A STOCHASTIC MODEL OF LANGUAGE CHANGE THROUGH SOCIAL STRUCTURE AND PREDICTION-DRIVEN INSTABILITY

W. GARRETT MITCHENER

Abstract. We develop a new stochastic model of language learning and change that incorporates variable speech and age structure. Children detect correlations between age and speech, predict how the language is changing, and speak according to that prediction. Although the population tends to be dominated by one idealized grammar or another, the learning process amplifies chance correlation between age and speech, generating prediction-driven instability and spontaneous language change. The model is formulated as a Markov chain, and a stochastic differential equation. It exhibits the abrupt monotonic shifts among equilibria characteristic of language changes. Fundamental results are proved.

1. The paradox of language change

Language has both discrete and continuous characteristics. On the discrete side, most sentences are clearly either grammatical or ungrammatical. An *idealized grammar* is a formalism that identifies correct utterances, and much of the research on how children acquire their native language focuses on how they might choose an idealized grammar from many innate possibilities on the basis of example sentences from

Key words and phrases. language change; prediction-driven instability; population dynamics; stochastic differential equation; noise-activated transitions; AMS subjects 34K50, 37H10, 60H15, 60J20, 60J70 .

the surrounding society [1, 27, 28]. Many models of child language acquisition operate on a discrete space of idealized grammars more or less as follows: The learner holds one hypothesis at a time, changes it in response to example sentences, and in the end chooses a single idealized grammar [2, 3, 6, 7, 9, 10, 17–20, 22–26, 28]. Often the input is assumed to be from a single idealized grammar, with no noisy or incorrect utterances.

From the perspective of idealized grammar, language change is paradoxical: Children acquire their native language accurately, yet over time, the language can change. Previously unacceptable sentences become grammatical, and other constructions fall out of use and become ungrammatical. English word order has changed drastically over the centuries, for example. Some changes may be attributed to an external event, such as political upheaval, but not every instance of language change seems to have an external cause.

An alternative perspective that can resolve the language change paradox is to approach grammaticality continuously: Utterances can fall along a range from clearly correct to clearly incorrect, and may be classified as more or less archaic or innovative. Such an approach can consider the variation present in natural speech [12, 13]. For example, in a corpus of late Middle and early Modern English, each manuscript uses a combination of verb-raising syntax and *do*-support syntax in forming questions and negative statements[11, 29]:

(1) Know you what time it is? (verb raising)

(2) Do you know what time it is? (*do*-support)

(3) I know not what time it is. (verb raising)

(4) I don't know what time it is. (*do*-support)

The speech pattern of such an individual can be described by a *stochastic grammar,* that is, a collection of similar idealized grammars, each of which is used randomly at a particular rate. Over time, the usage rate of verb-raising syntax gradually dropped to zero among all speakers, and *do*-support usage rate increased to 100%.

From this continuous perspective, language change is no longer a paradox, but acquisition requires more than selecting a single idealized grammar compatible with the community's speech. Instead, children must learn multiple idealized grammars, plus the usage rates and whatever conditions affect them.

Despite their variability, languages maintain considerable short-term stability and consistently accept and reject large classes of sentences for centuries. A challenge for modeling language learning and use is to capture this meta-stability.

As we will see in Section 2, a mean-field model in which children learn from the entire population equally does not lead to spontaneous change, even in the presence of random variation. However, Section 3 describes an improved model, in which children can detect age-correlated patterns in variation. When subject to random fluctuations, this model does exhibit meta-stability. The population tends to hover near a state where one idealized grammar is highly preferred. However, children occasionally detect accidental correlations between age and speech, predict that the population is about to undergo a language change, and accelerate the change. This feature is called *prediction-driven instability.*

Once the age-structure model is formulated with a finite population as a discrete Markov chain, we will consider the limit of an infinitely large population and reformulate it as a martingale problem. Focusing on a low dimensional case, we will rewrite it as a system of stochastic differential equations (SDEs), show that it has a unique solution for all initial values, and show that paths of the Markov chain converge weakly to solutions of the SDEs.

2. An unstructured mean-field model

2.1. **An ordinary differential equation.** Let us suppose, for the sake of simplicity, that individuals have the choice between two similar idealized grammars G_1 and G_2. Each simulated agent uses G_2 in forming an individual-specific fraction of spoken sentences, and G_1 in forming the rest. Assume that children are always able to acquire both idealized grammars and the only challenge is learning the usage rates.

Consider a continuous mean-field model, that is, an infinitely large unstructured population, in which children learn from all individuals equally and therefore hear essentially the mean usage rate of G_2. Using a time scale under which the birth rate is 1, the simplest population-learning model for $m(t) =$ the time-dependent mean usage rate of G_2 in the population is

$$\dot{m} = q(m) - m \tag{5}$$

where the *learning function* $q(m)$ is the mean usage rate of children learning from a population with a mean rate m. The $q(m)$ term represents birth and learning, and the $-m$ term represents death. If learning were perfect, that is, children exactly reproduced the population's average usage rates, then q would be the identity function. Instead, the learning function must be S-shaped to ensure that there are two stable equilibrium states, representing populations dominated by one grammar or the other. The phase portrait of (5) will then consist of two stable fixed points separated by an unstable fixed point, as in Figure 1. In general, q is assumed to be strictly increasing, twice continuously differentiable, with one inflection point, and $0 < q(0) < 1/4$ and $3/4 < q(1) < 1$.

For the rest of the paper, all pictures and calculations will be based on the specific learning function

$$q(m) = \frac{1}{32} + \frac{3600}{751}\left(\frac{33m}{1280} + \frac{161m^2}{320} - \frac{m^3}{3}\right). \tag{6}$$

Its graph is shown in Figure 1. This particular function is slightly asymmetric so that pictures in this paper will not exhibit atypical symmetry.

This curved learning function means that the more commonly used idealized grammar becomes even more commonly used, until the other grammar all but disappears. This tendency is in agreement with the observation that children regularize a language: A growing body of evidence [8] indicates that for the task of learning a language with multiple ways to say something, adults tend to use all the options and

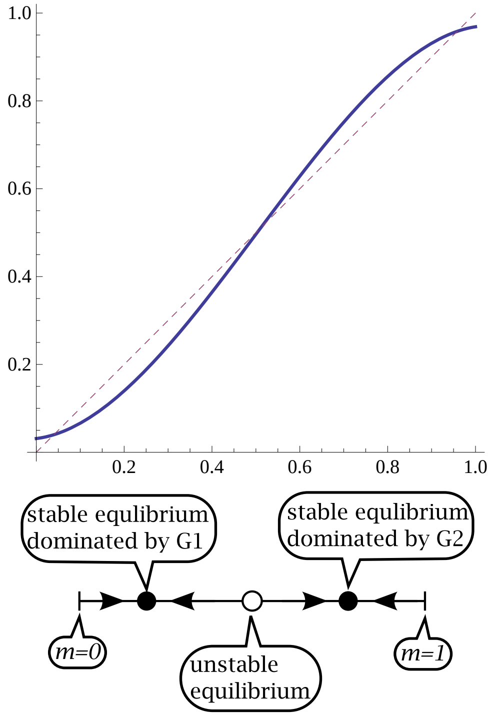

Figure 1. A plot of $q(m)$ from (6) as a function of m, and the phase portrait of m from (5).

match the usage rates in the given data, but children prefer to pick one option and stick with it.

The dynamical system (5) is too simple to model ongoing language change. It is deterministic and one-dimensional, so there is no way for the population to spontaneously switch grammars.

2.2. **A Markov chain.** To add the possibility of a language change, we reformulate the model as a Markov chain, thereby adding random fluctuations. We assume that the population consists of N adults, each of which is one of $K+1$ types, numbered 0 to K, where type k means that the individual uses G_2 at a rate k/K. The state of the chain at step j is a vector A where $A_n(j)$ is the state of the n-th agent. We also define a count vector C where $C_k(t)$ is the number of agents of type k,

$$C_k(j) = \sum_n \mathbf{1}\left(A_n(j) = k\right).$$

Dividing the count vector by the population size yields the distribution vector $Z = C/N$ such that an agent selected at random from the population uniformly at time j is of type k with probability $Z_k(j)$. The mean usage rate of G_2 at step j is therefore

$$M(j) = \sum_{k=0}^{K} \left(\frac{k}{K}\right) Z_k(j) \tag{7}$$

The transition process from step j to $j+1$ is as follows. Two parameters are required, a birth-and-death rate r_D and a resampling rate r_R. Either the count vector $C(j)$ or the distribution vector $Z(j)$ suffices to represent the state of the chain, but it is simpler to state the transition function in terms of what happens to each individual agent:

- With probability $p_D = r_D/N$ it is removed to simulate death.
- With probability r_R it is resampled.
- With probability $1 - p_D - r_R$ it is unchanged.

Each time step is interpreted as $1/N$ years. The lifespan of an individual in time steps has a geometric distribution with parameter p_D. The average life span is therefore $1/p_D$ time steps or $1/r_D$ years.

When an agent dies, a replacement agent is created and its type is selected at random based on a discrete distribution vector $Q(M(j))$. That is, $Q_k(m)$ is the probability that a child learning from a population with mean usage rate m is of type k, and therefore uses G_2 at rate k/K. For the purposes of this article, $Q(m)$ will be the mass function for a binomial distribution with parameters $q(m)$ and K,

$$Q_k(m) = \binom{K}{k} q(m)^k (1 - q(m))^{K-k}.$$

This Markov chain model maintains the mean-field assumption because the population influences language acquisition only through the mean usage rate of G_2.

When an agent is resampled, its new state is copied from another agent picked uniformly at random. The average time an agent spends between resamplings is $1/r_R$ time steps. This feature of the transition process incorporates the fact that as an adult, an individual's language is not entirely fixed. Furthermore, as will be seen in Section 3.2, without this resampling feature, the random fluctuations of the Markov chain diminish to 0 in the limit as $N \to \infty$, which would defeat the purpose of developing a stochastic model.

It is possible, though unlikely, for all agents to die off and return to state 0 in a single step. Therefore, this Markov chain is ergodic, meaning that it must visit every possible state eventually. It spends most of its time hovering near an equilibrium dominated by one grammar or the other, but it must eventually exhibit spontaneous language change by switching to the other equilibrium.

However, computer experiments show that under this model, a population takes an enormous amount of time to switch dominant grammars. See Figure 2 for a graph of the mean usage rate of G_2 as a function of time for a typical run of this Markov chain, which shows no sign of any

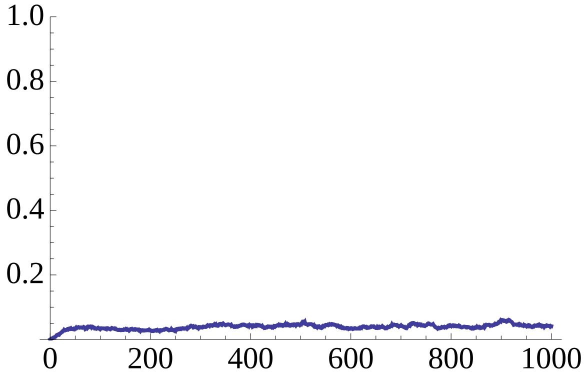

FIGURE 2. A plot of the mean usage rate $M(t)$ of G_2 from a sample path of the basic Markov chain over 1000 years.

transition despite running for 1000 years. For this example, the population consists of 1000 agents, and the replacement rate is $r_D = 1/40$. Therefore, the mean lifespan of an agent is 40 years. The resampling rate is $r_R = 0.0001$. There are 6 types of agents, representing speech patterns that use G_2 for a fraction $0, 1/5, \ldots, 1$ of spoken sentences.

This model is therefore unsuitable for simulating language change on historical time scales. A further undesirable property is that if a population does manage to shift to an intermediate state, it is just as likely to return to the original grammar as to complete the shift to the other grammar. Historical studies [11, 30] show that language changes typically run to completion monotonically and do not reverse themselves partway through (but see [29] for some evidence to the contrary), so again this model is unsuitable.

3. An age-structured model

One way to remedy the weaknesses of these mean-field models is to introduce social structure into the population. According to sociolinguistics, ongoing language change is reflected in variation, so there is reason to believe children are aware of socially correlated speech variation and use it during acquisition [12].

There are many ways to formulate a socially structured population, and not all formulations apply to all societies. For simplicity, we assume that there are two age groups, roughly representing youth and their parents, and that children can detect systematic differences in their speech. We also assume that there are social forces leading children to avoid sounding out-dated.

3.1. **Adapting the mean-field Markov chain.**

Let us adapt the Markov chain from Section 2 to include age structure. To represent the population at time j, we let $U_n(j)$ be the state of the n-th youth and $V_n(j)$ be the state of the n-th parent. Define $C_k(j)$ to be the number of youth of type k, and define $D_k(j)$ to be the number of parents of type k. The total number of youth and the total number of parents are fixed at N. We also assume that apart from age, children make no distinction among individuals. Thus, they learn essentially from the mean usage rates of the two generations,

$$\begin{aligned} M_C(j) &= \sum_{k=0}^{K} \left(\frac{k}{K}\right)\left(\frac{C_k(j)}{N}\right) \\ M_D(j) &= \sum_{k=0}^{K} \left(\frac{k}{K}\right)\left(\frac{D_k(j)}{N}\right) \end{aligned} \tag{8}$$

We have modified the mean-field assumption by expanding the influence of the population on a child to two aggregate quantities. The modified transition process from time j to $j+1$ is as follows. Each adult is examined:

- With probability $p_D = r_D/N$ it is removed to simulate death.
- With probability r_R it is resampled from the adult population.
- With probability $1 - p_D - r_R$ it is unchanged.

Each youth is examined:

- With probability $p_D = r_D/N$ it is removed to simulate aging.
- With probability $p_R = r_R/N$ it is resampled from the youth population.
- With probability $1 - p_D - p_R$ it is unchanged.

Each time step is interpreted as $1/N$ years. The number of time steps spent by an individual in each age group has a geometric distribution with parameter p_D. The average time spent as an adult and as a youth is therefore $1/p_D$ time steps or $1/r_D$ years, so the average life span is now $2/r_D$.

When an agent is resampled, its new state is copied from another agent from the same generation selected uniformly at random. As before, resampling leaves the mean behavior unchanged while introducing volatility.

When an adult is removed, a new adult is created by copying a youth at random. When a youth ages, a new youth is created based on the discrete probability vector $R(M_C(t), M_D(t))$. Here, $R(x, y)$ represents the acquisition process, together with prediction: Children hear that the younger generation uses G_2 at a rate x, and the older generation uses a rate y. Based on x and y and any trend those numbers indicate, they predict a rate that their generation should use, and learn based on that predicted target value. Let the prediction be given by a function $r(x, y)$ that is increasing with respect to x, decreasing with respect to y, and satisfies $x < y$ implies $r(x, y) < x$ and $x > y$ implies $r(x, y) > x$. Then, our assumptions on learning based on prediction can be incorporated into the mathematics by setting $R(x, y) = Q(r(x, y))$.

For a specific example, let us consider a population of 1000 agents, 500 in each age group, with a replacement rate of $r_D = 1/20$. Therefore, the mean lifespan of an agent is 40 years. The resampling rate is $r_R = 0.0001$. There are 6 types of agents, representing speech patterns that use G_2 for a fraction $0, 1/5, \dots, 1$ of spoken sentences. The learning distribution $Q(m)$ is a binomial distribution with parameters $q(m)$ and 5. The prediction function $r(x, y)$ is based on an exponential sigmoid, as in Figure 3. Given $\sigma(t) = 1/(1 + \exp(-t))$, define $t_1 = \sigma^{-1}(x)$ and $t_2 = \sigma^{-1}(y)$. Then $h = t_1 - t_2$ is a measure of the trend between the generations. A scale factor α is applied to h, and the scaled trend is added to t_1. After some simplification,

$$\begin{aligned} r(u, v) &= \sigma(t_1 + \alpha h) \\ &= \frac{1}{1 + \left(\frac{1-x}{x}\right)^{\alpha+1}\left(\frac{y}{1-y}\right)^{\alpha}} \end{aligned} \tag{9}$$

For the example calculations in this paper, $\alpha = 3$. The behavior of r at the boundaries of the phase space will be important later in the paper.

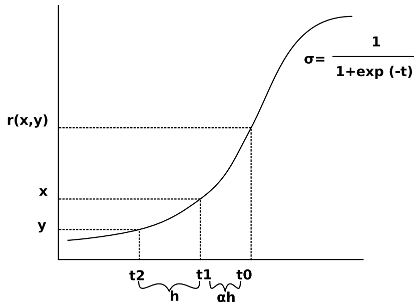

FIGURE 3. An illustration of the prediction function r.

Since r is a rational function, the following results follow immediately:

$$\begin{aligned} &\forall y \in (0,1) \qquad \lim_{x\to 0} r(x,y) = 0 \text{ and } \lim_{x\to 1} r(x,y) = 1 \\ &\forall x \in (0,1) \qquad \lim_{y\to 0} r(x,y) = 1 \text{ and } \lim_{y\to 1} r(x,y) = 0 \end{aligned} \tag{10}$$

This means that r is discontinuous only at the corners $x = y = 0$ and $x = y = 1$.

This model turns out to exhibit the desired properties. The population can spontaneously change from one language to the other and back within a reasonable amount of time, and once initiated the change runs to completion without turning back. See Figure 4 for a graph of the mean usage rate of G_2 among the younger age group as a function of time for a typical run of this Markov chain.

3.2. **Diffusion limit.** To better understand why spontaneous change happens in this model, we approximate the Markov chain by a continuous time stochastic process governing the speech distributions $X = C/N$ and $Y = D/N$ of the younger and older generations. In the limit as the population size N increases without bound, the Markov chain

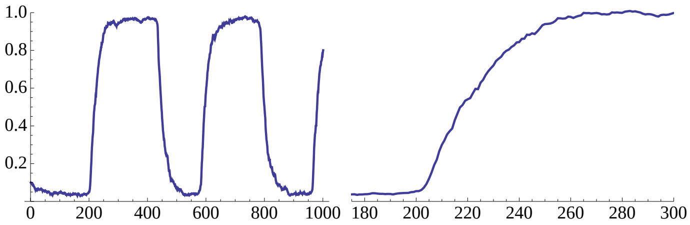

Figure 4. Trajectory of the mean usage rate $M_C(t)$ of G_2 in the young generation from a sample path of the age-structured Markov chain. Left: The path from time 0 to 1000 years, showing several changes between G_1 (low) and G_2 (high). Right: The path from time 175 to 300 years, showing a single grammar change.

$(X(j), Y(j))$ ought to converge to the solution $(\xi(t), \zeta(t))$ of a martingale problem. To formulate it, we must calculate the infinitesimal drift and covariance functions.

3.2.1. *Some helpful definitions.* To reduce notational clutter in this subsection, all time-dependent quantities at time j will be written without a time index, as in U_n, V_n, C_k, D_k, X_k, and Y_k. The learning distribution $Q(r(M_C, M_D))$ will be written as just Q with mean q. Time-dependent quantities at time $j+1$ will be written with a prime mark, as in U'_n, V'_n, C'_k, D'_k, X'_k, and Y'_k. Expectations and variances with a j subscript are conditioned on the information available at time step j.

3.2.2. *Infinitesimal mean and variance.* Conditioning on time step j, $\mathbf{1}\left(U'_n = k\right)$ is a Bernoulli random variable that takes on the value 1 with probability $g(n,k) = (1 - p_D - r_R)\mathbf{1}\left(U_n = k\right) + p_D Q_k + r_R X_k$. With this observation, the mean and variance of C'_k conditioned on information known at time step j can be calculated as follows.

$$\begin{aligned}\mathbb{E}_j\left(C'_k\right) &= \sum_n g(n,k) \\ &= (1-p_D)C_k + p_D N Q_k\end{aligned} \tag{11}$$

$$\begin{aligned}\mathrm{Var}_j\left(C'_k\right) &= \sum_n g(n,k) - g(n,k)^2 \\ &= (1-p_D-r_R)C_k + p_D N Q_k + r_R N X_k \\ &\quad - (1-p_D-r_R)^2 C_k \\ &\quad - 2(1-p_D-r_R)C_k(p_D Q_k + r_R X_k) \\ &\quad - N(p_D Q_k + r_R X_k)^2\end{aligned} \tag{12}$$

If $h \neq k$ and $m \neq n$ then for all j, $\mathbf{1}\left(U_n(j) = k\right)\mathbf{1}\left(U_n(j) = h\right) = 0$ and $U_m(j)$ and $U_n(j)$ are independent. That implies

$$\begin{aligned}\mathrm{Cov}_j\left(C'_k, C'_h\right) &= \sum_n \mathrm{Cov}_j\left(\mathbf{1}\left(U'_n = k\right), \mathbf{1}\left(U'_n = h\right)\right) \\ &= -\sum_n g(n,k)g(n,h) \\ &= -\big((1-p_D-r_R)C_k(p_D Q_h + r_R X_h) \\ &\qquad + (1-p_D-r_R)C_h(p_D Q_k + r_R X_k) \\ &\qquad + N(p_D Q_k + r_R X_k)(p_D Q_h + r_R X_h)\big)\end{aligned} \tag{13}$$

It follows that

$$\begin{aligned}\mathbb{E}_j\left(X'_k\right) &= \frac{1}{N}\,\mathbb{E}_j\left(C_k\right) \\ &= (1-p_D)X_k + p_D Q_k\end{aligned} \tag{14}$$

which can be rearranged into

$$\mathbb{E}_j\left(\frac{X'_k - X_k}{1/N}\right) = r_D(Q_k - X_k) \tag{15}$$

That gives the infinitesimal drift component for a martingale problem. To get the infinitesimal covariance, we need a $O\left(1/N\right)$ approximation of the covariance matrix for X.

$$\begin{aligned} \operatorname{Var}_j\left(X'_k\right) &= \frac{1}{N^2}\operatorname{Var}_j\left(C'_k\right) \\ &= \frac{1}{N}\left((2r_R - r_R^2)(X_k - X_k^2)\right) + O\left(\frac{1}{N^2}\right) \end{aligned} \tag{16}$$

For the last equation, we need the fact that $p_D = r_D/N = O\left(1/N\right)$.

$$\begin{aligned} \operatorname{Cov}_j\left(X'_k, X'_h\right) &= \frac{1}{N^2}\operatorname{Cov}_j\left(C'_k, C'_h\right) \\ &= -\frac{1}{N}\left((2r_R - r_R^2)X_k X_h\right) + O\left(\frac{1}{N^2}\right) \end{aligned} \tag{17}$$

Similar drift and covariance formulas can be derived for Y in the same way.

The infinitesimal drift and covariance functions are the leading terms from the preceding approximations. As a further simplification, we can rescale time by a factor of r_D. This finally yields the infinitesimal drift function

$$\mathbf{b}\begin{pmatrix}\xi \\ \zeta\end{pmatrix} = \begin{pmatrix}Q - \xi \\ \xi - \zeta\end{pmatrix} \tag{18}$$

and the infinitesimal covariance function

$$\mathbf{a}\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \varepsilon^2 \begin{pmatrix} \xi_1-\xi_1^2 & -\xi_1\xi_2 & \dots & & & \\ -\xi_1\xi_2 & \ddots & & & 0 & \\ \vdots & & & & & \\ & & & \zeta_1-\zeta_1^2 & -\zeta_1\zeta_2 & \dots \\ & 0 & & -\zeta_1\zeta_2 & \ddots & \\ & & & \vdots & & \end{pmatrix} \tag{19}$$

where

$$\varepsilon = \sqrt{\frac{2r_R - r_R^2}{r_D}} = \sqrt{\frac{1-(1-r_R)^2}{r_D}} \tag{20}$$

Note that ξ_0 and ζ_0 are omitted from the dynamics because of the population size constraint, that is,

$$\xi_0 = 1 - \xi_1 - \dots - \xi_K$$

$$\zeta_0 = 1 - \zeta_1 - \dots - \zeta_K$$

Furthermore, if the resampling feature is removed by setting $r_D = 0$, then $\varepsilon = 0$ and the dynamics become deterministic. The resampling feature can also be removed from just the older generation by zeroing out the lower right quadrant of $\mathbf{a}$, or from just the younger generation by zeroing out the upper left quadrant.

Using the matrix square-root for example, is possible to find infinitesimal standard deviation functions $\boldsymbol{\sigma}$ such that $\boldsymbol{\sigma}\boldsymbol{\sigma}^T = \sqrt{\mathbf{a}}$ and use $\boldsymbol{\sigma}$ to write a system of stochastic differential equations for ξ and ζ. However, there does not seem to be any suitable $\boldsymbol{\sigma}$ with a simple closed form.

4. Theory for a 2-dimensional case

Rather than treat the full $2(K+1)$ variable system, we will continue by restricting our attention to the case of $K = 1$. That is, simulated individuals use G_2 exclusively or not at all, and X_0 is the fraction of the young generation that never uses G_2 and X_1 is the fraction that always uses G_2. Since $X_0 + X_1 = 1$, it is only necessary to deal with X_1. As a further simplification of the notation, an X with no subscript will refer to X_1. Likewise, a Y with no subscript will refer to Y_1. The learning process Q simplifies as well. The mean usage rates of G_2 among the younger and older generations are X and Y respectively, so $Q_1 = q(r(X, Y))$ and $Q_0 = 1 - Q_1$. The mean of a random number distributed according to Q is Q_1.

The time associated with step j of the Markov chain is $t = j/N$, so to properly express the convergence of the Markov chain to a process in continuous time and space, we need the auxiliary processes $\bar{X}$ and $\bar{Y}$ that map continuous time to discrete steps,

$$\begin{aligned} \bar{X}(t) &= X(\lfloor Nt \rfloor) \\ \bar{Y}(t) &= Y(\lfloor Nt \rfloor) \end{aligned} \tag{21}$$

The covariance function (19) reduces to a 2-by-2 diagonal matrix so it has a very simple square-root. With these definitions, we claim

Proposition 1. *Suppose* $(\bar{X}(0), \bar{Y}(0))$ *converges to* (ξ_0, ζ_0) *as* $N \to \infty$. *For sufficiently small* $\varepsilon > 0$, *the process* $(\bar{X}(t), \bar{Y}(t))$ *converges weakly* *as $N \to \infty$ to the solution $(\xi(t), \zeta(t))$ of*

$$
\begin{aligned}
d\xi &= (q(r(\xi,\zeta)) - \xi)dt + \varepsilon\sqrt{\xi(1-\xi)}dB^{\xi} \\
d\zeta &= (\xi - \zeta)dt + \varepsilon\sqrt{\zeta(1-\zeta)}dB^{\zeta} \\
\xi(0) &= \xi_0 \\
\zeta(0) &= \zeta_0
\end{aligned}
\tag{22}
$$

where B^{ξ} and B^{ζ} are independent one-dimensional Brownian motions.

Proof. We will show that the results of Chapter 8 of Durrett [4] apply, specifically theorem 7.1, which implies this result.

The first step is to verify that the step-to-step drift, variances, and covariances of the Markov chain converge to the functions in the SDE as the time step size $1/N$ goes to zero. The calculations (15), (16), and (17) in Section 3.2 verify that these conditions (listed as (a), (b), and (c) in [4] lemma 8.2) hold.

The remaining condition to check is Durrett's hypothesis (A) for theorem 7.1, which is that the martingale problem associated to (22) is well posed. The proof, which forms the rest of this section, begins with some observations about the deterministic dynamical system derived by setting $\varepsilon = 0$. After a change of variables to move the system from the square phase space $(0, 1) \times (0, 1)$ to $\mathbb{R}^2$, standard results imply well posedness. □

4.1. **Deterministic limit.** In the deterministic limit $\varepsilon = 0$, the SDE (22) becomes

$$\begin{aligned} \dot{\xi} &= q(r(\xi,\zeta)) - \xi \\ \dot{\zeta} &= \xi - \zeta \end{aligned} \tag{23}$$

Intuitively, the phase space of this dynamical system is a square, and it happens to have two stable equilibria representing populations where both generations are dominated by one grammar or the other. Each such equilibrium has a basin of attraction. The separatrix forming the boundary between the two basins passes very close to the stable equilibria. See Figure 5. Under the stochastic dynamics, the population will hover near one equilibrium or the other, but a random fluctuation may cause it to stray across the separatrix, where it will be blown toward the other equilibrium.

Under these deterministic dynamics, the square phase space is a trapping region. The vector field points inward all the way along the boundary. With random fluctuations, it is more difficult to guarantee that solutions stay within the phase space. In preparation for that proof, we will sketch the null-clines and map out the edges of the phase space. The $\dot{\zeta}$ null-cline is simply the diagonal line $\xi = \zeta$. The $\dot{\xi}$ null-cline is more complicated. The graph of the prediction function $r(\xi, \zeta)$ is primarily a cliff along the line $\xi = \zeta$. Composing with the learning function q sharpens the cliff and moves the plateau down from 1 and the floor up from 0. Thus, the plane given by the graph of $(\xi, \zeta) \mapsto \xi$ is slightly below the graph of $q(r(\xi, \zeta))$ on the left edge and slightly below

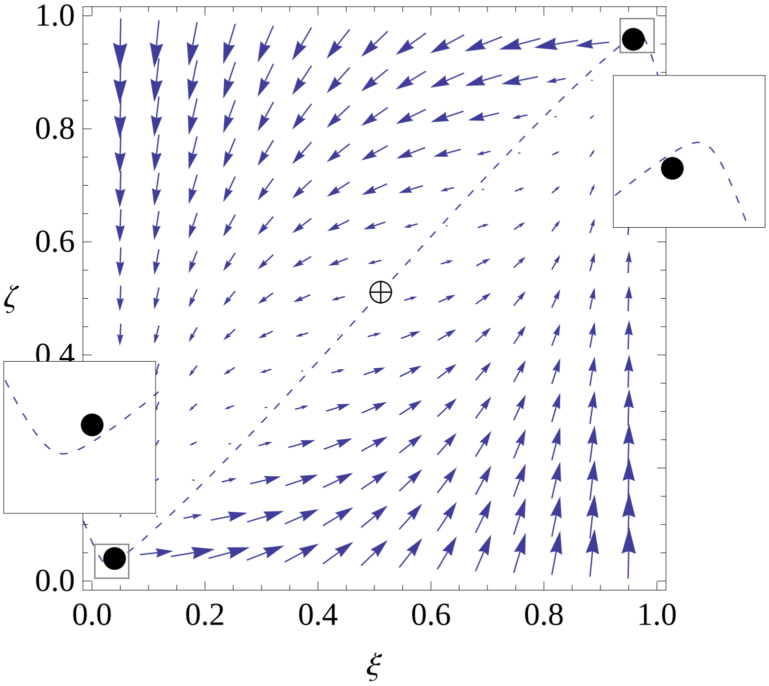

Figure 5. Phase portrait for (23). The crossed dot ⊕ is a saddle point, the two dots • are sinks, and the dashed curve is the separatrix between their basins of attraction. The arrows indicate the direction of the vector field as given by (23). The inset boxes show magnified pictures of the areas around the sinks.

on the right, and it intersects with the cliff in a diagonal line, as shown in Figure 6. This means that the $\dot{\xi}$ null-cline, which lies under that intersection, is a sharp zig-zag like a reversed N. A schematic diagram of the two null-clines is shown in Figure 7 which makes the structure of the vector field easier to see near the edges of the phase space.

4.2. **Well-posedness of the SDE.** The SDE (22) has pathwise-unique strong solutions, as we will prove. This implies uniqueness in distribution [4, theorem 4.1 §5.4] which implies that the martingale problem is well posed [4, theorem 4.5 §5.4].

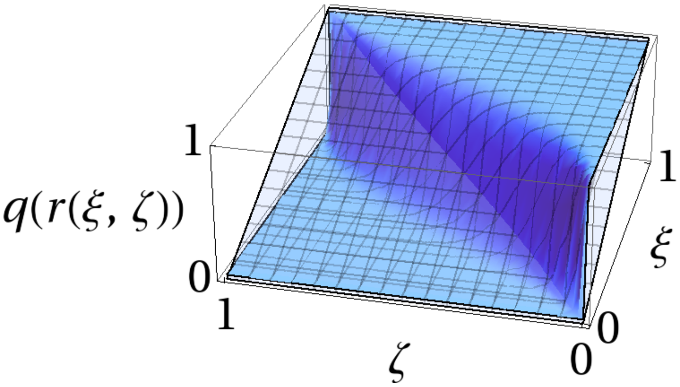

FIGURE 6. The learning function $q(r(\xi, \zeta))$ and the plane given by the graph of $(\xi, \zeta) \mapsto \xi$.

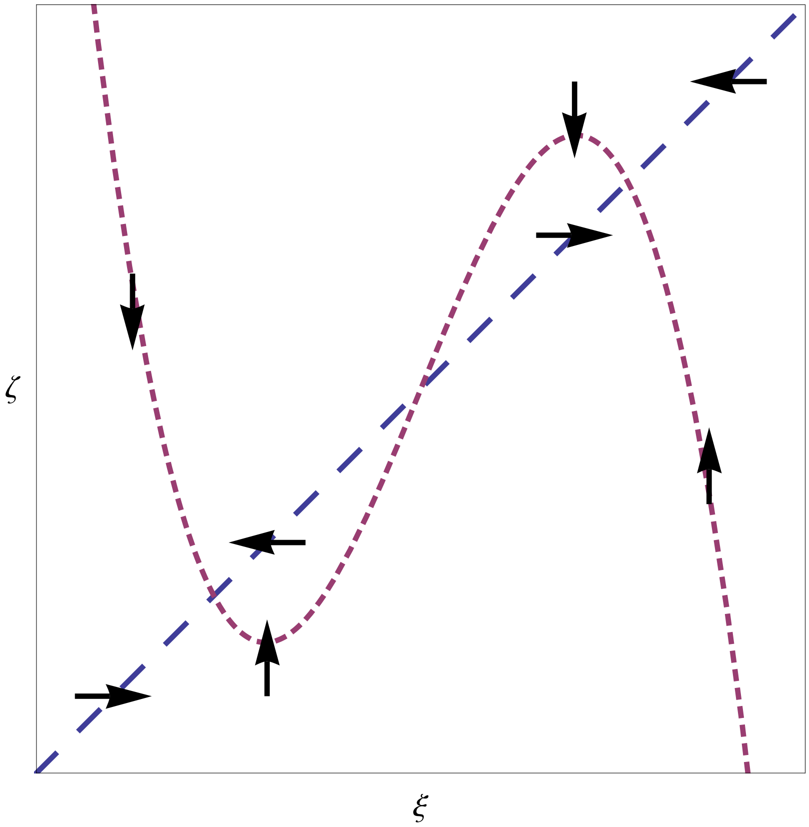

FIGURE 7. Null-clines drawn schematically. Arrows indicate the vector field along the null-clines.

Lemma 2. *There is a number* $\varepsilon_0 > 0$ *such that if* $\varepsilon < \varepsilon_0$ *then the following holds: Given an initial condition* $(\xi, \zeta) \in (0,1) \times (0,1)$*, the system* (22) *has a pathwise-unique strong solution taking values in* $(0,1)\times(0,1)$ *almost surely for all* $t \geq 0$*.*

Proof. We must deal with some details concerning the boundary of the phase space. The semi-circle function $\sqrt{x(1-x)}$ in the diffusion terms is not Lipschitz continuous: At 0 and 1 the derivative is unbounded. Furthermore, the prediction function r is discontinuous at two of the corners. We therefore change variables so as to push the boundary of the phase space off to infinity.

The new variables and their relationships to ξ and ζ are

$$\begin{aligned} & u = 2\xi - 1, u \in (-1,1) \\ & v = 2\zeta - 1, v \in (-1,1) \\ & \kappa = \frac{u}{\sqrt{1-u^2}}, \kappa \in (-\infty, \infty) \\ & \lambda = \frac{v}{\sqrt{1-v^2}}, \lambda \in (-\infty, \infty) \end{aligned} \tag{24}$$

This particular change of variables recenters the square phase space on the origin and blows it up to occupy the whole plane. Applying Itō's formula,

$$\begin{aligned} d\kappa &= \left(\overbrace{2(1+\kappa^2)^{3/2}(q(r(\xi,\zeta))-\xi)+\frac{3}{2}\varepsilon^2\kappa(1+\kappa^2)}^{b_1}\right)dt \\ &\quad+\overbrace{\varepsilon(1+\kappa^2)^{1/2}}^{\sigma_1}dB^{\xi} \\ d\lambda &= \left(\overbrace{2(1+\lambda^2)^{3/2}(\xi-\zeta)+\frac{3}{2}\varepsilon^2\lambda(1+\lambda^2)}^{b_2}\right)dt \\ &\quad+\overbrace{\varepsilon(1+\lambda^2)^{1/2}}^{\sigma_2}dB^{\zeta} \end{aligned} \tag{25}$$

The change of variables from (ξ, ζ) to (κ, λ) is order-preserving in both directions and maintains the general shape of the null-clines.

The standard theorem concerning the existence and uniqueness of solutions [4, §5.3] requires the drift and standard deviation terms to be locally Lipschitz. In this case, both are continuously differentiable as functions of κ and λ, so they automatically satisfy a local Lipschitz inequality.

The standard theorem also requires a growth constraint formulated as follows. Define

$$\beta = 2\kappa b_1 + 2\lambda b_2 + \sigma_1^2 + \sigma_2^2. \tag{26}$$

Then there must be a positive constant A such that for all κ and λ,

$$\beta < A(1 + \kappa^2 + \lambda^2). \tag{27}$$

The difficulty is that β contains terms that grow like κ^4 and λ^4. Luckily, they turn out to be negative for sufficiently large κ and λ because the vector field as in (23) points toward zero out on the edges of the (ξ, ζ) phase space. To prove (27), we begin with the series

$$(1+x^2)^{3/2} = |x|^3 \left(1 + \frac{3}{2x^2} + \frac{3}{8x^4} + O\left(x^{-6}\right)\right)$$

and expand β into powers of κ and λ.

$$\begin{aligned}
\beta = {} & \overbrace{\left(4(q(r(\xi,\zeta)) - \xi)(\operatorname{sgn}\kappa) + 3\varepsilon^2\right)\kappa^4 + \left(4(\xi - \zeta)(\operatorname{sgn}\lambda) + 3\varepsilon^2\right)\lambda^4}^{\beta_4} \\
& + \overbrace{\left(6(q(r(\xi,\zeta)) - \xi)(\operatorname{sgn}\kappa) + 4\varepsilon^2\right)\kappa^2 + \left(6(\xi - \zeta)(\operatorname{sgn}\lambda) + 4\varepsilon^2\right)\lambda^2}^{\beta_2} \\
& + \overbrace{\left(\frac{3}{2}(q(r(\xi,\zeta)) - \xi)(\operatorname{sgn}\kappa) + \varepsilon^2\right) + \left(\frac{3}{2}(\xi - \zeta)(\operatorname{sgn}\lambda) + \varepsilon^2\right)}^{\beta_0} \\
& + \overbrace{O\left(\kappa^{-2}\right) + O\left(\lambda^{-2}\right)}^{\beta_{-2}}.
\end{aligned}$$

We begin with β_4. For each sector of the plane, a slightly different argument guarantees that if ε is sufficiently small and κ and λ are appropriately constrained, then $\beta_4 < 0$. See Figure 8.

Northeast. In the northeast sector, marked NE, $\lambda > \kappa > 0$. Define λ_{NE} such that the line segment from $(0, \lambda_{NE})$ to $(\lambda_{NE}, \lambda_{NE})$ lies above the null-cline. Let $-\delta_{NE}$ be the maximum of $q(r(\xi,\zeta)) - \xi$ on that line segment. Increasing ζ decreases $r(\xi,\zeta)$ which decreases $q(r(\xi,\zeta))$, so for all points in this northeast region, $-\delta_{NE}$ is an upper bound on

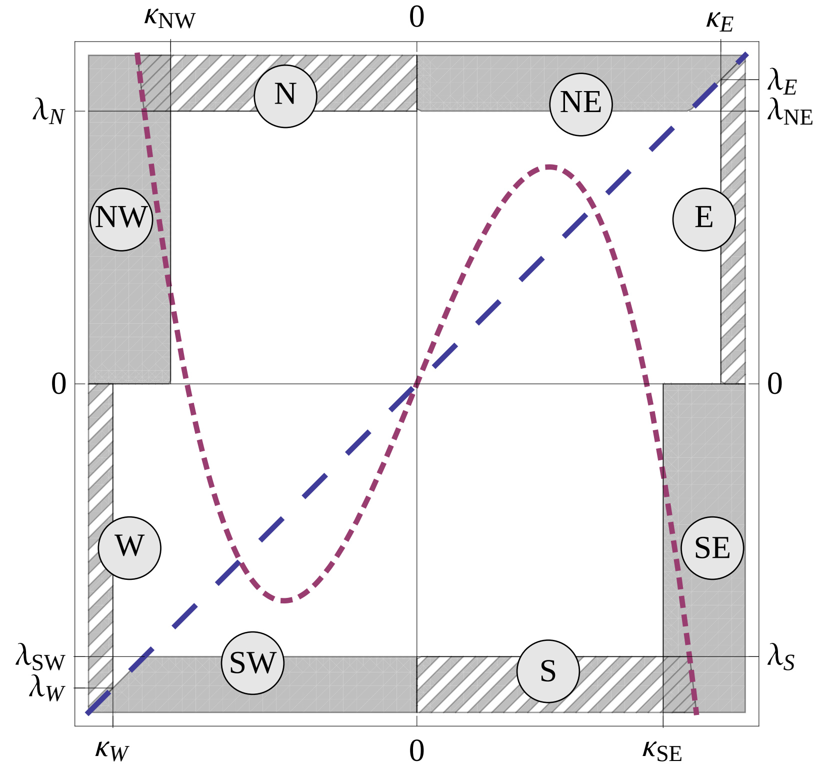

FIGURE 8. Regions of the plane for estimating β_4.

$q(r(\xi, \zeta)) - \xi$. Also, $\xi - \zeta \leq 0$. Thus,

$$\beta_4 < (-4\delta_{NE} + 6\varepsilon^2)\lambda^4.$$

We therefore require

$$\varepsilon^2 < \frac{2\delta_{NE}}{3}. \tag{28}$$

Southwest. The southwest sector, marked SW, is handled similarly, yielding the constraint

$$\varepsilon^2 < \frac{2\delta_{SW}}{3}. \tag{29}$$

where δ_{SW} is the maximum of $q(r(\xi,\zeta)) - \xi$ on a line segment from $(\lambda_{SW}, \lambda_{SW})$ to $(0, \lambda_{SW})$ that lies below the null-cline.

East. In the east sector, marked E, two sub-cases are required, one where λ is large and a second where it is small. Let $\zeta_E = (q(1)+1)/2$ and $\xi_E = (q(1)+3)/4$, and let λ_E and κ_E be the corresponding values of λ and κ. Note that $1 > \xi_E > \zeta_E > q(1)$.

For the sub-case of $\lambda \geq \lambda_E$, we may use the fact that $\kappa \geq \lambda \geq \lambda_E > 0$ to get

$$\begin{aligned}\beta_4 &\leq (4(q(1) - \zeta_E) + 6\varepsilon^2)\kappa^4 \\ &\leq (-4(1 - q(1)) + 6\varepsilon^2)\kappa^4.\end{aligned}$$

We therefore impose the constraint

$$\varepsilon^2 < \frac{2(1 - q(1))}{3} \tag{30}$$

so that $\beta_4 < 0$.

For the sub-case $1/2 \leq \lambda \leq \lambda_E$, the general fact that $q(r(\xi,\zeta)) \leq q(1)$ and the constraint $\kappa \geq \kappa_E > \lambda_E \geq \lambda$ guarantee that

$$\begin{aligned}\beta_4 &\leq (4(q(1) - \xi_E) + 3\varepsilon^2)\kappa^4 + (4(1 - \zeta_E) + 3\varepsilon^2)\lambda_E^4 \\ &\leq (-1 - 2q(1) + 6\varepsilon^2)\kappa^4.\end{aligned}$$

We therefore impose the constraint

$$\varepsilon^2 < \frac{1 + 2q(1)}{6} \tag{31}$$

so that $\beta_4 < 0$.

West. The west sector, marked W, works out similarly, requiring

$$\varepsilon^2 < \frac{2q(0)}{3} \tag{32}$$

and

$$\varepsilon^2 < \frac{1 + 2(1 - q(0))}{6} \tag{33}$$

Southeast. For the southeast sector, marked SE, let $\xi_{SE} = (q(1) + 3/4)/2 < q(1)$ and let κ_{SE} be the corresponding value of κ. The region is defined by $\lambda \leq 0$, and $\kappa \geq \kappa_{SE}$. It follows that

$$\begin{aligned}\beta_4 &\leq \left(4\left(q(1) + \frac{1}{2} - 2\xi_{SE}\right) + 6\varepsilon^2\right) \max\{\kappa^4, \lambda^4\} \\ &< \left(-1 + 6\varepsilon^2\right) \max\{\kappa^4, \lambda^4\}\end{aligned}$$

We therefore impose the constraint

$$\varepsilon^2 < \frac{1}{6} \tag{34}$$

to guarantee that $\beta_4 < 0$.

Northwest. A similar argument handles the northwest sector, marked NW, also requiring

$$\varepsilon^2 < \frac{1}{6}. \tag{35}$$

South. In the south sector, marked S, define $\zeta_S = (q(1) + 1)/2$ and let λ_S be the corresponding value of λ. Within this region, $\lambda \leq \lambda_S$ and

$0 \le \kappa \le -\lambda$.

$$\begin{aligned}\beta_4 &< (4(q(1) - 1 + \zeta_S) + 6\varepsilon^2)\lambda^4 \\ &< (2(1 - q(1)) + 6\varepsilon^2)\lambda^4\end{aligned}$$

We therefore impose the constraint

$$\varepsilon^2 < \frac{1 - q(1)}{2} \tag{36}$$

to guarantee that $\beta_4 < 0$.

North. The north region, marked N, works out similarly, requiring

$$\varepsilon^2 < \frac{q(0)}{2} \tag{37}$$

To summarize, in all eight sectors of the (κ, λ) plane, there is a constraint on ε which, if satisfied, guarantees that outside of a bounded set, $\beta_4 < 0$. Specifically, let

$$\varepsilon_0 = \sqrt{\min\left\{\frac{2\delta_{NE}}{3}, \frac{2\delta_{SW}}{3}, \frac{1}{6}, \frac{1 - q(1)}{2}, \frac{q(0)}{2}\right\}} \tag{38}$$

For each $0 < \varepsilon < \varepsilon_0$, all the constraints on ε in the preceding sector-by-sector analysis hold. This means there is a compact region in the (κ, λ) plane (the white area in Figure 8) outside of which β_4 is negative. Therefore, there exists a number R_4 such that if $\kappa^2 + \lambda^2 > R_4$ then $\beta_4 < 0$. Returning to the other terms, $\beta_{-2} + \beta_0$ is clearly less than, say, 5 when $\kappa^2 + \lambda^2$ exceeds some R_0. Furthermore,

$$|\beta_2| \le 10(\kappa^2 + \lambda^2).$$

Let $R = \max\{R_0, R_4\}$. Let A_0 be the maximum of β for $\kappa^2+\lambda^2 \leq R$. Let $A = A_0 + 15$. Putting all the pieces together, if $\kappa^2 + \lambda^2 > R$ then $\beta < A(\kappa^2+\lambda^2)$ and if $\kappa^2+\lambda^2 \leq R$ then $\beta < A$. Thus, $\beta < A(1+\kappa^2+\lambda^2)$, and with this inequality, standard results [4, theorems 3.1 and 3.2 in §5.3] imply the existence and uniqueness of solutions to (22). □

5. Discussion and conclusion

The first goal of this project is to build a mathematical model that can represent spontaneous language change in a population between two meta-stable states, representing populations dominated by one idealized grammar or another. Language is represented as a mixture of the idealized grammars to reflect the variability of speech seen in manuscripts and social data. A Markov chain that includes age structure has all the desired properties. The population can switch spontaneously from one language to the other and the transition is monotonic. Intuitively, the mechanism of these spontaneous changes is that every so often, children pick up on an accidental correlation between age and speech, creating the beginning of a trend. The prediction step in the learning process amplifies the trend, and moves the population away from equilibrium, which suggests the term *prediction-driven instability* for this effect.

Another way to understand this form of instability is to use an integrating factor to rewrite the $d\zeta$ equation from (22). By Itō's formula,

$$d(e^t\zeta) = e^t\xi\, dt + \varepsilon e^t\sqrt{\zeta(1-\zeta)}dB^\zeta$$

or in integral form

$$\begin{aligned}(39)\quad \zeta(t) = e^{-t}\zeta_0 + \int_0^t e^{-(t-s)}\xi(s)dt \\ + \varepsilon \int_0^t e^{-(t-s)}\sqrt{\zeta(s)(1-\zeta(s))}dB^{\zeta}(s)\end{aligned}$$

This can be interpreted as saying that ζ is an average of ξ over its past, with an exponential kernel giving greater weight to the recent past, plus some random noise. A further simplification is to apply the resampling step from the Markov chain (§3.1) only to the younger generation, which removes the random term from $d\zeta$ in (22) but not from $d\xi$. This yields a stochastic delay-functional equation for ξ alone

$$(40)\qquad d\xi(t) = \big(q(r(\xi(t), K_t\xi)) - \xi(t)\big)dt + \varepsilon\sqrt{\xi(t)(1-\xi(t))}dB$$

where the delay appears through convolution with a memory kernel

$$K_t f = e^{-t}f(0) + \int_0^t e^{-(t-s)}f(s)ds.$$

The age structure serves to give the population a memory, so that the speech pattern ξ of the young generation changes depending on how the current young generation deviates from its recent past average.

Since this is a new model, some fundamental results were proved. Specifically, in the limit as the number of agents goes to infinity, sample paths of the Markov chain converges weakly to solutions to a system of well-posed SDEs. These have the form of drift terms plus a small stochastic perturbation. Looking at the limit of zero perturbation, the prediction-driven instability comes from the proximity of stable sinks to the separatrix of their basins of attraction. Concrete formulas were given for q, r, and Q, but the interesting behavior is limited to these examples. The instability comes from the general geometry of the phase space as in Figure 5, and the proof that the system of SDEs is well-posed relies only on general properties of q, r, and Q.

A related problem is the FitzHugh-Nagumo model for a spiking neuron [14, 21], which is a general family of two-variable dynamical systems. The geometric structure is similar to Figure 7 except that the zigzag null-cline is shifted downward so there is only one fixed point which represents a resting neuron. A disturbance causes the neuron's state to stray away from that rest state and go on a long excursion known as an action potential or spike. The language change model examined here differs from the stochastic FitzHugh-Nagumo model in several ways. It is derived as a continuous limit of a Markov chain rather than from adding noise to an existing dynamical system. It has two stable equilibria rather than one (although it is conceivable that some linguistic phenomenon might exhibit the single stable equilibrium). It is naturally confined to a square, where FitzHugh-Nagumo models occupy an entire plane. The random term added to a FitzHugh-Nagumo model is normally Brownian motion multiplied by a small constant. The change of variables $\theta = \arcsin(2\xi - 1)$, $\phi = \arcsin(2\zeta - 1)$ transforms (22) to that form but the system remains confined to a square, and the change of variables to (25) on the whole plane has a non-constant coefficient on the Brownian motion.

Future studies of this model will include adapting and applying techniques for studying noise-activated transitions among meta-stable states, including exit problems [5, 15, 16]. For example, it is possible to numerically estimate the time between transitions using a partial differential equation or a variational technique.

This research was funded in part by a grant from the National Science Foundation (0734783).

References

[1] David Adger. *Core Syntax: A minimalist approach.* Oxford University Press, Oxford, 2003.

[2] E. J. Briscoe. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296, 2000.

[3] E. J. Briscoe. Grammatical acquisition and linguistic selection. In E. J. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models.* Cambridge University Press, 2002. URL `http://www.cl.cam.ac.uk/users/ejb/creo-evol.ps.gz`.

[4] Richard Durrett. *Stochastic Calculus: A Practical Introduction.* CRC Press, New York, 1996.

[5] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren der mathematischen Wissenschaften.* Springer Verlag, New York, 1984.

[6] E. Gibson and Kenneth Wexler. Triggers. *Linguistic Inquiry*, 25: 407–454, 1994.

[7] E. Mark Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[8] Carla L. Hudson Kam and Elissa L. Newport. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2): 151–195, 2005.

[9] Simon Kirby. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2): 102–110, 2001.

[10] Natalia L. Komarova, Partha Niyogi, and Martin A. Nowak. The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–59, 2001.

[11] Anthony Kroch. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244, 1989.

[12] William Labov. *Principles of Linguistic Change: Internal Factors*, volume 1. Blackwell, Cambridge, MA, 1994.

[13] William Labov. *Principles of Linguistic Change: Social Factors*, volume 2. Blackwell, Cambridge, MA, 2001.

[14] B. Lindner and L. Schimansky-Geier. Analytical approach to the stochastic FitzHugh-Nagumo system and coherence resonance. *Physical Review E*, 60(6):7270–7276, 1999.

[15] Robert S. Maier and Daniel L. Stein. A scaling theory of bifurcations in the symmetric weak-noise escape problem. *Journal of Statistical Physics*, 83(3):291–357, 1996. doi: 10.1007/BF02183736.

[16] Robert S. Maier and Daniel L. Stein. Limiting exit location distributions in the stochastic exit problem. *SIAM Journal on Applied Mathematics*, 57(3):752–790, 1997. doi: 10.1137/S0036139994271753.

[17] W. Garrett Mitchener. Bifurcation analysis of the fully symmetric language dynamical equation. *Journal of Mathematical Biology*, 46:265–285, March 2003.

[18] W. Garrett Mitchener. Game dynamics with learning and evolution of universal grammar. *Bulletin of Mathematical Biology*, 69 (3):1093–1118, April 2007. doi: 10.1007/s11538-006-9165-x.

[19] W. Garrett Mitchener and Martin A. Nowak. Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93, January 2003.

[20] W. Garrett Mitchener and Martin A. Nowak. Chaos and language. *Proceedings of the Royal Society of London, Biological Sciences*, 271(1540):701–704, April 2004. doi: 10.1098/rspb.2003.2643.

[21] James D. Murray. *Mathematical Biology*, volume I. Springer-Verlag, New York, 2002.

[22] Partha Niyogi. *The Informational Complexity of Learning*. Kluwer Academic Publishers, Boston, 1998.

[23] Partha Niyogi and Robert C. Berwick. A language learning model for finite parameter spaces. *Cognition*, 61:161–193, 1996.

[24] Partha Niyogi and Robert C. Berwick. A dynamical systems model for language change. *Complex Systems*, 11:161–204, 1997. URL `ftp://publications.ai.mit.edu/ai-publications/1500-1999/AIM-1515.ps.Z`.

[25] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291(5501):114–118, 2001.

[26] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. *Nature*, 417 (6889):611–617, June 2002.

[27] Andrew Radford. *Minimalist Syntax: Exploring the structure of English*. Cambridge University Press, Cambridge, 2004.

[28] Bruce Tesar and Paul Smolensky. *Learnability in Optimality Theory*. MIT Press, 2000.

[29] Anthony Warner. Why DO dove: Evidence for register variation in Early Modern English negatives. *Language Variation and Change*, 17:257–280, 2005. doi: 10.1017/S0954394505050106.

[30] Charles D. Yang. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, 2002.