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

One of the primary tools in the field of linguistics is the *idealized grammar*, that is, a formalism that identifies correct utterances (1, 2). 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 the surrounding society (3–5). From the perspective of idealized grammar, language change is paradoxical: Children acquire their native language accurately and communicate with adults from preceding generations, yet over time, the language can change drastically. 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. Despite their variability, languages do maintain considerable short-term stability, consistently accepting and rejecting large classes of sentences

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

for centuries. The primary challenge addressed by the new model discussed in this article is to capture this meta-stability.

A number of models focus on characterizing stable properties of languages. For example, naming games and other lexical models are focused on the ability of a population to form a consensus on a vocabulary, and how effective that vocabulary is at representing meanings (6–12). Related models focus on the structure of stable lexicons or phoneme inventories (13).

Several researchers have proposed and analyzed algorithms for learning idealized grammars, some designed to reproduce specific features of human language and associated changes (5, 14–19). These focus on details of the acquisition process and follow the probably-almost-correct (PAC) learning framework (20), in which the learner's input is a list of grammatically correct utterances (perhaps generated by agents using different idealized grammars) and the learner is required to choose a single idealized grammar from a limited set that is most consistent with that input. The input may be from a single individual (18, 19) or an unstructured population (14). Pearl and Weinberg (21) and Lightfoot (22, 23) analyze learning models that use fragments of the input and can exclude rarely occurring constructions as noise. These models assume the input contains no data on social structure, and they typically have sink states from which the learner or population cannot escape. In addition, there is evidence that the PAC framework and the so-called subset principle do not accurately reproduce certain features of child language acquisition (24, 25).

Many language models are adapted from deterministic, continuous, biological population models and represent language by communication games. These focus on stable behavior in an infinite homogeneous population, although some exhibit ongoing fluctuations (26–36). Some are designed to represent a single change (37). In these models, children learn from an average of speech patterns, and except for (35), these do not model the origins of language changes directly. Instead, an external event must disturb the system and push it from one stable state to another.

As we will see in Section 2, a general 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. It appears that spontaneous changes can only arise from random fluctuations in combination with some sort of momentum driven by social structure.

Labov (38) proposes a model in which females naturally change their individual speech over time, a force called *incrementation*. A semi-structured

approach as in (39) assumes a fully interconnected finite population but agents vary in their influence on learners. There is no age information available during learning. These models approximate the time course of a single change, in qualitative agreement with data, but neither addresses the origin of the change.

Some models use network dynamics rather than a mean-field assumption and allow learners to collect input disproportionately from nearby members of the population (40, 41). These models incorporate observations made by Labov (38, 42, 43) and others that certain individuals tend to lead the population in adopting a new language variant, and the change spreads along the friendship network.

In contrast, the new model analyzed in this article is built from an alternative perspective that can resolve the language change paradox. 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 (37, 38, 42, 44) and model it as a *stochastic grammar,* that is, a collection of similar idealized grammars, each of which is used randomly at a particular rate. From this continuous perspective, language change is no longer a paradox, but acquisition requires more than selecting a single idealized grammar as in the PAC framework. Instead, children must learn multiple idealized grammars, plus the usage rates and whatever conditions affect them.

Crucially, instead of limiting learners' input to example sentences, the new model assumes that children also know the age of the speaker and prefer not to sound outdated. They bias their speech against archaic forms by incorporating a prediction step into their acquisition of a stochastic grammar, which introduces incrementation without directly imposing it as in (38). The author is unaware of any other models with this prediction feature. It introduces momentum into the dynamics, which generates the desired metastability: 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 will be called *prediction-driven instability.*

The majority of the language modeling literature does not focus on the formal aspects of mathematical models, such as confirming that the dynamics are well-posed or deriving a continuous model as the limit of a discrete model, even though such details are known to be significant models (45).

The author has performed numerical simulations of the discrete form of model to confirm that it has the desired behavior (46) but its continuous form has yet to be placed on a sound theoretical foundation. So in this article, we formulate the age-structure model as a Markov chain, then 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

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.

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 type of the n-th agent. We also define a count vector C where $C_k(t)$ is the number of agents of type k,

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

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

$$M(j) = \sum_{k=0}^{K} \left(\frac{k}{K}\right) Z_k(j) \tag{1}$$

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. At each time step, each individual agent is examined:

- 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,

$$Q_k(m) = \binom{K}{k} q(m)^k (1 - q(m))^{K-k}. \tag{2}$$

The *mean learning function* $q(m)$ is the mean usage rate of children learning from a population with a mean rate m,

$$q(m) = \sum_{k=0}^{K} \left(\frac{k}{K}\right) Q_k(m). \tag{3}$$

The mean learning function must be S-shaped to ensure that there are two equilibrium states, representing populations dominated by one grammar or the other. In general, q is assumed to be smooth, strictly increasing, with one inflection point, and $0 < q(0) < 1/4$ and $3/4 < q(1) < 1$.

A curved mean 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 language: A growing body of evidence (47) indicates that for the task of learning a language with multiple ways to say something, adults tend to use all the options and match the usage rates in the given data, but children prefer to pick one option and stick with it. Beyond these general properties, this learning model makes no attempt to directly represent the neurological details of language acquisition, although researchers are exploring this area (24, 48–51).

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.

This Markov chain is regular. Although it spends most of its time hovering near a state dominated by one idealized grammar or the other, it must eventually exhibit spontaneous language change by switching to the other. However, computer experiments show that under this model, a population takes an enormous amount of time to switch dominant grammars. This model is therefore unsuitable for simulating language change on historical time scales. A further undesirable property is that when 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 (37, 48) show that language changes typically run to completion monotonically and do not reverse themselves partway through (but see (44) 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 (42).

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 outdated.

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 type of the n-th youth and $V_n(j)$ be the type 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,

$$M_C(j) = \sum_{k=0}^{K} \left(\frac{k}{K}\right)\left(\frac{C_k(j)}{N}\right)$$
$$M_D(j) = \sum_{k=0}^{K} \left(\frac{k}{K}\right)\left(\frac{D_k(j)}{N}\right) \tag{4}$$

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 r_R it is resampled from the youth population.
- With probability $1 - p_D - r_R$ 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 predicted mean usage rate 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 $0, 1/5, \ldots, 1$ of spoken sentences. The learning distribution $Q(m)$ is a binomial distribution with parameters $q(m)$ and 5, where

$$q(m) = \frac{1}{32} + \frac{3600}{751}\left(\frac{33m}{1280} + \frac{161m^2}{320} - \frac{m^3}{3}\right) \tag{5}$$

The prediction function $r(x, y)$ is based on an exponential sigmoid, as in Figure 2. 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,

$$r(x, y) = \sigma(t_1 + \alpha h) = \frac{1}{1 + \left(\frac{1-x}{x}\right)^{\alpha+1}\left(\frac{y}{1-y}\right)^{\alpha}} \tag{6}$$

For the example calculations in this paper, $\alpha = 3$. Observe that r is a rational function, and is continuous everywhere in $(0, 1) \times (0, 1)$ except at the corners $(0, 0)$ and $(1, 1)$.

The combined mean learning-prediction function $q(r(x, y))$ is plotted in Figure 1. An important feature is that since $q(0) > 0$ and $q(1) < 1$, the graph is slightly above the plane given by $(x, y) \mapsto x$ along the edge where $x = 0$, and is slightly below that plane along the edge where $x = 1$. This means that given an initial condition where one of the idealized grammars is not used at all, there is a non-zero probability that it will appear spontaneously.

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 3 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 develop a continuous limit for the Markov chain governing the speech distributions $X = C/N$ and $Y = D/N$ of the younger and older generations, which are points in the simplex,

$$\mathcal{S}^n = \left\{ (x_0, \ldots, x_n) \;\middle|\; x_k \in [0, 1], \sum x_k = 1 \right\}.$$

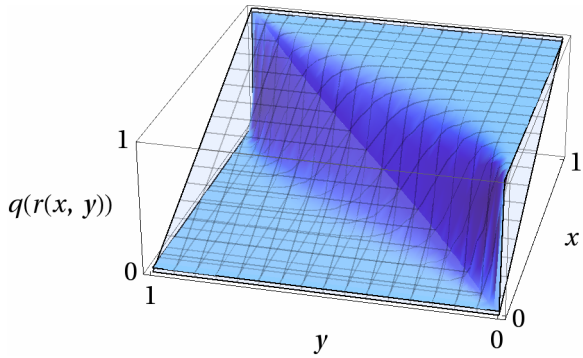

FIGURE 1. The learning-prediction function $q(r(x,y))$ and the plane given by the graph of $(x,y) \mapsto x$.

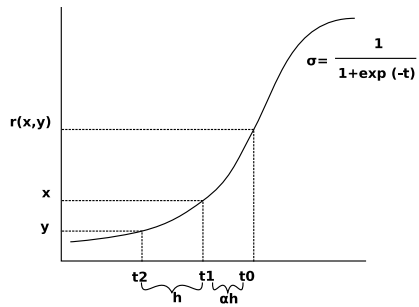

FIGURE 2. An illustration of the prediction function r.

In the limit as the population size N increases without bound, the Markov chain $(X(j), Y(j)) : \mathbb{N} \to \mathcal{S}^n \times \mathcal{S}^n$ ought to converge to the solution $(\xi(t), \zeta(t)) : [0, \infty) \to \mathcal{S}^n \times \mathcal{S}^n$ of a martingale problem. To formulate it, we must calculate the infinitesimal drift and covariance functions.

3.2.1. *Notation.* 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. 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

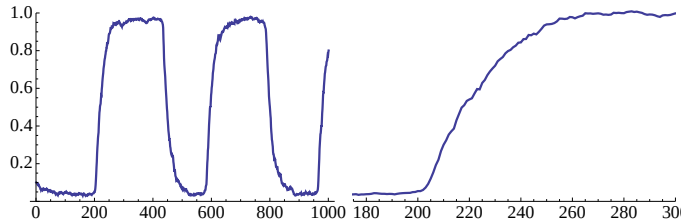

FIGURE 3. Trajectory of the mean usage rate $M_C(t)$ of G_2 in the young generation from a sample path of the age-structured Markov chain. Left: The path from time 0 to 1000 years, showing several changes between G_1 (low) and G_2 (high). Right: The path from time 175 to 300 years, showing a single grammar change.

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

$$\mathbb{E}_j\left(C'_k\right) = \sum_n g(n,k) = (1-p_D)C_k + p_D N Q_k \tag{7}$$

$$\begin{aligned}
\mathrm{Var}_j\left(C'_k\right) &= \sum_n g(n,k) - g(n,k)^2 \\
&= (1-p_D-r_R)C_k + p_D N Q_k + r_R N X_k \\
&\quad - (1-p_D-r_R)^2 C_k \\
&\quad - 2(1-p_D-r_R)C_k(p_D Q_k + r_R X_k) \\
&\quad - N(p_D Q_k + r_R X_k)^2
\end{aligned} \tag{8}$$

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

$$\begin{aligned}\mathrm{Cov}_j\left(C'_k, C'_h\right) &= \sum_n \mathrm{Cov}_j\left(\mathbf{1}(U'_n = k), \mathbf{1}(U'_n = h)\right) \\ &= -\sum_n g(n,k)g(n,h) \\ &= -\big((1-p_D-r_R)C_k(p_DQ_h + r_RX_h) \\ &\quad + (1-p_D-r_R)C_h(p_DQ_k + r_RX_k) \\ &\quad + N(p_DQ_k + r_RX_k)(p_DQ_h + r_RX_h)\big)\end{aligned} \tag{9}$$

It follows that

$$\mathbb{E}_j\left(\frac{X'_k - X_k}{1/N}\right) = r_D(Q_k - X_k), \tag{10}$$

which gives the infinitesimal drift component for a martingale problem. We also need an estimate of the covariance matrix for X:

$$\mathrm{Var}_j\left(X'_k\right) = \frac{1}{N}\left((2r_R - r_R^2)(X_k - X_k^2)\right) + O\left(\frac{1}{N^2}\right) \tag{11}$$

$$\mathrm{Cov}_j\left(X'_k, X'_h\right) = -\frac{1}{N}\left((2r_R - r_R^2)X_kX_h\right) + O\left(\frac{1}{N^2}\right) \tag{12}$$

Similar drift and covariance formulas can be derived for Y.

As a further simplification, we can rescale time by a factor of r_D. This finally yields the infinitesimal drift function

$$\mathbf{b}\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \begin{pmatrix}Q-\xi\\ \xi-\zeta\end{pmatrix} \text{ where } \xi, \zeta, Q \in \mathcal{S}^n \tag{13}$$

and the infinitesimal covariance function

$$\mathbf{a}\begin{pmatrix}\xi\\ \zeta\end{pmatrix} = \varepsilon^2 \begin{pmatrix} \xi_1 - \xi_1^2 & -\xi_1\xi_2 & \cdots & & & \\ -\xi_1\xi_2 & \ddots & & & 0 & \\ \vdots & & & & & \\ & & & \zeta_1 - \zeta_1^2 & -\zeta_1\zeta_2 & \cdots \\ & 0 & & -\zeta_1\zeta_2 & \ddots & \\ & & & \vdots & & \end{pmatrix} \tag{14}$$

where

$$\varepsilon = \sqrt{\frac{2r_R - r_R^2}{r_D}} = \sqrt{\frac{1-(1-r_R)^2}{r_D}} \text{ and } \xi, \zeta \in \mathcal{S}^n \tag{15}$$

Note that ξ_0 and ζ_0 are omitted from the dynamics because of the population size constraints

$$\xi_0 = 1 - \xi_1 - \cdots - \xi_K$$
$$\zeta_0 = 1 - \zeta_1 - \cdots - \zeta_K$$

Furthermore, if the resampling feature is removed by setting $r_R = 0$, 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.

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 distribution 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 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,

$$\bar{X}(t) = X(\lfloor Nt \rfloor)$$
$$\bar{Y}(t) = Y(\lfloor Nt \rfloor) \tag{16}$$

The covariance function (14) reduces to a 2-by-2 diagonal matrix so it has a very simple square-root. With these definitions, we claim

Proposition 4.1. *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 $(\xi(t), \zeta(t)) : [0, \infty) \to (0,1) \times (0,1)$ of

$$
\begin{aligned}
\mathrm{d}\xi &= (q(r(\xi,\zeta)) - \xi)\mathrm{d}t + \varepsilon\sqrt{\xi(1-\xi)}\mathrm{d}B^{\xi} \\
\mathrm{d}\zeta &= (\xi - \zeta)\mathrm{d}t + \varepsilon\sqrt{\zeta(1-\zeta)}\mathrm{d}B^{\zeta} \\
\xi(0) &= \xi_0 \text{ and } \zeta(0) = \zeta_0
\end{aligned}
\tag{17}
$$

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

Proof. We will show that the results of Chapter 8 of Durrett (52) apply, specifically theorem 7.1, which implies this result.

The calculations (10), (11), and (12) in Section 3.2 verify that the step-to-step drift, variances, and covariances of the Markov chain converge to the corresponding functions in the SDE as the time step size $1/N$ goes to zero. The remaining condition to check is Durrett's hypothesis (A) for theorem 7.1, which is that the martingale problem associated to (17) is well posed. That proof forms the rest of this section. 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. **Well-posedness of the SDE.** The SDE (17) has pathwise-unique strong solutions, as we will prove. This implies uniqueness in distribution (52, §5.4 theorem 4.1) which implies that the martingale problem is well posed (52, §5.4 theorem 4.5).

Lemma 4.2. *There is a number $\varepsilon_0 > 0$ such that if $\varepsilon < \varepsilon_0$ then the following holds: Given an initial condition $(\xi_0, \zeta_0) \in (0,1) \times (0,1)$, the system* (17) *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: Near 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

$$\begin{aligned} u &= 2\xi - 1 \in (-1, 1) \\ v &= 2\zeta - 1 \in (-1, 1) \\ \kappa &= \frac{u}{\sqrt{1-u^2}} \in (-\infty, \infty) \\ \lambda &= \frac{v}{\sqrt{1-v^2}} \in (-\infty, \infty) \end{aligned} \tag{18}$$

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,

$$\begin{aligned} \mathrm{d}\kappa &= \left(\overbrace{2(1+\kappa^2)^{3/2}(q(r(\xi,\zeta)) - \xi) + \frac{3}{2}\varepsilon^2\kappa(1+\kappa^2)}^{b_1=} \right) \mathrm{d}t \\ &\quad + \overbrace{\varepsilon(1+\kappa^2)^{1/2}}^{\sigma_1=} \mathrm{d}B^\xi \\ \mathrm{d}\lambda &= \left(\overbrace{2(1+\lambda^2)^{3/2}(\xi - \zeta) + \frac{3}{2}\varepsilon^2\lambda(1+\lambda^2)}^{b_2=} \right) \mathrm{d}t \\ &\quad + \overbrace{\varepsilon(1+\lambda^2)^{1/2}}^{\sigma_2=} \mathrm{d}B^\zeta \end{aligned} \tag{19}$$

Both the drift and standard deviation are continuously differentiable as functions of κ and λ, so they automatically satisfy a local Lipschitz inequality, as required by the standard theorem concerning the existence and uniqueness of solutions (52, §5.3).

The standard theorem also requires a growth constraint formulated as follows. Let us generalize the usual big-O notation, using

$$f(\kappa, \lambda) = g(\kappa, \lambda) + \mathcal{O}^2$$

to mean that there exists a constant $A > 0$ such that for all κ and λ,

$$f(\kappa, \lambda) - g(\kappa, \lambda) < A(1 + \kappa^2 + \lambda^2).$$

The required growth constraint is $\beta = \mathcal{O}^2$ where

$$\beta = 2\kappa b_1 + 2\lambda b_2 + \sigma_1^2 + \sigma_2^2. \tag{20}$$

For many stochastic differential equations, this bound is straightforward to verify because the corresponding β has no terms of degree higher than 2. The difficulty here is that β contains degree 4 terms, so we must confirm that they are negative for sufficiently large κ and λ. 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. Eliding terms of degree 2 and below and leaving the argument to q implicit,

(21)

$$\beta = \overbrace{\left(4(q-\xi)(\operatorname{sgn}\kappa) + 3\varepsilon^2\right)\kappa^4 + \left(4(\xi-\zeta)(\operatorname{sgn}\lambda) + 3\varepsilon^2\right)\lambda^4}^{\beta_4 =} + \mathcal{O}^2$$

For each sector of the plane (Figure 4) a slightly different argument guarantees that if ε is sufficiently small, then $\beta_4 < 0$ when κ and λ are large, which implies $\beta = \mathcal{O}^2$.

The southeast sector, marked SE, is defined by $0 \le -\lambda \le \kappa$, $7/8 \le \xi < 1$, and $1 - \xi \le \zeta \le 1/2$. Replacing λ^4 with κ^4, ζ with $1/2$, $-\xi$ with $-7/8$, and q with 1 yields

$$\beta_4 \le \left(-1 + 6\varepsilon^2\right)\kappa^4.$$

It therefore suffices to require $\varepsilon^2 < 1/6$. A similar argument applied to the northwest sector, marked NW, yields the same constraint.

The south sector, marked S, is defined by $0 \le \kappa \le -\lambda$, $0 < \zeta \le 1 - q(1)/2$, and $1/2 \le \xi \le 1 - \zeta$. Replacing κ^4 with λ^4, $-\xi$ with $-1/2$, q with $q(1)$, and ζ with $1 - q(1)/2$ yields

$$\beta_4 \le (2(-1 + q(1)) + 6\varepsilon^2)\lambda^4.$$

It therefore suffices to require $\varepsilon^2 < (1-q(1))/3$. A similar argument applied to the north sector, marked N, yields the constraint $\varepsilon^2 < q(0)/3$.

Define $\zeta_E = (1 + q(1))/2$ and let λ_E be the corresponding value of λ. The northeast sector, marked NE, is defined by $0 \le \kappa$, $1/2 \le \xi < 1$, $0 < \lambda_E \le \lambda$, and $\zeta_E \le \zeta < 1$. Replacing κ^4 with λ^4, q with $q(1)$, and $-\zeta$ with $-\zeta_E$, and canceling ξ yields

$$\beta_4 \le (-2(1 - q(1)) + 6\varepsilon^2)\lambda^4.$$

It therefore suffices to require $\varepsilon^2 < (1-q(1))/3$. A similar argument applied to the southwest sector, marked SW and defined by $\zeta_W = q(0)/2$, $0 < \zeta \le \zeta_W$, and $0 < \xi \le 1/2$, yields the constraint $\varepsilon^2 < q(0)/3$.

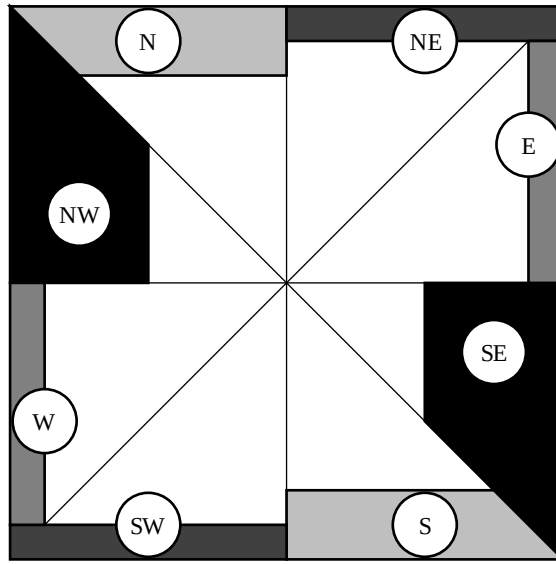

FIGURE 4. Schematic drawing of the sectors of the (κ, λ) plane for estimating β_4. The thin lines represent points satisfying $\kappa = 0$, $\lambda = 0$, $\lambda = \kappa$, and $\lambda = -\kappa$; or equivalently $\xi = 1/2$, $\zeta = 1/2$, $\zeta = \xi$, and $\zeta = 1 - \xi$.

The east sector, marked E, is defined by $0 \le \lambda \le \lambda_E \le \kappa$, and $\zeta \le \zeta_E \le \xi < 1$. Using the upper bound on λ and factoring out κ^4,

$$\beta_4 \le \kappa^4 \left(4(q - \xi) + 3\varepsilon^2 + \left(\frac{\lambda_E}{\kappa} \right)^4 \left(4(\xi - \zeta) + 3\varepsilon^2 \right) \right)$$

The right hand term becomes insignificant as $\kappa \to \infty$. Specifically, Let $\kappa_E = \lambda_E / \sqrt{\varepsilon}$. Combining $\kappa \ge \kappa_E$ with the already-derived constraint that $\varepsilon^2 < 1/6$ and $0 \le \xi - \zeta \le 1/2$, the inequality for β_4 simplifies to

$$\beta_4 \le \kappa^4 (4(q - \xi) + 6\varepsilon^2).$$

Replacing q with $q(1)$ and $-\xi$ with $-\zeta_E$ yields

$$\beta_4 \le \kappa^4 (-2(1 - q(1)) + 6\varepsilon^2).$$

It therefore suffices to require $\varepsilon^2 < (1-q(1))/3$. A similar argument applied to the west sector, marked W, yields the constraint $\varepsilon^2 < q(0)/3$.

Combining these constraints on ε and taking the strongest, define

$$\varepsilon_0 = \sqrt{\min\left\{\frac{1-q(1)}{3}, \frac{q(0)}{3}\right\}} \tag{22}$$

For each $0 < \varepsilon < \varepsilon_0$, all the constraints on ε in the preceding sector-by-sector analysis hold. Thus, $\beta = \mathcal{O}^2$, and standard results (52, §5.3 theorems 3.1 and 3.2) imply the existence and uniqueness of solutions to (17). □

5. Discussion

5.1. **Comparison to the deterministic limit.** In the deterministic limit $\varepsilon = 0$, the dynamical system (17) has two stable equilibria representing populations where both generations are dominated by one grammar or the other. The separatrix forming the boundary between the two basins of attraction 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 random fluctuations cause it to stray across the separatrix, where it will be blown toward the other equilibrium. These separatrix-crossing events generate spontaneous monotonic language changes separated by reasonably long intervals of temporary stability.

5.2. **Memory kernel form.** Another way to understand this form of instability is to express ζ as an average of ξ over its past, with an exponential kernel giving greater weight to the recent past. This is accomplished by first making the simplification of applying the resampling step from the Markov chain only to the younger generation, which removes the random term from $\mathrm{d}\zeta$ in (17) but not from $\mathrm{d}\xi$. This yields a linear ordinary equation for ζ with ξ acting as an inhomogeneity

$$\frac{\mathrm{d}\zeta}{\mathrm{d}t} = \xi - \zeta, \text{ with solution } \zeta(t) = \mathrm{e}^{-t}\zeta_0 + \int_0^t \mathrm{e}^{-(t-s)}\xi(s)\mathrm{d}s.$$

With this simplification, the dynamics for ξ take the form of a stochastic functional-delay differential equation

$$\mathrm{d}\xi(t) = \big(q(r(\xi(t), K_t\xi)) - \xi(t)\big)\mathrm{d}t + \varepsilon\sqrt{\xi(t)(1-\xi(t))}\mathrm{d}B \tag{23}$$

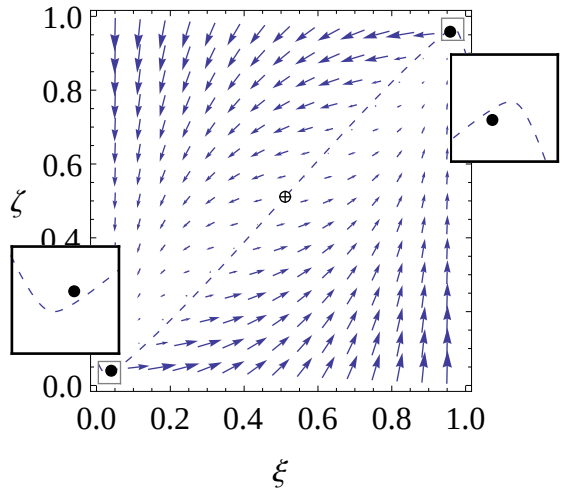

FIGURE 5. Phase portrait for (17) with $\varepsilon = 0$. The crossed dot $\oplus$ is a saddle point, the two dots $\bullet$ are sinks, and the dashed curve is the separatrix between their basins of attraction. The arrows indicate the direction of the vector field. Inset boxes show magnified pictures of the areas around the sinks.

where the delay appears through convolution with a memory kernel

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

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. Chance deviations of sufficient size are amplified when children detect them and predict that the trend will continue, yielding prediction-driven instability.

5.3. **Comparison to other biological models.** The discrete and continuous models as described Sections 2 and 3.1 are based on the Wright-Fisher model of population genetics as described in (52), which is formulated as a Markov chain and its limit as a stochastic differential equation for an infinite population. The Wright-Fisher model does not incorporate age structure or forces

such as learning and prediction that are not present in biological selection-mutation processes.

A related dynamical system is the FitzHugh-Nagumo model for a spiking neuron (53, 54), which is a general family of two-variable dynamical systems. Its structure is similar to Figure 5 except that it has only the lower left stable equilibrium, 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 as long as ε is sufficiently small (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 (17) to that form but the system remains confined to a square, and the change of variables to (19) on the whole plane has a non-constant coefficient on the Brownian motion. Thus, the theory of FitzHugh-Nagumo models must be adapted before it can be applied to this language model.

6. Conclusion

The main goal of this article is to build a mathematical model that can represent spontaneous language change in a population between two metastable 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.

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, which have the form of drift terms plus a small stochastic perturbation. Looking at the limit of zero noise, the prediction-driven instability comes from the proximity of stable sinks to the separatrix of their basins of attraction. The instability comes from the general geometry of the phase space as in Figure 5. Alternatively, the prediction process may be understood as comparing the current state of the population to an average emphasizing its recent past, and chance sudden changes trigger the instability. Concrete formulas were given for q, r, and Q, but the interesting behavior is not limited to these examples. The proof that the system of SDEs is well-posed relies only on general properties of q, r, and Q.

Remaining work on this model includes the extension of the proofs in Section 4 to any number of dimensions. Calculations show that such an extension is potentially useful for modeling spontaneous transitions among multiple possible languages (46).

Future studies of this model will include adapting and applying techniques for studying noise-activated transitions among meta-stable states, including exit time problems (55–57). 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] Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, November 2002.

[2] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

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

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

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

[6] Natalia L. Komarova and Martin A. Nowak. The evolutionary dynamics of the lexical matrix. *Bulletin of Mathematical Biology*, 63(3):451–485, 2001.

[7] Peter E. Trapa and Martin A. Nowak. Nash equilibria for an evolutionary language game. *Journal of Mathematical Biology*, 41:172–1888, 2000.

[8] Joshua Plotkin and Martin A. Nowak. Language evolution and information theory. *Journal of Theoretical Biology*, 205:147–159, 2000.

[9] Martin A. Nowak, Joshua Plotkin, and V. A. A. Jansen. Evolution of syntactic communication. *Nature*, 404(6777):495–498, 2000.

[10] Martin A. Nowak, Joshua Plotkin, and David C. Krakauer. The evolutionary language game. *Journal of Theoretical Biology*, 200:147–162, 1999.

[11] Martin A. Nowak, D. C. Krakauer, and A. Dress. An error limit for the evolution of language. *Proceedings of the Royal Society of London, Series B*, 266:2131–2136, 1999.

[12] Felipe Cucker, Steve Smale, and Ding-Xuan Zhou. Modeling language evolution. *Foundations of Computational Mathematics*, 4(3):315–343, July 2004.

[13] Willem Zuidema and Bart de Boer. The evolution of combinatorial phonology. *Journal of Phonetics*, 37:125–144, 2009. doi: 10.1016/j.wocn.2008.10.003.

[14] Partha Niyogi. *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge, MA, 2006.

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

[16] 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, Cambridge, UK, 2002. URL `http://www.cl.cam.ac.uk/users/ejb/creo-evol.ps.gz`.

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

[18] Simon Kirby and James R. Hurford. The emergence of structure: An overview of the iterated learning model. In Angelo Cangelosi and Dominico Parisi, editors, *Simulating the Evolution of Language*, pages 121–148. Springer-Verlag, New York, 2002.

[19] 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.

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

[21] Lisa Pearl and Amy Weinberg. Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development*, 3(1):43–72, 2007.

[22] David Lightfoot. *The Development of Language: Acquisition, Changes and Evolution*. Blackwell, Malden, MA, 1999.

[23] David Lightfoot. *How to Set Parameters: Arguments from Language Change*. MIT Press, Cambridge, MA, 1991.

[24] William Garrett Mitchener and Misha Becker. Computational models of learning the raising-control distinction. *Research on Language and Computation*, April 2011. ISSN 1570-7075. doi: 10.1007/s11168-011-9073-6. available on-line; in print soon.

[25] Misha Becker. There began to be a learnability puzzle. *Linguistic Inquiry*, 37(3):441–456, 2006. doi: 10.1162/ling.2006.37.3.441.

[26] Martin A. Nowak. *Evolutionary Dynamics: Exploring the equations of life*. Harvard University Press, 2006.

[27] W. Garrett Mitchener and Martin A. Nowak. Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93, January 2003. doi: 10.1006/bulm.2002.0322.

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

[29] K. M. Page and M. A. Nowak. Unifying evolutionary dynamics. *Journal of Theoretical Biology*, 219:93–98, 2002.

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

[31] Martin A. Nowak and Natalia L. Komarova. Towards an evolutionary theory of language. *Trends in Cognitive Sciences*, 5(7):288–295, July 2001.

[32] 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.

[33] W. Garrett Mitchener. Bifurcation analysis of the fully symmetric language dynamical equation. *Journal of Mathematical Biology*, 46:265–285, March 2003. doi: 10.1007/s00285-002-0172-8.

[34] W. Garrett Mitchener. A mathematical model of the loss of verb-second in Middle English. In Nikolaus Ritt, Herbert Schendl, Christiane Dalton-Puffer, and Dieter Kastovsky, editors, *Medieval English and its Heritage: Structure, Meaning and Mechanisms of Change*, volume 16 of *Studies in English Medieval Language and Literature*. Peter Lang, Frankfurt, 2006. Proceedings of the 13th International Conference on English Historical Linguistics.

[35] 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.

[36] W. Garrett Mitchener. Mean-field and measure-valued differential equation models for language variation and change in a spatially distributed population. *SIAM Journal on Mathematical Analysis*, 42(5): 1899–1933, January 2010. doi: 10.1137/07069955X.

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

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

[39] W. Garrett Mitchener. Inferring leadership structure from data on a syntax change in english. In Carlos Martín-Vide, editor, *Scientific Applications of Language Methods*, volume 2 of *Mathematics, Computing, Language, and Life: Frontiers in Mathematical Linguistcs and Language Theory*, chapter 13, pages 633–662. Imperial College Press, London, 2011.

[40] Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079, August 2010. ISSN 0024-3841. doi: 10.1016/j.lingua.2010.02.001.

[41] Samarth Swarup and Les Gasser. Unifying evolutionary and network dynamics. *Physical Review E*, 75(6):066114, June 2007. doi: 10.1103/PhysRevE.75.066114.

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

[43] William Labov. Transmission and diffusion. *Language*, 83(2):344–387, June 2007.

[44] 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.

[45] Richard Durrett and Simon Levin. The importance of being discrete (and spatial). *Theoretical Population Biology*, 46(3):363–394, December 1994.

[46] W. Garrett Mitchener. A mathematical model of prediction-driven instability: How social structure can drive language change. *Journal of Logic, Language and Information*, 2011. to appear.

[47] 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.

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

[49] Amy Perfors, Joshua B. Tennenbaum, and Elizabeth Wonnacott. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(Special Issue 03):607–642, 2010. doi: 10.1017/S0305000910000012.

[50] Afra Alishahi and Suzanne Stevenson. A Computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834, 2008. doi: 10.1080/03640210801929287.

[51] Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 10(3):307–321, 2007. doi: 10.1111/j.1467-7687.2007.00585.x.

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

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

[54] 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.

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

[56] 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.

[57] 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.