicit mathematical model in our sense. If our interpretation is accurate, then we have also provided a precise mathematization of Kroch's theory.

2 The Acquisition-Based Model of Language Change

To begin, 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, computing 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.,

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

Remark. This formulation is neutral with respect to grammatical theories and languages, as long as they can be enumerated in the usual way. Thus, e.g., generalized phrase structure grammar or even more radical departures from conventional generative approaches, like Hare and Elman's neural network representations, may be readily accomodated.

Remark. Similarly, the formulation is neutral with respect to acquisition algorithms. One may use a simple error-detection algorithm, as we do below; a genetic algorithm (as in Clark and Roberts, 1993); network-type learning (as in Hare and Elman, 1995); or simulated annealing.

Remark. In contrast with previous models of language change such as those of Clark and Roberts, Kroch, or Hare and Elman, this formulation makes explicit the difference between individuals and a population of individual speakers, a distinction that has ramifications for the dynamical systems as discussed in the sequel.

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

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

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

The learning algorithm $\mathcal{A}$ uses the linguistic data (n examples, denoted 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:

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

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, $P_{pop,t} \longrightarrow^{\mathcal{A}} P_{pop,t+1}$.

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.

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). As perhaps the simplest explicit model for a parameterized grammatical system, we 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 and gradient ascent (hill climbing) learning algorithm called the TLA (Triggering Learning Algorithm). We stress that we have adopted this particular learning algorithm because it is rather simple visualize. For a similar analysis that instead uses a maximum likelihood estimate model (pick the 'most likely grammar' given the evidence), see Niyogi and Berwick (1996).

For concreteness, we next 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). It is also essentially single-step hill climbing: the system tries to find a single parameter change that will let it parse the current sentence being analyzed. There are variants of the TLA that we will consider later in this article.

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 draw on the following results (the first is straightforward; see Niyogi, 1994; most of the others follow straightforwardly from Markov theory):

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.*

Of course, it is hardly surprising that a memoryless learning algorithm can be modeled by a first order Markov chain. The value 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. We again stress that if one adopts another learning algorithm, one can still carry out the dynamical system analysis, though the solution may become more difficult; for an example that uses a maximum likelihood estimator, see Niyogi and Berwick (1996). In the case of the TLA, however, we obtain:

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

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

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

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 system update rule. Thus, we obtain our required dynamical system for parameter-based theories and memoryless acquisition algorithms. The system 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 examine the simplest possible case in part by analogy with methodology in mathematical modeling generally. We shall see that even here the mathematics becomes nontrivial. Following the detailed analysis of the one parameter case, to pursue richer, more concrete examples, 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

We first take up the the following simple scenario for language change, perhaps the simplest one possible, where there are just two languages. We then progressively examine more complex cases.
$\mathcal{G}$: Imagine that due to universal grammatical constraints there are only two possible grammars parameterized by one boolean valued parameter, yielding two possible languages, 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 the following two values a and b, which give the distribution of sentences that will be produced common to both languages, given that the true target is either L_1 or L_2. These values are critical to the learning algorithm, since it is driven by the *set difference* between L_1 and L_2, or $1-a$ and $1-b$, respectively:[6]

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

[6] Under alternative learning algorithms, e.g., a maximum likelihood algorithm, these values may also be readily calculated, though with some greater effort.

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.
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 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 associated Markov 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 above matrices. 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 linguistic 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 a proportion $1-p$ of the population will speak L_2.

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

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 remarks concerning this dynamical system are in order:
Remark 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 We note that Cavalli-Sforza and Feldman (1981), using a different formulation, also obtain a quadratic map in such cases for the example of general 'vertical' cultural change.

Remark 2. The scenario $a \neq b$ is much more likely to occur in practice — consequently, we are more likely to see logistic change rather than exponential change.

Remark 3. We obtain 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.

Remark 4. Logistic maps are known to be chaotic. However, 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 last result naturally raises the question whether nonchaotic behavior holds true for all grammatical dynamical systems, specifically the linguistically "natural" cases or whether there are linguistic systems where chaos will manifest itself. It is suggestive that Cavalli-Sforza's Mendelian model obtains similar results when interpreted as acting within reasonable learnability constraints: that is, assuming that the 'transmission' probability for learning language L after being exposed to an adequate sample of L is high, say, 0.9 — that is, that the learning system usually succeeds — then chaotic regimes are impossible for a 1-parameter system. Further research on this subject is planned.

4.2 Other Choices of Maturation Time, N:

To isolate the effect of maturation time, let us now consider the case where the maturation time, N, is equal to 3 (or some other small integer value). All other assumptions remain the same, 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, with the order dependent on maturation time.[7] 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 an arbitrary amount of time (samples) to mature and attain its linguistic hypothesis. It is possible to prove:

[7] This result is more general than that of Cavalli-Sforza, who obtain only quadratic maps because they assume a Mendelian-type model without maturation time effects.

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

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

In this case of indefinitely long maturation time a number of observations are worthwhile. First, note that if $a = b$, we find that $p_{n+1} = p_n$ for all n. In other words, if the initial proportion of the two languages is equal, the population *never* changes its linguistic characteristics from generation to generation. However, if $a < b$ it can be shown that p_n tends to 1 as n tends to infinity from *all* initial conditions: all learners tend to wind up internalizing the first language L_1. This makes sense; recall that $1 - a$ measures the set difference between the two languages, and hence the 'ease' with which L_1 is picked out. If a is small relative to b, then $1 - a$ is is large relative to $1 - b$, and L_1 is easier to acquire than L_2. Consequently over generations we find that all speakers tend to acquire L_1. Symmetrically, the reverse is true when $a > b$; then p_n tends to 0 as n tends to infinity, that is, all learners eventually wind up speaking L_2.

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 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 Learning Algorithm, $\mathcal{A}$:

Let us now return to the situation where $N = 2$. As we discussed, one obtains a variety of quadratic maps in this case. How these quadratic maps vary with the parameters a, b depends upon the details of the learning algorithm. Here we provide a brief account of the effect of the learning algorithm on the evolution. Space prohibits a more detailed consideration.

Consider the following three variations in the learning algorithm: (1) **Algorithm 1:** Same as TLA except the initial hypothesis is always g_1; (2) **Algorithm 2:** Choose g_2 unless all examples are not analyzable by g_2, otherwise choose g_1; (3) **Algorithm 3:** Run Algorithm 2 and flip the result with probability η.

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. 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 — flipping between one language choice and the next — and in the limit, the populations are always homogeneous except that generations alternately speak L_1 and L_2. A more systematic investigation of the different evolutionary consequences of different language learning algorithms is beyond the scope of the current paper.

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). The aim here is to develop a miniature, but more realistic grammatical system. To this end, the parameters aim to describe some of the major properties of grammar, plus one that has been often implicated in language change.

Specifically, let us 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 Specifier parameter essentially says whether Subjects come before Verbs (generalized to other lexical categories); the Head parameter says whether Objects follow or precede verbs (postfix or prefix); and the Verb second parameter is one that has been found descriptively useful to account for changes among Germanic languages, and Germanic-Romance interactions. We take the learning algorithm to be 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.** We 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 growth); but sometimes not, as in fig. 5.[8]

4. **In many cases a homogeneous population splits into stable linguistic groups.** As shown in figure 5, the dynamical system's (stable) fixed point may be a *mix* of two different languages, despite initial homegeneity. This is a joint property of the parameter space and the learning algorithm.

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

6. **Different initial population mixes lead to phase space plots with different 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, con-

[8] As remarked earlier, it is possible to extend the maturation time notion to incorporate a 'bilingual' model for language change, as seemingly suggested by Kroch. The point about the logistic still stands.

sider 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 summarize *all* the trajectories corresponding to different initial conditions, exhibited in fig. 5.

It remains to precisely characterize the fixed points in such settings, for different parameterizations.

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 (empty) 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 patterns 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, foreign 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 that the 3-parameter system's behavior did not match historically observed patterns with respect to the loss of Verb second, but the 5-parameter system did. In this way the dynamical system model supports the 5-parameter linguistic system to explain some diachronic changes in French. We have also greatly sharpened the informal notions of the time course of linguistic change and grammatical stability, 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 article, 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 mathematically. For example, although the total adult population might be half Spanish speaking (L) and half English speaking (l), 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 straightforward modification is to assume there are three possible language source types that learners are exposed to: LL, corresponding to a pure Spanish source; Ll, corresponding to some 'contact' distribution over Spanish and English; and ll, corresponding to a pure English source. In this simplest case, we again obtain a quadratic map; we omit the details here.

A second simplifying assumption made was that of nonoverlapping 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 be 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. 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, again along the lines advanced by Cavalli-Sforza and Feldman (1981) (though our model is different from theirs in the sense of explicitly including a maturational learning component, and exploring concrete linguistic parameterizations).

Finally, 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, David Lightfoot, and Dick Lewontin 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.

Cavalli-Sforza, L. and M. Feldman. (1973) Cultural versus biological inheritance: Phenotypic transmission from parent to children. *American Journal of Human Genetics*, *25*, pages 618–637.

Cavalli-Sforza, L. and M. Feldman. (1981) *Cultural Transmission and Evolution: a Quantitative Approach*, Princeton University Press, Princeton, New Jersey.

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.

Hare, M., and J.Elman (1995).Learning and morphological change. *Cognition*, 56, pages 61–98.

Λ**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.

P. Niyogi and R. C. Berwick. (1996) Populations of learners: the case of Portuguese language change. Paper presented at the Brazilian workshop on language change, University of Sao Paolo, Sao Sebastio, Brazil, August 1996.

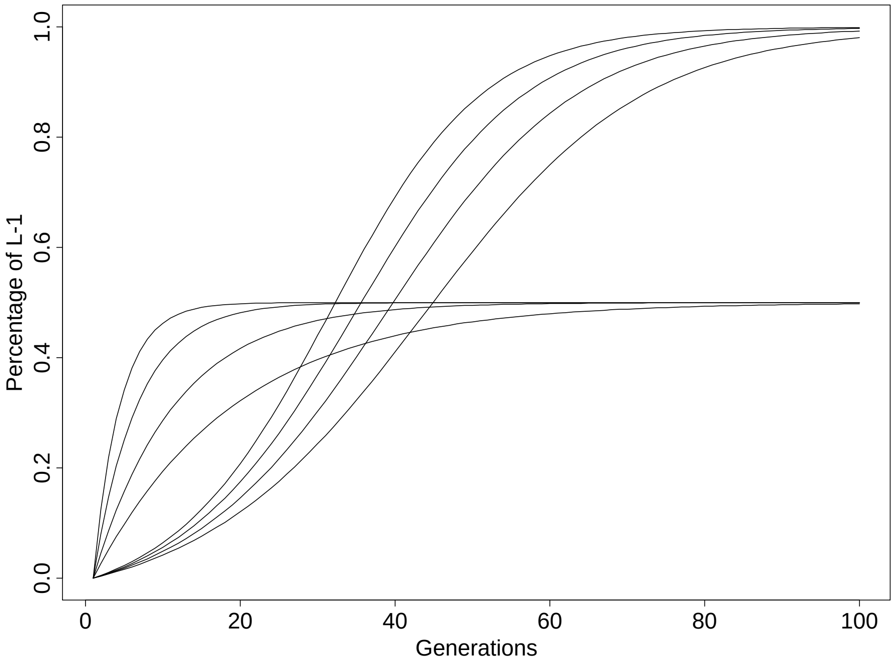

Figure 1: Evolution of linguistic populations whose speakers differ only in the $V2$ parameter setting. This reduces to a one parameter model as discussed. Note the exponential growth when $a = b$. The different exponential curves are obtained by varying the value $a = b$. When a is not equal to b, the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

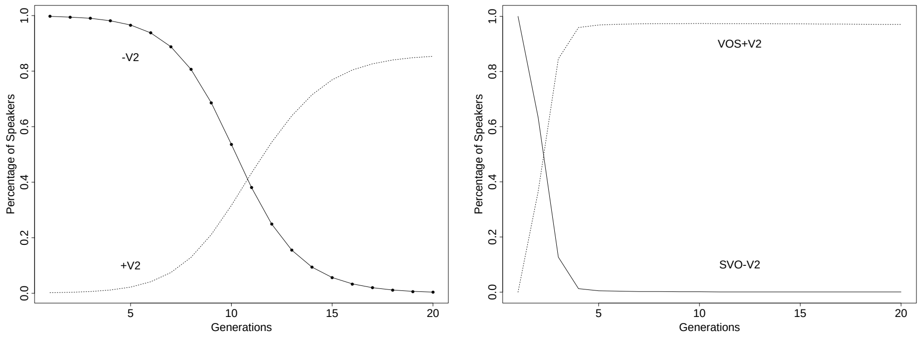

Figure 2: **Left:** Percentage of the population speaking languages of the basic forms V(erb) O(bject) S(ubject) with and without Verb second. The evolution has been shown upto 20 generations, as the proportions do not vary significantly thereafter. **Right:** Percentage of the population speaking languages 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.

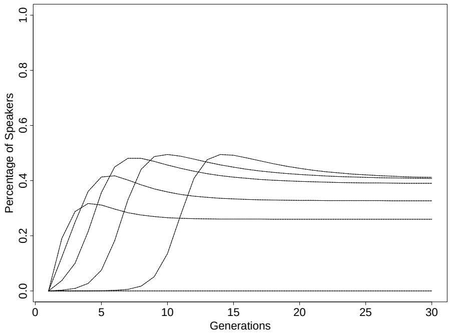

Figure 3: Time evolution of linguistic composition for the situations where the learning algorithm used is gradient ascent. Only the percentage of people speaking V(erb) O(bject) S(ubject) (+Verb second) is shown. The initial population is homogeneous and speaks V O S (−V2). The maturational time (number 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.

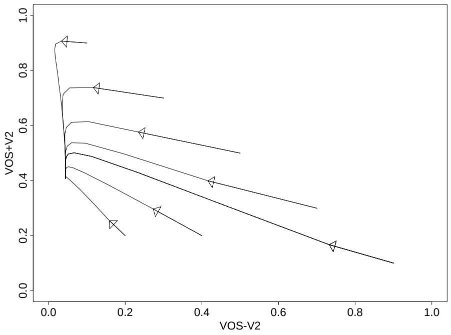

Figure 4: Subspace of a Phase-space plot. The plot shows the number of speakers of V(erb) O(bject) S(ubject) (−Verb second and +Verb second) as t varies. The learning algorithm was single step, gradient ascent. The different curves correspond to grammatical trajectories for different initial conditions.

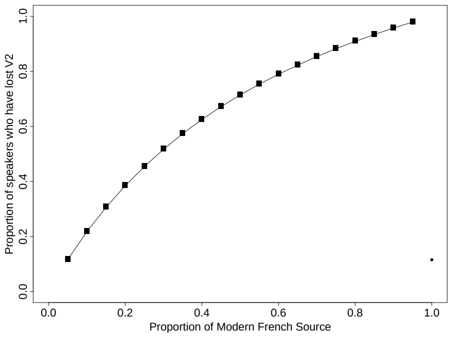

Figure 5: Tendency to lose Verb second as a result of new word orders introduced by Modern French sources in the dynamical systems model.