Combining data and mathematical models of language change

Morgan Sonderegger
University of Chicago
Chicago, IL, USA.
morgan@cs.uchicago.edu

Partha Niyogi
University of Chicago
Chicago, IL, USA.
niyogi@cs.uchicago.edu

Abstract

English noun/verb (N/V) pairs (*contract*, *cement*) have undergone complex patterns of change between 3 stress patterns for several centuries. We describe a longitudinal dataset of N/V pair pronunciations, leading to a set of properties to be accounted for by any computational model. We analyze the dynamics of 5 dynamical systems models of linguistic populations, each derived from a model of learning by individuals. We compare each model's dynamics to a set of properties observed in the N/V data, and reason about how assumptions about individual learning affect population-level dynamics.

1 Introduction

The fascinating phenomena of language evolution and language change have inspired much work from computational perspectives in recent years. Research in this field considers populations of linguistic agents, and asks how the population dynamics are related to the behavior of individual agents. However, most such work makes little contact with empirical data (de Boer and Zuidema, 2009).[1] As pointed out by Choudhury (2007), most computational work on language change deals with data from cases of change either not at all, or at a relatively high level.[2]

Recent computational work has addressed "real world" data from change in several languages (Mitchener, 2005; Choudhury et al., 2006; Choudhury et al., 2007; Pearl and Weinberg, 2007; Daland et al., 2007; Landsbergen, 2009). In the same spirit, we use data from an ongoing stress shift in English noun/verb (N/V) pairs. Because stress has been listed in dictionaries for several centuries, we are able to trace stress longitudinally and at the level of individual words, and observe dynamics significantly more complicated than in changes previously considered in the computational literature. In §2, we summarize aspects of the dynamics to be accounted for by any computational model of the stress shift. We also discuss proposed sources of these dynamics from the literature, based on experimental work by psychologists and linguists.

In §3–4, we develop models in the mathematical framework of dynamical systems (DS), which over the past 15 years has been used to model the interaction between language learning and language change in a variety of settings (Niyogi and Berwick, 1995; Niyogi and Berwick, 1996; Niyogi, 2006; Komarova et al., 2001; Yang, 2001; Yang, 2002; Mitchener, 2005; Pearl and Weinberg, 2007).

We interpret 6 aspects of the N/V stress dynamics in DS terms; this gives a set of 6 desired properties to which any DS model's dynamics can be compared. We consider 5 models of language learning by individuals, based on the experimental findings relevant to the N/V stress shift, and evaluate the population-level dynamics of the dynamical system model resulting from each against the set of desired properties. We are thus able to reason about which theories of the source of language change — considered as hypotheses about how individuals learn — lead to the population-level patterns observed in change.

2 Data: English N/V pairs

The data considered here are the stress patterns of English homographic, disyllabic noun/verb pairs (Table 1); we refer to these throughout as "N/V pairs". Each of the N and V forms of a pair can have initial ($\acute{\sigma}\sigma$: *cónvict*, n.) or final ($\sigma\acute{\sigma}$: *convíct*,

[1] However, among language evolution researchers there has been significant recent interest in behavioral experiments, using the "iterated learning" paradigm (Griffiths and Kalish, 2007; Kalish et al., 2007; Kirby et al., 2008).

[2] We do not review the literature on computational studies of change due to space constraints; see (Baker, 2008; Wang et al., 2005; Niyogi, 2006) for reviews.

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1019–1029,
Uppsala, Sweden, 11-16 July 2010.

	N	V	
$\{1,1\}$	$\acute{\sigma}\sigma$	$\acute{\sigma}\sigma$	(exile, anchor, fracture)
$\{1,2\}$	$\acute{\sigma}\sigma$	$\sigma\acute{\sigma}$	(consort, protest, refuse)
$\{2,2\}$	$\sigma\acute{\sigma}$	$\sigma\acute{\sigma}$	(cement, police, review)

Table 1: Attested N/V pair stress patterns.

v.) stress. We use the notation {Nstress,Vstress} to denote the stress of an N/V pair, with 1=$\acute{\sigma}\sigma$, 2=$\sigma\acute{\sigma}$. Of the four logically possible stress patterns, all current N/V pairs follow one of the 3 patterns shown in Table 1: {1,1}, {1,2}, {2,2}.[3] No pair follows the fourth possible pattern, {2,1}.

N/V pairs have been undergoing variation and change between these 3 patterns since Middle English (ME, c. 1066-1470), especially change to {1,2}. The vast majority of stress shifts occurred after 1570 (Minkova, 1997), when the first dictionary listing English word stresses was published (Levens, 1570). Many dictionaries from the 17th century on list word stresses, making it possible to trace change in the stress of individual N/V pairs in considerable detail.

2.1 Dynamics

Expanding on dictionary pronunciation data collected by Sherman (1975) for the period 1570–1800, we have collected a corpus of pronunciations of 149 N/V pairs, as listed in 62 British dictionaries, published 1570–2007. Variation and change in N/V pair stress can be visualized by plotting *stress trajectories*: the moving average of N and V stress vs. time for a given pair. Some examples are shown in Fig. 1. The corpus is described in detail in (Sonderegger and Niyogi, 2010); here we summarize the relevant facts to be accounted for in a computational model.[4]

Change Four types of clear-cut change between the three stress patterns are observed:

{2,2}→{1,2} (Fig.1(a)) {1,2}→{1,1}
{1,1}→{1,2} (Fig. 1(b)) {1,2}→{2,2}

However, change to {1,2} is much more common than change from {1,2}; in particular, {2,2}→{1,2} is the most common change. When change occurs, it is often fairly sudden, as in Figs. 1(a), 1(b). Finally, change never occurs *directly* between {1,1} and {2,2}.

Stability Previous work on stress in N/V pairs (Sherman, 1975; Phillips, 1984) has emphasized change, in particular {2,2}→{1,2} (the most common change). However, an important aspect of the diachronic dynamics of N/V pairs is stability: most N/V pairs do *not* show variation or change.

The 149 N/V pairs, used both in our corpus and in previous work, were chosen by Sherman (1975) as those most likely to have undergone change, and thus are not suitable for studying how stable the three attested stress patterns are. In a *random* sample of N/V pairs (not the set of 149) in use over a fixed time period (1700–2007), we find that only 12% have shown variation or change in stress (Sonderegger and Niyogi, 2010). Most pairs maintain the {1,1}, {2,2}, or {1,2} stress pattern for hundreds of years. A model of the diachronic dynamics of N/V pair stress must explain how it can be the case both that some pairs show variation and change, and that many do not.

Variation N/V pair stress patterns show both synchronic and diachronic variation.

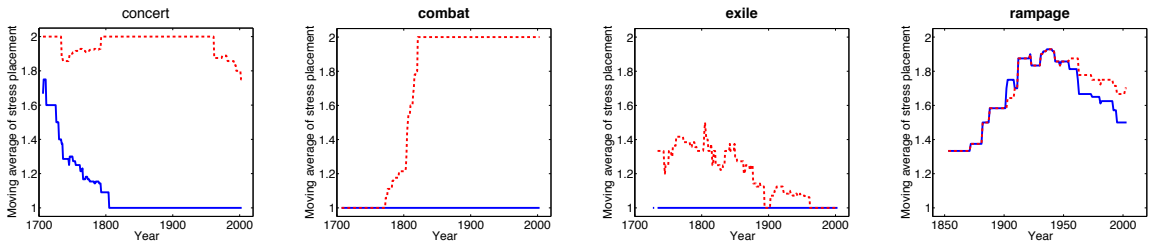
Synchronically, there is variation at the population level in the stress of some N/V pairs at any given time; this is reflected by the inclusion of more than one pronunciation for some N/V pairs in many dictionaries. An important question for modeling is whether there is variation within *individual* speakers. We show in (Sonderegger and Niyogi, 2010) that there is, for present-day American English speakers, using a corpus of radio speech. For several N/V pairs which have currently variable pronunciation, 1/3 of speakers show variation in the stress of the N form. Metrical evidence from poetry suggests that individual variation also existed in the past; the best evidence is for Shakespeare, who shows variation in the stress of over 20 N/V pairs (Kökeritz, 1953).

Diachronically, a relevant question for modeling is whether all variation is short-lived, or whether *stable variation* is possible. A particular type of stable variation is in fact observed relatively often in the corpus: *either* the N or V form stably vary (Fig. 1(c)), but not both at once. Stable variation where both N and V forms vary almost never occurs (Fig. 1(d)).

Frequency dependence Phillips (1984) hypoth-

[3] However, as variation and change in N/V pair stress is ongoing, a few pairs (e.g. *perfume*) currently have variable stress. By "stress", we always mean "primary stress". All present-day pronunciations are for British English, from CELEX (Baayen et al., 1996).

[4] The corpus is available on the first author's home page (currently, `people.cs.uchicago.edu/~morgan`).



(a) concert (b) combat (c) exile (d) rampage

Figure 1: Example N/V pair stress trajectories. Moving averages (60-year window) of stress placement (1=$\acute{\sigma}\sigma$, 2=$\sigma\acute{\sigma}$). Solid lines=nouns, dashed lines=verbs.

esizes that N/V pairs with lower frequencies (summed N+V word frequencies) are more likely to change to {1,2}. Sonderegger (2010) shows that this is the case for the most common change, {2,2}→{1,2}: among N/V pairs which were {2,2} in 1700 and are either {2,2} or {1,2} today, those which have undergone change have significantly lower frequencies, on average, than those which have not. In (Sonderegger and Niyogi, 2010), we give preliminary evidence from real-time frequency trajectories (for <10 N/V pairs) that it is not lower frequency *per se* which triggers change to {1,2}, but *falling* frequency. For example, change in *combat* from {1,1}→{1,2} around 1800 (Fig. 1(b)) coincides with falling word frequency from 1775–present.

2.2 Sources of change

The most salient facts about English N/V pair stress are that (a) change is most often to {1,2} (b) the {2,1} pattern never occurs. We summarize two types of explanation for these facts from the experimental literature, each of which exemplifies a commonly-proposed type of explanation for phonological change. In both cases, there is experimental evidence for biases in present-day English speakers reflecting (a–b). We assume that these biases have been active over the course of the N/V stress shift, and can thus be seen as possible sources of the diachronic dynamics of N/V pairs.[5]

Analogy/Lexicon In historical linguistics, analogical changes are those which make "...related forms more similar to each other in their phonetic (and morphological) structure" (Hock, 1991).[6] Proposed causes for analogical change thus often involve a speaker's production and perception of a form being influenced by similar forms in their lexicon.

The English lexicon shows a broad tendency, which we call *Ross' generalization*, which could be argued to be driving analogical change to {1,2}, and acting against the unobserved stress pattern {2,1}: "primary stress in English nouns is farther to the left than primary stress in English verbs" (Ross, 1973). Change to {1,2} could be seen as motivated by Ross' generalization, and {2,1} made impossible by it.

The argument is lent plausibility by experimental evidence that Ross' Generalization is reflected in production and perception. English listeners strongly prefer the typical stress pattern (N=$\acute{\sigma}\sigma$ or V=$\sigma\acute{\sigma}$) in novel English disyllables (Guion et al., 2003), and process atypical disyllables (N=$\sigma\acute{\sigma}$ or V=$\acute{\sigma}\sigma$) more slowly than typical ones (Arciuli and Cupples, 2003).

Mistransmission An influential line of research holds that many phonological changes are based in asymmetric transmission errors: because of articulatory or perceptual factors, listeners systematically mishear some sound α as β, but rarely mishear β as α.[7] We call such effects *mistransmission*. Asymmetric mistransmission (by individu-

[5] This type of assumption is necessary for any hypothesis about the sources of a completed or ongoing change, based on present-day experimental evidence, and is thus common in the literature. In the case of N/V pairs, it is implicitly made in Kelly's (1988 *et seq*) account, discussed below. Both biases discussed here stem from facts about English (Ross' Generalization; rhythmic context) that we believe have not changed over the time period considered here (≈1600–present), based on general accounts of English historical phonology during this period (Lass, 1992; MacMahon, 1998). We leave more careful verification of this claim to future work.

[6] "Forms" here means any linguistic unit; e.g. sounds, words, or paradigms, such as an N/V pair's stress pattern.

[7] A standard example is final obstruent devoicing, a common change cross-linguistically. There are several articulatory and perceptual reasons why final voiced obstruents could be heard as unvoiced, but no motivation for the reverse process (final unvoiced obstruents heard as voiced) (Blevins, 2006).

als) is argued to be a necessary condition for the change $\alpha \rightarrow \beta$ at the population level, and an explanation for why the change $\alpha \rightarrow \beta$ is common, while the change $\beta \rightarrow \alpha$ is rarely (or never) observed. Mistransmission-based explanations were pioneered by Ohala (1981, *et seq.*), and are the subject of much recent work (reviewed by Hansson, 2008)

For English N/V pairs, M. Kelly and collaborators have shown mistransmission effects which they propose are responsible for the directionality of the most common type of N/V pair stress shifts ($\{1,1\}$, $\{2,2\} \rightarrow \{1,2\}$), based on "rhythmic context" (Kelly, 1988; Kelly and Bock, 1988; Kelly, 1989). Word stress is misperceived more often as initial in "trochaic-biasing" contexts, where the preceding syllable is weak or the following syllable is heavy; and more often as final in analogously "iambic-biasing" contexts. Nouns occur more frequently in trochaic contexts, and verbs more frequently in iambic contexts; there is thus pressure for the V forms of $\{1,1\}$ pairs to be misperceived as $\sigma\acute{\sigma}$, and for the N forms of $\{2,2\}$ pairs to be misperceived as $\acute{\sigma}\sigma$.

3 Modeling preliminaries

We first describe assumptions and notation for models developed below (§4).

Because of the evidence for within-speaker variation in N/V pair stress (§2.1), in all models described below, we assume that what is learned for a given N/V pair are the *probabilities* of using the $\sigma\acute{\sigma}$ form for the N and V forms.

We also make several simplifying assumptions. There are discrete generations G_t, and learners in G_t learn from G_{t-1}. Each example a learner in G_t hears is equally likely to come from any member of G_{t-1}. Each learner receives an identical number of examples, and each generation has infinitely many members.

These are idealizations, adopted here to keep models simple enough to analyze; the effects of relaxing some of these assumptions have been explored by Niyogi (2006) and Sonderegger (2009). The infinite-population assumption in particular makes the dynamics fully deterministic; this rules out the possibility of change due to *drift* (or *sample variation*), where a form disappears from the population because no examples of it are encountered by learners in G_t in the input from G_{t-1}.

Notation For a fixed N/V pair, a learner in G_t hears N_1 examples of the N form, of which k_1^t are $\sigma\acute{\sigma}$ and (N_1-k_1^t) are $\acute{\sigma}\sigma$; N_2 and k_2^t are similarly defined for V examples. Each example is sampled i.i.d. from a random member of G_{t-1}. The N_i are *fixed* (each learner hears the same number of examples), while the k_i^t are random variables (over learners in G_t). Each learner applies an algorithm $\mathcal{A}$ to the N_1+N_2 examples to learn $\hat{\alpha}_t, \hat{\beta}_t \in [0,1]$, the probabilities of *producing* N and V examples as $\sigma\acute{\sigma}$. α_t, β_t are the expectation of $\hat{\alpha}_t$ and $\hat{\beta}_t$ over members of G_t: $\alpha_t = E(\hat{\alpha}_t)$, $\beta_t = E(\hat{\beta}_t)$. $\hat{\alpha}_t$ and $\hat{\beta}_t$ are thus random variables (over learners in G_t), while $\alpha_t, \beta_t \in [0,1]$ are numbers.

Because learners in G_t draw examples at random from members of G_{t-1}, the distributions of $\hat{\alpha}_t$ and $\hat{\beta}_t$ are determined by $(\alpha_{t-1}, \beta_{t-1})$. (α_t, β_t), the expectations of $\hat{\alpha}_t$ and $\hat{\beta}_t$, are thus determined by $(\alpha_{t-1}, \beta_{t-1})$ via an *iterated map* f:

$$f : [0,1]^2 \rightarrow [0,1]^2, \quad f(\alpha_t, \beta_t) = (\alpha_{t+1}, \beta_{t+1}).$$

3.1 Dynamical systems

We develop and analyze models of populations of language learners in the mathematical framework of (discrete) dynamical systems (DS) (Niyogi and Berwick, 1995; Niyogi, 2006). This setting allows us to determine the diachronic, population-level consequences of assumptions about the learning algorithm used by individuals, as well as assumptions about population structure or the input they receive.

Because it is in general impossible to solve a given iterated map as a function of t, the dynamical systems viewpoint is to understand its long-term behavior by finding its fixed points and *bifurcations*: changes in the number and stability of fixed points as system parameters vary.

Briefly, α_* is a *fixed point* (FP) of f if $f(\alpha_*) = \alpha_*$; it is *stable* if $\lim_{t \rightarrow \infty} \alpha_t = \alpha_*$ for α_0 sufficiently near α_*, and *unstable* otherwise; these are also called *stable states* and *unstable states*. Intuitively, α_* is stable iff the system is stable under small perturbations from α_*.[8]

In the context of a linguistic population, change from state α (100% of the population uses $\{1,1\}$) to state β (100% of the population uses $\{1,2\}$) corresponds to a bifurcation, where some system parameter (N) passes a critical value (N_0). For

[8] See (Strogatz, 1994; Hirsch et al., 2004) for introductions to dynamical systems in general, and (Niyogi, 2006) for the type of models considered here.

$N<N_0$, α is stable. For $N>N_0$, α is unstable, and β is stable; this triggers change from α to β.

3.2 DS interpretation of observed dynamics

Below, we describe 5 DS models of linguistic populations. To interpret whether each model has properties consistent with the N/V dataset, we translate the observations about the dynamics of N/V stress made above (§2.1) into DS terms. This gives a list of desired properties against which to evaluate the properties of each model.

1. $^*\{2,1\}$: $\{2,1\}$ is not a stable state.
2. *Stability of* $\{1,1\}$, $\{1,2\}$, $\{2,2\}$: These stress patterns correspond to stable states (for some system parameter values).
3. *Observed stable variation*: Stable states are possible (for some system parameter values) corresponding to variation in the N or V form, but not both.
4. *Sudden change*: Change from one stress pattern to another corresponds to a bifurcation, where the fixed point corresponding to the old stress pattern becomes unstable.
5. *Observed changes*: There are bifurcations corresponding to each of the four observed changes ($\{1,1\} \rightleftharpoons \{1,2\}$, $\{2,2\} \rightleftharpoons \{1,2\}$).
6. *Observed frequency dependence*: Change to $\{1,2\}$ corresponds to a bifurcation in frequency (N), where $\{2,2\}$ or $\{1,1\}$ loses stability as N is decreased.

4 Models

We now describe 5 DS models, each corresponding to a learning algorithm $\mathcal{A}$ used by individual language learners. Each $\mathcal{A}$ leads to an iterated map, $f(\alpha_t, \beta_t) = (\alpha_{t+1}, \beta_{t+1})$, which describes the state of the population of learners over successive generations. We give these evolution equations for each model, then discuss their dynamics, i.e. bifurcation structure. Each model's dynamics are evaluated with respect to the set of desired properties corresponding to patterns observed in the N/V data. Derivations have been mostly omitted for reasons of space, but are given in (Sonderegger, 2009).

The models differ along two dimensions, corresponding to assumptions about the learning algorithm ($\mathcal{A}$): whether or not it is assumed that the stress of examples is possibly mistransmitted (Models 1, 3, 5), and how the N and V probabilities acquired by a given learner are *coupled*. In Model 1 there is no coupling ($\hat{\alpha}_t$ and $\hat{\beta}_t$ learned independently), in Models 2–3 coupling takes the form of a hard constraint corresponding to Ross' generalization, and in Models 4–5 different stress patterns have different prior probabilities.[9]

4.1 Model 1: Mistransmission

Motivated by the evidence for asymmetric misperception of N/V pair stress (§2.2), suppose the stress of N=$\sigma\acute{\sigma}$ and V=$\acute{\sigma}\sigma$ examples may be misperceived (as N=$\acute{\sigma}\sigma$ and V=$\sigma\acute{\sigma}$), with *mistransmission probabilities* p and q.

Learners are assumed to simply probability match: $\hat{\alpha}_t = k_1^t/N_1$, $\hat{\beta}_t = k_2^t/N_2$, where k_1^t is the number of N and V examples *heard* as $\sigma\acute{\sigma}$ (etc.) The probabilities $p_{N,t}$ & $p_{V,t}$ of hearing an N or V example as final stressed at t are then

$$p_{N,t} = \alpha_{t-1}(1-p),\ p_{V,t} = \beta_{t-1} + (1-\beta_{t-1})q \quad (1)$$

k_1^t and k_2^t are binomially-distributed:

$$P_B(k_1^t, k_2^t) \equiv \binom{N_1}{k_1^t} {p_{N,t}}^{k_1^t} (1-p_{N,t})^{N_1-k_1^t} \times \binom{N_2}{k_2^t} {p_{V,t}}^{k_2^t} (1-p_{V,t})^{N_2-k_2^t} \quad (2)$$

α_t and β_t, the probability that a random member of G_t produces N and V examples as $\sigma\acute{\sigma}$, are the ensemble averages of $\hat{\alpha}_t$ and $\hat{\beta}_t$ over all members of G_t. Because we have assumed infinitely many learners per generation, α_t=$E(\hat{\alpha}_t)$ and β_t=$E(\hat{\beta}_t)$. Using (1), and the formula for the expectation of a binomially-distributed random variable:

$$\alpha_t = \alpha_{t-1}(1-p) \quad (3)$$
$$\beta_t = \beta_{t-1} + (1-\beta_{t-1})q \quad (4)$$

these are the *evolution equations* for Model 1. Due to space constraints we do not give the (more lengthy) derivations of the evolution equations in Models 2–5.

Dynamics There is a single, stable fixed point of evolution equations (3–4): $(\alpha_*, \beta_*) = (0, 1)$, corresponding to the stress pattern $\{1,2\}$. This model thus shows none of the desired properties discussed in §3.2, except that $\{1,2\}$ corresponds to a stable state.

[9] The sixth possible model (no coupling, no mistransmission) is a special case of Model 1, resulting in the identity map: $\alpha_{t+1} = \alpha_t$, $\beta_{t+1} = \beta_t$.

4.2 Model 2: Coupling by constraint

Motivated by the evidence for English speakers' productive knowledge of Ross' Generalization (§2.2), we consider a second learning model in which the learner attempts to probability match as above, but the $(\hat{\alpha}_t, \hat{\beta}_t)$ learned must satisfy the constraint that $\sigma\acute{\sigma}$ stress be more probable in the V form than in the N form.

Formally, the learner chooses $(\hat{\alpha}_t, \hat{\beta}_t)$ satisfying a quadratic optimization problem:

$$\text{minimize } [(\alpha - \frac{k_1^t}{N_1})^2 + (\beta - \frac{k_2^t}{N_2})^2] \text{ s.t. } \alpha \leq \beta$$

This corresponds to the following algorithm, $\mathcal{A}_2$:

1. If $\frac{k_1^t}{N_1} < \frac{k_2^t}{N_2}$, set $\hat{\alpha}_t = \frac{k_1^t}{N_1}$, $\hat{\beta}_t = \frac{k_2^t}{N_2}$.
2. Otherwise, set $\hat{\alpha}_t = \hat{\beta}_t = \frac{1}{2}(\frac{k_1^t}{N_1} + \frac{k_2^t}{N_2})$

The resulting evolution equations can be shown to be

$$\alpha_{t+1} = \alpha_t + \frac{A}{2}, \quad \beta_{t+1} = \beta_t - \frac{A}{2} \qquad (5)$$

where $A = \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P_B(k_1^t, k_2^t)(\frac{k_1^t}{N_1} - \frac{k_2^t}{N_2})$.

Dynamics Adding the equations in (5) gives that the (α_t, β_t) trajectories are lines of constant $\alpha_t + \beta_t$ (Fig. 2). All $(0, x)$ and $(x, 1)$ ($x \in [0, 1]$) are stable fixed points. This model thus has stable FPs corresponding to {1,1}, {1,2}, and {2,2}, does not have {2,1} as a stable FP (by construction), and allows for stable variation in exactly one of N or V. It does not have bifurcations, or the observed patterns of change and frequency dependence.

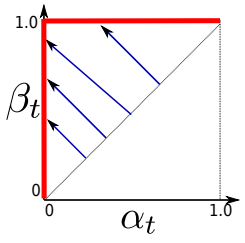


Figure 2: Dynamics of Model 2

4.3 Model 3: Coupling by constraint, with mistransmission

We now assume that each example is subject to mistransmission, as in Model 1; the learner then applies $\mathcal{A}_2$ to the *heard* examples. The evolution equations are thus the same as in (5), but with α_{t-1} and β_{t-1} changed to $p_{N,t}$, $p_{V,t}$ (Eqn. 1).

Dynamics There is a single, stable fixed point, corresponding to stable variation in both N and V. This model thus shows none of the desired properties, except that {2,1} is not a stable FP (by construction).

4.4 Model 4: Coupling by priors

The type of coupling assume in Models 2–3 — a constraint on the relative probability of $\sigma\acute{\sigma}$ stress for N and V forms — has the drawback that there is no way for the rest of the lexicon to affect a pair's N and V stress probabilities: there can be no influence of the stress of other N/V pairs, or in the lexicon as a whole, on the N/V pair being learned. Models 4–5 allow such influence by formalizing a simple intuitive explanation for the lack of $\{2, 1\}$ N/V pairs: learners cannot hypothesize a $\{2, 1\}$ pair because there is no support for this pattern in their lexicons.

We now assume that learners compute the probabilities of each possible N/V pair *stress pattern*, rather than separate probabilities for the N and V forms. We assume that learners keep two sets of probabilities (for $\{1, 1\}$, $\{1, 2\}$, $\{2, 1\}$, $\{2, 2\}$):

1. *Learned probabilities*:
   $\vec{P}$=$(P_{11}, P_{12}, P_{22}, P_{21})$, where

$$P_{11} = \frac{N_1 - k_1^t}{N_1}\frac{N_2 - k_2^t}{N_2}, \quad P_{12} = \frac{N_1 - k_1^t}{N_1}\frac{k_2^t}{N_2}$$
$$P_{22} = \frac{k_1^t}{N_1}\frac{k_2^t}{N_2}, \quad P_{21} = \frac{k_1^t}{N_1}\frac{N_2 - k_2^t}{N_2}$$

2. *Prior probabilities*: $\vec{\lambda} = (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})$, based on the support for each stress pattern in the lexicon.

The learner then produces N forms as follows:

1. Pick a pattern $\{n_1, v_1\}$ according to $\vec{P}$.
2. Pick a pattern $\{n_2, v_2\}$ according to $\vec{\lambda}$
3. Repeat 1–2 until n_1=n_2, then produce N=n_1.

V forms are produced similarly, but checking whether $v_1 = v_2$ at step 3. Learners' production of an N/V pair is thus influenced by both their learning experience (for the particular N/V pair) and by how much support exists in their lexicon for the different stress patterns.

We leave the exact interpretation of the λ_{ij} ambiguous; they could be the percentage of N/V pairs already learned which follow each stress pattern, for example. Motivated by the absence of {2,1} N/V pairs in English, we assume that $\lambda_{21} = 0$.

By following the production algorithm above, the learner's probabilities of producing N and V forms as $\sigma\acute{\sigma}$ are:

$$\hat{\alpha}_t = \tilde{\alpha}(k_1^t, k_2^t) = \frac{\lambda_{22}P_{22}}{\lambda_{11}P_{11} + \lambda_{12}P_{12} + \lambda_{22}P_{22}} \quad (6)$$

$$\hat{\beta}_t = \tilde{\beta}(k_1^t, k_2^t) = \frac{\lambda_{12}P_{12} + \lambda_{22}P_{22}}{\lambda_{11}P_{11} + \lambda_{12}P_{12} + \lambda_{22}P_{22}} \quad (7)$$

Eqns. 6–7 are undefined when (k_1^t, k_2^t)=$(N_1, 0)$; in this case we set $\tilde{\alpha}(N_1, 0) = \lambda_{22}$ and $\tilde{\beta}(N_1, 0) = \lambda_{12} + \lambda_{22}$.

The evolution equations are then

$$\alpha_t = E(\hat{\alpha}_t) = \sum_{k_1=0}^{N_1}\sum_{k_2=0}^{N_2} P_B(k_1, k_2)\tilde{\alpha}(k_1, k_2) \quad (8)$$

$$\beta_t = E(\hat{\beta}_t) = \sum_{k_1=0}^{N_1}\sum_{k_2=0}^{N_2} P_B(k_1, k_2)\tilde{\beta}(k_1, k_2) \quad (9)$$

Dynamics The fixed points of (8–9) are $(0, 0)$, $(0, 1)$, and $(1, 1)$; their stabilities depend on N_1, N_2, and $\vec{\lambda}$. Define

$$R = \left(\frac{N_2}{1 + (N_2 - 1)\frac{\lambda_{12}}{\lambda_{11}}}\right)\left(\frac{N_1}{1 + (N_1 - 1)\frac{\lambda_{12}}{\lambda_{22}}}\right) \quad (10)$$

There are 6 regions of parameter space in which different FPs are stable:

1. $\lambda_{11}, \lambda_{22} < \lambda_{12}$: $(0, 1)$ stable
2. $\lambda_{22} > \lambda_{12}$, $R < 1$: $(0, 1)$, $(1, 1)$ stable
3. $\lambda_{11} < \lambda_{12} < \lambda_{22}$, $R > 1$: $(1, 1)$ stable
4. $\lambda_{11}, \lambda_{22} > \lambda_{12}$: $(0, 0)$, $(1, 1)$ stable
5. $\lambda_{22} < \lambda_{12} < \lambda_{11}$, $R > 1$: $(0, 0)$ stable
6. $\lambda_{11} > \lambda_{12}$, $R < 1$: $(0, 0)$, $(0, 1)$ stable

The parameter space is split into these regimes by three hyperplanes: λ_{11}=λ_{12}, λ_{22}=λ_{12}, and R=1. Given that λ_{21}=0, $\lambda_{12} = 1 - \lambda_{11} - \lambda_{22}$, and the parameter space is 4-dimensional: $(\lambda_{11}, \lambda_{22}, N_1, N_2)$. Fig. 3 shows An example phase diagram in $(\lambda_{11}, \lambda_2)$, with N_1 and N_2 fixed.

The bifurcation structure implies all 6 possible changes between the three FPs ({1,1}⇌{1,2}, {1,2}⇌{2,2}, {2,2}⇌{1,2}). For example, suppose the system is at stable FP $(1, 1)$ (corresponding to {2,2}) in region 2. As λ_{22} is decreased, we move into region 1, $(1, 1)$ becomes unstable, and the system shifts to stable FP $(0, 1)$. This transition corresponds to change from {2,2} to {1,2}.

Note that change to {1,2} entails crossing the hyperplanes λ_{12}=λ_{22} and λ_{12}=λ_{11}. These hyperplanes do not change as N_1 and N_2 vary, so change to {1,2} is *not* frequency-dependent. However, change from {1,2} entails crossing the hyperplane R=1, which does change as N_1 and N_2 vary (Eqn. 10), so change from {1,2} is frequency-dependent. Thus, although there is frequency dependence in this model, it is not as observed in the diachronic data, where change *to* {1,2} is frequency-dependent.

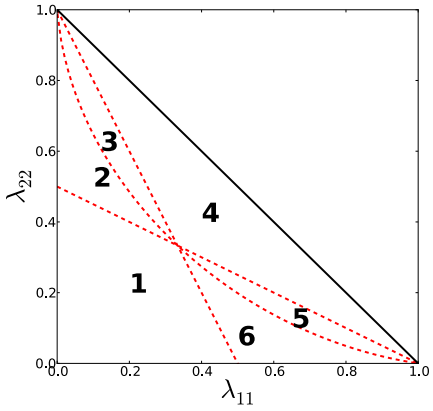


Figure 3: Example phase diagram in $(\lambda_{11}, \lambda_{22})$ for Model 4, with $N_1 = 5$, $N_2 = 10$. Numbers are regions of parameter space (see text).

Finally, no stable variation is possible: in every stable state, all members of the population categorically use a single stress pattern. {2,1} is never a stable FP, by construction.

4.5 Model 5: Coupling by priors, with mistransmission

We now suppose that each example from a learner's data is possibly mistransmitted, as in Model 1; the learner then applies the algorithm from Model 4 to the *heard* examples (instead of using k_1^t, k_2^t) . The evolution equations are thus the same as (8–9), but with α_{t-1} and β_{t-1} changed to $p_{N,t}$, $p_{V,t}$ (Eqn. 1).

Dynamics $(0, 1)$ is always a fixed point. For some regions of parameter space, there can be one fixed point of the form $(\kappa, 1)$, as well as one fixed point of the form $(0, \gamma)$, where $\kappa, \gamma \in (0, 1)$. Define $R' = (1 - p)(1 - q)R$, $\lambda'_{12} = \lambda_{12}$, and

$$\lambda'_{11} = \lambda_{11}(1 - q\frac{N_2}{N_2 - 1}), \quad \lambda'_{22} = \lambda_{22}(1 - p\frac{N_1}{N_1 - 1})$$

There are 6 regions of parameter space corresponding to different stable FPs, identical to the 6 regions in Model 4, with the following substitu-

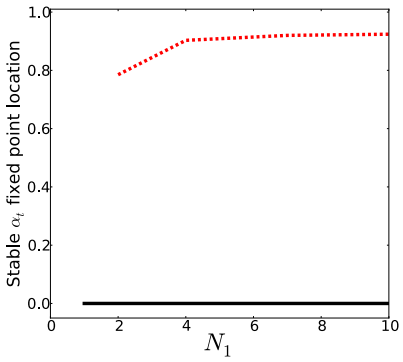


Figure 4: Example of falling N_1 triggering change from $(1,1)$ to $(0,1)$ for Model 5. Dashed line = stable FP of the form $(\gamma,1)$, solid line = stable FP $(0,1)$. For $N_1 > 4$, there is a stable FP near $(1,1)$. For $N_1 < 2$, $(0,1)$ is the only stable FP. $\lambda_{22} = 0.58$, $\lambda_{12} = 0.4$, $N_2 = 10$, $p = q = 0.05$.

tions made: $R \rightarrow R'$, $\lambda_{ij} \rightarrow \lambda'_{ij}$, $(0,0) \rightarrow (0,\kappa)$, $(1,1) \rightarrow (\gamma,1)$.

The parameter space is again split into these regions by three hyperplanes: λ'_{11}=λ'_{12}, λ'_{22}=λ'_{12}, and R'=1. As in Model 4, the bifurcation structure implies all 6 possible changes between the three FPs. However, change to {1,2} entails crossing the hyperplanes λ'_{11}=λ'_{12} and λ'_{2}=λ'_{12}, and is thus now frequency dependent.

In particular, consider a system at a stable FP $(\gamma,1)$, for some N/V pair. This FP becomes unstable if λ'_{22} becomes smaller than λ'_{12}. Assuming that the λ_{ij} are fixed, this occurs only if N_1 falls below a critical value, $N_1^* = (1 - \frac{\lambda_{22}}{\lambda_{12}}(1-p))^{-1}$; the system would then transition to $(0,1)$, the only stable state. By a similar argument, falling frequency can lead to change from $(0,\kappa)$ to $(0,1)$. Falling frequency can thus cause change to {1,2} in this model, as seen in the N/V data; Fig. 4 shows an example.

Unlike in Model 4, stable variation of the type seen in the N/V stress trajectories — one of N or V stably varying, but not both — is possible for some parameter values. $(0,0)$ and $(1,1)$ (corresponding to {1,1} and {2,2}) are technically never possible, but effectively occur for FPs of the form $(\kappa,0)$ and $(\gamma,1)$ when κ or γ are small. {2,1} is never a stable FP, by construction.

This model thus arguably shows all of the desired properties seen in the N/V data.

Property	Model				
	1	2	3	4	5
*{2,1}	✓	✓	✓	✓	✓
{1,1}, {1,2}, {2,2}	✗	✓	✗	✓	✓
Obs. stable variation	✗	✓	✗	✗	✓
Sudden change	✗	✗	✗	✓	✓
Observed changes	✗	✗	✗	✓	✓
Obs. freq. depend.	✗	✗	✗	✗	✓

Table 2: Summary of model properties

4.6 Models summary, observations

Table 2 lists which of Models 1–5 show each of the desired properties (from §3.2), corresponding to aspects of the observed diachronic dynamics of N/V pair stress.

Based on this set of models, we are able to make some observations about the effect of different assumptions about learning by individuals on population-level dynamics. Models including asymmetric mistransmission (1, 3, 5) generally do not lead to stable states in which the entire population uses {1,1} or {2,2}. (In Model 5, stable variation very near {1,1} or {2,2} is possible.) However, {1,1} and {2,2} are diachronically very stable stress patterns, suggesting that at least for this model set, assuming mistransmission in the learner is problematic. Models 2–3, where analogy is implemented as a hard constraint based on Ross' generalization, do not give most desired properties. Models 4–5, where analogy is implemented as prior probabilities over N/V stress patterns, show crucial aspects of the observed dynamics: bifurcations corresponding to the changes observed in the stress data. Model 5 shows change to {1,2} triggered by falling frequency, a pattern observed in the stress data, and an *emergent* property of the model dynamics: this frequency effect is not present in Models 1 or 4, but is present in Model 5, where the learner combines mistransmission (Model 1) with coupling by priors (Model 4).

5 Discussion

We have developed 5 dynamical systems models for a relatively complex diachronic change, found one successful model, and were able to reason about the source of model behavior. Each model describes the diachronic, population-level consequences of assuming a particular learning algorithm for individuals. The algorithms considered

were motivated by different possible sources of change, from linguistics and psychology (§2.2). We discuss novel contributions of this work, and future directions.

The dataset used here shows more complex dynamics, to our knowledge, than in changes previously considered in the computational literature. By using a detailed, longitudinal dataset, we were able to strongly constrain the desired behavior of a computational model, so that the task of model building is not "doomed to success". While all models show *some* patterns observed in the data, only one shows all such properties. We believe detailed datasets are potentially very useful for evaluating and differentiating between proposed computational models of change.

This paper is a first attempt to integrate detailed data with a range of DS models. We have only considered some schematic properties of the dynamics observed in our dataset, and used these to qualitatively compare each model's predictions to the dynamics. Future work should consider the dynamics in more detail, develop more complex models (for example, by relaxing the infinite-population assumption, allowing for stochastic dynamics), and *quantitatively* compare model predictions and observed dynamics.

We were able to reason about how assumptions about individual learning affect population dynamics by analyzing a range of simple, related models. This approach is pursued in more depth in the larger set of models considered in (Sonderegger, 2009). Our use of model comparison contrasts with most recent computational work on change, where a small number (1–2) of very complex models are analyzed, allowing for much more detailed models of language learning and usage than those considered here (e.g. Choudhury et al., 2006; Minett & Wang, 2008; Baxter et al., 2009; Landsbergen, 2009). An advantage of our approach is an enhanced ability to evaluate a range of proposed causes for a particular case of language change.

By using simple models, we were able to consider a range of learning algorithms corresponding to different explanations for the observed diachronic dynamics. What makes this a useful exercise is the fundamentally non-trivial map, illustrated by Models 1–5, between individual learning and population-level dynamics. Although the type of individual learning assumed in each model was chosen with the same patterns of change in mind, and despite the simplicity of the models used, the resulting population-level dynamics differ greatly. This is an important point given that proposed explanations for change (e.g., mistransmission and analogy) operate at the level of individuals, while the phenomena being explained (patterns of change, or particular changes) are aspects of the population-level dynamics.

Acknowledgments

We thank Max Bane, James Kirby, and three anonymous reviewers for helpful comments.

References

J. Arciuli and L. Cupples. 2003. Effects of stress typicality during speeded grammatical classification. *Language and Speech*, 46(4):353–374.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. *CELEX2 (CD-ROM)*. Linguistic Data Consortium, Philadelphia.

A. Baker. 2008. Computational approaches to the study of language change. *Language and Linguistics Compass*, 2(3):289–307.

G.J. Baxter, R.A. Blythe, W. Croft, and A.J. McKane. 2009. Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change*, 21(2):257–296.

J. Blevins. 2006. A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics*, 32(2):117–166.

M. Choudhury, A. Basu, and S. Sarkar. 2006. Multi-agent simulation of emergence of schwa deletion pattern in Hindi. *Journal of Artificial Societies and Social Simulation*, 9(2).

M. Choudhury, V. Jalan, S. Sarkar, and A. Basu. 2007. Evolution, optimization, and language change: The case of Bengali verb inflections. In *Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 65–74.

M. Choudhury. 2007. *Computational Models of Real World Phonological Change*. Ph.D. thesis, Indian Institute of Technology Kharagpur.

R. Daland, A.D. Sims, and J. Pierrehumbert. 2007. Much ado about nothing: A social network model of Russian paradigmatic gaps. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 936–943.

B. de Boer and W. Zuidema. 2009. Models of language evolution: Does the math add up? ILLC Preprint Series PP-2009-49, University of Amsterdam.

T.L. Griffiths and M.L. Kalish. 2007. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480.

S.G. Guion, J.J. Clark, T. Harada, and R.P. Wayland. 2003. Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech*, 46(4):403–427.

G.H. Hansson. 2008. Diachronic explanations of sound patterns. *Language & Linguistics Compass*, 2:859–893.

M.W. Hirsch, S. Smale, and R.L. Devaney. 2004. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, Amsterdam, 2nd edition.

H.H. Hock. 1991. *Principles of Historical Linguistics*. Mouton de Gruyter, Berlin, 2nd edition.

M.L. Kalish, T.L. Griffiths, and S. Lewandowsky. 2007. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, 14(2):288.

M.H. Kelly and J.K. Bock. 1988. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):389–403.

M.H. Kelly. 1988. Rhythmic alternation and lexical stress differences in English. *Cognition*, 30:107–137.

M.H. Kelly. 1989. Rhythm and language change in English. *Journal of Memory & Language*, 28:690–710.

S. Kirby, H. Cornish, and K. Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

S. Klein, M.A. Kuppin, and K.A. Meives. 1969. Monte Carlo simulation of language change in Tikopia & Maori. In *Proceedings of the 1969 Conference on Computational Linguistics*, pages 1–27. ACL.

S. Klein. 1966. Historical change in language using monte carlo techniques. *Mechanical Translation and Computational Linguistics*, 9:67–82.

S. Klein. 1974. Computer simulation of language contact models. In R. Shuy and C-J. Bailey, editors, *Toward Tomorrows Linguistics*, pages 276–290. Georgetown University Press, Washington.

H. Kökeritz. 1953. *Shakespeare's Pronunciation*. Yale University Press, New Haven.

N.L. Komarova, P. Niyogi, and M.A. Nowak. 2001. The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–60.

F. Landsbergen. 2009. *Cultural evolutionary modeling of patterns in language change: exercises in evolutionary linguistics*. Ph.D. thesis, Universiteit Leiden.

R. Lass. 1992. Phonology and morphology. In R.M. Hogg, editor, *The Cambridge History of the English Language*, volume 3: 1476–1776, pages 23–156. Cambridge University Press.

P. Levens. 1570. *Manipulus vocabulorum*. Henrie Bynneman, London.

M. MacMahon. 1998. Phonology. In S. Romaine, editor, *The Cambridge History of the English Language*, volume 4: 1476–1776, pages 373–535. Cambridge University Press.

J.W. Minett and W.S.Y. Wang. 2008. Modelling endangered languages: The effects of bilingualism and social structure. *Lingua*, 118(1):19–45.

D. Minkova. 1997. Constraint ranking in Middle English stress-shifting. *English Language and Linguistics*, 1(1):135–175.

W.G. Mitchener. 2005. Simulating language change in the presence of non-idealized syntax. In *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition*, pages 10–19. ACL.

P. Niyogi and R.C. Berwick. 1995. The logical problem of language change. AI Memo 1516, MIT.

P. Niyogi and R.C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition*, 61(1-2):161–193.

P. Niyogi. 2006. *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge.

J.J. Ohala. 1981. The listener as a source of sound change. In C.S. Masek, R.A. Hendrick, and M.F. Miller, editors, *Papers from the Parasession on Language and Behavior*, pages 178–203. Chicago Linguistic Society, Chicago.

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

B.S. Phillips. 1984. Word frequency and the actuation of sound change. *Language*, 60(2):320–342.

J.R. Ross. 1973. Leftward, ho! In S.R. Anderson and P. Kiparsky, editors, *Festschrift for Morris Halle*, pages 166–173. Holt, Rinehart and Winston, New York.

D. Sherman. 1975. Noun-verb stress alternation: An example of the lexical diffusion of sound change in English. *Linguistics*, 159:43–71.

M. Sonderegger and P. Niyogi. 2010. Variation and change in English noun/verb pair stress: Data, dynamical systems models, and their interaction. Ms. To appear in A.C.L. Yu, editor, *Origins of Sound Patterns: Approaches to Phonologization*. Oxford University Press.

M. Sonderegger. 2009. Dynamical systems models of language variation and change: An application to an English stress shift. Masters paper, Department of Computer Science, University of Chicago.

M. Sonderegger. 2010. Testing for frequency and structural effects in an English stress shift. In *Proceedings of the Berkeley Linguistics Society 36*. To appear.

S. Strogatz. 1994. *Nonlinear Dynamics and Chaos*. Addison-Wesley, Reading, MA.

W.S.Y. Wang, J. Ke, and J.W. Minett. 2005. Computational studies of language evolution. In C. Huang and W. Lenders, editors, *Computational Linguistics and Beyond*, pages 65–108. Institute of Linguistics, Academia Sinica, Taipei.

C. Yang. 2001. Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250.

C. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press.