MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1516
C.B.C.L. Paper No. 115
July, 1995

The Logical Problem of Language Change

Partha Niyogi and Robert C. Berwick

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

This paper considers the problem of language change. Linguists must explain not only how languages are learned but also how and why they have evolved along certain trajectories and not others. While the language learning problem has focused on the behavior of individuals and how they acquire a particular grammar from a class of grammars $\mathcal{G}$, here we consider a *population* of such learners and investigate the emergent, global population characteristics of linguistic communities over several generations. We argue that language change follows logically from specific assumptions about grammatical theories and learning paradigms. In particular, we are able to transform parameterized theories and memoryless acquisition algorithms into grammatical dynamical systems, whose evolution depicts a population's evolving linguistic composition. We investigate the linguistic and computational consequences of this model, showing that the formalization allows one to ask questions about diachronic that one otherwise could not ask, such as the effect of varying initial conditions on the resulting diachronic trajectories. From a more programmatic perspective, we give an example of how the dynamical system model for language change can serve as a way to distinguish among alternative grammatical theories, introducing a formal diachronic adequacy criterion for linguistic theories.

This report describes research done at the Center for Biological and Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC–9217041. Robert C. Berwick was supported by a Vinton-Hayes Fellowship.

1 Introduction

As is well known, languages change over time. Language scientists have long been occupied with describing phonological, syntactic, and semantic change, often appealing to the analogy between language change and evolution. Some even suggest that language itself is a complex adaptive system (see Hawkins and Gell-Mann, 1989). For example, 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...In population biology and linguistic change there is constant flux..... If one views a language as a totality, as historians often do, one sees a *dynamic system*." Indeed, entire books have been devoted to the description of language change using the terminology of population biology: genetic drift, clines, and so forth[1] However, these analogies have rarely been pursued beyond casual and descriptive accounts.[2] In this paper we formalize these intuitions, to the best of our knowledge for first time, as a concrete, computational, dynamical systems model, and investigating the consequences of this formalization.

In particular, we show that a model of language change emerges as a logical consequence of language acquisition, a point made by Lightfoot (1991). We shall see that Lightfoot's intuition that languages could behave just as though they were dynamical systems is essentially correct, as is his proposal for turning language acquisition models into language change models. We can provide concrete examples of both "gradual" and "sudden" syntactic changes, occurring over time periods of many generations to just a single generation.[3]

Many interesting points emerge from the formalization, some programmatic:

- *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.
- We derive explicit dynamical systems corresponding to parametrized linguistic theories (e.g., the Head First/Final parameter in head-driven phrase structure grammars or government-binding grammars) and memoryless language learning algorithms (e.g., gradient ascent in parameter space).
- 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 in the literature does not seem to permit this historical change, while another does. We can more accurately model the time course of language change. In particular, in contrast to Kroch (1989) and others, who mimic population biology models by imposing S-shaped logistic curves on possible language changes by assumption, we derive the time course of language change from more basic assumptions, and show that it need not be S-shaped; rather, an S-shape can emerge from more fundamental properties of the underlying dynamical system.
- We examine by simulation and traditional phase-space plots the form and stability of possible "diachronic envelopes" given varying alternative language distributions, language acquisition algorithms, parameterizations, input noise, and sentence distributions. The results bear on models of language "mixing"; so-called "wave" models for language change; and other proposals in the diachronic literature.
- As topics for future research, the dynamical system model provides a novel possible source for explaining several linguistic changes including: (a) the evolution of modern Greek metrical stress assignment from proto-Indo-European; and (b) Bickerton's (1990) "creole hypothesis," concerning the striking fact that all creoles, irrespective of linguistic origin, have exactly the same grammar. In the latter case, the "universality" of creoles could be due a parameterization corresponding to a common condensation point of a dynamical system, a possibility not considered by Bickerton.

2 An Acquisition-Based Model of Language Change

How does the combination of a grammatical theory and learning algorithm lead to a model of language change? We first note that just as with language acquisition, there is a seeming paradox in language change: it is generally assumed that children acquire their caretaker (target) grammars without error. However, if this were always true, at first glance grammatical changes within a population could seemingly never occur, since generation after generation children would successfully acquire the grammar of their parents.

Of course, Lightfoot and others have pointed out the obvious solution to this paradox: the possibility of slight misconvergence to target grammars could, over several generations, drive language change, much as speciation occurs in the population biology sense:

> As somebody adopts a new parameter setting, say a new verb-object order, the output of that person's grammar often differs from that of other people's. This in turn affects the linguistic environment, which may then be more likely to trigger the new parameter setting in younger people. Thus a chain reaction may be created. (Lightfoot, 1991, p. xxx)

We pursue this point in detail below. Similarly, just

[1] For a recent example, see Nichols (1992), *Linguistic Diversity in Space and Time*.

[2] Some notable exceptions are Kroch (1990) and Clark and Roberts (1993).

[3] Lightfoot 1991 refers to these sudden changes acting over a single generation as "catastrophic" but in fact this term usually has a different sense in the dynamical systems literature.

as in the biological case, some of the most commonly observed changes in languages seem to occur as the result of the effects of surrounding populations, whose features infiltrate the original language.

We begin our treatment by arguing that the problem of language acquisition at the individual level leads logically to the problem of language change at the group or population level. Consider a population speaking a particular language[4]. This is the target language—children are exposed to primary linguistic data (PLD) from this source, typically in the form of sentences uttered by caretakers (adults). The logical problem of language acquisition is how children acquire this target language from their primary linguistic data—to come up with an adequate learning theory. We take a learning theory to be simply a mapping from primary linguistic data to the class of grammars, usually effective, and so an algorithm. For example, in a typical inductive inference model, given a stream of sentences, an acquisition algorithm would simply update its grammatical hypothesis with each new sentence according to some preprogrammed procedure. An important criterion for learnability (Gold, 1967) is to require that the algorithm converge to the target as the data goes to infinity (identification in the limit).

Now suppose that we fix an adequate grammatical theory and an adequate acquisition algorithm. There are then essentially two means by which the linguistic composition of the population could change over time. First, if the primary linguistic data presented to the child is altered (due to any number of causes, perhaps to presence of foreign speakers, contact with another population, disfluencies, and the like), the sentences presented to the learner (child) are no longer consistent with a single target grammar. In the face of this input, the learning algorithm might no longer converge to the target grammar. Indeed, it might converge to some other grammar (g_2); or it might converge to g_2 with some probability, g_3 with some other probability, and so forth. In either case, children attempting to solve the acquisition problem using the same learning algorithm could internalize grammars different from the parental (target) grammar. In this way, in one generation the linguistic composition of the population can change.[5]

Second, even if the PLD comes from a single target grammar, the actual data presented to the learner is truncated, or finite. After a finite sample sequence, children may, with non-zero probability, hypothesize a grammar different from that of their parents. This can again lead to a differing linguistic composition in succeeding generations.

In short, the diachronic model is this: Individual children 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 successive generations, the linguistic composition evolves as a dynamical system.

On this view, language change is a logical consequence of specific assumptions about:

1. the *grammar hypothesis space*—a particular parametrization, in a parametric theory;
2. the *language acquistion device*—the learning algorithm the child uses to develop hypotheses on the basis of data;
3. the *primary linguistic data*—the sentences presented to the children of any one generation.

If we specify (1) through (3) for a particular generation, we should, in principle, be able to compute the linguistic composition for the next generation. In this manner, we can compute the evolving linguistic composition of the population from generation to generation; we arrive at a dynamical system. We now proceed to make this calculation precise. We first review a standard language acquisition framework, and then show how to derive a dynamical system from it.

2.1 The Language Acquisition Framework

Let us state our assumptions about grammatical theories, learning algorithms, and sentence distributions.

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

2. Denote by P a distribution on Σ^* according to which sentences are drawn and presented to the learner. Note that if there is a well defined target, g_t, and only positive examples from this target are presented to the learner, then P will have all its measure on $L(g_t)$, and zero measure on sentences outside Suppose n examples are drawn in this fashion, one can then let $\mathcal{D}_n = (\Sigma^*)^n$ be the set of all n-example data sets the learner might be presented with. 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 acquisition algorithm that children use to hypothesize a grammar on the basis of input data. $\mathcal{A}$ can be regarded as a mapping from $\mathcal{D}_n$ to $\mathcal{G}$. Thus, acting upon a particular presentation sequence $d_n \in \mathcal{D}_n$, the learner posits a hypothesis $\mathcal{A}(d_n) = h_n \in \mathcal{G}$. Allowing for the possibility of randomization, the learner could, in general, posit $h_i \in \mathcal{G}$ with probability p_i for such a presentation sequence d_n. The standard (stochastic version) learnability criterion (Gold, 1967) can then be stated as follows:

For every target grammar, $g_t \in \mathcal{G}$, with positive-only examples presented according to P as above, the learner must converge to the target with probability 1, i.e.,

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

[4]In our analysis this implies that all the adult members of this population have internalized the same grammar (corresponding to the language they speak).

[5]Sociological factors affecting language change, affect language acquisition in exactly the same way, yet are abstracted away from the formalization of the logical problem of language acquisition. In this same sense, we similarly abstract away such causes here.

For an analysis of learnability issues for memoryless algorithms in finite parameter spaces, consult Niyogi (1995) .

2.2 From Language Learning to Popuation Dynamics

The framework for language learning has learners attempting to infer grammars on the basis of linguistic data. At any point in time, n, (i.e., after hearing n examples) the learner has a current hypothesis, h, with probability $p_n(h)$. What happens when there is a *population* of learners? Since an arbitrary learner has a probability $p_n(h)$ of developing hypothesis h (for every $h \in \mathcal{G}$), it follows that a fraction $p_n(h)$ of the population of learners internalize the grammar h after n examples. We therefore have a *current state* of the population after n examples. This state of the population might well be different from the state of the parent population. Assume for now that after n examples, maturation occurs, i.e., after n examples the learner retains the grammatical hypothesis for the rest of its life. Then one would arrive at the state of the mature population for the next generation.[6] This new generation now produces sentences for the following generation of learners according to the distribution of grammars in its population. Then, the process repeats itself and the linguistic composition of the population evolves from generation to generation.

We can now define a discrete time dynamical system by providing its two necessary components:

A State Space: a set of system states, $\mathcal{S}$. Here the state space is the space of possible linguistic compositions of the population. Each state is described by a distribution P_{pop} on $\mathcal{G}$ describing the language spoken by the population.[7] At any given point in time, t, the system is in exactly one state $s \in \mathcal{S}$;

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}[8]

For example, a typical linear dynamical system might consist of state variables $\mathbf{x}$ (where $\mathbf{x}$ is a k-dimensional state vector) and a system of differential equations $\mathbf{x}' = Ax$ (A is a matrix operator) which characterize the evolution of the states with time. RC circuits are a simple example of linear dynamical systems. The state (current) evolves as the capcitor discharges through the resistor. Population growth models (for example, using logistic equations) provide other examples.

[6]Maturation seems to be a reasonable hypothesis in this context. After all, it seems even more unreasonable to imagine that learners are forever wandering around in hypothesis space. There is evidence from developmental psychology to suggest that this is the case, and that after a certain point children mature and retain their current grammatical hypotheses forever.

[7]As usual, one needs to be able to define a σ-algebra on the space of grammars, and so on. This is unproblematic for the cases considered in this paper because the set of grammars is finite.

[8]In general, this mapping could be fairly complicated. For example, it could depend on previous states, future states, and so forth; for reasons of space we do not consider all possibilities here. For reference, see Strogatz, 1993.

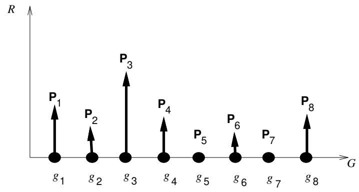

Figure 1: A simple illustration of the state space for the 3-parameter syntactic case. There are 8 grammars. A probability distribution on these 8 grammars, as shown above, can be interpreted as the linguistic composition of the population. Thus, a fraction P_1 of the population have internalized grammar, g_1, and so on.

As as linguistic example, consider the three parameter syntactic space described in Gibson and Wexler (1994). This defines 8 possible "natural" grammars. Thus $\mathcal{G}$ has 8 elements. We can picture a distribution on this space as shown in fig. 1. In this particular case, the state space is

$$\mathcal{S} = \{\mathbf{P} \in R^8 | \sum_{i=1}^{8} \mathbf{P}_i = 1\}$$

Here we interpret the *state* as the linguistic composition of the population.[9] For example, a distribution that puts all its weight on grammar g_1 and 0 everywhere else indicates a homogeneous population that speaks a language corresponding to grammar g_1. Similarly, a distribution that puts a probability mass of 1/2 on g_1 and 1/2 on g_2 denotes a population (nonhomogeneous) with half its speakers speaking a language corresponding to g_1 and half speaking a language corresponding to g_2.

To see in detail how the update rule may be computed, consider the acquisition algorithm, $\mathcal{A}$. For example, given the state at time t, $(P_{pop,t})$, the distribution of speakers in the parental population, one can obtain the distribution with which sentences from $\Sigma*$ will be presented to the learner. To do this, imagine that the i^{th} linguistic group in the population, speaking language L_i, produces sentences with distribution P_i. Then for any $\omega \in \Sigma^*$, the probability with which ω is presented to the learner is given by

$$P(\omega) = \sum_i P_i(\omega) P_{pop,t}(i)$$

This fixes the distribution with which sentences are presented to the learner. The logical problem of language acquisition also assumes some success criterion for attaining the mature target grammar. For our purposes, we take this as being one of two broad possibilities: either (1) the usual Gold scenario of identification in the limit, what we shall call the *limiting sample* case; or (2)

[9]Note that we do not allow for the possibility of a *single* learner having more than one hypothesis at a time; an extension to this case, in which individuals would more closely resemble the "ensembles" of particles in a thermodynamic system is left for future research.

identification in a fixed, finite time, what we shall call the *finite sample* case.[10]

Consider case (2) first. Here, one draws n example sentences according to distribution P, and the acquisition algorithm develops hypotheses ($\mathcal{A}(d_n) \in \mathcal{G}$). One can, in principle, compute the probability with which the learner will posit hypothesis h_i after n examples:

$$\textbf{Finite Sample: } Prob[\mathcal{A}(d_n) = h_i] = p_n(h_i) \qquad (1)$$

The finite sample situation is always well defined—the probability p_n always exists.[11].

Now turn to case (1), the limiting case. Here learnability requires $p_n(g_t)$ to go to 1, for the unique target grammar, g_t, if such a grammar exists. However, in general there need not be a unique target grammar since the linguistic population can be nonhomogeneous. Even so, the following limiting behavior might still exist:

$$\textbf{Limiting Sample: } \lim_{n\to\infty} Prob[\mathcal{A}(d_n) = h_i] = p(h_i) \qquad (2)$$

Turning from the individual child to the population, since the individual child internalizes grammar $h_i \in \mathcal{G}$ with probability $p_n(h_i)$ in the "finite sample" case or with probability $p(h_i)$ "in the limit", in a *population* of such individuals one would therefore expect a proportion $p_n(h_i)$ or $p(h_i)$ respectively to have internalized grammar h_i. In other words, the linguistic composition of the *next* generation is given by $P_{pop,t+1}(h_i) = p_n(h_i)$ for the finite sample case and by $P_{pop,t+1}(h_i) = p(h_i)$ in the limiting sample case . In this fashion,

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

Remarks. 1. For a Gold-learnable family of languages and a limiting sample assumption, homogeneous populations are always stable. This is simply because each child and therefore the entire population always eventually converges to a single target grammar, generation after generation.

2. However, finite sample case is different from the limiting sample case. Suppose we have solved the maturation problem, that is, we know roughly the time, or number of examples N the learner takes to develop its mature (adult) hypothesis. In that case $p_N(h)$ is the probability that a child internalizes the grammar h, and $p_N(h)$ is the percentage of speakers of L_h in the next generation. Note that under this finite sample analysis, even for a homogeneous population with all adults speaking a particular language (corresponding to grammar, g, say), $p_N(g)$ will not be 1—that is, there will be a small percentage of learners who have misconverged. This percentage could blow up over several generations, and we therefore have potentially unstable languages.

3. The formulation is very general. Any $\{\mathcal{A}, \mathcal{G}, \mathcal{P}_{\}}\}$ triple yields a dynamical system.[12]. In short:

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

4. The formulation also does not assume any particular linguistic theory, learning algorithm, or distribution with which sentences are drawn. Of course, we have implicitly assumed a learning model, i.e., positive examples are drawn in i.i.d. fashion and presented to the learner. Our dynamical systems formalization follows as a logical consequence of this learning framework. One can conceivably imagine other learning frameworks—these would potentially give rise to other kinds of dynamical systems—but we do not formalize them here.

This completes the abstract formulation of the dynamical system model. Next, we choose specific linguistic theories and learning paradigms to model particular kinds of language changes, with the goal of answering the following questions:

- Can we really compute all the relevant quantities to specify the dynamical system?
- Can we evaluate the behavior (phase-space characteristics) of the resulting dynamical system?
- Does the dynamical system model—the formalization—shed light on diachronic models and linguistic theories generally?

In the remainder of this paper, we give some concrete answers to these questions within the principles and parameters theory of modern linguistics.

3 Language Change in Parametric Systems

In previous works (Niyogi and Berwick, 1993, 1994a, 1994b; Niyogi, 1995), we investigated the problem of learnability within parametric systems. In particular, we showed that the behavior of any memoryless algorithm can be modeled as a Markov chain. This analysis allows us to solve equations 1 and 2, and thus obtain the update equations of the associated dynamical system. Let us now show how to derive such models in detail. We first provide the particular $\mathcal{G}, \mathcal{A}, \{P_i\}$ triple, and then give the update rule.

The learning system triple.

1. $\mathcal{G}$: Assume there are n parameters—this leads to a space $\mathcal{G}$ with 2^n different grammars.
2. $\mathcal{A}$: Let us imagine that the child learner follows some memoryless (incremental) algorithm to set parameters. For the most part, we will assume that

[10]Of course, a variety of other success criteria, e.g., convergence within some epsilon, or polynomial in the size of the target grammar, are possible; each leads to potentially different language change model. We do not pursue these alternatives here.

[11]This is easy to see for deterministic algorithms, $\mathcal{A}_{det}$. Such an algorithm would have a precise behavior for every data set of n examples drawn. In our case, the examples are drawn in i.i.d. fashion according to a distribution P on Σ^*. It is clear that $p_n(h_i) = P[\{d_n|\mathcal{A}_{det}(d_n) = h_i\}]$. For randomized algorithms, the case is trickier, though tedious, but the probability still exists because all the finite choice paths over all sequences of length n is enumerable. Previous work (Niyogi and Berwick, 1993,1994a,1994b) shows how to compute p_n for randomized memoryless algorithms.

[12]Note that this probability could evolve with generations as well. That will complete all the logical possibilites. However, for simplicity, we assume that this does not happen.

the algorithm is the "triggering learning algorithm" or TLA (the single step, gradient-ascent algorithm of Gibson and Wexler, 1994) or one of the variants discussed in Niyogi and Berwick (1993).

3. $\{P_i\}$: Let speakers of the ith language, L_i, in the population produce sentences according to the distribution P_i. For the most part we will assume in our simulations that this distribution is uniform on degree-0 (unembedded) sentences, exactly as in the learnability analysis of Gibson and Wexler 1994 or Niyogi and Berwick 1993.

The update rule. We can now compute the update rule associated with this triple. Suppose the state of the parental population is $P_{pop,n}$ on $\mathcal{G}$. Then one can obtain the distribution P on the sentences of Σ^* according to which sentences will be presented to the learner. Once such a distribution is obtained, then given the Markov equivalence established earlier, we can compute the transition matrix T according to which the learner updates its hypotheses with each new sentence. From T one can finally compute the following quantities, one for the "finite sample" case and one for the "limiting sample" case:

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

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

Similary, making use of the limiting distributions of Markov chains (Resnick, 1992) one can obtain the following (where ONE is a $\frac{1}{2^n} \times \frac{1}{2^n}$ matrix with all ones).

$$Prob[\text{ Learner's hypothesis } = h_i \text{"in the limit"}]$$

$$= (1,\ldots,1)'(I - T + ONE)^{-1}$$

These expressions allow us to compute the linguistic composition of the population from one generation to the next according to our analysis of the previous section.

Remark. The limiting distribution case is more complex than the finite sample case and requires some careful explanation. There are two possibilities. If there is just a single target grammar, then, by definition, the learners all identify the target correctly in the limit, and there is no further change in the linguistic composition from generation to generation. This case is essentially uninteresting. If there are two or more target grammars, then recalling our analysis of learnability (Niyogi and Berwick, 1994), there can be no absorbing states in the Markov chain corresponding to the parametric grammar family. In this situation, a single learner will oscillate between some set of states in the limit. In this sense, learners will not converge to any single, correct target grammar. However, there is a sense in which we can characterize limiting behavior for learners: although a given learner will visit each of these states infinitely often in the limit, it will visit some more often than others. The exact percentage the learner will be in a particular state is given by equation 3 above. Therefore, since we know the fraction of the time the learner spends in each grammatical state in the limit, we assume that this is the probability with which it internalizes the grammar corresponding to that state in the Markov chain.

Summarizing, we provide the basic computational framework for modeling language change:

1. Let $\pi_{\mathbf{1}}$ be the initial population mix, i.e., the percentage of different language speakers in the community. Assuming that the i^{th} group of speakers produces sentences with probability P_i, we can obtain the probability P with which sentences in Σ^* occur for the next generation of learners.
2. From P we can obtain the transition matrix T for the Markov learning model and the limiting distribution of the linguistic composition $\pi_{\mathbf{2}}$ for the next generation.
3. The second generation now has a population mix of $\pi_{\mathbf{2}}$. We repeat step 1 and obtain $\pi_{\mathbf{3}}$. Continuing in this fashion, in general we can obtain $\pi_{\mathbf{i+1}}$ from $\pi_{\mathbf{i}}$.

We next turn to specific applications of this model. We begin with a simple 3-parameter system as our first example, considering variations on the learning algorithm, sentence distributions, and sample size available for learning. We then consider a different, 5-parameter system already presented in the literature (Clark and Roberts, 1993) as one intended to partially characterize the change from Old French to Modern French.

4 Example 1: A Three Parameter System

The previous section developed the necessary mathematical and computational tools to completely specify the dynamical systems corresponding to memoryless algorithms operating on finite parameter spaces. In this example we investigate the behavior of these dynamical systems. Recall that every choice of $(\mathcal{G}, \mathcal{A}, \{P_i\})$ gives rise to a unique dynamical system. We start by making specific choices for these three elements:

1. $\mathcal{G}$: This is a 3-parameter syntactic subsystem described in Gibson and Wexler (1994). Thus $\mathcal{G}$ has exactly 8 grammars, generating languages from L_1 through L_8, as shown in the appendix of this paper (taken from Gibson and Wexler, 1994).
2. $\mathcal{A}$: The memoryless algorithms we consider are the TLA, and variants by dropping either or both of the single-valued and greediness constraints.
3. $\{P_i\}$: For the most part, we assume sentences are produced according to a uniform distribution on the degree-0 sentences of the relevant language, i.e., P_i is uniform on (degree-0 sentences of) L_i.

Ideally of course, a complete investigation of diachronic possibilities would involve varying $\mathcal{G}$, $\mathcal{A}$, and $\mathcal{P}$ and characterizing the resulting dynamical systems by their phase space plots. Rather than explore this entire space, we first consider only systems evolving from homogeneous initial populations, under four basic variants of the learning algorithm $\mathcal{A}$. This will give us an

Initial Language	Change to Language?
$(-V2)$ 1	2 (0.85), 6 (0.1)
$(+V2)$ 2	2 (0.98); stable
$(-V2)$ 3	6 (0.48), 8(0.38)
$(+V2)$ 4	4 (0.86); stable
$(-V2)$ 5	2 (0.97)
$(+V2)$ 6	6 (0.92); stable
$(-V2)$ 7	2 (0.54), 4(0.35)
$(+V2)$ 8	8 (0.97); stable

Table 1: Language change driven by misconvergence from a homogeneous initial linguistic population. A finite-sample analysis was conducted allowing each child learner 128 examples to internalize its grammar. After 30 generations, initial populations drifted (or not, as shown in the table) to different final linguistic compositions.

initial grasp of how linguistic populations can change. Indeed, linguistic change has been studied before; even the dynamical system metaphor itself has been invoked. Our computational paradigm lets us say much more than these previous descriptions: (1) we can say precisely what the rates of change will be; (2) we can determine what diachronic population curve changes will look like, without stipulating in advance that they must be S-shaped (sigmoid) or not, and without curve fitting to a pre-defined functional form.

4.1 Homogeneous Initial Populations

First we consider the case of a homogeneous population—no noise or confounding factors like foreign target languages. How stable are the languages in the 3-parameter system in this case? To determine this, we begin with a finite-sample analysis with $n = 128$ example sentences (recall by the analysis of Niyogi and Berwick (1993,1994a,1994b) that learners converge to target languages in the 3-parameter system with high probability after hearing this many sentences). Some small proportion of the children misconverge; the goal is to see whether this small proportion can drive language change—and if so, in what direction. To give the reader some idea of the possible outcomes, let us consider the four possible variations in the learning algorithm (±Single-step, ±Greedy)holding fixed the sentence distributions and learning sample.

4.1.1 Variation 1: $\mathcal{A}$ = TLA (+Single Step, +Greedy); P_i = Uniform; Finite Sample = 128

Suppose the learning algorithm is the triggering learning algorithm (TLA). The table below shows the language mix after 30 generations. Languages are numbered from 1 to 8. Recall that +V2 refers to a language that has the verb second property, and −V2 one that does not.

Observations. Some striking patterns regarding the resulting population mixes can be noted.

1. *First, all the $+V2$ languages are relatively stable*, i.e., the linguistic composition did not vary significantly over 30 generations. This means that every succeeding generation acquired the target parameter settings and no parameter drifts were observed over time.

2. *In contrast, populations speaking $-V2$ languages all drift to $+V2$ languages.* Thus a population speaking L_1 winds up speaking mostly L_2 (85%). A population speaking language L_7 gradually shifts to a population with 54 percent speaking L_2 and 35 percent speaking L_4 (with a smattering of other speakers) and apparently remains basically stable in this mix thereafter. Note that the relative stability of +V2 languages and the tendency of −V2 languages to drift to +V2 is exactly contrary to evidence in the linguistic literature. Lightfoot (1991), for example, claims that the tendency to lose V2 dominates the reverse tendency in the world's languages. Certainly, both English and French lost the V2 parameter setting—an empirically observed phenomenon that needs to be explained. Immediately then, we see that our dynamical system does not evolve in the expected manner. The reason could be due to any of the assumptions behind the model: the the parameter space, the learning algorithm, the initial conditions, or the distributional assumptions about sentences presented to learners. Exactly which is in error remains to be seen, but nonetheless our example shows concretely how assumptions about a grammatical theory and learning theory can make evolutionary, diachronic predictions—in this case, incorrect predictions that falsify the assumptions.

3. *The* **rates** *at which the linguistic composition changes vary significantly from language to language.* Consider for example the change of L_1 to L_2. Figure 2 below shows the gradual decrease in speakers of L_1 over successive generations along with the increase in L_2 speakers. We see that over the first 6 or seven generations very little change occurs, but over the next 6 or seven generations the population changes at a much faster rate. Note that in this particular case the two languages differ only in the V2 parameter, so the curves essentially plot the gain of V2. In contrast, consider figure 3 which shows the decrease of L_5 speakers and the shift to L_2. Here we note a sudden change: over a space of just 4 generations, the population shifts completely. Analysis of the time course of language change has been given some attention in linguistic analyses of diachronic syntax change, and we return to this issue below.

4. *We see that in many cases a homogeneous population splits up into different linguistic groups*, and seems to remain stable in that mix. In other words, certain combinations of language speakers seem to asymptote towards equilibrium (at least through 30 generations). For example, a population of L_7 speakers shifts over 5–6 generations to one with 54 percent speaking L_2 and 35 percent speaking L_4 and remains that way with no shifts in the distri-

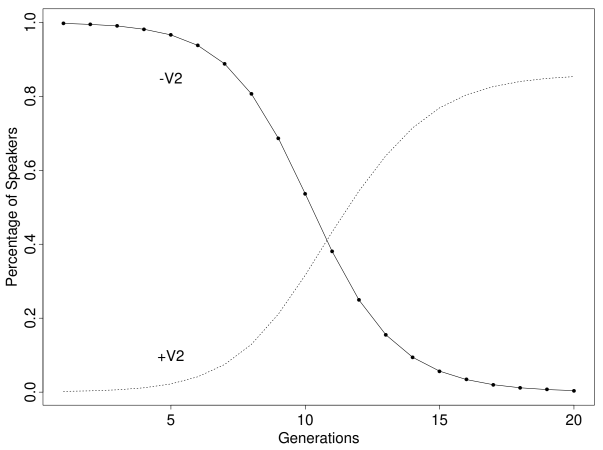

Figure 2: Percentage of a population speaking languages L_1 and L_2, measured on the Y-axis, as the population evolves over some number of generations, measured on the X-axis. The plot has been shown only up to 20 generations, as the proportions of L_1 and L_2 speakers do not vary significantly thereafter. Note that this curve is "S" shaped. Kroch(1989) *imposes* such a shape using models from population biology, while we derive this shape as an emergent property of our dynamical model. L_1 and L_2 differ only in the V2 parameter setting.

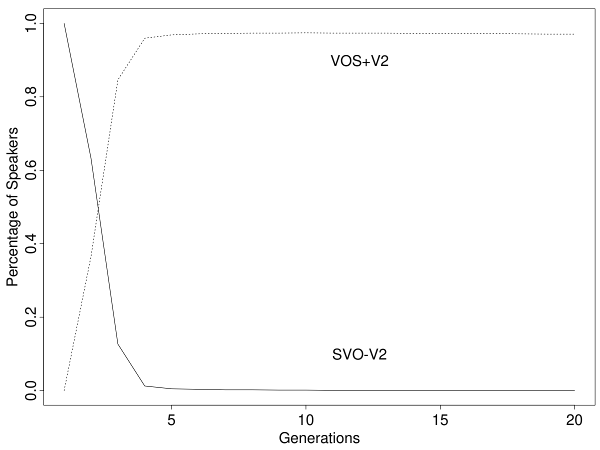

Figure 3: Percentage of the population speaking languages L_5 and L_2 as the population evolves over a number of generations. Note that a complete shift from L_5 to L_2 occurs over just 4 generations.

bution of speakers. Of course, we do not know for certain whether this is really a stable mixture. It could be that the population mix could suddenly shift after another 100 generations. What we really need to do is characterize the stable points or "limit cycles" of these dynamical systems. Other linguistic mixes can be inherently unstable; they might drift systematically to stable situations, or might shift dramatically (as with language L_1).

5. *It seems that the observed instability and drifts are to a large extent an artifact of the learning algorithm.* Remember that the TLA suffers from the problem of local maxima.[13] We note that those languages whose acquisition is not impeded by local maxima (the +V2 languages) are stable over time. Languages that have local maxima are unstable; in particular they drift to the local maxima over time. Consider L_7. If this is the target language, then there are two local maxima (L_2 and L_4) and these are precisely the states to which the system drifts over time. The same is true for languages L_5 and L_3. In this respect, the behavior of L_1 is quite unusual since it actually does not have any local maxima, yet it tends to flip the V2 parameter over time.

Now let us consider a different learning algorithm from the TLA that does not suffer from local maxima problems, to see whether this changes the dynamical system results.

4.1.2 Variation 2: $\mathcal{A}$ = +Greedy, −Single value; P_i = Uniform; Finite Sample = 128

Consider a simple variant of the TLA obtained by dropping the single valued constraint. This implies that the learner is no longer constrained to change just one parameter at a time: on being presented with a sentence it cannot analyze, it chooses any of the alternative grammars and attempts to analyze the sentence with it. Greediness is retained; thus the learner retains its original hypothesis if the new one is also not able to analyze the sentence. Given this new learning algorithm, and retaining all the other original assumptions, Table 2 shows the distribution of speakers after 30 generations.

Observations. In this situation there are no local maxima, and the evolutionary pattern takes on a very different nature. There are two distinct observations to be made.

1. *All homogeneous populations eventually drift to a strikingly similar population mix, irrespective of what language they start from.* What is unique about this mix? Is it a stable point (or attractor)? Further simulations and theoretical analyses are needed to resolve this question; we leave these as open questions.

2. *All homogeneous populations drift to a population mix of only +V2 languages.* Thus, the V2 parameter is gradually set over succeeding generations by all people in the community (irrespective of which language they speak). In other words, as before,

[13] We regard local maxima of a language L_i to be alternative absorbing states (sinks) in the Markov chain for that target language. This formulation differs slightly from the conception of local maxima in Gibson and Wexler (1994), a matter discussed at some length in Niyogi and Berwick (1993). Thus according to our definition L_4 is not a local maxima for L_5 and consequently no shift is observed.

Initial Language	Change to Language?
$-V2$ 1	2 (0.41), 4 (0.19), 6 (0.18), 8 (0.13)
$+V2$ 2	2 (0.42), 4 (0.19), 6 (0.17), 8 (0.12)
$-V2$ 3	2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)
$+V2$ 4	2 (0.41), 4 (0.19), 6 (0.18), 8 (0.13)
$-V2$ 5	2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)
$+V2$ 6	2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)
$-V2$ 7	2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)
$+V2$ 8	2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)

Table 2: Language change driven by misconvergence. A finite-sample analysis was conducted allowing each child learner (following the TLA with single-value dropped) 128 examples to internalize its grammar. Initial populations were linguistically homogeneous, and they drifted to different linguistic compositions. The major language groups after 30 generations have been listed in this table. Note how all initially homogeneous populations tend to the same composition.

there is a tendency to gain V2 rather than lose V2, contrary to the empirical facts.

As an example, fig. 4 shows the changing percentage of the population speaking the different languages starting off from a homogeneous population speaking L_5. As before, learners who have not converged to the target in 128 examples are the driving force for change here. Note again the time evolution of the grammars. For about 5 generations there is only a slight decrease in the percentage of speakers of L_5. Then the linguistic patterns switch rapidly over the next 7 generations to a relatively stable mix.

4.1.3 Variations 3 & 4: $-$Greedy, $\pm$Single Value constraint; P_i =Uniform; Finite Sample = 128

Having dropped the single value constraint, we consider the next obvious variation in the learning algorithm: dropping greediness while varying the single value constraint. Again, our goal is to see whether this makes any difference in the resulting dynamical system. This gives rise to two different learning algorithms: (1) allow the learning algorithm to pick any new grammar at most one parameter value away from its current hypothesis (retaining the single-value constraint, but without greediness, that is, the new grammar does not have to be able to parse the current input sentence); (2) allow the learning algorithm to pick any new grammar at each step (no matter how far away from its current hypothesis).

In both cases, the population mix after 30 generations is the same irrespective of the initial language of the homogeneous population. These results are shown in table 3.

Observations:

1. *Both algorithms yield dynamical systems that arrive at the same population mix after 30 generations.* The path by which they arrive at this mix is, however, not the same (see figure 5).

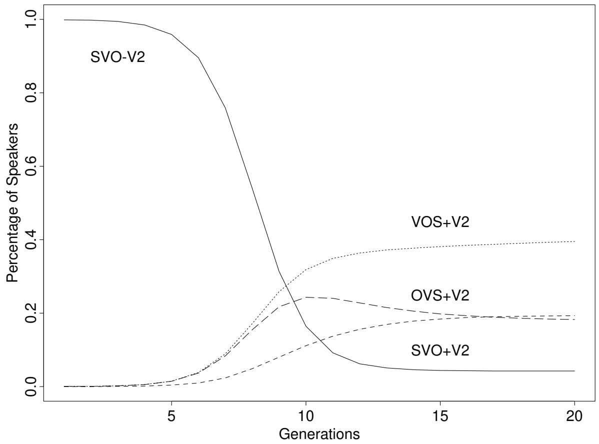

Figure 4: Time evolution of grammars using a greedy learning algorithm with no single value constraint in place.

Initial Language	Change to Language?
Any Language	1 (0.11), 2 (0.16), 3 (0.10), 4 (0.14)
(Homogeneous)	5 (0.12), 6 (0.14), 7 (0.10), 8 (0.13)

Table 3: Language change driven by misconvergence, using two different acquisition algorithms that do not obey a local gradient-ascent rule (a greediness constraint). A finite-sample analysis was conducted with the learning algorithm following a random-step algorithm or else a single-step algorithm, along with 128 examples to internalize its grammar. Initial populations were linguistically homogeneous, and they drifted to different linguistic compositions. The major language groups after 30 generations have been listed in this table. Note that all initially homogeneous populations converge to the same final composition.

2. *The final population mix contains all languages in significant proportion.* This is in distinct contrast to the previous situations, where we saw that $-$V2 languages were eliminated over time.

4.2 Modeling Diachronic Trajectories

With a basic notion of how diachronic systems can evolve given different learning algorithms, we turn next to the question of population trajectories. While we can already see that some evolutionary trajectories have a "linguistically classical" S-shape, their smoothness can vary. However, our formalization allows us to say much more than this. Unlike the previous work in diachronic linguistics that we are familiar with, we can explore the space of possible trajectories, examining factors that affect their evolutionary time course, without assuming an *a priori* S-shape.

For example, Bailey (1973) proposed a "wave" model of linguistic change: linguistic replacements follow an S-shaped curve over time. In Bailey's own words (taken

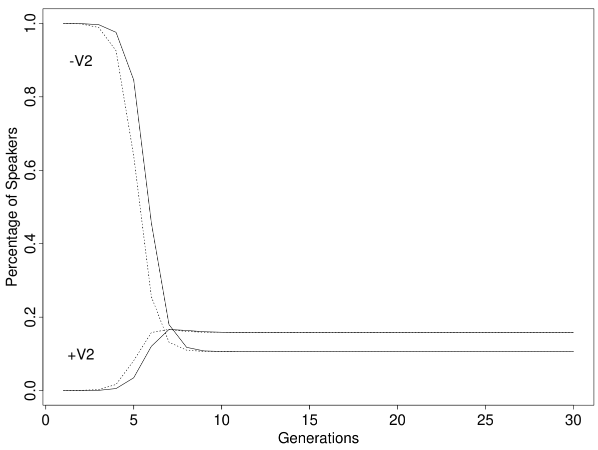

Figure 5: Time evolution of linguistic composition for the situations where the learning algorithm is $-$Greedy, +Single Value constraint (dotted line), and $-$Greedy, $-$Single Value (solid line). Only the percentage of people speaking L_1 (-V2) and L_2 (+V2) are shown. The initial population is homogeneous and speaks L_1. The percentage of L_1 speakers gradually decreases to about 11 percent. The percentage of L_2 speakers rises to about 16 percent from 0 percent. The two dynamical systems converge to the same population mix; however, their trajectories are not the same—the rates of change are different, as shown in this plot.

from Kroch, 1990):

> A given change begins quite gradually; after reaching a certain point (say, twenty percent), it picks up momentum and proceeds at a much faster rate; and finally tails off slowly before reaching completion. The result is an S-curve: the statistical differences among isolects in the middle relative times of the change will be greater than the statistical differences among the early and late isolects.

The idea that linguistic changes follow an S-curve has also been proposed by Osgood and Sebeok (1954) and Weinreich, Labov, and Herzog (1968). More specific logistic forms have been advanced by Altmann (1983) and Kroch (1982,1989). Here, the idea of a logistic functional form is borrowed from population biology where it is demonstrable that the logistic governs the replacement of organisms and of genetic alleles that differ in Darwinian fitness. However, Kroch (1989) concedes that "unlike in the population biology case, no mechanism of change has been proposed from which the logistic form can be deduced."

Crucially, in our case, we suggest a specific mechanism of change: an acquisition-based model where the combination of grammatical theory, learning algorithms, and distributional assumptions on sentences drive change. The specific form might or might not be S-shaped, and might have varying rates of change.[14]

Among the other factors that affect evolutionary trajectories are maturation time—the number of sentences available to the learner before it internalizes its adult grammar—and the distributions with which sentences are presented to the learner. We examine these in turn.

4.2.1 The Effect of Maturation Time or Sample Size

One obvious factor influencing the evolutionary trajectories is the maturational time, i.e., the number (N) of sentences the child is allowed to hear before forming its mature hypothesis. This was fixed at 128 in all the systems shown so far (based in part on our explicit computation for the Markov convergence time in this situation). Figure 6 shows the effect of varying N on the evolutionary trajectories. As usual, we plot only a subspace of the population. In particular, we plot the percentage of L_2 speakers in the population with each succeeding generation. The initial composition of the population was homogeneous (with people speaking L_1).

Observations.

1. *The initial rate of change of the population is highest when the maturation time is smallest*, i.e., the learner is allowed the least amount of time to develop its mature hypothesis. This is not surprising. If the learner were allowed access to a lot of examples to make its mature hypothesis, most learners would reach the target grammar. Very few would misconverge, and the linguistic composition would change little over the next generation. On the other hand, if the learner were allowed very few examples to develop its hypothesis, many would misconverge, possibly causing great change over one generation.

2. *The "stable" linguistic compositions seem to depend upon maturation time.* For example, if learners are allowed only 8 examples, the percentage of L_2 speakers rises quickly to about 0.26. On the other hand, if learners are allowed 128 examples, the percentage of L_2 speakers eventually rises to about 0.41.

3. *Note that the trajectories* **do not** *have an S-shaped curve* in contrast to the results of Kroch (1989).

4. *The maturation time is related to the order of the dynamical system.*

[14] Of course, we do not mean to say that we can simulate *any* possible trajectory—that would make the formalism empty. Rather, we are exploring the initial space of possible trajectories, given some example initial conditions that have been already advanced in the literature. Because the mathematics for dynamical systems is in general quite complex, at present we cannot make general statements of the form, "under these particular initial conditions the trajectory will be sigmoidal, and under these other conditions it will not be." We have conducted only very preliminary investigations demonstrating that potentially at least, reasonable, distinct initial conditions can lead to demonstrably different trajectories.

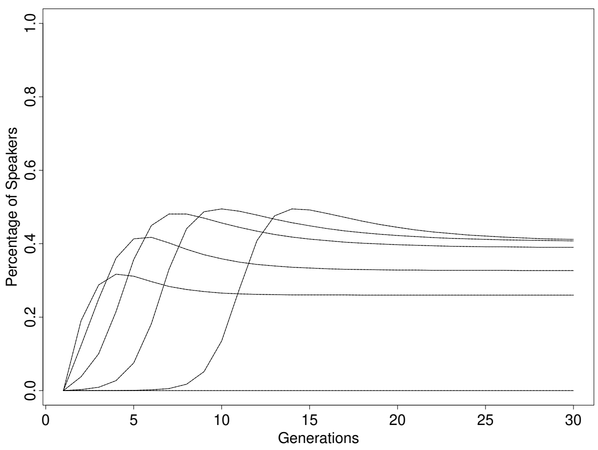

Figure 6: Time evolution of linguistic composition when varying maturation time (sample size). The learning algorithm used is the +Greedy, −Single value. Only the percentage of people speaking L_2 (+V2) is shown. The initial population is homogeneous and speaks L_1. The maturation time was varied through 8, 16, 32, 64, 128, and 256, giving rise to the six curves shown. The curve with the highest initial rate of change corresponds to 8 examples for maturation time. The initial rate of change decreases as the maturation time N increases. The value at which these curves asymptote also seems to vary with the maturation time, and increases monotonically with it.

4.2.2 The Effect of Sentence Distributions (P_i :)

Another important factor influencing evolutionary trajectories is the distribution P_i with which sentences of the ith language, L_i, are presented to the learner. In a certain sense, the grammatical space and the learning algorithm jointly determine the order of the dynamical system. On the other hand, sentence distributions are much like the parameters of the dynamical system (see sec. 4.3.2). Clearly the sentence distributions affect rates of convergence within one generation. Further, by putting greater weight on certain word forms rather than others, they might influence systemic evolution in certain directions. While this is again an obvious point, the model lets us consider the alternatives precisely.

To illustrate the idea, consider the following example: the interaction between L_1 and L_2 speakers in the community as the sentence distributions with which these speakers produce sentences changes. Recall that so far we have assumed that all speakers produce sentences with uniform distributions on degree-0 sentences of their respective languages. Now we consider alternative distributions, parameterized by a value p:

1. Let $L_{1,2} = L_1 \cap L_2$.
2. P_1 : Speakers of L_1 produce sentences so that all degree-0 sentences of $L_{1,2}$ are equally likely and their total probability is p. Further, sentences of $L_1 \setminus L_{1,2}$ are also equally likely, but their total proability is $1 - p$.
3. P_2 : Speakers of L_2 produce sentences so that all degree-0 sentences of $L_{1,2}$ are equally likely and their total probability is p. Further, sentences of $L_2 \setminus L_{1,2}$ are also equally likely, but their total proability is $1 - p$.
4. Other P_i's are all uniform over degree-0 sentences.

The parameter p determines the weight on the sentence patterns in common between the languages L_1 and L_2. Figure 7 shows the evolution of the L_2 speakers as p varies. Here the learning algorithm is +Greedy, +Single value (TLA, or local gradient ascent) and the initial population is homogeneous, 100% L_1; 0% L_2. Note that the system moves in different ways as p varies. When p is very small (0.05), that is, sentences common to L_1 and L_2 occur infrequently, in the long run the percentage of L_2 speakers *does not* increase; the population stays put with L_1. However, as p grows, more strings of L_2 occur, and the dynamical system changes so that the long-term percentage of L_1 speakers decreases and that of L_2 speakers increases. When p reaches 0.75 the initial population evolves into a completely L_2 speaking community. After this, as p increases further, we notice (see $p = 0.95$) that the L_2 speakers increase but can never rise to 100 percent of the population; there is still a residual L_1 speaking component. This is to be expected, because for such high values of p, many strings common to L_1 and L_2 occur frequently. This means that a learner could sometimes converge to L_1 just as well as L_2, and some learners indeed begin to do so, increasing the number of the L_1 speakers.

This example shows us that if we wanted a homogeneous L_1 speaking population to move to a homogeneous L_2 speaking population, by choosing our distributions appropriately, we could drive the grammatical dynamical system in the appropriate direction. It suggests another important application of the dynamical system approach: one can work backwards, and examine the conditions needed to generate a change of a certain kind. By checking whether such conditions could have possibly existed historically, we can falsify a grammatical theory or a learning paradigm. Note that this example showed the effect of sentence distributions, and how to alter them to obtain desired evolutionary envelopes. One could, in principle, alter the grammatical theory or the learning algorithm in the same fashion—-leading to a tool to aid the search for an adequate linguistic theory.[15]

4.3 Nonhomogeneous Populations: Phase-Space Plots

For our three-parameter system, we have been able to characterize the update rules for the dynamical systems corresponding to a variety of learning algorithms. Each

[15] Again, we stress that we obviously do not want so weak a theory that we can arrive at *any* possible initial conditions simply by carrying out reasonable changes to the sentence distributions. This may, of course, be possible; we have not yet examined the general case.

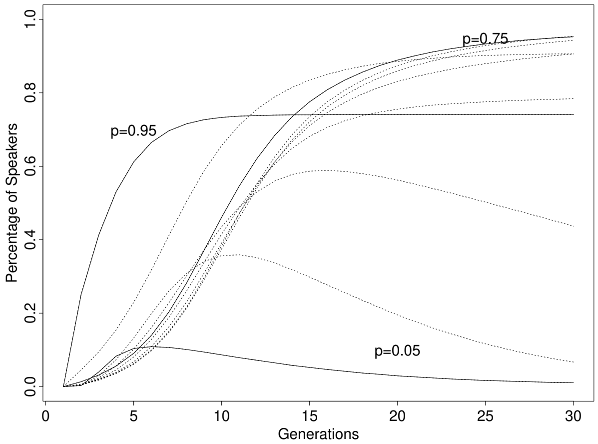

Figure 7: The evolution of L_2 speakers in the community for various values of p (a parameter related to the sentence distributions P_i, see text). The algorithm used was the TLA, the inital population was homogeneous, speaking only L_1. The curves for $p = 0.05, 0.75$, and 0.95 have been plotted as solid lines.

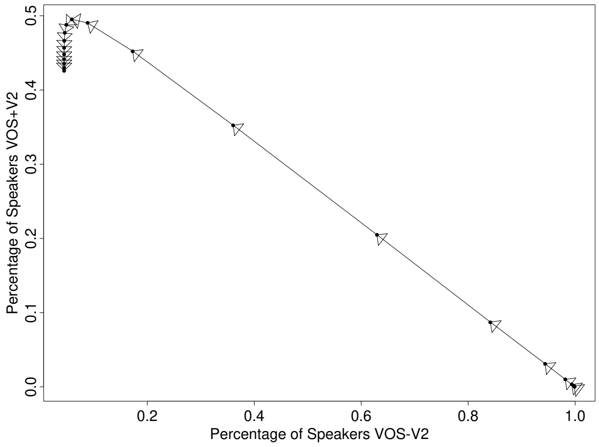

Figure 8: Subspace of a phase-space plot. The plot shows $(\pi_1(t), \pi_2(t))$ as t varies, i.e., the proportion of speakers speaking languages L_1 and L_2 in the population. The initial state of the population was homogeneous (speaking language L_1). The algorithm used was +Greedy −Single value.

dynamical system has a specific update procedure according to which the states evolve from some *homogeneous* initial population. A more complete characterization of the dynamical system would be achieved by obtaining *phase-space* plots of this system. Such phase-space plots are pictures of the state-space $\mathcal{S}$ filled with *trajectories* obtained by letting the system evolve from various initial points (states) in the state space.

4.3.1 Phase-Space Plots: Grammatical Trajectories

We have described earlier the relationship between the state of the population in one generation and the next. In our case, let Π denote an 8-dimensional vector variable (state variable). Specifically, $\Pi = (\pi_1, \ldots, \pi_8)'$ (with $\sum_{i=1}^{8} \pi_i$) as we discussed before. The following schema reiterates the chain of dependencies involved in the update rule governing system evolution. The state of the population at time t (in generations), allows us to compute the transition matrix T for the Markov chain associated with the memoryless learner. Now, depending upon whether we want (1) an asymptotic analysis or (2) a finite sample analysis, we compute (1) the limiting behavior of T^m as m (the number of examples) goes to infinity (for an asymptotic analysis), or (2) the value of T^N (where N is the number of examples after which maturation occurs). This allows us to compute the next state of the population. Thus $\Pi(t+1) = g(\Pi(t))$ where g is a complex non-linear relation.

$$\Pi(t) \Longrightarrow P \text{ on } \Sigma* \Longrightarrow T \Longrightarrow T^m \Longrightarrow \Pi(t+1)$$

If we choose a certain initial condition Π_1, the system will evolve according to the above relation and one can obtain a trajectory of Π in the 8 dimensional space over time. Each initial condition yields a unique trajectory and one can then plot these trajectories obtaining a phase-space plot. Each such trajectory corresponds to a line in the 8-dimensional plane given by $\sum_{i=1}^{8} \pi_i = 1$. One cannot directly display such a high dimensional object, but we plot in figure 8 the projection of a particular trajectory onto a two dimensional subspace given by $(\pi_1(t), \pi_2(t))$ (the proportion of speakers of L_1 and L_2) at different points in time.

As mentioned earlier, with a different initial condition we get a different grammatical trajectory. The complete state space picture is thus filled with all the different trajectories corresponding to different initial conditions. Fig. 9 shows this.

4.3.2 Stability Issues

The phase-space plots show that many initial conditions yield trajectories that seem to converge to a single point in the state space. In the dynamical systems terminology, this corresponds to a fixed point of the system—a population mix that stays at the same composition. Many natural questions arise at this stage. What are the conditions for stability? How many fixed points are there in a given system? How can we solve for them? These are interesting questions but detailed answers are not within the scope of the current paper. In lieu of a more complete analysis we state here a fixed point theorem that allows one to characterize the stable population mixes.

First, some notational preliminaries. As before, let P_i be the distribution on the sentences of the ith language L_i. From P_i, we can construct T_i, the transition matrix whose elements are given by the explicit procedure documented in Niyogi and Berwick (1993, 1994a, 1994b). The matrix T_i models a +Greedy −Single value learner if the target language is L_i (with sentences from

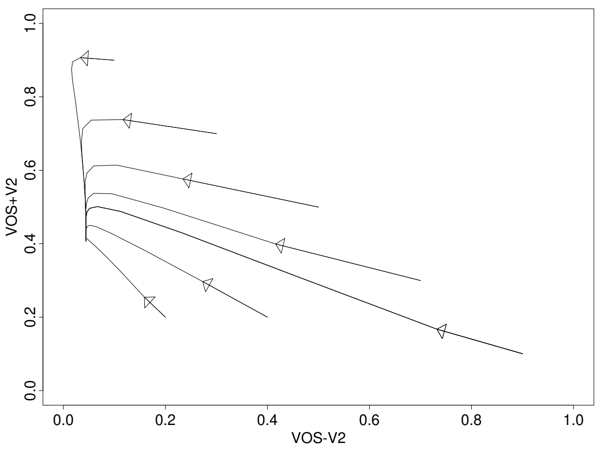

Figure 9: Subspace of a Phase-space plot. The plot shows $(\pi_1(t), \pi_2(t))$ as t varies for different nonhomogeneous initial population conditions. The algorithm used was +Greedy −Single value.

the target produced with P_i). Similarly, one can obtain the matrices for other learning variants. Note that fixing the P_i's fixes the T_i's and in so the P_i's are a different sort of "parameter" that characterize how the dynamical system evolves.[16] If the state of the parent population at time t is $\Pi(t)$, then it is possible to show that the (true) transition matrix for ±Greedy ±Single value learners is $T = \sum_{i=1}^{8} \pi_i(t)T_i$. For the finite case analysis, the following theorem holds:

Theorem 1 (Finite Case) *A fixed point (stable point) of the grammatical dynamical system (obtained by a ±Greedy ±Single value learner operating on the 8 parameter space with k examples to choose its final hypothesis) is a solution of the following equation:*

$$\Pi' = (\pi_1, \ldots, \pi_8) = (1, \ldots, 1)'(\sum_{i=1}^{8} \pi_i T_i)^k$$

Proof (Sketch): This equation is obtained simply by setting $\Pi(t+1) = \Pi(t)$. Note however, that this is an example of a nonlinear multidimensional iterated function map. The analysis of such dynamical systems is nontrivial, and our theorem by no means captures all the possibilities. ∎

We can similarly state a theorem for the limiting (asymptotic) case analysis.

Theorem 2 (Limiting or Asymptotic Analysis) *A fixed point (stable point) of the grammatical dynamical system (obtained by a ±Greedy ±Single value learner operating on the 8 parameter space (given infinite examples to choose its mature hypothesis) is a solution of the following equation:*

$$\Pi' = (\pi_1, \ldots, \pi_8) = (1, \ldots, 1)'(I - \sum_{i=1}^{8} \pi_i T_i + ONE)^{-1}$$

where ONE is the 8×8 matrix with all its entries equal to 1.

Proof: Again this is trivially obtained by setting $\Pi(t+1) = \Pi(t)$. The expression on the right provides an analytical expression for the update equation in the asymptotic case. See Resnick (1992) for details. All the caveats mentioned in the proof section of the previous theorem apply here as well. ∎

Remark. We have just touched the surface as far as the theoretical characterization of these grammatical dynamical systems are concerned. The main purpose of this paper is to show that these dynamical systems exist as a logical consequence of assumptions about the grammatical space and an acquisition theory. We have exhibited only some preliminary simulations with these systems. From a theoretical perspective, it would be much more valuable to have complete characterizations of such systems. Strogatz (1993) suggests that nonlinear multidimensional mappings with greater than 3 dimensions are likely to be chaotic. It is also interesting to note that iterated function maps define fractal sets . Such investigations are beyond the scope of this paper, and might well be a fruitful area for further research.

5 Example 2: From Old French to Modern French; Clark and Roberts Analysis Revisited

So far, our examples have been based on a 3-parameter linguistic theory for which we derived several different dynamical systems. Our goal was to concretely instantiate our philosophical arguments, sketching the factors that influence evolutionary trajectories. In this section, we briefly consider a different parametric linguistic system studied by Clark and Roberts, 1993. The historical context in which Clark and Roberts advanced their linguistic proposal is the evolution of Modern French from Old French. Their parameters are intended to capture some, but of course not all, of this change. They too use a learning algorithm—in their case, a genetic algorithm—to account for historical change but do not analyze their model from the dynamical systems viewpoint. Here we adopt their parameterization, with all its strengths and weaknesses, but consider an alternative learning paradigm and the dynamical systems approach.

Extensive simulations in the earlier section reveal that while the learnability problem of the 3-parameter space can be solved by stochastic hill climbing algorithms, the long term evolution of these algorithms have a behavior that is at variance with the diachronic change actually observed in historical linguistics. In particular, we saw how there was a tendency to gain rather than lose the V2 parameter setting. While this could well be an artifact of

[16] There are thus two *distinct* kinds of parameters in our model: first, parameters that define the 2^n languages and define the state-space of the system; and second, the P_i's the characterize the way in which the system evolves and are therefore the parameters of the complete grammatical dynamical system.

the class of learning algorithms considered, a more likely explanation is that loss of V2 (observed in many of the world's languages like French, English, and so forth) is due to an interaction of parameters and triggers other than those considered in the previous section. We investigate this possibility and begin by first reviewing Clark and Roberts' alternative parametric theory.

5.1 The Parametric Subspace and Data

We now consider a syntactic space involving the with 5 (boolean-valued) parameters. We do not attempt to describe these parameters. The interested reader should consult Haegeman (1991) for details and Clark and Roberts (1993) for details.

1. p_1: Case assignment under agreement ($p_1 = 1$) or not ($p_1 = 0$).
2. p_2: Case assignment under government ($p_2 = 1$) or not (($p_2 = 0$). Relevant triggers for this parameter include "Adv V S", "S V O".
3. p_3: Nominative clitics.
4. p_4: Null Subject. Here relevant triggers would include "wh V S O".
5. p_5: Verb-second V2. Triggers include "Adv V S" , and "S V O".

These 5 parameters define a 32 grammar space. Each grammar in this parametrized system can be represented by a string of 5 bits depending upon the values of $p_1, \ldots, p_5$, for instance, the first bit position corresponds to case assignment under agreement. We can now look at the surface strings (sentences) generated by each such grammar. For the purpose of explaining how Old French changed to Modern French, Clark and Roberts consider the following key sentences. The parameter settings required to generate each sentence are provided in brackets; an asterisk is a "doesn't matter" value and an "X" means any phrase.

The Relevant Data

adv V S	[*1**1]
SVO	[*1**1] or [1***0]
wh V S O	[*1***]
wh V S O	[**1**]
X (pro) V O	[*1*11] or [1**10]
X V s	[**1*1]
X s V	[**1*0]
X S V [1***0]	
(S) V Y [*1*11]	

The parameter settings provided in brackets set the grammars which generate the sentence. For example, the sentence form "adv V S" (corresponding to *quickly ran John*), an incorrect word order in English) is generated by all grammars that have case assignment under government (the second element of the array set to 1, $p_2 = 1$) and verb second movement ($p_5 = 1$). The other parameters can be set to any value. Clearly there are 8 different grammars that can generate (alternatively parse) this sentence. Similarly there are 16 grammars that generate the form S V O (8 corresponding to parameter settings of [*1**1] and 8 corresponding to parameter settings of [1***0]) and 4 grammars that generate ((s) V Y).

Remark. Note that the sentence set Clark and Roberts considered is only a subset of the the total number of degree-0 sentences generated by the 32 grammars in question. In order to directly compare their model with ours, we have not attempted to expand the data set or fill out the space any further. As a result, all the grammars do not have unique extensional properties, i.e., some generate the same set of sentences.

5.2 The Case of Diachronic Syntax Change in French

Continuing with Clark and Roberts' analysis, within this parameter space, it is historically observed that the language spoken in France underwent a parametric change from the twelfth century to modern times. In particular, they point out that both V2 and prodrop are lost, illustrated by examples like these:
Loss of null subjects: pro-drop

(1) (Old French; +pro drop)
Si firent (pro) grant joie la nuit
'thus (they) made great joy the night'

(2) (Modern French; −pro drop)
∗ Ainsi s'amusaient bien cette nuit
'thus (they) had fun that night'

Loss of V2

(3) (Old French; +V2)
Lors oirent ils venir un escoiz de tonoire
'then they heard come a clap of thunder'

(4) (Modern French; −V2)
∗ Puis entendirent-ils un coup de tonerre. 'then they heard a clap of thunder'

Clark and Roberts observe that it has been argued this transition was brought about by the introduction of new word orders during the fifteenth and sixteenth centuries resulting in generations of children acquiring slightly different grammars and eventually culminating in the grammar of modern French. A brief reconstruction of the historical process (after Clark and Roberts, 1993) runs as follows.

Old French; setting [11011] The language spoken in the twelfth and thirteenth centuries had verb-second movement and null subjects, both of which were dropped by the twentieth century. The sentences generated by the parameter settings corresponding to Old French are:

Old French

adv V S –	[*1**1]
S V O –	[*1**1] or [1***0]
wh V S O	[*1***]
X (pro) V O	[*1*11] or [1**10]

Note that from this data set it appears that both the Case agreement and nominative clitics parameters remain ambiguous. In particular, Old French is in a subset-superset relation with another language (generated by the parameter settings of 11111). In this case, possibly some kind of subset principle (Berwick, 1985) could be used by the learner; otherwise it is not clear how the data would allow the learner to converge to the Old French grammar in the first place. None of the ±Greedy, ±Single value algorithms would converge uniquely to the grammar of Old French.

The string (X)VS occurs with frequency 58% and SV(X) occurs with 34% in Old French texts. I t is argued that this frequency of (X)VS is high enough to cause the V2 parameter to trigger to +V2.

Middle French In Middle French, the data is not consistent with any of the 32 target grammars (equivalent to a heterogenous population). Analysis of texts from that period reveal that some old forms (like Adv V S) decreased in frequency and new forms (like Adv S V) increased. It is argued in Clark and Roberts that such a frequency shift causes "erosion" of V2, brings about parameter instability and ultimately convergence to the grammar of Modern French. In this transition period (i.e. when Middle French was spoken/written) the data is of the following form:

adv V S [*1**1]; SVO [*1**1] or [1***0]; wh V S O [*1***]; wh V s O [**1**]; X (pro)V O [*1*11] or [1**10]; X V s [**1*1]; X s V [**1*0]; X S V [1***0]; (s)VY [*1*11]

Thus, we have old sentence patterns like Adv V S (though it decreases in frequency and becomes only 10%), SVO, X (pro)V O and whVSO. The new sentence patterns which emerge at this stage are adv S V (increases in frequency to become 60%), X subjclitic V, V subjclitic (pro)V Y (null subjects) , whV subjclitic O.

Modern French [10100] By the eighteenth century, French had lost both the V2 parameter setting as well as the null subject parameter setting. The sentence patterns consistent with Modern French parameter settings are SVO [*1**1] or [1***0], X S V [1***0], V s O [**1**]. Note that this data, though consistent with Modern French, will not trigger all the parameter settings. In this sense, Modern French (just like Old French) is not uniquely learnable from data. However, as before, we shall not concern ourselves overly with this, for the relevant parameters (V2 and null subject) are uniquely set by the data here.

5.3 Some Dynamical System Simulations

We can obtain dynamical systems for this parametric space, for a TLA (or TLA-like) algorithm in a straightforward fashion. We show the results of two simulations conducted with such dynamical systems.

5.3.1 Homogeneous Populations [Initial–Old French]

We conducted a simulation on this new parameter space using the Triggering Learning Algorithm. Recall that the relevant Markov chain in this case has 32 states. We start the simulation with a homogeneous population speaking Old French (parameter setting = 11011). Our goal was to see if misconvergence alone, could drive Old French to Modern French.

Just as before, we can observe the linguistic composition of the population over several generations. It is observed that in one generation, 15 percent of the children converge to grammar 01011; 18 percent to grammar

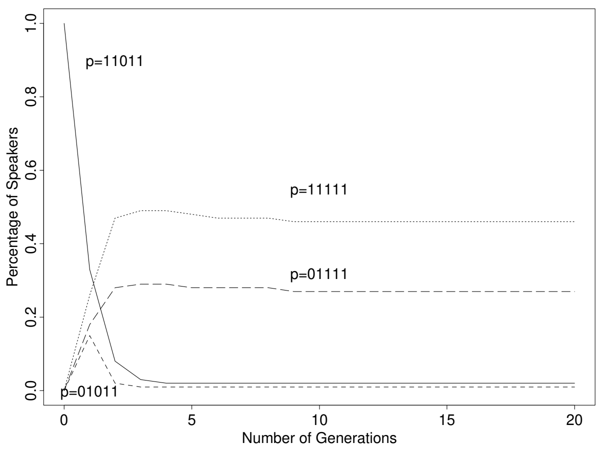

Figure 10: Evolution of speakers of different languages in a population starting off with speakers only of Old French.

01111; 33 percent to grammar 11011 (target) and 26 percent to grammar 11111 with very few having converged to other grammars. Thereafter, the population consists mostly of speakers of these 4 languages, with one important difference: 15 percent of the speakers eventually *lose V2*. In particular, they have acquired the grammar 11110. Shown in fig. 10 are the percentage of the population speaking the 4 languages mentioned above as they evolve over 20 generations. Notice that in the space of a few generations, the speakers of 11011, and 01011 have dropped out altogether. Most of the population now speaks language 1111 (46 percent) and 01111 (27 percent). Fifteen percent of the population speaks 11110 and there is a smattering of other speakers. The population remains roughly stable in this configuration thereafter.

Observations:

1. On examining the four languages to which the system converges after one generation, we noice that they share the same settings for the principles [Case assignemnt under government], [pro drop], and [V2]. These correspond to the three parameters which are uniquely set by data from Old French. The other two parameters can take on any value. Consequently 4 languages are generated all of which satisfy the data from Old French.

2. Recall our earlier remark that due to insufficient data, there were equivalent grammars in the parameter system. It turns out that in this particular case, the grammars (01011) and (11011) are identical as far as their extensional properties are concerned; as are the grammars (11111) and (01111).

3. There is subset relation between the two sets described in (2). The grammar (11011) is in a subset relation with (11111). This explains why after a few generations most of the population switches to either (11111) or (01111) (the superset grammars).

4. An interesting feature of the simulation is that 15 percent of the population eventually acquires the gram-

mar (11110), i.e., they have lost the V2 parameter setting. This is the first sign of instability of V2 that we have seen in our simulations so far (for greedy algorithms which are psychologically preferred). Recall that for such algorithms, the V2 parameter was very stable in our previous example.

5.3.2 Heterogenous Populations (Mixtures)

The earlier section showed that with no new (foreign) sentence patterns the grammatical system starting out with only Old French speakers showed some tendency to lose V2. However, the grammatical trajectory did not terminate in Modern French. In order to more closely duplicate this historically observed trajectory, we examine alternative inital conditions. We start our simulations with an initial condition which is a mixture of two sources; data from Old French and data from New French (reproducing in this sense, data similar to that obtained from the Middle French period). Thus children in the next generation observe new surface forms. Most of the surface forms observed in Middle French are covered by this mixture.

Observations:

1. On performing the simulations using the TLA as a learning algorithm on this parameter space, an interesting pattern is observed. Suppose the learner is exposed to sentences with 90 percent generated by Old French grammar (11011) and 10 percent by Modern French grammar (10100), within one generation 22 percent of the learners have converged to the grammar (11110) and 78 percent to the grammar (11111). Thus the learners set each of the parameter values to 1 except the V2 parameter setting. Now Modern French is a non-V2 language; and 10 percent of data from Modern French is sufficient to cause 22 percent of the speakers to lose V2. This is the behaviour over one generation. The new population (consisting of 78 percent speaking grammar (11111) and 22 percent speaking grammar (11110)) remains stable for ever.

2. Fig. 11 shows the proportion of speakers who have lost V2 after one generation, as a function of the proportion of sentences from the Modern French Source. The shape of the curve is interesting. For small values of the proportion of the Modern French source, the slope of the curve is greater than 1. Thus there is a greater tendency of speakers to lose V2 than to retain it. Thus 10 percent of novel sentences from the Modern French source causes 20 percent of the population to lose V2; similarly 20 percent of novel sentences from the Modern French source causes 40 percent of the speakers to lose V2. This effect wears off later. This seems to capture computationally the intuitive notion of many linguists that a small change in inputs provided to children could drive the system towards larger change.

3. Unfortunately, there are several shortcomings of this particular simulation. First, we notice that mixing Old and Modern French sources does not cause the desired (historically observed) grammatical trajectory from Old to Modern French (corresponding in our system to movement from state (11011) to state (10100) in our Markov Chain). Although we find that a small injection

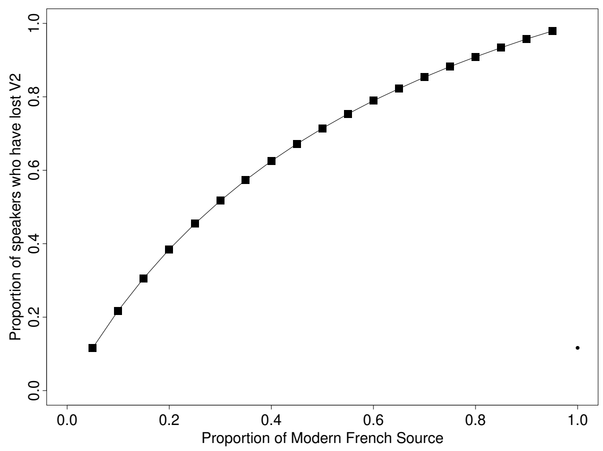

Figure 11: Tendency to lose V2 as a result of new word orders introduced by Modern French source in our Markov Model.

of sentences from Modern French causes a larger percentage of the population to lose V2 and gain subject clitics (which are historically observed phenomena), nevertheless, the entire population retains the null subject setting and case assignment under government. It should be mentioned that Clark and Roberts argue that the change in case assignment under government is the driving force which allows alternate parse-trees to be formed and causes the parametric loss of V2 and null subject. In this sense, it is a more fundamental change.

4. If the dynamical system is allowed to evolve, it ends up in either of the two states (11111) or (11110). This is essentially due to the subset relations these states (languages) have with other languages in the system. Another complication in the system is the equivalence of several different grammars (with respect to their surface extensions) e.g. given the data we are considering, the grammars (01011) and (11011) (Old French) generate the same sentences. This leads to multiplicity of paths, convergence to more than one target grammar and general inelegance of the state-space description.

Future Directions: There are several possibilities to consider here.

1. Using more data and filling out the state-space might yield greater insight. Note that we can also study the development of other languages like Italian or Spanish within this framework and that might be useful.

2. TLA-like hill climbing algorithms do not pay attention to the subset principle explicitly. It would be interesting to explicitly program this into the learning algorithm and observe the evolution thereafter.

3. There are often cases when several different grammars generate the same sentences or atleast equally well fit the data. Algorithms which look only at surface strings are unable then to distinguish between them resulting in convergence to all of them with different probabilities in our stochastic setting. We saw an example of this for convergence to four states earlier. Clark

and Roberts suggest an elegance criterion by looking at the parse-trees to decide between these grammars. This difference between *strong* generative capacity and *weak* generative capacity can easily be incorporated into the Markov model as well. The transition probabilites, now, will not depend upon the surface properties of the grammars alone, but also upon the elegance of derivation for each surface string.

4. Rather than the evolution of the population, one could look at the evolution of the distribution of words. One can also obtain bounds on frequencies with which the new data in the Middle French Period must occur so that the correct drift is observed.

6 Conclusions and Directions for Future Research

In this paper, we have argued that any combination of (grammatical theory, learning paradigm) leads to a model of grammatical evolution and diachronic change. A learning theory (paradigm) attempts to account for how children (the individual child) solve the problem of language acquisition. By considering a population of such "child learners", we have arrived at a model of the *emergent*, global, population behavior. The key point is that such a model is a logical consequence of grammatical, and learning theories. Consequently, whenever a linguist suggests a new grammatical, or learning theory, they are also suggesting a particular evolutionary theory—and the consequences of this need to be examined.

Historical Linguistics and Diachronic Criteria

From a programmatic persepective, this paper has two important consequences. First, it allows us to take a formal, analytic view of historical linguistics. Most accounts of language change have tended to be descriptive in nature (though significant exceptions are the work of Lightfoot, Kroch, Clark and Roberts, among others). In contrast, we place the study of historical linguistics (diachronic phenomena) on a scientific[17] platform. In this sense, our conception of historical linguistics is closest in spirit to evolutionary theory and population biology[18] (which attempts to describe the origin and changing patterns of life) and cosmology (which attempts to describe the origin and evolution of the physical universe).

Second, it allows us to formally pose a *diachronic* criterion for the adequacy of grammatical theories. A significant body of work in learning theory, has already sharpened the *learnability* criterion for grammatical theories—in other words, the class of grammars $\mathcal{G}$ must be learnable by some psychologically plausible algorithm from primary linguistic data. Now we can go one step further. The class of grammars $\mathcal{G}$ (along with a proposed learning algorithm $\mathcal{A}$) can be reduced to a dynamical system whose evolution must match that of the true evolution of human languages (as reconstructed from historical data).

We have attempted to lay the framework for the development of research tools to study historical phenomena. To concretely demonstrate that the grammatical dynamical systems need not be impossibly difficult to compute (or simulate), we explicitly showed how to transform parametrized theories, and memoryless learning algorithms to dynamical systems. The specific simulations of this paper are far too incomplete to have any long term linguistic implications, though, we hope, it certainly forms a starting point for research in this direction. Nevertheless, there were certain interesting results obtained.

1. We saw that the V2 parameter was more stable in the 3-parameter case, than it was in the 5 parameter case. This suggests that the loss of V2 (actually observed in history) might have more to do with the choice of parametrizations than learning algorithms, or primary linguistic data (though, we suggest great caution, before drawing strong conclusions on the basis of this study).

2. We were able to shed some light on the time course of evolution. In particular, we saw how this was a derivative of more fundamental assumptions about initial population conditions, sentence distributions, and learning algorithms.

3. We were able to formally develop notions of system stability. Thus, certain parameters could change with time, others might remain stable. This can now be measured, and the conditions for stability or change can be investigated.

4. We were able to demonstrate how one could tinker with the system (by changing the algorithm, or the sentence distributions, or maturational time) to allow evolution in certain directions. This would suggest the kinds of changes needed in linguistics for greater explanatory adequacy.

Further Research

This has been our first attempt to define the boundaries of the problem. There are several directions of further research.

1. From a linguistic perspective, the most interesting thing to do, would perhaps be the examination of alternative parametrized theories, and to track the change of certain languages in the context of these theories (much like our attempt to track the change of French in this paper). Some worthwhile attempts would include a) the study of parametric stress systems (Halle and Idsardi, 1992)–and in particular, the evolution of modern Greek stress patterns from proto-Indo European; b) the investigation of the possibility that creoles correspond to fixed points in parametric dynamical systems, a possibility which might explain the striking fact that all creoles (irrespective of the linguistic origin, i.e., initial linguistic composition of the population) have the same grammar; c) the evolution of modern Urdu, with Hindi syntax, and Persian vocabulary.

2. From a mathematical perspective, one could take this research in many directions including a) the formal-

[17] By scientific, we mean, the construction of models with explanatory, and predictive powers– models which can be falsified in the sense of Popper.

[18] Indeed, most previous attempts to model language change, like that of Clark and Roberts (1993), and Kroch (1990) have been influenced by the evolutionary models.

ization of the update rule for other grammatical theories and learning algorithms, and the characterization of the dynamical systems implied therein b) the investigation of stability issues more closely, and characterizing better the phase-space plots c) recall that our dynamical systems are multi-dimensional non-linear iterated function mappings—a recipe for chaotic behaviour, and a possibility to investigate further.

It is our hope that research in this line will mature to make useful contributions, both to linguistics, and in view of the unusual nature of the dynamical systems involved, to the study of such systems from a mathematical perspective.

References

[1] R. C. Berwick. *The Acquisition of Syntactic Knowledge.* MIT Press, 1985.

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

[3] G. Altmann et al. A law of change in language. In B. Brainard, editor, *Historical Linguistics*, pages 104–115, Studienverlag Dr. N. Brockmeyer., 1982. Bochum, FRG.

[4] E. Gibson and K. Wexler. Triggers. *Linguistic Inquiry*, 25, 1994.

[5] L. Haegeman. *Introduction to Government and Binding Theory.* Blackwell: Cambridge, USA, 1991.

[6] A. S. Kroch. Grammatical theory and the quantitative study of syntactic change. In *Paper presented at NWAVE 11, Georgetown Universtiy*, 1982.

[7] A. S. Kroch. Function and gramar in the history of english: Periphrastic "do.". In Ralph Fasold, editor, *Language change and variation.* Amsterdam:Benjamins. 133-172, 1989.

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

[9] D. Lightfoot. *How to Set Parameters.* MIT Press, Cambridge, MA, 1991.

[10] B. Mandelbrot. *The Fractal Geometry of Nature.* New York, NY: W. H. Freeman and Co., 1982.

[11] P. Niyogi. *The Informational Complexity of Learning From Examples.* PhD thesis, Massachussetts Institute of Technology, Cambridge, MA, 1994.

[12] P. Niyogi and R. C. Berwick. Formal models for learning finite parameter spaces. In P. Broeder and J. Murre, editors, *Models of Language Learning: Inductive and Deductive Approaches*, chapter 14. MIT Press; to appear, Cambridge, MA.

[13] P. Niyogi and R. C. Berwick. Formalizing triggers: A learning model for finite parameter spaces. Tech. Report 1449, AI Lab., M.I.T., 1993.

[14] P. Niyogi and R. C. Berwick. A markov language learning model for finite parameter spaces. In *Proceedings of 32nd meeting of Association for Computational Linguistics*, 1994.

[15] C. Osgood and T. Sebeok. Psycholinguistics: A survey of theory and research problems. *Journal of Abnormal and Social Psychology*, 49(4):1–203, 1954.

[16] S. Resnick. *Adventures in Stochastic Processes.* Birkhauser, 1992.

[17] S. Strogatz. *Nonlinear Dynamics and Chaos.* Addison-Wesley, 1993.

[18] U. Weinreich, W. Labov, and M. I. Herzog. Empirical foundations for a theory of language change. In W. P. Lehmann, editor, *Directions for historical linguistics: A symposium.*, pages 95–195. Austin: University of Texas Press, 1968.

A The 3-parameter system of Gibson and Wexler (1994)

The 3-parameter system discussed in Gibson and Wexler (1994) includes two parameters from X-bar theory. Specifically, they relate to specifier-head relations, and head-complement relations in phrase structure. The following parmetrized production rules denote this:

$$XP \rightarrow SpecX'(p_1 = 0) \text{ or } X'Spec(p_1 = 1)$$
$$X' \rightarrow CompX'(p_2 = 0) \text{ or } X'Comp(p_2 = 1)$$
$$X' \rightarrow X$$

A third parameter is related to verb movement. In German, and Dutch root declarative clauses, it is observed that the verb occupies exactly the second position. This Verb-Second phenomenon might or might not be present in the world's languages, and this variation is captured by means of the V2 parameter.

The following table provides the unembedded (degree-0) sentences from each of the 8 grammars (languages) obtained by setting the 3 parameters of example 1 to different values. The languages are referred to as L_1 through L_8.

Language	Spec	Comp	V2	Degree-0 unembedded sentences
L_1	1	1	0	"V S" "V O S" "V O1 O2 S" "AUX V S" "AUX V O S" "AUX V O1 O2 S" "ADV V S" "ADV V O S" "ADV V O1 O2 S" "ADV AUX V S" "ADV AUX V O S" "ADV AUX V O1 O2 S"
L_2	1	1	1	"S V" "S V O" "O V S" "S V O1 O2" "O1 V O2 S" "O2 V O1 S" "S AUX V" "S AUX V O" "O AUX V S" "S AUX V O1 O2" "O1 AUX V O2 S" "O2 AUX V O1 S" "ADV V S" "ADV V O S" "ADV V O1 O2 S" "ADV AUX V S" "ADV AUX V O S" "ADV AUX V O1 O2 S"
L_3	1	0	0	"V S" "O V S" "O2 O1 V S" "V AUX S" "O V AUX S" "O2 O1 V AUX S" "ADV V S" "ADV O V S" "ADV O2 O1 V S" "ADV V AUX S" "ADV O V AUX S" "ADV O2 O1 V AUX S"
L_4	1	0	1	"S V" "O V S" "S V O" "S V O2 O1" "O1 V O2 S" "O2 V O1 S" "S AUX V" "S AUX O V" "O AUX V S" "S AUX O2 O1 V" "O1 AUX O2 V S" "O2 AUX O1 V S" "ADV V S" "ADV V O S" "ADV V O2 O1 S" "ADV AUX V S" "ADV AUX O V S" "ADV AUX O2 O1 V S"
L_5 (English, French)	0	1	0	"S V" "S V O" "S V O1 O2" "S AUX V" "S AUX V O" "S AUX V O1 O2" "ADV S V" "ADV S V O" "ADV S V O1 O2" "ADV S AUX V" "ADV S AUX V O" "ADV S AUX V O1 O2"
L_6	0	1	1	"S V" "S V O" "O V S" "S V O1 O2" "O1 V S O2" "O2 V S O1" "S AUX V" "S AUX V O" "O AUX S V" "S AUX V O1 O2" "O1 AUX S V O2" "O2 AUX S V O1" "ADV V S" "ADV V S O" "ADV V S O1 O2" "ADV AUX S V" "ADV AUX S V O" "ADV AUX S V O1 O2"
L_7 (Bengali, Hindi)	0	0	0	"S V" "S O V" "S O2 O1 V" "S V AUX" "S O2 O1 V AUX" "ADV S V" "ADV S O V" "ADV S O2 O1 V" "ADV S V AUX" "ADV S O V AUX" "ADV S O2 O1 V AUX"
L_8 (German, Dutch)	0	0	1	"S V" "S V O" "O V S" "S V O2 O1" "O1 V S O2" "O2 V S O1" "S AUX V" "S AUX O V" "O AUX S V" "O1 AUX S O2 V" "O2 AUX S O1 V" "ADV V S" "ADV V S O" "ADV V S O2 O1" "ADV AUX S V" "ADV AUX S O V" "S AUX O2 O1 V" "S O V AUX" "ADV AUX S O2 O1 V"