

PROJECT DESCRIPTION FOR W. G. MITCHENER

1. INTRODUCTION

Mitchener proposes to study stochastic processes and dynamical systems that model language learning and change in a population, supported by statistical analyses of manuscript data.

Some fundamental questions in linguistics are: When can a language change be attributed to natural fluctuations from learning and usage, and when must an extralinguistic event be invoked? What is the evolution of a language change? Is the set of known languages a well-distributed sample of the set of all possible human languages? How many possibilities are yet to be seen? What are the respective influences of spatial structure, social structure, and literacy on language change? How does language evolve in the biological sense? The PI proposes to express and analyze these problems in terms of probability and stochastic processes, and develop the mathematical machinery that will enable linguists to address these important questions that would otherwise be impossible to state and answer precisely.

These questions in linguistics give rise to interesting and difficult questions in mathematics [1, 5–13, 18, 19, 23, 24, 27–29, 31, 33–36, 38, 39, 52–64, 68, 70–73, 77]. The following subsections summarize the PI’s past and proposed research projects. Details follow in Sections 2 and 3.

1.1. Past research. Section 2 contains a summary of the PI’s past research into language modeling: an evolutionary model based on population game dynamics [30] with learning, and a model of word order change in Middle English driven by contact between regional dialects. Both are dynamical systems models that represent language change through bifurcations. Both are deterministic and make the strong assumption that speakers exclusively use one grammar from a discrete set of possibilities. The PI’s proposed research introduces random behavior and relaxes this strong assumption.

1.2. Deterministic dynamics of speech distributions. The PI has developed an infinite dimensional ODE for the time-dependent density $u(t, z)$ of the part of a population whose speech consists of samples from a distribution parameterized by z , described in Section 3.1. The dynamics for u then form a Banach-space valued differential equation, $u' = \beta(Q(u) - u)$, where Q is the learning algorithm, and β is a birth-death rate. The PI conjectures that there are unique solutions for all time that can be understood in some cases in terms of a finite dimensional projection. Under these deterministic dynamics, the size of the basin of attraction of each stable steady state gives information on how big a perturbation must be to shift the population to another basin, thereby modeling how drastic an extralinguistic event must be to trigger a language change.

Spoken language varies according to context and chance. Many previous models [23, 24, 36, 36, 44, 48–50, 52–55, 61] assume that an individual’s speech is described by one of a discrete set of idealized grammars. This simplification is inspired by advances in the theory of syntax, based on the assumption of perfectly regular speech under ideal circumstances. However, manuscript data shows that true speech is a mixture of different constructions. This model is designed to address this variability in a mathematically tractable way.

1.3. Markov chain model for a finite population. To step away from the simplification of an infinite population and verify the circumstances under which such a simplification is correct, the PI has formulated a Markov chain model of a finite population with arbitrary social structure and proposes to explore it through theory and computation. The details are in Section 3.2. The chain almost satisfies a monotonicity property that would allow for the use of a perfect sampling algorithm called coupling from the past [66] to sample from its stationary distribution. The PI conjectures that coupling from the past can be generalized to give estimates for the mixing time of almost monotonic Markov chains, which can then be applied to the question of how well the set of known languages represents the set of all possible languages.

1.4. Stochastic differential equation derived from the Markov chain. The PI has developed a stochastic differential equation by taking a limit of the Markov chain as the number of individuals becomes infinite, generalizing the Wright-Fisher process from population genetics [16]. See Section 3.3. The deterministic model of Section 3.1 turns out to be dominated by stable equilibria. Stochastic dynamics allow the population to hover around one state then change spontaneously. The PI proposes to study this hovering behavior and apply the results to questions of spontaneous language change and the time course of a change.

1.5. Generalizations: spatial structure, social structure, literacy. The most interesting mathematical problem comes from modeling literacy. Literacy allows the past to directly influence the present through the written word, thereby naturally introducing time delays. Delay dynamics give rise to difficult and important theoretical problems in ODEs, PDEs, and SDEs [26, 32, 43, 51, 76]. In this case, the delay is especially challenging because learning can take place from an average of the population’s entire history. The PI proposes to find conditions under which the proposed ODEs and SDEs with delay have unique solutions for all time, to investigate stability and hovering behavior as in the non-delay models, and to apply these results to manuscript data: If an ancient culture had a strong literary standard, changes in its language might appear later in the written record than they appeared in speech.

Additionally, all of these models can be generalized to include spatial and social structure. For example, the population can be represented as a network of well-mixed compartments representing cities, regions, or social classes, with restricted flow between them.

1.6. Modeling acquisition and the subset problem. The PI has developed a model of language acquisition based on Bayesian inference [22] as part of the Markov chain model: Hearers attempt to parse each sample sentence with a randomly selected idealized grammar. If the parse succeeds and if a learning heuristic indicates that the sentence is informative, the hearer uses Bayes’ rule to update its knowledge of the population’s speech, which is stored as a list of beta distributions. The naive heuristic of learning from every sentence fails.

A more intelligent heuristic only learns from sentences that cannot be parsed by changing parameters in the idealized grammar, and this succeeds. The PI proposes to further investigate this disparity, to prove theorems explaining why it happens, and to draw conclusions about what general properties learning algorithms must have to yield linguistic consensus.

Linguists argue from certain changes [41, 42] that children determine that their language should have certain constructions by hearing “unambiguous” cue sentences that clearly require those constructions. However, the exact meaning of “unambiguous” is imprecise as almost all sentences can be generated by several idealized grammars, and the set of possible grammars is not well-understood. Thus, the learning model must approximate learning from unambiguous data, and the PI’s model acquisition algorithm is a step toward a precise formulation of this linguistic concept.

The subset principle is an ongoing debate in the linguistics community [3, 23, 52, 67, 75]. Children only learn from positive evidence, that is, they generalize from sample sentences that are assumed to be grammatical, so how they determine that certain constructions are ungrammatical is a long-standing puzzle. One proposed resolution is that statistical patterns provide implicit negative evidence. The PI proposes to test this theory on simulated data and corpus data by developing a hierarchical Bayesian model [22] in which meaning types are generated by a certain distribution, and surface word orders are generated by distributions conditioned on the meaning type. The amount of data required for this inference should indicate whether statistical patterns are a plausible source of implicit negative evidence. The PI has been in contact with linguist Misha Becker at UNC Chapel Hill and plans to collaborate with her to test this model on data concerning the acquisition of raising and control verbs, as explained in Section 3.5.

2. PAST RESEARCH

2.1. The language dynamical equation. This section outlines the continuous model for the interaction of language learning and natural selection from the PI's dissertation under Martin A. Nowak [48], followed by summaries of the PI's related publications.

Consider a group of individuals that speak one of a finite set of grammars G_1, G_2, \dots, G_n . Denote by Q_{ij} the probability that a child learner will acquire grammar G_j when exposed to sample sentences generated by a parent speaking G_i . The linguistic data available to the child and the acquisition algorithm determine Q . Imperfect learning means that $Q_{ii} < 1$ for at least some i , which implies that sometimes the learner will acquire a different grammar. As a simplifying assumption, the Q matrix is taken to be constant in time. Hence, it most accurately reflects scenarios such as an isolated population or one subject to a constant level of contact with other languages, and it is of limited use when learning probabilities are fluctuating due to contact or other forces.

Consider a large, well-mixed population. The fraction of the population that speaks G_i is denoted x_i , with $\sum_i x_i = 1$. Individuals derive a benefit from communicating successfully with each other, according to a payoff matrix B , where B_{ij} is the benefit to a speaker of G_i from an encounter with a speaker of G_j . The entries of this matrix may include effects such as the benefit of correct communication, cost of ambiguity, and so forth. A natural assumption is that people communicate best with others who have the same grammar. In this case, B is diagonally dominant, which implies that each grammar is a strict Nash equilibrium. With perfect learning, each language would then be an evolutionarily stable equilibrium.

The fitness associated with grammar G_j is the weighted average payoff $F_j = \sum_k B_{jk}x_k$. Here we make the simplifying assumption that the communication game is the dominant source of fitness, thereby incorporating selection in favor of individuals who communicate well.

The average fitness of the population is given by $\phi = \sum_j F_j x_j$. The language dynamical equation is given by

$$(1) \quad \dot{x}_j = \sum_{i=1}^n F_i x_i Q_{ij} - \phi x_j, \quad j = 1 \dots n.$$

The language dynamical equation exhibits a rich variety of behavior, which is to be expected since it has two n by n matrices of parameters to be picked.

2.1.1. Bifurcation analysis of the fully symmetric language dynamical equation. [44] The PI solves (1) in a symmetric case in which one number q represents learning accuracy for all grammars. Prior work on this case [36] located two classes of fixed point, one symmetric where all grammars are equally represented, and n asymmetric ones with a dominant grammar. For small values of q , the symmetric fixed point is the only fixed point and it is stable. For intermediate values of q , the asymmetric fixed points come into existence and are stable. For larger values of q , the symmetric fixed point becomes unstable.

This paper completes the analysis. All fixed points are identified, including a large number of previously unknown saddles created through a complex sequence of bifurcations as q increases. The dynamical system in question is symmetric but taken in an arbitrary number of dimensions, so rigorously accounting for all fixed points, their stabilities, and bifurcations is non-trivial. The PI proves that trajectories generically converge toward one of the fixed points.

The linguistic interpretation is that under normal circumstances, learning is reliable, so q is large, the population has a single dominant grammar, and the language is stable. If an external event such as contact with another language introduces a linguistic disturbance, then q decreases. A large enough decrease in q could destabilize the single-grammar state of the population, allowing it to switch to another dominant grammar once the disruption dissipates.

2.1.2. *Competitive exclusion and coexistence of universal grammars.* [49] This paper extends (1) to include genetic variation in the form of multiple universal grammars (UGs).¹ We completely analyze the case of a single UG with two grammars, G_1 and G_2 . It exhibits either a single stable fixed point to which all populations converge, or a pair of stable fixed points separated by an unstable fixed point. If mutation introduces a second UG that admits only G_1 , it takes over if G_1 dominates the population, and dies out if G_2 dominates. More generally, we show that a multi-grammar UG is unstable when competing directly against single-grammar UGs. However, if G_1 is very ambiguous (meaning there is high probability of miscommunication) it cannot invade a multi-grammar UG with sufficiently reliable learning.

These results suggests that human UG may have once admitted many more grammars and the current state may represent the influence of mutations limiting UG to fewer grammars.

2.1.3. *Chaos and language.* [50] Many language models are analyzed with a focus on equilibrium behavior, representing the fact that human languages are indeed stable on time scales of about a century. In this paper, the PI discusses an instance of the language dynamical equation with chaotic behavior. The Q matrix creates this chaos by encoding a natural flow among 5 possible grammars. As an error rate parameter increases, the system undergoes period doubling bifurcations that lead to chaos, something similar to Šilnikov’s mechanism [25].

On longer time scales, language is known to change unpredictably and is sometimes very sensitive to perturbations caused by contact with other languages. Hence, chaotic oscillations are a qualitatively accurate model.

2.1.4. *Game dynamics with learning and evolution of UG.* [47] This paper continues the analysis of the multi-UG model from [49] by considering a case of two UGs, each of which admits two grammars. The PI proves that if the payoff matrix B obeys certain inequalities, then independent of the learning matrix Q , the two UGs experience competitive exclusion: One or the other will dominate the population and neither can invade the other. This result holds even if Q is allowed to change in time. An interesting example called “accidental stability” is presented that does not fall under this result; it has the property that each UG can invade the other UG if a certain one of its grammars dominates the population, but not if the other dominates. Hence, the historical accident of which grammar dominates a population determines whether a later mutation takes over or dies out. This means that there is no straightforward notion of relative fitness for UG.

2.2. A model of word order change in Middle English. This section summarizes some of the PI’s work as a research associate at Duke University under the support of the VIGRE grant. It was presented at several math conferences, and the International Conference for English Historical Linguistics (ICEHL) in Vienna in August, 2004 and will appear as [46].

We consider a population split into northern and southern regions, where individuals use one of two grammars, either G_1 or G_2 . Define x_N to be the fraction of northerners with G_1 and x_S to be the fraction of southerners with G_1 . After making certain equilibrium assumptions about the population and rescaling time, the result is a system of differential equations

$$(2) \quad \begin{aligned} \dot{x}_N &= q(x_N) - (1 + \alpha)x_N + \alpha x_S \\ \dot{x}_S &= q(x_S) - (1 + \beta)x_S + \beta x_N \end{aligned}$$

where q is a sigmoid-shaped learning function, and α and β are parameters that represent mixing between the two regions. For small α and β corresponding to little or no mixing, there is a stable fixed point representing a split population with $x_N \approx 1$ and $x_S \approx 0$ corresponding to the historical

¹*Universal grammar* or *UG* is the set of innate hints and limitations that children use to properly generalize from a set of sample sentences to a complete adult language. For this model, UG consists of a set of allowed grammars and a learning process for picking one.

state. If α and β increase, that fixed point vanishes in a saddle-node bifurcation and a split population will tend to $x_N = 0, x_S = 0$ representing the extinction of the northern dialect.

This dynamical system models a change in Middle English word order. Old English included a rule called verb-second or V2 that scrambled top-level sentences by moving the tensed verb to the front, and a topic in front of the verb. The topic can be any phrase and receives a certain emphasis. Middle English had at least two regional dialects with slightly different forms of V2 [20, 37, 40]. The southern dialect retained the Old English form of V2 and the special nature of its subject pronouns. The northern dialect was influenced by Old Norse and adopted its form of V2. Around 1400, the use of V2 began to disappear from both dialects, and is present in Modern English only in questions and certain idiomatic expressions.

Lightfoot [41] proposes that children acquire the northern V2 rule only if they hear enough *cue* sentences whose word order clearly exposes the fact that they have been re-organized from the underlying word order. The southern dialect allowed pronouns to appear just before the finite verb, resulting in numerous sentences where the verb is third. The theory is that contact between the northern and southern dialects caused northern children to hear too few cue sentences, hence they did not acquire the proper V2 rule. The northern grammar died out, giving way to a third non-V2 grammar like Modern English.²

Lightfoot’s proposed cue-based learning process can be translated directly into mathematics. Assume that there are two grammars, G_1 representing the northern V2 grammar, and G_2 representing the modern non-V2 grammar. Assume that children acquire G_1 if some fraction of the sentences they hear exhibit the proper cue, otherwise they select G_2 . Assume that a fraction x of the people around the child use G_1 , and that speakers of G_1 produce cue sentences at a rate p . If a speaker is selected at random, the resulting sentence is a cue with probability px . Assume that children hear n sentences, and that they must hear m or more cues to select G_1 . Then the probability that a child picks G_1 is

$$(3) \quad q(x) = \mathbb{P}(\text{child picks } G_1) = \sum_{j=m}^n \binom{n}{j} (px)^j (1 - px)^{n-j}$$

The result is a polynomial sigmoid function $q(x)$ for (2).

The main criticism to this model raised by Kroch is that the Middle English manuscripts indicate that writers were not confined to a single grammar, and exhibit considerable grammatical variety. There is reason to believe that speakers use multiple grammars at a variety of frequencies. Thus a change from one grammar to another is a population-wide shift in usage frequencies, rather than the disappearance of exclusive G_1 speakers. Section 3.1 discusses a modification to this model that addresses Kroch’s concern.

3. CURRENT AND PROPOSED RESEARCH

3.1. A continuous model of population-level language learning. A very simple model of learning dynamics is to assume a well-mixed population, where everyone uses one of two options exclusively: G_1 or G_2 . If x is the fraction of agents using G_1 , then learning can be expressed by

$$(4) \quad q(x) = \mathbb{P}(\text{child learns } G_1)$$

²There is some disagreement in the linguistics literature about the formulation V2 in Middle English. Lightfoot [41] proposes that the southern grammar was not actually V2, but was instead closer to the modern grammar. Other experts, notably Anthony Kroch and Fischer et al. [20], disagree, and claim that the southern dialect maintains the Old English V2 rule. The description here is an effort to synthesize the two proposals.

as in Section 2.2, and with a birth-death rate of β , the population should follow the dynamics³

$$x' = \beta(q(x) - x).$$

This one-dimensional dynamical system is easily solved for reasonable q : Populations tend toward one or more stable fixed points \bar{x} satisfying $q(\bar{x}) - \bar{x} = 0$. For example, under Lightfoot's cue-based learning as in (3), the function $q(x)$ is a polynomial sigmoid and there are three fixed points, two stable ones near 0 and 1, and an unstable one in the middle. This behavior is typical of observed language dynamics, where when two grammars are present in a population, one or the other generally (but not always) takes over.

However, analysis of manuscripts shows that individual writers use mixtures of the idealized grammars studied by linguists. To capture this variability mathematically while remaining in the realm of continuous models, we represent the population at time t by a density $u(t, z)$ of individuals using G_1 at rate z and G_2 at rate $1 - z$. Analogously, the dynamics of u should be

$$(5) \quad \frac{\partial u(t, z)}{\partial t} = \beta(Q(u(t, \cdot), z) - u(t, z)).$$

Here, $Q(f, z)$ is the density of children who learn to use G_1 at rate z , given that they are learning from speakers distributed according to density f . We can think of (5) as an infinite dimensional dynamical system, with $u(t)$ taking values in some Banach space X of functions $[0, 1] \rightarrow \mathbf{R}$,

$$(6) \quad u' = \beta(Q(u) - u).$$

3.1.1. *Well-posedness.* The PI has partial results (building from [14, 78]) concerning existence of unique solutions for all time when working in the space $X = L^1([0, 1] \rightarrow \mathbf{R})$, and proposes to complete these results and re-prove them in L^2 . Preliminary calculation suggest that L^1 is the obvious choice for X because u , being a probability density, should have fixed mass. The projected dynamics in Section 3.1.3 suggest that Hilbert space methods will give similar results in L^2 .

3.1.2. *Restriction to mean dynamics.* If we assume that the population is well-mixed, we may restrict Q to depend on u through its mean m , as in $Q(u) = q(m)$. Taking the mean of (6) yields simple one-dimensional dynamics,

$$(7) \quad m' = \beta(q(m) - m).$$

The PI proposes to prove that the mean dynamics control the complete dynamics of u : If m converges to a fixed point \bar{m} , then $q(m)$ should converge. Then for fixed z , the original PDE (5) becomes a one-dimensional dynamical system where $u(t, z)$ converges to $q(\bar{m}, z)$. Thus, almost all populations under the u dynamics converge to a fixed point.

3.1.3. *Restriction to projected dynamics.* Let us generalize the mean dynamics. Let $\phi_1, \phi_2, \dots, \phi_n$ be functions $[0, 1] \rightarrow \mathbf{R}$, and define projections k_j of u onto ϕ_j as

$$(8) \quad k_j(t) = \int_0^1 \phi_j(z)u(t, z)dz.$$

We now consider the restriction of (6) to finite dimensions, where $Q(u) = q(k_1, \dots, k_n)$. An ODE for k_j is obtained by multiplying (6) by ϕ_j and integrating, which results in

$$(9) \quad k_j' = \beta(q(k) - k_j).$$

The PI conjectures that such finite-dimensional dynamical systems can in general capture all the behavior of (6) given the restricted form of Q .

³We could rescale time to eliminate the birth rate as in Section 2.2, however in the formulation of a stochastic differential equation model, it will be simpler to keep time in its original scaling so as to clearly indicate the proper scaling between the drift and noise terms.

3.1.4. *Multiple grammars.* The PI proposes to extend this model to allow for arbitrary mixtures of any number of idealized grammars, not just two. That is, u will be a density on an n -vertex simplex. The application of this generalization is that some linguistic accounts of the loss of V2 in Middle English require three grammars—two regional dialects with different forms of V2 plus the modern grammar—so an extension to multiple grammars is required to properly simulate those accounts.

3.1.5. *Revised Middle English model.* The PI proposes to extend this model to include regional dialects by splitting the population into north and south compartments, thereby creating an improvement of the model from Section 2.2 for the loss of verb-second in Middle English.

3.2. A Markov chain for population-level language learning. The PI has developed a Markov chain model of a population of learning linguistic agents [45], and implemented its transition function, together with a perfect sampling algorithm, as a computer program. This simulation improves on other such simulations [6, 53] in several ways, based on ideas in [77]: (1) Agents may be arranged in an arbitrary social network, with edges connecting those that interact; (2) Agents can use any mixture of idealized grammars; (3) Learning requires the inference of a probability distribution on idealized grammars rather than the identification of a single idealized grammar.

An important difference from [53] and [6] is that the simulated learning algorithm is success driven rather than error driven. That is, rather than hold a hypothesis and move away from it when given a sample sentence that it cannot parse, it holds many hypotheses and strengthens the ones that can parse given sample sentences. This choice is part of the PI’s attempt to address the subset problem (see Section 3.5).

The simulated UG consists of all grammars determined by a fixed number of binary parameters. Each agent maintains a list of beta distributions, each of which represents its knowledge (prior distribution) of the usage frequency for one parameter. When presented with a sentence, the agent attempts to parse it with an idealized grammar, selected at random according to its prior distribution. If the parse succeeds, the agent picks a binary parameter and uses a heuristic to determine whether the sentence is informative about that parameter’s setting. If it is, then the agent updates the corresponding beta distribution so that its new prior is more likely to generate the same setting for the next parse.

The program currently uses two heuristics. One, called LEARNALWAYS, always reinforces the parameter’s setting. The other, called PARAMETERCRUCIAL, tries a second parse with the parameter in the opposite setting. If the sentence cannot be parsed, the parameter’s original setting is declared “crucial” to the parse and therefore the sentence is informative. Otherwise the sentence is discarded as being too ambiguous and no learning takes place.

In each round of the simulation, an agent is selected at random to be the hearer. One of its neighbors is selected to generate a sentence, and the hearer performs one round of the learning algorithm on the speaker’s sentence. An agent’s speech is generated randomly according to its prior, with an optional step to bias its speech toward the most likely parameter settings (according to its prior). Although the simulation can be adapted for an arbitrary social network, preliminary experiments have been limited to a loop. As a final detail, each hearer has a probability of dying and being replaced by a newborn (with all uniform priors) when selected.

A specific set of languages is required for the simulation. For now, it uses four languages defined by two binary syntactic options: The simulated languages have either plain subject-verb-object (abbreviated SVO) or SVO+V2 for their word order and are either pro-drop⁴ or not. Out of all the syntactic parameters identified by linguists, these two are the most relevant to studying the loss of V2 in Middle English and Old French.

⁴A *pro-drop* language such as Italian or Spanish allows subject pronouns to be dropped. The pronoun can be inferred from the verb ending.

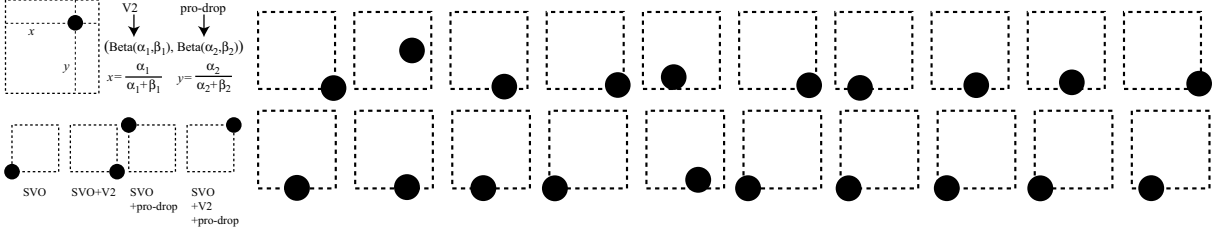


FIGURE 1. MCFTP samples from Markov chain simulation, using `LEARNALWAYS` (top) and `PARAMETERCRUCIAL` (bottom) for 10 agents configured in a loop. Each agent speaks to its left and right neighbors.

Mathematically, the simulation is an aperiodic, irreducible Markov chain with a huge number of states. With an appropriate ordering on its state space as in [45], this Markov chain is approximately monotonic, so it is possible to use a perfect sampling algorithm called monotone coupling from the past (MCFTP) as a non-rigorous heuristic to generate samples approximately according to its stationary distribution. See Figure 1 for examples.

The following problems remain to be solved concerning this simulation.

3.2.1. *Approximation by a stochastic differential equation.* The PI proposes to develop a stochastic differential equation approximating this simulation that should provide further understanding of its behavior in a more mathematically tractable setting. This large multi-stage project is described in Section 3.3. The SDE model and this simulation will be checked against each other for consistency.

3.2.2. *Revised sampling algorithm and mixing time estimate.* Since the Markov chain is only approximately monotonic, MCFTP serves as a non-rigorous heuristic for how many steps to take in the chain to arrive at a random population state that is close to the stationary distribution. The PI proposes to generalize MCFTP to draw samples and estimate the mixing time for an almost monotonic Markov chain.

Suppose X_t is a discrete time, aperiodic, irreducible Markov chain on a finite state space S with a transition function $\phi : R \times S \rightarrow S$. Let U_t be a sequence of independent identically distributed random variables in a probability space R such that $X_{t+1} = \phi(U_t, X_t)$. (For example, let $R = [0, 1]$ and require each U_t to be uniformly distributed.) *Monotonicity* means the state space has a partial ordering \succeq with maximal and minimal elements \hat{x} and \hat{x} such that $x \succeq y$ implies that for all $u \in R$, $\phi(u, x) \succeq \phi(u, y)$. MCFTP assumes monotonicity, and it can generate samples from the Markov chain’s stationary distribution and estimate its mixing time. However, the language simulation satisfies only the weaker condition that $\phi(U, x) \succeq \phi(U, y)$ with probability $\geq 1 - \varepsilon$. The PI proposes to develop an algorithm that can sample from the stationary distribution under this weaker assumption, and apply it to the simulation.

For the language simulation, the mixing time would provide an estimate of the population’s memory, that is, how long it takes to forget its initial state. An estimate of the mixing time of actual human languages would be very useful to linguists: A current practice in linguistics is to argue based on the distribution of the world’s languages as they stand now. However, it is unclear how big the space of possible human languages is, and how well the set of known languages represents this space. A mixing time estimate would allow one to determine whether human language is old enough to have forgotten its initial state, and therefore whether the known languages are at all a representative sample.

3.2.3. *Miscalibration.* Results from the model in Section 3.1 imply that if a learning algorithm is to lead to a population dominated by one idealized grammar, it must be miscalibrated to drive speakers to overestimate the population’s overall usage frequency of the most frequently used grammar. In the Markov chain simulation, the `LEARNALWAYS` heuristic results in improper miscalibration:

populations are not led to states dominated by one V2 setting or the other. PARAMTERCRUCIAL is properly miscalibrated. The PI conjectures that LEARNALWAYS is improperly miscalibrated for a wide range of language learning scenarios and proposes to prove this.

The fact that LEARNALWAYS is improperly miscalibrated has important applications, as it confirms a theory put forth by many linguists that the acquisition process is aware that sample sentences may be compatible with many idealized grammars and compensates by learning primarily from unambiguous cue sentences. The meaning of “unambiguous” in this sense is vague in the linguistics literature, because the complete set of valid human grammars is unknown and it is not known whether there are any truly unambiguous sentence types that occur in only one. LEARNALWAYS makes no attempt to filter out ambiguous sentences and has trouble learning V2. PARAMTERCRUCIAL is an easily-computed approximate measure of ambiguity, and its success in this simulation is an important proof of concept for the unambiguous cue theory.

3.2.4. *Data collection.* The PI proposes to run the simulation from a variety of initial conditions to determine how quickly it leaves the (presumed unstable) state where everyone uses mostly SVO+V2+pro-drop, in comparison to the other extreme states (SVO, SVO+V2, SVO+pro-drop). If the SVO+V2+pro-drop combination is noticeably less stable than the others, that would support the hypothesis that the combination of V2 and special properties of pronouns contributed to the loss of V2 in Middle English and Old French.

3.2.5. *Initial prior.* It is commonly assumed in linguistics that when languages have a choice between two options, such as pro-drop or non-pro-drop, one option is considered *marked*, meaning dispreferred, and the other is considered *unmarked*, meaning preferred or default. When a feature of a particular language is rare, it is considered marked, and more common features are considered unmarked. However, these notions are vague and experts frequently disagree on which options are marked and unmarked.⁵ The simulation can incorporate such markedness by using an informative prior. For example, to indicate that V2 is a marked option, newborn agents might start with a beta distribution skewed toward the non-V2 setting. It would then take significant evidence to shift the child to using V2. The PI proposes to use this simulation to make further explorations into learning marked and unmarked options, and to find evidence from simulation either supporting the need for the marked/unmarked distinction in linguistic theory, or showing that it is not necessary, or perhaps not the cleanest system for explaining asymmetry in the world’s languages.

3.3. **Population-level learning dynamics in a diffusion limit.** This section formulates a discrete stochastic model of a population of agents whose speech consists of arbitrary mixtures of two similar idealized grammars G_1 and G_2 . Thus, language acquisition is reduced to learning a frequency with which to use each grammar. From there, we derive a continuous stochastic process representing the limiting behavior for a large number of individuals, and describe questions to be investigated concerning that process and various generalizations thereof.

3.3.1. *Setup and fundamental results.* Consider a population that at each time m consists of N agents, each belonging to one of the K types $0, 1, 2, \dots, K - 1$. The intent is to interpret type k as meaning that the individual uses G_1 with frequency q_k and G_2 with frequency $1 - q_k$, with $q_k = k/(K - 1)$ for example. The restriction to a finite number of types is so that the population may be approximated by a finite dimensional stochastic differential equation.

Let $Z_k(m)$ be the number of individuals of type k at time m , and define $Y(m) = Z(m)/N$ to be the population distribution of the different types as a vector. To form $Z(m + 1)$ from $Z(m)$, copy

⁵For example, one might assume that children’s grammar initially consists of all default settings. Children seem to begin with a pro-drop grammar and only later develop non-pro-drop grammar when learning English, which suggests that pro-drop is unmarked. Alternatively, creoles result when children construct a complete human language from an artificial pidgin, and creoles are thought to represent all default settings. However, most creole languages are non-pro-drop, which suggests that pro-drop is marked.

each individual with probability $1 - \beta/N$, but with probability β/N replace the individual with a new individual of type θ , with each replacement selected independently according to the dynamic distribution $\mathbb{P}(\theta = k) = Q_k$, where Q is a vector that depends on $Z(m)$ and Q_k is the probability that a child acquires usage frequency p_k .

The following can be determined:

$$(10) \quad \mathbb{E}(Y(m+1) - y | Y(m) = y) = \frac{\beta}{N}(Q - y)$$

$$(11) \quad \text{Var}(Y_n(m+1) | Y(m) = y) = \frac{1}{N}(P_n - P_n^2)$$

where

$$P_n = \frac{\beta}{N}Q_n + \left(1 - \frac{\beta}{N}\right)y_n.$$

$$(12) \quad \text{Cov}(Y_n(m+1), Y_r(m+1) | Y(m) = y) = -\frac{1}{N} \left(\left(\frac{\beta}{N}\right)^2 Q_n Q_r + \frac{2\beta}{N^2} \left(1 - \frac{\beta}{N}\right) Q_n Q_r \right)$$

for $n \neq r$. With these results, we may take the limit as the number of individuals $N \rightarrow \infty$ of the process

$$(13) \quad X^N(t) = Y([Nt]).$$

Note that P_n converges to $Y(m, n)$ and we must think of Q as a function of X . Furthermore, the covariances between different components of Y and therefore of X are very small, $o(1/N)$, so in the limit, they will be zero. Following the derivation of the Wright-Fisher diffusion model of population genetics [16, 17] leads to the following stochastic differential equation for the n -th component of X :

$$(14) \quad X_n(t) = X_n(0) + \int_0^t \beta(Q_n(X(s)) - X_n(s)) dt + \int_0^t \sqrt{X_n(s)(1 - X_n(s))} dW_n(s).$$

Here, $W(s)$ is a standard multi-dimensional Brownian motion, that is, a Wiener process, with each component $W_n(t)$ an independent one-dimensional Wiener process, and the second integral is to be interpreted in the Itô sense.

Neglecting the stochastic integral (thereby removing noise from the mathematical model), leaves a system of ordinary differential equations

$$(15) \quad X' = \beta(Q(X) - X)$$

similar to (6), and under such dynamics with a reasonable learning algorithm for Q , populations should tend to some stable fixed point, representing an invariant distribution of the learning process. With the stochastic integral left in, the population should hover around one of these fixed points, and occasionally escape to hover around another.

A fully general analysis of (15) with arbitrary Q is all but impossible. However, if we assume that new agents learn by sampling speech from all members of the population equally, then it is reasonable to restrict our attention to learning functions that depend on X only through limited information. For example, we may restrict Q to depend only on the mean usage frequency of G_1 :

$$\mu(X) = \sum_k p_k X_k$$

where an agent of type k uses G_1 a fraction p_k of the time. A specific learning algorithm is to sample from a Bernoulli distribution with parameter $\mu(X)$, then use Bayesian inference to estimate $\mu(x)$ in terms of a beta distribution. Then $Q(X)$ would be a function of the inferred beta distribution. The PI will leave Q as general as possible in the following sub-projects, and restrict to $Q(X) = q(\mu(X))$ as needed.

3.3.2. *Rigorous derivation of the limiting process.* The above derivation for (14) is informal. The PI proposes to rigorously derive the continuous dynamics of X as the limit of the discrete dynamics of Y using theorems and techniques from [16], to prove that solutions have a density, and to solve the Kolmogorov PDEs for the density. The long term behavior of the density and the stability of any steady states are of primary interest.

3.3.3. *Confinement.* The components of X are supposed to be positive and sum to 1, so each $X(t)$ is contained in a simplex. With stochastic processes, there is always the risk that the noise might kick the population outside of its bounds. The PI conjectures that there are conditions on X that guarantee that the process is almost surely properly bounded, to be proved using speed measures and Feller theory, as in [16].

3.3.4. *Hovering behavior.* Assuming that X starts at a stable fixed point \bar{x} (that is, $Q(\bar{x}) = \bar{x}$ and \bar{x} is asymptotically stable under (15)), what is the distribution of the time $X(t)$ spends near \bar{x} ? More mathematically, if we define the hitting time for a ball of radius ε to be

$$(16) \quad \tau(\varepsilon, \bar{x}) = \inf_t \{|X(t) - \bar{x}| = \varepsilon\}$$

then what is the distribution of $\tau(\varepsilon, \bar{x})$? For example the mean of τ might be a power law in ε . The PI plans to address these questions by interpreting (14) as a stochastic perturbation of (6) and using techniques as in [21].

3.3.5. *Transition behavior.* Given $X(0) = x_0$, two fixed points \bar{x}_1 and \bar{x}_2 , and ε_1 and ε_2 , what is the probability that the population goes to \bar{x}_1 first? That is,

$$(17) \quad \mathbb{P}(\tau(\varepsilon_1, \bar{x}_1) < \tau(\varepsilon_2, \bar{x}_2)) = ?$$

Also, on what time scale does X go to \bar{x}_1 ? The PI plans to address these questions by interpreting (14) as a stochastic perturbation of (6) and using techniques as in [21].

3.3.6. *Generalization to continuous types.* The PI proposes to replace the finite number of types of individuals with a continuum. Thus, the population state $X(t, z)$ will be a density on $z \in [0, 1]$ for the part of the population that uses G_1 with usage frequency z at time t . That formulation will require an infinite dimensional stochastic differential equation analogous to (6), so some care must be taken in deriving it and assigning the proper interpretation to infinite dimensional Brownian motion.

3.3.7. *Application to historical events.* An important question in understanding a language change is to determine whether the change should be attributed to chance or to an external force. For example, a change in word order might be driven by a change in pronunciation or a social event such as contact with a foreign language. The PI proposes to connect theoretical results concerning hovering time and transition behavior to manuscript data concerning word order changes in Middle English, and make quantitative statements about how likely it is for such changes to happen spontaneously.

3.4. **Generalizations.** The PI proposes to add the following generalizations to the proposed population models after the fundamental questions have been answered. These model factors that are known or suspected to have an influence on language change. In each case, the generalization should improve the model's ability to represent qualitative and quantitative observations about language change, and lead to conclusions about how strongly the added features influence language change.

3.4.1. *Literacy.* The PI proposes to incorporate literacy into these models to find ways to measure the literary inertia present in the manuscript record, and the influence of literary tradition on the spoken language. For example, the learning function Q from (6) and (14) could incorporate literacy through dependence on the population's history averaged against a distribution $d\rho(s)$ for the level of influence of the speech or writing from time s ago on the present:

$$(18) \quad u'(t) = \beta \left(Q \left(\int_0^\infty u(t-s) d\rho(s) \right) - u(t) \right)$$

$$(19) \quad dX(t) = \beta \left(Q \left(\int_0^\infty X(t-s) d\rho(s) \right) - X(t) \right) + \sqrt{X(t)(1-X(t))} dW(t)$$

The PI proposes to prove that solutions to (18) and (19) exist for all time under appropriate hypotheses, and to explore methods, such as the projected dynamics from Section 3.1.3, for understanding their behavior.

Delay dynamics are notoriously difficult [2, 26, 32, 43, 51, 76]. Equations (18) and (19) are especially challenging as learning depends on the entire past history. The PI expects analysis of population learning dynamics to contribute important and fundamental results to the theory of non-linear SDEs with delay.

These delay models will allow linguists to explore the influence of literacy on language and to better understand the written record of language change. For example, some linguistic studies of Old English suggest that due to a strong literary tradition, the written form of late Old English might have been quite different from the spoken form. That is, the time course of syntactic changes observed in manuscripts at the end of the Old English period is distorted because the strong literary standard delayed the appearance of spoken innovations in the written record. Middle English manuscripts, written after the social upheaval of the Norman conquest, appear to be closer to the spoken language and the literary standard is not as strong.

3.4.2. *Social and spatial structure.* As described in [46], bifurcations occur in deterministic models as mixing between compartments increases. The PI proposes to add compartments to the models in Sections 3.1, 3.2, and 3.3, and generalize the bifurcation results from [46]. Linguistic research shows that language varies across social and spatial boundaries, suggesting a multi-compartment model where each compartment is governed by well-mixed dynamics and a mixing process moves individuals from one compartment to another. The compartments can be interpreted as spatial or social communities. Bifurcations that arise from changing the mixing process can model dialect creation and extinction due to changes in social structure.

A further generalization for the Markov chain simulation would be to join this linguistic model to a clustering process, such as small world graphs [69, 74], preferential attachment [15] or a Chinese restaurant process [4], thereby making the compartment structure variable. In preferential attachment, a network grows such that new vertices attach to existing vertices with preference for those that already have more edges. In a Chinese restaurant process, individuals arrive one at a time and either sit at an empty table or join an existing table with preference for tables with more people.

3.4.3. *Prediction-driven instability.* During a language change, old constructions become associated with the older generation of speakers. It seems that children sometimes accelerate the change by observing that age association, and preferring constructions favored by younger speakers. Thus they are predicting where the population as a whole is heading, and leaping closer to the conclusion. There are several possible routes for modeling this effect. One is to extend any of the population models to include age structure and learning algorithms that prefer to match the speech of younger adults. Another is to allow the learning algorithm to depend on the first derivative of the population state, thereby giving it an infinitesimal peek into the future. Prediction-driven instability has been

studied in the economics literature, in explaining bubbles for example [65]. The PI proposes to adapt this material to language modeling.

3.5. The subset problem in language acquisition. Mathematics has contributed to linguistics through the framework of formal languages, abstract machines, and the theory of computability. However, Gold’s *inductive inference* (II) approach to learning formal languages [24], which is frequently used in modeling, turns out to be somewhat ineffective as a model of language learning. In particular, a problem that continues to puzzle linguists and mathematical modelers is the so-called *subset problem*, which the PI proposes to address through statistical methods more like Bayesian inference and the *probably almost correct* (PAC) framework [72].

It is generally accepted that children only use positive evidence during language acquisition, that is, they are given a sample of grammatical sentences and derive a grammar from that. If there exist two possible grammars G_1 and G_2 such that the sentences that G_2 generates are a superset of those that G_1 generates, then a child learning from a G_1 environment can never receive evidence that G_2 is wrong. For example, G_2 might have freer word order than G_1 . For G_1 to be learnable, UG must include some rule that causes children to assume that the subset language is correct until convinced otherwise by sentences that it cannot produce. This is called the *subset principle*. The subset problem is that there are instances where children seem to begin with a superset language and somehow narrow it down to a subset language, which cannot be accomplished from positive evidence alone.

3.5.1. Implicit negative evidence from statistical patterns. Several linguists have suggested that children may use statistical patterns in the sentences they hear. For instance, a child who hears a single word order consistently concludes that no other word order is grammatical despite the fact that there is no evidence in the II sense that other word orders are unavailable. Child language acquisition is robust enough that the occasional deviation can be memorized as a special case or discarded as noise. The PI proposes to simulate the acquisition of grammar in the presence of subset languages using Bayesian inference. Children will have a set of n possible grammars and attempt to infer the usage frequency of each based on sample input. They must simultaneously infer the frequency of m meaning types, such as negation, whether the verb takes one, two, or three arguments, and so on. The mathematical problem is to infer a distribution on a product space of two simplices $\mathcal{S}^n \times \mathcal{S}^m$ representing the probabilities of using each of the n grammars and m sentence types, given some sentence data. Assuming that there is no degeneracy in the problem (such as two identical grammars) there should be no problem inferring the correct distribution, but the question is how much data and computational resources does it take. If the computation can be done with a reasonable amount of memory and limited data, it provides proof of concept that statistical patterns can provide implicit negative evidence, and that this might be part of the resolution of the subset problem.

The first stage will be to apply the statistical learner to synthetic data from the Markov chain simulation from Section 3.2. The next stages will be to apply it to actual data as described in the next two sub-projects.

3.5.2. Inferring per-manuscript V2 rates in Middle English. As described in Section 2.2, Old and Middle English had SVO+V2 word order that eventually gave way to the modern SVO word order. Middle English manuscripts vary considerably in the frequencies at which they use different sentence types, suggesting that both grammars are present in the minds of all speakers at a variety of usage frequencies. The PI proposes to apply the model of implicit negative evidence from statistical patterns to infer the usage frequencies of SVO and SVO+V2 grammars used for each manuscript in the Pennsylvania Parsed Corpus of Middle English (PPCME), and the frequencies of using the subject as fronted topic, an object as topic, an adverb as topic, and so on. It should be possible to map the overall shift from SVO+V2 to SVO over the course of the Middle English period. The

raw manuscript data is sparse and noisy, so naive statistical methods are unlikely to be of much use. Thus, a more elaborate model such as the one described here is required.

The connection to the subset problem is that SVO is a subset of SVO+V2 (or almost a subset, depending on the linguist’s preferred formalism and whether certain additional information is available to the hearer). Many simulations [6, 23, 53] that use II learning have the unrealistic property that all populations develop a V2 grammar and can never change back because they have no way to change to a subset language. Thus, this statistical model is a vital step in producing a realistic learning function than can connect the abstract population models from Sections 3.1, 3.2, and 3.3 to linguistic applications.

3.5.3. *Work with Misha Becker on raising and control verbs.* The PI has begun collaborating with Misha Becker, a linguist at UNC Chapel Hill. Her work [3] focuses on child acquisition of raising and control verbs.⁶ The subset principle suggests that children hearing an unfamiliar verb, as in ‘John gorphs to eat sushi,’ should initially assume that it is a control verb, as these have the most limited use. However, Becker’s work with child language acquisition suggests that young children assume that control verbs can be used with a raising interpretation, while older children and adults do not allow this interpretation. She hypothesizes that the animacy of the subject, the activeness of the infinitive, and statistical patterns may be sufficient explanation for this behavior and for the acquisition of raising and control verbs. The PI proposes to assist her by adapting the model of implicit negative evidence from statistical patterns to her specific problem. The model will be tested on data from the Brown Corpus of parsed modern English text and the CHILDES corpus of adult speech to children, to see if there are strong enough statistical patterns in common usage to support this hypothesis.

4. BROADER IMPACT

4.1. **Applications in linguistics.** The primary broad impact of this research is to develop mathematical tools for linguists to use in understanding the human language faculty. With these tools, linguists can investigate questions that are difficult to address without mathematics. Mathematics has turned out to be extremely useful in biology, and the PI believes the same will prove to be true of the social, cognitive, and behavioral sciences.

Since this research is inherently interdisciplinary, the PI has contacted several linguists and is currently consulting with Anthony Kroch at the University of Pennsylvania (concerning Old and Middle English manuscripts, and corpus data), Misha Becker at UNC Chapel Hill (concerning the subset problem), and Lisa Pearl, a linguistics graduate student at the University of Maryland at College Park (concerning models and simulations of Old English word order).

The proposed mathematical research can be applied to the following questions in linguistics.

- Why do languages change and diversify? Do the fluctuations inherent in normal usage and child language acquisition suffice to explain changes and diversity or must an extralinguistic force be invoked? Specific cases include word order changes in Middle English, which are thought to be the result of contact between regional dialects and Old Norse.
- What dynamics govern language change and diversification? How do literacy, spatial structure, and social structure affect the speed of a language change? Specific cases to be explored include

⁶Raising verbs raise their subject from an embedded infinitive, as in ‘John seems to be happy,’ where John is originally the subject of ‘to be,’ and only appears as the syntactic subject of ‘seem’ because sentences in English must have an overt subject. Raising verbs can be used in constructions where the subject is not raised, such as ‘It seems that John is happy.’ Control verbs look similar, but they select their subject directly and the infinitive has an implicit pronoun for a subject, as in ‘John wants to be happy.’ Control verbs cannot be used in constructions such as ‘*It wants that John is happy.’ Some verbs like ‘begin’ can be used both ways. Thus a sample sentence using a given verb with an expletive subject indicates that it can be used in raising constructions, but it does not rule out the possibility that it might also be used with control constructions.

Middle English word order and a change in the case system in Icelandic that seems to be driven by children reinforcing one another's acquisition mistakes in school.

- Can statistical patterns in the primary linguistic data provide the implicit negative evidence needed to resolve the subset problem? That is, is it possible for a child to hypothesize that a construction is grammatical, and reject that hypothesis on the basis of statistical evidence, thereby moving to a grammar that is a subset of the original hypothesis?

4.2. Broad dissemination. The PI posts papers on a personal homepage on the Duke mathematics web site, and at a site hosted by the University of Illinois at Urbana-Champaign for the on-line message group `langev@yahoo.com` for language evolution and computational biology. The PI has started a focused message group `mathling@googlegroups.com` for himself and his collaborators.

The PI has presented at several non-mathematical conferences, including the 2004 International Conference on English Historical Linguistics, the 2005 meeting of the Association for Computational Linguistics, the spring colloquium for the linguistics department of the University of North Carolina at Chapel Hill, and the computational linguistics colloquium at the University of Maryland at College Park. Linguists have generally been open to hearing about this research, even though it is unusual by their standards.

Michael Lavine, editor of *Chance*, asked the PI to submit an article on the Markov chain population model. *Chance* is a non-research journal for applications of probability and statistics, intended for a broad, scientifically literate audience.

4.3. Education. During the summer of 2004, the PI supervised an undergraduate research project at Duke, in which the PI and Adam Chandler constructed a model of how a language change spreads, using work by Herold [29] on a phonological change spreading through English as spoken in Pennsylvania. This research project was part of Duke's successful PRUV program (Practical Research for Undergraduates with VIGRE) funded by the mathematics department's VIGRE grant. The application of mathematical models to linguistics is a relatively new and open field, so it motivates many projects accessible to undergraduate mathematics majors.

During the spring 2005 semester, the PI taught the department's mathematical modeling course, with an emphasis on modeling and linguistics. It attracted 12 undergraduates from a variety of majors, including mathematics (9, some with second majors in computer science), computer science (1), economics (1), and political science (1), plus an auditor majoring in Russian and linguistics. Assignments included a guided lab based on [39], and open-ended projects requiring papers and presentations. Guest speakers included Michael Reed from the Duke mathematics department, who spoke on models of the neurophysiology of bat hearing, and Don Burdick from Metametrics, a developer of the Lexile text metric.

Also during the spring 2005 semester, the PI supervised undergraduate Ashleigh Price in an independent study of stochastic calculus leading to the Black-Scholes formula for option pricing. Price wished to pursue a career in finance, and asked the PI to teach this material.

The PI mentors the department's teams for the mathematical contest in modeling (MCM), an annual competition sponsored by COMAP. A Duke team won an Outstanding rating for their paper on the toll booth problem in 2005. Team members Adam Chandler and Pradeep Baliga are continuing with the project as an independent study under the PI for graduation with distinction.

REFERENCES

- [1] K. Aoki and Marcus W. Feldman. Toward a theory for the evolution of cultural communication: Coevolution of signal transmission and reception. *Proceedings of the National Academy of Sciences, USA*, 84:7164–7168, 1987.
- [2] Yuri Bakhtin and Jonathan C. Mattingly. Stationary solutions of stochastic differential equation with memory and stochastic partial differential equations. To appear; online at arXiv:math.PR/0509166 v1.
- [3] Misha Becker. Learning verbs without arguments: The problem of raising verbs. *Journal of Psycholinguistic Research*, 34(2):173–199, March 2005.
- [4] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 conference*. MIT Press, 2004.
- [5] Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors. *Probabilistic Linguistics*. MIT Press, Cambridge, MA, 2003.
- [6] E. J. Briscoe. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296, 2000.
- [7] E. J. Briscoe, editor. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press, 2002.
- [8] E. J. Briscoe. Grammatical acquisition and linguistic selection. In *Linguistic Evolution through Language Acquisition: Formal and Computational Models* Briscoe [7]. URL <http://www.cl.cam.ac.uk/users/ejb/creo-evol.ps.gz>.
- [9] Aangelo Cangelosi and Dominico Parisi, editors. *Simulating the Evolution of Language*. Springer-Verlag, 2001.
- [10] Angelo Cangelosi and Dominico Parisi. The emergence of a “language” in an evolving population of neural networks. *Connection Science*, 10(2):83–97, 1998.
- [11] Luigi Luca Cavalli-Sforza and Marcus W. Feldman. *Cultural transmission and evolution: a quantitative approach*. Princeton University Press, Princeton, NJ, 1981.
- [12] Brady Z. Clark. *A Stochastic Optimality Theory Approach to Syntactic Change*. PhD thesis, Stanford University, 2004.
- [13] Felipe Cucker, Steve Smale, and Ding-Xuan Zhou. Modeling language evolution. *Foundations of Computational Mathematics*, 4(3):315–343, July 2004.
- [14] J. Dieudonné. *Foundations of Modern Analysis*. Academic Press, New York, 1960.
- [15] Konstantinos Drakakis. *A detailed mathematical study of several aspects of the Internet*. PhD thesis, Princeton University, 2003.
- [16] Richard Durrett. *Stochastic Calculus: A Practical Introduction*. CRC Press, New York, 1996.
- [17] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York, 1986.
- [18] Ramon Ferrer i Cancho and Richard V. Solé. Two regimes in the frequency of words and the origin of complex lexicons: Zipf’s law revisited. *Journal of Quantitative Linguistics*, 8:165–173, 2001.
- [19] Ramon Ferrer i Cancho and Richard V. Solé. The small world of human language. *Proceedings of the Royal Society of London B*, 268:2261–2266, 2001.
- [20] Olga Fischer, Ans van Kemenade, Willem Koopman, and Wim van der Wurff. *The Syntax of Early English*. Cambridge University Press, 2000.
- [21] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren der mathematischen Wissenschaften*. Springer Verlag, New York, 1984.
- [22] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.

- [23] E. Gibson and Kenneth Wexler. Triggers. *Linguistic Inquiry*, 25:407–454, 1994.
- [24] E. Mark Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.
- [25] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, 1990.
- [26] Jack K. Hale and Sjoerd M. Verduyn Lunel. *Introduction to Functional Differential Equations*, volume 99 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1993.
- [27] Marc D. Hauser. *The Evolution of Communication*. Harvard University Press, Cambridge, MA, 1996.
- [28] 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.
- [29] Ruth Herold. Solving the actuation problem: Merger and immigration in eastern pennsylvania. *Language Variation and Change*, 9(2):165–189, 1997.
- [30] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.
- [31] James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors. *Approaches to the Evolution of Language*. Cambridge University Press, 1998.
- [32] A. V. Kim. *Functional Differential Equations: Application of i -smooth calculus*. Kluwer Academic Publishers, Boston, 1999.
- [33] 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.
- [34] Natalia L. Komarova and Martin A. Nowak. The evolutionary dynamics of the lexical matrix. *Bulletin of Mathematical Biology*, 63(3):451–485, 2001.
- [35] Natalia L. Komarova and Martin A. Nowak. Natural selection of the critical period for language acquisition. *Proceedings of the Royal Society of London, Series B*, 268:1189–1196, 2001.
- [36] Natalia L. Komarova, Partha Niyogi, and Martin A. Nowak. The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–59, 2001.
- [37] Ekkehard König and Johan van der Auwera, editors. *The Germanic Languages*. Routledge, New York, 1994.
- [38] David C. Krakauer. Kin imitation for a private sign system. *Journal of Theoretical Biology*, 213:145–157, 2001.
- [39] Anthony Kroch. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244, 1989.
- [40] Anthony Kroch, Ann Taylor, and Donald Ringe. The middle english verb-second constraint: A case study in language contact and language change. In Susan Herring, Pieter van Reenen, and Lene Schøsler, editors, *Textual Parameters in Older Languages*, pages 353–391, Philadelphia, 2000. John Benjamins Publishing Company.
- [41] David Lightfoot. *The Development of Language: Acquisition, Changes and Evolution*. Blackwell Publishers, 1999.
- [42] David Lightfoot. *How to Set Parameters: Arguments from Language Change*. MIT Press, Cambridge, MA, 1991.
- [43] Michael C. Mackey and Irina G. Nechaeva. Solution moment stability in stochastic differential delay equations. *Physical Review E*, 52(4), October 1995.
- [44] W. Garrett Mitchener. Bifurcation analysis of the fully symmetric language dynamical equation. *Journal of Mathematical Biology*, 46:265–285, March 2003.
- [45] W. Garrett Mitchener. Simulating language change in the presence of non-idealized speech. In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*, pages 10–19. Association for Computational Linguistics, 2005.

- [46] W. Garrett Mitchener. A mathematical model of the loss of verb-second in Middle English. In *Proceedings of the 13th International Conference on English Historical Linguistics*, 2004. To appear.
- [47] W. Garrett Mitchener. Game dynamics with learning and evolution of universal grammar. *Submitted*, 2005.
- [48] W. Garrett Mitchener. *A Mathematical Model of Human Languages: The interaction of game dynamics and learning processes*. PhD thesis, Princeton University, 2003.
- [49] W. Garrett Mitchener and Martin A. Nowak. Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93, January 2003.
- [50] 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.
- [51] S-E A Mohammed. *Stochastic functional differential equations*, volume 99 of *Research Notes in Mathematics*. Pitman Publishing Inc., Marshfield, MA, 1984.
- [52] Partha Niyogi. *The Informational Complexity of Learning*. Kluwer Academic Publishers, Boston, 1998.
- [53] Partha Niyogi and Robert C. Berwick. A language learning model for finite parameter spaces. *Cognition*, 61:161–193, 1996.
- [54] Partha Niyogi and Robert C. Berwick. Evolutionary consequences of language learning. *Linguistics and Philosophy*, 20:697–719, 1997.
- [55] 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>.
- [56] Martin A. Nowak and Natalia L. Komarova. Towards an evolutionary theory of language. *Trends in Cognitive Sciences*, 5(7):288–295, July 2001.
- [57] Martin A. Nowak and David C. Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences, USA*, 96:8028–8033, 1999.
- [58] 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.
- [59] Martin A. Nowak, Joshua Plotkin, and David C. Krakauer. The evolutionary language game. *Journal of Theoretical Biology*, 200:147–162, 1999.
- [60] Martin A. Nowak, Joshua Plotkin, and V. A. A. Jansen. Evolution of syntactic communication. *Nature*, 404(6777):495–498, 2000.
- [61] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291(5501):114–118, 2001.
- [62] Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617, June 2002.
- [63] Daniel Osherson, Michael Stob, and Scott Weinstein. *Systems That Learn*. MIT Press, Cambridge, MA, 1986.
- [64] Joshua Plotkin and Martin A. Nowak. Language evolution and information theory. *Journal of Theoretical Biology*, 205:147–159, 2000.
- [65] David P. Porter and Vernon L. Smith. Stock market bubbles in the laboratory. *Journal of Behavioral Finance*, 4(1):7–20, 2003.
- [66] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(2):223–252, 1996.
- [67] William Gregory Sakas. A word-order database for testing computational models of language acquisition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 415–422, July 2003.
- [68] Luc Steels. Self-organizing vocabularies. In *Proceedings of the Artificial Life Conference*, volume 5. MIT Press, 1996.

- [69] Steven H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.
- [70] Bruce Tesar and Paul Smolensky. *Learnability in Optimality Theory*. MIT Press, 2000.
- [71] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:436–445, 1984.
- [72] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
- [73] Tandy Warnow. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences, USA*, 94:6585–6590, June 1997.
- [74] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, June 1998.
- [75] Lydia White. *Universal Grammar and Second Language Acquisition*, volume 1 of *Language Acquisition & Language Disorders*. John Benjamins Publishing Company, 1989.
- [76] Jianhong Wu. *Theory and Applications of Partial Functional Differential Equations*, volume 119 of *Applied Mathematical Sciences*. Springer-Verlag, 1996.
- [77] Charles D. Yang. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, 2002.
- [78] Eberhard Zeidler. *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*. Springer-Verlag, New York, 1986.