The proper treatment of language acquisition and change in a population setting

Partha Niyogi[a] and Robert C. Berwick[b,1]

[a]Department of Computer Science and Department of Statistics, University of Chicago, Chicago, IL 60637; and [b]Department of Electrical Engineering and Computer Science and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by A. Noam Chomsky, Massachusetts Institute of Technology, Cambridge, MA, April 13, 2009 (received for review December 27, 2008)

Language acquisition maps linguistic experience, primary linguistic data (PLD), onto linguistic knowledge, a grammar. Classically, computational models of language acquisition assume a single target grammar and one PLD source, the central question being whether the target grammar can be acquired from the PLD. However, real-world learners confront populations with variation, i.e., multiple target grammars and PLDs. Removing this idealization has inspired a new class of population-based language acquisition models. This paper contrasts 2 such models. In the first, iterated learning (IL), each learner receives PLD from one target grammar but different learners can have different targets. In the second, social learning (SL), each learner receives PLD from possibly multiple targets, e.g., from 2 parents. We demonstrate that these 2 models have radically different evolutionary consequences. The IL model is dynamically deficient in 2 key respects. First, the IL model admits only linear dynamics and so cannot describe phase transitions, attested rapid changes in languages over time. Second, the IL model cannot properly describe the stability of languages over time. In contrast, the SL model leads to nonlinear dynamics, bifurcations, and possibly multiple equilibria and so suffices to model both the case of stable language populations, mixtures of more than 1 language, as well as rapid language change. The 2 models also make distinct, empirically testable predictions about language change. Using historical data, we show that the SL model more faithfully replicates the dynamics of the evolution of Middle English.

dynamical system model | learnability | phase transitions | social learning | iterated learning

At least from the seminal work of Gold (1), there has been a tradition of inquiry into computational models of language acquisition. Much of this work has considered a learner acquiring a target grammar from its linguistic experience interacting with speakers of this grammar, generally positing that this experience is consistent with a single target grammar (2, 3). Such models assume an idealized, homogeneous speaker–hearer population. Over the past 15 years computational models have relaxed this homogeneity assumption to confront the reality of language variation.* Clark and Roberts (7) and Niyogi and Berwick (8) are representative early attempts. This revision is marked by 2 developments. First, learning occurs in a population setting with potential variation in the attained grammars or languages of its members. Second, language acquisition is the mechanism by which language is transmitted from the speakers of one generation to the next. We call this newer formulation the population view of language acquisition and change, and the resulting models social learning or (SL) models.

This shift from single-source grammar acquisition to a population view parallels the epistemological shift in Darwin's introduction of population level reasoning in *Origin of Species*. Darwin argued that variation was essential to biological evolution on 2 levels—variation in the parental generation and variation in the offspring generation (9). We claim this as well for language variation and change. However, there are differences. Instead of trait inheritance from parents alone, we model language as possibly acquired from the population at large. This formalization is not identical to conventional mathematical population biology, because it generalizes inheritance to the notion of an acquisition algorithm, roughly, any computable function from data to grammars. (We can still use classical population biology as a special case.) We note that while Cavalli-Sforza and Feldman (10) have also formulated similar extended models of inheritance, allowing offspring traits to be acquired from non-parents (including the possibility of "horizontal" transmission, that is, from learners in the same generation), their model varies in certain crucial ways from what we propose here, and we obtain distinct results, most notably because we use the mechanism of an acquisition algorithm rather than just an extended Mendelian inheritance scheme, as we elaborate in more detail in *From Language Acquisition to Language Evolution*.

It follows immediately that models not admitting multiple learning sources, assuming instead a single source of learning input and then iterating that single learner over multiple generations, do not embrace the full Darwinian variational picture or our extension of it. Because such single-learner, or so-called iterated learning (IL), models have gained some currency [as developed in a series of papers by Kirby (11) and Griffiths, Dowman, and Kirby (12)], this distinction is of some importance to point out.†

The goal of this article is to examine the importance of the population view, using as a touchstone the theoretical and empirical contrasts between the SL and IL models. In the IL models, individual learners are immersed in a population and each such individual learns from only 1 person. In the SL models, each learner is exposed to data from multiple individuals in a community. There are important theoretical and empirical differences between both models. When individuals learn from a single source, the evolutionary dynamics that results is necessarily linear, leading to a single stable equilibrium from all initial conditions. In contrast, in SL models the dynamics is potentially nonlinear. Bifurcations arise only in SL models. These are linguistically interpretable and as described in *Empirical Linguistic Applications and Discussion* provide an explanatory construct for the patterns of rapid change often observed in historical linguistics and not apparent from either previous linguistic models or the IL models. In comparison, although ref. 10 also yields quadratic dynamical systems that are in principle capable of modeling bifurcations, it appears that in many natural linguistic contexts and parameter regimes of interest, bifurcations do not arise. If so, then the models in ref. 10 also do not properly accommodate rapid historical language

Author contributions: P.N. and R.C.B. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

*Of course, variation has always been a central concern in the sociolinguistics tradition; see refs. 4–6 for typical examples.

†The term "iterated learning model" has come to be associated only with a certain type of dynamical model, as in ref. 11, in which one considers a chain of learners, each of whom learns from the output of the previous one. It is worth noting that in fact all models of language change and evolution "iterate" the learners from one generation to the next. However, for the purposes of this article we will continue to use the term IL to refer to the kinds of models described in ref. 11.

[1]To whom correspondence should be addressed. E-mail: berwick@csail.mit.edu.

change (for more discussion on this point, see ref. 13). Finally, SL models, but not IL models, appear to make the right empirical predictions about observable cases of language change.

In our mathematical developments, we make certain idealizations (infinite populations, random mixing, and the like). These idealizations provide simplifying assumptions that allow us to reason coherently about the subtle interplay between learning by individuals and the evolution of populations. These assumptions are made in the same spirit as in conventional population genetics, where the 1-gene, 2-allele infinite population random mating models provide a starting point for the analysis of Mendelian inheritance, Hardy–Weinberg laws, and other basic results. Here too we present the basic results, noting that the assumptions must, of course, be relaxed as our understanding deepens.

Language Acquisition: A Population Framework

To formalize the IL and SL models, let $\mathcal{G}$ be a family of grammatical systems where each $g \in \mathcal{G}$ is a possible target grammar. Each such $g \in \mathcal{G}$ yields a potentially infinite set of expressions $L_g \in \Sigma^*$, generated by g. Not all expressions are produced with the same frequency by users of g. We denote by P_g a probability distribution over the set L_g such that $s \in L_g$ is produced with probability $P_g(s)$ by speakers of g.

Consider the state of a potentially heterogeneous population at time t. This may be characterized via a probability distribution $P^{(t)}$ on $\mathcal{G}$ where $P^{(t)}(g)$ is the proportion of the population using grammar g. Thus a homogeneous population where everyone uses grammar h places probability mass 1 on h and 0 elsewhere. A heterogeneous population with equal numbers of g and h users could similarly be characterized by a distribution $P^{(t)}$ where $P^{(t)}(h) = P^{(t)}(g) = \frac{1}{2}$. Finally, consider a child/learner immersed in such a population. The learner is an acquisition algorithm that maps linguistic experience onto linguistic knowledge, expressed formally as

$$\mathcal{A} : \cup_{i=1}^{\infty}(\Sigma^*)^i \to \mathcal{G}. \qquad [1]$$

Thus if the acquisition algorithm receives a sequence of expressions $s_1, s_2, s_3, \ldots, s_n$, where each $s_i \in \Sigma^*$, then linguistic experience may be expressed as the ordered n-tuple $D = (s_1, \ldots, s_n) \in (\Sigma^*)^n$, belonging to the domain of the map $\mathcal{A}$, with $\mathcal{A}(D)$ being the grammar acquired after exposure to this experience.[‡]

This framework admits considerable generality. The linguistic experience D could come from either a single source or multiple sources. Similarly, since $\mathcal{A}$ can be any algorithm and $\mathcal{G}$ any family of computational systems, the framework models a wide range of learning options and most current linguistic theories. For example, it accommodates the class of probabilistic context-free grammars (with $\mathcal{G}$ a family of probability distributions), along with probabilistic learning techniques, including Bayesian learning procedures and current 'corpus based' methods. With respect to linguistic theories, setting $\mathcal{G} = \{g_1, \ldots, g_N\}$ yields a finite number N of possible target grammars; this is compatible with the current phonological approach known as Optimality theory (14), positing a universal, fixed finite number of "ranking constraints" underlying all human sound systems. It is also compatible with the so-called "principles and parameters" theory of syntax known as Government-Binding theory (15). More generally, this framework is compatible with nearly every current parameterized linguistic theory, from head-driven phrase structure grammar (16) to lexical–functional grammar (17).

From Language Acquisition to Language Evolution

When we iterate the map $\mathcal{A}$ we can explicitly calculate how language will evolve over generational time as learners acquire their language(s) from the primary linguistic data (PLD) given by each previous generation. We now discuss 2 distinct scenarios in which this question may be posed.

Iterated Learning. First, one can assume that the PLD comes only from a single source. IL begins with a single agent and a grammar $g_1 \in \mathcal{G}$. The agent then produces n example sentences, the linguistic experience $D = (s_1, \ldots, s_n)$ for the learner; the learner applies the map $\mathcal{A}$ to attain a mature grammar $\mathcal{A}(D) = g_2 \in \mathcal{G}$ and proceeds to produce D for a single agent in the next generation, iteratively, yielding the sequence $g_1 \to g_2 \to g_3 \to \ldots$. It is easily seen that this yields a single trajectory corresponding to an underlying Markov chain whose state space is $\mathcal{G}$, with the chain's state denoted by $g_t \in \mathcal{G}$ at each point in time t. We can characterize the transition matrix[§] of this chain: For any $g \in \mathcal{G}$ and $h \in \mathcal{G}$, the probability of mapping from g to h is given by

$$T[g,h] = prob[g \to h] = prob[\mathcal{A}(D) = h | D \text{ generated by } P_g], \qquad [2]$$

where $T[g,h]$ is the probability the learner would acquire h given data D provided to it by an agent using the grammar g and generating the primary linguistic data D according to P_g.

IL has been considered in a series of papers (18), most recently in a setting where the algorithm $\mathcal{A}$ uses a Bayesian estimation scheme (12). However, irrespective of the particular learning algorithm $\mathcal{A}$, i.e., whether it is Bayesian estimation, parametric learning, etc., the evolutionary trajectory of IL is always characterized by a Markov chain under fairly general assumptions. The exact values of the transition probabilities will of course depend on the nature of the learning algorithm.

Furthermore, there is a natural population interpretation of IL. Consider a population whose initial state is provided by $P^{(0)}$ (a distribution on $\mathcal{G}$) such that $P^{(0)}(g)$ is the proportion of g grammar users in generation 0. Then according to the Markov dynamics given above, this distribution will evolve as

$$P^{(t+1)}(h) = \sum_{g \in \mathcal{G}} P^{(t)}(g) T[g,h]. \qquad [3]$$

According to this equation, in the IL model the distribution of speakers of different grammars in the population must inevitably evolve according to a linear rule. Although each learner is immersed in a potentially heterogeneous population, each learns only from a single individual, never reflecting the population variation. Different learners, of course, learn from potentially different individuals.[¶]

There are three further critical conclusions to be drawn from the Markovian property of IL:

1. IL's linear dynamics converges to a single stable equilibrium from all initial conditions, given by the leading left eigenvector

[‡] Following the mainstream literature in language acquisition models, especially in the generative linguistics tradition, we have assumed that children learn from positive examples alone and that the effects of feedback and active linguistic experimentation are minimal. Furthermore, note that ultimately we will need to characterize the probability with which different PLD D will be experienced by the learner. Here we will assume that $P(D)$ may be factored into the product $P(D) = \prod_{i=1}^{n} P(s_i)$, i.e., each speaker of g produces sentences independently and identically distributed (i.i.d) according to P_g. Of course, P_g itself might assign more probability mass to "simpler" sentences according to, perhaps, a "Zipfian" law, although we make no particular assumptions about the nature of P_g. This i.i.d. assumption allows us to capture the intuition that not all sentences are equally likely, and this nonuniformity in sentence probabilities will in turn affect the likelihood of different kinds of linguistic experiences. One may rightly question the i.i.d. assumption but we argue that it is a simple and natural base assumption that captures the first-order frequency effects that behaviorists have discussed and is further consistent with the base assumptions of statistical learning theory as well as many prominent probabilistic models of language acquisition that provide a frame of reference for this paper. The consequences of more complex assumptions about input distributions may then be compared against this base assumption.

[§] If $|\mathcal{G}| = \infty$, then this corresponds to an operator on an infinite space.

[¶] Note that this population version of iterated learning from a single teacher is also a special case of the language evolution equation of Nowak, Komarova, and Niyogi with no fitness function (19).

of the T matrix in the usual way. If the Markov chain is non-ergodic, conventional Markov chain theory provides further characterization (20).
2. Such linear dynamics precludes the possibility of any bifurcations. To the extent that bifurcations are assumed to be a necessary part of modeling language change, the IL model is thus dynamically insufficient.
3. Conversely, the IL framework cannot properly model what one might think to be a common empirical linguistic situation, namely, language stability. A simple example illustrates this point. Suppose there are just 2 possible languages (equivalently grammars), L_1 and L_2, e.g., English and French, and assume any possible initial proper distribution of speakers over the two, e.g., from 99.9% speaking English and only 0.1% speaking French, to the reverse. Suppose each language is acquired with some small probability of error, ϵ. The transition matrix entry $T(2,1)$ gives the probability of an individual being exposed to language 2 but learning language 1 instead, $(=\epsilon)$ while $T(1,2)$ is the converse. Referring to equations 10 and 13 below, the fixed point of this dynamical system is $\epsilon/(\epsilon+\epsilon) = 1/2$. That is, the resulting steady-state will have 50% English speakers and 50% French speakers. *Although each language is effectively learnable (with probability* $1-\epsilon$*), a homogeneous community cannot be maintained and degenerates to a mixture of grammatical types over time.*

Social Learning. We turn now to a second approach to learning in a population, where the PLD for a learner can come from multiple speakers/parents, or what we call SL, as first outlined in (21). In this framework, the evolutionary dynamics of linguistic populations may be derived as follows. Let $P^{(t)}$ be the linguistic state of the population at time t. Consider a typical child learner in this population, provided with data drawn from a distribution μ_t given by the following equation:

$$\mu_t = \sum_g P^{(t)}(g) P_g. \qquad [4]$$

Here we assume that the population is "perfectly mixed" with no network effects, i.e., the typical child is exposed to different types of speakers with probabilities proportional to the fractions of these types in the population at large. The probability that a typical such learner acquires the grammar h is then given by

$$\text{prob}[\mathcal{A}(D) = h | D \text{ drawn according to } \mu_t]. \qquad [5]$$

If this is the probability with which a typical learner acquires h, then this is the proportion of h speakers in the next generation. Thus,

$$P^{(t+1)}(h) = \text{prob}[\mathcal{A}(D) = h | D \text{ drawn according to } \mu_t], \qquad [6]$$

which yields the map

$$f_{\mathcal{A}} : \mathcal{S} \to \mathcal{S}, \qquad [7]$$

where $\mathcal{S}$ is the state-space of possible linguistic populations. Each $s \in \mathcal{S}$ corresponds to a probability distribution over $\mathcal{G}$ with $s(g)$ denoting the proportion of the population using g. This model is critically distinct from IL:

1. In contrast to IL's linear dynamics, here the iterated map $s_{t+1} = f(s_t)$ is generically nonlinear and therefore can exhibit a far richer set of possible outcomes.
2. *In particular, as parameters change continuously, there may be discontinuous changes (bifurcations) in the dynamics that lead to qualitatively different regimes of equilibria.* Multiple stable states are possible. Shared grammatical systems in a homogenous population may go from stability to instability and vice versa.
3. For every learning algorithm $\mathcal{A}$, there exists a corresponding evolutionary dynamics $f_{\mathcal{A}}$ (the converse is not true; for every dynamical map, there does not necessarily exist a learning algorithm). Thus different learning algorithms may have different evolutionary consequences and may be distinguished from each other on this basis. Thus our framework is general and the empirical content arises from choosing different particular learning algorithms, and comparing their predicted evolutionary trajectories against the historically observed trajectories of change.

Comparing Iterated and Social Learning

We now proceed to illustrate the distinctions between these two very different kinds of evolutionary dynamics through a concrete example. We will assume $\mathcal{G} = \{g_1, g_2\}$, i.e., just 2 languages (or grammars). While seemingly oversimplified, this permits analytical results and corresponding insight, while retaining applicability to real language change, an idealization analogous to that of 1-gene 2-allele models in population genetics.

Given this simplification, it suffices to characterize the state-space of the linguistic population by a number $\alpha \in [0,1]$ denoting the proportion of g_1 users. The SL model then leads to a map $\alpha_{t+1} = f_{\mathcal{A}}(\alpha_t)$ that is typically nonlinear, may have linguistically interpretable bifurcations, and depends on the learning algorithm $\mathcal{A}$. The linear dynamics corresponding to the IL model can now be derived from this more general map $f_{\mathcal{A}}$ as follows. The entries of the transition matrix T leading from one generation to the next are given by

$$T(1,1) = f_{\mathcal{A}}(1) \text{ and } T(2,2) = 1 - f_{\mathcal{A}}(0) \qquad [8]$$

Because T is a stochastic matrix (with rows summing to 1), the other entries are immediately specified, with the corresponding linear dynamics given by the following equation:

$$\alpha_{t+1} = (T(1,1) - T(2,1))\alpha_t + T(2,1). \qquad [9]$$

These dynamics converge to a fixed point given by

$$\alpha_* = \frac{T(2,1)}{T(2,1) + (1 - T(1,1))}. \qquad [10]$$

Using this derivation, we can now describe precisely the differences between the IL and SL models given a particular choice of a learning algorithm.

A Cue-Based Learner. To complete a comparison between the IL and SL models and actually carry out a computation of the resulting dynamical systems, we must specify a particular learning algorithm, $\mathcal{A}$. As noted, there are many plausible choices for $\mathcal{A}$. For concreteness, let us posit a particular algorithm $\mathcal{A}$, known as cue-based learning, which has been advanced by linguists and psycholinguists, in particular refs. 22 and 23. A cue-based learner examines the PLD for surface evidence of a particular linguistic parameter setting, for example, whether a language is head-initial or verb-initial (like English, with objects following verbs), or verb-final (like German or Japanese). Such structured examples, particular analyzed sentences of a language that steer the learner towards one language or another, are called cues. For example, the (structurally analyzed) sequence *eat ice-cream*, where *eat* is a verb and begins a verb phrase, and *ice-cream* is an object of *eat*, roughly in the form [$_{VP}$ Verb Object], can be said to be a cue for languages that follow verb–object order.

To formalize this intuition in a general way, let a set $C \subseteq (L_1 \backslash L_2)$ be a set of examples that are cues to the learner that L_1 is the target language. If such cues occur often enough in the learner's data set, the learner will choose L_1, otherwise the learner chooses L_2. In particular, let the learner receive K examples. Out of the K examples, say k are from the cue set. Then, if $\frac{k}{K} > \tau$ the learner

chooses L_1, otherwise the learner chooses L_2 (following ideas in ref. 22).

Evolutionary Dynamics. The evolutionary dynamics of a population based on this $\mathcal{A}$ can then be computed as follows. Let $P_1(C) = p$; p is the probability with which an L_1 user produces a cue. If a proportion α_t of adults use L_1, then the probability with which a cue is presented to a learner is $p\alpha_t$, so the probability with which $k > K\tau$ is as follows:[‖]

$$\sum_{K\tau \le i \le K} \binom{K}{i} (p\alpha_t)^i (1 - p\alpha_t)^{(K-i)}. \qquad [11]$$

Therefore,

$$\alpha_{t+1} = \sum_{K\tau \le i \le K} \binom{K}{i} (p\alpha_t)^i (1 - p\alpha_t)^{(K-i)}, \qquad [12]$$

where α_t is the proportion of L_1 users in the tth generation. The evolutionary trajectory is defined by an iterated map that is a polynomial of degree K and the behavior depends on the value of the parameter p. An analysis reveals the following (proof omitted here):

1. For $p \in [0, 1]$, there are never more than three fixed points in $[0, 1]$. For small values of p, $\alpha = 0$ is the only fixed point of the population and it is stable.
2. As p increases, eventually a bifurcation occurs as two new fixed points arise. There are then three fixed points in all: $\alpha = 0$, which remains stable; $\alpha = \alpha_1$, which is unstable; and $\alpha = \alpha_2 > \alpha_1$, which is stable.
3. For $p = 1$, there are two stable fixed points ($\alpha = 0$ and $\alpha = 1$) and there is exactly one unstable fixed point in $(0, 1)$.

See Fig. 1 for examples of phase digrams for different values of K. Let us reflect on the bifurcation and the directionality of change this dynamical model implies.

First, note that if the target grammar is g_2 it will always be acquired. As a result, a population composed entirely of g_2 speakers will remain stable for all time for all values of p.

Second, consider a community composed entirely of g_1 users. As one can see from Fig. 1, there is a regime $p > p_{critical}$ where the population moves from a completely homogeneous one to a stable equilibrium composed mostly of g_1 users, with a small proportion of g_2 users thrown in. However, when $p < p_{critical}$, then this equilibrium rather abruptly disappears and the population moves to one composed entirely of g_2 users. *Thus, a tiny difference in the frequency with which speakers produce cues can dramatically alter the stability profile of a linguistic community.* This extreme sensitivity to parameter values is the hallmark of bifurcations and does not follow straightforwardly without a formal analysis. As we have seen, it also does not arise in the IL model.

It is also possible to show that a similar bifurcation arises if one holds p constant but changes K, the total size of the PLD. A decrease in the value of K leads to effects similar to that of the frequency (p) with which speakers produce cues.

In addition to sensitivity, we also see directionality. A population of g_1 speakers can change to a population of g_2 speakers; a change in the reverse direction is not possible.

Given $f_{\mathcal{A}}$ for the SL model where $\mathcal{A}$ is the cue based learner, the IL dynamics for the same learner is easily characterized.

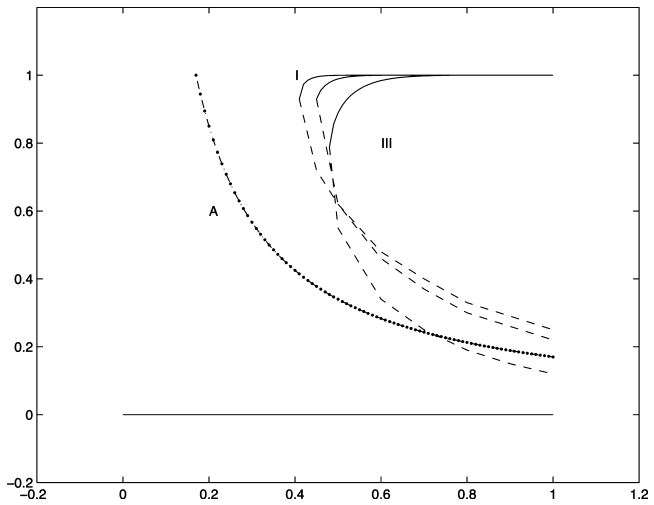

Fig. 1. Phase diagram curves indicating how the fixed points for the proportion of V2 speakers vary given different cue-frequencies p in the SL model and how these relate to the historically assumed number of cues available (17% or more) and the possible loss of Verb-second. The horizontal x axis denotes the frequency p of cues for a V2 language. The vertical y axis denotes the fraction of the population speaking a V2 language. Curves I–III depict the corresponding stable and unstable fixed points for the percentage of V2 speakers as a function of cue-frequency p for three different choices of K, the total number of sentences in the PLD. The leftmost curve (labeled I) corresponds to $K = 70$, while curves further to the right depict decreasing values for K, with the rightmost, third (III) curve corresponding to $K = 10$; the curve lying between I and III (II; label omitted) corresponds to $K = 30$. In each case the solid portions of the curve denote stable fixed points, while the dashed portions are unstable. For example, given $K = 70$, as soon as p falls below ≈0.5 then the percentage of V2 speakers begins to decline. Note how for small values of p there is only one stable point at $y = 0$. As p increases a new pair of fixed points arises—one of which is unstable (dotted) and the other stable (solid). The curve labeled A denotes the historically attested value where the number of cues is at 17%, i.e., $yp = 0.17$. Note that all three stability curves (I–III) lie almost entirely above and to the right of curve A. Because curve A lies in the basin of attraction of $y = 0$, we see that if one is on this curve at any point in time, the system will evolve to $y = 0$, i.e., V2 will be eliminated.

$$T(1, 1) = f_{\mathcal{A}}(1) < 1 \text{ and } T(2, 2) = 1 - f_{\mathcal{A}}(0) = 1. \qquad [13]$$

As expected, the result is a linear dynamics, and, from all initial conditions and for all p, the population evolves to a community of all g_2 users. There are no bifurcations and g_1 can never be stably maintained in the population under any circumstances. Thus this model results in a more restricted range of possible outcomes; as we shall see, this limits the IL model in its empirical adequacy, a matter to which we turn next in the concluding section.

Empirical Linguistic Applications and Discussion

Turning from mathematical modeling to empirical data, in this concluding section we illustrate how to use the modeling paradigm described previously to sharpen our understanding of language learning and change, providing an explanatory account for certain linguistic phenomena. We develop the data and analysis according to our own reading of ref. 22.[**]

Our key finding is that the SL model with a cue-based learning algorithm correctly predicts an observed "phase transition" that took place in the case of an historically verified two language contact situation in the development of early English, resulting in the rapid loss of a certain type of syntactic construction and a corresponding shift to an apparently stable equillibrium lacking this

[‖] Note here that we make crucial use of the i.i.d. assumption regarding the PLD. Although this might at first seem unrealistic, it is worth remarking that in the analysis, sentences are grouped into equivalence classes based on structural (syntactic) properties, i.e., whether they are cues or not. What is really being assumed is that structures are i.i.d. This is almost always the case in any real linguistic application of our analysis. It seems plausible at face value that, although surface forms of sentences instantiated by the lexical choices may be correlated, the structural forms are less so and in fact are independent.

[**] We are aware that there are alternative linguistic accounts (notably as in refs. 24 and 25, among others). The general approach presented here may be used to formally model these as well but a full treatment is beyond the scope of the current paper.

construction. In contrast, as far as we can determine, the IL model with its linear dynamics cannot successfully make this prediction.

The Data: Loss of V2 in English. The particular historical data and analysis is as follows. English was once what linguists term a "verb second" language, (abbreviated V2), just as Dutch and German are today, but English is no longer V2, stably so. We outline the basic linguistics. Superficially, a V2 language allows verbal elements bearing tense in exactly the second structural position in a sentence (besides its normal position), while a non-V2 language does not. Examples such as the following illustrate, with verbs highlighted by -Verb. The examples are taken from ref. 22.

1. Wij zagen-Verb vel studenten in Amsterdam ("We saw many students in Amsterdam")
2. Vele studenten zagen-Verb wij in Amsterdam ("Many students saw we in Amsterdam"); note that *zagen* is in exactly the second position after the phrase "many students".
3. In Amsterdam zagen-Verb wij vele studentent ("In Amsterdam saw we many students"); again *zagen* is exactly in the second position after the first phrase "in Amsterdam".

In contrast, in a non-V2 language the second position may be filled by a nonverbal element as in modern English and French, and an example like example 2 above is not attested, for example, "In Amsterdam we saw many students", where "we" occupies the second structural position in the sentence.

Our reading of ref. 22, covering the basic historical facts and linguistic analysis relating to the change in an older form of English, Middle English, from a V2 language to a non-V2 language during roughly the 13th century is as follows.

1. There were two English-like languages in this historical contact situation: (*i*) Northern (Middle) English, and (*ii*) Southern (Middle) English. We label the 2 sets of sentences relevant for the cue-based distinction between these 2 languages L_{north} and L_{south}, structurally analyzed as needed by the cue-based approach.
2. Northern (Middle) English was a V2 language; limited to subjects with pronouns (we, she…), Southern (Middle) English was a non-V2 language.
3. The grammars for the 2 languages generated different cues. We denote the cues appearing only in the Northern, V2 language by the set-difference $L_{\text{north}} \setminus L_{\text{south}}$. These include sentences with initial phrases of any category apart from subjects, followed by verbal elements in second position, then pronouns, a sequence not possible in a non-V2 language, and not found in modern English. Such cue sentences include the following, glossed in modern English: "In Amsterdam saw we many students."
4. The remainder of the model depends on sentences in the intersection $L_{\text{north}} \cap L_{\text{south}}$ and $L_{\text{south}} \setminus L_{\text{north}}$, that is, sentences in both V2 and non-V2 languages (e.g., "In Amsterdam saw John the students") and cue sentences only in the Southern language ("In Amsterdam we saw many students").[††]
5. According to ref. 22, we assume that if cues occur with a sufficient frequency threshold, τ, children acquire the V2 grammar, otherwise a non-V2 grammar. On this account, when speakers of Northern Middle English came into contact with speakers of Southern Middle English, the Northern speakers stopped hearing a sufficient number of cues, triggering a phase transition.

Let us now establish this claim. Based on statistics from modern Dutch, ref. 22 argues that the learning threshold for non-initial subject sentence cues is approximately 30%. Based on work of Ans van Kemenade from *Sawles Warde*, an early-13th-century text, ref. 22 estimates the actual cue percentage at 17%, noting this "evidence suggests that 17% of initial non-subjects does not suffice to trigger a verb second grammar, but 30% is enough; somewhere between 17 and 30% is a phase transition." We now proceed to confirm this hypothesis formally by using our cue-based model.

Mapping to a Dynamical Model from Cue-Based Learning. Let us adopt the conclusion in ref. 22 that cues for V2 must occur at least 30% of the time for correct acquisition. We therefore examine how the SL cue-based model behaves with $\tau = 0.3$. For a fixed K and $\tau = 0.3$, we obtain the diagram in Fig. 1, where $x=$ the frequency p of V2 cues and $y =$ the fraction of the population speaking a V2 language (so that yp is the actual percentage of V2 cues). Fig. 1 displays bifurcation curves for three different possible numbers of sentences in the PLD, $K = 10$, 30, and 70, respectively, labeled I–III (the intermediate curve II corresponding to $K = 30$ lies between I and III; its label is omitted to save space).

Now posit a community where $p = 0.44$ and 99% of the population uses a V2 grammar, corresponding to the point (x,y) in Fig. 1 with $x = 0.44$, and $y = 0.99$, which is on the curve labeled I, corresponding to exposure to 70 sentences. This is a stable point, and the V2 grammar would therefore be maintained over time. With this in mind, we examine 3 different scenarios in which a stable V2 community such as this might lose V2 over time.

Scenario 1: K decreases. The total number of input examples decreases. Then the bifurcation curves shift to the right e.g., to curve III. Here, at $p = 0.44$ there is only one stable fixed point, with $y = 0$. For this K regime, the entire population loses V2 over time.

Scenario 2: p decreases. The frequency of cue production decreases. Then the bifurcation diagram illustrates how the stability profile of the dynamics changes (for each K) as a function of p. Notably, there exists a $p_*(\tau, K)$ such that if $p < p_*(\tau, K)$, the only stable mode is $y = 0$, i.e., a homogeneous population with a non-V2 grammar. Now imagine the population were stable at a V2 level with a value of $p > p_*(\tau, K)$. If the speakers of the V2 language started producing fewer cues (a drift in cue production due to any cause, whether random or sociological) in a manner such that p drifted across the critical point p_*, then the V2 grammar would no longer be stably maintained in the population.

Scenario 3: Jumps from one attractor basin to another. In the parameter space where $p > p_*(\tau, K)$, there are 2 stable equilibria; a population can be stable with a preponderantly V2 grammar or with a preponderantly non-V2 grammar. However, each stable point has its own attractor basin. If $p > p_*(\tau, K)$, then there exists a $y_*(p, K, \tau)$ s.t. if one begins with any initial condition $y \in (y_*, 1)$, the population moves to a largely V2 grammar, while from an initial condition $y \in (0, y_*)$, the population moves to a non-V2 grammar. One can imagine a population stable at a mostly V2 grammar. Given language contact, if a sufficient number of non-V2 speakers entered the population, the population mix might be shifted, due to migration from the basin of attraction of one stable attractor to the attractor basin of the other. Thus, the population might move from a largely V2 grammar to one having a non-V2 grammar.

Given this analysis, we return to the statistics cited earlier from ref. 22: in the early 13th century, $\tau = 17\%$. There are 2 possible interpretations for this value.

First, the text was entirely written by a single author using a V2 grammar, representative of the general population, and so the value of p at this historical time point is also 0.17. Referring to Fig. 1, in such a regime there is only one equilibrium point, with $y = 0$, and the V2 grammar would be lost over time.

[††] Note this means there is an inherent asymmetry or markedness principle in that the learner does not scan for cues for non-V2 languages or compares cues for V2 languages against cues for non-V2 languages.

Second, the 17% value corresponds to the percentage input cues provided to typical children of the next generation. Because we do not know the actual fraction of V2 speakers in the population of the 13th century, y, in this case we could only conclude that $yp = 0.17$. Then the true values of y and p would lie on the curve $yp = 0.17$, plotted as the dotted line in Fig. 1, and this curve lies entirely in the attractor basin for $y = 0$ (for $K = 30$, 70 and generally for large K). Thus, although we cannot fix the exact values of y or p, we may still plausibly conclude that the (y,p) pair is in the basin of attraction of $y = 0$ and such a system given this regime would lose V2 over time, as with the first interpretation.

It is important to see why this model allows us to make an empirically testable prediction. From a single snapshot in time, we can only measure the fraction of V2 speakers (y) along with the fraction of cues p used by them, or more accurately, the proportion of cues in the PLD of typical children. From a single point in time, how are we to know whether y will increase, decrease, or remain the same in the future? Note that if one could sample y at successive time points one might be able to estimate trends and make an educated guess. However, all we have is a single point in time. In the absence of a model, no prediction is possible. However, in the context of the SL model and its assumptions, if we are able to position the possible (y,p) points with respect to a bifurcation diagram, one can in fact make a prediction. One can then compare this prediction against the historically observed trajectory. This population level evolutionary dynamics is predicted from the individual level acquisition mechanisms.

In this way we are able to make good on the call in ref. 22 for a formal phase transition analysis that is otherwise only informally available, arriving at a qualitative prediction that can be verified from the historical record. Crucially, this is possible given the dynamical regimes delimited by the SL model, but not by the IL model. In this sense one can add a new dimension of explanatory adequacy to models of linguistic change, a notion of historical explanatory adequacy previously absent from linguistic accounts, as well as more generally demonstrating that the notion of true population learning may be essential to the historical explanation of linguistic change.

ACKNOWLEDGMENTS. We thank Michael Coen, A. N. C., Morris Halle, and anonymous reviewers for helpful comments that improved the article.

1. Gold E (1967) Language identification in the limit. *Information and Control* 10:447–474.
2. Wexler K, Culicover P (1980) *Formal Principles of Language Acquisition* (MIT Press, Cambridge, MA).
3. Tesar B, Smolensky P (2000) *Learnability in Optimality Theory* (MIT Press, Cambridge, MA).
4. Labov W (1994) *Principles of Linguistic Change* (Blackwells, Cambridge, MA).
5. Mufwene SS (2001) *The Ecology of Language Evolution* (Cambridge Univ Press, Cambridge, UK).
6. Kroch A (1989) Reflexes of grammar in patterns of language change. *Language Variat Change* 1:199–244.
7. Clark R, Roberts I (1993) A computational model of language learnability and language change. *Linguistic Inq* 24:299–345.
8. Niyogi P, Berwick R (1996) A language learning model for finite parameter spaces. *Cognition* 61:161–193.
9. Lewontin R (1974) *The Genetic Basis of Evolutionary Change* (Columbia Univ Press, New York, NY).
10. Cavalli-Sforza L, Feldman M (1981) *Cultural Transmission and Evolution* (Princeton Univ Press, Princeton, NJ).
11. Kirby S (2000) in *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed Knight, C (Cambridge Univ Press, Cambridge, England), pp 303–323.
12. T. Griffiths MD, Kirby S (2007) Innateness and culture in the evolution of language. *Proc Natl Acad Sci USA* 104:5241–5245.
13. Niyogi P (2006) *The Computational Nature of Language Learning and Evolution* (MIT Press, Cambridge, MA).
14. Prince A, Smolensky P (2004) *Optimality Theory: Constraint Interaction in Generative Grammar* (Wiley–Blackwell, Hoboken, NJ).
15. Chomsky N (1980) *Lectures on Government and Binding* (Foris, Dordrecht, The Netherlands).
16. Sag IA, Wasow T, Bender EM (2003) *Syntactic Theory: A Formal Introduction* (CSLI, Stanford).
17. Bresnan J (2000) *Lexical-Functional Syntax* (Blackwell, Malden, MA).
18. Griffiths T, Kalish M (2007) Language evolution by iterated learning with Bayesian agents. *Cognit Sci* 31:441–480.
19. Nowak M, Komarova N, Niyogi P (2001) Evolution of universal grammar. *Science* 291:114–118.
20. Isaacson DL, Madsen RW (1976) *Markov Chains: Theory and Applications* (Wiley, New York).
21. Niyogi P, Berwick R (1997) Evolutionary consequences of language learning. *J Complex Sys* 11:161–204.
22. Lightfoot D (1998) *The Development of Language* (Blackwells, Malden, MA).
23. Sakas W, Fodor J (2001) in *Language Acquisition and Learnability*, ed Bertolo, S (Cambridge Univ Press, Cambridge, UK), pp 172–233.
24. Kroch A, Taylor A (1997) *Verb movement in Old and Middle English: Dialect variation and language contact* (Cambridge Univ Press, Cambridge), pp 297–325.
25. Roberts I (2007) *Diachronic Syntax* (Oxford Univ Press, Oxford).