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

Partha Niyogi^a and Robert C. Berwick^{b,1}

^aDepartment of Computer Science and Department of Statistics, University of Chicago, Chicago, IL 60637; and ^bDepartment 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.

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

of the T matrix in the usual way. If the Markov chain is non-ergodic, conventional Markov chain theory provides further characterization (20).

- 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.
- 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:

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

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

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

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,

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

which yields the map

$$f_{\mathcal{A}} : \mathcal{S} \rightarrow \mathcal{S}, \quad [7]$$

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:

- 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.
- 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 homogeneous population may go from stability to instability and vice versa.

- 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

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

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:

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

These dynamics converge to a fixed point given by

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

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 $C \subseteq (L_1 \setminus L_2)$ 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

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.