Noname manuscript No.
(will be inserted by the editor)

A Mathematical model of prediction-driven instability

How social structure can drive language change

W. Garrett Mitchener

the date of receipt and acceptance should be inserted later

Abstract I discuss a stochastic model of language learning and change. During a syntactic change, each speaker makes use of constructions from two different idealized grammars at variable rates. The model incorporates regularization in that speakers have a slight preference for using the dominant idealized grammar. It also includes incrementation: The population is divided into two interacting generations. Children can detect correlations between age and speech. They then predict where the population's language is moving and speak according to that prediction, which represents a social force encouraging children not to sound out-dated. Both regularization and incrementation turn out to be necessary for spontaneous language change to occur on a reasonable time scale and run to completion monotonically. Chance correlation between age and speech may be amplified by these social forces, eventually leading to a syntactic change through prediction-driven instability.

Keywords language variation · language change · incrementation · mathematical model · social structure · prediction-driven instability

1 Introduction

Languages change, and although many measurements of change and statistical models thereof have been studied, the underlying forces and mechanisms are not well understood. In this article, based on a presentation at the Mathematics of Language conference of 2007, I will formulate and discuss a mathematical model of language change driven by the fundamental forces of regularization and incrementation together with natural variation of speech and random fluctuations.

By an *idealized grammar,* I mean a formalism that specifies a spoken form for each meaning, perhaps conditioned on the context of the conversation (Adger 2003; Radford 2004; Tesar and Smolensky 2000). The *probably almost correct* family of model learning algorithms operate on a discrete space of idealized grammars and the learner ultimately chooses a single one for its speech (Briscoe 2000, 2002; Gibson and Wexler 1994; Gold 1967; Kirby 2001; Komarova et al 2001; Mitchener 2003; Mitchener and Nowak 2003,

College of Charleston, Charleston, SC, USA E-mail: MitchenerG@cofc.edu

2004; Mitchener 2007; Niyogi 1998; Niyogi and Berwick 1996, 1997; Nowak et al 2001, 2002; Tesar and Smolensky 2000). However, there is considerable variation present in natural speech that this approach does not capture (Labov 1994, 2001; Kroch 1989). The speech pattern of an individual can be better described by a *stochastic grammar,* that is, a collection of similar idealized grammars, each of which is used to form a fraction of the speaker's utterances. That fraction will be called the *usage rate* of the idealized grammar.

Let us suppose that individuals have the choice between two similar idealized grammars, G_1 and G_2, when forming sentences, and that each individual has particular fixed usage rates, that is he uses G_2 in forming a fraction of spoken sentences, and G_1 in forming the rest. As a specific example, consider the syntax of questions in Late Middle and Early Modern English. We take G_1 to be idealized English grammar with verb-raising syntax, and G_2 to be a similar grammar but with *do*-support:

(1) Know you what time it is? (verb-raising, Middle English, G_1)

(2) Do you know what time it is? (*do*-support, Modern English, G_2)

In a well-studied corpus of late Middle and early Modern English, each manuscript uses a combination of verb-raising syntax and *do*-support (Ellegård 1953; Kroch 1989; Warner 2005). In light of the manuscript and sociolinguistic data, it is clear that language acquisition requires more than selecting a single idealized grammar compatible with the primary linguistic data. Instead, children must learn multiple idealized grammars, plus the usage rates. Since verb-raising and *do*-support both exhibit stability over a certain time scale, we should seek a model of learning within a population that has two stable states, one representing populations that prefer G_1 and a second representing populations that prefer G_2. To represent a language change from G_1 to G_2, the model must be able to switch from one stable state to the other over large time scales, while remaining steady over short time scales.

Chance fluctuations in such variation might be enough to trigger a language change, but they are not sufficient to cause it to run to completion. If children learned and used the population-wide average usage rate perfectly, every mixture of idealized grammars would be marginally stable. The average usage rate might drift at random but will show no definite tendency to drive one variant or the other to extinction. Such stable variation does seem to occur for some grammatical features, such as the choices of object syntax for certain ditransitive verbs in English (Bresnan and Nikitina 2007). For variation that does eventually settle, the population must experience *regularization,* in which children prefer to use one favorite variant of those in use. Psycholinguistic studies show that this is a general property of child language acquisition, and that in contrast, adults are more likely to use all the available variants at approximately the usage rates they hear (Hudson Kam and Newport 2005). Regularization causes populations to tend to extreme states where some variants go extinct. It accounts for the fact that languages have many features that are well described as regular rules plus their few exceptions.

An additional force is required to cause changes to be monotonic and to run to completion within a reasonable time. Historical studies show that language changes typically do not reverse themselves partway through (Kroch 1989; Yang 2002) (but see (Warner 2005) for some evidence to the contrary). This means that learners must be able to identify the idealized grammars present in their population's language, estimate their usage rates, and determine which variants are becoming obsolete. Some sort of collective momentum or memory is required, and the resulting force is known as *incrementation:* children can detect and advance changes in progress. As a specific example, Labov (1994, 2001) discusses observations of vowel shifts that show striking gender and age correlations. Labov concludes that very young females initially match their caregiver's speech, but eventually begin shifting at an approximately constant rate for several years, presumably due to increased contact with speakers outside their immediate family (Labov 2001, Chap. 14). A girl entering the shifting phase somehow identifies the correct direction and amplifies partially completed shifts. It is conceivable that all shifts have an innately specified direction, but Labov concludes that although some directions are more likely than others, no directions of vowel shifting are forbidden (Labov 1994, p. 116). Rather than relying on an innate direction, is plausible that children discover the direction by comparing the speech patterns of people of different ages, for example, peers and caregivers. Under some circumstances, there appear to be social forces driving children to avoid sounding conformist or out-dated; Labov (2001, p. 383) characterizes some leaders of sound changes as distinctly non-conformist in their attitudes and speech.

Based on these observations, I will assume that children can detect age-correlated variation in speech patterns and target their speech to where they predict the population will eventually be. Furthermore, I will assume that this process is not limited to phonetics, and that usage rates of idealized grammars can be detected as well. Incrementation can be simulated in an age-structured population model by including a prediction step in the acquisition process that uses information about speech patterns detected in two generations. Regularization is incorporated by distorting usage rates in favor of the dominant idealized grammar. The result is a mathematical model that can exhibit ongoing spontaneous language changes: random fluctuations generate an accidental correlations between age and speech, which causes children to infer that the population is experiencing a language change and to amplify it. This process, which I will call *prediction-driven instability,* pushes the population away from one stable state toward another.

2 Learning with regularization and incrementation

Initially, we might consider an infinitely large unstructured population, in which children learn from all individuals equally and therefore hear essentially the mean usage rate. The simplest dynamic model with the desired bi-stability is a differential equation for the time-dependent mean usage rate $m(t)$ of G_2 in the population,

$$\dot{m} = q(m) - m \tag{3}$$

where the *learning function* $q(m)$ is the mean usage rate of children learning from a population that uses G_2 with a mean rate m. The usage rate of G_1 is $1-m(t)$, so there is no need for a separate variable for it. The $q(m)$ term represents birth and learning, and the $-m$ term represents death. The term *mean field* refers to the fact that the population's influence on an individual is represented by a single aggregate property, in this case, the mean usage rate of G_2. If q is the identity function, then children learn the exact mean usage rate of the entire population and no change is possible. Every mixture is stable, and there is no tendency to favor one of the idealized grammars at

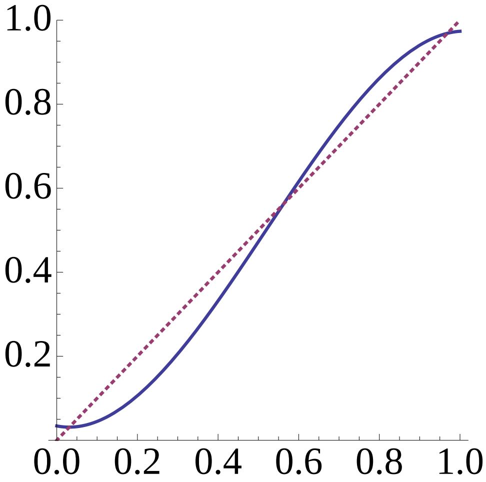

Fig. 1 The regularizing learning function $q(m)$

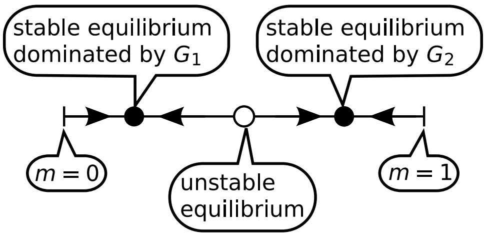
stable equilibrium dominated by G_1
stable equilibrium dominated by G_2
$m = 0$
unstable equilibrium
$m = 1$

Fig. 2 Phase portrait for (3)

the expense the other. If instead q is a sigmoid, for example

$$q(m) = m - 2\left(z - \frac{1}{32}\right)\left(z - \frac{9}{16}\right)\left(z - \frac{31}{32}\right) \tag{4}$$

as shown in Figure 1, then the dynamics of (3) consist of two stable fixed points separated by an unstable fixed point, as in Figure 2. This learning function generates bistability, however, there is no way for a population to spontaneously switch grammars in this model. Even the addition of random noise to (3) can produce spontaneous change only on astronomically long time scales.

To incorporate incrementation, some age information must be available. Assume that there are two age groups, roughly representing youth and their parents, and that children can detect systematic differences in their speech. Assume further that there are social forces leading children to avoid sounding out-dated. Rather than a single mean usage rate m, assume that children hear the younger generation use G_2 at a rate v, and the older generation use a rate w. Based on v and w and any trend those numbers indicate, they predict a rate that their generation should use, and learn based on that predicted target value. Thus, the prediction should be given by a function $r(v, w)$ that satisfies

$$\begin{aligned} &v < w \text{ implies } r(v, w) < v, \text{ and} \\ &v > w \text{ implies } r(v, w) > v. \end{aligned} \tag{5}$$

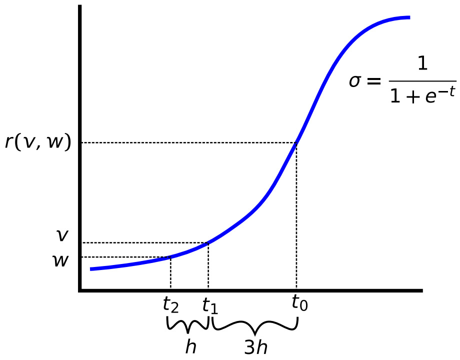

Fig. 3 The prediction function $r(v, w)$

That is, if youth are less likely on average to use G_2 than parents, the prediction is that future generations will use it even less frequently. If youth are more likely on average to use G_2, then future generations will use it even more frequently. We will use a specific prediction function with these properties defined by finding points (t_1, w) and (t_2, v) on the graph of an exponential sigmoid $\sigma(t) = 1/(1 + e^{-t})$. Then $t_0 = t_2 + 3(t_2 - t_0)$ and $r(v, w) = \sigma(t_0)$. See Figure 3.

To add the possibility of spontaneous language change, we formulate the model as a Markov chain rather than a deterministic differential equation. The population consists of N youth and N parents, each of which is one of $K + 1$ types, numbered 0 to K, where type j means that the individual uses G_2 at a rate j/K. These states represent a set of possible stochastic grammars formed by combining two idealized grammars. Examples in this article will use $K = 5$ and $N = 500$.

To represent the population at time t, define $V_j(t)$ to be the number of youth of type j, and define $W_j(t)$ to be the number of parents of type j. We assume that apart from age, children make no distinction among individuals. Thus, they learn essentially from the mean usage rates of the two generations,

$$\begin{aligned} M_V(t) &= \sum_{j=0}^{K} \left(\frac{j}{K}\right) \left(\frac{V_j(t)}{N_V}\right) \\ M_W(t) &= \sum_{j=0}^{K} \left(\frac{j}{K}\right) \left(\frac{W_j(t)}{N_W}\right) \end{aligned} \tag{6}$$

The transition process from $(V(t), W(t))$ to $(V(t + 1), W(t + 1))$ is as follows; see Figure 4. Each adult is examined, and dies with probability p_D. A replacement individual is selected according to the distribution of youth to simulate aging. Similarly, each youth is removed with probability p_D to simulate aging, and is replaced by a new youth whose state is drawn from a discrete probability vector $Q_2(M_V(t), M_W(t))$. The vector $Q_2(v, w)$ is defined to be a binomial distribution over the $K + 1$ possible states with mean $q(r(v, w))$. Thus $Q_2(v, w)$ represents the acquisition process, including regularization and incrementation. The lifetime of an individual follows a geometric distribution with mean of $2/p_D$ time steps, half in each generation. The replacement parameter is set to $p_D = 1/20$ for a mean life span of 40 steps.

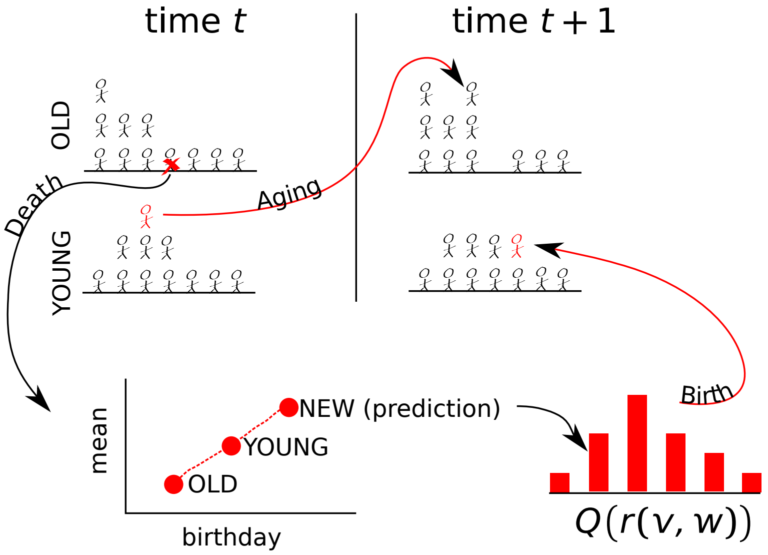

Fig. 4 Diagram of the transition function for the age-structured Markov chain

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 5 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.

To understand geometrically why spontaneous change happens in this model, we approximate the Markov chain by a system of deterministic differential equations governing the mean usage rates v and w of the two generations,

$$\begin{aligned} \dot{v} &= q(r(v,w)) - v \\ \dot{w} &= v - w \end{aligned} \tag{7}$$

The phase space of this dynamical system is a square, and it happens to have two stable equilibria representing populations where both generations are dominated by one grammar or the other. Each such equilibrium has a basin of attraction. Populations in the basin flow toward the equilibrium and settle there. The boundary between the two basins is called the *separatrix,* and in this case, the separatrix passes very close to the stable equilibria. See Figure 6. The population hovers near one equilibrium or the other, but due to random fluctuations, it is possible for the population state to stray across the separatrix, where it will be blown toward the other equilibrium.

3 Change among more than two options

Language changes frequently involve several features of grammar that interact. The age-structured Markov chain model can be extended to any number of interacting

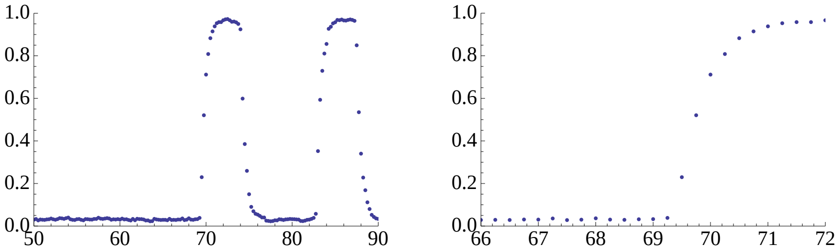

Fig. 5 Trajectory of the mean usage rate $M_V(t)$ of G_2 in the young generation from a sample path of the age-structured Markov chain; left: the path from time 50 to 90, showing several changes between G_1 (low) and G_2 (high); right: the path from time 66 to 72, showing a single grammar change

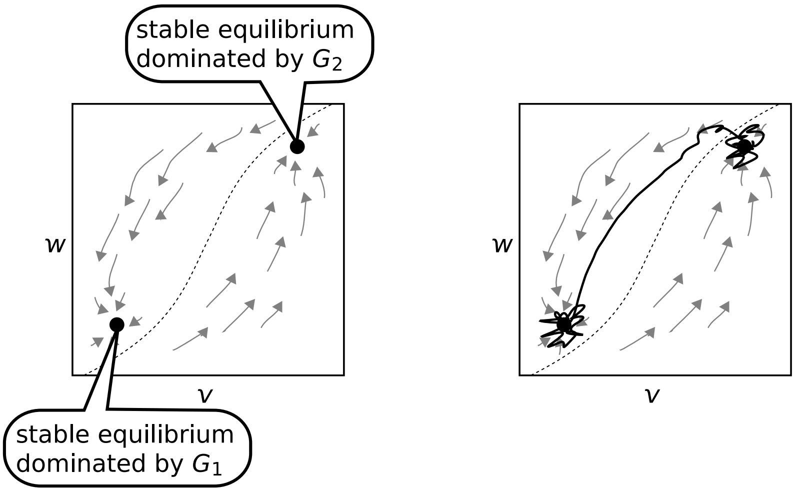
stable equilibrium dominated by G_2
w
v
stable equilibrium dominated by G_1
w
v

Fig. 6 Phase portrait for (7); dots: stable equilibria; dashed curve: the separatrix between their basins of attraction; right: phase portrait with sample trajectory in the presence of random fluctuations

parameters. This section treats the case of two binary parameters, so there are up to four possible grammars. The type of each individual must now be represented as a pair (j_1, j_2) which means that the individual sets parameter 1 with probability j_1/K and sets parameter 2 with probability j_2/K. The element $Q_{j_1,j_2}(m_1, m_2)$ of the joint learning distribution indicates the probability that a child with target mean usage rates m_1 and m_2 for the two parameters grows up to be of type (j_1, j_2). To incorporate prediction, let $M_{V1}(t)$ and $M_{W1}(t)$ be the mean usage rates of constructions with parameter 1 set by youth and parents respectively at time t, and let $M_{V2}(t)$ and $M_{W2}(t)$ be the mean usage rates of constructions with parameter 2 set by youth and parents respectively at time t. Using the prediction function $r(v, w)$, children for time step $t+1$ are drawn from the distribution $Q(r(M_{V1}(t), M_{W1}(t)), r(M_{V2}(t), M_{W2}(t)))$.

For specific results, we must specify $Q(m_1, m_2)$. The easiest case is for four idealized grammars determined by two independent binary parameters. That is, children use the same prediction and learning algorithm as in Section 2 to determine how often to set parameter 1 based on the mean usage rates of parameter 1 among adults and youth, and

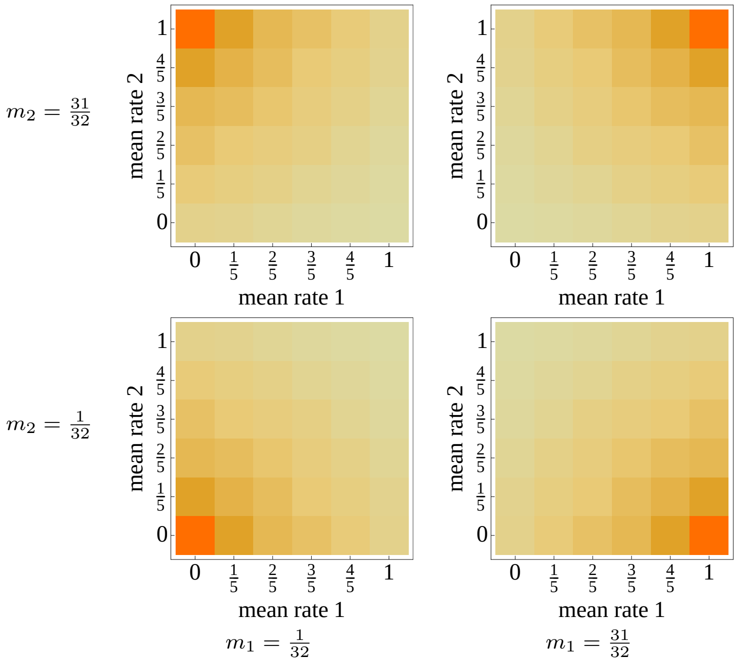

Fig. 7 Learning distribution function $Q(m_1, m_2)$ for two independent parameters, at four values of (m_1, m_2). Darker coloring indicates higher probability.

similarly for parameter 2, with the assumption that the usage rate of each parameter has no bearing on the learning or use of the other parameter.

Figure 7 illustrates a joint learning distribution $Q(m_1, m_2)$ for four different pairs (m_1, m_2). For each (m_1, m_2), $Q(m_1, m_2)$ is a probability distribution over pairs of integers, so it is represented as an array of colored squares. The joint mean learning function is a product, $q(m_1, m_2) = q(m_1)q(m_2)$ with the same q as above. The joint distribution $Q(m_1, m_2)$ is a product of independent binomial distributions with means $q(m_1)$ and $q(m_2)$. Figure 8 shows a sample trajectory of the Markov chain, where the mean speech pattern of the younger generation is plotted as a function of time. The horizontal axis indicates the mean usage rate of idealized grammars with parameter 1 set, and the vertical axis indicates the rate for parameter 2. As time passes, the population spontaneously switches to states dominated by each of the four idealized grammars.

Alternatively, the parameters might be dependent. For example, a sentence cannot use a verb-second construction (as in Old and Middle English) if it does not also use verb raising.[1] Children will not set parameter 2 (verb-second) if parameter 1 (verb raising) is not set. A different learning distribution models this situation: $Q(m_1, m_2)$ is defined such that the marginal distribution for the usage rate of the first parameter is binomial with mean $q(m_1)$, and the conditional distribution for the usage rate of the

[1] It is conceivable that a language could combine verb-second with *do*-support and not have verb raising. Something like this may be happening in modern Dutch (personal communication with a native Dutch speaker). However, in Middle English the verb-second construction was lost before verb raising, so we need not consider *do*-support at this point.

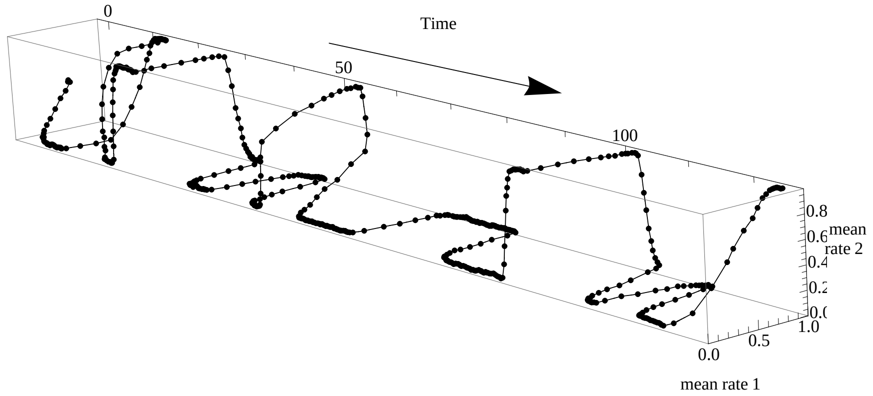

Fig. 8 Sample trajectory for an age structured population learning two independent parameters, mean usage rates of parameters 1 and 2 among the young generation as a function of time

second parameter given the first is binomial with mean $q(m_2)$ if the first parameter is set and zero if it is not. See Figure 9, which shows Q for four different values of (m_1, m_2).

A sample trajectory for two dependent parameters is plotted in Figure 10. The population switches among states that prefer three of the four possible parameter settings, but since the fourth results in an invalid grammar, the population never prefers it. Also observe that the simulated population generally unsets parameter 2 before unsetting parameter 1. This is visible in Figure 10 in that the trajectory tends to go down before it goes to the left. The simulation sometimes unsets both parameters almost simultaneously, going down and left at the same time, but it never tries to unset parameter 1 while parameter 2 is set, which would take it to the left across the top. For reasons that are not clear, when the population prefers a grammar with both parameters unset, it tends to set both parameters simultaneously rather than sequentially. Thus, this example makes clear that dependence among parameters can influence the order in which syntactic changes occur, and that several syntactic changes can occur at the same time and reinforce each other.

4 Discussion and conclusion

We set out to build a mathematical model that can represent spontaneous language change in a population. The model was required to have two semi-stable states, representing populations dominated by one idealized grammar or another. To represent language change on historical time scales, the model was required to hover near one stable state on short time scales, but to spontaneously switch to the other after a reasonable amount of time. Language is represented as a mixture of the idealized grammars to reflect the variability of speech seen in manuscripts and social data.

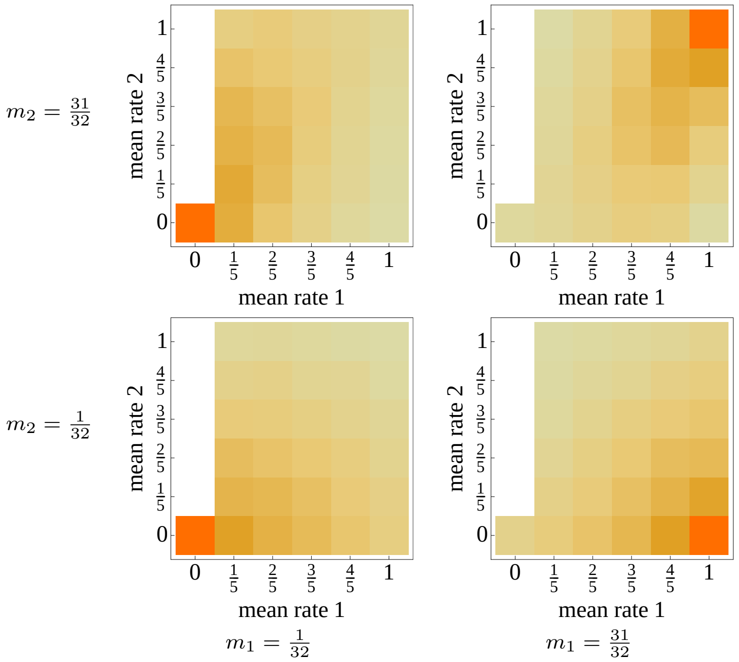

Fig. 9 Learning distribution function $Q(m_1, m_2)$ for two dependent parameters, at four values of (m_1, m_2); darker coloring indicates higher probability

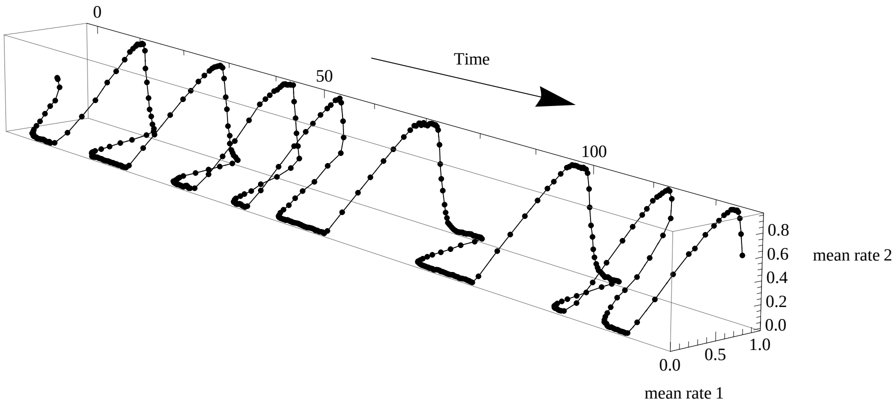

Fig. 10 Sample trajectory for an age structured population learning two dependent parameters, mean usage rates of parameters 1 and 2 among the young generation as a function of time

A Markov chain model that includes age structure, regularization, and incrementation 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. The prediction step in the acquisition process amplifies the correlation, and moves the population away from equilibrium. I therefore coin the term *prediction-driven instability* for this effect. The age-structured Markov chain has reasonable behavior for languages consisting of mixtures of two idealized grammars, and for mixtures of several idealized grammars specified by independent or dependent parameter settings.

More detailed analysis of Markov chain models like the ones described in this article are given in (Mitchener 2009b). A technical analysis of deterministic mean-field population language dynamics with spatial effects is described in (Mitchener 2009a). This research suggests that some social structure is necessary in a model so that it may accurately represent the qualitative features of spontaneous language change. A further project would be to fit the parameters of the age-structured Markov chain to manuscript data and obtain quantitative results as well.

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

References

Adger D (2003) Core Syntax: A minimalist approach. Oxford University Press, Oxford

Bresnan J, Nikitina T (2007) The gradience of the dative alternation. In: Uyechi L, Wee LH (eds) Reality Exploration and Discovery: Pattern Interaction in Language and Life, CSLI Publiscations, Stanford, URL http://www.stanford.edu/~bresnan/publications/index.html

Briscoe EJ (2000) Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. Language 76(2):245–296

Briscoe EJ (2002) Grammatical acquisition and linguistic selection. In: Briscoe EJ (ed) Linguistic Evolution through Language Acquisition: Formal and Computational Models, Cambridge University Press, URL http://www.cl.cam.ac.uk/users/ejb/creo-evol.ps.gz

Ellegård A (1953) The Auxiliary do: The Establishment and Regulation of Its Use in English, Gothenburg Studies in English, vol II. Almqvist and Wiksell

Gibson E, Wexler K (1994) Triggers. Linguistic Inquiry 25:407–454

Gold EM (1967) Language identification in the limit. Information and Control 10:447–474

Hudson Kam CL, Newport EL (2005) Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development 1(2):151–195

Kirby S (2001) 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

Komarova NL, Niyogi P, Nowak MA (2001) The evolutionary dynamics of grammar acquisition. Journal of Theoretical Biology 209(1):43–59

Kroch A (1989) Reflexes of grammar in patterns of language change. Language Variation and Change 1:199–244

Labov W (1994) Principles of Linguistic Change: Internal Factors, vol 1. Blackwell, Cambridge, MA

Labov W (2001) Principles of Linguistic Change: Social Factors, vol 2. Blackwell, Cambridge, MA

Mitchener WG (2003) Bifurcation analysis of the fully symmetric language dynamical equation. Journal of Mathematical Biology 46:265–285, DOI 10.1007/s00285-002-0172-8

Mitchener WG (2007) Game dynamics with learning and evolution of universal grammar. Bulletin of Mathematical Biology 69(3):1093–1118, DOI 10.1007/s11538-006-9165-x

Mitchener WG (2009a) Mean-field and measure-valued differential equation models for language variation and change in a spatially distributed population, submitted

Mitchener WG (2009b) A stochastic model of language change through social structure and prediction-driven instability, submitted

Mitchener WG, Nowak MA (2003) Competitive exclusion and coexistence of universal grammars. Bulletin of Mathematical Biology 65(1):67–93, DOI 10.1006/bulm.2002.0322

Mitchener WG, Nowak MA (2004) Chaos and language. Proceedings of the Royal Society of London, Biological Sciences 271(1540):701–704, DOI 10.1098/rspb.2003.2643

Niyogi P (1998) The Informational Complexity of Learning. Kluwer Academic Publishers, Boston

Niyogi P, Berwick RC (1996) A language learning model for finite parameter spaces. Cognition 61:161–193

Niyogi P, Berwick RC (1997) A dynamical systems model for language change. Complex Systems 11:161–204, URL ftp://publications.ai.mit.edu/ai-publications/1500-1999/AIM-1515.ps.Z

Nowak MA, Komarova NL, Niyogi P (2001) Evolution of universal grammar. Science 291(5501):114–118

Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. Nature 417(6889):611–617

Radford A (2004) Minimalist Syntax: Exploring the structure of English. Cambridge University Press, Cambridge

Tesar B, Smolensky P (2000) Learnability in Optimality Theory. MIT Press

Warner A (2005) Why DO dove: Evidence for register variation in Early Modern English negatives. Language Variation and Change 17:257–280, DOI 10.1017/S0954394505050106

Yang CD (2002) Knowledge and Learning in Natural Language. Oxford University Press, Oxford