Dynamical systems models of language variation and change: An application to an English stress shift

Morgan Sonderegger

October 2009

Abstract

Both variation and change are widespread in natural languages. Most variation does not lead to change, but variation between two forms is a necessary condition for change from one to the other to occur. Under what conditions does variation lead to change? We combine two existing approaches to this question: building and making observations from historical datasets, and building mathematical models of linguistic populations. We describe the diachronic dynamics of an English stress shift, based on a historical dataset (1600-2000) of 149 words as listed in 76 dictionaries. This dataset shows several common aspects of variation and change: long-term stability followed by rapid change, multiple stable states, long-term stable variation, and word frequency effects. We translate each of these into into dynamical systems terms, as statements about fixed points and bifurcations. We then describe a range of dynamical systems models of populations of language learners, based on several theories from linguistics and cognitive science on the causes of change. We find the fixed points and bifurcations of these models to determine their dynamics as system parameters are varied. We examine which model properties lead to dynamics consistent with our dataset, and with observations about variation and change generally. One generalization which emerges is that successful models incorporate some form of bias in the data learners receive, as well as bias in the algorithm learners apply to data.

Contents

1 Introduction

One of the most striking facts about language is its heterogeneity across space and time. Linguistic variation, the use of more than one linguistic form for the same linguistic object, is widespread. Language change is constantly occurring in every language. The interaction between variation and change is key to understanding language change because of a simple observation: every linguistic change begins with variation, but not all variation leads to change. What determines whether, in a given linguistic population, a pattern of variation leads to change or not?

In the influential characterization of Weinreich et al. (1968), this is the *actuation problem*: why do linguistic changes begin? We can restate the actuation problem as follows:

1. Why does language change occur at all?
2. Why does it arise from variation?
3. What determines whether a pattern of variation is stable or unstable (leads to change)?

This thesis addresses these questions by combining two approaches to studying the general problem of why language change occurs: first, building and making observations from detailed datasets, in the tradition of sociolinguists and historical linguists; second, building mathematical models of linguistic populations, to model the diachronic consequences of assumptions about the process of language learning (Niyogi and Berwick, 1995; Niyogi, 2006).

Specifically, we describe the diachronic dynamics of an English stress shift, based on a diachronic dataset (1600–2000) which shows both variation and change. This stress shift has several interesting properties which must be accounted for by any computational model. We then build a variety of models of populations of linguistic learners, based on proposals from several theoretical viewpoints on the causes of change. We examine these models' diachronic dynamics, with the goal of determining which model properties lead to dynamics consistent with the stress data, and with observations about variation and change more generally.

1.1 Previous work, Current goals

Computational studies of language change have mushroomed over the past 15 years. A recent review (Baker, 2008b) lists over 50 papers, mostly written in this time period, and along with (Niyogi, 2006) provides a useful review of this literature. This thesis builds on previous work, but also addresses a new set of questions and methods.

Our main contribution is to connect three approaches to the study of language change, practiced largely by different communities. We build and make observations from a relatively detailed, word-level dataset, inspired by the practice of sociolinguists and historical linguists. We consider a range of proposed explanations for language change proposed by phonologists, psychologists, and cognitive scientists, and examine their bearing on the change represented in our dataset. Finally, we build mathematical models of linguistic populations, to model the effect of assumptions about the process of language learning on diachronic, population-level dynamics. Our goal is to go back and forth between data and models, by using modeling to explore which types of models lead to properties observed in our dataset, and using properties observed in the data to inform our choice of models.

We differ from many computational models of language change in assuming that the elementary objects speakers learn are probabilities of using one form versus another. With some exceptions (Harrison et al., 2002; Yang, 2002; Mitchener, 2005; Niyogi, 2006; Daland et al., 2007; Troutman

et al., 2008), most computational models have assumed that the learner's task is to choose one form to categorically use, based on their input data. While this idealization may sometimes be appropriate, e.g. for models of parameter setting, variation is widespread in much of language, as has been best shown for phonetics and phonology. We are thus posing the actuation problem in a broader context than usual: when and how does variation lead to change, with variation both at the individual and population levels?

Finally, our modeling approach is somewhat different than usual. Many computational studies of change consider 1–5 models, often by simulation. Our approach is complementary. Our emphasis is not on finding a single model that explains a set of facts, but on developing a wide range of models and comparing their dynamics. We believe there is intrinsic value in exploring a "landscape" of models, in particular different algorithmic implementations of the same idea (about learning), to get a better sense of the source of observed model dynamics. By building simpler models, for example idealizing social network and lexicon structure, we can analyze their properties analytically and consider a larger model set. By considering a landscape of models, we hope to connect model and dataset properties.

1.2 Variation and change facts

We first outline the broad findings of linguists to be accounted for in modeling the interaction between variation and change.

1.2.1 Types of variation

Variation within individual speakers between discrete forms for the same linguistic object is widespread.[1] A classic example is English final t/d-deletion. Because sociolinguists have focused on the factors which condition an individual's use of different forms, such variation is often called "style shifting" or "stylistic variation" (e.g. Schilling-Estes, 2003). Since the source of the variation is not important here, we call it *intraspeaker variation*. We call the case where each individual uses one form exclusively *interspeaker variation*. Both kinds of variation are heavily influenced by both social (class, gender, social context) and internal (phonetic context, word frequency) factors, and therefore must be learned. One aim of this thesis is to understand the diachronic consequences of varying how this learning takes place and what type of variation is assumed.

That an individual's use of different forms is finely conditioned by a variety of factors can obscure an essential fact: when the linguistic and social contexts are fixed, individuals still show extensive variation across a variety of linguistic variables. That is, any computational model of language change must deal with the intraspeaker variation as well as the interspeaker variation case.

1.2.2 Stability of variation

If both intraspeaker and interspeaker variation exist, a relevant question for modeling is whether they are stable: do multiple forms survive in a population over long periods of time without one of them moving towards elimination, or is all variation temporary? Because we are used to thinking of a language's past as a series of completed changes, an intuitive answer would be that variation is

[1] We use "discrete" here to mean a choice between two or more forms, rather than "continuous" variation in some parameter (such as formant values).

never (or at least rarely) stable However, there is significant evidence to the contrary, to the extent that William Labov (2000, p. 75) writes:

> Stable, long-term variation that persists over many centuries in much the same form is perhaps more common than changes which go to completion.

Labov cites several sociolinguistic variables in English which have shown stable variation for several centuries, including the *-in*/*-ing* alternation. In the English N/V dataset discussed below, many words show stable variation in either the N or V forms (though rarely both). Nevalainen and Raumolin-Brunberg (2003, p. 78) show that variation between *-one* and *-body* indefinite pronouns has existed since the 1400s, although a third class (*-man*) has disappeared.[2] Outside of English, variable aspiration and deletion of Spanish /s/ has been observed in at least 12 dialects, and shows no sign of recent diachronic change (Labov, 2000, p. 86), while Brink and Lund (1975) report variation for several Danish sociolinguistic variables over periods of 100–250 years.

1.3 Motivation

What do we gain by formalizing theories of variation and change computationally? Different theories of learning, all of which seem intuitively plausible, can have quite different diachronic consequences. This is because change in a population accumulates by many transmissions over time, and the outcome of iterated transmission can be surprisingly subtle.

As an example, simple models of populations of learners with intraspeaker variation and of populations with interspeaker variation are worked out in §5.1–5.2. Both sections consider models which differ only in the target of learning: in the interspeaker variation case, learners choose the form they hear most often in their learning data; for intraspeaker variation, learners probability match. With this simple difference, the resulting diachronic, population-level dynamics turn out to be quite different.

However, differences between models need not lead to different dynamics. For several classes of models considered below, we consider both the case where learners all receive the same number N of examples, and the case where learners receive a random, Poisson-distributed number of examples, with mean N. Provided N is not very small, the (diachronic, population-level) dynamics are essentially the same in both cases.

The upshot is that varying some assumptions about the learning process changes the diachronic results, while varying others does not. Using computational models, we can tease apart the relative diachronic contributions of all aspects of any model of learning.

[2] Actually, Nevalainen and Raumolin-Brunberg show that all three types were used 1400–1700, and over this period the percentage of *-one* and *-body* forms increased while the use of *-man* forms decreased. We extrapolate to the present without proof.

2 Data

The data considered here are English disyllabic noun-verb pairs such as *convict*, *concrete*, *exile*. N/V pairs are a productive class (*YouTube*, *google*). To get a rough count of N/V pairs, the overlap of the CMU Dictionary (American English) with the British National Corpus word-frequency list (Leech et al., 2001) gives 3185 pairs, of which 1783 have N or V frequency > 0, and 647 have both N and V frequency > 0.

Out of the four logically possible stress patterns, all current N/V pair pronunciations follow one of three patterns (e.g. Fudge, 1984, p. 166):[3]

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

As discussed below, the $\{2,1\}$ pattern is also never observed diachronically. At any given time, variation exists in the pronunciation of some N/V pairs, e.g. *research*, *address*, *perfume* in current American English.

Variation and change in the stress of N/V pairs have a long history. Change in N/V pair stress was first studied in detail by Sherman (1975), and subsequently by Phillips (1984). Sherman (1975) found that many words have shifted stress since the first dictionary listing stress appeared (1570), largely to $\{1,2\}$.[4] On the hypothesis that this was lexical diffusion (Wang, 1969) to $\{1,2\}$, he counted 149 pairs where $\{1,2\}$ was a possible pronunciation in two contemporary dictionaries, one British and one American, and examined when the shift for each N/V pair took place. We call these 149 words List 1 (Appendix A). Sherman found the stress of all words in List 1 for all dictionaries listing stress information published before 1800, and concluded that while many words were $\{1,2\}$ by 1800, those that were not must have shifted at some point by 1975. We will reexamine Sherman's interpretation below after examining the dynamics of an expanded dataset.

Stability of {1,1}, {2,2}, {1,2} Because Sherman's study only considers N/V pairs which are known to have changed to {1,2} by 1975, it does not tell us about the stability of the {1,1}, {2,2}, and {1,2} pronunciations in general. Over a random set of N/V pairs in use over a fixed time period, is it the case that most pairs pronounced {1,1} and {2,2} shift stress to {1,2}?

List 2 (App. B) is a set of 110 N/V pairs chosen at random from all pairs which (a) have both N and V frequency of at least 1 per million in the British National Corpus (b) have both N and V forms listed in a dictionary from 1700 (Boyer, 1700) (c) have both N and V forms listed in a dictionary from 1847 (James and Molé, 1847). These criteria serve as a rough check that the N and V forms of each word have been in use since 1700.

In List 2, Only 11.8% of the words have changed stress at all from 1700–2007, and as shown in Table 1, the proportion of $\{1,1\}$, $\{1,2\}$, and $\{1,2\}$ words has changed little over time.

Those stress shifts observed are mostly as described by Sherman, from $\{2,2\}$ to $\{1,2\}$, and mostly for words from List 1. It is also clear that stress shifts have occurred largely for words

[3] We abbreviate N=noun, V=verb, N/V=noun-verb throughout, and use curly brackets to denote N and V stress, where 1=initial stress, 2=final stress.

[4] However, most words are not first listed until 1700 or later.

Dict	{1, 1}	{1, 2}	{2, 1}	{2, 2}
B1700	0.613	0.099	0	0.279
J1847	0.616	0.125	0	0.241
C2007	0.617	0.144	0.002	0.210

Table 1: Percentage of words from List 2 for different stress patterns in 3 dictionaries, 1700–2007. When variation is listed, all possible forms are counted with equal probability. (For example, N=1/2 and V=2 would contribute 1/2 count to both {1,2} and {2,2}.)

	1550-1699	1700-99	1800-99	1900-	**Sum**
British	8	25	15	14	**62**
American	0	0	4	10	**14**

Table 2: Distribution of dictionaries by date and dialect.

beginning with a morphological prefix.[5] But this quick look suggests that when the set of *all* N/V pairs is sampled from over a 300 year period, most words do not change stress: $\{1, 1\}$, $\{1, 2\}$, and $\{2, 2\}$ are all "stable states", to a first approximation. From this perspective, both sides of the actuation problem are equally puzzling for the dataset: why do the large majority of N/V pairs not change, and what causes change in those that do?

2.1 Diachronic: Dictionary data

To get a better idea of the diachronic dynamics, Sherman's data on N/V stress for List 1 words from 33 British dictionaries were extended to the present using 29 additional British and 14 additional American dictionaries, published 1800–2003. Words from List 1 were used rather than a list of N/V pairs controlled for first attestation and non-zero frequency (such as List 2) for two reasons. First, we wish to use the large dataset already collected by Sherman for List 1 pronunciations up to 1700. Second, we are interested in the dynamics of *change*, and would therefore like to focus on words which have changed by the present. Because most pairs do not change stress over time and most change is to {1,2}, List 1 will include most pairs which have undergone a stress shift.

All dictionaries are listed in App. A, and are referred to by the codes listed there (e.g. B1700 for (Boyer, 1700)) from here on to avoid confusion with non-dictionary references. Table 2 shows the distribution of dictionaries. Relatively few dictionaries (listing stress) were published pre-1700.

For the 149 N/V pairs of List 1 in 76 dictionaries, each of N and V was recorded as 1 (initial stress), 2 (final stress) 1/2 (both listed, 1 first) 2/1 (both listed, 2 first), 1.5 (level stress) 0 (not listed).[6] It is often assumed that English words have a unique primary stress, so that "level stress" is not possible. We interpret the 113 reports of level stress as equal preference for initial and final primary stress.

We denote the data as a 149×76 matrix A, where $A_{word,dict} = (\text{pron}_N, \text{pron}_V)$. A is 37% sparse ($(0, 0)$ entries), reflecting the fact that at a given time, the N or V forms for many words in List 1 are rare, archaic, or not yet in use. The combination N=2, V=1 is never reported.

[5] This is in line with an observation by Fudge (1984, p. 32): "Certain words exhibit different stress patterns depending on whether they are nouns or verbs. In the majority of cases the structure of such words is prefix + root..."

[6] A few extra possibilities, such as 1/2/1.5 occur, but are rare. We only use the first two pronunciations when 3 are listed.

	V=1	V=var	V=2
N=1	7.1	4.6	57.1
N=var	0	2.2	7.1
N=2	0	0	21.8

Table 3: Distribution (percentage) of data with both N and V stresses listed. "Var" means 1.5, 1/2, or 2/1.

2.1.1 Stress vs. time trajectories

Changes in individual N/V pairs' pronunciations can be visualized by plotting the moving average of their N and V stresses. To represent averages of reported stresses on a scale, we need to map reported stresses s as numbers $f(s)$ in $[1, 2]$. We use

$$f(1) = 1, \quad f(2) = 2, \quad f(1/2) = f(2/1) = f(1.5) = 1.5$$

This measure overestimates variation between 1 and 2 by interpreting 1/2 and 2/1 as meaning equal variation between 1 and 2. In fact, dictionary authors often state that the first listed pronunciation is "primary", meaning 1/2 really means 1 is heard more in the population than 2. In practice, 1/2 and 2/1 are uncommon enough that trajectories plotted using $f(1/2) = 1.25$, $f(2/1) = 1.75$ look similar.

For a word w at time t, the average of pronunciations reported in the time window $(t-25, t+25)$ (years) was plotted if ≥ 2 dictionaries in this time window listed pronunciation data for w. So that the trajectories would reflect change in one dialect of English, only data from British dictionaries were used. All 149 stress vs. time trajectories are in App. F; a subset of the trajectories are reproduced in Figs. 1–2 for convenience.

Four types of complete stress shift, defined as a trajectory moving from one endpoint ($\{1, 1\}$, $\{1, 2\}$, or $\{2, 2\}$) to another, are observed (ordered by decreasing frequency):

1. $\{2, 2\} \to \{1, 2\}$ (*concert*)
2. $\{1, 1\} \to \{1, 2\}$ (*combat*)
3. $\{1, 2\} \to \{1, 1\}$ (*collect*)
4. $\{1, 2\} \to \{2, 2\}$ (*cement*)

A sample of each type is shown in Fig. 1.

Change directly between $\{1, 1\}$ and $\{2, 2\}$ never occurs, suggesting that change occurs one form (N or V) at a time. This is borne out impressionistically by examination of all trajectories, and by the percentages of dictionary entries reporting variation in N, V, or neither, shown in Table 3. While variation is reported in either N or V in 13.9% of entries which list both N and V pronunciations, variation in both N and V forms at once is reported in only 2.2% of such entries.[7]

Examining all trajectories, we can make some impressionistic observations about the diachronic behavior of the observed variation. Variation near endpoints (*concert*, Fig. 1(c)) is relatively common, but long-term variation away from endpoints (*exile*, Fig. 2(b)) is rare. Long-term variation in *both* N and V forms at once (*rampage*, Fig. 2(a)), is very rare (2 trajectories).

{2,1} is never observed in the dataset, and we argue it is in fact "unstable" in the following sense. Entries "near" {2,1}, such as (N=2/1,V=1/2) are very rare (9 instances), and are scattered

[7]However, reporting variation in a N and reporting variation in its associated V co-occur significantly more often than would be expected by chance (O/E=3.48, Pearson's X^2=227, $p = 0$).

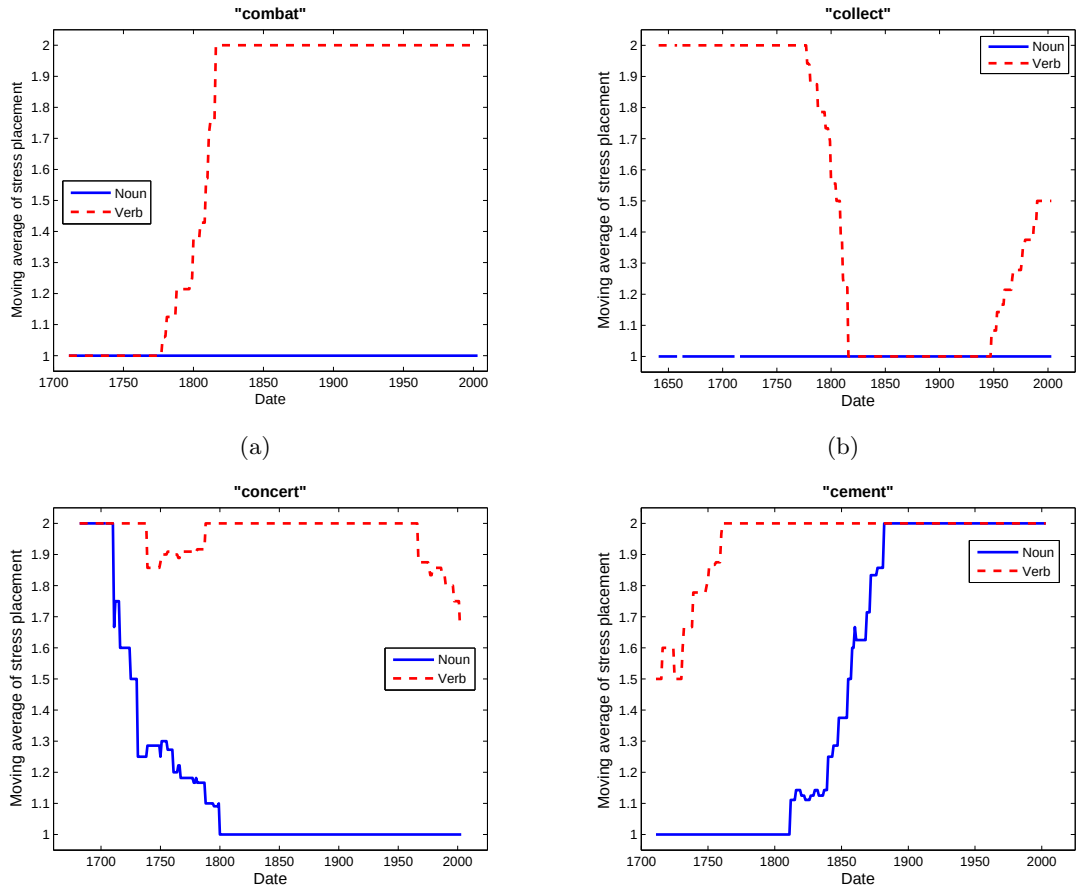


Figure 1: Sample trajectories 1: change between endpoints.

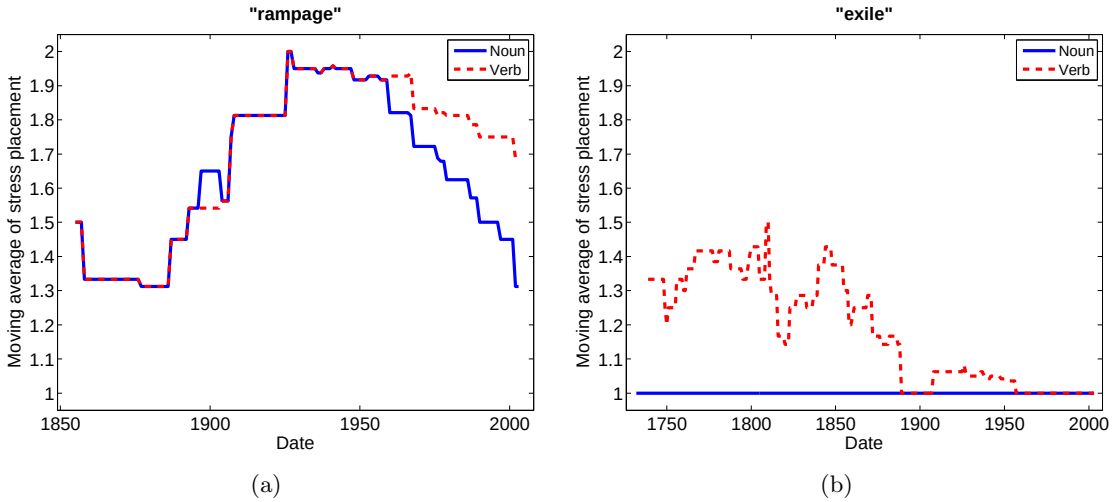


Figure 2: Sample trajectories 2: rare trajectory types

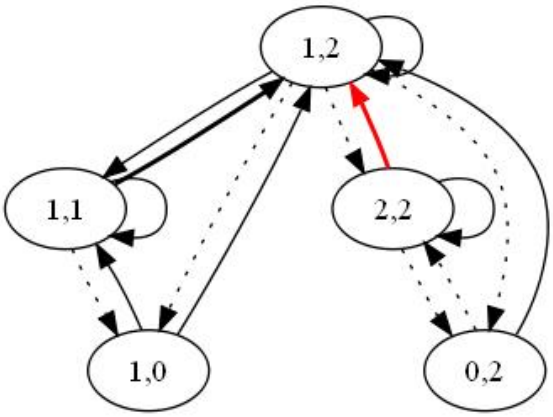


Figure 3: Schematic of observed changes. Self-loops indicate that $\{1,1\}$, $\{1,2\}$, $\{2,2\}$ are stable states. $\{1,0\}$ and $\{0,2\}$ are disyllabic words without V and N forms (respectively). Line thicknesses indicate transitions' relative frequencies.

across different words and dictionaries. This means that the few times a trajectory drifts towards the region $\text{pron}_N > \text{pron}_V$, it quickly moves away. In the language of dynamical systems (Sec. 4), this suggests the region contains an unstable fixed point (one which repels trajectories), {2,1}.

We can summarize the observed diachronic facts as follows:

1. {1,1}, {1,2}, {2,2} are "stable states", but variation around them often occurs. Long-term variation away from these states is rare.
2. Variation usually occurs in only one of N or V at a time.
3. Trajectories largely lie on or near a 1D axis in the 2D $(\text{pron}_N, \text{pron}_V)$ space: $\{1,1\} \leftrightarrow \{1,2\} \leftrightarrow \{2,2\}$. Both variation and change take place along this axis.
4. Changes to {1,2} are much more common than changes from {1,2}.
5. {2,1} never occurs, and is an "unstable state".

Returning to the question of what kind of change is taking place, we see that to a first approximation and restricted to List 1, Sherman was correct: most change takes place to $\{1,2\}$. But taking into account that change from $\{1,2\}$ also occurs, and that most words in stable states never change, the diachronic picture is more completely schematized as in Fig. 3. The observed dynamics are thus more complicated than diffusion to $\{1,2\}$. To understand their origin, we consider below (Sec. 3) proposed mechanisms driving stress shift.

2.2 Synchronic: Radio data

We can infer from the dictionary data that significant population-level variation exists in the pronunciation of many N/V pairs at a given time. However, to build models, we must also know whether pronunciation variation exists in individuals or not: do individuals learn gradient (a probability $\alpha \in [0,1]$ of using one form versus another) or categorical (each speaker uses one form exclusively) forms? As above (§1.2.1), we call these options *intraspeaker* and *interspeaker* variation.

Word	# N=1	# Var	# N=2
research	9	6	2
perfume	2	3	4
address	1	1	2

Table 4: Summary of radio pronunciation data from App. D. Number of speakers who used exclusively initial stress, exclusively final stress, or both for the noun form of a given word. Details in text.

One place to check the type of variation is on the radio, by observing how an individual speaker pronounces different tokens of words known to show variation at the population level. For a sample of 34 ratio stories, mostly from National Public Radio, Table 4 lists the number of speakers (31 total, 18 male) who pronounced the noun form of *research*, *address*, or *perfume*, exclusively with initial stress, exclusively with final stress, or used both. Each speaker listed for a word used it at least 5 times. The recordings are a mixture of conversations, broadcast speech, interviews, and call-ins to talk shows. Apps. C–D give a list of recordings used, as well as observed pronunciations for all speakers.

Intraspeaker variation thus does occur for N/V pairs, at least in this relatively small dataset. Though tentative, this finding has important consequences for modeling. As has been pointed out in both dynamical systems (Niyogi, 2006) and other computational models of language change (e.g. Liberman, 2000; Troutman et al., 2008), the choice of whether learners' target is a gradient or categorical form profoundly affects the population-level dynamics.

Based on the radio data, we can also make an observation about the structure of intraspeaker variation for modeling: although intraspeaker variation exists, 2/3 of speakers show no variation at all. This suggests learners cannot simply be probability matching (assuming their input includes both N=1 and N=2 examples), and that the learning procedure can terminate in gradient *or* categorical output, given gradient input.

3 Mechanisms of change

We turn from description of variation and change in the dataset to proposed mechanisms and motivations for change, which will inform the models we build below. After some background on English lexical stress, we discuss four types of explanations proposed for language change: mistransmisison, word frequency, regularization, and analogy.

3.1 Background: English stress

English word stress is complex and has long been of interest to phonologists (e.g. Chomsky and Halle, 1968; Halle and Keyser, 1971; Ross, 1972; Fudge, 1984; Hammond, 1999). We briefly discuss some very general facts which are relevant to modeling patterns observed in the N/V data and understanding experimental results discussed below.

English has a significant initial-stress bias, inherited from Germanic. By both type and token counts, a large majority of English content words have initial primary stress (Cutler and Carter, 1987), and more polysyllabic words (of 2–4 syllables) have initial primary stress than all other locations combined (Clopper, 2002).

The distribution of primary stress location also differs by part of speech, especially the broad tendency which we call *Ross' generalization*: "Primary stress in English nouns is farther to the left than primary stress in English verbs" (Ross, 1973, p. 168). For example, over the Kucera-Francis word list, 89% of nouns have initial primary stress, versus 46% of verbs. As discussed above, there are no English N/V pairs with final/initial primary stress, either at present or in all historical sources examined.

Finally, English stress assignment is quantity-sensitive. Roughly, syllables can be either heavy or light, where heavy syllables are those containing a long (tense) vowel, a coda, or both. Every heavy syllable must bear either secondary or primary stress, and heavy syllables are more likely than light syllables to bear primary stress. The interaction between syllable structure and lexical class has been observed by several authors, and is summarized by Albright (2008) as follows:

	Complex nucleus	Complex coda	Neither
Noun	✓		
Verb	✓	✓	

Here, the columns refer to final syllable structure, and "✓" denotes "usually receives final stress."

Although syllable weight is fundamental to English stress, it may be less important than lexical class for the stress shift considered here. There is experimental evidence, discussed below, that for (present-day) English speakers, lexical class is more important than syllable structure in stress assignment to novel disyllabic words. From the diachronic perspective, we observe from the N/V data that finally-stressed nouns with complex nuclei often shift to initial stress (*concert*, *decrease*, *detail*,...), while the reverse process is almost unattested.

Productivity There is good evidence that English speakers have productive knowledge of the factors influencing English stress, particularly lexical class. By introspection, stress is shifted leftwards in nominalized verb phrases: "He failed to follow thróugh" vs. "his fóllow-through was weak" (Kelly and Bock, 1988).[8]

[8]Since most experiments involving stress assignment have been done by American researchers, "English" in this section should be read "American English".

Less anecdotally, several experiments have addressed stress assignment to novel English words. Several early studies (Ladefoged and Fromkin, 1968; Walch, 1972; Baker and Smith, 1976; Baptista, 1984) focused on how well novel stress assignment was predicted by the rules laid out in *The Sound Pattern of English* (Chomsky and Halle, 1968). Unfortunately, these studies all involved written stimuli or writing tasks, an important confound given the opacity of English orthography.

A recent study (Guion et al., 2003) tested the effect of three variables on stress assignment to novel assignments to novel disyllabic words by native English speakers:

1. Lexical class (N vs. V)
2. Syllabic structure (CVV.CVCC, CV.CVCC, CV.CVC, CV.CVVC)
3. The stress of phonologically-similar words ("analogy")

Words were (importantly) presented aurally, as two isolated stressed syllables, and responses (to production and perception tasks) were given orally. In both production and perception tasks, all three variables significantly affected stress assignment, with the following effect sizes (odds ratios):

	N/V	Syl. Structure	Analogy
Production	4.6	2.8	1.7
Perception	2.0	1.8	1.6

In both production and perception, the N/V asymmetry was more influential in novel stress assignment than syllabic structure or analogy, but the effect is less pronounced in perception than in production.

Kelly (1988a) gives several experimental results consistent with the hypothesis that speakers use the N/V asymmetry productively. In one experiment, after hearing nonsense disyllables in isolation, speakers were more likely to use trochees than iambs as nouns, and vice versa for verbs. In another, speakers explicitly prompted to extend nouns to verb usages (and vice versa) were more likely to shift iambic nouns (than trochaic nouns) to verb usages (and vice versa for noun-to-verb shifts).

3.2 Mistransmission

The most commonly-proposed explanations for sound change are transmission errors, either in perception or production. The most prominent theories in this vein, reviewed by Hansson (2008),[9] are J. Ohala's listener-based theory of sound change (Ohala 1981 et seq.), and the Evolutionary Phonology framework of J. Blevins and collaborators (e.g. Blevins and Garrett, 1998; Blevins, 2004). Proposed transmission errors can be based in either articulation (coarticulation, reduction, assimilation), or perception (perceptual confusability, failure to compensate for coarticulation). We call all such errors *mistransmission*. Some examples:

- *Final devoicing*: The loss of a word-final contrast between voiced and voiceless obstruents is a common sound change. Several articulatory reasons (why voicing is weaker in final position on obstruents) and perceptual reasons (why voicing is harder to perceive on final obstruents) for this have been advanced (summarized by Blevins, 2006).
- *Unconditioned place shifts*: Sounds changes such as /θ/ → /f/, /x/ → /h/, or /k^{w}/ → /p/, where a consonant shifts place in all environments. Such changes are hypothesized to result from perceptual similarity between the new and old sounds.

[9] From which this discussion borrows

- *Palatalization*: A common sound change is /t/ → /t͡ʃ/ or /t/ → /t͡s/ before high vowels and glides.[10] This change has been argued (Ohala, 1983) to have acoustic-phonetic roots.

3.2.1 The N/V case

In the domain of N/V pair stress, there is significant experimental evidence for perception and production biases in English listeners consistent with the most commonly-observed diachronic shifts ({2,2}, {1,1}→{1,2}). English listeners prefer the typical stress pattern (N=1 or V=2) in novel English disyllables (Guion et al., 2003), and show higher decision times and error rates (in a grammatical category assignment task) for atypical (N=2 or V=1) than for typical disyllables (Arciuli and Cupples, 2003).

Stress perception is also strongly affected by words' rhythmic contexts. English shows a strong tendency to alternate strong and weak syllables. For disyllables, 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.[11] This effect is more pronounced for nouns than for verbs; and nouns occur more frequently in trochaic contexts (Davis and Kelly, 1997; Kelly, 1988b, 1989; Kelly and Bock, 1988) M. Kelly and collaborators have argued these facts are responsible for both the N/V stress asymmetry and the directionality of the N/V pair stress shifts described by Sherman (1975).

3.3 Frequency

The possibility that sound change proceeds faster for some words than others, contrary to the Neogrammarian hypothesis, has remained controversial since it was proposed by Schuchardt (1885). More recent work argues for a relationship between word frequency and sound change in a variety of cases.[12] Proponents argue that in some changes (e.g. English yod-dropping: Phillips (1981)), low-frequency words change first (on average); in others (e.g. English schwa-deletion: Hooper (1976)) high-frequency words change first.[13]

3.3.1 The N/V case: Low-frequency first?

Although the role of frequency in sound change remains contested, frequency is commonly invoked in other types of changes, such as analogy and paradigmatic change. Because the N/V stress shift seems to fall into these categories rather than a classical sound change,[14] it would not be surprising to find that word frequency plays a role in which N/V pairs shift stress.

Phillips (1984) claims that Sherman's N/V dataset is an example of low-frequency-first (analogical) change, by the following analysis. Phillips compares lists of N/V pairs given by Sherman which (a) were originally $\{2, 2\}$, and have undergone change to $\{1, 2\}$ (b) were originally $\{2, 2\}$,

[10] In American English, this process operates as an optional reduction rule: *don't you*/*doncha*. In British English, it optionally applies in words such as *tuna*.

[11] Sample trochaic-biasing and iambic biasing contexts are "Use the colvane proudly" and "The proud colvane refused".

[12] For various perspectives, see Labov (2000, Part D), Bybee and Hopper (2001), Hock (1991).

[13] This line of work argues changes which are "analogical" or which involve "grammatical analysis" are low-frequency-first, while those which are "phonetically-motivated" are high-frequency-first. However, De Schryver et al. (2008) have recently shown that in the ongoing sound change of Dutch fricative devoicing, lower-frequency words are affected first.

[14] We thank Brian Joseph for pointing this out.

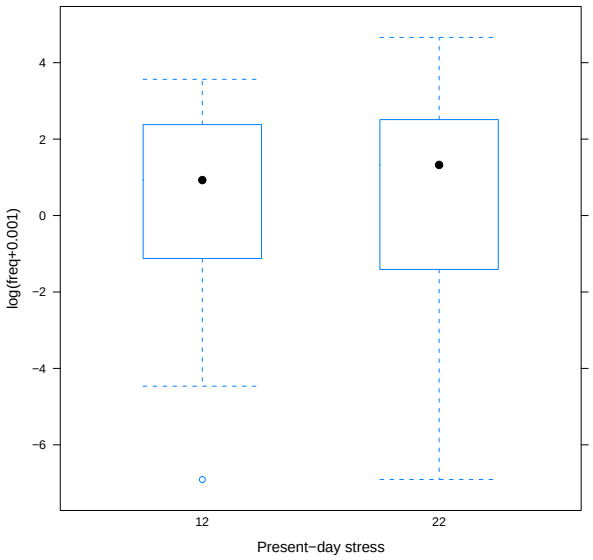


Figure 4: Boxplot of log-transformed frequencies vs. current pronunciation ({1,2} or {2,2}) as reported in (Phillips, 1984) (all words).

and have not changed to $\{2, 2\}$; and finds that within a given prefix class, the average frequency of group (b) is higher than that of group (a).

Unfortunately, this analysis suffers from an important statistical problem: it is not tested whether the frequency differences observed are statistically significant.[15] To visualize the distribution of frequencies reported by Phillips, Fig. 4 and Fig. 5 give boxplots of frequencies for {1,2} and {2,2} words (collapsed across all prefixes and sorted by prefix, respectively). The plots suggest that while the mean frequency of {1,2} words is reliably higher than the mean frequency of {2,2} words, both within prefix classes and across all words, the differences in means are small relative to the within-group variances.

This intuition is borne out by testing the significance of these differences. Within each prefix class, as well as across all words, the frequencies of {2,2} and {1,2} words are not significantly different ($p > 0.2$, two-sided Mann-Whitney tests).

While the frequency differences observed by Phillips are in fact not significant, they are consistently in the direction predicted by Phillips' hypothesis ("changed" word frequencies < "unchanged" word frequencies), suggesting that the negative result may be due to the methodological problems discussed above. To give Phillips' hypothesis a fair test, we constructed a dataset which addresses methodological issues.

[15]There are also several methodological issues: (1) A contemporary American English word frequency list is used for British pronunciation data from 1700–present. (2) Word frequencies are not log-transformed. (3) Sherman's list of words "likely to undergo change" is used as the control group (of unchanged $\{2, 2\}$ words), but it is not clear how Sherman obtained this list, and thus why it should be used. (4) N/V pairs are not controlled for how long they have been in use. (A high-frequency pair first used in 1700 which has not changed to $\{1, 2\}$ presumably should count more than an unchanged, high-frequency pair where either the N or V form only recently appeared.)

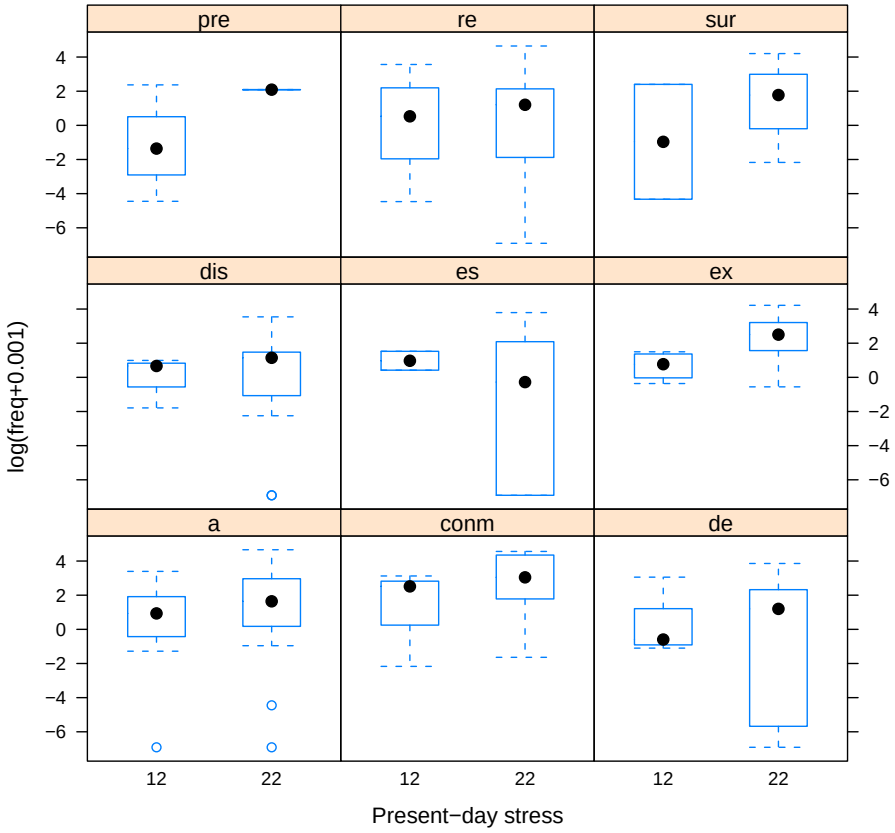


Figure 5: Boxplots of log-transformed frequencies vs. current pronunciation ({1,2} or {2,2}) as reported in (Phillips, 1984) (sorted by prefix class).

Testing on a new dataset List 3 (App. B) is a set of disyllabic N/V pairs which meet two conditions, based on (British) pronunciations listed in a French-English dictionary from 1700 (Boyer, 1700), as well as modern pronunciations (Cambridge Advanced Learner's Dictionary, OED).

1. Each word is listed as {2,2} in 1700, and as either {2,2} or {1,2} today.
2. Each word's initial syllable is *a*, *com/n*, *de*, *dis*, *ex*, *pro*, or *re*.

List 4 (App. B) is an analogous list of pairs listed in both an 1847 French-English dictionary (James and Molé, 1847) and the modern dictionaries.

For each N/V pair in List 2, log-transformed frequencies were found based on two corpora: the British National Corpus, and the 1700–1800 portion of the OED Quotes Database.[16] These two frequency measures are approximations of the frequencies we actually want: frequencies of N and V lemmas from 1700 to the present in British sources, which cannot be easily found using currently available corpora. The two measures deviate from this ideal in different ways. The BNC frequencies are taken from a state-of-the-art corpus and are lemmatized, but reflect current usage. The OED frequencies are taken from a smaller, unlemmatized corpus which was not purpose-built for linguistic use, but (importantly) reflect 18th-century usage.

Despite these drawbacks, the new dataset does not have most of the methodological problems discussed above. Most importantly, Lists 3 and 4 only include words known to be (1) {2,2} in a historic dictionary (2) listed in present-day dictionaries.

We can now check whether Phillips' hypothesis holds for this dataset. We consider four overlapping subsets of Lists 3–4:

1. G_{1700}: Words listed as {2,2} in 1700 and today.
2. G'_{1700}: Words listed as {2,2} in 1700 and {1,2} today.
3. G_{1847}: Words listed as {2,2} in 1847 and today.
4. G'_{1847}: Words listed as {2,2} in 1847 and {1,2} today.

Of interest is whether words in G_{1700} have greater frequencies than words in G'_{1700}, and similarly for G_{1847} and G'_{1847}. Table 5 summarizes the results of Mann-Whitney tests of these hypotheses, across all words and restricted to those prefixed with *re-*, the only prefix for which enough words changed for a comparison between changed and unchanged word frequencies to be meaningful.[17] Fig. 6 gives boxplots comparing changed and unchanged words for the two frequency measures for both 1700 and 1847 data. All 8 tests listed are marginally significant ($p < 0.1$), and of each pair of tests using BNC and OED frequencies, at least one is significant ($p < 0.05$).

Real-time frequency trajectories We have thus confirmed Phillips' hypothesis, at least for the most common subset of the N/V stress shift, {2,2}→{1,2}. However, one can entertain at least two hypothesis for why low-frequency words change (on average) earlier:

1. Words' relative frequencies stay approximately constant diachronically. In a given year, word a is more likely than word b to change if a is less frequent than b.

[16] BNC frequencies were taken to be log(`Nfreq`+`Vfreq`+0.25), where Nfreq and Vfreq are the frequencies (per million) of the N and V lemmas (Leech et al., 2001). OED frequencies were queried from Mark Davies' web interface (`http://corpus.byu.edu/oed/`) to the OED quote database, and were taken to be log(`freq`+0.1), where `freq` is the number of hits per million for the uninflected form. Hoffmann (2004) discusses the merits of using the OED quote database as a corpus.

[17] Specifically, the only class in which > 3 words changed.

	1700	1847
BNC	$W = 568$, $p < 0.05$	$W = 513$, $p < 0.05$
OED	$W = 580$, $p < 0.1$	$W = 514$, $p < 0.05$

	1700	1847
BNC	$W = 54$, $p < 0.1$	$W = 39.5$, $p < 0.1$
OED	$W = 34$, $p < 0.05$	$W = 24$, $p < 0.05$

Table 5: Left: Summary of one-sided Mann-Whitney tests of whether frequencies of G_{1700} words are greater than G'_{1700} words (and similarly for G_{1847} vs. G'_{1847} words), for BNC and OED frequency measures. Right: Same for *re-* words only. (Details in text.)

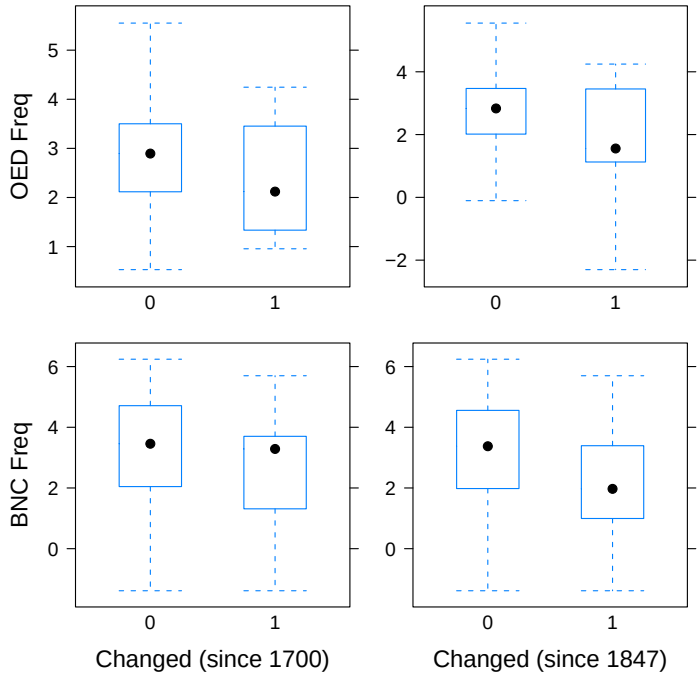


Figure 6: OED (top row) and BNC (bottom row) frequencies (log-transformed) for words which were {2,2} in 1700 (left column) and {2,2} in 1847 (right column). Changed=0 means current pronunciation={2,2}; changed=1 means current pronunciation={1,2}.

2. A given word changes when its frequency drops below a (possibly word-specific) critical value.

Under hypothesis 2, the reason present-day frequencies are on average lower for words which have changed is that their frequencies have decreased diachronically.

We can begin to differentiate between these hypotheses by examining diachronic frequency trajectories for N/V pairs which have changed, and checking whether they show negative trends. Real-time frequency trajectories were found for 6 N/V pairs (*combat*, *decrease*, *dictate*, *perfume*, *progress*, *protest*) which have shifted stress since 1700. Fig. 7 shows frequency trajectories alongside pronunciation trajectories for these pairs.

Frequencies were found by sampling from prose written by British authors in the Literature Online (LiOn) database, then normalizing against frequency trajectories for a set of reference words.[18] All words show negative correlations between year and N+V frequency, 5/6 of which are significant.[19] Although these correlations are weak, they lend support to Hypothesis 2, and rule out the assumption that the frequency trajectories for N/V pairs show no long-term trends. We thus adopt the working hypothesis that change occurs in an N/V pair when its frequency drops below a critical level.

3.4 Regularization

Recent experimental and computational studies have raised the general idea of *regularization* as a potential source of language change, defined as any bias in learners against gradience, as opposed to categorical forms. Under this hypothesis, although learners can learn variation between multiple forms, they produce less variation than was present in their learning data.

This possibility was suggested in a seminal case study (Singleton and Newport, 2004) of Simon, a deaf child learning ASL from deaf parents who learned ASL late in life, and who has received no native ASL input. Simon and his parents were prompted to give signs corresponding to a set of verbs of motion, which in ASL are morphologically complex and expressed by a combination of attributes. For individual verbs, Simon's parents usually showed variation between the correct form and incorrect forms for a given attribute, with the correct form most frequent. For most verb attributes, Simon produced the correct form with higher probability than his parents, and in *all* cases with higher probability than the lower-scoring parent; Singleton and Newport call this *frequency boosting*. This behavior is striking because Simon has no native input to indicate which forms of verb attributes are correct, and must therefore have a bias towards correct forms on the basis (at least in part) of their higher frequency.

Similar behavior has been observed in experimental settings by Hudson Kam and Newport (2005, 2009), who found that children learning artificial VSO languages containing variation often showed frequency boosting, while adult subjects usually did not.

From a different perspective, two simulation-based studies of linguistic populations incorporating intraspeaker variation (Liberman, 2000; Troutman et al., 2008) each tested several models, and found that only those including a regularization bias give dynamics resembling language change.

[18] `lion.chadwyck.com`. Only 6 words were considered because finding trajectories is time-intensive.

[19] Alphabetically: $r = -0.78$ ($p < 0.001$), $r = -0.78$ ($p < 0.1$), $r = -0.79$ ($p < 0.01$), $r = -0.32$ ($p > 0.25$), $r = -0.76$ ($p < 0.05$), $r = -0.74$ ($p < 0.01$)

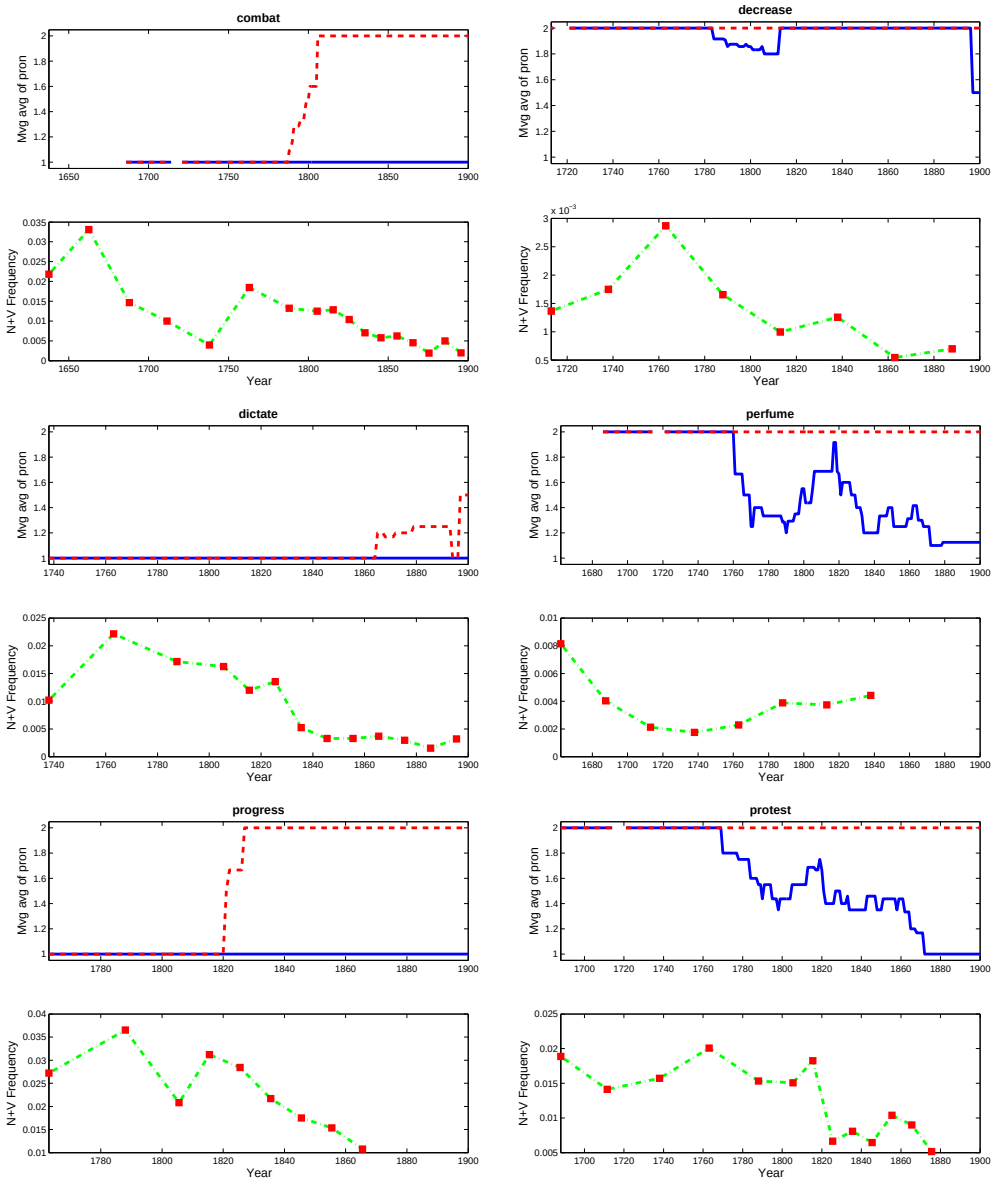


Figure 7: Frequency (bottom) and pronunciation (top) trajectories for *combat*, *decrease*, *dictate*, *perfume*, *progress*, *protest*.

3.5 Analogy

We use "analogy" here as a cover term for any time some sort of similarity between linguistic forms plays a role (or is hypothesized to) in change.

As used by historical linguists, analogy includes any change in which a set of linguistic objects becomes somehow more similar. Analogical change is common in language, but traditionally analogical explanations are only adopted as a last resort (once a researcher is satisfied that no regular change could have accounted for a given pattern) because they typically are not falsifiable. Put otherwise, one can always postulate that two forms became more similar because of analogy, but this does not explain why most possible analogical changes, even very similar ones, do *not* happen.

In discussions of change by non-historical linguists, analogy encompasses any learning algorithm in which a learner's treatment of data pertaining to some form (sound, word, paradigm) is conditioned by similar forms in their lexicon (e.g. Bybee and Hopper, 2001; Bybee, 2003; Daland et al., 2007; De Schryver et al., 2008).

We show below evidence that analogy (in the first sense) plays a role in the diachronic stress trajectories of N/V pairs. When building models (§7.4), we consider the population-level effects of including analogy (in the second sense) in individuals' learning algorithm.

3.6 The N/V case: Analogy within prefix classes

Impressionistically, over all N/V pair trajectories, those for pairs sharing a prefix often seem more similar than would be expected by chance. For example, many *re-* pairs were historically {2,2}, then began to change sometime 1875–1950. We would like a principled way to test the hypothesis of coupling between the trajectories for words in the same prefix class; for this, we need a way to test how much two words "change like" each other, or how similar their trajectories are. We use a simple distance metric over trajectories' dissimilarity ("distance"), $d(w, w')$, for N/V pairs w and w'.

We propose one simple d. Assume there are n words and m dictionaries. Order the dictionaries by publication date, and let w_i be denoted by i. Map reported stresses to $[0, 1]$ as above (2.1.1). Define

$$d(w, w') = \lambda D_1 + (1 - \lambda) D_2$$

where $\lambda \in [0, 1]$, D_1 is the normalized sum of $f_1(A_{d,w}) - f_1(A_{d,w'})$ over all dictionaries d which give pronunciations for both w and w', and D_2 is the normalized sum of the difference between the first differences of $f(A_{d,w})$ and $f(A_{d,w'})$ whenever this is defined, i.e. the empirical estimate of the difference between derivatives.

This d (1) penalizes w_1 and w_2 for each dictionary d such that $A_{w_1,d} \neq A_{w_2,d}$ (2) penalizes w_1 and w_2 for each timestep in which their rates of change are different. The relative amount (1) and (2) are penalized is determined by λ.

Finding $d(w, w')$ for all possible word pairs defines a graph $G(d)$ with nodes $w_1, \ldots, w_{149}$, and edges $d(w_i, w_j)$ equal to the distance between w_i and w_j's trajectories. This structure suggests a way of testing whether, given a group of words which are linguistically related, their trajectories are similar: check the goodness of the cluster formed by their vertices in G. For a subset of vertices

$C \in [n]$ of $G = (V, E)$, define

$$R(C) = \frac{\sum_{(i,j):i,j \in C} d(i,j)}{\binom{|C|}{2}} - \frac{\sum_{(i,j):i \in C, j \notin C} d(i,j)}{\binom{n}{2} - \binom{|C|}{2} - \binom{n-|C|}{2}}$$

This is the mean in-degree of C minus the mean out-degree of C. $R(C)$ will be high if most vertices of C are on average closer to each other than to vertices in $V \setminus C$.[20]

As a measure of the goodness of a cluster C, let $p(C) \in [0, 1]$ be the empirical p-value, defined as the location of $R(C)$ on the distribution of R for all communities of size $|C|$ in G:

$$p(C) = \frac{|S \in [n] : |S| = |C|, R(S) < R(C)|}{\binom{n}{|C|}}$$

The closer the value of $p(C)$ to 0, the more similar the trajectories for words in C are, compared to a random set of words of size $|C|$

This setup can be used to test whether words in List 1 which share a prefix have similar trajectories. Table 6 shows $p(C)$ for all prefix classes of size $|C| > 2$, with $\lambda = 0.5$. Considering $\lambda \in [0, 1]$, the results are broadly similar for $\lambda \in (0.3, 1)$.[21]

C	$\lvert C\rvert$	$p(C)$
a-	10	0.270
com-	5	0.067
comp-	3	0.032
con-	17	0.001
cont-	4	0.266
conv-	4	0.033
com-/con-	22	0.0005
de-	7	0.285
de- w/o des-	5	0.050
dis-	5	0.746
ex-	6	0.981
im-	4	0.021
in-	12	0.029
im-/in-	16	0.004

C	$\lvert C\rvert$	$p(C)$
out-	10	0.055
per	3	0.263
pre	5	0.065
pro	4	0.078
re-	24	0.011
re- (bound)[a]	8	0.576
re- (unbound)	16	0.0017
sub-	3	0.710
trans-	3	0.173
up-	7	0.196

[a] "bound": re-μ, where μ is a bound morpheme

Table 6: Prefix class $p(C)$ values, $|C| > 2$, $\lambda = 0.5$.

Many potential prefix classes have small $p(C)$, confirming the initial intuition that N/V pairs sharing a prefix tend to have more similar trajectories. The *com-/con-*, *im-/in-*, and *re-* categories

[20] This quantity is adapted from a common metric for finding community structure in networks (Newman and Girvan, 2004), with the important difference that here we are only evaluating *one* hypothesized community rather than a partitioning of G into communities , so the heuristics for evaluating global community structure do not apply.

[21] For $\lambda \in (0, 0.3)$ the results below quickly break down, suggested that the difference in approximated derivatives alone is not a good measure of similarity.

are particularly interesting because they suggest that it is a shared morphological prefix rather than simply a shared initial syllable which correlates with trajectory similarity. $p(C)$ for combined *com-* and *con-* is lower than either alone, and the same holds for *im-*/*in-*; this makes sense under the assumption that in- and *im-* are allophones of a single underlying prefix (*in-*). For *re-*, the difference between the productive (before bound morphemes) and unproductive (before unbound) versions is clearly reflected in $p(C)$.

We also find that larger classes have lower $p(C)$: there is a significant negative relationship between $|C|$ and $\log(p(C))$ ($r = -0.72$, $p < 0.0001$) for the data in Table 6. This relationship is interesting given the main criticism of explanations invoking analogy, its arbitrariness: what determines whether a given set of forms which are somehow similar become more similar? In the current case, we can answer that analogy is more likely to affect larger sets of similar words. This makes intuitive sense: if analogy can affect any two or three related words as much as large classes, we would expect much less regularity in language change than is observed.

4 Dynamical systems

We derive models in the dynamical systems framework, 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, 1996; Niyogi, 2006; Komarova et al., 2001; Yang, 2001, 2002; Mitchener, 2005; Pearl and Weinberg, 2007). This framework is not a theory of language change, but a formalism to test theories of how change occurs. It 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 received by learners.

Our models are *discrete dynamical systems*, or *iterated maps* (e.g. Strogatz, 1994; Hirsch et al., 2004). Given a domain X, an iterated map is a function $f : X \to X$ that "iterates" the system by one step. If a system has value $\alpha_t \in X$ at step t, it has value $\alpha_{t+1} = f(\alpha_t) \in X$ at step $t + 1$. In models considered here, $X = I$ (Sec. 5, 6), I^2, or I^3 (Sec. 7), where $I = [0, 1]$.

Example: Let $I = [0, 1]$, and let $f(x) = x^a$, where $a > 0$, so that

$$\alpha_{t+1} = \alpha_t^a$$

Given initial state α_0, we can solve explicitly: $\alpha_t = {\alpha_0}^{a^n}$ Depending on the initial state α_0 and a, we have:

- $\alpha_0 = 0$ or $\alpha_0 = 1$: $\alpha_t = \alpha_0$ for all $t > 0$.
- $0 < a < 1$, $\alpha_0 > 0$: $\alpha_t \to 1$ as $t \to \infty$.
- $a > 1$, $\alpha_0 > 0$: $\alpha_t \to 0$ as $t \to \infty$

Unlike in this example, it is usually impossible to explicitly solve for α_t as a function of α_0. The dynamical systems viewpoint is to instead look at the long-term behavior of the system by looking for (1) fixed points (2) changes in their number and stability, called bifurcations.

Definition: $\alpha_* \in X$ is a *fixed point* of f if $\alpha_* = f(\alpha_*)$.

In Example 4, 0 and 1 are fixed points. However, when a is fixed, there is a qualitative difference between them. Fix $a = 2$. Then for any $\alpha_0 \neq 0$, $\alpha_t \to 1$. 0 is "unstable" in the sense that perturbing α_0 from 0 gives a different α_t as $t \to \infty$, while 1 is "stable" in the sense that perturbing α_0 from 1 does not. This notion is equivalent to a condition on f which allows a concise definition of stability.

Definition: Let f' denote the derivative of f if $X = I$, or the Jacobian of f if $X = I^2$ or I^3. A fixed point α_* is *stable* if $|f'(\alpha_*)| < 1$, *unstable* if $|f'(\alpha_*)| > 1$, and *neutrally stable* if $f(\alpha_*) = 1$.

As intuitive justification, let $X = I$ and consider a point $\alpha_1 = \alpha_* + \epsilon$ near a fixed point α_*. The ratio of the distance from α_* before and after applying f at this point is

$$\left| \frac{f(\alpha_1) - \alpha_*}{\alpha_1 - \alpha_*} \right| = \left| \frac{f(\alpha_* + \epsilon) - \alpha_*}{\epsilon} \right|$$

As $\epsilon \to 0$, this is the definition of the derivative of f at α_*. So f maps a point near α_* nearer to α_* if $|f'(\alpha_*)| < 1$, and further from α_* if $|f'(\alpha_*)| > 1$.

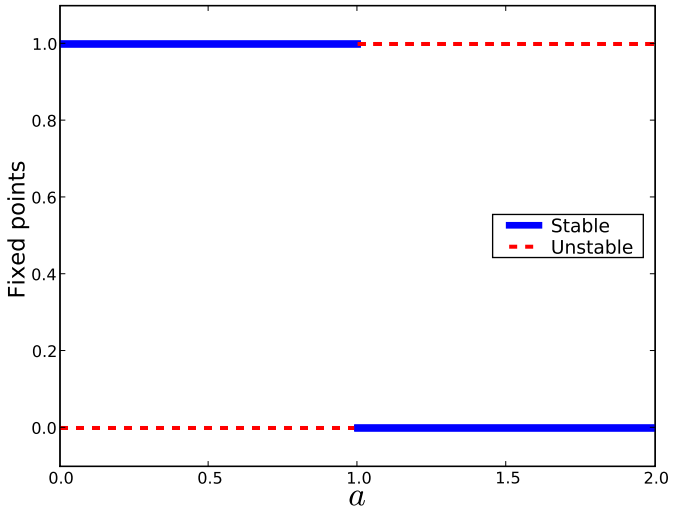


Figure 8: Bifurcation diagram for Example 4. The fixed points at 0 and 1 change stability at $a = 0$.

Definition: A *bifurcation* occurs when the number or stability of fixed points changes as a system parameter is changed.

Information about fixed points as a function of system parameters can be summarized in a *bifurcation diagram*. The bifurcation diagram in Fig. 8 shows the bifurcation at $a = 1$ in Example 4, where the fixed points exchange stabilities.

One more definition will be useful for talking about fixed points of dynamical systems on I^n:

Definition: $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is an *interior point* if each $x_i \in (0, 1)$.

4.1 Dynamical systems interpretation of variation/change data

The remainder of the thesis describes a range of dynamical systems models of linguistic populations. To interpret whether a given model has properties consistent with the N/V dataset, and with observations about variation and change more generally, we must translate the empirical facts into the language of dynamical systems. We will then have a list of desired properties against which to evaluate the dynamics of a given model. We argue this list is as follows:

- **Bifurcations**: A central insight of the dynamical systems approach (Niyogi and Berwick, 1995; Niyogi, 2006) is that the abruptness which characterizes many linguistic changes can be understood as bifurcations in the dynamical systems describing linguistic populations.

  In the context of the N/V data, we interpret {1,1}, {1,2}, and {2,2} as fixed points, and a stress shift as a bifurcation where a fixed point becomes unstable. A model's bifurcations should be consistent with the endpoint-to-endpoint changes observed in the data. For example, a bifurcation from {1,1} and {1,2} as stable fixed points to {1,2} as the only stable fixed point would correspond to change from {1,1} to {1,2}, which is observed in the data. Change from {1,1} to {2,2} is not observed, so a bifurcation from stable fixed points {1,1} and {2,2} to a single stable fixed point {2,2} would not be consistent with the data.

- **Frequency dependence**: A system parameter involved in these bifurcations should be an N/V pair's frequency, by the hypothesis that a drop in word frequency (N) below a critical value (N_0) is what triggers a word's stress to shift from {2,2} or {1,1} to {1,2}.

- **Stable and unstable non-interior fixed points**: {1,1}, {1,2}, and {2,2} are interpreted as stable fixed points where 100% of the population uses the same pronunciation, and {2,1} as an unstable fixed point.

- **Stable interior points**: There exists a range of system parameter values where an interior point is a stable state. In linguistic populations, a stable interior point corresponds to stable variation, which has been argued to be widespread in languages (§1.2.2).

  In the context of the N/V data, stable variation in both N and V at once is very rare, but is common in one of N or V at a time. If we observe, for example, that N/V trajectories can stay near (α_*,2) for decades, where $\alpha_* \in (0,1)$, a successful model should predict a stable state of this form for some parameter values. Because we very rarely observe stable variation of the form (α_*, α'_*), a successful model will not predict such stable states.

- **Multistability**: There exists a range of system parameter values with multiple stable states (*bistability*, *tristability*, etc.). This means that there are multiple states which different populations with identical (or similar) system parameters could remain in indefinitely. In linguistic populations, we interpret multistability as an explanation for dialects of the same language, which are often broadly extremely similar, except for significant differences in particular linguistic features. Intuitively, if two different populations *start* in different stable states, they stay in them.

  In the N/V data, for example, different dialects of English assign different stress to some N/V pairs, despite the fact that they (a priori) differ little in the various "system parameters" which may underly observed stress patterns: perceptual errors due to Ross' generalization, the relative frequency of the {1,1}/{1,2}/{2,2} stress patterns, the syllabic structure of N/V pairs, and word frequencies.[22] For example, if a community of speakers stresses *cement* as {1,2}, another community speaking a similar dialect stresses it as {2,2}, and neither stress shifts diachronically, a successful model should predict that these stress patterns can *both* be stable states for some system parameter settings.

Each of these properties seems very broad, and indeed many models considered below satisfy each one. What is more interesting is how few models satisfy *all* three properties, and can thus be used to model variation and change.

4.2 Models: Outline

In the remainder of this thesis, we consider three general types of models. The properties of all models considered, with respect to the desiderata discussed above, are summarized in Tables 7 and 8.

In Section 5, we consider models where the object to be learned is a single form, and learners are unbiased. By a single form, we mean that the learner chooses one of two options present in their learning data (interspeaker variation), or learns a probability of producing one option versus

[22] c.f. British vs. American 'research', 'perfume', Indian vs. American 'hexagon', 'delay', etc.

the other (intraspeaker variation). By unbiased, we mean that the learner is not biased towards learning some probabilities over others.[23] In Section 6, we consider models where the object to be learned is a single form, but learners are somehow biased.

In Section 7, we consider models where the object to be learned is a pair of forms, such as the probabilities of initial and final stress for the N and V forms of an N/V pair. We call these forms "coupled".

4.2.1 Model assumptions

We assume the following for all models:

- *Discrete generations*: Learners in generation n learn from generation $n-1$.
- *Full connectivity*: Each example a learner in generation n hears is equally likely to come from any member of generation $n-1$.
- *Infinite populations*: Each generation has infinitely many members.

These are idealizations, adopted here to keep models simple enough to analyze. Future work should examine the effects of overlapping generations, social network structure, and finite populations in more detail; Niyogi (2006) gives preliminary explorations in these directions. That learners can also learn within the lifespan should also be explored. But these idealizations are not unrealistic as a limiting case: variation and change should still occur in strongly-connected, large populations, and the time between generations can be made as short as desired.

Fixed vs. Poisson input We consider two types of input to learners:

- *Fixed input*: Each learner receives the same number of examples, N.
- *Poisson input*: Each learner receives a number of examples drawn from a Poisson distribution with mean N.

Most dynamical systems models in the literature assume fixed input, which may be appropriate in some settings, for example a cue-based learning model in which a learner generalizes over a fixed number of forms based on generalizations from the input. But Poisson input is in general more appropriate, since hearing relevant input is a Poisson process: the number of times an infrequent event (receiving relevant data) occurs over a fixed period of time (acquisition).

Why can fixed-input ever be used, then? Informally, Poisson input "looks like" fixed input for large N in the following sense: the mean and standard deviation of a Poisson distribution are $\mu = N$, $\sigma = \sqrt{N}$. Thus a measure of the "sharpness" of the distribution, $\sigma/\mu = 1/\sqrt{N}$, goes to 0 as $N \to \infty$. So for large N, a Poisson distribution looks increasingly similar to a fixed-input distribution, the delta function $\delta(N)$, where all probability mass is at N.

However, this convergence happens slowly (as a function of $\sqrt{N}$), and the N/V data considered here includes many low-frequency words. We therefore consider both the Poisson and fixed input cases whenever feasible.

[23] E.g. "regularization", discussed above (§6.1).

5 Models I: Individual forms, unbiased learners

We now consider a variety of models for variation and change for a single form in populations of learners.

5.1 Base models

We first describe the simplest models for the interspeaker and intraspeaker variation cases, introducing notation as we go.

In both cases, we assume speaker i of generation $t+1$ learns and stores a number $\hat{\alpha}_{i,t}$ corresponding to their probability of using form 2. We also assume fixed input.

5.1.1 Interspeaker variation

Assume two forms, and assume each speaker uses either form 1 or form 2 exclusively (interspeaker variation). Let α_t be the percentage of form 2 users in generation t. In generation $t+1$, learner i hears N examples, of which k examples are form 2 and $N-k$ are form 1, and sets their $\hat{\alpha}_{i,t}$ as follows:

$$\hat{\alpha}_{i,t} = \begin{cases} \hat{\alpha}_{i,t} = 1 & : k > rN \\ \hat{\alpha}_{i,t} = 0 & : \text{otherwise} \end{cases}$$

where $r \in (0,1)$ is a constant. For simplicity, let $r = 0.5$.

The probability of drawing k form 2 examples is binomially distributed:

$$P(k) = \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k}$$

Let $\hat{\alpha}_t$ be the random variable corresponding to the $\hat{\alpha}_{i,t}$ of learners in generation $t+1$, learning from generation t. That is, every draw of $\hat{\alpha}_t$ picks a random i and gives $\hat{\alpha}_{i,t}$. The probability of setting $\hat{\alpha}_t = 1$ is then[24]

$$P(\hat{\alpha}_t = 1) = \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k}$$

α_{t+1} is then the expectation value of $\hat{\alpha}_t$:

$$\begin{aligned} \alpha_{t+1} &= E(\hat{\alpha}_t) \\ &= 1 \cdot P(\hat{\alpha}_t = 1) + 0 \cdot P(\hat{\alpha}_t = 0) \\ &= \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k} f(\alpha_t) \end{aligned}$$

and the evolution equation is

$$\alpha_{t+1} = \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k} \tag{1}$$

Call this function $f(\alpha_t) = \alpha_{t+1}$, examples of which are plotted in Fig. 9. We are looking for the

[24] $\lceil x \rceil$ is the smallest integer $\geq x$.

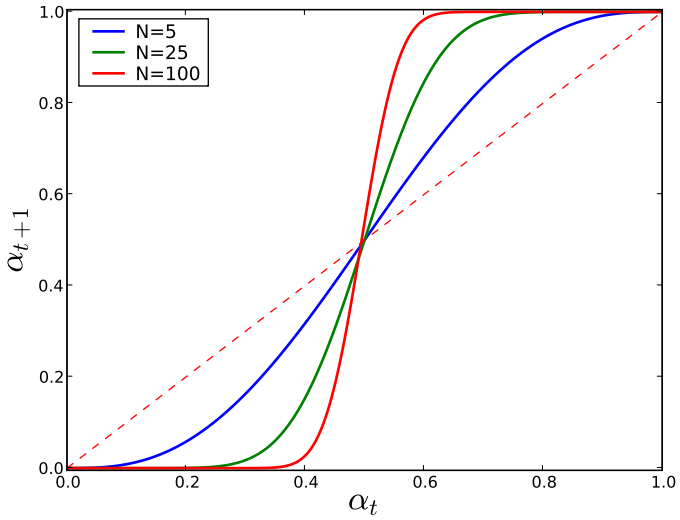


Figure 9: Evolution equation 1 for $N = 5, 25, 100$.

fixed points of f, the partial sum of a binomial distribution, which can be written as a known continuous function, the *regularized incomplete beta function* $I(x; A, B)$:

$$I(x; A, B) = \sum_{j=A}^{A+B-1} \binom{A+B-1}{j} x^j (1-x)^{A+B-1-j} \tag{2}$$

Here we need only the following properties:

- $I(0; A, B) = 0$, $I(1; A, B) = 1$
- $I'(0; A, B) = I'(1; A, B) = 0$
- $I(x; A, B)$ has only one inflection point in $[0, 1]$.

Since $f(x) = I(x; N/2, N/2+1)$, these facts guarantee that f has fixed points at 0, some $x_* \in (0, 1)$, and 1. Since $|f'(0)| < 1$, $|f'(x_*)| > 1$, $|f'(1)| < 1$, they are stable, unstable, and stable.

Since the same reasoning holds for any $r \in (0, 1)$ with a change in the location of x_*, x_* is determined by r and N, and we write $x_*(r, N)$. The location of the three fixed points means that if a population's initial state is $x_0 < x_*(r, N)$, then $x_t \to 0$ (as t increases), while if $x_0 > x_*(r, N)$, then $x_t \to 1$.

Result Stable fixed points at 0 and 1 for any r, N. One unstable interior ($\in (0, 1)$) fixed point $x_*(r, N)$. No bifurcations.

5.1.2 Intraspeaker variation

Suppose learners now probability match: learner i in generation $t+1$ who hears k form 2 and $N-k$ form 1 examples sets $\hat{\alpha}_{i,t} = \frac{k}{N}$. Although the random variable $\hat{\alpha}_t$ corresponding to members of generation $t + 1$ now can have any of $k + 1$ values, by the full connectivity assumption, all that

matters for calculating $\hat{\alpha}_{t+1}$ is $E(\hat{\alpha}_t)$, the average probability of producing form 2 over members of generation t. Defining $\alpha_t \equiv E(\hat{\alpha}_{t-1})$ as in the interspeaker variation case, $\hat{\alpha}_t$ is binomially distributed:

$$P(\hat{\alpha}_t = \frac{k}{N}) = P(k) = \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k}$$

α_{t+1} is then

$$\begin{aligned} \alpha_{t+1} = E(\hat{\alpha}_t) &= \sum_{k=0}^{N} \frac{k}{N} P(\hat{\alpha}_t = \frac{k}{N}) \\ &= \sum_{k=0}^{N} \binom{N}{k} {\alpha_t}^k (1-\alpha_t)^{N-k} \frac{k}{N} \end{aligned} \tag{3}$$

Since the mean of a binomial distribution with parameters α, N is $N\alpha$, (3) simplifies to the evolution equation

$$\alpha_{t+1} = \alpha_t \tag{4}$$

so that *all* $\alpha \in [0,1]$ are fixed points. Since $f'(\alpha) = 1$ for all $\alpha \in [0,1]$, these are neutrally stable fixed points, meaning they neither repel nor attract solutions with nearby initial points.

Result All $\alpha \in [0,1]$ are (neutrally stable) fixed points. No bifurcation.

5.2 Mistransmission

As discussed above (§3.2), many proposed explanations for sound changes are based on mistransmission, in which the form intended by the speaker is different from that perceived by the listener.

We use a simple model of mistransmission errors here: define a and b as *mistransmission probabilities*:

$$a = P(1 \text{ heard} \,|\, 2 \text{ intended}), \quad b = P(2 \text{ heard} \,|\, 1 \text{ intended})$$

For the purposes of our models, all that matters is the probability that form 2 is heard when form 1 is intended (and vice versa). For the N/V case, mistransmission probabilities correspond to the psycholinguistic evidence (§3.2) that (English) speakers mishear iambic, bisyllabic nouns as trochaic more than vice versa, while the opposite pattern holds for bisyllabic verbs.

5.2.1 Interspeaker variation, mistransmission

With mistransmission, the probability that an example from generation t is heard as form 2 is

$$p_{2,t} = \alpha_t(1-a) + (1-\alpha_t)b$$

and α_{t+1} is

$$\begin{aligned} \alpha_{t+1} &= E(\hat{\alpha}_t) \\ &= 1 \cdot P(\hat{\alpha}_t = 1) + 0 \cdot P(\hat{\alpha}_t = 0) \\ &= \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{k} {p_{2,t}}^k (1-p_{2,t})^{N-k} \\ &\equiv f(\alpha_t) \end{aligned} \tag{5}$$

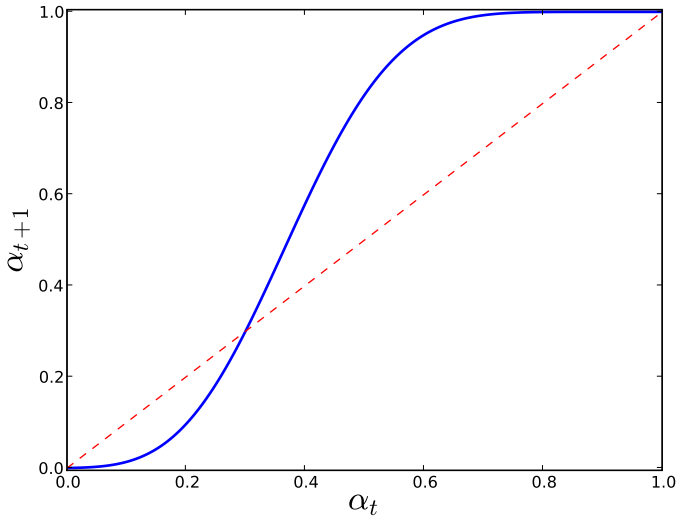


Figure 10: Plot of evolution equation (5) for $N = 15$, $a = 0$, $b = 0.1$. Fixed points at $3.4 \cdot 10^{-5}$ (stable), 0.414 (unstable), 1 (stable).

where

$$f(\alpha) = I(\alpha(1-a) + (1-\alpha)b; \lceil \frac{N}{2} \rceil, \lceil \frac{N+1}{2} \rceil)$$

We assume that $a < 0.5$ and $b < 0.5$, meaning mistransmission probabilities are never worse than chance.[25] With this restriction, there are (by simulation) two cases, based on the values of a, b, N:

1. One fixed point (stable)
2. Three fixed points (stable, unstable, stable).

The fixed point locations cannot be solved for analytically, but we can make some observations:

- $f(0) > 0 \Longleftrightarrow b > 0$ and $f(1) < 1 \Longleftrightarrow a > 0$. If there is *any* mistransmission probability for form 1, 100% form 1 ($\alpha = 0$) is not a stable state. Similarly, if there is any mistransmission probability for form 2, 100% form 2 is not a stable state.
- When $a = 0$, $0 < b < 0.5$, it is not necessary the case that a stable "mostly form 1" state (α near 0) does not exist (Fig. 10). It is true is that for $a = 0$, $0 < b < 0.5$, there is a "mostly form 2" stable state, but it may not be the only one.

In this model asymmetric mistransmission is not enough to explain the complete disappearance of a form. Conversely, if a form has any mistransmission probability, its competing form cannot be eliminated.

Result Either one fixed point (stable) or three fixed points (stable, unstable, stable), locations depend on N, a, b. Bifurcation.

[25] We make this assumption for all models with mistransmission. If $a > 0.5$ or $b > 0.5$, it is possible to have *no* fixed points in $[0, 1]$.

5.2.2 Intraspeaker variation, mistransmission

Now consider the intraspeaker variation case. Defining a and b as above, α_{t+1} is now

$$\begin{aligned}\alpha_{t+1} = E(\hat{\alpha}_t) &= \sum_{k=0}^{N} \frac{k}{N} P(\hat{\alpha}_t = \frac{k}{N}) \\ &= \sum_{k=0}^{N} \binom{N}{k} {p_{2,t}}^k (1-p_{2,t})^{N-k} \frac{k}{N} \\ &= p_{2,t}\end{aligned}$$

(For the last step, the mean of a binomial distribution with parameters $p_{2,t}$, N is $p_{2,t}N$.) The evolution equation is then

$$f(\alpha_t) \equiv \alpha_{t+1} = \alpha_t(1-a) + (1-\alpha_t)b \tag{6}$$

Since $f(\alpha_t)$ is a line, $f(0) = b < 0.5$, and $f(1) = (1-a) > 0.5$, there is just one fixed point (stable) at $\alpha_* = \frac{b}{a+b}$, found by solving $f(\alpha_*) = \alpha_*$. Note that

1. The location of α_* does not depend on N.
2. Unlike for interspeaker variation with mistransmission (§ 5.2), when there is no probability of mistransmission of a form ($a = 0$), its competitor form is eliminated in the stable state ($\alpha_* = 1$)

Result One fixed point (stable) at $\frac{b}{a+b}$. No bifurcations.

5.3 Summary: Interspeaker vs. intraspeaker variation models

Up to this point, we have considered both intraspeaker and interspeaker variation models. For the Base (§5.1) and Mistransmission (§5.2) models, the dynamics of the intraspeaker and interspeaker variation models are qualitatively different, either because (a) variation between forms was eliminated in the interspeaker case and not in the intraspeaker case (b) the interspeaker case shows a bifurcation, while the intraspeaker case does not. This result illustrates a simple but important point: the type of variation assumed in modeling a linguistic population profoundly affects model dynamics.

From here on we only consider intraspeaker variation models, motivated by the evidence that our N/V case study shows intraspeaker variation (§2.2) and by the general goal of understanding when variation within individuals can lead to change at the population level (§1.1).

5.4 Poisson input, default strategies

As discussed above (§4.2.1), the assumption of fixed input is often unrealistic.

In Poisson input, we assume each learner receives N examples, where N is Poisson-distributed with mean λ:

$$P(N) = \frac{e^{-\lambda}N^{\lambda}}{N!}$$

A learner now can receive *no* examples with non-zero probability ($e^{-\lambda}$), and must have a *default strategy* for this case.[26] The default strategy turns out to have a significant influence on system

[26] More generally, a default strategy is needed for any learning algorithm where a learner may receive no informative examples, as in the "discarding" learners discussed later on.

dynamics, especially for small λ (when the chance of receiving no examples is non-negligible). We assume a simple default strategy: set $\hat{\alpha} = r$, where $r \in [0, 1]$.

5.4.1 Mistransmission, Poisson input

We introduce Poisson input for the intraspeaker case with mistransmission, which turns out to be solvable analytically.

Define a, b, $p_{2,t}$, α_t, $\hat{\alpha}_t$ as above (§5.2.2). Each learner now receives $N \sim \text{Poisson}(\lambda)$ examples, of which k are heard as form 2, and sets

$$\hat{\alpha} = \begin{cases} r & : N = 0 \\ \frac{k}{N} & : N \neq 0 \end{cases}$$

Then

$$\begin{aligned} \alpha_{t+1} = E[\hat{\alpha}_t] &= rP(N=0) + \sum_{N=1}^{\infty} P(N) \sum_{k=0}^{N} \frac{k}{N} P(k \mid N) \\ &= re^{-\lambda} + \sum_{N=1}^{\infty} \frac{e^{-\lambda} N^{\lambda}}{N!} \sum_{k=0}^{N} \frac{k}{N} \binom{N}{k} {p_{2,t}}^k (1 - p_{2,t})^{N-k} = e^{-\lambda} r + e^{-\lambda} \sum_{N=1}^{\infty} p_{2,t} \frac{N^{\lambda}}{N!} \\ &= re^{-\lambda} + p_{2,t}(1 - e^{-\lambda}) \end{aligned}$$

where the last step uses $\sum_{N=0}^{\infty} \frac{N^{\lambda}}{N!} = e^{\lambda}$.

Substituting in for $p_{2,t}$ and simplifying, the evolution equation is

$$\alpha_{t+1} = b + (r - b)e^{-\lambda} + \alpha_t(1 - a - b)(1 - e^{-\lambda}) \tag{7}$$

Setting $\alpha_{t+1} = \alpha_t$ in (7) gives the fixed point

$$\alpha_*(a, b, \lambda) = \frac{r + b(e^{\lambda} - 1)}{1 + (a + b)(e^{\lambda} - 1)} \tag{8}$$

Fig. 11 plots $\alpha_*(a, b, \lambda)$ for $b > a$. Note that

- As $\lambda \to \infty$, $\alpha_* \to \frac{b}{a+b}$, the fixed point for the fixed input case.
- As $\lambda \to 0$, $\alpha_* \to r$, the default value.
- As suggested by Fig. 11, there is only one inflection point of $\alpha_*(a, b, \lambda)$ considered as a function of λ; taking derivatives of (8) shows it is at

$$\lambda_C = \ln(\frac{1}{a + b} - a)$$

  Roughly, for $\lambda < \lambda_C$, the dynamics are determined by the default strategy, and for $\lambda > \lambda_C$ they are determined by mistransmission probabilities.

Result One fixed point (stable), no bifurcation.

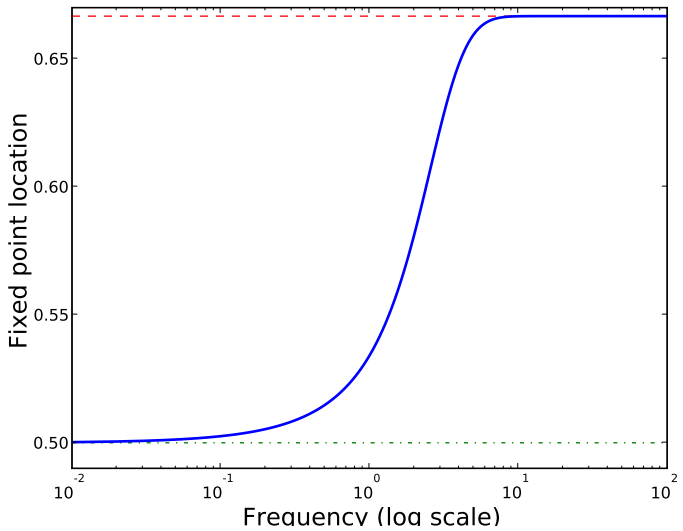


Figure 11: Plot of (8), fixed point location vs. mean frequency (log scale), with $a = 0.05$, $b = 0.1$, $r = 0.5$. Non-solid lines are r, $\frac{a}{a+b}$.

5.5 Discarding

We have assumed so far that each learner hears N examples, every one of which is heard as form 1 or form 2. However, it seems plausible that learners might discard some examples, for example sentences with several possible parses, words where the distinction between primary and secondary stress is not clear, or non-native speech. We call these examples *discarded*, and the case where learners have a chance of receiving such examples *discarding*.[27] What are the consequences of discarding on the dynamics for the types of learners considered so far?

5.5.1 Discarding, fixed input

Consider first the simplest case: no mistransmission and fixed input (N examples). Assume the probabilities form 1 or form 2 examples are discarded are constant between generations, and define

$$r_1 = P(\text{discarded} \,|\, 1 \text{ intended}), \quad r_2 = P(\text{discarded} \,|\, 2 \text{ intended})$$

Let $p_{2,t}$, $p_{1,t}$ be the probabilities an example produced by generation t is heard as form 2 or form 1; the probability it is discarded is then $p_{3,t} = 1 - p_{2,t} - p_{1,t}$. Let k_2, k_1 be the number of form 2 and form 1 examples heard; these are random variables, as is $N - k_1 - k_2$, the number of discarded examples heard. As above, α_t is the probability form 2 is intended in an example from generation

[27] This concept is related to "input filtering" of the sort considered in a series of computational studies by L. Pearl and collaborators (Pearl, 2007; Pearl and Weinberg, 2007; Pearl, 2008), but sufficiently different that we have chosen the term "discarding" instead. Pearl has shown that certain parametric systems can be learned by learners when the parameters are acquired in some orders, but not others. In this sense, "input filtering" to only consider input relevant for setting certain parameters at each stage of the learning process is what makes learning feasible. Input filtering is a form of discarding: learners do not consider input which is uninformative about the parameter they are currently setting.

t. Then

$$p_{2,t} = \alpha_t(1-r_2), \quad p_{1,t} = (1-\alpha_t)(1-r_1), \quad p_{3,t} = 1 - p_{1,t} - p_{2,t}$$

Learners set

$$\hat{\alpha}_t = \begin{cases} r & : k_1 + k_2 = 0 \\ \frac{k_2}{k_1+k_2} & : k_1 + k_2 \neq 0 \end{cases} \tag{9}$$

where the default strategy when all examples are discarded is to set $\hat{\alpha}_t = r$ for some (fixed) $r \in [0,1]$. α_{t+1} is then

$$\begin{aligned} \alpha_{t+1} = E[\hat{\alpha}_t] &= P(k_1+k_2=0)r + P(k_2 \neq 0, k_1 = 0) \cdot 1 + \sum_{k_2=1}^{N} \sum_{k_1=1}^{N-k_2} P(k_1,k_2) \frac{k_2}{k_1+k_2} \\ &= (1-p_{1,t}-p_{2,t})^N r + \sum_{k_2=1}^{N} \binom{N}{k_2} {p_{2,t}}^{k_2} (1-p_{1,t}-p_{2,t})^{N-k_2} \\ &\quad + \sum_{k_2=1}^{N} \sum_{k_1=1}^{N-k_2} \binom{N}{k_1,k_2} {p_{1,t}}^{k_1} {p_{2,t}}^{k_2} (1-p_{1,t}-p_{2,t})^{N-k_1-k_2} \frac{k_2}{k_1+k_2} \\ &\vdots \\ &= \frac{p_{2,t}}{p_{1,t}+p_{2,t}} (1-(1-p_{1,t}-p_{2,t})^N) + (1-p_{1,t}-p_{2,t})^N r \end{aligned}$$

where the last step is shown in App. E.1. The evolution equation is then

$$\alpha_{t+1} = (1-{p_{3,t}}^N) \frac{p_{2,t}}{p_{1,t}+p_{2,t}} + {p_{3,t}}^N r \tag{10}$$

Call Eqn. 10 $g_N(\alpha_t)$. We examine the dynamics of $g_N(\alpha_t)$ as a function of N.

Case 1: Large N For N large, ${p_{3,t}}^N \to 0$ and (10) reduces to

$$g_\infty(\alpha_t) = \alpha_{t+1} = \frac{\alpha_t(1-r_2)}{(1-r_1)+\alpha_t(r_1-r_2)} \tag{11}$$

Setting $\alpha_{t+1} = \alpha_t$ gives two fixed points $\alpha_- = 0$, $\alpha_+ = 1$, and differentiating (11) gives

$$\left.\frac{\partial \alpha_{t+1}}{\partial \alpha_t}\right|_{\alpha_t=0} = \frac{1-r_2}{1-r_1}, \qquad \left.\frac{\partial \alpha_{t+1}}{\partial \alpha_t}\right|_{\alpha_t=1} = \frac{1-r_1}{1-r_2}$$

There is a bifurcation at $r_1 = r_2$: for $r_1 < r_2$, α_- is stable, α_+ is unstable; for $r_2 < r_1$, α_- is unstable, α_+ is stable. Intuitively, the form with a higher probability of being discarded is eliminated.

Eqn. 11 is plotted for both cases in Fig. 12.

Case 2: Finite N For finite N, note that

$$g_N(\alpha_t) - g_\infty(\alpha_t) = {p_{3,t}}^N \left(r - \frac{p_{2,t}}{p_{1,t}+p_{2,t}}\right)$$

Since $\frac{p_{2,t}}{p_{1,t}+p_{2,t}}$ increases monotonically with α_t, $g_N(\alpha_t) > g_\infty(\alpha_t)$ for $\frac{p_{2,t}}{p_{1,t}+p_{2,t}} < r$ and $g_N(\alpha_t) < g_\infty(\alpha_t)$ for $\frac{p_{2,t}}{p_{1,t}+p_{2,t}} > r$. There is thus now only one fixed point (stable). Essentially, as N increases, the fixed-point plot quickly looks increasingly like a bifurcation at the $r_1 = r_2$ line (the large N case); this is illustrated in Fig. 15, discussed below.

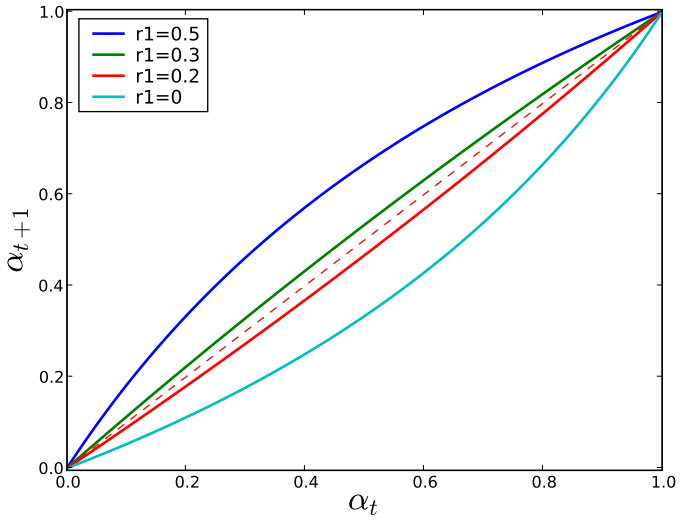


Figure 12: Evolution equation (11) plotted with r_1 varied, $r_1 + r_2 = 0.5$.

Result One stable, interior fixed point for finite N, bifurcation at $r_1 = r_2$ as $N \to \infty$.

5.5.2 Discarding, mistransmission, fixed input

Now consider the case with both discarding and mistransmission. Define a, b, r_1, r_2 as follows, with H=heard, I=intended:

$$P(H = 1 \,|\, I = 1) = 1 - b, \quad P(H = 2 \,|\, I = 1) = br_1, \quad P(\text{discarded} \,|\, I = 1) = b(1 - r_1)$$
$$P(H = 2 \,|\, I = 2) = 1 - a, \quad P(H = 1 \,|\, I = 2) = ar_2, \quad P(\text{discarded} \,|\, I = 2) = a(1 - r_2)$$

a, b now denote "$H \neq I$", while r_1 and r_2 determine how likely a mistransmitted example (heard, but as the wrong form) is versus a discarded example (not heard). We now have

$$p_{2,t} = \alpha_t(1 - a) + (1 - \alpha_t)br_1, \quad p_{1,t} = \alpha_t ar_1 + (1 - \alpha_t)(1 - b), \quad p_{3,t} = 1 - p_{1,t} - p_{2,t} \tag{12}$$

The evolution equation is (10),

$$\alpha_{t+1} = (1 - {p_{3,t}}^N)\frac{p_{2,t}}{p_{1,t} + p_{2,t}} + {p_{3,t}}^N r \tag{13}$$

but now with $p_{1,t}$, $p_{2,t}$, $p_{3,t}$ from (12).

Case 1: Large N The evolution equation simplifies to (11). Substituting from (12), and solving for $\alpha_{t+1} = \alpha_t$ gives two fixed points:

$$\alpha^*_{\pm} = \frac{a(1 - r_2) - b(1 - r_2) - (br_1 + ar_2) \pm \sqrt{(a - b)^2 + 4abr_1r_2}}{2[a(1 - r_2) - b(1 - r_1)]} \tag{14}$$

By simulation, for a,b,r_1, r_2 given, exactly one of the two is stable and $\in [0, 1]$. Eqn. 14 is too large to give insight into the fixed points' behavior. Consider instead two special cases:

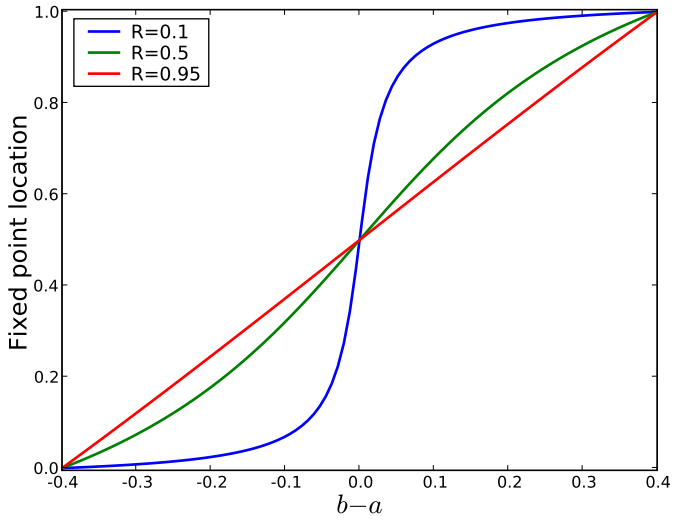


Figure 13: Location of α_* vs. $b-a$, $a+b=0.5$, for different values of R, from (14)

1. $a=b$: the equilibrium points are

$$\alpha_\pm^* = \sqrt{r_2}\frac{\sqrt{r_2} \mp \sqrt{r_1}}{r_2 - r_1}$$

   In this case, despite symmetric mistransmission errors the fixed point is an interior point.

2. $r_1 = r_2 = R$: the equilibrium points are

$$\alpha_\pm^* = \frac{(b-a)(1-R) - R(a+b) \pm \sqrt{(b-a)^2 + 4abR^2}}{2(1-R)(b-a)} \tag{15}$$

   Call the stable fixed point $\alpha_*(a,b,R)$. When $R=0$, only discarding occurs, and changing $a-b$ is equivalent to changing the discarding parameter difference $r_1 - r_2$, giving a bifurcation at $a=b$ (§ 5.5.1). When $R=1$, only mistransmission occurs, and α_* changes smoothly with $a-b$ (§ 5.2.2). Thus, plotting α_* against $a-b$ shows how "bifurcation like" α_* behaves for a given value of R (Fig. 13).

Case 2 : Finite N As in the fixed-input case (§ 5.5.1), the finite N and large-N dynamics are qualitatively similar: there is one stable fixed point, and its exact location depends now on N as well as a, b, R.

Result One stable fixed point in $[0,1]$, location depends on a, b, r_1, r_2. Bifurcation at $a=b$ when $r_1 = r_2 = 0$. As r_1, r_2 are increased, fixed point location as a function of $a-b$ becomes smoother (less "bifurcation-like").

5.5.3 Discarding, mistransmission, Poisson input

Define $p_{1,t}$, $p_{2,t}$, $p_{3,t}$, a, b, r_1, r_2 as above, and let each learner hear $N \sim \text{Poisson}(\lambda)$ examples, of which k_2 are form 2, k_1 are form 1, and sets $\hat{\alpha}$

$$\hat{\alpha}_t = \begin{cases} r & : k_1 + k_2 = 0 \text{ or } N = 0 \\ \frac{k_2}{k_1+k_2} & : k_1 + k_2 \neq 0 \end{cases}$$

Then we can show (App. E.2) that

$$E[\hat{\alpha}_t] = \frac{p_{2,t}}{p_{1,t} + p_{2,t}} + e^{-\lambda p_{3,t}}(r - \frac{p_{2,t}}{p_{1,t} + p_{2,t}})$$

Equating $E[\hat{\alpha}_t]$ with α_{t+1} and re-arranging, the evolution equation is

$$\alpha_{t+1} = \frac{p_{2,t}}{p_{1,t} + p_{2,t}}(1 - (e^{-p_{3,t}})^\lambda) + r(e^{-p_{3,t}})^\lambda \tag{16}$$

Compare the fixed-input evolution equation:

$$\alpha_{t+1} = \frac{p_{2,t}}{p_{1,t} + p_{2,t}}(1 - {p_{3,t}}^\lambda) + r{p_{3,t}}^\lambda \tag{17}$$

The functions $p_{3,t}^N$ and $e^{-p_3 N}$ are both $\in [0, 1]$, but are rough opposites: $p_{3,t}^N$ attains its minimum at $p_{3,t} = 0$, its maximum at $p_{3,t} = 1$, and has positive slope in $[0, 1]$. ${e^{-p_{3,t}}}^N$ attains its maximum at $p_{3,t} = 0$, its minimum at $p_{3,t} = 1$, and has negative slope in $[0, 1]$.

This means that the Poisson and fixed-input cases show different frequency-dependent behavior: we expect the transition to bifurcation-like behavior to occur at higher λ in the Poisson case than in the fixed-input case.[28]

We illustrate by an example for the discarding-only case where $p_{3,t}$ is small but non-zero.

Example: Suppose there are discarding probabilities a, b, but no mistransmission ($r_1 = r_2 = 0$), such that at t,

$$p_{1,t} = (1 - \alpha_t)(1 - b), \quad p_{2,t} = \alpha_t(1 - a), \quad p_{3,t} = \alpha_t(a - b) + b$$

In this case, the Poisson-input evolution equation is (from Eqn. 16)

$$\alpha_{t+1} = \frac{\alpha_t(1 - a)}{\alpha_t(b - a) + 1 - b}(1 - e^{-\lambda(\alpha_t(b-a)+1-b)}) + re^{-\lambda(\alpha_t(b-a)+1-b)} \tag{18}$$

and the fixed-input evolution equation is

$$\alpha_{t+1} = \frac{\alpha_t(1 - a)}{\alpha_t(b - a) + 1 - b}(1 - (\alpha_t(b - a) + 1 - b)^\lambda) + r(\alpha_t(b - a) + 1 - b)^\lambda \tag{19}$$

By simulation, (18) and (19) each have a unique, stable fixed point, plotted in Figs. 14, 15 as a function of $b - a$ with $a + b$ fixed (at 0.25) and $r = 0.5$. In the Poisson case, the dynamics are

[28] Intuitively, this is because assuming $p_{3,t}$ evaluated near α_* is small, the second term of (16) goes to 0 as λ is increased slower than the second term of (17).

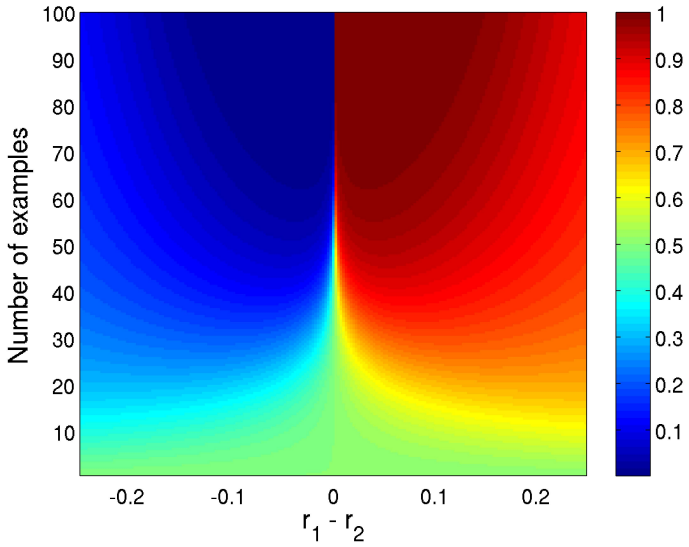


Figure 14: Pseudocolor plot of fixed point location for the Poisson case (18) vs. λ and $b - a$, with $a + b = 0.25$. Note the different y-axis scale in Fig. 15

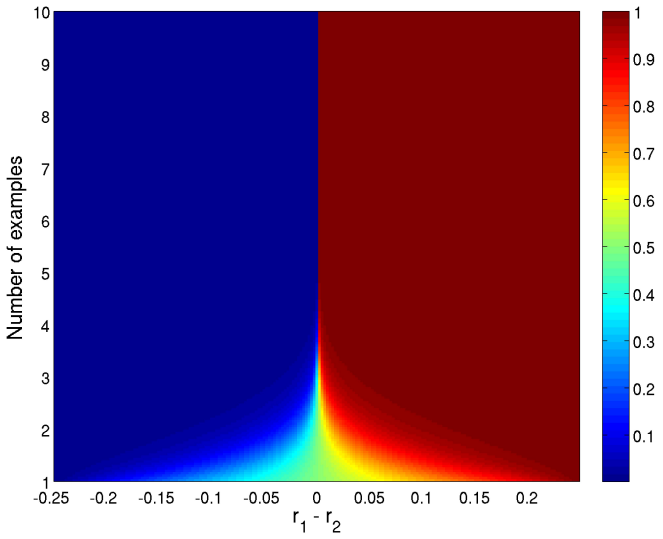


Figure 15: Pseudocolor plot of fixed point location for the fixed-input case (19) vs. λ and $b - a$, with $a + b = 0.25$. Note the different y-axis scale in Fig. 14

frequency-dependent. For N small, there is essentially stable variation near r (0.5). As N increases, the dynamics look increasingly like a bifurcation at $b = a$ between fixed points near 0 and 1. This transition to bifurcation-like behavior is still not complete when $\log(N)$ increases by 3.

In the fixed-input case, there is essentially no frequency dependence: the dynamics at any $N \geq 2$ look like a bifurcation at $b = a$ between fixed points very near 0 and 1.

In both the Poisson and fixed-input cases, the transition to bifurcation-like behavior as N increases is faster for smaller $a + b$.

□

Beyond this example, it turns out the fixed-point profile for the Poisson-input case (16) is qualitatively the same as for the fixed-input (17) case.

Result One stable fixed point, location depends on a, b, r_1, r_2, λ. Transition to bifurcation-like behavior is now dependent on frequency (λ) as well as r_1, r_2.

5.6 Interpretation

This section has considered the effect of mistransmission, discarding, and varying input type (Poisson vs. fixed) on the population-level dynamics. To summarize, we have found:

- Adding mistransmission shifts fixed-point locations, but does not introduce bifurcations.
- Adding discarding introduces bifurcations, or a transition to bifurcation-like behavior as other system parameters are varied.
- Using Poisson input, there is a sharp transition as a function of λ between dynamics reflecting the default strategy (low λ) and dynamics similar to the fixed-input case (high λ).

All models considered in this section are summarized in Table 7.

6 Models II: Individual forms, biased learners

We have explored the possibility that learners discard some examples in setting $\hat{\alpha}_t$, the probability of producing form 2 (instead of form 1). Put otherwise, learners only make an inference about $\hat{\alpha}_t$ based on a subset of their input data.

Another property a proposed learner could have is *regularization* (see 6.1) a bias of some sort towards setting $\hat{\alpha} = 0$ or $\hat{\alpha}_t = 1$ when not all examples were heard as one form or the other. In this case, it is the learner's inference procedure itself which is biased, rather than the input they receive.

To examine the effects of regularization on the dynamics, an explicit model of how learners regularize must be given. We consider the effects of three implementations of regularization on the dynamics of the types of models considered so far.

6.1 Regularization I: Thresholding

We consider a simple implementation of regularization (*thresholding*): upon receiving k form 2 and $N - k$ form 1 examples, a learner sets

$$\hat{\alpha}_t = \begin{cases} 0 & : \frac{k}{N} < \epsilon_1 \\ 1 & : \frac{k}{N} > 1 - \epsilon_2 \\ \frac{k}{N} & : \text{otherwise} \end{cases} \tag{20}$$

where $\epsilon_1, \epsilon_2 \in [0, 1]$ and $\epsilon_1 \leq 1 - \epsilon_2$. Note that because learners can now probability match or not depending on the examples received, there are no longer distinct intraspeaker-variation and interspeaker-variation cases. Otherwise, what is the impact of thresholding on the types of models considered so far?

6.1.1 Fixed input, no mistransmission, no discarding

First consider the fixed-input case with no mistransmission or discarding. $E[\hat{\alpha}_t]$ is

$$\begin{aligned} E[\hat{\alpha}_t] &= P(k < \epsilon_1 N) \cdot 0 + P(k > \epsilon_2 N) \cdot 1 + \sum_{\epsilon_1 N \leq k \leq (1-\epsilon_2)N} P(k) \frac{k}{N} \\ &= \sum_{(1-\epsilon_2)N < k \leq N} \binom{N}{k} {\alpha_t}^k (1 - \alpha_t)^{N-k} + \sum_{\epsilon_1 N \leq k \leq (1-\epsilon_2)N} \binom{k}{N} {\alpha_t}^N (1 - \alpha_t)^{k-N} \frac{k}{N} \end{aligned}$$

After some algebra, the evolution equation is

$$\alpha_{t+1} = \alpha_t + \sum_{k < \epsilon_2 N} \binom{k}{N} {\alpha_t}^N (1 - \alpha_t)^{k-N} - \sum_{k < \epsilon_1 N} \binom{k}{N} {\alpha_t}^N (1 - \alpha_t)^{k-N} \tag{21}$$

Fig 16 shows examples of (21). It is clear that 0 and 1 are fixed points of (21). Considering the first and second derivatives of (16), we find that more generally,

- $\epsilon_1 = 0$, $\epsilon_2 > 0$: Fixed points 0 (unstable), 1 (stable)
- $\epsilon_1 > 0$, $\epsilon_2 = 0$: Fixed points 0 (stable), 1 (unstable)

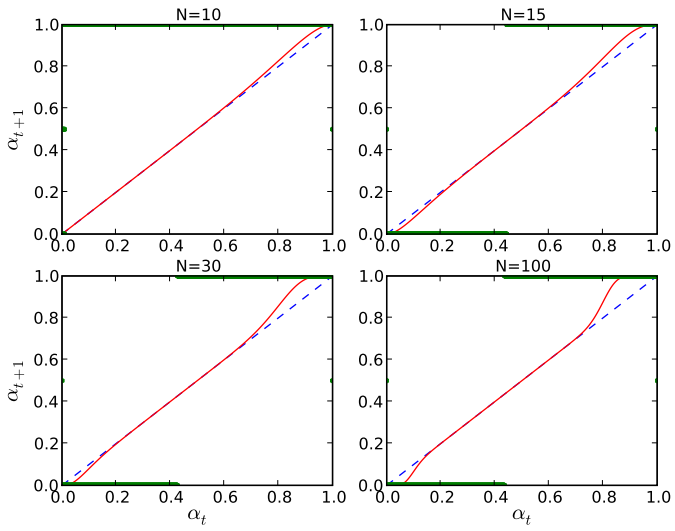


Figure 16: Evolution Eqn. 21, varying N for $\epsilon_1 = 0.1$, $\epsilon_2 = 0.2$.

- $\epsilon_1 > 0$, $\epsilon_2 > 0$: Fixed points 0 (stable), $\alpha_*(\epsilon_1, \epsilon_2)$ (unstable), 1 (stable), where $\alpha_* \in (0, 1)$.

($\epsilon_1 = \epsilon_2 = 0$ is the intraspeaker variation case from § 5.1.2) There are thus bifurcations on the lines $\epsilon_1 = 0$ and $\epsilon_2 = 0$. Qualitatively, when $\epsilon_1, \epsilon_2 > 0$, around $\alpha_*(\epsilon_1, \epsilon_2)$ there is a region where $\alpha_{t+1} \approx \alpha_t$, i.e. the fixed-input, non-thresholding case from § 5.1.2. As N increases, this region sharpens and the stable fixed points become "stronger" in the sense that nearby trajectories reach them in fewer generations.

6.1.2 Fixed input, mistransmission

Now hold the learner's algorithm constant, but introduce mistransmission probabilities a, b. By the same steps as above, the evolution equation becomes

$$\alpha_{t+1} = p_{2,t} + \sum_{k<\epsilon_2 N} \binom{k}{N} {p_{2,t}}^N (1 - p_{2,t})^{k-N} - \sum_{k<\epsilon_1 N} \binom{k}{N} {p_{2,t}}^N (1 - p_{2,t})^{k-N} \tag{22}$$

where $p_{2,t} = \alpha_t(1 - a) + (1 - \alpha_t)b$ is the probability of an example being heard as form 2 at t. The dynamics are now more interesting and complicated: sample plots, with N, ϵ_1, ϵ_2 held fixed, are shown in Fig. 17. By simulation, thresholding introduces two new inflection points. Along with the fact that α_{t+1} is a monotonically increasing function of α_t, there are now three possible fixed point profiles,[29] with corresponding parameter regions:[30]

- Tristability (Upper left): Stable fixed points 0, α_*, 1.
- Bistability: Stable fixed points 0 and 1 (upper right), stable fixed points 0 and α_* (lower right) or 1 and α_* (not shown).

[29] That there are three possible fixed-point profiles follows from having three inflection points; that all three occur when ϵ_1, ϵ_2 are varied was confirmed by simulation.

[30] Where α_* is an interior point $\alpha_*(\epsilon_1, \epsilon_2, a, b) \in (0, 1)$

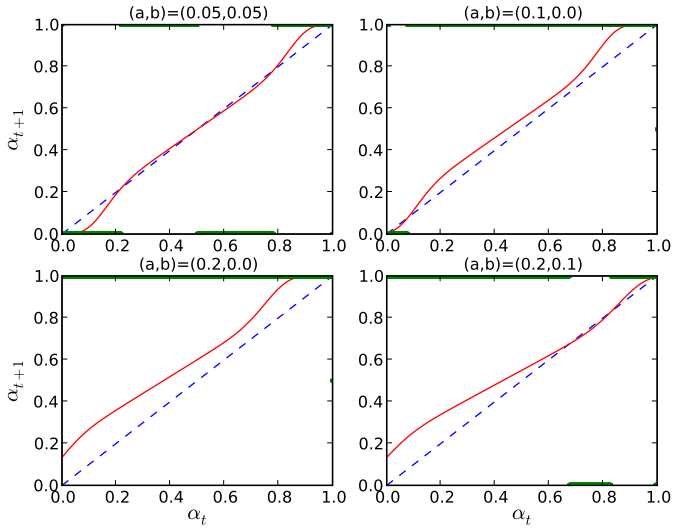


Figure 17: Evolution Eqn. 22, varying N for $\epsilon_1 = 0.1$, $\epsilon_2 = 0.2$.

- Monostability: Stable fixed point 0 (lower left), 1, or α_* (not shown)

Results One, two, or three stable fixed points, with bifurcations between these cases. In the bistable case at least one of 0 and 1 is a stable fixed point, in the tristable case both are.

6.1.3 Poisson input

Now assume each learner hears $N \sim \text{Poisson}(\lambda)$ examples, of which k are form 2 and $N - k$ form 1 examples, and sets:

$$\hat{\alpha}_t = \begin{cases} r & : N = 0 \\ 0 & : N > 0,\ \frac{k}{N} < \epsilon_1 \\ 1 & : N > 0,\ \frac{k}{N} > 1 - \epsilon_2 \\ \frac{k}{N} & : \text{otherwise} \end{cases} \tag{23}$$

No mistransmission $E[\hat{\alpha}_t]$ is

$$\begin{aligned} E[\hat{\alpha}_t] &= P(N=0)\cdot r + \sum_{N=0}^{\infty} P(N)[\sum_{k>(1-\epsilon_2)N} P(k\,|\,N) + \sum_{\epsilon_1 N \le k \le (1-\epsilon_2)N} P(k\,|\,N)\frac{k}{N}] \\ &= e^{-\lambda} r + \sum_{N=0}^{\infty} \frac{\lambda^N e^{-\lambda}}{N!}[\sum_{k>(1-\epsilon_2)N} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} \\ &\quad + \sum_{\epsilon_1 N \le k \le (1-\epsilon_2)N} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} \frac{k}{N}] \end{aligned}$$

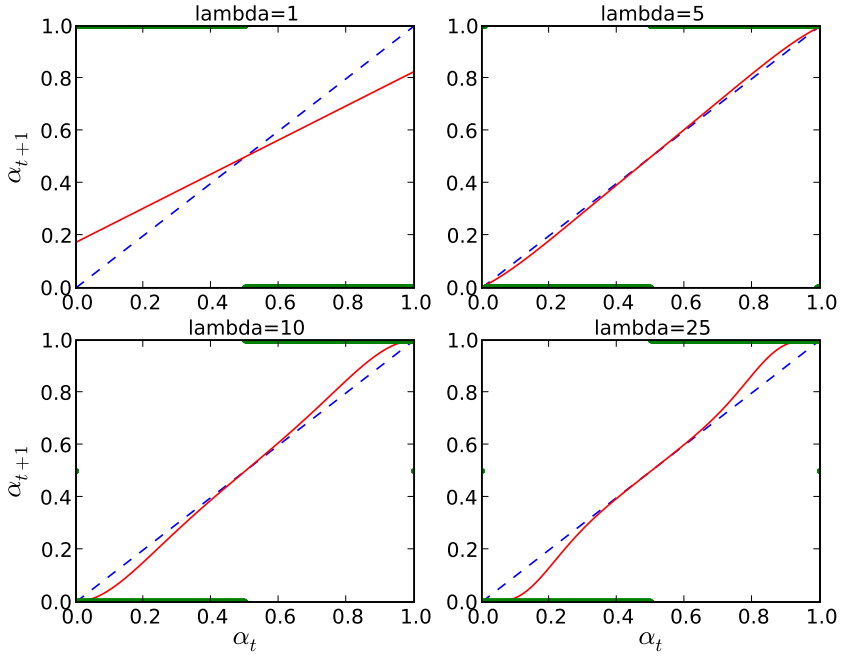


Figure 18: Evolution Eqn. 24, varying λ for $\epsilon_1 = \epsilon_2 = 0.2$, $r = 0.5$.

The evolution equation is

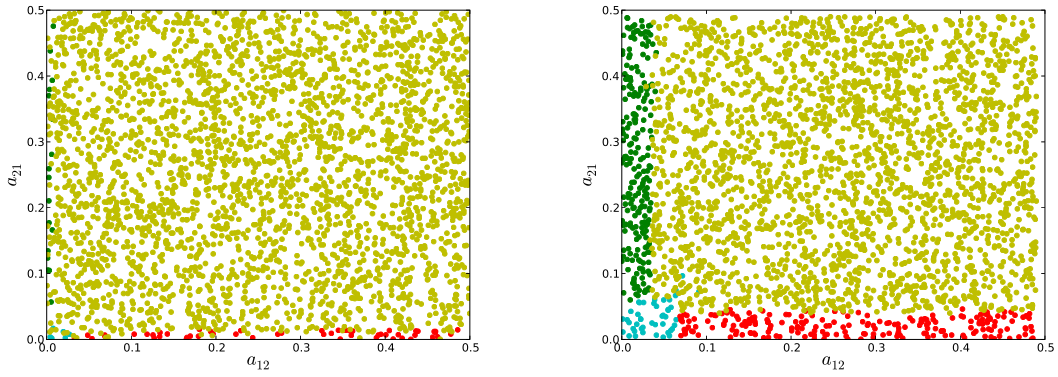
$$\alpha_{t+1} = e^{-\lambda} r + \sum_{N=0}^{\infty} \frac{\lambda^N e^{-\lambda}}{N!} [\sum_{k>(1-\epsilon_2)N} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} + \sum_{\epsilon_1 N \leq k \leq (1-\epsilon_2)N} \binom{N}{k} \alpha_t^k (1-\alpha_t)^{N-k} \frac{k}{N}] \tag{24}$$

a weighted sum over (21) for $N > 0$ with a flat ($\alpha_{t+1} = r$) contribution for $N = 0$. When ϵ_1 and ϵ_2 are fixed, as $\lambda \to 0$, Eqn. 24 becomes a flat line, and (by simulation), as λ increases (past $\lambda \approx 5 - 10$), (24) looks like (21): three inflection points and stable fixed points near 0 and 1. For fixed ϵ_1 and ϵ_2, there is thus a bifurcation (from 1 to 2 stable fixed points) as λ is varied, illustrated in Fig. 18.

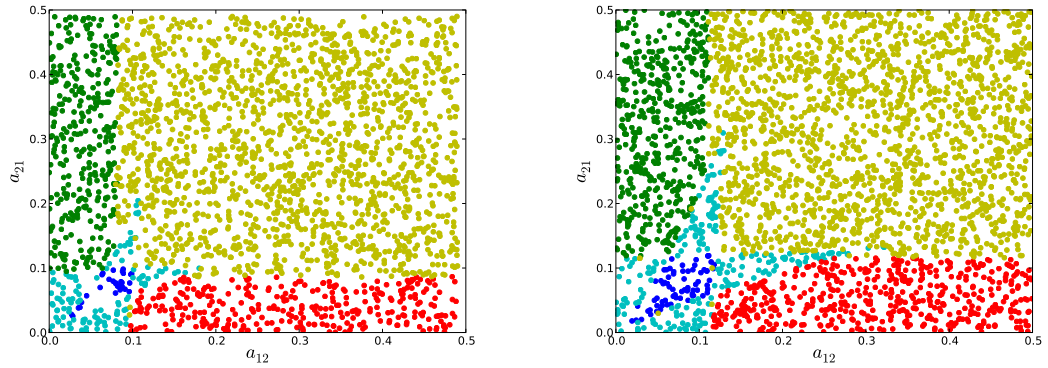
Mistransmission The evolution equation is now

$$\alpha_{t+1} = e^{-\lambda} r + \sum_{N=0}^{\infty} \frac{\lambda^N e^{-\lambda}}{N!} [\sum_{k>(1-\epsilon_2)N} \binom{N}{k} p_{2,t}^k (1-p_{2,t})^{N-k} + \sum_{\epsilon_1 N \leq k \leq (1-\epsilon_2)N} \binom{N}{k} p_{2,t}^k (1-p_{2,t})^{N-k} \frac{k}{N}] \tag{25}$$

where $p_{2,t} = \alpha_t(1-a) + (1-\alpha_t)b$ is the probability of an example being heard as form 2 at t. By simulation, the dynamics are as expected: there can be one, two, or three stable fixed points; as



(a) $\lambda = 5$ (b) $\lambda = 10$



(c) $\lambda = 25$ (d) $\lambda = 50$

Figure 19: Phase diagrams in (a, b), varying λ, for evolution Eqn. 25, $\epsilon_1 = \epsilon_2 = 0.2$ fixed. Regions: one stable FP < 0.01 (red), > 0.99 (green), in $(0.01, 0.99)$ (yellow), two stable FPs (cyan), three stable FPs (blue). See text.

λ increases, a learner's change of drawing $N = 0$ decreases exponentially, and the dynamics look like the fixed-input case (22); as $\lambda \to 0$, the dynamics look increasingly like the line $\alpha_{t+1} = r$. To see the effects of mistransmission and λ, Fig. 6.1.3 shows phase diagrams in (a, b) space for several values of N, when ϵ_1, ϵ_2 are fixed.

Four qualitative regions are shown:

1. Monostable: One stable FP near 0 or 1.
2. Monostable with variation: One stable, interior FP.
3. Bistable: Two stable FPs, separated by an unstable FP.
4. Tristable: Three stable FPs, separated by two unstable FPs.

Bifurcations occur between regions with different numbers of fixed points, and as $\lambda \to 0$, region 2 expands to fill the whole space.

Mistransmission, discarding The dynamics in this case are similar to the previous one (mistransmission, no discarding) and are omitted for brevity.

Results Qualitatively similar to the fixed-input case: one, two, or three stable fixed points, with bifurcations between these cases. However, the parameter regions now depend on λ as well as $a, b, \epsilon_1, \epsilon_2$.

6.1.4 Discussion

Adding thresholding has striking results. In the simplest model, if *any* thresholding towards one of the endpoints is introduced ($\epsilon_i > 0$), that endpoint becomes a stable state, so that for $\epsilon_1 > 0$ and $\epsilon_2 > 0$, both 0 and 1 are stable. When mistransmission is added, tristable states, as well as additional bistable states, are possible. We thus have multistability, which we have argued is observed frequently in the N/V data, for even very small thresholding biases.

When mistransmission is added, a stable interior point $\alpha_* \in (0, 1)$, i.e. stable variation, can coexist with 0, 1, or both as fixed points. We can thus have bifurcations from an endpoint to stable variation or vice versa, which we have argued is observed frequently in the N/V data.[31]

Finally, the dynamics of both fixed-input and Poisson-input models are frequency dependent.

6.2 Regularization II: frequency boosting

As motivation for another model of regularization, consider the work on "frequency boosting" in artificial grammar learning discussed above (§6.1)

Our thresholding model does not describe this sort of regularization, where learners seem to be *reducing* variation rather than eliminating it (when below some threshold). We describe a simple frequency-boosting learning algorithm and its effect on system dynamics.

6.2.1 Frequency boosting as weighting

Say learners receive N examples, of which k are form 2, then weight their observed probability k/N towards an endpoint determined by some threshold. That is, given $r, w_1, w_2 \in (0, 1)$, the learner sets

$$\begin{cases} \frac{k}{N}(1 - w_1) & : \frac{k}{N} \leq r \\ \frac{k}{N}(1 - w_2) + w_2 & : \frac{k}{N} > r \end{cases}$$

For simplicity, assume $r = 0.5$ (the learner weights towards the nearest endpoint) and $w_1 = w_2 = w$. We examine the fixed and Poisson-input cases.

Fixed input $E[\hat{\alpha}_t]$ is

$$\begin{aligned} E[\hat{\alpha}_t] &= \sum_{k \leq \frac{N}{2}} \binom{N}{k} {\alpha_t}^k (1 - \alpha_t)^{N-k} \frac{k}{N}(1 - w) + \sum_{k > \frac{N}{2}} \binom{N}{k} {\alpha_t}^k (1 - \alpha_t)^{N-k} [\frac{k}{N}(1 - w) + w] \\ &= \alpha_t + w \sum_{k > \frac{N}{2}} \binom{N}{k} {\alpha_t}^k (1 - \alpha_t)^{N-k} \end{aligned}$$

[31] For one of N or V at a time.

Assume N is even; then we can rewrite the sum to get the evolution equation

$$\alpha_{t+1} = \alpha_t(1-w) + wI(\alpha_t; \frac{N}{2}+1, \frac{N}{2}) \tag{26}$$

where I is the incomplete beta function. By the same reasoning as in §5.1.1, Eqn. 26

- Has only one inflection point in $(0,1)$
- Has stable fixed points at 0 and 1, an unstable interior fixed point, and no other fixed points.

This is the same fixed point profile as the simplest fixed-input interspeaker variation model (where the learner chooses whichever of 0 and 1 is closest to $\frac{k}{N}$: §5.1.1). So adding *any* frequency boosting gives bistability.

Poisson input Now assume $N \sim \text{Poisson}(\lambda)$ and learners choose $\hat{\alpha}_t = 0.5$ if $N = 0$. Then by a similar derivation, assuming N even, the evolution equation is now

$$\alpha_{t+1} = \frac{e^{-\lambda}}{2} + (1-w)(1-e^{-\lambda})\alpha_t + w\sum_{N=1}^{\infty} \frac{e^{-\lambda}\lambda^N}{N!} I(\alpha_t; \frac{N}{2}+1, \frac{N}{2}) \tag{27}$$

By simulation, (27) has either one (stable) or three (stable, unstable, stable) fixed points, with a bifurcation between these two cases.

Results For fixed input, stable fixed points at 0 and 1 and one interior fixed point. For Poisson input, bifurcation (in λ) between one stable fixed point and three (stable, unstable, stable) fixed points.

6.3 Regularization III: Bayesian learners

Another possible class of biased learners are Bayesian learners: each learner in a population begins with an identical prior over the parameter to be estimated (here $\hat{\alpha}_t$), receives evidence during the learning process, and updates the prior to obtain a posterior distribution for $\hat{\alpha}_t$. We consider two types of learners which differ by how they produce forms given the posterior. One samples over the posterior (*posterior mean*) each time a form is produced, the other chooses the maximum value of $\hat{\alpha}_t$ over the posterior – the *maximum a posteriori* (MAP) estimate – and produces form 2 with this probability.

6.3.1 Preliminaries

Formally: learners receive N examples, of which k are form 2. Let α_t be the probability of hearing form 2 from generation t, and let $\hat{\alpha}_t$ be the random variable corresponding to hypotheses of learners learning from this population, so that $\alpha_{t+1} = E(\hat{\alpha}_t)$.

Each learner has a prior $\pi(\theta)$ over the probability of producing form 2. For binomial likelihood (as here), this prior is often assumed to be a Beta distribution with parameters A and B ($A, B \geq 0$), denoted $\mathcal{B}_\theta(A,B)$.[32] That is,

$$\pi(\theta) \propto \theta^{A-1}(1-\theta)^{B-1}$$

[32]This is because the Beta distribution is the "conjugate prior" to the binomial distribution.

In Bayesian analyses, care is often taken to choose a maximally "non-informative prior." In this vein, three standard non-informative priors from the Beta family (Gelman et al., 2004, p. 63) are $A = B = 1$ (flat prior), $A = B = 0.5$ (Jeffrey prior), $A = B = 0$ (flat prior in natural parameter). However, because we are in part interested in the effect of the prior's shape on the dynamics, we deal with the general A and B cases below.

We need some properties of $\mathcal{B}_\theta(A, B)$:

$$E(\mathcal{B}_\theta(A,B)) = \frac{A}{A+B} \tag{28}$$

$$\text{mode}(\mathcal{B}_\theta(A,B)) = \begin{cases} 0 & : A < 1, B > 1 \\ 1 & : A > 1, B < 1 \\ \frac{A-1}{A+B-2} & : A > 1, B > 1 \end{cases} \tag{29}$$

After hearing k examples, a learner updates their prior using Bayes' rule:

$$\begin{aligned} \pi(\theta \,|\, k, n) &= P(k, n \,|\, \theta)\pi(\theta) \\ &\propto \theta^{A-1}(1-\theta)^{B-1}\theta^k(1-\theta)^{n-k} \\ &= \theta^{(A+k)-1}(1-\theta)^{(B+n-k)-1} \\ &\propto \mathcal{B}(A+k, B+n-k) \end{aligned}$$

6.3.2 Posterior mean

Suppose each learner stores their prior, and when called upon to produce a form, first chooses θ from the prior, then produces form 2 with probability θ. Then

$$P(\text{form } 2 \,|\, k, n) = \int \theta P(\theta \,|\, k, n) d\,\theta = \frac{A+k}{A+B+n}$$

from (28) and (29)

$$\alpha_{t+1} = E(\hat{\alpha}_t) = \sum_{k=0}^{n} \binom{n}{k} {\alpha_t}^k (1-\alpha_t)^{n-k} P(\text{form } 2 \,|\, k, n) \tag{30}$$

$$\alpha_{t+1} = \frac{A}{A+B+n} + \frac{n\alpha_t}{A+B+n} \tag{31}$$

Eqn. 31 has a unique fixed point at

$$\alpha_* = \frac{A}{A+B}$$

so the population "regresses to the prior" over time. For the three non-informative priors, $\alpha_* = 0.5$.

Posterior mean with mistransmission Let a and b be mistransmission probabilities, so that the probability of hearing form 2 from a member of generation t is

$$p_{2,t} = \alpha_t(1-a) + (1-\alpha)b$$

Then by a similar derivation to above, there is a unique fixed point at

$$\frac{A' + b(A'+B'+1)}{A'+B'+(a+b)}$$

where $A' = A/n$, $B' = B/n$. This is the same as the non-Bayesian result as $A', B' \to 0$, meaning that the system's behavior looks less Bayesian as either $A, B \to 0$ or as n (frequency) increases. This means the non-Bayesian case corresponds to the improper prior, which is maximally weighted towards $\theta = 0$ and $\theta = 1$ over all possible (Beta) priors.

6.3.3 MAP estimate

Now suppose each learner does not store their prior, but chooses the MAP likelihood (the mode of the posterior distribution) and produces form 2 with that probability.

1. $A > 1$, $B > 1$: In this case the derivation is similar, but now using the mode rather than the mean of the posterior:

$$\alpha_{t+1} = \sum_{k=0}^{n} \binom{n}{k} {\alpha_t}^k (1-\alpha_t)^{n-k} \frac{A-1+k}{A+B+n-2}$$

   which gives the evolution equation

$$\alpha_{t+1} = \frac{A-1}{A+B+n-2} + \alpha_t \frac{n}{A+B+n-2} \tag{32}$$

   which has unique fixed point

$$\alpha_* = \frac{A-1}{A+B-2}$$

   Derivations for the remaining three cases are similar, and given in App. E.3
2. $A < 1$, $B > 1$: The evolution equation is

$$\alpha_{t+1} = \frac{1}{A+B+n-2}[n\alpha_t + (1-a)((1-\alpha_t)^n - 1)] \tag{33}$$

   Since 0 is a fixed point and one can check that $\alpha_t - \alpha_{t+1}$ is negative for $\alpha_t > 0$, 0 is the unique fixed point.
3. $A > 1$, $B < 1$: 1 is the unique fixed point.
4. $A > 1$, $B > 1$: The evolution equation is

$$\alpha_{t+1} = \frac{1}{A+B+n-2}\left[n\alpha_t + (A-1)(1-(1-\alpha_t)^n) + (B-1){\alpha_t}^n\right] \tag{34}$$

   One can show that 0 and 1 are the only stable fixed points of (34).

Qualitatively, these cases correspond to four solution regions:

1. $A > 1$, $B > 1$: Stable FP at $\frac{A-1}{A+B-2}$.
2. $A < 1$, $B > 1$: Stable FP at 0.
3. $A > 1$, $B < 1$: Stable FP at 1.
4. $A < 1$, $B < 1$: Stable FPs at 0 and 1.

There are thus two bifurcations corresponding to the lines $\{A < 1, B = 1\}$ and $\{A = 1, B < 1\}$: the first determines whether 1 is a stable FP; the second determines whether 0 is a stable FP.[33]

[33] Note that the transitions between region 1 and 2 and 1 and 3 are not bifurcations: the location of the single fixed point changes smoothly.

MAP estimate with mistransmission Define a and b, $p_{2,t}$ as above. For the four cases depending on the signs of $A-1$ and $B-1$, we now find:

1. $A > 1$, $B > 1$: The evolution equation is (32) with $\alpha_t \to p_{2,t}$:

$$\alpha_{t+1} = \frac{A-1}{A+B+n-2} + \alpha_t \frac{n}{A+B+n-2} \tag{35}$$

   which has a unique fixed point at

$$\alpha_* = \frac{(A-1)+bn}{(A-1)+(B-1)+n(a+b)} \tag{36}$$

   Note that (36) looks like the non-Bayesian mistransmission case ($\alpha_* = \frac{b}{a+b}$) as $n \to \infty$ and like the no-mistransmission case above (6.3.3) as $n \to 0$.

2. $A < 1$, $B > 1$: The evolution equation is (33) with $\alpha_t \to p_{2,t}$:

$$\alpha_{t+1} = \frac{1}{A+B+n-2}[np_{2,t} + (1-a)((1-p_{2,t})^n - 1)] \tag{37}$$

   Assuming $a < 0.5$ and $n > 2$, there is 1 stable FP in $[0,1]$.[34]

3. $A > 1$, $B < 1$: By similar reasoning, there is still only one stable fixed point.

4. $A < 1$, $B < 1$: The evolution equation is (34) with $\alpha_t \to p_{2,t}$:

$$\alpha_{t+1} = \frac{1}{A+B+n-2}\left[np_{2,t} + (A-1)(1-(1-p_{2,t})^n) + (B-1){p_{2,t}}^n\right] \tag{38}$$

   In this case, one can show that α_{t+1} has either 3 or 1 FP(s) in $[0,1]$.[35]

In case 4, There can therefore be a bifurcation between 2 stable FPs and 1 stable FP, as for the non-mistransmission case (§6.3.3). Fig. 20 shows a sample phase diagram.

From simulations, it seems the size of the yellow (two stable FPs) region decreases as a and b increase (holding $n = 10$, by $a = b = 0.1$ it has disappeared) and as n increases (holding $a = b = 0.025$, by $n = 50$ it has disappeared). Adding mistransmission thus makes the dynamics frequency-dependent, in that the bifurcation disappears as n is increased.

6.4 Discussion

Adding frequency boosting has striking results: if learners boost at all ($w > 0$), the dynamics have the same fixed point profiles as analogous interspeaker variation cases, in particular showing the desirable property of multiple stable fixed points.

Thresholding and frequency boosting share an important property: in both cases, adding *any* regularization leads to qualitatively different dynamics which can give multistability and bifurcations. This is encouraging given the evidence that intraspeaker variation and bifurcations in fixed point stability coexist (§4.1).

[34] $\alpha'_{t+1}(0) < 0$, $\alpha_{t+1}(0) > 0$, and α_{t+1} is concave up for $\alpha_t \in [0,1]$, which imply there are either 1 or 0 stable FPs in $[0,1]$. Considering $\alpha_{t+1} - \alpha_t$ shows there is 1.

[35] Because $\alpha''_{t+1}(0) > 0$, $\alpha''_{t+1}(1) < 0$, and α_{t+1} has only one inflection point in $[0,1]$.

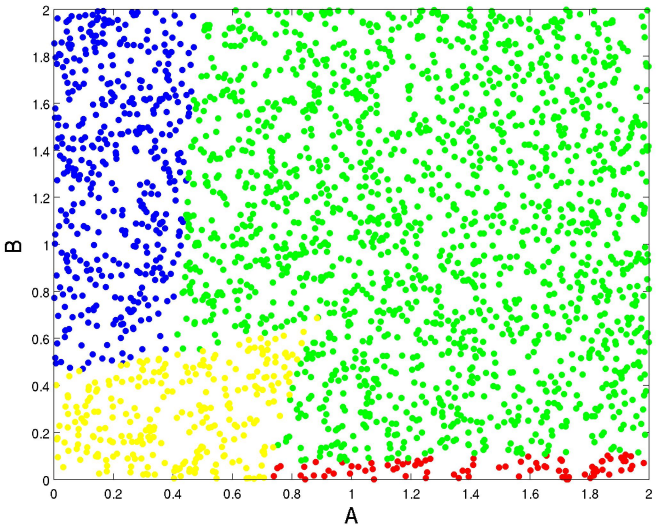


Figure 20: Phase diagram (in prior parameters A, B) for stable FPs of evolution equation 38, with $a = 0.03$, $b = 0.01$, $n = 10$. Regions are 1 stable FP $\alpha_* < 0.01$ (blue), 1 stable FP $\alpha_* > 0.99$ (red), 1 stable FP $\alpha_* \in (0.01, 0.99)$ (green), 2 stable FPs (yellow).

The effect of regularization can be contrasted with the effect of mistransmission, another proposed source of sound change. In intraspeaker variation models considered here, mistransmission never leads to bifurcations or multistability; instead, it tends to shift the location of stable fixed points.

It can also be contrasted with the effect of discarding. In intraspeaker variation models considered here, discarding gives bifurcations, but not multistability, which we have argued is a desirable property to explain dialectical differences in N/V pair stresses (§4.1).

We also found that the dynamics of a population of simple Bayesian learners differs significantly depending on how learners sample from their posterior to produce input to the next generation. Assuming learners sample at random from the posterior, there is a unique fixed point determined by system parameters. Assuming learners choose the most likely $\hat{\alpha}$ (MAP estimation) and produce form 2 examples with this probability, the dynamics show bifurcations.

Previous work on "iterated learning" by Bayesian learners (Kirby et al., 2007; Griffiths and Reali, 2008; Griffiths and Kalish, 2007) is in a similar vein, but differs in one crucial respect: generations in iterated learning models are generally assumed to consist of one speaker, while we consider generations to consist of a large number of speakers ("social learning"). The two cases in general lead to very different diachronic dynamics (Niyogi and Berwick, 2009).

All regularization models are summarized in Table 7.

Model type		Input	Sec.	Bifurc.?	Bif.-like?	Freq-dep?	Stable fixed points
Prob. Matching		F	5.1.2				All $\alpha_* \in [0,1]$
Mistransmission		F	5.2.2				$\{\alpha_*\}$
		P	5.4.1			✓	$\{\alpha_*\}$
Discarding	Large λ	F	5.5.1	✓			$\{0\}$ or $\{1\}$
	Finite λ				✓	✓	$\{\alpha_*\}$
	$\lambda < 15$	P	5.5.3				$\{\alpha_*\}$
	$\lambda > 15$				✓	✓	$\{\alpha_*\}$
Discarding + Mistrans.	D>M	F	5.5.2	✓		✓	$\{\alpha_*\}$
	D<M					✓	$\{\alpha_*\}$
Thresholding	+Mistrans.	F	6.1.1	✓			$\{0\}$, $\{1\}$, or $\{0,1\}$
			6.1.2	✓		✓	$\{0\}$, $\{1\}$, $\{\alpha_*\}$, $\{0,1\}$, $\{0,\alpha_*\}$, $\{\alpha_*,1\}$, or $\{0,\alpha_*,1\}$
	+Mistrans.	P	6.1.3	✓		✓	$\{\alpha_*\}$ or $\{\epsilon, 1-\epsilon\}$
				✓		✓	$\{\epsilon\}$, $\{\bar{\epsilon}\}$, $\{\alpha_*\}$, $\{\epsilon,\bar{\epsilon}\}$, $\{\epsilon,\alpha_*\}$, $\{\alpha_*,\bar{\epsilon}\}$, or $\{\epsilon,\alpha_*,\bar{\epsilon}\}$
Frequency boosting		F	6.2.1				$\{0,1\}$
		P		✓		✓	$\{\alpha_*\}$ or $\{\epsilon,\bar{\epsilon}\}$
Bayesian learners: Posterior mean	+Mistrans.	F	6.3.2				$\{\alpha_*\}$
						✓	$\{\alpha_*\}$
Bayesian learners: MAP estimate	+Mistrans.	F	6.3.3	✓			$\{0\}$, $\{\alpha_*\}$, $\{1\}$, or $\{0,1\}$
				✓		✓	$\{\epsilon\}$, $\{\alpha_*\}$, $\{\bar{\epsilon}\}$, $\{\epsilon,\bar{\epsilon}\}$

Table 7: Summary of single-form, intraspeaker variation models. Abbreviations: F=Fixed input, P=Poisson input. Bif.-like=bifurcation-like (rapid change in stable FP location(s)). Freq-dep=Frequency-dependent FP location(s). $\{0,1\}$="stable fixed points at 0 and 1"; "$\{0\}$ or $\{1\}$"="bifurcation between 1 stable fixed point at 0 and 1 stable fixed point at 1." α_*=interior point ($\in (0,1)$). ϵ="fixed point near 0", $\bar{\epsilon}$="fixed point near 1."

7 Models III: Coupling between forms

All models considered so far deal with variation in individual forms. However, the N/V data analyzed above show significant interaction, or *coupling*, between the N and V pronunciations of individual N/V pairs, as well as the pronunciations of N/V pairs sharing a morphological prefix. The most salient facts about the N/V trajectories were that

- Stable variation in both N and V of an N/V pair's pronunciation is rare; stable variation in N or V individually is more common.
- Words with the same prefix have more similar N/V trajectories than random pairs of words, provided the prefix class is not small.
- {2,1} never occurs.
- Falling word frequency is associated with change in N/V pairs.

We describe several types of coupling models and examine which of these observed properties occur in each.

There is a crucial difference between the learning tasks for individual and coupled forms. In the single form case, learners can (under the simplest hypothesis) probability match: a priori, no inference is required. This is not the case for coupled forms: learners never hear N/V pairs, only individual N or V examples, so that given the evidence for coupling between N and V forms, some inference *must* take place. Put otherwise, learners hear forms, not {N,V} grammars, yet their algorithm to produce forms depends on information about grammars.

Our coupling models are based on maximum-likelihood inference by learners under different assumptions about the variables being estimated. Because even moving from one form to two coupled forms complicates things significantly, to get an idea of the dynamics of different coupling models we will usually stick to the fixed-input case, and at times to the high-frequency (λ) case. In addition, we always assume intraspeaker variation, which we have shown occurs for N/V pairs (§2.2).

7.1 Coupling by grammar I

One possibility, suggested by Yang (2001, 2002), is that learners store probabilities for multiple grammars, then choose which one to use in production probabilistically. A learner's task is to learn a probability for each grammar. We would like to see whether in such a system, building in a bias against learning {2,1} causes it to be eliminated diachronically.

Assume each member i of the population keeps probabilities α_i, β_i, γ_i, δ_i for the grammars $\{1,1\}$, $\{1,2\}$, $\{2,1\}$, $\{2,2\}$, with $\alpha_i + \beta_i + \gamma_i + \delta_i = 1$, corresponding to the probability they use each grammar in a given utterance. Using similar notation to above, let $\hat{\alpha}_t$ be the random variable corresponding to values of α_i in the population at t, and let $\alpha_{t+1} = E[\hat{\alpha}_t]$, so that α_t is the probability "grammar" $\{1,1\}$ is used by a member of generation t. Define β_t, γ_t, δ_t, etc. similarly.

Suppose learner i hears N_1 noun examples, of which k_1 have initial stress, and N_2 verb examples, of which k_2 have initial stress. The simplest strategy would be to assume that

$$\alpha_i + \beta_i = \frac{k_1}{N_1}, \quad \alpha_i + \gamma_i = \frac{k_2}{N_2}, \quad \alpha_i + \beta_i + \gamma_i + \delta_i = 1 \tag{39}$$

and solve. But there are four unknowns and three equations, so a unique solution for $(\alpha_i, \beta_i, \gamma_i, \delta_i)$ is not specified, and (for a unique solution) another constraint is needed. Motivated by Ross'

generalization, one possibility is to minimize the probability of $\{2,1\}$ (γ_i). The learner's algorithm then turns out to be (App. E.4)

- If $\frac{k_1}{N_1} > \frac{k_2}{N_2}$, set $(\alpha, \beta, \gamma) = (\frac{k_2}{N_2}, \frac{k_1}{N_1} - \frac{k_2}{N_2}, 0)$.
- If $\frac{k_1}{N_1} < \frac{k_2}{N_2}$, set $(\alpha, \beta, \gamma) = (\frac{k_1}{N_1}, 0, \frac{k_2}{N_2} - \frac{k_1}{N_1})$.

and we find that under this algorithm,

$$\alpha_{t+1} + \beta_{t+1} = \alpha_t + \beta_t \tag{40}$$

$$\alpha_{t+1} + \gamma_{t+1} = \alpha_t + \gamma_t \tag{41}$$

Note that the effective probabilities of hearing N=1 and V=1 at time t are

$$\alpha_{N,t} = \alpha_t + \beta_t, \quad \alpha_{V,t} = \alpha_t + \gamma_t$$

so the dynamics are the identity map on (α_N, α_V).

Interestingly, then, although we built in a constraint explicitly disfavoring $\{2,1\}$, there is no tendency for α_N to increase or for α_V to decrease over time. The "minimize the probability of {2,1}" constraint is apparently too weak to affect the dynamics – a constraint against {1,2}, the diachronically-favored form, would give the same result.

7.2 Coupling by grammar II: Mistransmission

Another strategy for estimating $(\alpha_i, \beta_i, \gamma_i, \delta_i)$, given they are underdetermined by (39), is to assume they come from a two-parameter subfamily:

$$\alpha_i = pq, \quad \beta_i = p(1-q), \quad \gamma_i = (1-p)q, \quad \delta_i = (1-p)(1-q) \tag{42}$$

Intuitively, the learner is making an independence assumption, that p and q are weights for "N=1" and "V=1", and the probability of the grammar {1,1} is proportional to the product of the weights for N=1 and V=1 (etc.)

To estimate p and q, consider the likelihood of observing (k_1, k_2) given p, q, N_1, N_2:

$$\begin{aligned} P(k_1, k_2 \,|\, p, q, N_1, N_2) &= \binom{N_1}{k_1} (\alpha_i + \beta_i)^{k_1} (1 - \alpha_i - \beta_i)^{N_1 - k_1} \times \\ &\qquad \binom{N_2}{k_2} (\alpha_i + \gamma_i)^{k_2} (1 - \alpha_i - \gamma_i)^{N_2 - k_2} \\ &= \binom{N_1}{k_1} p^{k_1} (1-p)^{N_1 - k_1} \binom{N_2}{k_2} q^{k_2} (1-q)^{N_2 - k_2} \end{aligned} \tag{43}$$

Viewed as a function of (p, q), (43) defines the likelihood:

$$\ell(p, q | k_1, k_2) = p^{k_1} (1-p)^{N_1 - k_1} q^{k_2} (1-q)^{N_2 - k_2}$$

Setting $\frac{\partial \ell}{\partial p} = \frac{\partial \ell}{\partial q} = 0$ then gives the maximum-likelihood estimate, $(\hat{p}, \hat{q}) = (\frac{k_1}{N_1}, \frac{k_2}{N_2})$.

Substituting back into (42), the learner estimates:

$$\alpha_i = \frac{k_1 k_2}{N_1 N_2}, \quad \beta_i = \frac{k_1 (N_2 - k_2)}{N_1 N_2}, \quad \gamma_i = \frac{(N_1 - k_1) k_2}{N_1 N_2}, \quad \delta_i = \frac{(N_1 - k_1)(N_2 - k_2)}{N_1 N_2}.$$

Mistransmission In the previous model learners were biased against $\{2,1\}$ by minimizing γ_i. Since this is not the case here, and some mechanism by which V=1 and/or N=2 is needed, assume we have mistransmission probabilities:

- Nouns: $a = P(2 \text{ heard} \,|\, 1\text{meant})$, $b = P(1 \text{ heard} \,|\, 2\text{meant})$
- Verbs: $c = P(2 \text{ heard} \,|\, 1\text{meant})$, $d = P(1 \text{ heard} \,|\, 2\text{meant})$

The probabilities $P_{N,t}$, $P_{V,t}$ of hearing N and V examples as 1 at t are

$$P_{N,t} = (\alpha_t + \beta_t)(1-a) + (\delta_t + \gamma_t)b, \quad P_{V,t} = (\alpha_t + \gamma_t)(1-c) + (\beta_t + \delta_t)d$$

Since N and V examples are independently-occurring events,$k_1 \sim \text{Bin}(P_{N,t}, N_1)$ and $k_2 \sim \text{Bin}(P_{V,t}, N_2)$, and so $E(k_1) = P_{N,t}N_1$ and $E(k_{2,t}) = P_{V,t}N_2$. The evolution of α_t is then given by

$$\begin{aligned}
\alpha_{t+1} = E(\hat{\alpha}_t) = E\left(\frac{k_1 k_2}{N_1 N_2}\right) &= \frac{E(k_1)E(k_2)}{N_1 N_2} \\
&= \frac{(N_1 P_{N,t})(N_2 P_{V,t})}{N_1 N_2} \\
&= P_{N,t}P_{V,t} \\
&= [(\alpha_t + \beta_t)(1-a) + (\delta_t + \gamma_t)b] * [(\alpha_t + \gamma_t)(1-c) + (\beta_t + \delta_t)d].
\end{aligned}$$

Similar calculation gives the whole set of evolution equations,

$$\alpha_{t+1} = [(\alpha_t + \beta_t)(1-a) + (\delta_t + \gamma_t)b] * [(\alpha_t + \gamma_t)(1-c) + (\beta_t + \delta_t)d] \tag{44}$$

$$\beta_{t+1} = [(\alpha_t + \beta_t)(1-a) + (\delta_t + \gamma_t)b] * [(\alpha_t + \gamma_t)c + (\beta_t + \delta_t)(1-d)] \tag{45}$$

$$\delta_{t+1} = [(\alpha_t + \beta_t)a + (\delta_t + \gamma_t)(1-b)] * [(\alpha_t + \gamma_t)c + (\beta_t + \delta_t)(1-d)] \tag{46}$$

$$\gamma_{t+1} = [(\alpha_t + \beta_t)a + (\delta_t + \gamma_t)(1-b)] * [(\alpha_t + \gamma_t)(1-c) + (\beta_t + \delta_t)d] \tag{47}$$

The fixed-point condition $(\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \delta_{t+1}) = (\alpha_t, \beta_t, \gamma_t, \delta_t)$ can be solved for analytically (App. E.5), and gives three cases depending on the mistransmission probabilities:

- **Case 1**: $a + b > 0$, $c + d > 0$: There is a unique, stable fixed point

$$(\alpha^*, \beta^*, \delta^*, \gamma^*) = \frac{1}{(a+b)(c+d)}(bd, bc, ac, ad)$$

- **Case 2**: $a + b > 0$, $c + d = 0$ or $a + b = 0$, $c + d > 0$: There is a line of fixed points (in $(\alpha, \beta, \gamma, \delta)$ space).
- **Case 3**: $a + b = 0$, $c + d = 0$: There is a two-dimensional surface of fixed points.

These cases can be interpreted as (1) N and V mistransmission (2) N *or* V mistransmission, not both (3) no mistransmission. Though it is interesting that the fixed points in each case form a different mathematical object, no case shows any of the properties suggested by the N/V data.

7.3 Coupling by constraint

In the previous two models, learners' N and V forms were coupled because probabilities of using different (N,V) grammars were stored, rather than separate probabilities for N and V forms. In this and the next coupling model, we assume learners store probabilities $\hat{\alpha}$ and $\hat{\beta}$ of producing N and V forms of a word with final (2) stress. Say each learner hears N_1 N and N_2 V examples, of which k_1, k_2 are heard as 2. Define α_t, $\hat{\alpha}_t$, β_t, $\hat{\beta}_t$ analogously to above, so $\alpha_{t+1} = E[\hat{\alpha}_t]$, etc.

Learners set $\hat{\alpha}$ and $\hat{\beta}$ as:

- If $\frac{k_1}{N_1} < \frac{k_2}{N_2}$, set $\hat{\alpha} = \frac{k_1}{N_1}$, $\hat{\beta} = \frac{k_2}{N_2}$.
- Otherwise, choose $\hat{\alpha}$ and $\hat{\beta}$ to satisfy the optimization problem

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

Using Lagrange multipliers gives

$$\hat{\alpha} = \hat{\beta} = \frac{1}{2}(\frac{k_1}{N_1} + \frac{k_2}{N_2})$$

in the second case.

The thinking behind this algorithm is that one way for $\{2, 1\}$ to never arise is if learners never hypothesize $\hat{\beta} < \hat{\alpha}$, in accordance with Ross' generalization for English (§3.1), that stress in nouns is further left than in verbs. If learners' best guess at $(\hat{\alpha}, \hat{\beta})$ violates this constraint, they pick the closest $(\hat{\alpha}, \hat{\beta})$ that does not violate it to their observed values.

Assuming no mistransmission or discarding, k_1 and k_2 are independent and binomially distributed:

$$P(k_1, k_2) = \binom{N_1}{k_1}\binom{N_2}{k_2}\alpha^{k_1}(1-\alpha)^{N_1-k_1}\beta^{k_2}(1-\beta)^{N_2-k_2}$$

Then we can show (App. E.8) the evolution equations are

$$\alpha_{t+1} = \alpha_t - \frac{1}{2}\sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) \tag{48}$$

$$\beta_{t+1} = \beta_t + \frac{1}{2}\sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) \tag{49}$$

Since the sum terms are non-negative, we have $\alpha_{t+1} \leq \alpha_t$, $\beta_{t+1} > \beta_t$, and $\alpha_{t+1} + \beta_{t+1} = \alpha_t + \beta_t$. The (α_t, β) trajectories are thus the lines of constant $\alpha_t + \beta_t$, moving in the direction of the lines $(\alpha_t, \beta_t) = (0, k)$ and $(\alpha_t, \beta_t) = (k, 1)$ $(k \in [0, 1])$. All points on these lines are stable equilibria.

Result Every point on the lines from $(0, 0)$ to $(0, 1)$ and from $(0, 1)$ to $(1, 1)$ is a stable fixed point, no bifurcations.

7.3.1 Mistransmission

Remembering that only $(0,0)$, $(0,1)$, and $(1,1)$,[36] or points near them , seem to be true stable fixed points in the N/V data, this first pass at coupling by constraint gives too many stable fixed points. What happens when mistransmission is added, pushing perception towards these endpoints?

Define mistransmission probabilities

$$\begin{aligned} P(\text{hear N as } 2 \,|\, 1 \text{ intended}) &= a_{12}, \quad P(\text{hear N as } 1 \,|\, 2 \text{ intended}) = a_{21} \\ P(\text{hear V as } 2 \,|\, 1 \text{ intended}) &= b_{12}, \quad P(\text{hear V as } 1 \,|\, 2 \text{ intended}) = b_{21} \end{aligned}$$

The probabilities $p_{N,t}$, $p_{V,t}$ of hearing an N or V example as final stressed at t are then

$$\begin{aligned} p_{N,t} &= \alpha_t(1-a_{21}) + (1-\alpha_t)a_{12} = \alpha_t(1-a_{12}-a_{21}) + a_{12} \\ p_{V,t} &= \beta_t(1-b_{21}) + (1-\beta_t)b_{12} = \beta_t(1-b_{12}-b_{21}) + b_{12} \end{aligned}$$

Then the probabilities of hearing k_1, k_2 N and V examples is

$$P_t(k_1,k_2) = \binom{N_1}{k_1}\binom{N_2}{k_2} p_{N,t}^{k_1}(1-p_{N,t})^{N_1-k_1} p_{V,t}^{k_2}(1-p_{v,t})^{N_2-k_2}$$

Letting $\hat{\alpha}_t$, $\hat{\beta}_t$ be determined by the same algorithm as above, similar derivations give

$$\alpha_{t+1} = a_{12} + \alpha_t(1-a_{12}-a_{21}) - \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P_t(k_1,k_2)\left(\frac{k_1}{N_1} - \frac{k_2}{N_2}\right) \tag{50}$$

$$\beta_{t+1} = b_{12} + \beta_t(1-b_{12}-b_{21}) + \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P_t(k_1,k_2)\left(\frac{k_1}{N_1} - \frac{k_2}{N_2}\right) \tag{51}$$

It turns out (App. E.6) these evolution equations have a unique, stable fixed point which lies on the line

$$\alpha(a_{12}+a_{21}) + \beta(b_{12}+b_{21}) = a_{12}+b_{12}$$

7.3.2 Discarding, large N_1, N_2

Now assume there is no mistransmission, and assume there is discarding, with probabilities

$$\begin{aligned} P(\text{N discarded} \,|\, 2 \text{ intended}) &= r_1 \quad P(\text{N discarded} \,|\, 1 \text{ intended}) = r_2 \\ P(\text{V discarded} \,|\, 2 \text{ intended}) &= s_1 \quad P(\text{V discarded} \,|\, 1 \text{ intended}) = s_2 \end{aligned}$$

Define α_t, β_t, N_1, N_2, k_1, k_2 as above, and let l_1 and l_2 be the numbers of N and V examples heard as 1, so the number of discarded N and V examples are $N_1-k_1-l_1$, $N_2-k_2-l_2$. Assume large N_1, N_2, so that the probability of receiving only discarded examples $\to 0$ A learner then sets $\hat{\alpha}_t$, $\hat{\beta}_t$ as

[36] Corresponding to $\{1,1\}$, $\{1,2\}$, $\{2,2\}$

- $\frac{k_1}{k_1+l_1} < \frac{k_2}{k_2+l_2}$: set $\hat{\alpha} = \frac{k_1}{l_1+l_1}$, $\hat{\beta} = \frac{k_2}{k_2+l_2}$.
- Otherwise: Choose $\hat{\alpha}$ and $\hat{\beta}$ to satisfy the optimization problem

$$\begin{aligned} &\min[(\hat{\alpha} - \frac{k_1}{k_1+l_1})^2 + (\hat{\beta} - \frac{k_2}{k_2+l_2})^2] \\ &\text{s.t. } \hat{\alpha} \leq \hat{\beta} \end{aligned}$$

Using Lagrange multipliers gives

$$\hat{\alpha} = \hat{\beta} = \frac{1}{2}\,(\frac{k_1}{k_1+l_1} + \frac{k_2}{k_2+l_2})$$

in this case.

The probabilities $p_{N,2}(t)$, $p_{V,2}(t)$, $p_{N,1}(t)$, $p_{V,1}(t)$ of hearing N and V examples as 1 or 2 from generation t are then

$$\begin{aligned} p_{N,2}(t) = \alpha_t(1-r_1),\ p_{V,2} = \beta_t(1-s_1) \\ p_{N,1}(t) = (1-\alpha_t)(1-r_2),\ p_{V,1} = (1-\beta_t)(1-s_2) \end{aligned}$$

which leads to (as in §5.5.3)

$$\begin{aligned} E\left(\frac{k_1}{k_1+l_1}\right) &= \frac{\alpha_t(1-r_1)}{(1-r_2)+\alpha_t(r_2-r_1)} \\ E\left(\frac{k_2}{k_2+l_2}\right) &= \frac{\beta_t(1-s_1)}{(1-s_2)+\beta_t(s_2-s_1)} \end{aligned}$$

The probability of receiving (k_1, k_2, l_1, l_2) is then

$$\begin{aligned} P_t \equiv P_t(k_1,k_2,l_1,l_2) &= \binom{N_1}{k_1,l_1} p_{N,2}(t)^{k_1} p_{N,1}(t)^{l_1} (1-p_{N,2}(t)-p_{N,1}(t))^{N_1-k_1-l_1} \\ &\quad \times \binom{N_2}{k_2,l_2} p_{N,2}(t)^{k_2} p_{N,1}(t)^{l_2} (1-p_{N,2}(t)-p_{N,1}(t))^{N_2-k_2-l_2} \end{aligned}$$

Then

$$\begin{aligned} \alpha_{t+1} = E(\hat{\alpha}_t) &= \sum_{\frac{k_1}{k_1+l_1} < \frac{k_2}{k_2+l_2}} P_t \left(\frac{k_1}{k_1+l_1}\right) + \sum_{\frac{k_1}{k_1+l_1} > \frac{k_2}{k_2+l_2}} \frac{P_t}{2} \left(\frac{k_1}{k_1+l_1} + \frac{k_2}{k_2+l_2}\right) \\ \beta_{t+1} = E(\hat{\beta}_t) &= \sum_{\frac{k_1}{k_1+l_1} < \frac{k_2}{k_2+l_2}} P_t \left(\frac{k_2}{k_2+l_2}\right) + \sum_{\frac{k_1}{k_1+l_1} > \frac{k_2}{k_2+l_2}} \frac{P_t}{2} \left(\frac{k_1}{k_1+l_1} + \frac{k_2}{k_2+l_2}\right) \end{aligned}$$

and a similar derivation as for the no-mistransmission case gives evolution equations

$$\alpha_{t+1} = E(\hat{\alpha}_t) = \frac{\alpha_t(1-r_1)}{(1-r_2)+\alpha_t(r_2-r_1)} - \frac{1}{2} \sum_{\frac{k_1}{k_1+l_1} > \frac{k_2}{k_2+l_2}} P_t \left(\frac{k_1}{k_1+l_1} - \frac{k_2}{k_2+l_2}\right) \tag{52}$$

$$\beta_{t+1} = E(\hat{\beta}_t) = \frac{\beta(1-s_1)}{(1-s_2)+\beta(s_2-s_1)} + \frac{1}{2} \sum_{\frac{k_1}{k_1+l_1} > \frac{k_2}{k_2+l_2}} P_t \left(\frac{k_1}{k_1+l_1} - \frac{k_2}{k_2+l_2}\right). \tag{53}$$

Substituting into (52) and (53) gives that $(0,0)$, $(0,1)$, $(1,1)$ are fixed points, while $(1,0)$ is not. Define

$$R_N = \frac{1-r_1}{1-r_2}, \quad R_V = \frac{1-s_1}{1-s_2},$$

the relative probabilities that form 2 vs form 1 is discarded, for nouns and verbs. $R_N, R_V \in (0,\infty)$, and it turns out (App. E.7) that under the large N_1, N_2 assumption, there are four fixed point regions:

1. $R_N < R_V$, $R_N R_V < 1$: $(0,0)$, $(0,1)$ stable.
2. $R_N > R_V$, $R_N R_V < 1$: $(0,0)$ stable.
3. $R_N < R_V$, $R_N R_V > 1$: $(0,1)$, $(1,1)$ stable.
4. $R_N > R_V$, $R_N R_V > 1$: $(1,1)$ stable.

There are two bifurcations: at $R_N = R_V$, corresponding to whether $(0,1)$ is stable or unstable, and $R_N R_V = 1$, corresponding to which of $(0,0)$ and $(1,1)$ is stable.

Although this model gives bifurcations between the fixed points observed in our data, it is missing the crucial property of bifurcations *to* $(0,1)$. Because there is no parameter region where $(0,1)$ alone is stable, there is no reason a population in $(0,0)$ or $(0,1)$ will ever transition to $(0,1)$ as system parameters (R_N and R_V) are changed. Put otherwise, this model has the right behavior if the data we observed showed lexical diffusion *from* $(0,1)$ to $(0,0)$ and $(1,1)$.

Results Similar to fixed-input models considered in the non-coupling case: lines of stable fixed points without mistransmission or discarding; which collapse to a unique fixed point when mistransmission is added (assuming there is mistransmission for both nouns and verbs); or bifurcations (in discarding probabilities) between stability of $(0,0)$, $(0,1)$, $(1,1)$ with discarding.

7.4 Coupling by the lexicon

The coupling by constraint model just considered assumes that learners receive data, make a hypothesis, and revise it if it violates a constraint on the relative frequency of N and V forms being pronounced as 1 and 2. This model has the drawback that there is no way for the rest of the lexicon to affect a pair's N and V stress probabilities, i.e. no coupling between the N/V pair being learned and (a) stress on other N/V pairs (b) stress in the lexicon as a whole. We consider another model that incorporates these factors into learning through the general notion of *lexical support*, which formalizes 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 model this idea by assuming that rather than one set of (learned) probabilities for each possible N/V pair pronunciation, learners keep two sets of probabilities (for $\{1,1\}$, $\{1,2\}$, $\{2,1\}$, $\{2,2\}$):

1. *Learned probabilities*: $\mathbf{\Lambda} = (pq, p(1-q), (1-p)q, (1-p)(1-q))$, as in §7.2: $(p,q) = (\frac{k_1}{N_1}, \frac{k_2}{N_2})$.
2. *Prior probabilities*: $\mathbf{\Sigma} = (a_{11}, a_{12}, a_{21}, a_{22})$, e.g. based on the support for each pattern in the lexicon, based on the pronunciation of N/V pairs already learned.

The learner then produces N forms as follows:

1. Pick a grammar $\{n_1, v_1\}$ according to $\mathbf{\Lambda}$ (pick {1,1} with probability pq, etc.)

2. Pick a grammar $\{n_2, v_2\}$ according to $\mathbf{\Sigma}$
3. If $n_1 = n_2$, produce N=n_1. Otherwise repeat steps 1-2.

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 different {N,V} patterns. We "build in" the fact that {2,1} does not occur in the learner's lexicon by setting $a_{2,1} = 0$.

We leave the actual interpretation of prior probabilities deliberately vague: they could be computed only over words with the same morphological prefix, over the whole lexicon, etc.

Define N_1, N_2, k_1, k_2 as above for an N/V pair, and let a_{11}, a_{22}, a_{12} be the prior probabilities for $\{1,1\}$, $\{2,2\}$, $\{1,2\}$, with $a_{11} + a_{22} + a_{12} = 1$. Given input $I = (k_1, k_2, N_1, N_2)$, a learner's probabilities of *producing* N=2 and V=2 are

$$\hat{\alpha}(k_1, k_2) = \frac{a_{22} k_1 k_2}{a_{11}(N_1 - k_1)(N_2 - k_2) + a_{22} k_1 k_2 + a_{12}(N_1 - k_1) k_2} \tag{54}$$

$$\hat{\beta}(k_1, k_2) = 1 - \frac{a_{11}(N_1 - k_1)(N_2 - k_2)}{a(N_1 - k_1)(N_2 - k_2) + a_{22} k_1 k_2 + a_{12}(N_1 - k_1) k_2} \tag{55}$$

$\hat{\alpha}$ and $\hat{\beta}$ are undefined for $(k_1, k_2) = (N_1, 0)$, so let the learning algorithm be

- $(k_1, k_2) \neq (N_1, 0)$: Use (54), (55).
- Otherwise: $\hat{\alpha} = a_{22}$, $\hat{\beta} = a_{22} + a_{12}$

In this algorithm, the learner's default strategy is to guess $\hat{\alpha}$, $\hat{\beta}$ based on the prior probabilities.

k_1 and k_2 are binomially distributed. Define

$$P_t(k_1, k_2) = \binom{N_1}{k_1} \binom{N_2}{k_2} \alpha_t^{k_1} (1 - \alpha_t)^{N_1 - k_1} \beta_t^{k_2} (1 - \beta_t)^{N_2 - k_2}$$

Then the evolution equations are

$$\alpha_{t+1} = E[\hat{\alpha}_t] = a_{22} \alpha_t^{N_1} (1 - \beta_t)^{N_2} + \sum_{k_1=1}^{N_1} \sum_{k_2=1}^{N_2} P_t(k_1, k_2) \frac{a_{22} k_1 k_2}{D(k_1, k_2)} \tag{56}$$

$$\beta_{t+1} = E[\hat{\beta}_t] = (a_{22} + a_{12}) \alpha_t^{N_1} (1 - \beta_t)^{N_2} + \sum_{k_1=0}^{N_1} \sum_{k_2=1}^{N_2} P_t(k_1, k_2) \frac{a_{22} k_1 k_2 + a_{12}(N_1 - k_1) k_2}{D(k_1, k_2)} \tag{57}$$

We can show (App. E.9) that the only fixed points are $(0,0)$, $(0,1)$, $(1,1)$, and that defining

$$B = \left(\frac{a_{22} N_1}{a_{22} - a_{12} + a_{12} N_1} \right), \quad A = \left(\frac{a_{11} N_2}{a_{11} - a_{12} + a_{12} N_2} \right) \tag{58}$$

there are 6 solution regions:

1. $a_{11}, a_{22} < a_{12}$: $(0,1)$ stable
2. $a_{22} > a_{12}$, $AB < 1$: $(0,1)$, $(1,1)$ stable
3. $a_{11} < a_{12} < a_{22}$, $AB > 1$: $(1,1)$ stable

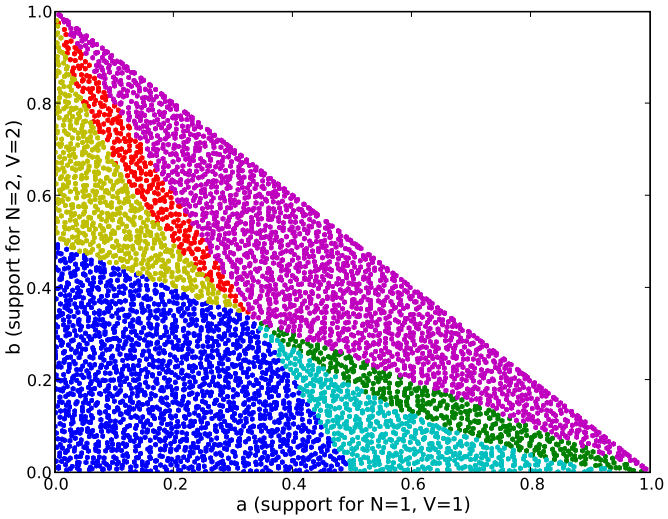


Figure 21: Phase diagram for stable fixed points of evolution equations 54, 55 as a function of prior probabilities a, b ($a_{12} = 1 - a - b$) for $N_1 = 5$, $N_2 = 10$. Dark blue: only $(0, 1)$ stable. Gold: $(0, 1)$ and $(1, 1)$ stable. Green: only $(0, 0)$ stable. Purple: $(0, 0)$ and $(1, 1)$ stable. Red: only $(1, 1)$ stable. Light blue: $(0, 0)$ and $(0, 1)$ stable.

4. $a_{11}, a_{22} > a_{12}$: $(0, 0)$, $(1, 1)$ stable
5. $a_{22} < a_{12} < a_{11}$, $AB > 1$: $(0, 0)$ stable
6. $a_{11} > a_{12}$, $AB < 1$: $(0, 0)$, $(0, 1)$ stable

A sample phase diagram of these regions is shown in Fig. 21.

As N_1 and N_2 are varied, the $a_{22} = a_{12}$ and $a = a_{12}$ lines do not change, but the relative sizes of regions 2 and 3 and regions 5 and 6 do. This is because

$$\frac{\partial(AB)}{\partial N_1} = \left(\frac{AB}{N_1}\right)^2 (1 - \frac{a_{12}}{a_{22}}), \quad \frac{\partial(AB)}{\partial N_2} = \left(\frac{AB}{N_2}\right)^2 (1 - \frac{a_{12}}{a_{11}}),$$

and so the curve $AB = 1$ shifts as N_1, N_2 change.

Under f, a population in a stable state can be either in regions 1, 3, 5, which permit only one stable state, or in the bistable regions 2, 4, 6. In these regions is possible for two populations to be in different stable states, yet have similar (a_{11}, a_{22}, a_{12}), perhaps because they entered the region with different (α, β) values. As proposed above (§4.1), this situation could correspond to populations speaking different dialects of English, which broadly speaking have similar stress patterns (here, similar prior probabilities for $\{1, 1\}$, $\{2, 2\}$, $\{1, 2\}$), yet often stress N/V pairs of the type considered here differently.

7.4.1 Mistransmission

What happens to the dynamics when mistransmission is added? Let p, q be $\in (0, 0.5)$ mistransmission probabilities:[37]

$$p = P(\text{N heard as } 1 \,|\, 2 \text{ intended}), \quad q = P(\text{V heard as } 2 \,|\, 1 \text{ intended})$$

Then the probability of receiving k_1 final-stressed N examples and k_2 final-stressed V examples is

$$P_t(k_1, k_2) = \binom{N_1}{k_1}\binom{N_2}{k_2}\alpha_t'^{k_1}(1-\alpha_t')^{N_1-k_1}\beta_t'^{k_2}(1-\beta_t')^{N_2-k_2}$$

where $\alpha_t' = \alpha_t(1-p)$, $\beta_t' = \beta_t + q(1-\beta_t)$. The evolution equations are then

$$\alpha_{t+1} = E[\hat{\alpha}_t] = a_{22}\alpha_t'^{N_1}(1-\beta_t')^{N_2} + \sum_{k_1=1}^{N_1}\sum_{k_2=1}^{N_2} P_{\alpha_t',\beta_t'}(k_1,k_2)\frac{a_{22}k_1k_2}{D(k_1,k_2)} \tag{59}$$

$$\beta_{t+1} = E[\hat{\beta}_t] = (a_{22}+a_{12})\alpha_t'^{N_1}(1-\beta_t')^{N_2} + \sum_{k_1=0}^{N_1}\sum_{k_2=1}^{N_2} P_{\alpha_t',\beta_t'}(k_1,k_2)\frac{a_{22}k_1k_2 + a_{12}(N_1-k_1)k_2}{D(k_1,k_2)} \tag{60}$$

It can be shown (App. E.10) that, with A, B defined by (58), letting

$$a_{11}' = a_{11}(1 - q\frac{N_2}{N_2-1}), \quad a_{22}' = a_{22}(1 - p\frac{N_1}{N_1-1}), \quad a_{12}' = a_{12}$$

be the *effective prior probabilities* for $\{1,1\}$, $\{2,2\}$, $\{1,2\}$, and letting $(0,\lambda)$ and $(\kappa,1)$ mean "some stable fixed point of this form", there are six parameter regions, analogous to those found above

1. $a_{11}', a_{22}' < a_{12}$: $(0,1)$ stable
2. $a_{22}' > a_{12}'$, $AB < \frac{1}{(1-p)(1-q)}$: $(0,1)$, $(\kappa,1)$ stable
3. $a_{11}' < a_{12}' < a_{22}'$, $AB > \frac{1}{(1-p)(1-q)}$: $(\kappa,1)$ stable
4. $a_{11}', a_{22}' > a_{12}'$: $(0,\lambda)$, $(\kappa,1)$ stable
5. $a_{22}' < a_{12}' < a_{11}'$, $AB > \frac{1}{(1-p)(1-q)}$: $(0,\lambda)$ stable
6. $a_{11}' > a_{12}'$, $AB < \frac{1}{(1-p)(1-q)}$: $(0,\lambda)$, $(0,1)$ stable

Fig. 22 shows an example phase diagram. Note that these regions reduce to those of the no-mistransmission case (above) when $p = q = 0$. Examining the mistransmission case parameter regions for the effect of p and q, we see that

- When $q > 0$, the effective prior probabilities for $(0,\lambda)$ is decreased and becomes frequency dependent. As N_2 decreases and q increases, more lexical support (prior probability mass) is needed for $(0,\lambda)$ relative to $(0,1)$ (a_{12}) for $(0,\lambda)$ to remain stable.
- In particular, when $a_{12} \geq (1 - q\frac{N_2}{N_2-1})$, no $(0,\lambda)$ fixed point exists.
- as p and q move from 0, the region of stability for $(0,1)$ expands.
- as N_1, N_2 decrease (for $N_1, N_2 \geq 2$), the region of stability for $(0,1)$ expands.

7.5 Discussion

All coupling models considered are summarized in Table 8.

[37] As above (§5.2.1), so that the probability of hearing the intended form is better than chance.

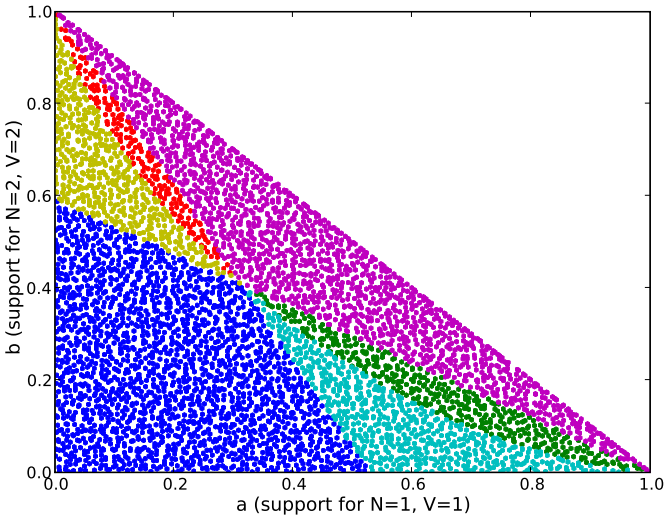


Figure 22: Phase diagram for stable fixed points of evolution equations 54, 55 as function of prior probabilities a_{11}, a_{22} ($a_{12} = 1 - a_{11} - a_{22}$) for $N_1 = 5$, $N_2 = 10$. Colors as in Fig. 21. Note expanded range of blue region (only $(0, 1)$ stable) relative to Fig. 21.

Model type		Sec.	{1,2}	{1,1}/{2,2}	*{2,1}	Bifurc?	Obs bif.?	Freq-dep?	Stable FPs
Coupling by grammar I		7.1							All $(\alpha_*, \beta_*) \in [0,1]^2$
Coupling by grammar II		7.2			✓				Unique (α_*, β_*)
Coupling by constraint		7.3	✓	✓	✓				All $[0, \beta_*]$, $[\alpha_*, 1]$
	+Mistrans.	7.3.1			✓			✓	Unique (α_*, β_*)
	+Discarding (large N_1,N_2)	7.3.2	✓	✓	✓	✓			{(0,0)}, {(0,0),(0,1)}, {(0,1),(1,1)} or {(1,1)}
Coupling by the lexicon		7.4	✓	✓	✓*	✓	✓	✓	6 possibilities
	+Mistrans.	7.4.1	✓	✓	✓*	✓	✓	✓	(see text)

Table 8: Summary of coupling models. Abbreviations:
{1,2}: {1,2} is a stable fixed point in a parameter region
{1,1}/{2,2}: each of {1,1} and {2,2} is a stable FP in some parameter region
*{2,1}: {2,1} ruled out
Obs bif= bifurcations from stable {1,1} to stable {1,2} and from stable {2,2} to stable {1,2} occur.
Freq-dep: frequency-dependent FP locations.
✓*: vacuously true (built into model).
α_*, β_*: Some point $\in (0,1)$ (interior point).

The last model (coupling by the lexicon with mistransmission) comes the closest of all models considered here to accounting for patterns seen in the N/V data:

- {1,2}, {1,1}, and {2,2} are fixed points for some values of system parameters, corresponding to (0,1), (0,λ), and (κ,1).
- Bifurcations occur where {1,1} or {2,2} lose stability, and the population moves to 100% {1,2}; but there are no bifurcations where the reverse happens ({1,2} becomes unstable, and the population moves to 100% {1,1} or {2,2}).
- Keeping lexical support parameters fixed, the loss of stability of {1,2} occurs as frequency (N_1, N_2) decreases.
- Stable variation is possible in one of the N or V form at once (depending on the values of λ and κ), but not both.

However, this model does not account for the fact that the less-common changes – {1,2}→{1,1} and {1,2}→{2,2} – can occur. It also does not give {2,1} as an unstable state; rather, {2,1} is simply ruled out by fiat (by setting $a_{21} = 0$).

This last point is interesting when compared with the results of all the coupling models. In each case, {2,1} is somehow disfavored (in some models not allowed at all); but how this dispreference is implemented in the model has consequences for the rest of the dynamics.

In coupling by grammar and coupling by constraint models, adding any mistransmission in a direction biased against {2,1} causes the dynamics to have only one fixed point, and thus no bifurcations. However, in lexical coupling models, adding mistransmission brings the dynamics closer to patterns observed in the data.

In the coupling by constraint model without mistransmission, constraining the learner to never hypothesize $\hat{\beta} > \hat{\alpha}$ causes trajectories to move along lines of constant $\alpha_t + \beta_t$, a pattern not observed in the data, and there are no bifurcations. When coupling by constraint is combined with discarding, the dynamics show bifurcations, but not those observed in the data.

8 Conclusions

The preceding discussion of coupling models considered shows that different properties of proposed learners interact in non-trivial ways in the population-level dynamics. We make no claim that the final model considered is what actually goes on in populations of English learners. It is meant as an "existence proof" in two senses. Most patterns seen in a relatively detailed dataset can be accounted for by a relatively simple learning model, inspired by a combination of linguistic hypotheses about why change occurs; and most patterns seen in the dataset are not accounted for by several other such models. Put otherwise, the task of linking model and data properties is not doomed to success.

The single-form models considered above lead to a similar conclusion. Bifurcations, multistability, frequency dependence, and the existence of stable interior points are the dynamical systems interpretation of fundamental aspects of many linguistic changes. Table 7, which summarizes the dynamics of single-form models, shows that while any one criterion is met by several models' dynamics, few models meet all four.

Implications for theories of change Because there are many ways any hypothesis about the causes of change (such as mistransmission or regularization) can be written down in a model, we cannot claim to have proved or disproved particular hypotheses in any sense. However, we can make some general observations based on the range of models considered.

Consider first mistransmission, our implementation of the most widely proposed type of explanation for sound change. Mistransmission alone gives only a single fixed point (§5.2), and hence no bifurcations. More generally, for a range of types of learners, adding mistransmission changes the dynamics, but not qualitatively so. As in §6.3.3, the boundary between fixed point regimes may change, but the fixed point regimes themselves do not. As in §7.4.1, adding mistransmission may change whether or not a particular bifurcation occurs as frequency is changed, but does not add new bifurcations. In sum, mistransmission usually does not change the number and type of bifurcations, or *bifurcation structure*, of the model. This is not to say that mistransmission is not important; for example, if adding mistransmission introduces frequency-dependence, the new model has an extra property observed in the data.

What frequently does determine a model's bifurcation structure is what the learner does with the data once they receive it. However, in most models considered, this is not enough for realistic dynamics. As in §6.1.2, it may be the case that stable variation is not possible unless mistransmission is added; or as in §5.5.3, that frequency dependence is not possible unless mistransmission is added.

Generalizing, we hypothesize that realistic models (in the sense of replicating patterns seen in linguistic change) are those which include both bias in the data (mistransmission) and bias in the learner (discarding, regularization, etc.) Interestingly, this distinction more or less corresponds to the two types of sources of typology and change proposed in recent years, *channel bias* and *analytic bias* (Moreton, 2008). Based on our models, we speculatively hypothesize that the debate between channel and analytic bias is misplaced: they are responsible for largely different properties of change (as observed in diachronic data), and a successful model must include both.

Finally, in line with previous work in the dynamical systems approach, as well as recent computational models of population-level change (Liberman, 2000; Daland et al., 2007; Baker, 2008a; Troutman et al., 2008), we note that the relationship between how individuals learn and the

population-level dynamics can be subtle and unexpected. Different proposed causes for change at the individual level, each of which seems plausible a priori, can have very different diachronic consequences. The non-trivial map from learning to population dynamics provides a potentially useful tool for testing theories of the causes of language change. It also has an important corollary: population-level models are necessary to evaluate any theory of why language change occurs.

Acknowledgements

I would like to thank Partha Niyogi, my advisor and collaborator on much of this work, for guidance, insight, and patience. I have also benefited from discussions with Max Bane, Adam Albright, John Goldsmith, Jason Riggle, and Alan Yu, and feedback from audiences at the University of Chicago, LabPhon 11, and Northwestern University.

References (non-dictionary)

Albright, A. (2008). How many grammars am I holding up? Discovering phonological differences between word classes. In Chang, C. and Haynie, H., editors, *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pages 1–20. Cascadilla.

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

Baker, A. (2008a). Addressing the actuation problem with quantitative models of sound change. *Penn Working Papers in Linguistics*, 14(1):1–13.

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

Baker, R. and Smith, P. (1976). A psycholinguistic study of English stress assignment rules. *Language and Speech*, 19(1):9–27.

Baptista, B. (1984). English stress rules and native speakers. *Language and Speech*, 27(3):217–33.

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press, Cambridge, UK.

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

Blevins, J. and Garrett, A. (1998). The origins of consonant-vowel metathesis. *Language*, 74(3):508–56.

Brink, L. and Lund, J. (1975). *Dansk rigsmål : lydudviklingen siden 1840 med særligt henblik på sociolekterne i København*. Gyldendal.

Bybee, J. (2003). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14(03):261–290.

Bybee, J. and Hopper, P., editors (2001). *Frequency and the emergence of linguistic structure*. Benjamins, Amsterdam.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Clopper, C. (2002). Frequency of stress patterns in English: A computational analysis. *Indiana University Linguistics Club Working Papers*, 2.

Cutler, A. and Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2:133–142.

Daland, R., Sims, A. D., and Pierrehumbert, J. (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.

Davis, S. and Kelly, M. (1997). Knowledge of the English noun–verb stress difference by native and nonnative speakers. *Journal of Memory & Language*, 36(3):445–460.

De Schryver, J., Neijt, A., Ghesquière, P., and Ernestus, M. (2008). Analogy, frequency, and sound change. The case of Dutch devoicing. *Journal of Germanic Linguistics*, 20(2):159–195.

Fudge, E. (1984). *English Word-stress*. Allen & Unwin.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition.

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

Griffiths, T. and Reali, F. (2008). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. In Love, B., McRae, K., and Sloutsky, V., editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 229–234. Cognitive Science Society.

Guion, S., Clark, J., Harada, T., and Wayland, R. (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.

Halle, M. and Keyser, S. (1971). *English stress*. Harper & Row.

Hammond, M. (1999). *The Phonology of English: A Prosodic Optimality-theoretic Approach*. Oxford University Press.

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

Harrison, K., Dras, M., and Kapicioglu, B. (2002). Agent-based modeling of the evolution of vowel harmony. In Hirotani, M., editor, *Proceedings of the Northeast Linguistic Society (NELS) 32*.

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

Hock, H. (1991). *Principles of historical linguistics*. Mouton de Gruyter.

Hoffmann, S. (2004). Using the OED quotations database as a corpus–a linguistic appraisal. *ICAME Journal*, 28(4):17–30.

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In Christie, W., editor, *Current Progress in Historical Linguistics*, pages 95–105. North-Holland.

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

Hudson Kam, C. and Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59:30–66.

Kelly, M. (1988a). Phonological biases in grammatical category shifts. *Journal of Memory & Language*, 27(4):343–358.

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

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

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

Kirby, S., Dowman, M., and Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.

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

Labov, W. (2000). *Principles of linguistic change. Vol. 2: Social factors*. Blackwell.

Ladefoged, P. and Fromkin, V. (1968). Experiments on competence and performance. *IEEE Transactions on Audio and Electroacoustics*, 16:130–136.

Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman.

Liberman, M. (2000). The 'lexical contract': modeling the emergence of word pronunciations. Ms., University of Pennsylvania.

Mitchener, W. (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. Association of Computational Linguistics.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, 25(01):83–127.

Nevalainen, T. and Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Prentice Hall.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

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

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

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

Niyogi, P. and Berwick, R. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106(25):10124–10129.

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

Ohala, J. (1983). The origin of sound patterns in vocal tract constraints. In MacNeilage, P., editor, *The Production of Speech*, pages 189–216. Springer, New York.

Pearl, L. (2007). *Necessary Bias in Language Learning*. PhD thesis, University of Maryland.

Pearl, L. (2008). Putting the emphasis on unambiguous: The feasibility of data filtering for learning English metrical phonology. In Chan, H., Jacob, H., and Kapia, E., editors, *BUCLD 32: Proceedings of the 32nd annual Boston University Conference on Child Language Development*, pages 390–401. Cascadilla.

Pearl, L. and Weinberg, A. (2007). Input Filtering in Syntactic Acquisition: Answers From Language Change Modeling. *Language Learning and Development*, 3(1):43–72.

Phillips, B. (1981). Lexical diffusion and Southern tune, duke, news. *American Speech*, 56(1):72–78.

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

Ross, J. (1972). A reanalysis of English word stress. In Brame, M., editor, *Contributions to Generative Phonology*, pages 229–323. University of Texas Press.

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

Schilling-Estes, N. (2003). Investigating stylistic variation. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 375–401. Wiley-Blackwell.

Schuchardt, H. (1885). *Über die Lautgesetze. Gegen die Junggrammatiker*. R. Oppenheim, Berlin.

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

Singleton, J. and Newport, E. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49(4):370–407.

Strogatz, S. (1994). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Addison-Wesley.

Troutman, C., Goldrick, M., and Clark, B. (2008). Social networks and intraspeaker variation during periods of language change. *Penn Working Papers in Linguistics*, 14(1):325–338.

Walch, M. (1972). Stress rules and performance. *Language and Speech*, 15(3):279–87.

Wang, W. (1969). Competing changes as a cause of residue. *Language*, 45(1):9–25.

Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W. and Malkiel, Y., editors, *Directions for historical linguistics*, pages 95–188. University of Texas Press.

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

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

Dictionaries

Anonymous (1753). *A pocket dictionary or complete English expositor...* London.

Anonymous (1763). *An universal dictionary of the English language...* Alexander Donaldson and John Reid, Edinburgh.

Anonymous (1991). *Longman Dictionary of the English Language*. Longman, Harlow, 2nd edition.

Anonymous (c2000). *The American Heritage dictionary of the English language*. Houghton Mifflin, Boston, 4th edition.

Ash, J. (1775). *The new and complete dictionary of the English language...* London.

Bailey, N. (1735). *An universal etymological English dictionary...* London, 7th edition.

Bailey, N., Gordon, G., and Miller, P. (1736). *Dictionarium Britannicum: or a more compleat universal etymological English dictionary...* London, 2nd edition.

Barclay, J. (1774). *A complete and universal English dictionary: on a new plan...* London.

Barclay, J. (1812). *A complete and universal dictionary of the English language*. Brightly & Childs, Bungay.

Boag, J. (1848). *The Imperial lexicon of the English language : exhibiting the pronunciation...* A. Fullarton & co., Edinburgh.

Boyer, A. (1700). *The royal dictionary: in two parts...* London.

Boyer, A. (1791). *Boyer's royal dictionary abridged...* London, 17th edition. Carefully corrected and improved ... by J. C. Prieur.

Boyer, A. (1819). *Boyer's royal dictionary abridged : Containing the greatest number of words...* London, 23rd edition.

British Broadcasting Corporation (1992). *BBC English dictionary.* BBC English/Harper Collins, London.

Burn, J. (1786). *A pronouncing dictionary of the English language.* Alex. Adam, Glasgow, 2nd edition.

Butler, C. (1634). *The English grammar, or the institution of letters, syllables, and woords in the English tung...* William Turner, Oxford.

Care, H. (1687). *The tutor to true English, or, Brief and plain directions...* George Larkin, London.

Chambers, W. and Chambers, R. (1872). *Chambers's English Dictionary...* W. & R. Chambers, London & Edinburgh.

Chambers, W., Chambers, R., and Macdonald, A. M. (1972). *Chambers twentieth century dictionary.* Chambers, Edinburgh, rev. edition.

Cooper, C. (1687). *The English teacher, or, The discovery of the art of teaching and learning the English tongue...* John Richardson, London.

Davidson, T. (1901). *Chambers's Twentieth Century Dictionary of the English Language...* W. & R. Chambers, London/Edinburgh.

Dyche, T. and Pardon, W. (1735). *A new general English dictionary...* London.

Emery, H. G. and Brewster, K. G. (c1927). *The new century dictionary of the English language...* Century, New York/London. 3 v.

Fenning, D. (1763). *The royal English dictionary: or, a treasury of the English language...* London, 2nd edition.

Flint, J. M. (1740). *Prononciation de la langue angloise...* Didot, Paris.

Fowler, H. W. and Fowler, F. G. (1911). *The concise Oxford dictionary of current English.* Clarendon, Oxford.

Fowler, H. W. and Fowler, F. G. (1951). *The concise Oxford dictionary of current English.* Clarendon, Oxford, 4th rev. edition.

Fulton, G. and Knight, G. (1833). *A dictionary of the English language, greatly improved...* Stirling and Kenney, Edinburgh.

Funk, I. K., editor (1893–1895). *A standard dictionary of the English language...* New York. 2 v.

Funk, I. K. (1958). *Funk and Wagnalls standard dictionary of the English language...* Encyclopaedia Brittanica, Chicago, International edition.

Geddie, W. (1952). *Chambers's Twentieth Century Dictionary.* W. & R. Chambers, Edinburgh/London, midcentury edition.

Hayward, A. L. and Sparkes, J. J. (1962). *Cassell's English Dictionary ...* Cassell, London.

Hunter, R. (1879–1888). *The encyclopaedic dictionary...* London. 7 v.

James, W. and Molé, A. (1847). *Dictionary of the English and French languages...* Leipzig.

Johnson, S. (1755). *A dictionary of the English language: in which the words are deduced from their originals...* W. Strahan, for J. and P. Knapton, London.

Johnson, S. (1756). *A dictionary of the English language...* London, 1st 8vo edition.

Johnson, S. (1775). *A dictionary of the English language...* Dublin, 4th edition.

Jones, D. (1917). *An English pronouncing dictionary (on strictly phonetic principles).* Dent, London.

Jones, D. and Gimson, A. C. (1977). *Everyman's English pronouncing dictionary...* Dent, London, 14th edition.

Jones, D., Roach, P., Hartman, J., and Setter, J. (c2003). *Cambridge English pronouncing dictionary.* Cambridge University Press, Cambridge.

Jones, S. (1798). *Sheridan improved: A general pronouncing and explanatory dictionary of the English language...* London, 3rd edition.

Jonson, B. (1640). *The English Grammar made by B. Johnson...* London.

Kenrick, W. (1773). *A new dictionary of the English language...* London.

Kenyon, J. S. and Knott, T. A. (1944). *A pronouncing dictionary of American English.* Merriam, Springfield, Mass.

Knowles, J. (1835). *A pronouncing and explanatory dictionary of the English language...* F. de Porquet and Cooper, London.

Levens, P. (1570). *Manipulus vocabulorum: A dictionarie of English and Latine wordes...* Henrie Bynneman, London.

Ludwig, C. (1706). *A dictionary English, German and French...* Thomas Fritschen, Leipzig.

Ludwig, C. (1763). *A dictionary English, German and French...* Ulrich Christian Saalbach, Leipzig, 3rd edition.

Martin, B. (1749.). *Lingua Britannica reformata: or, a new English dictionary...* London.

Minsheu, J. (1617). *Ductor in linguas, the guide into tongues...* J. Broune, London.

Mulcaster, R. (1582). *The first part of the elementarie: vvhich entreateth chefelie of the right writing of our English tung...* Thomas Vautroullier, London.

Nares, R. (1784). *Elements of orthoepy: containing a distinct view...* London.

Ogilvie, J. and Annandale, C. (1882). *The imperial dictionary of the English language ...* Blackie, London.

Ogilvie, J. and Cull, R. (1862). *The comprehensive English dictionary, explanatory, pronouncing, & etymological...* Blackie and Son, London.

Perry, W. (1805). *Synonymous, etymological, and pronouncing English dictionary.* London,.

Price, O. (1665). *The vocal organ.* William Hall, Oxford.

Pryse, R. J. (1857). *An English and Welsh pronouncing dictionary : in which the pronunciation is given in Welsh letters...* T. Gee, Dinbych.

Reader's Digest Association (1985). *Reader's Digest great illustrated dictionary.* Reader's Digest, London, 2nd edition.

Reid, A. (1844). *A dictionary of the English language. : A vocabulary of the roots of English words...* Edinburgh.

Sheridan, T. (1780). *A general dictionary of the English language...* London.

Smith, A. H. and O'Loughlin, J. L. N. (1965). *Odhams dictionary of the English language...* Odham, London.

Spence, T. (1775). *The grand repository of the English language...* T. Saint, Newcastle upon Tyne.

Stein, J. and Urdang, L. (1967). *The Random House dictionary of the English language.* Random House, New York.

Stormonth, J. (1879). *Etymological and pronouncing dictionary of the English language...* W. Blackwood and sons, Edinburgh/London, 5th edition.

Walker, J. (1775). *A dictionary of the English language...* London.

Walker, J. (1791). *A critical pronouncing dictionary and expositor of the English language.* London, 1st edition.

Walker, J. (1802). *A critical pronouncing dictionary and expositor of the English language.* Oriental Press, London, 3rd edition.

Walker, J. and Smart, B. H. (1836). *Walker remodelled. A new critical pronouncing dictionary of the English language, adapted to the present state of literature and science...* London.

Webster, N. (1828). *An American dictionary of the English language...* S. Converse, New York.

Webster, N., Harris, W. T., and Allen, F. (1913.). *New international dictionary of the English language based on...* G. Bell, London.

Webster, N., Harris, W. T., and Porter, N. (1902). *International Dictionary of the English Language ...* G & C. Merriam & Co., London.

Webster, N., Neilson, W. A., Knott, T. A., and Carhart, P. W. (1934). *New international dictionary of the English language.* G. Bell & Sons, London,, 2nd edition.

Whitney, W. D. (1889–1909). *The Century Dictionary. An encyclopedic lexicon of the English language...* Century, New York.

Worcester, J. E. (1859). *A dictionary of the English language.* Sampson Low, Son and Marston, London.

Wyld, H. C. K. (1932). *The universal dictionary of the English language ...* Routledge, London.

Appendixes

A Dictionary List

Code	Reference	Brit/Am
L1570	Levens (1570)	B
M1582	Mulcaster (1582)	B
M1617	Minsheu (1617)	B
B1634	Butler (1634)	B
J1640	Jonson (1640)	B
P1665	Price (1665)	B
C1687	Cooper (1687)	B
Ca1687	Care (1687)	B
B1700	Boyer (1700)	B
L1706	Ludwig (1706)	B
B1735	Bailey (1735)	B
D1735	Dyche and Pardon (1735)	B
B1736	Bailey et al. (1736)	B
F1740	Flint (1740)	B
M1749	Martin (1749)	B
P1753	Anonymous (1753)	B
J1755	Johnson (1755)	B
J1756	Johnson (1756)	B
F1763	Fenning (1763)	B
L1763	Ludwig (1763)	B
U1763	Anonymous (1763)	B
K1773	Kenrick (1773)	B
B1774	Barclay (1774)	B
A1775	Ash (1775)	B
J1775	Johnson (1775)	B
S1775	Spence (1775)	B
W1775	Walker (1775)	B
S1780	Sheridan (1780)	B
N1784	Nares (1784)	B
B1786	Burn (1786)	B
W1791	Walker (1791)	B
B1791	Boyer (1791)	B
J1798	Jones (1798)	B
W1802	Walker (1802)	B
P1805	Perry (1805)	B
B1812	Barclay (1812)	B
B1819	Boyer (1819)	B
W1828	Webster (1828)	A
FK1833	Fulton and Knight (1833)	B
K1835	Knowles (1835)	B
S1836	Walker and Smart (1836)	B
R1844	Reid (1844)	B
B1847	Boag (1848)	B
P1857	Pryse (1857)	B
W1859	Worcester (1859)	A
O1864	Ogilvie and Cull (1862)	B
C1872	Chambers and Chambers (1872)	B
H1879	Hunter (1888)	B
S1879	Stormonth (1879)	B
I1882	Ogilvie and Annandale (1882)	B
C1889	Whitney (1909)	A
F1893	Funk (1895)	A
C1901	Davidson (1901)	B
W1902	Webster et al. (1902)	A
O1911	Fowler and Fowler (1911)	B
W1912	Webster et al. (1913)	A
J1917	Jones (1917)	B
C1927	Emery and Brewster (1927)	A
U1932	Wyld (1932)	B
W1934	Webster et al. (1934)	A
PD1944	Kenyon and Knott (1944)	A
O1951	Fowler and Fowler (1951)	B
C1952	Geddie (1952)	B
FW1958	Funk (1958)	A
C1962	Hayward and Sparkes (1962)	B
O1965	Smith and O'Loughlin (1965)	B
RH1967	Stein and Urdang (1967)	A
C1972	Chambers et al. (1972)	B
J1977	Jones and Gimson (1977)	B
R1984	Reader's Digest Association (1985)	B
RU1984	Reader's Digest Association (1985)	A
L1991	Anonymous (1991)	B
B1992	British Broadcasting Corporation (1992)	B
AH2000	Anonymous (2000)	A
CB2003	Jones et al. (2003)	B
CA2003	Jones et al. (2003)	A

B Word lists

List 1: Word list from Sherman (1975)

Script indicates first reported pronunciation: {1,1}, **{2,2}**, {*1,2*}

abstract	*compact*	**contrast**	**discharge**	*impact*	*invert*	**permit**	*rebel*	**rehash**	*torment*
accent	*compound*	*converse*	discord	**import**	*legate*	*pervert*	**rebound**	**reject**	transfer
addict	**compress**	*convert*	**discount**	**impress**	**misprint**	postdate	**recall**	*relapse*	**transplant**
address	**concert**	*convict*	**discourse**	*imprint*	*object*	**prefix**	**recast**	**relay**	*transport*
affect	*concrete*	*convoy*	*egress*	*incense*	*outcast*	**prelude**	**recess**	**repeat**	*transverse*
affix	*conduct*	**decoy**	**eject**	**incline**	*outcry*	*premise*	**recoil**	*reprint*	traverse
alloy	*confect*	**decrease**	**escort**	**increase**	*outgo*	presage	*record*	**research**	**undress**
ally	*confine*	**defect**	*essay*	**indent**	*outlaw*	*present*	**recount**	**reset**	*upcast*
annex	*conflict*	**defile**	**excerpt**	*infix*	**outleap**	*produce*	**redraft**	sojourn	*upgrade*
assay	*conscript*	**descant**	**excise**	*inflow*	*outlook*	progress	**redress**	*subject*	*uplift*
bombard	*conserve*	*desert*	*exile*	**inlay**	*outpour*	*project*	*refill*	sublease	upright
cement	*consort*	**detail**	**exploit**	*inlet*	*outspread*	**protest**	**refit**	**sublet**	**uprise**
collect	**content**	dictate	*export*	*insert*	*outstretch*	purport	**refund**	**surcharge**	*uprush*
combat	*contest*	**digest**	**extract**	*inset*	*outwork*	**rampage**	**refuse**	**survey**	*upset*
commune	**contract**	**discard**	*ferment*	*insult*	**perfume**	**rebate**	**regress**	**suspect**	

List 2: Sample of words in use 1700–2007

Script indicates 1700 pronunciation Boyer (1700), *=changed by 2007.[38]

abuse	bottom	*contest*	envy	harbour	measure	**proceed***	**repeal**	table	whistle
accent	breakfast	*contract*	*exile**	hollow	mention	**protest***	**repose**	tally	witness
advance	buckle	*convict*	**express**	**import***	merit	purchase	**reserve**	thunder	
affront	bundle	cover	favour	**increase***	motion	puzzle	**review**	title	
ally*	butter	**decrease***	ferret	interest	murder	quarry	rival	*torment*	
anchor	*cement**	**decree**	flourish	iron	muster	reason	saddle	travel	
arrest	challenge	diet	**forecast***	journey	order	**redress**	second	treble	
assault	channel	**digest***	forward	level	outlaw	**reform**	shiver	triumph	
assay	**command**	**dispatch**	gallop	levy	pepper	**regard**	shoulder	trouble	
attack	**concern**	**dissent**	glory	licence	plaster	**relapse***	squabble	value	
bellow	*conduct*	**distress**	hammer	license	*premise**	relish	stable	visit	
blunder	*consort*	double	handle	matter	*present*	**remark**	stomach	vomit	

List 3: Control set of words pronounced {2,2} in 1700

abuse	ally	attack	concern	decease	demand	disdain	dispose	excuse	rebuke
accord	amend	attaint	concert	decline	demise	disease	dispraise	exempt	recoil
account	appeal	attempt	consent	decoy	design	disgrace	dispute	exploit	record
address	approach	award	content	decrease	desire	disguise	dissent	express	recruit
advance	array	command	control	decree	despair	disgust	distaste	proceed	redoubt
affront	arrest	compare	debate	defeat	devise	dislike	distress	protest	redress
alarm	assault	compute	debauch	delay	discharge	dispatch(despise)	distrust	rebate	reflect
allay/alloy	assay	conceit	decay	delight	discourse	display	exchange	rebound	reform

[38] 2007 pronunciations from Cambridge Advanced Learner's Dictionary, OED. Words were randomly chosen from all N/V pairs which (a) have both N and V frequency of at least 1 per million in the BNC Leech et al. (2001) (b) have both N and V forms listed in a dictionary from 1700 Boyer (1700) (c) have both N and V forms listed in a dictionary from 1847 (James and Molé (1847)).
* indicates the pronunciations listed in the 1700, 1847, and 2007 dictionaries are not identical.

refrain	release	repair	repose	repute	resort	retreat	review
regard	remain(s)	repeal	reprieve	request	respect	return	revise
regret	remark	reply	reproach	reserve	result	revenge	revolt
relapse	remove	report	repulse	resolve	retort	reverse	reward

List 4: Control set of words pronounced {2,2} in 1847

abuse	assay	debate	defy	discourse	disport	excuse	reform	repeal	result
accord	attack	debauch	delay	disdain	dispose	exploit	refrain	reply	retort
account	attempt	decay	delight	disease	dispraise	express	refuse	report	retreat
advance	award	decease	demand	disgrace	dispute	proceed	regard	repose	return
affront	command	decline	demise	disguise	dissent	protract	regret	reproach	revenge
ally	compare	decoy	demur	disgust	distaste	rebound	relapse	repulse	reverse
annex	consent	decrease	design	dislike	distress	rebuke	release	request	review
appeal	conserve	decree	desire	dismay	distrust	recoil	remain(s)	reserve	revise
array	content	default	despair	dispatch(despise)	disuse	record	remark	resolve	revolt
arrest	control	defeat	despite	dispense	excerpt	recruit	remove	resort	reward
assault	control	defile	discharge	display	excise	redress	repair	respect	

C Radio stories

All stories streamed from `npr.org`, except those labeled 'BBC' (streamed from `bbc.co.uk`).

Word	Story Title	Date	Speakers
address	"Fake Addresses"	4/17/2003	F15
address	"AOL Sues Spam Distributors"	4/15/2003	M16
address	"Spam Blocker Failure"	12/6/2006	M15
address	"How to Get Red of Spam and Junk E-Mail"	5/9/2003	M17
increase	"Report: Being 'Unmarried with Children' Increasingly Popular"	12/11/2007	F13
increase	"New Focus on Homeless Attacks, Victims.."	7/25/2007	F13
increase	"Politics of Humanitarian Aid and Hamas.."	3/19/2006	F13
increase	"Youth Violence An Issue of Public Health?"	6/4/2007	F13
increase	"New Stamp Wouldn't Need A Rate Upgrade"	5/4/2006	F4
increase	"Watchdog Blasts Bush's National Parks Policy"	6/11/2003	F4
increase	"What are CEO's Worth?"	4/17/2006	F13
increase	"Nation's Health Care Bill Hits All-Time High"	1/8/2008	F1
increase	"'Marketplace' Report: Gas Tax"	1/15/2008	F14
increase	"U.S. Boosts Use of Airstrikes in Iraq.."	1/17/2008	F14
increase	"Cold Medicines Targets in Meth.."	9/26/06	F14
increase	"Pregnancy Discrimination Increases in Workplace"	5/22/2007	F14
increase	"Iraqis React to Effect of U.S. Troop.."	8/2/2007	F14
increase	"Teen Birth Rate Spikes After 14 Years of Decline"	12/18/2007	M12
perfume	"Perfume Gallery Preserves, Re-Creates Fragrances"	11/5/2006	F11, F7
perfume	"Pez Perfume"	1/9/2000	F12
perfume	"Christmas Boosts Perfume Sales" (BBC Video)	12/22/2006	F8
perfume	"In New York, Eau de Borough"	4/20/2004	F9
perfume	"Smells like.."	12/19/1998	F10
perfume	"Perfume Master"	5/5/2002	F10
perfume	"Love in the Days of Shalimar"	11/29/2004	M13
perfume	"The 'Times' Gets a Scent Critic"	8/28/2006	M12, M14
research	"Research Funding Cutoff"	7/19/2001	M8
research	"Scientist Admits Faking Data"	3/18/2005	F1
research	"Health Research Funding Call" (BBC)	11/12/2007	F8, M4
research	"Medic Defends Research Letter" (BBC)	11/11/2007	M9
research	"Funding Campus Research: Conflict of Interest?"	10/19/2007	M1, M5, M7, M10, F2
research	"The Next Horizon in Stem Cell Research"	11/30/2007	M1
research	"Study Finds Conflicts of Interest in Medical Research"	11/29/2006	F5, M7
research	"The Ethics of Medical Research on Children"	10/31/2006	M8
research	"Katrina's Effect on Scientific Research"	10/21/2005	M1, M11
research	"Access to Research Data"	3/5/1999	M1, F3, F6
research	"Opposition Research: Know Thine Enemies"	2/6/2007	F4
research	"Top Stem Cell Researcher Resigns After Ethical Lapse"	11/24/2005	M8
research	"Drug Companies Balk at Flu Vaccine"	2/10/2004	F5
research	"Treating Heart Disease"	12/6/2006	F5
research	"FDA's handling of diabetes drug reviewed"	1/6/2007	F1
research	"FDA criticized for diabetes drug Avandia"	5/22/2007	F1
research	"Hopes for 'good' cholesterol drug defy bad tests"	5/27/2007	F1
research	"Study: Tastes Form in Infancy"	4/5/2004	F4
research	"Opening Statements in Vioxx Wrongful Death"	7/14/2005	F5
research	"Part 1: Documents Suggest Merck Tried to Censor Vioxx Critics"	6/9/2005	F5
research	"Part 2: Did Merck Try to Censor Vioxx Critics?"	6/9/2005	F5
research	"Drawing the Line Between Science and Politics"	7/27/2007	M8, M10
research	"Government Science Advisory Committees"	1/10/2003	M1, M2, M6

D Radio pronunciation data

Pronunciations of "research", "perfume", "address", "increase" observed in radio stories. A=American, B=British, I=Indian, O=non-native.

Speaker	Dialect	Word	N=1	N=2	V=1	V=2
M1	A	research	25	0	0	0
F1	A	research	11	0	0	0
F2	A	research	13	0	1	0
F3	A	research	7	0	0	0
F4	A	research	10	0	3	0
M2	A	research	12	0	0	0
M3	B	research	6	0	0	0
M4	B	research	6	0	0	0
F5	I	research	15	0	1	0
M5	A	research	4	2	0	0
M6	A	research	4	2	0	0
M7	A	research	5	4	0	0
M8	A	research	4	9	0	0
M9	B	research	2	7	0	0
F6	A	research	3	15	0	0
M10	A	research	0	12	0	0
M11	A	research	0	5	0	0
F7	O	perfume	7	0	0	0
F8	B	perfume	5	0	0	0
F9	A	perfume	7	3	0	0
F10	A	perfume	6	6	0	0
M12	A	perfume	2	6	0	0
F11	A	perfume	0	8	0	0
F12	A	perfume	0	10	0	0
M13	A	perfume	0	5	0	0
M14	A	perfume	0	16	0	0
F13	A	increase	5	0	0	1
F14	A	increase	6	1	1	3
F15	A	address	8	0	0	0
M15	address	9	1			
M16	A	address	0	10	0	0
M17	A	address	0	7	0	0

E Proofs

E.1 Section 5.5.1

We show that

$$\sum_{k_2=1}^{N}\binom{N}{k_2}{p_{2,t}}^{k_2}(p_{3,t})^{N-k_2}+\sum_{k_2=1}^{N}\sum_{k_1=1}^{N-k_2}\binom{N}{k_1,k_2}{p_{1,t}}^{k_1}{p_{2,t}}^{k_2}(p_{3,t})^{N-k_1-k_2}\frac{k_2}{k_1+k_2}=\frac{p_{2,t}}{p_{1,t}+p_{2,t}}(1-{p_{3,t}}^{N}) \tag{61}$$

First,

$$\begin{aligned}\sum_{k_2=1}^{N}{p_{2,t}}^{k_2}{p_{3,t}}^{N-k_2} &= \sum_{k_2=0}^{N}{p_{2,t}}^{k_2}{p_{3,t}}^{N-k_2}-p_{3,t}^{N}\\ &= (p_{2,t}+p_{3,t})^{N}-p_{e,t}^{N}\end{aligned} \tag{62}$$

Next,

$$\begin{aligned}&\sum_{k_1=1}^{N}\sum_{k_2=1}^{N-k_1}\frac{k_2}{k_1+k_2}\binom{N}{k_1,k_2}{p_{1,t}}^{k_1}{p_{2,t}}^{k_2}{p_{3,t}}^{N-k_1-k_2}\\
=&\sum_{k_1+k_2=2}^{N}\sum_{k_2=1}^{(k_1+k_2)-1}\frac{k_2}{k_1+k_2}\binom{N}{k_1+k_2}\binom{k_1+k_2}{k_2}{p_{1,t}}^{(k_1+k_2)-k_2}{p_{2,t}}^{k_2}{p_{3,t}}^{N-(k_1+k_2)}\\
=&\sum_{k=2}^{N}\sum_{j=1}^{k-1}\frac{j}{k}\binom{N}{k}\binom{k}{j}{p_{1,t}}^{k-j}{p_{2,t}}^{j}{p_{3,t}}^{N-k}\\
=&\sum_{k=2}^{N}\frac{1}{k}\binom{N}{k}p_{3,t}^{N-k}\sum_{j-1=0}^{k-2}\frac{k}{k}j\binom{k-1}{j-1}{p_{1,t}}^{k-j}{p_{2,t}}^{j}\\
=&\,p_{2,t}\sum_{k=2}^{N}\binom{N}{k}{p_{3,t}}^{N-k}((p_{1,t}+p_{2,t})^{k-1}-p_{2,t}^{k-1})\\
&\vdots\\
=&\frac{p_{2,t}}{p_{1,t}+p_{2,t}}-\frac{p_{2,t}{p_{3,t}}^{N}}{(1-p_{3,t})}-{p_{3,t}}^{N-1}p_{2,t}-(p_{3,t}+p_{2,t})^{N}+{p_{3,t}}^{N}+{p_{3,t}}^{N-1}p_{2,t}\\
=&\frac{p_{2,t}}{p_{1,t}+p_{2,t}}(1-p_{3,t})^{N}+p_{3,t}^{N}-(p_{2,t}+p_{3,t})^{N}\end{aligned} \tag{63}$$

Adding (62) and (63) gives the result in (61).

E.2 Section 5.5.3

The evolution equation is

$$\begin{aligned}
E[\hat{\alpha}_t] &= P(N=0)r \\
&\quad + \sum_{N=1}^{\infty} P(N)[p(k_1+k_2=0)\cdot r + \underline{\sum_{k_2=1}^{N} {p_{2,t}}^{k_2}{p_{3,t}}^{N-k_2}\cdot 1} \\
&\quad \underline{+ \sum_{k_1=1}^{N}\sum_{k_2=1}^{N-k_1} \frac{k_2}{k_1+k_2}\binom{N}{k_1,k_2}{p_{1,t}}^{k_1}{p_{2,t}}^{k_2}{p_{3,t}}^{N-k_1-k_2}} \\
&= e^{-\lambda}r + \sum_{N=1}^{\infty}\frac{\lambda^N e^{-\lambda}}{N!}\left[{p_{3,t}}^N r + \frac{p_{2,t}}{p_{1,t}+p_{2,t}}(1-{p_{3,t}}^N)\right] \qquad (64) \\
&= e^{-\lambda}\left[r + \sum_{N=1}^{\infty}\frac{\lambda^N}{N!}[({p_{3,t}}^N r + \frac{p_{2,t}}{p_{1,t}+p_{2,t}}(1-{p_{3,t}}^N)\right] \\
&= e^{-\lambda}[r + (r - \frac{p_{2,t}}{p_{1,t}+p_{2,t}})\sum_{N=1}^{\infty}\frac{(\lambda p_{3,t})^N}{N!} + \frac{p_{2,t}}{p_{1,t}+p_{2,t}}\sum_{N=1}^{\infty}\frac{\lambda^N}{N!}] \\
&= e^{-\lambda}[r + (r - \frac{p_{2,t}}{p_{1,t}+p_{2,t}})(e^{\lambda p_{3,t}}-1) + \frac{p_{2,t}}{p_{1,t}+p_{2,t}}(e^{\lambda}-1)] \qquad (\text{using } \sum_{k=0}^{\infty}\frac{x^k}{k!} = e^x) \\
&\vdots \\
&= \frac{p_{2,t}}{p_{1,t}+p_{2,t}} + e^{-\lambda p_{3,t}}(r - \frac{p_{2,t}}{p_{1,t}+p_{2,t}})
\end{aligned}$$

where the underlined term is calculated above (App. E.2).

E.3 Section 6.3.3

- $A < 1$, $B > 1$: In this case the posterior mode when $k = 0$ is 0, so that

$$\begin{aligned}
\alpha_{t+1} &= \sum_{k=1}^{n}\binom{n}{k}{\alpha_t}^k(1-\alpha_t)^{n-k}\frac{A+k-1}{A+B+n-2} \\
&= \frac{A-1}{A+B+n-2} + \alpha_t\frac{n}{A+B+n-2} - (1-\alpha_t)^n\frac{A-1}{A+B+n-2}
\end{aligned}$$

  and the evolution equation is

$$\alpha_{t+1} = \frac{1}{A+B+n-2}[n\alpha_t + (1-a)((1-\alpha_t)^n - 1)]$$

  This is (33).

- $A > 1$, $B < 1$: By symmetry to the above case, 1 is the unique fixed point.
- $A > 1$, $B > 1$: In this case, the posterior mode is 1 when $k = n$ and 0 when $k = 0$, so that

$$\begin{aligned}
\alpha_{t+1} &= {\alpha_t}^n + \sum_{k=1}^{n}\binom{n}{k}{\alpha_t}^k(1-\alpha_t)^{n-k}\frac{A+k-1}{A+B+n-2} \\
&= {\alpha_t}^n + \frac{A-1}{A+B+n-2} + \alpha_t\frac{n}{A+B+n-2} - (1-\alpha_t)^n\frac{A-1}{A+B+n-2} - {\alpha_t}^n\frac{A+n-1}{A+B+n-2}
\end{aligned}$$

which gives the evolution equation

$$\alpha_{t+1} = \frac{1}{A+B+n-2}\left[n\alpha_t + (A-1)(1-(1-\alpha_t)^n) + (B-1){\alpha_t}^n\right]$$

This is (34).

E.4 Section 7.1

From (39), $\gamma_i = (\frac{k_2}{N_2} - \frac{k_1}{N_1}) + \beta_i$. Now,

- If $\frac{k_2}{N_2} - \frac{k_1}{N_1} > 0$: γ_i is minimized by taking $\beta_i = 0$ and $\gamma_i = \frac{k_2}{N_2} - \frac{k_1}{N_1} \Rightarrow \alpha_i = \frac{k_1}{N_1}$.
- If $\frac{k_2}{N_2} - \frac{k_1}{N_1} < 0$: γ is minimized by taking $\gamma_i = 0$ and $\beta_i = \frac{k_1}{N_1} - \frac{k_2}{N_2} \Rightarrow \alpha = \frac{k_2}{N_2}$.

which is the learning algorithm claimed.

Now, the probability of a learner in generation $t+1$ hearing k_1 examples as N=1 and $N_1 - k_1$ examples as N=2, and similarly for k_2, $N_2 - k_2$ for verbs, is

$$P(k_1,k_2) \equiv \binom{N_1}{k_1}\binom{N_2}{k_2}(\alpha_t+\beta_t)^{k_1}(1-\alpha_t-\beta_t)^{N_1-k_1}(\alpha_t+\gamma_t)^{k_2}(1-\alpha_t-\gamma_t)^{N_2-k_2}$$

The evolution equations are then:

$$\alpha_{t+1} = E[\hat{\alpha}_t] = \sum_{\frac{k_1}{N_1}<\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_2}{N_2} \tag{65}$$

$$\beta_{t+1} = E[\hat{\beta}_t] = \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) \tag{66}$$

$$\gamma_{t+1} = E[\hat{\gamma}_t] = \sum_{\frac{k_1}{N_1}<\frac{k_2}{N_2}} P(k_1,k_2)(\frac{k_2}{N_2} - \frac{k_1}{N_1}) \tag{67}$$

Adding (65) and (66) gives (40), and adding (66) and (67) gives (41).

E.5 Section 7.2

We find the fixed points of the evolution equations

$$\alpha_{t+1} = [(\alpha_t+\beta_t)(1-a) + (\gamma_t+\delta_t)b] * [(\alpha_t+\delta_t)(1-c) + (\beta_t+\gamma_t)d] \tag{68}$$

$$\beta_{t+1} = [(\alpha_t+\beta_t)(1-a) + (\gamma_t+\delta_t)b] * [(\alpha_t+\delta_t)c + (\beta_t+\gamma_t)(1-d)] \tag{69}$$

$$\gamma_{t+1} = [(\alpha_t+\beta_t)a + (\gamma_t+\delta_t)(1-b)] * [(\alpha_t+\delta_t)c + (\beta_t+\gamma_t)(1-d)] \tag{70}$$

$$\delta_{t+1} = [(\alpha_t+\beta_t)a + (\gamma_t+\delta_t)(1-b)] * [(\alpha_t+\delta_t)(1-c) + (\beta_t+\gamma_t)d] \tag{71}$$

The fixed points $(\alpha^*, \beta^*, \gamma^*, \delta^*)$ satisfy

$$\alpha^* = E(\hat{\alpha}), \quad \beta^* = E(\hat{\beta}), \quad \gamma^* = E(\hat{\gamma}), \quad \delta^* = E(\hat{\delta}).$$

Define the quantities $A = (\alpha\gamma - \beta\delta)$ and $A^* = \alpha^*\gamma^* - \beta^*\delta^*$. Some algebra with (68) then gives the equilibrium condition

$$A^*(1-a-b)(1-c-d) = (ac-a-c)\alpha^* + d(1-a)\beta^* + b(1-c)\delta^* + bd\gamma^* \tag{72}$$

Note that

$$\alpha^* + \beta^* + \gamma^* + \delta^* = 1 \tag{73}$$

Eqns. (69)-(71) are the same as (68) under these changes of parameters:

$$68 \to 69: \quad (\alpha,\beta,\gamma,\delta) \to (\beta,\alpha,\delta,\gamma), \quad (a,b,c,d) \to (a,b,d,c) \tag{74}$$

$$68 \to 70: \quad (\alpha,\beta,\gamma,\delta) \to (\gamma,\delta,\alpha,\beta), \quad (a,b,c,d) \to (b,a,d,c) \tag{75}$$

$$68 \to 71: \quad (\alpha,\beta,\gamma,\delta) \to (\delta,\gamma,\beta,\alpha), \quad (a,b,c,d) \to (b,a,c,d) \tag{76}$$

Under transformations (74) and (76), $A \to A$, while under (75), $A \to -A$. Thus, applying (74) and (75) to (69), (70), and A gives the 3 independent equilibrium conditions (the fourth, Eqn. 71, is eliminated because of the sum-to-one constraint (73))

$$A(1-a-b)(1-c-d) = (ac-a-c)\alpha^* + d(1-a)\beta^* + b(1-c)\delta^* + bd\,\gamma^* \tag{77}$$

$$-A(1-a-b)(1-c-d) = c(1-a)\alpha^* + (ad-a-d)\beta^* + b(1-d)\gamma^* + bc\,\delta^* \tag{78}$$

$$A(1-a-b)(1-c-d) = ac\,\alpha^* + a(1-d)\beta^* + (bd-b-d)\gamma^* + c(1-b)\delta^*. \tag{79}$$

Now, adding (78) and (79) gives

$$\begin{aligned} c(\alpha^* + \delta^*) &= d(\beta^* + \gamma^*) \\ \Rightarrow \beta^* + \gamma^* &= \frac{c}{c+d}, \end{aligned} \tag{80}$$

using (73). Similarly, adding (77) and (78) gives

$$\beta^* + \alpha^* = \frac{b}{a+b}. \tag{81}$$

From (73), (80), and (81), choosing β^* uniquely determines $(\alpha^*, \beta^*, \gamma^*, \delta^*)$. Choose β^* – substituting into the definition of A then gives, after some algebra,

$$A = \frac{bc}{(a+b)(c+d)} - \beta \tag{82}$$

Using (73), (80), (81), and (82) to write α^*, γ^*, and δ^* in terms of β^*, (77) eventually simplifies to

$$(1-a-b)(1-c-d)[bc - (a+b)(c+d)\beta] = -[bc - \beta(a+b)(c+d)] \tag{83}$$

Assuming that the probability of mistransmission is never 1, $|1-a-b| < 1$ and $|1-c-d| < 1$, and (83) has no solution if $[bc - (a+b)(c+d)\beta] \neq 0$. This quantity must therefore $= 0$, giving

$$\beta^* = \frac{bc}{(a+b)(c+d)}$$

and the unique equilibrium

$$(\alpha^*, \beta^*, \gamma^*, \delta^*) = \frac{1}{(a+b)(c+d)}(bd, bc, ac, ad) \tag{84}$$

This solution is only valid if two conditions hold: $a+b \neq 0$ and $c+d \neq 0$. This leads to four cases, described without proof:

1. $(a+b \neq 0)$, $(c+d \neq 0)$: Unique stable equilibrium point, given by (84).
2. Case 2: $(a+b=0)$, $(c+d \neq 0)$: Line of fixed points
3. Case 3: $(a+b \neq 0)$, $(c+d=0)$: Line of fixed points.
4. Case 4: $(a+b=0)$, $(c+d=0)$: Two-dimensional manifold of fixed points.

E.6 Section 7.3.1

We prove that the evolution equations

$$\alpha_{t+1} = a_{12} + \alpha_t(1 - a_{12} - a_{21}) - \frac{1}{2} \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P_t(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) \tag{85}$$

$$\beta_{t+1} = b_{12} + \beta_t(1 - b_{12} - b_{21}) + \frac{1}{2} \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P_t(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) \tag{86}$$

have a unique, stable fixed point, where

$$P_t(k_1, k_2) = \binom{N_1}{k_1}\binom{N_2}{k_2} p_{N,t}^{k_1}(1 - p_{N,t})^{N_1 - k_1} p_{V,t}^{k_2}(1 - p_{v,t})^{N_2 - k_2}.$$

Adding (85) and (86) gives

$$E(\hat{\alpha}_t + \hat{\beta}_t) - (\alpha_t + \beta_t) = a_{12} + b_{12} - \alpha_t(a_{12} + a_{21}) - \beta_t(b_{12} + b_{21}),$$

so the line

$$\alpha(a_{12} + a_{21}) + \beta(b_{12} + b_{21}) = a_{12} + b_{12} \tag{87}$$

is a nullcline for the direction $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Letting $nc_1 = E(\hat{\alpha}_t + \hat{\beta}_t) - (\alpha_t + \beta_t)$, some algebra shows that $nc_1(0,0) \geq 0$, $nc_1(1,1) \leq 0$, and $\frac{\partial nc_1}{\partial \alpha_t}$, $\frac{\partial nc_1}{\partial \beta_t} \leq 0$, so all trajectories tend toward the line (87).

Subtracting (86) from (85) gives

$$E(\hat{\alpha}_t - \hat{\beta}_t) - (\alpha_t - \beta_t) = a_{12} - b_{12} + \beta_t(b_{12} + b_{21}) - \alpha_t(a_{12} + a_{21}) \sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P_t(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2})$$

Let $nc_2 = E(\hat{\alpha}_t - \hat{\beta}_t) - (\alpha_t - \beta_t)$. Some algebra shows that $nc_2(1,0) \leq 0$, $nc_2(0,1) \geq 0$, $\frac{\partial nc_2}{\partial \alpha_t} < 0$, $\frac{\partial nc_2}{\partial \beta_t} > 0$, so nc_2 defines a nullcline (though it is not a line) $nc_2 = 0$, or

$$\sum_{\frac{k_1}{N_1} > \frac{k_2}{N_2}} P_t(k_1, k_2)(\frac{k_1}{N_1} - \frac{k_2}{N_2}) = a_{12} - b_{12} + \beta(b_{12} + b_{21}) - \alpha(a_{12} + a_{21}), \tag{88}$$

and all trajectories tend toward (88). There is thus a unique stable equilibrium point (α^*, β^*), located at the intersection of (87) and (88).

E.7 Section 7.3.2

To calculate the stability of these fixed points, first define $g_1(\alpha)$, $g_2(\beta)$ as

$$g_1(\alpha) = \frac{\alpha(1-r_1)}{(1-r_2)+\alpha(r_2-r_1)}, \quad g_2(\beta) = \frac{\beta(1-s_1)}{(1-s_2)+\beta(s_2-s_1)}, \quad f_\rho(\alpha,\beta) = \sum_{\frac{k_1}{k_1+l_1}>\frac{k_2}{k_2+l_2}} P_t\left(\frac{k_1}{k_1+l_1}-\frac{k_2}{k_2+l_2}\right)$$

where

$$\begin{aligned} P_t \equiv P_t(k_1,k_2,l_1,l_2) &= \binom{N_1}{k_1,l_1} p_{N,2}(t)^{k_1} p_{N,1}(t)^{l_1} (1-p_{N,2}(t)-p_{N,1}(t))^{N_1-k_1-l_1} \\ &\quad \times \binom{N_2}{k_2,l_2} p_{N,2}(t)^{k_2} p_{N,1}(t)^{l_2} (1-p_{N,2}(t)-p_{N,1}(t))^{N_2-k_2-l_2} \end{aligned}$$

Then $E(\hat{\alpha}) = g_1(\alpha) - \frac{f_\rho(\alpha,\beta)}{2}$, $E(\hat{\beta}) = g_2(\beta) + \frac{f_\rho(\alpha,\beta)}{2}$.

Taking derivatives of g_1 and g_2 gives

$$\begin{aligned} g_1'(0) &= \frac{1-r_1}{1-r_2}, \quad g_1'(1) = \frac{1-r_2}{1-r_1} \\ g_2'(0) &= \frac{1-s_1}{1-s_2}, \quad g_2'(1) = \frac{1-s_2}{1-s_1} \end{aligned}$$

We must also find $\frac{\partial f_\rho}{\partial \alpha}$ and $\frac{\partial f_\rho}{\partial \beta}$ for the fixed points $(\alpha,\beta) = (0,0)$, $(1,0)$, and $(1,1)$. Considering f_ρ and the consequences of taking single partial derivatives then evaluating at the fixed points, only one term ($k_1 = 1$, $l_1 = 0$, $k_2 = 0$, $l_2 = 1$) is possibly non-zero:

$$\sigma = N_1 N_2 (1-r_1)(1-s_2)\alpha(1-\beta)[\alpha(r_1-r_2)+r_2)]^{N_1-1}[\beta(s_1-s_2)+s_2)]^{N_2-1}$$

Taking partial derivatives of σ then gives:

$$\begin{aligned} \frac{\partial f_\rho}{\partial \alpha}(0,0) &= N_1 N_2 (1-r_1)(1-s_2) {r_2}^{N_1-1} {s_2}^{N_2-1}, \quad \frac{\partial f_\rho}{\partial \alpha}(0,1) = \frac{\partial f_\rho}{\partial \alpha}(0,1) = 0 \\ \frac{\partial f_\rho}{\partial \beta}(0,0) &= \frac{\partial f_\rho}{\partial \beta}(0,1) = 0, \quad \frac{\partial f_\rho}{\partial \beta}(1,1) = -N_1 N_2 (1-r_1)(1-s_2) {r_1}^{N_1-1} {s_1}^{N_2-1} \end{aligned}$$

The only non-zero terms are $\frac{\partial f_\rho}{\partial \alpha}(0,0)$, $\frac{\partial f_\rho}{\partial \beta}(1,1)$, but since both are exponential in N_1 or N_2, under the large N_1, N_2 assumption they are negligible, and can be disregarded in finding fixed point stabilities.

The Jacobian of the evolution equations evaluated at the fixed points is then

$$D(0,0) = R_N R_V, \quad D(0,1) = \frac{R_N}{R_V}, \quad D(1,1) = \frac{1}{R_N R_V}$$

where $R_N = \frac{1-r_1}{1-r_2}$, $R_V = \frac{1-s_1}{1-s_2}$ can be interpreted as the relative probabilities that form 2 vs form 1 is discarded, for nouns and verbs. $R_N, R_V \in (0,\infty)$, and there are four fixed point regions:

- $R_N < R_V$, $R_N R_V < 1$: $(0,0)$, $(0,1)$ stable.
- $R_N > R_V$, $R_N R_V < 1$: $(0,0)$ stable.
- $R_N < R_V$, $R_N R_V > 1$: $(0,1)$, $(1,1)$ stable.
- $R_N > R_V$, $R_N R_V > 1$: $(1,1)$ stable.

E.8 Section 7.3

α_{t+1} is given by

$$\begin{aligned}
\alpha_{t+1} = E[\hat{\alpha}_t] &= \sum_{\frac{k_1}{N_1}<\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{1}{2}(\frac{k_1}{N_1}+\frac{k_2}{N_2}) \\
&= 2\cdot\frac{1}{2}\sum_{\frac{k_1}{N_1}<\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_2}{N_2} \\
&= \frac{1}{2}\sum_{k_1,k_2} P(k_1,k_2)\frac{k_1}{N_1} + \frac{1}{2}[\sum_{\frac{k_1}{N_1}<\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_2}{N_2}] \\
&= \frac{1}{2}\sum_{k_1,k_2} P(k_1,k_2)\frac{k_1}{N_1} + \frac{1}{2}[\sum_{k_1,k_2} P(k_1,k_2)\frac{k_1}{N_1} - \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_1}{N_1} + \sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)\frac{k_2}{N_2}] \\
&= \alpha_t - \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)(\frac{k_1}{N_1}-\frac{k_2}{N_2}).
\end{aligned}$$

After a similar derivation for β_{t+1} (with appropriate signs reversed) the evolution equations are

$$\alpha_{t+1} = \alpha_t - \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)(\frac{k_1}{N_1}-\frac{k_2}{N_2})$$

$$\beta_{t+1} = \beta_t + \frac{1}{2}\sum_{\frac{k_1}{N_1}>\frac{k_2}{N_2}} P(k_1,k_2)(\frac{k_1}{N_1}-\frac{k_2}{N_2}).$$

E.9 Section 7.4

Consider the evolution equations

$$\alpha_{t+1} = E[\hat{\alpha}_t] = b\alpha_t^{N_1}(1-\beta_t)^{N_2} + \sum_{k_1=1}^{N_1}\sum_{k_2=1}^{N_2} P_{\alpha_t,\beta_t}(k_1,k_2)\frac{bk_1k_2}{D(k_1,k_2)} \tag{89}$$

$$\beta_{t+1} = E[\hat{\beta}_t] = (b+c)\alpha_t^{N_1}(1-\beta_t)^{N_2} + \sum_{k_1=0}^{N_1}\sum_{k_2=1}^{N_2} P_{\alpha_t,\beta_t}(k_1,k_2)\frac{bk_1k_2 + c(N_1-k_1)k_2}{D(k_1,k_2)} \tag{90}$$

Call this map $f(\alpha_t,\beta_t) \equiv (\alpha_{t+1},\beta_{t+1})$. $f(0,0) = (0,0)$, $f(0,1) = (0,1)$, and $f(1,1) = (1,1)$, so $(0,0)$, $(0,1)$, and $(1,1)$ are fixed points of f. By simulation, there are no other stable fixed points as a, b, c, N_1, N_2 are varied, provided $a \neq b \neq c$.

To check their stability, taking the Jacobian D of f gives:

$$D(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & \frac{cN_2}{c-a+aN_2} \end{pmatrix} \tag{91}$$

$$D(1,1) = \begin{pmatrix} \frac{cN_1}{c-b+bN_1} & 0 \\ 0 & 0 \end{pmatrix} \tag{92}$$

$$|D(0,1)| = \left(\frac{bN_1}{b-c+cN_1}\right)\left(\frac{aN_2}{a-c+cN_2}\right) \tag{93}$$

So f projects points near $(0,0)$ onto the β axis and points near $(1,1)$ onto the α axis, $(0,0)$ is stable when $a > c$, and $(1,1)$ is stable when $b > c$. Finally, defining

$$B = \left(\frac{bN_1}{b-c+cN_1}\right), \quad A = \left(\frac{aN_2}{a-c+cN_2}\right), \tag{94}$$

so $|D(0,1)| = AB$, gives the 6 solution regions in §7.4

E.10 Section 7.4.1

We examine the fixed points of

$$\alpha_{t+1} = E[\hat{\alpha}_t] = b{\alpha'_t}^{N_1}(1-\beta'_t)^{N_2} + \sum_{k_1=1}^{N_1}\sum_{k_2=1}^{N_2} P_t(k_1,k_2)\frac{bk_1k_2}{D(k_1,k_2)} \tag{95}$$

$$\beta_{t+1} = E[\hat{\beta}_t] = (b+c){\alpha'_t}^{N_1}(1-\beta'_t)^{N_2} + \sum_{k_1=0}^{N_1}\sum_{k_2=1}^{N_2} P_t(k_1,k_2)\frac{bk_1k_2 + c(N_1-k_1)k_2}{D(k_1,k_2)} \tag{96}$$

where

$$P_t(k_1,k_2) = \binom{N_1}{k_1}\binom{N_2}{k_2}{\alpha'_t}^{k_1}(1-\alpha'_t)^{N_1-k_1}{\beta'_t}^{k_2}(1-\beta'_t)^{N_2-k_2}$$

Call these evolution equations g. g is a function of p and q, $g(p,q)$ and in particular $f = g(0,0)$, where f is the no-mistransmission case considered above. When p or $q \neq 0$, $(0,1)$ is still a fixed point, and $(0,0)$ and $(1,1)$ give solution branches $x_1(q) = (0,\lambda(q))$ and $x_2(p) = (\kappa(p),1)$ (where $\lambda(0) = 0$, $\kappa(0) = 1$). Intuitively, this is because mistransmission only occurs along one axis near $(0,0)$ or $(1,1)$.[39]

By graphing the map $g(0,\beta_t)$ vs β_t as q is perturbed from 0 and noting that $g(0,1) = (0,1)$, it is clear that the fixed point $x_1(q)$ exists if $\frac{\partial g(0,\beta_t)}{\partial \beta_t}|_{(0,1)} > 1$ and is stable, while taking derivatives of the formula

$$g(0,\beta_t) = \sum_{k_2=1}^{N_2} P_{0,\beta_t}(0,k_2)\frac{ck_2}{a(N_2-k_2)+cN_2}$$

shows $x_1(q)$ is unique (when it exists). Similarly, $x_2(p)$ exists if $\frac{\partial g(1,\alpha)}{\partial \alpha}|_{(1,1)} > 1$, and is unique and stable if it exists.

[39]Formally, the center manifold at $(0,0)$ is the β-axis and the center manifold at $(1,1)$ is the α-axis; since the axes are invariant subspaces, the fixed points stay on them as p, q perturbed from 0.

The existence conditions work out to

$$x_1(q) \text{ exists if } a \geq \frac{c}{1 - q\frac{N_2}{N_2 - 1}}, \quad x_2(p) \text{ exists if } b \geq \frac{c}{1 - p\frac{N_1}{N_1 - 1}}$$

A similar derivation as for (93) gives that $(0, 1)$ is stable for

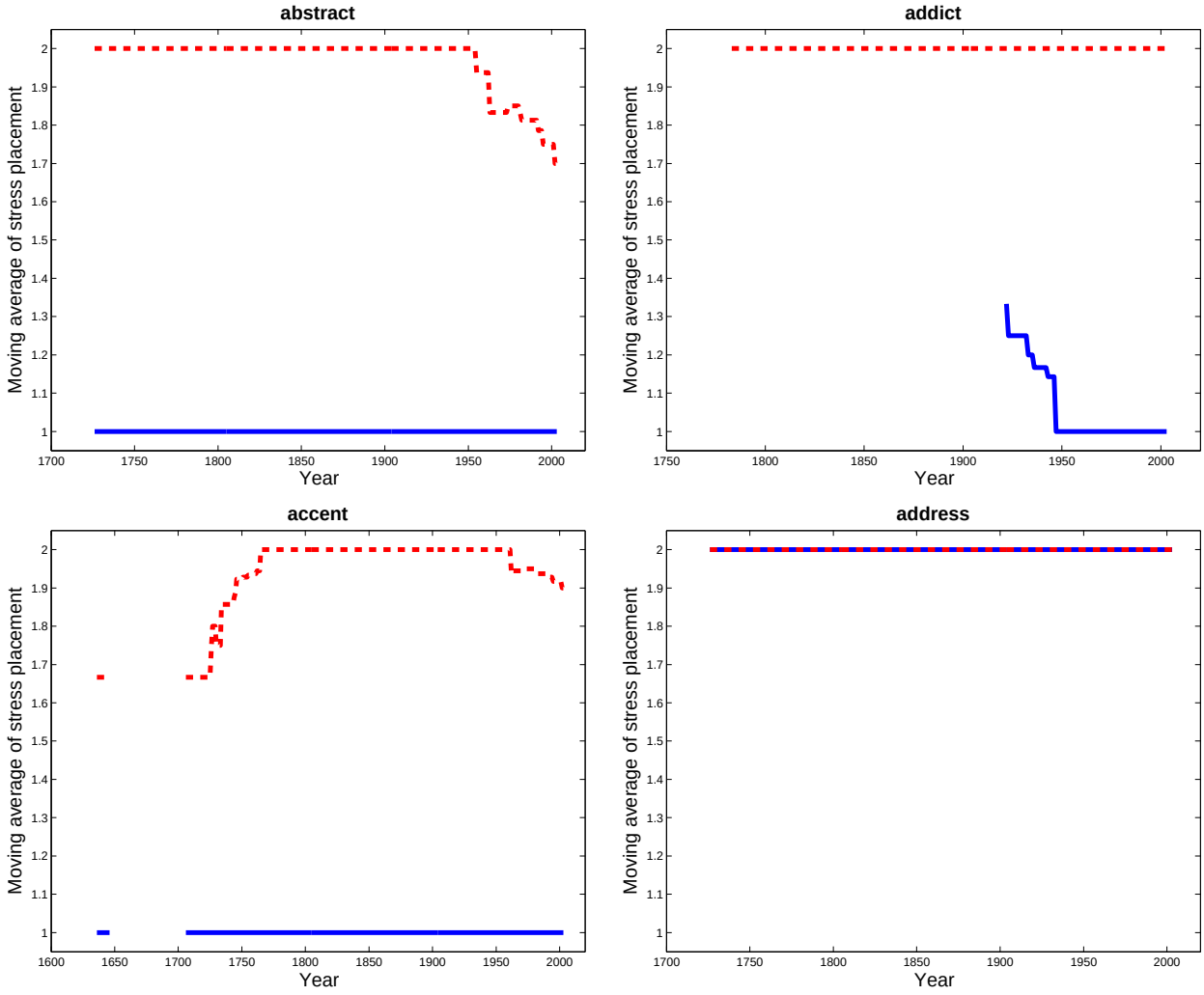
$$\frac{abN_1N_2}{(b - c + cN_1)(a - c + cN_2)} < \frac{1}{(1 - p)(1 - q)}$$

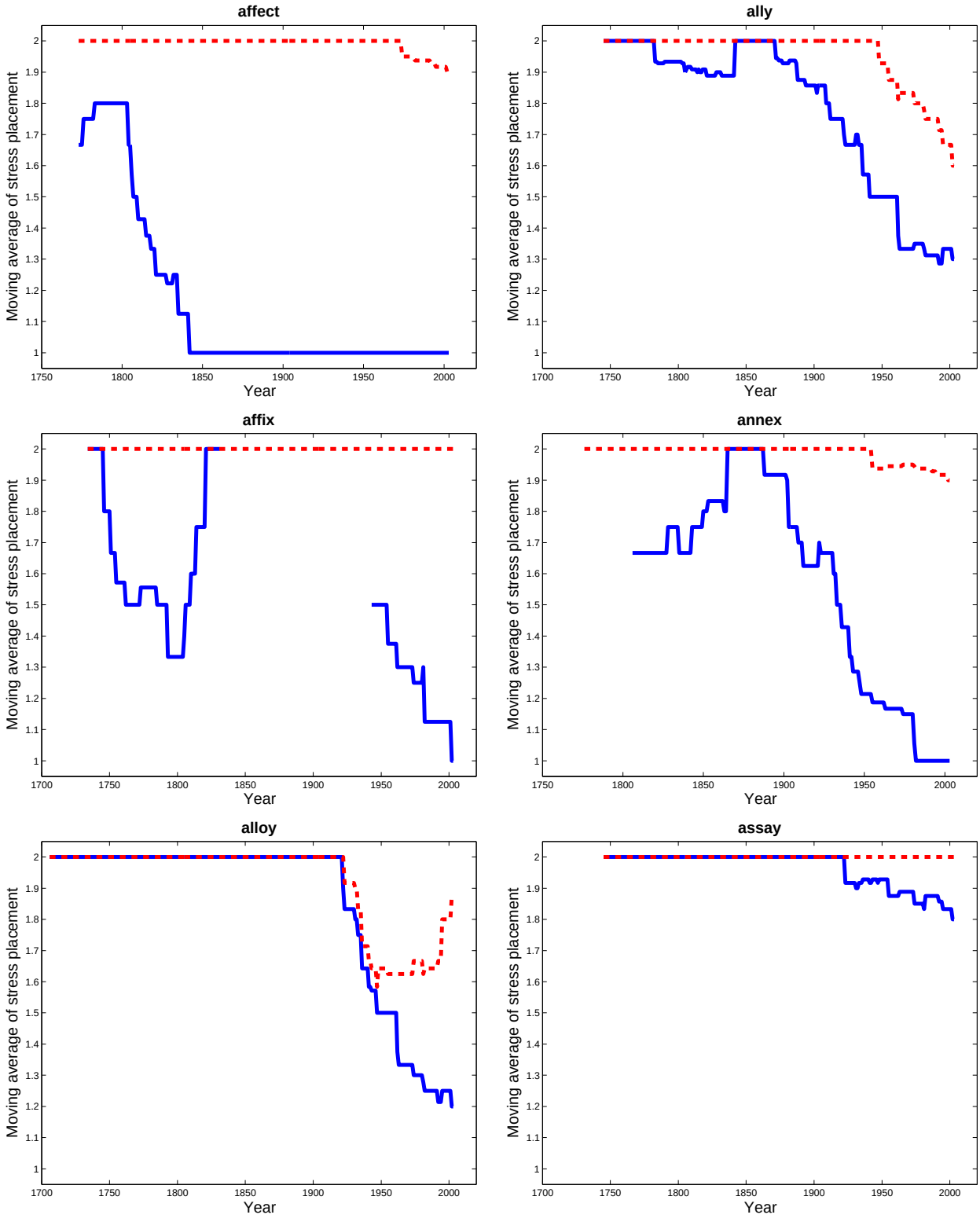
which is $AB < \frac{1}{(1-p)(1-q)}$, using (94).

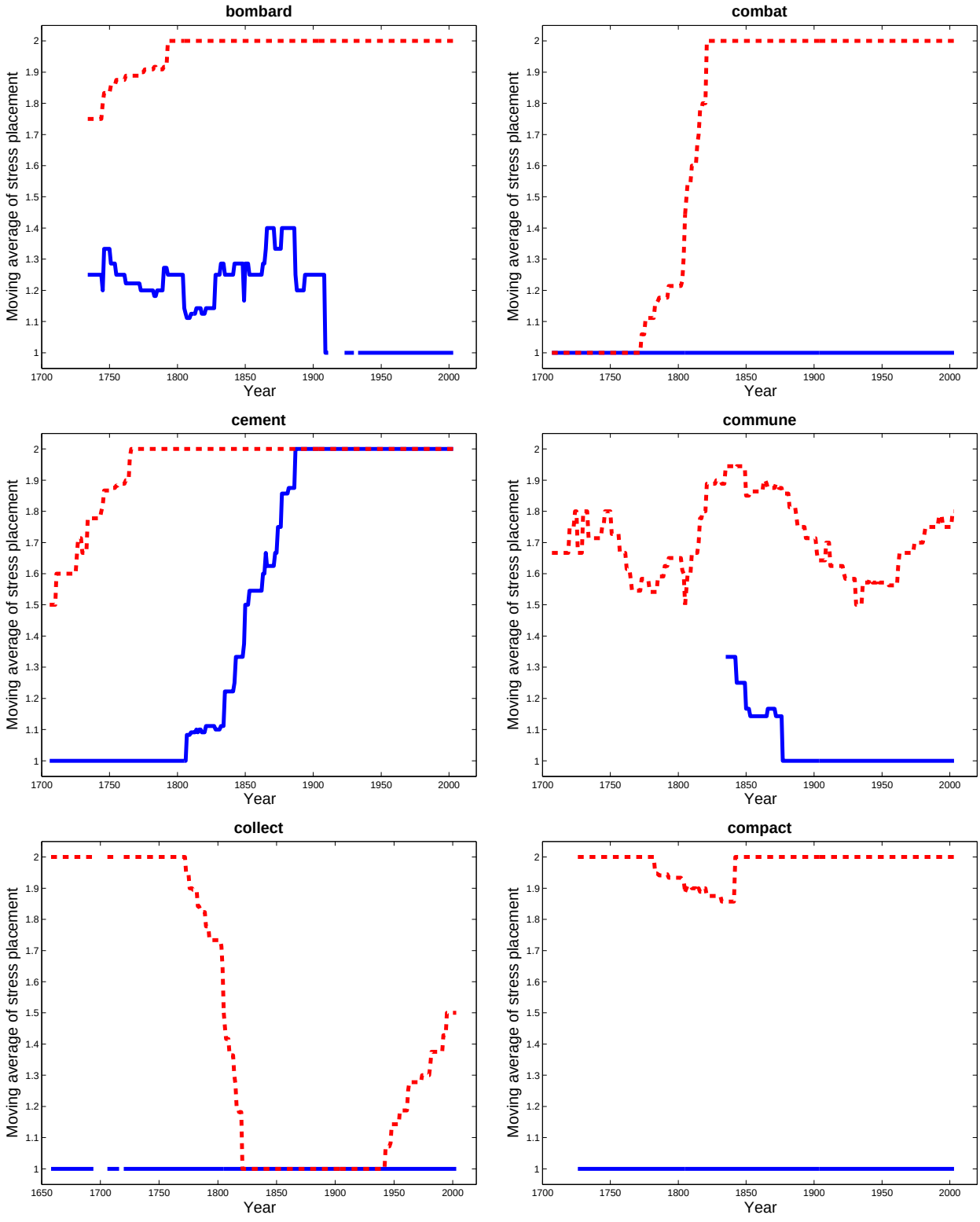
The upshot is that there are still 6 regions as in the no-mistransmission model.

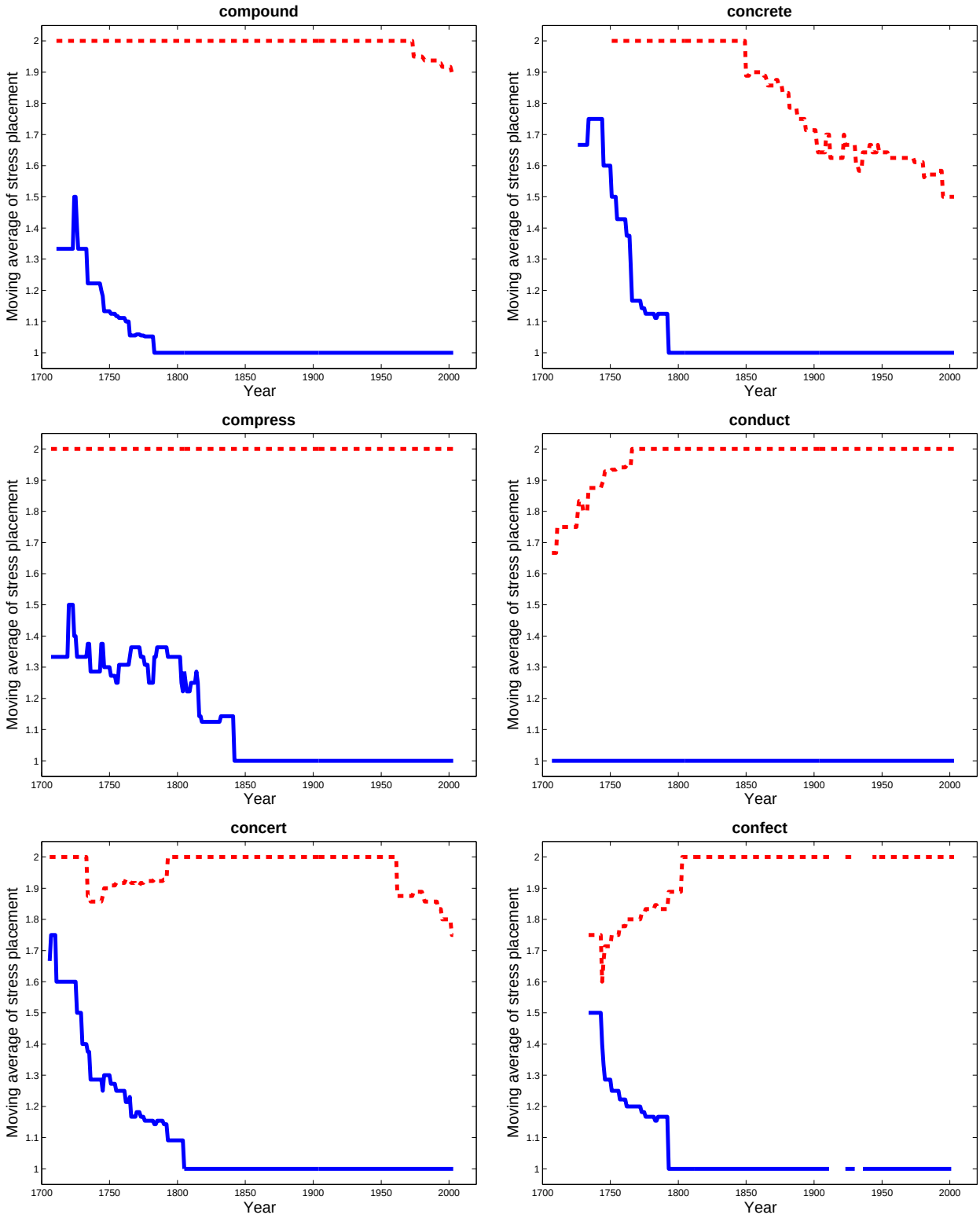
F Trajectories

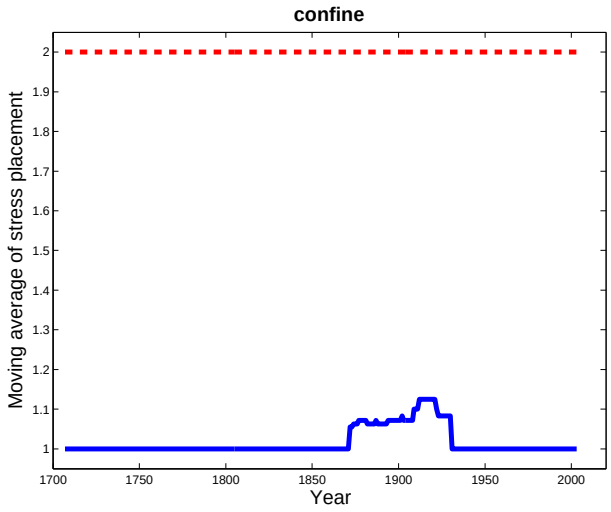
Noun trajectories are blue, verb trajectories are red. A point was included at time t for the N form of a N/V pair if 3 or more British dictionaries in the window $(t-30, t+30)$ listed it (and similarly for the V form).

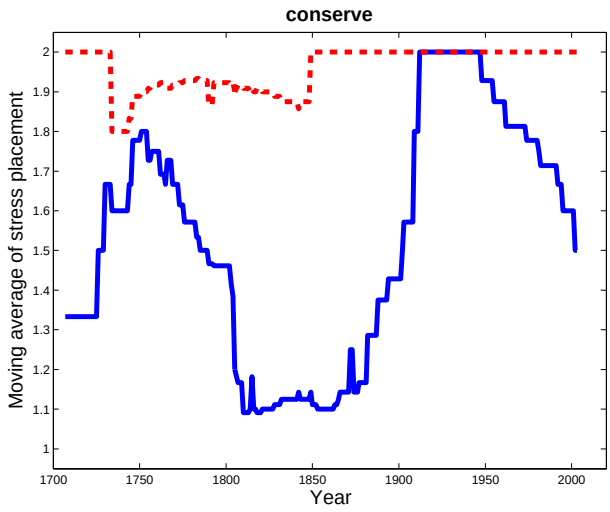


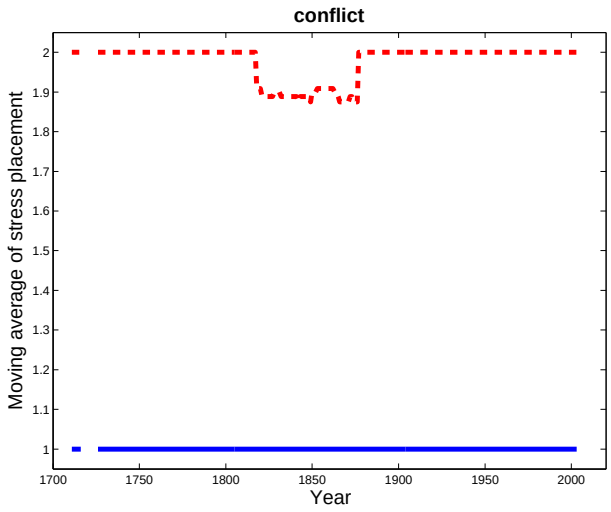

affect
ally
affix
annex
alloy
assay
Moving average of stress placement
Year

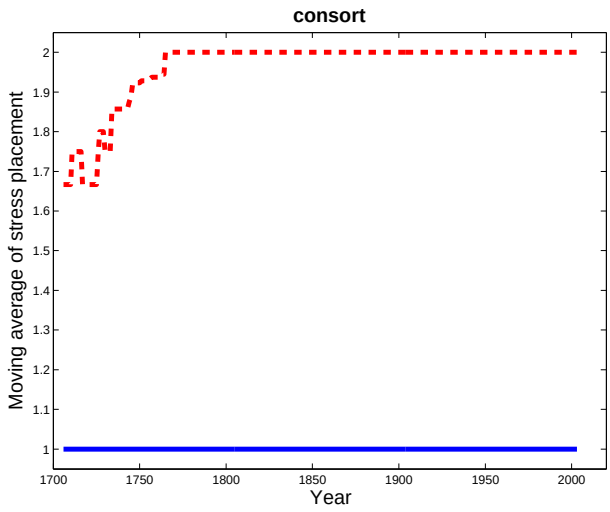

bombard
combat
cement
commune
collect
compact
Moving average of stress placement
Year

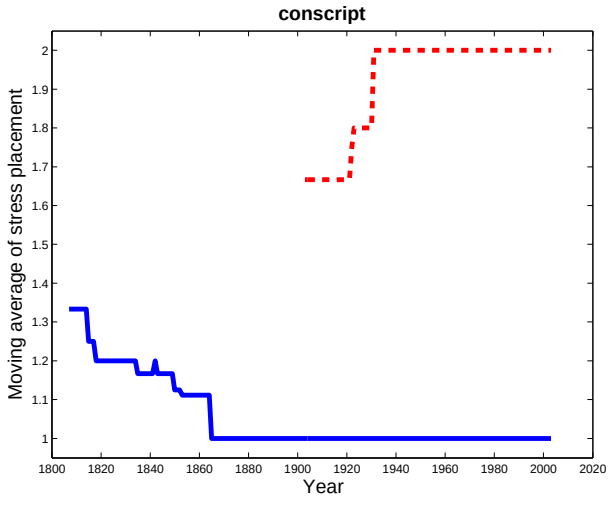

compound
concrete
compress
conduct
concert
confect
Moving average of stress placement
Year
1700
1750
1800
1850
1900
1950
2000
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2

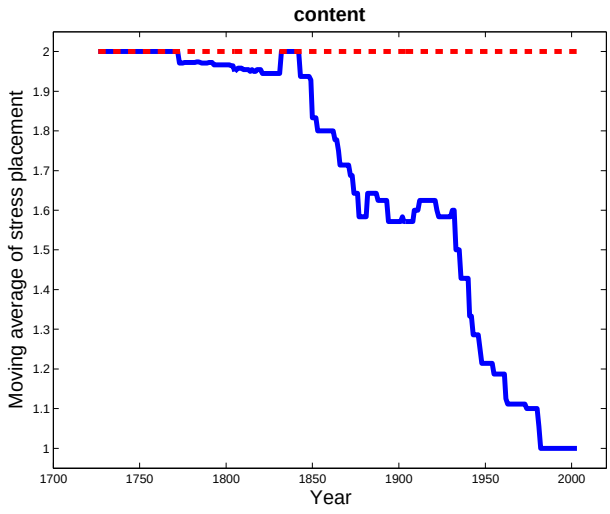

confine
Moving average of stress placement
Year

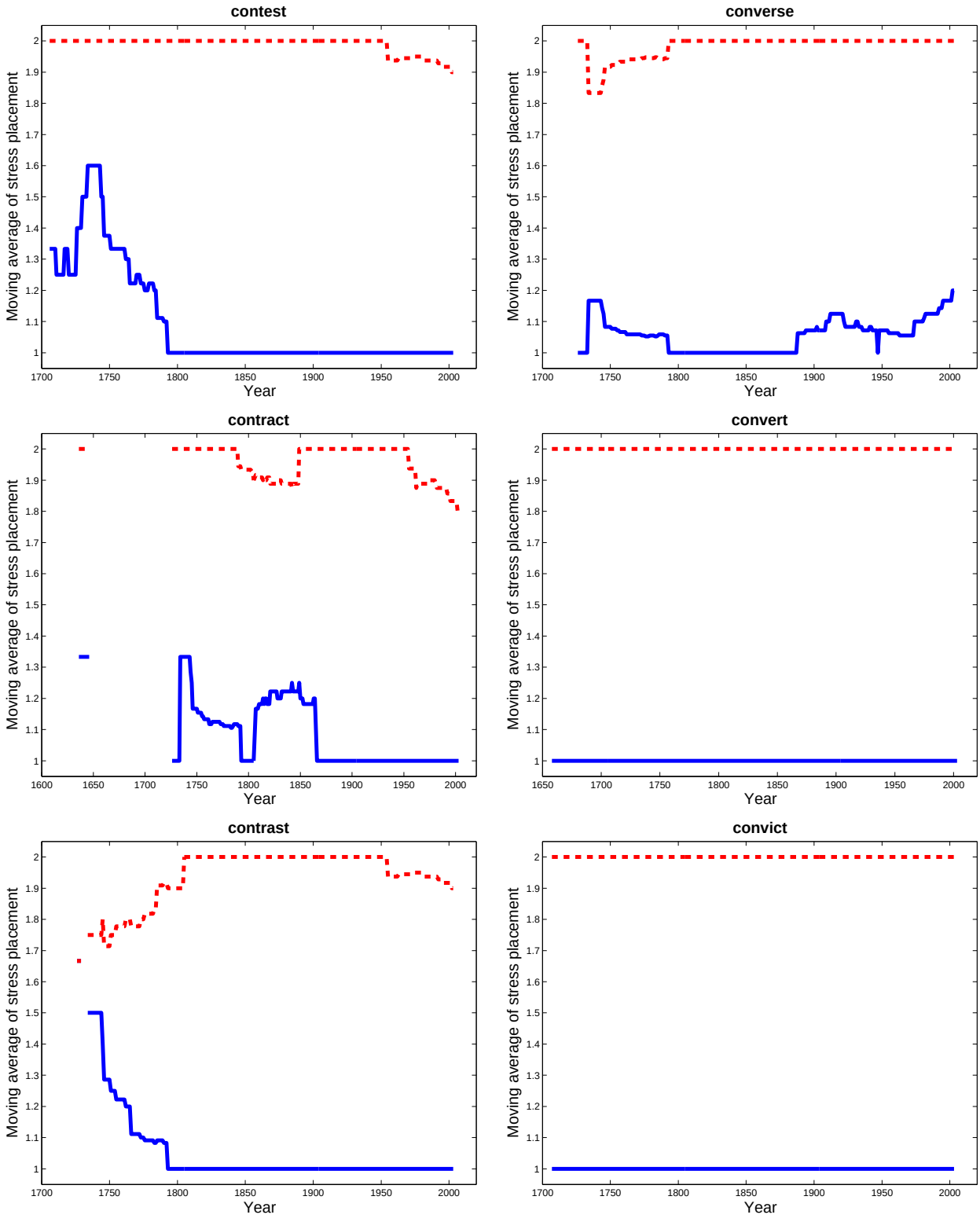

conserve
Moving average of stress placement
Year

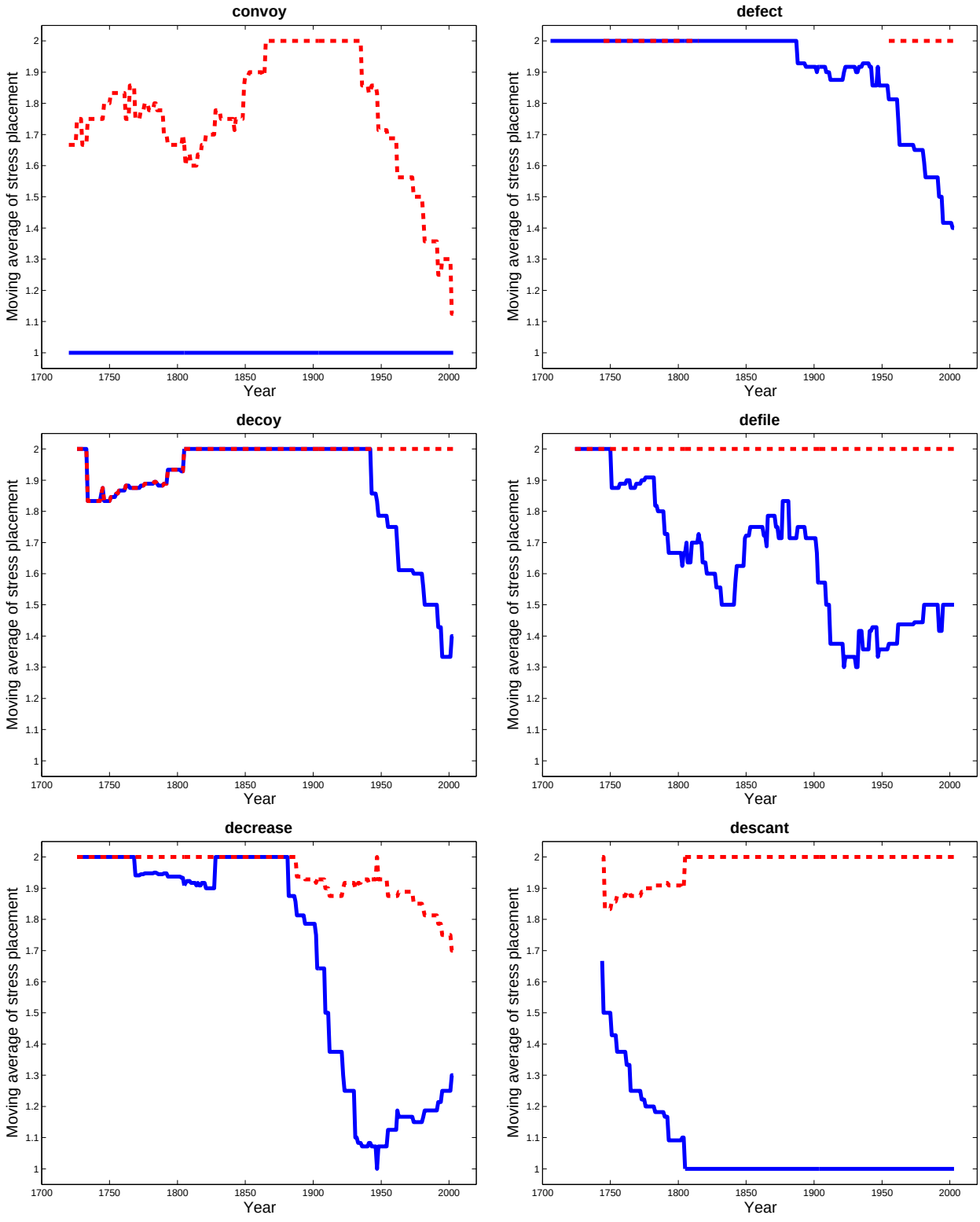

conflict
Moving average of stress placement
Year

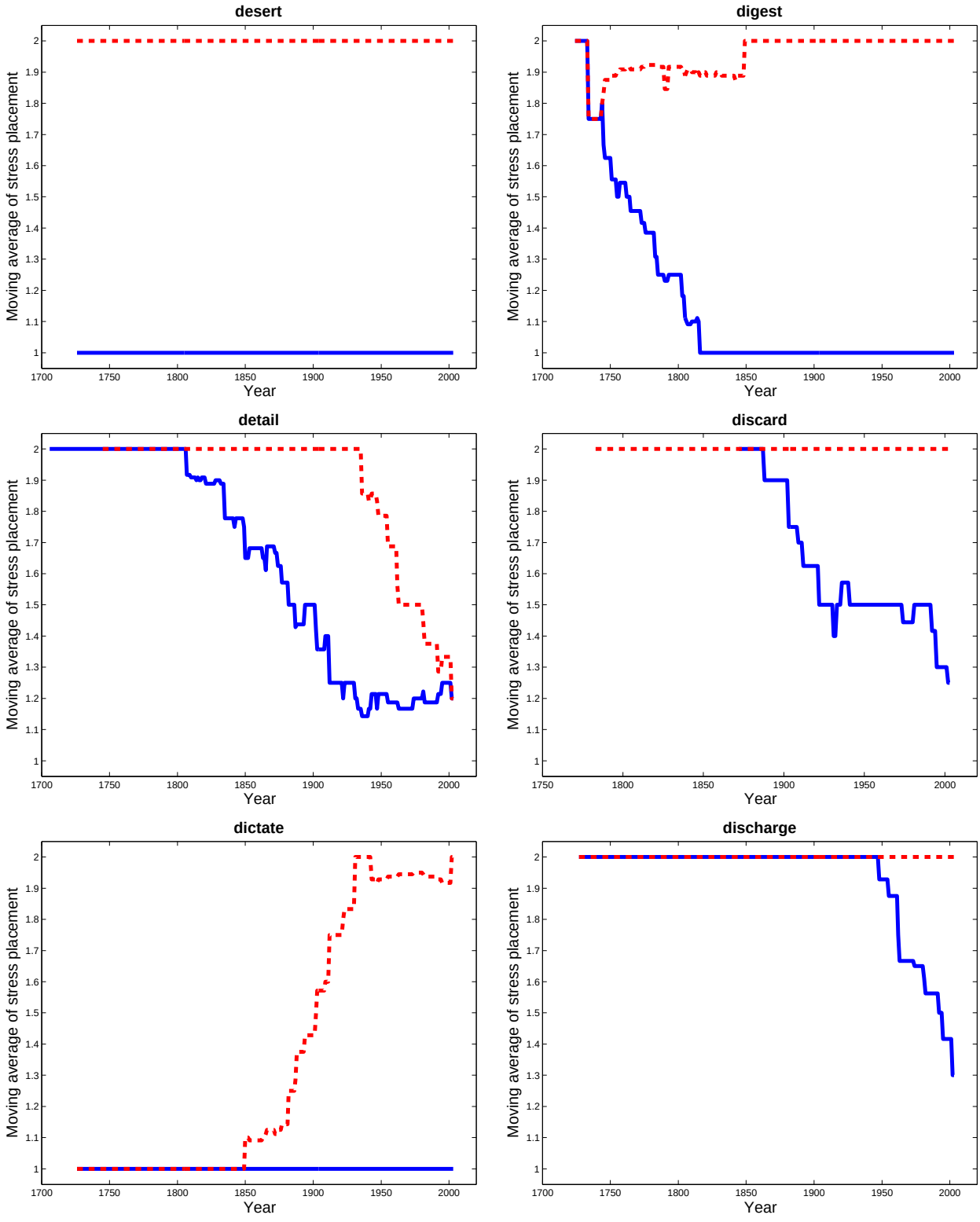

consort
Moving average of stress placement
Year

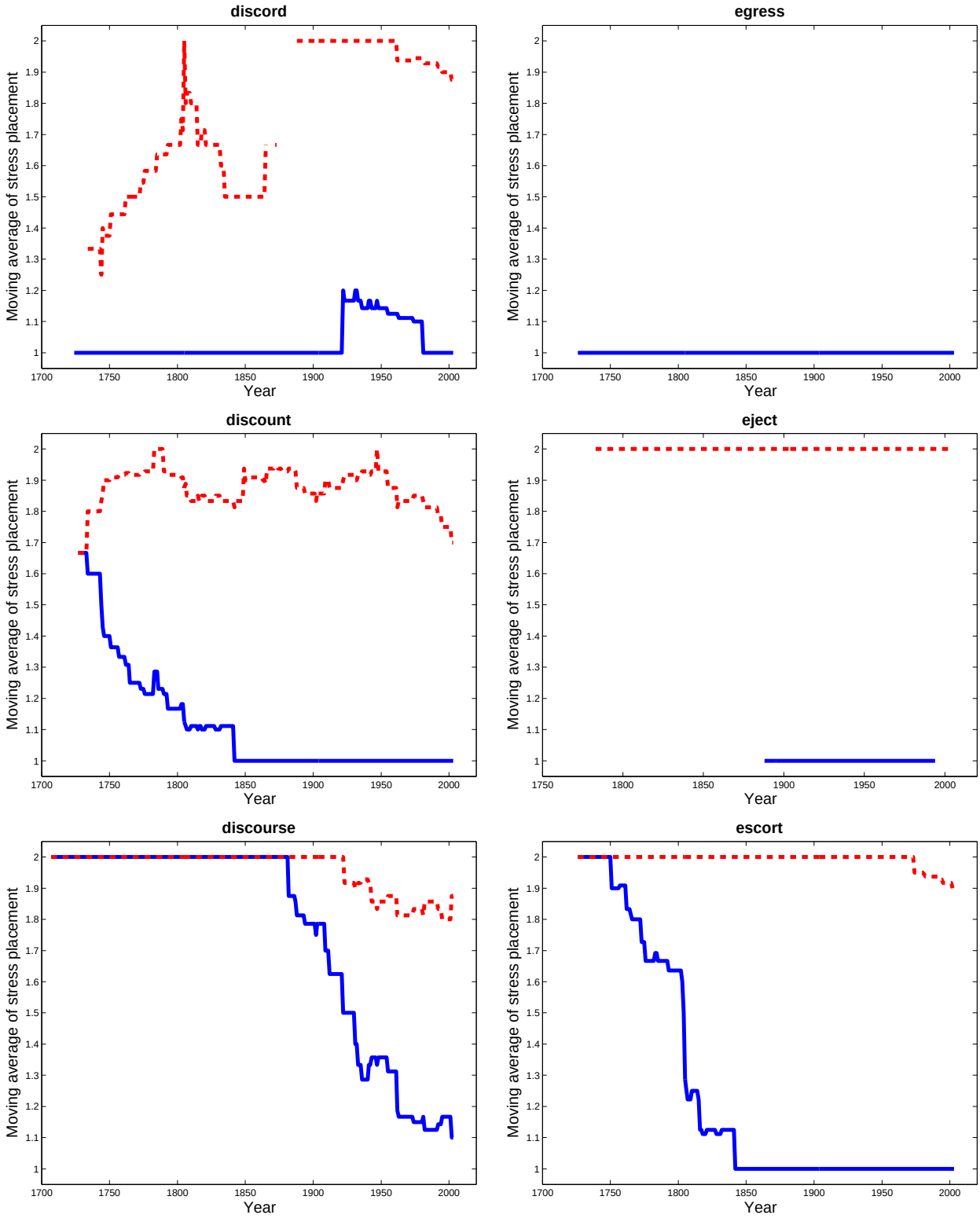

conscript
Moving average of stress placement
Year

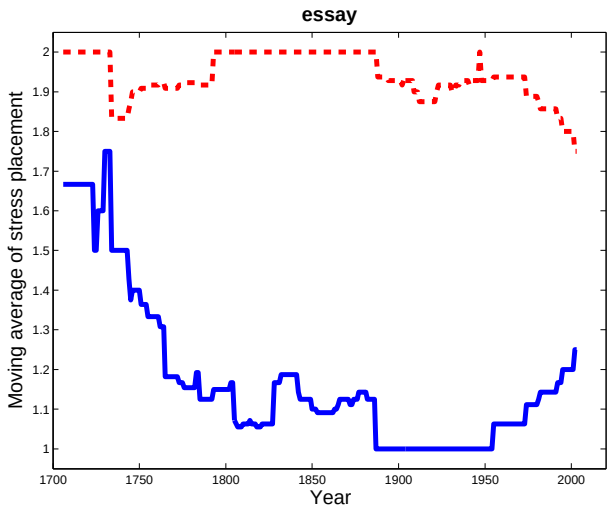

content
Moving average of stress placement
Year

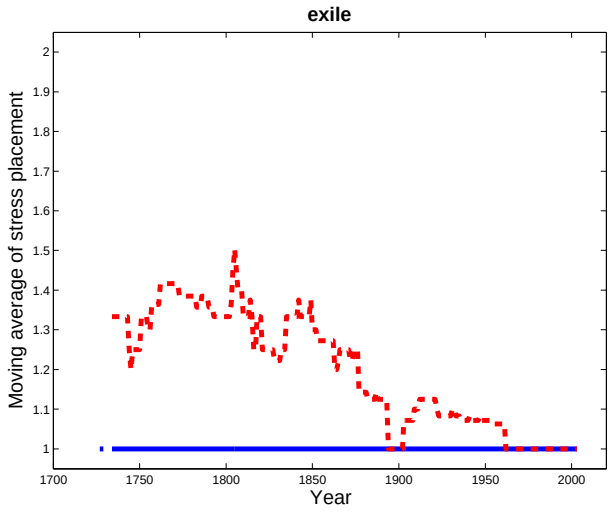

contest
converse
contract
convert
contrast
convict
Moving average of stress placement
Year

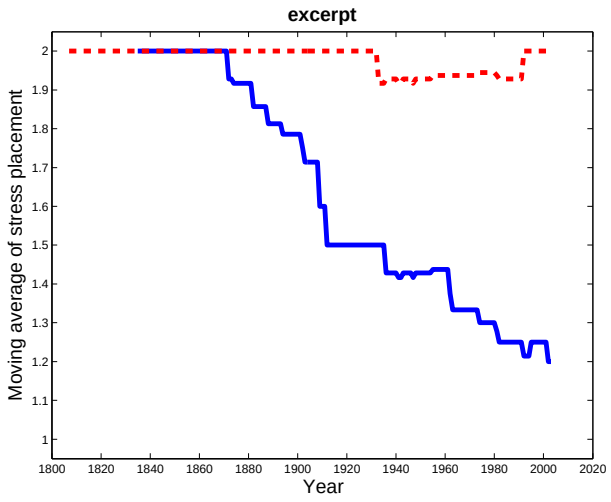

convoy
defect
decoy
defile
decrease
descant
Moving average of stress placement
Year
1700
1750
1800
1850
1900
1950
2000
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2

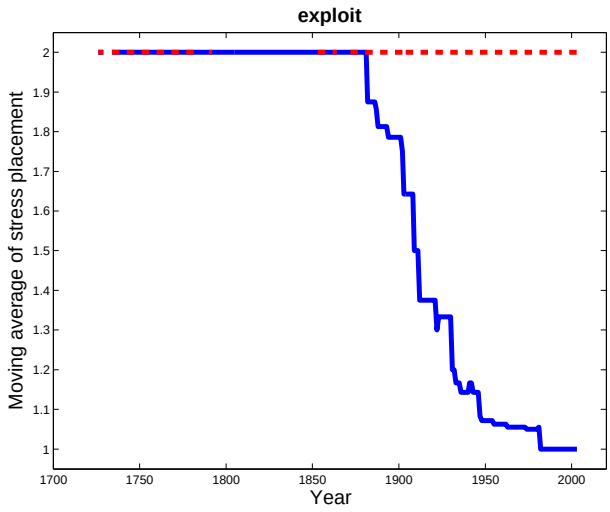

desert
digest
detail
discard
dictate
discharge
Moving average of stress placement
Year
1700
1750
1800
1850
1900
1950
2000
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2

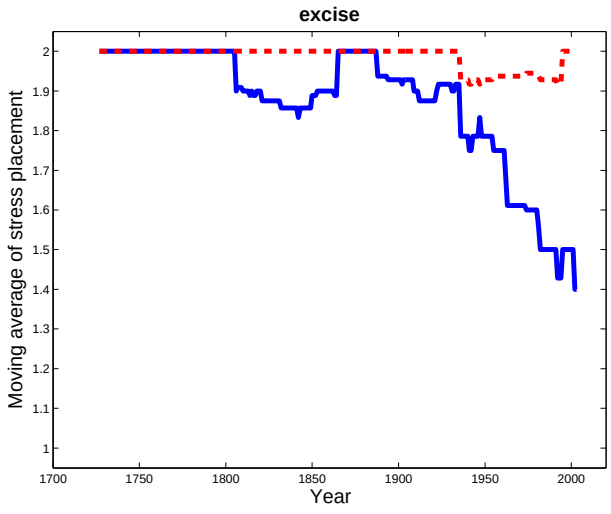

discord
egress
discount
eject
discourse
escort
Moving average of stress placement
Year

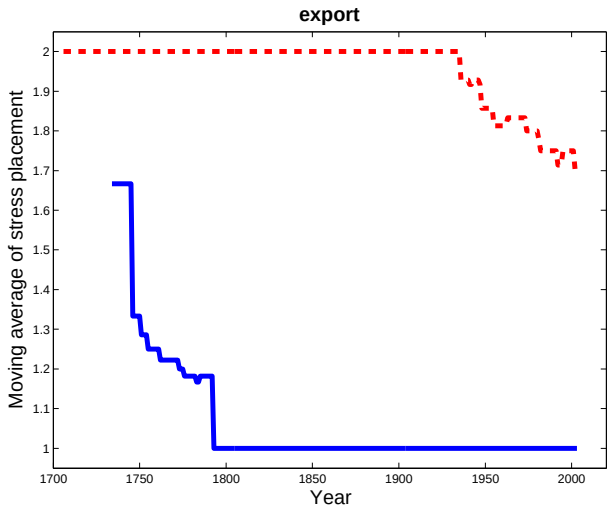

essay
Moving average of stress placement
Year

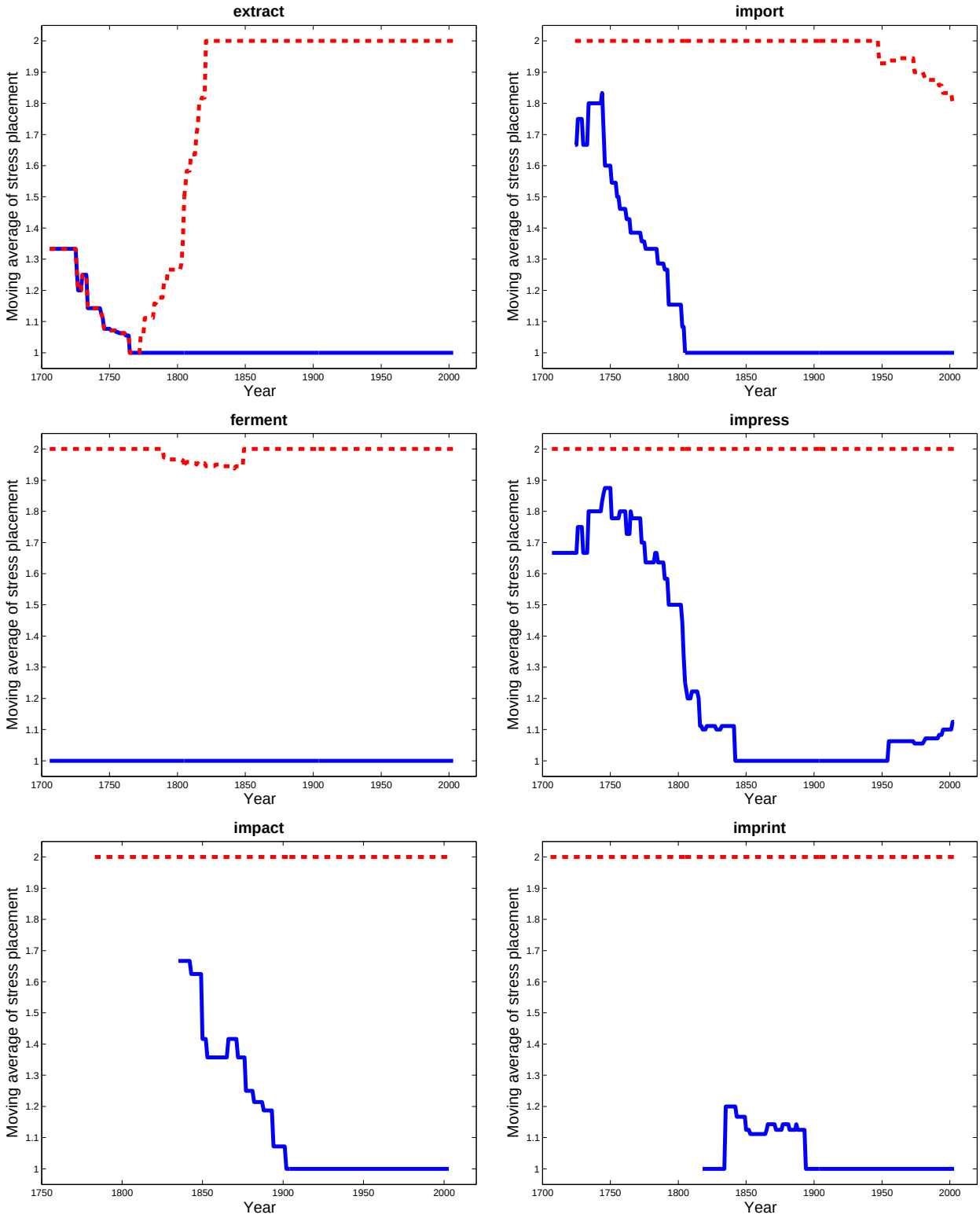

exile
Moving average of stress placement
Year

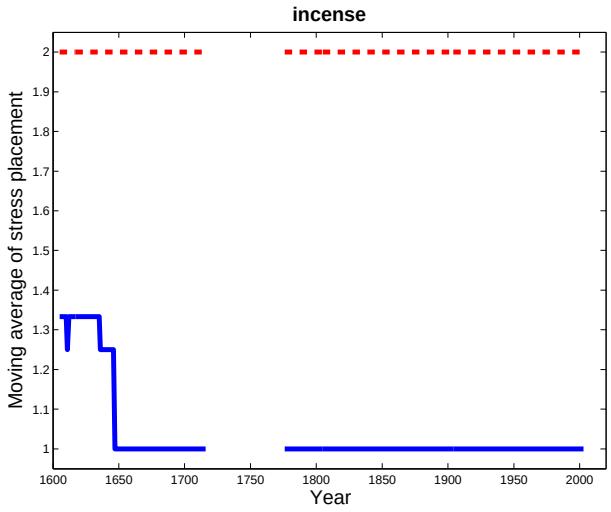

excerpt
Moving average of stress placement
Year

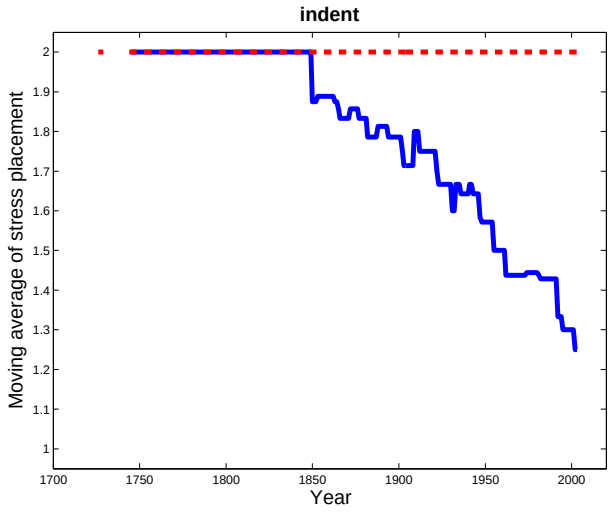

exploit
Moving average of stress placement
Year

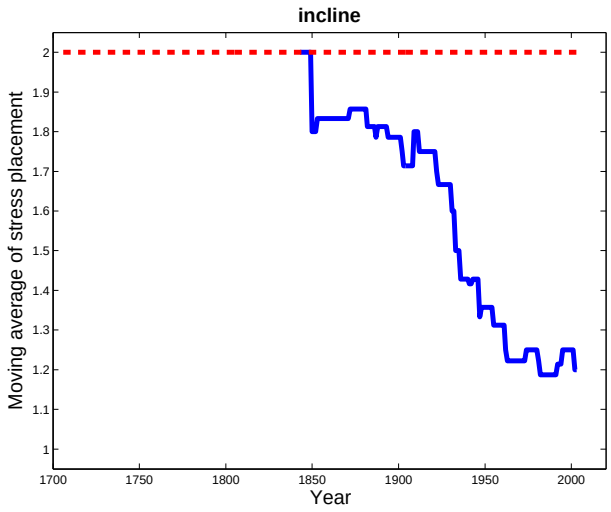

excise
Moving average of stress placement
Year

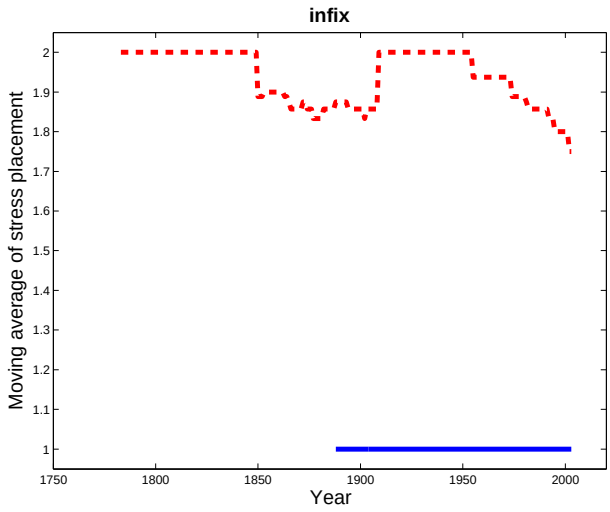

export
Moving average of stress placement
Year

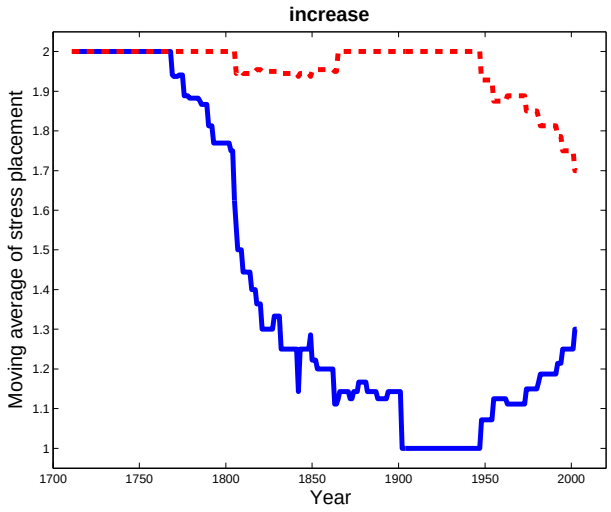

extract
import
ferment
impress
impact
imprint
Moving average of stress placement
Year

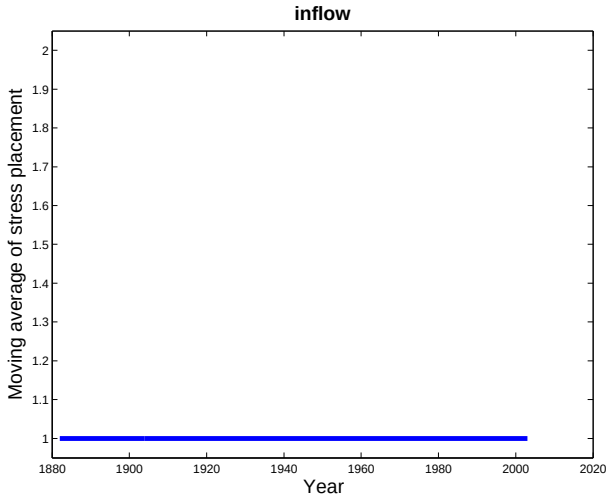

incense
Moving average of stress placement
Year

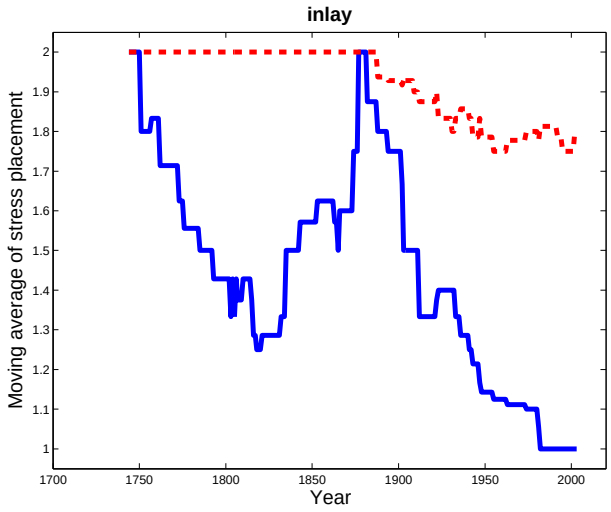

indent
Moving average of stress placement
Year

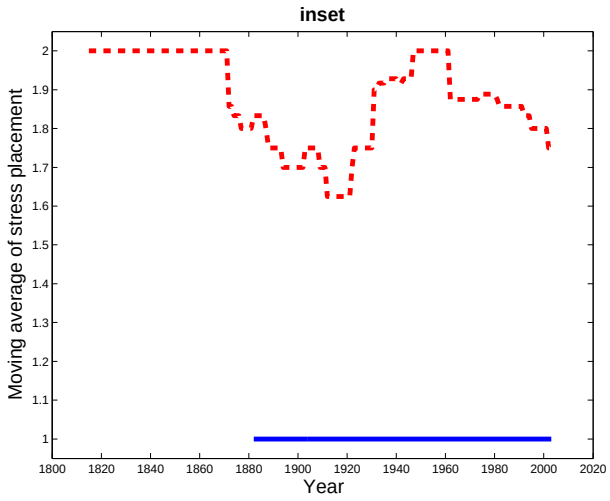

incline
Moving average of stress placement
Year

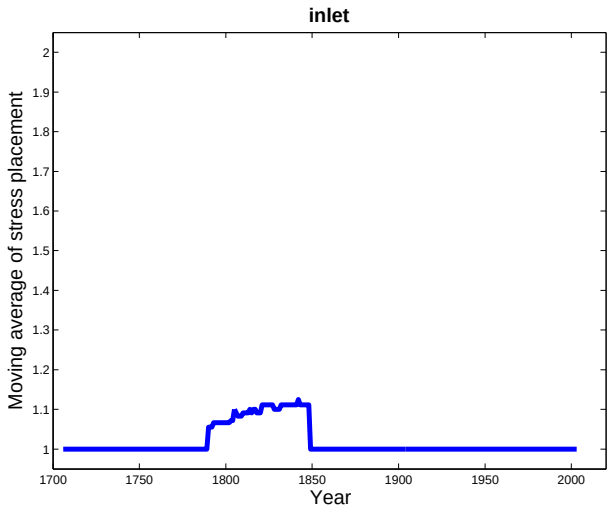

infix
Moving average of stress placement
Year

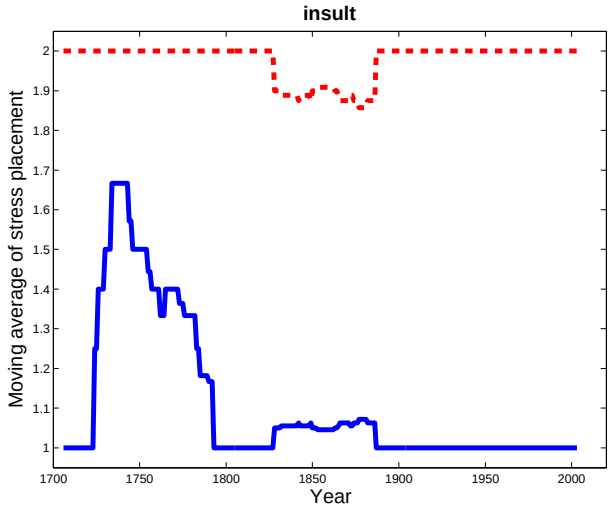

increase
Moving average of stress placement
Year

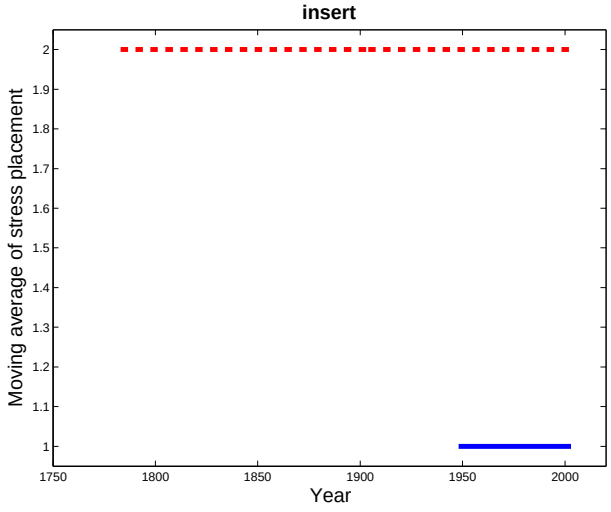

inflow
Moving average of stress placement
Year

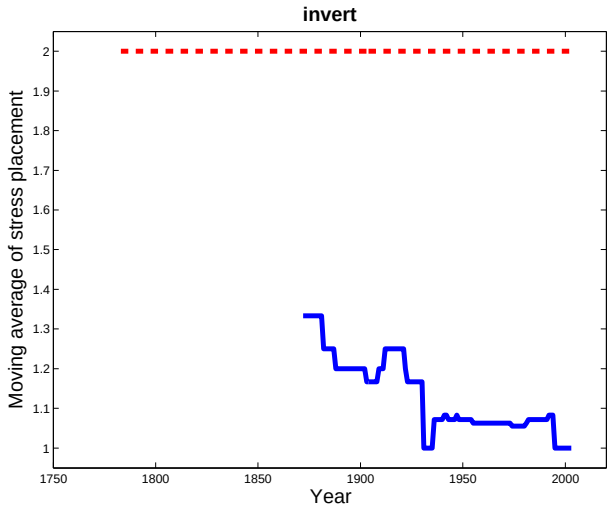

inlay
Moving average of stress placement
Year

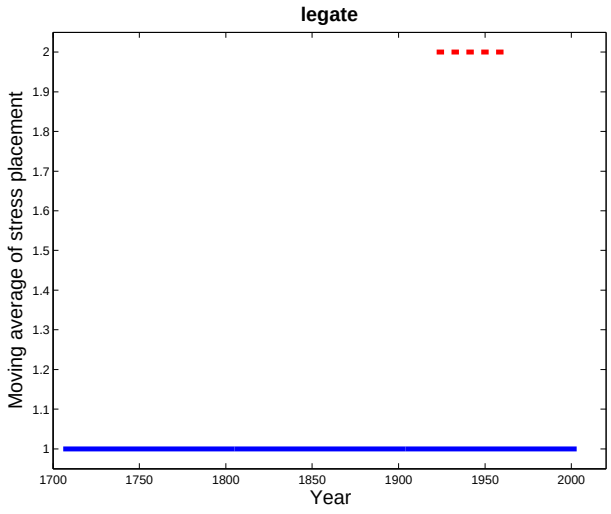

inset
Moving average of stress placement
Year

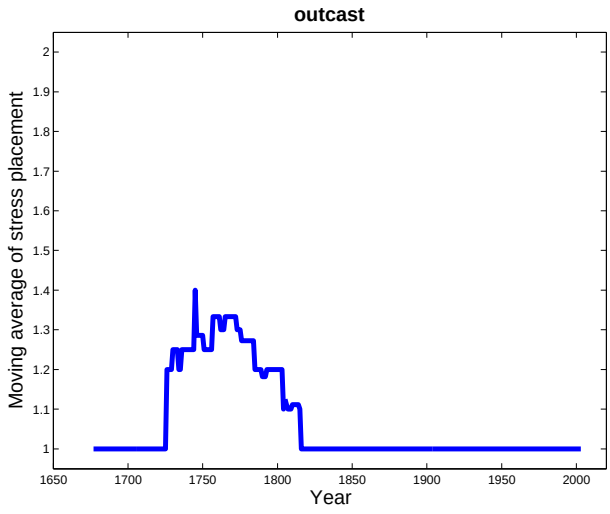

inlet
Moving average of stress placement
Year

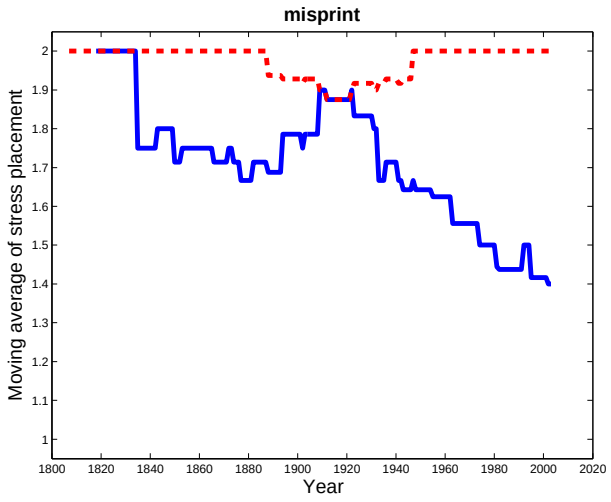

insult
Moving average of stress placement
Year

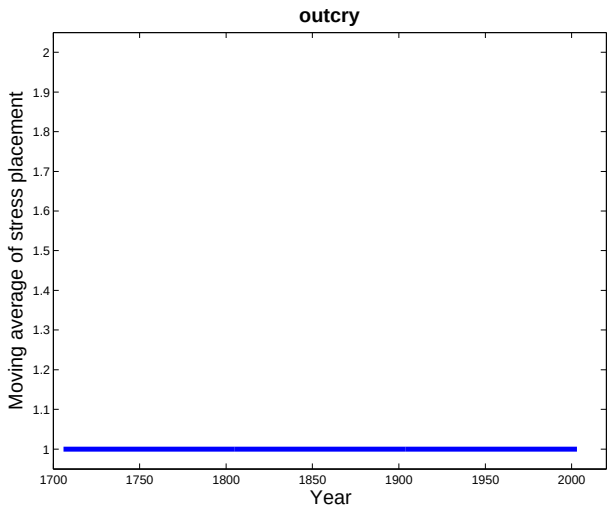

insert
Moving average of stress placement
Year

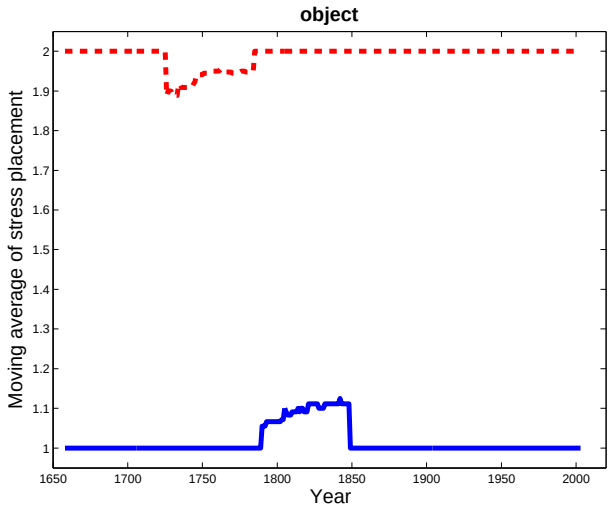

invert
Moving average of stress placement
Year

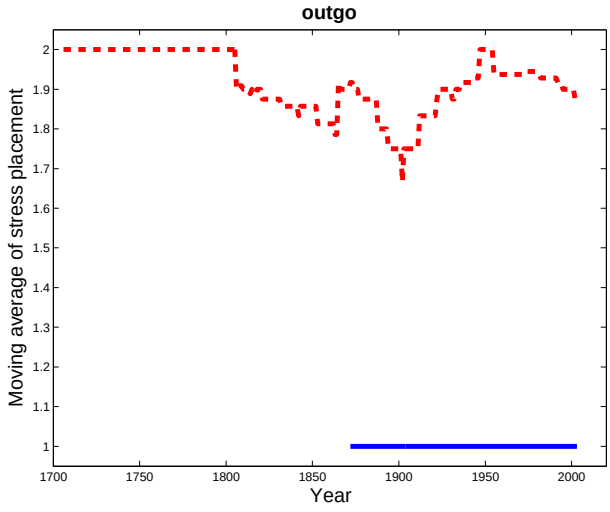

legate
Moving average of stress placement
Year

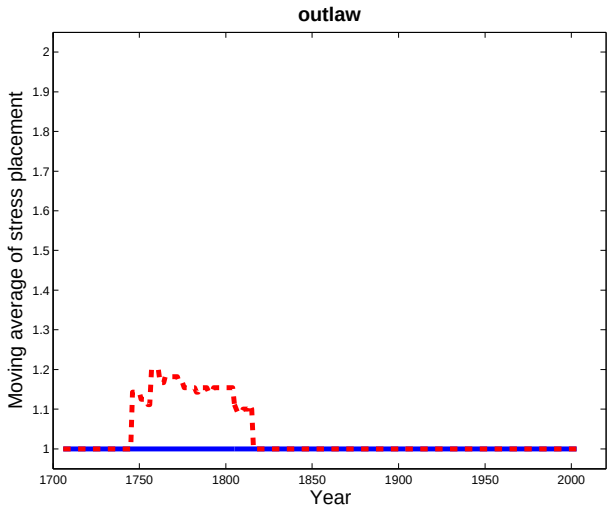

outcast
Moving average of stress placement
Year

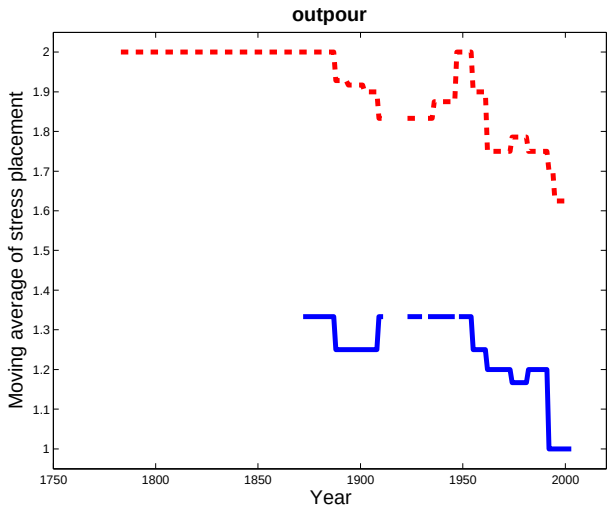

misprint
Moving average of stress placement
Year

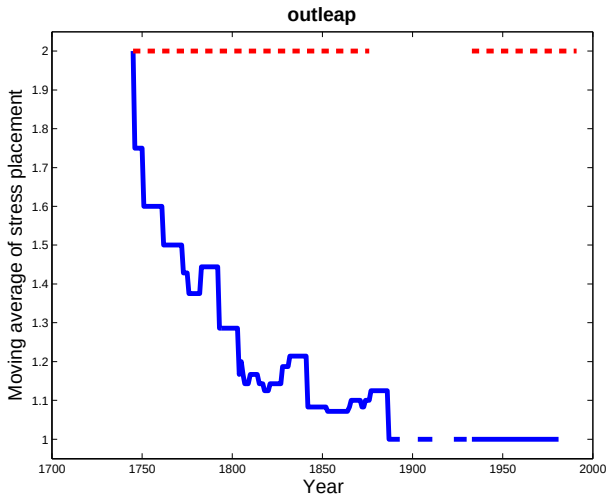

outcry
Moving average of stress placement
Year

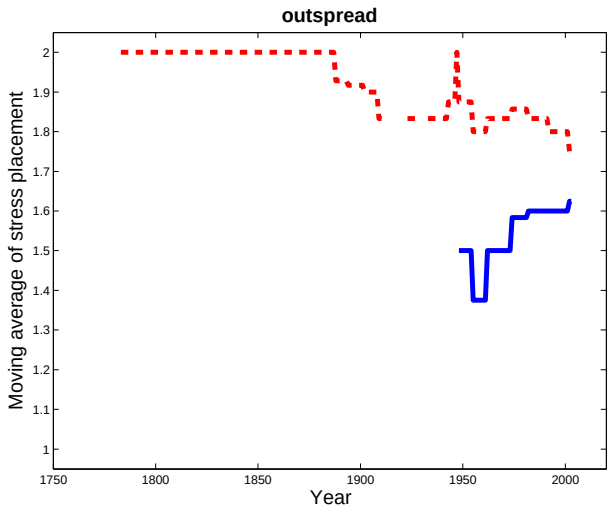

object
Moving average of stress placement
Year

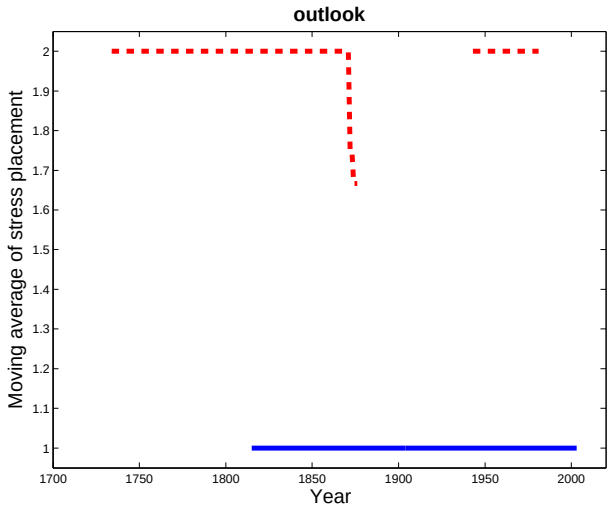

outgo
Moving average of stress placement
Year

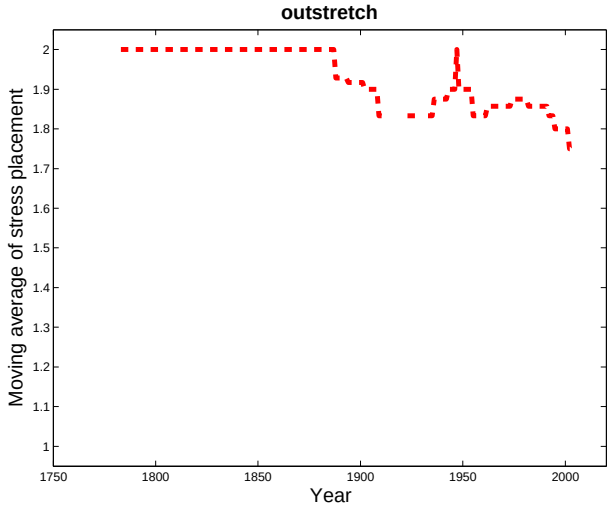

outlaw
Moving average of stress placement
Year

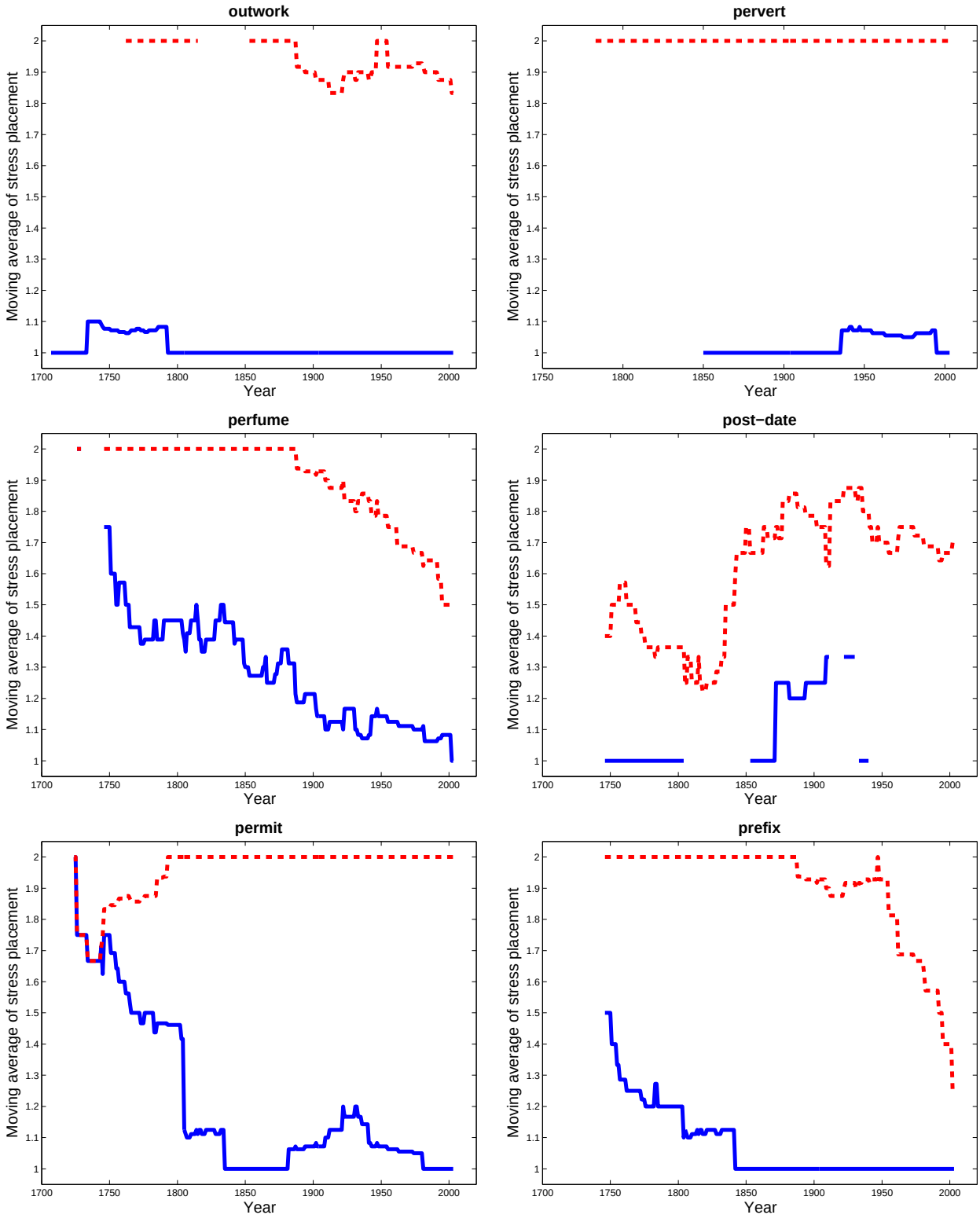

outpour
Moving average of stress placement
Year

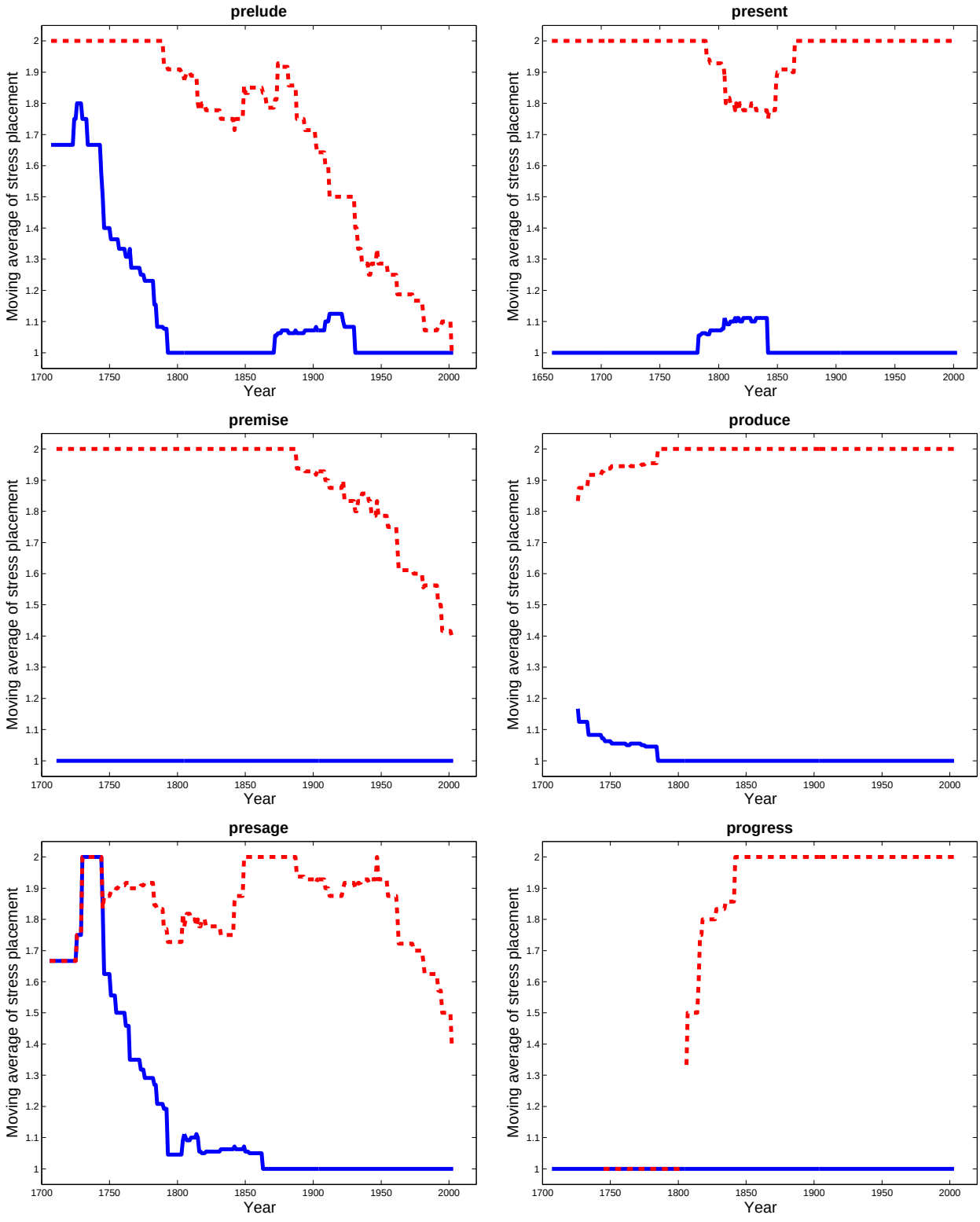

outleap
Moving average of stress placement
Year

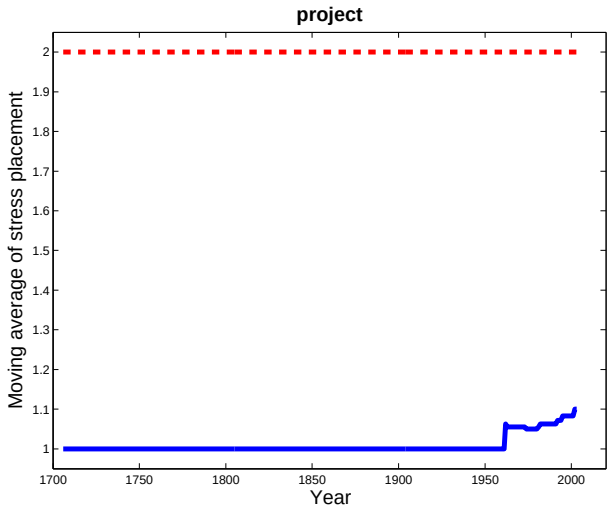

outspread
Moving average of stress placement
Year

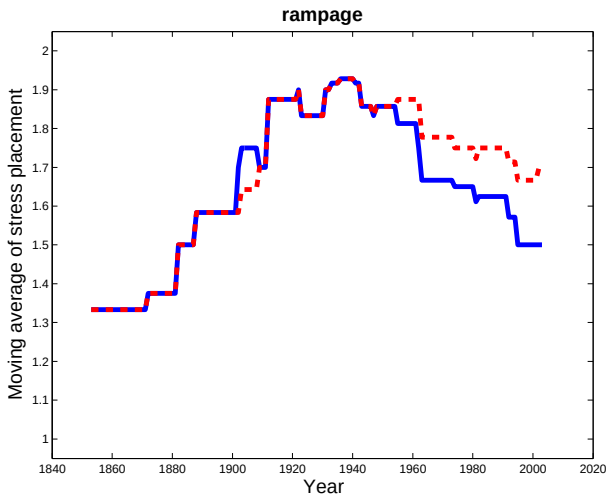

outlook
Moving average of stress placement
Year

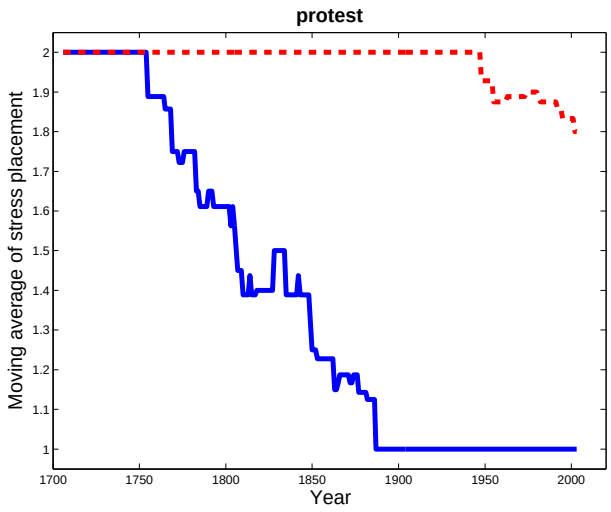

outstretch
Moving average of stress placement
Year

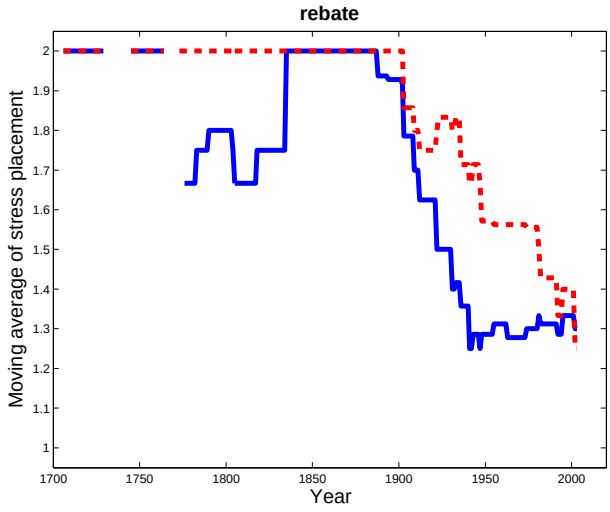

outwork
pervert
perfume
post−date
permit
prefix
Moving average of stress placement
Year

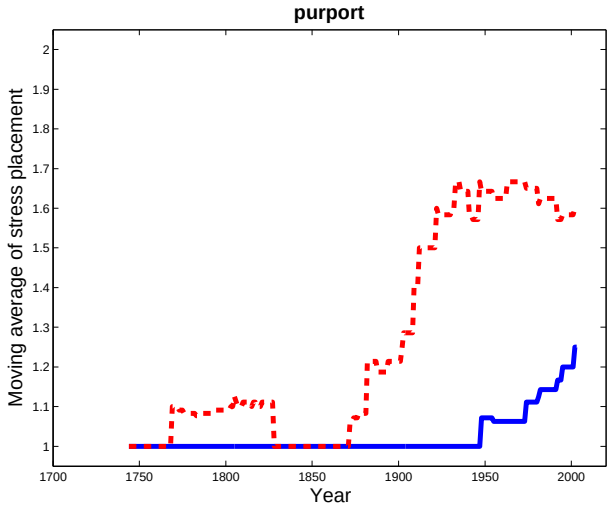

prelude
present
premise
produce
presage
progress
Moving average of stress placement
Year

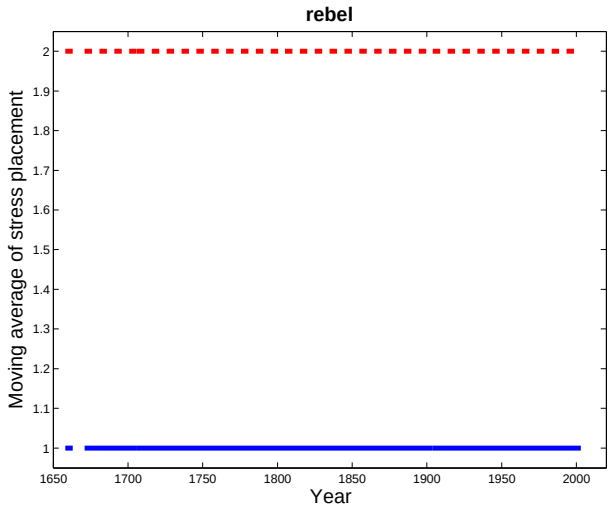

project
Moving average of stress placement
Year

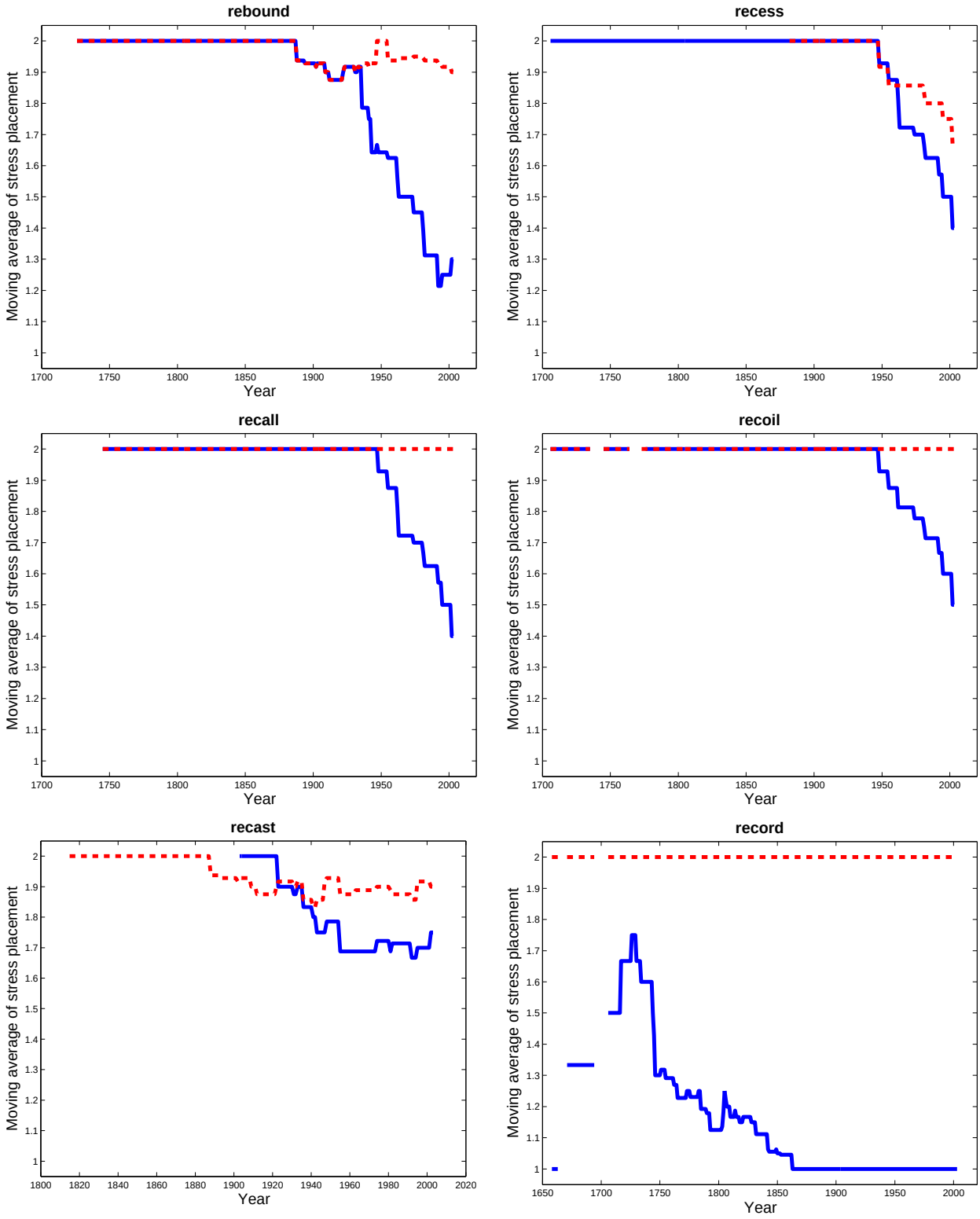

rampage
Moving average of stress placement
Year

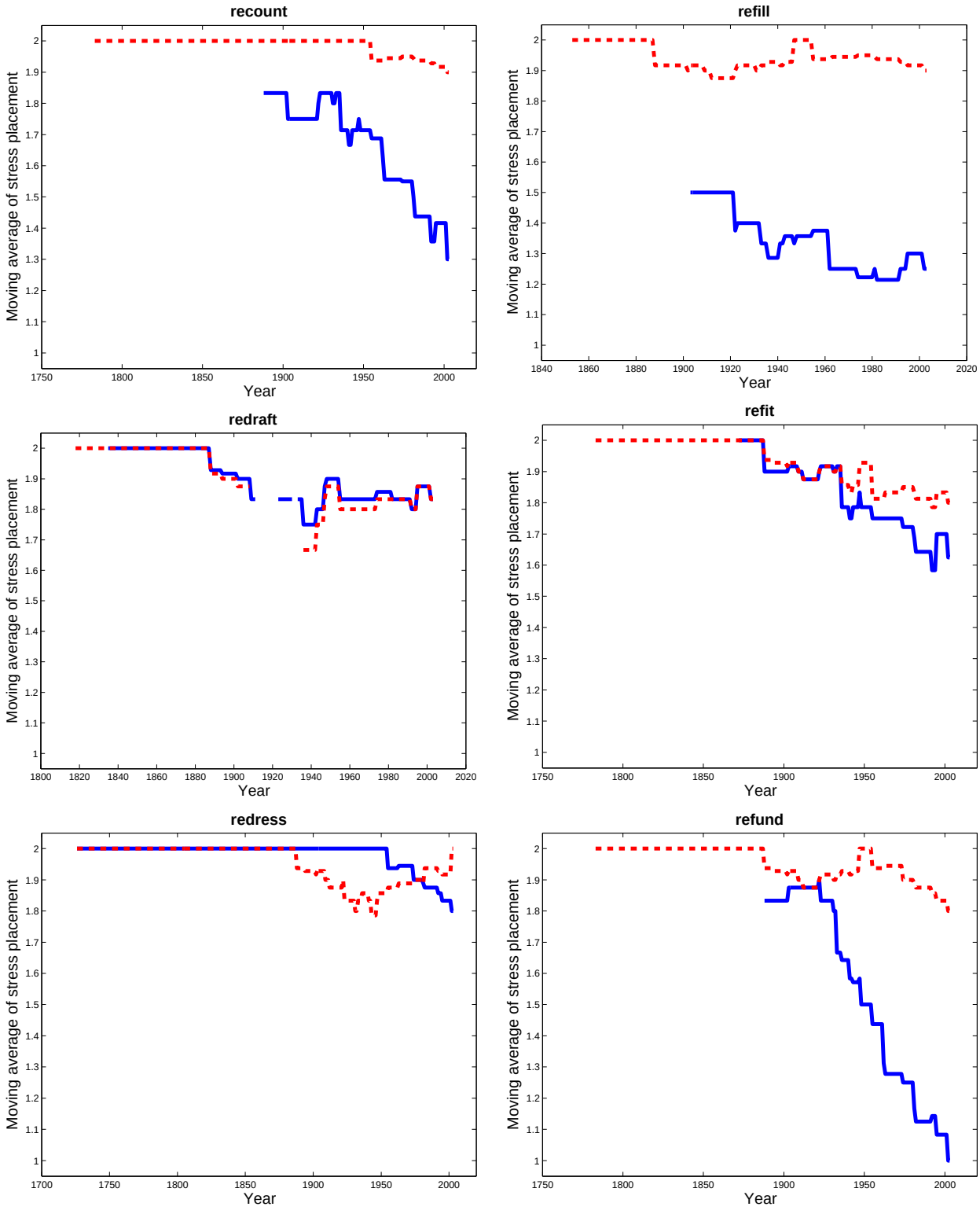

protest
Moving average of stress placement
Year

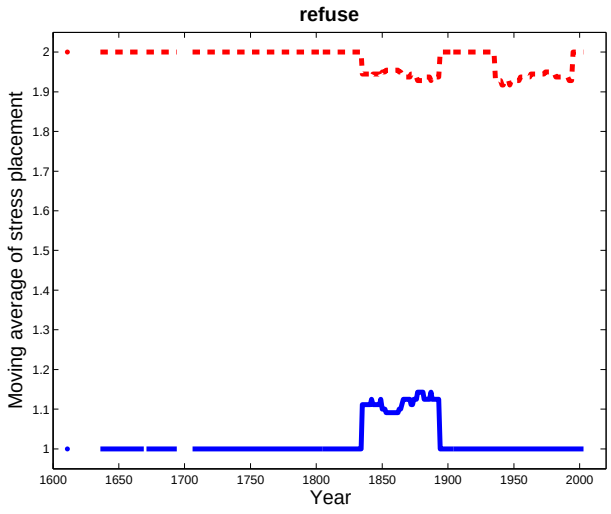

rebate
Moving average of stress placement
Year

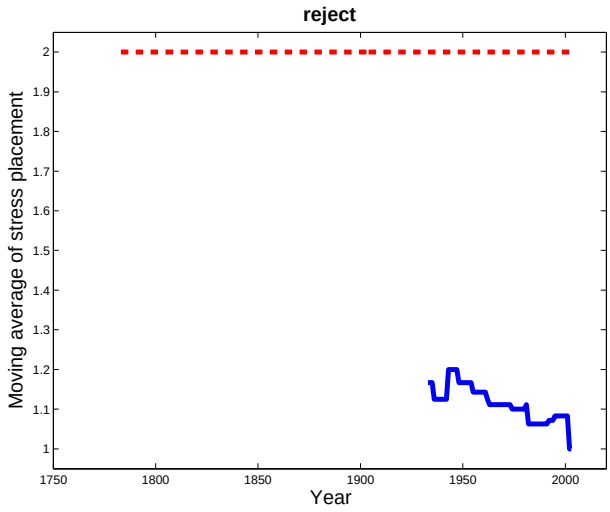

purport
Moving average of stress placement
Year

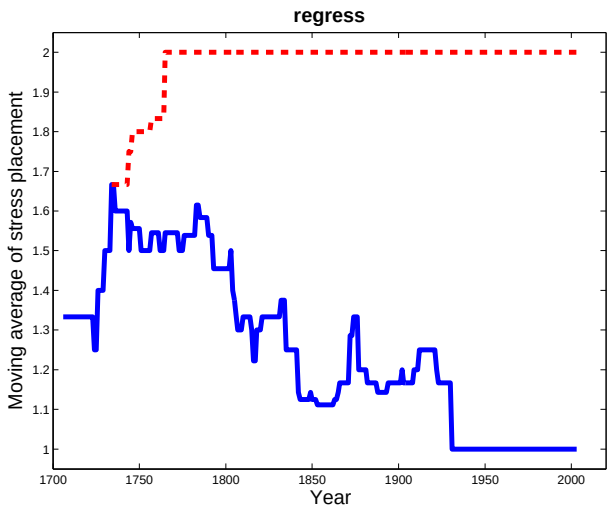

rebel
Moving average of stress placement
Year

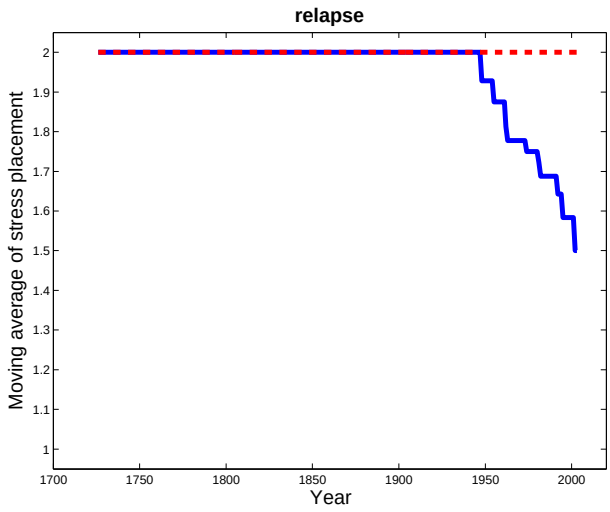

rebound
recess
recall
recoil
recast
record
Moving average of stress placement
Year

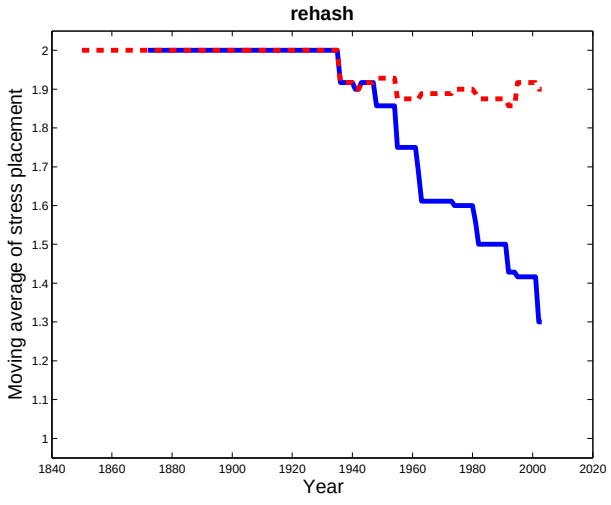

recount
refill
redraft
refit
redress
refund
Moving average of stress placement
Year

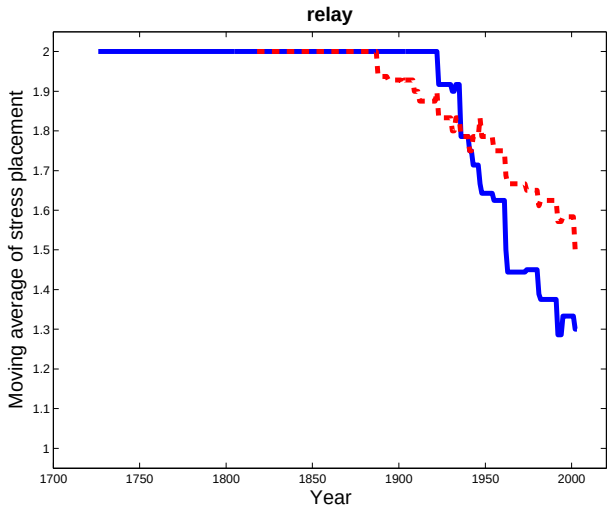

refuse
Moving average of stress placement
Year

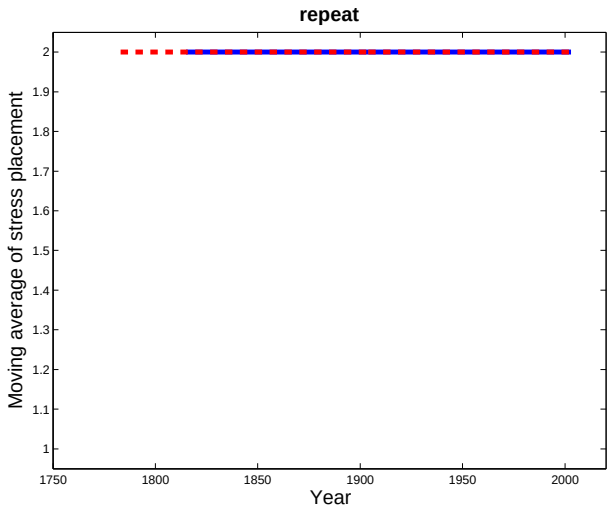

reject
Moving average of stress placement
Year

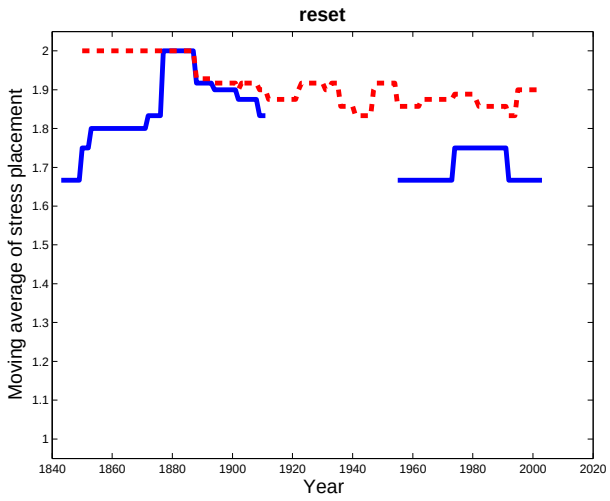

regress
Moving average of stress placement
Year

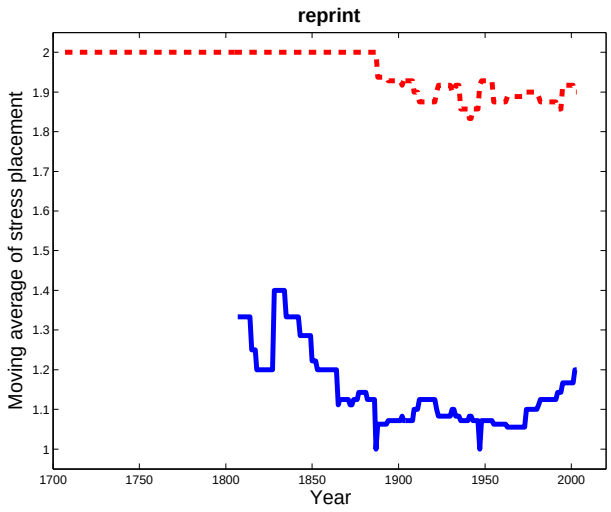

relapse
Moving average of stress placement
Year

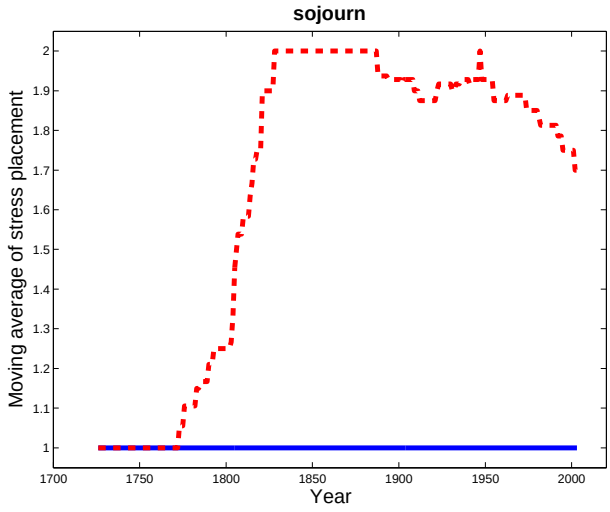

rehash
Moving average of stress placement
Year

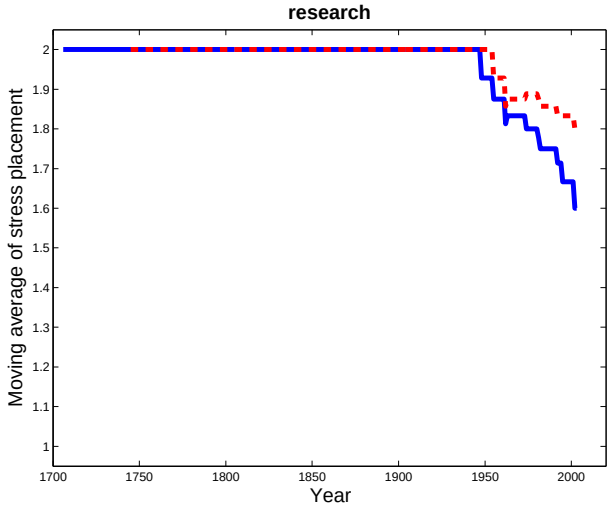

relay
Moving average of stress placement
Year

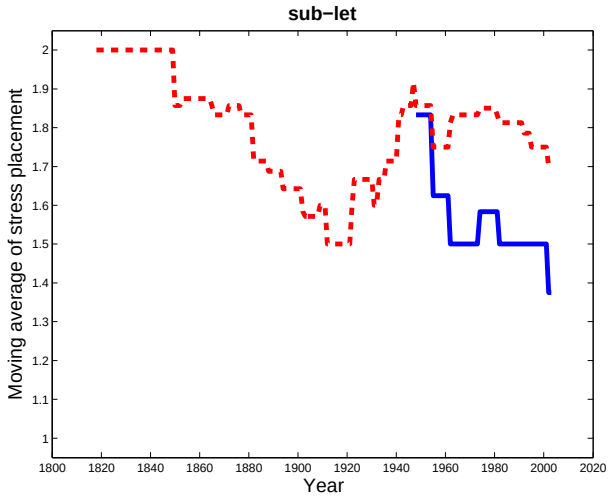

repeat
Moving average of stress placement
Year

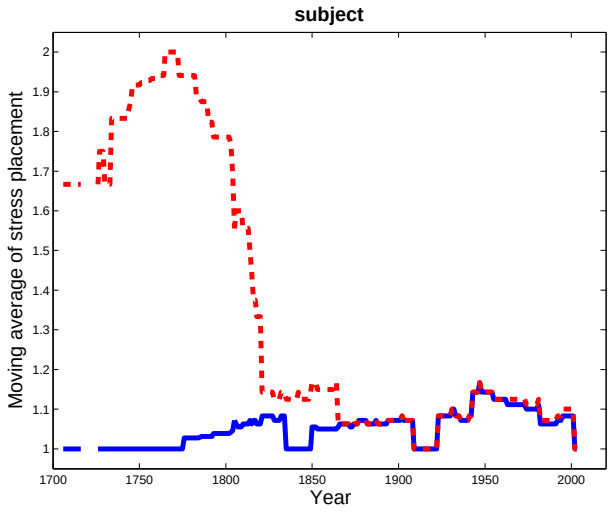

reset
Moving average of stress placement
Year

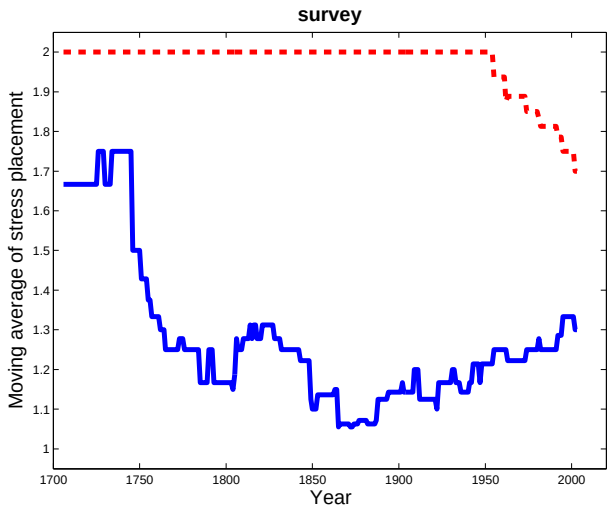

reprint
Moving average of stress placement
Year

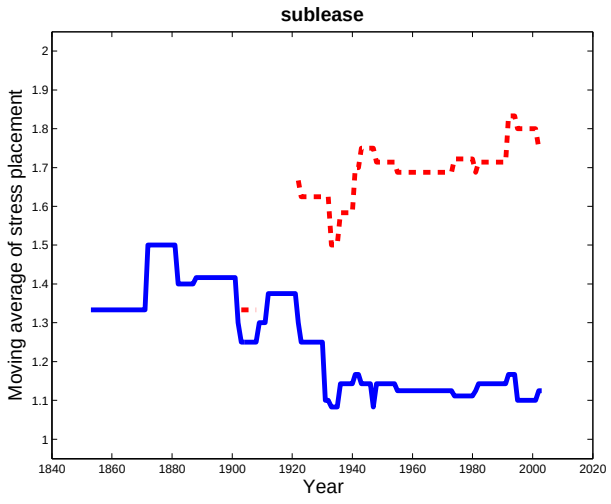

sojourn
Moving average of stress placement
Year

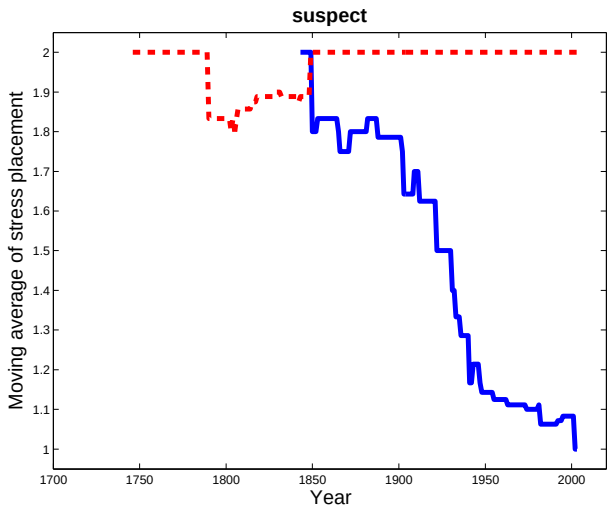

research
Moving average of stress placement
Year

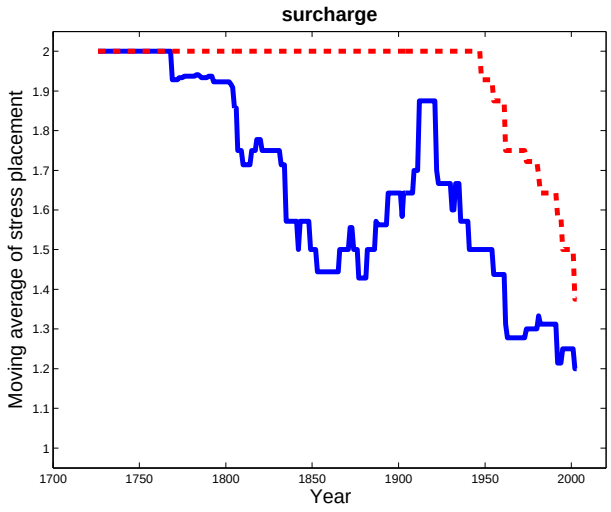

sub−let
Moving average of stress placement
Year


subject
Moving average of stress placement
Year

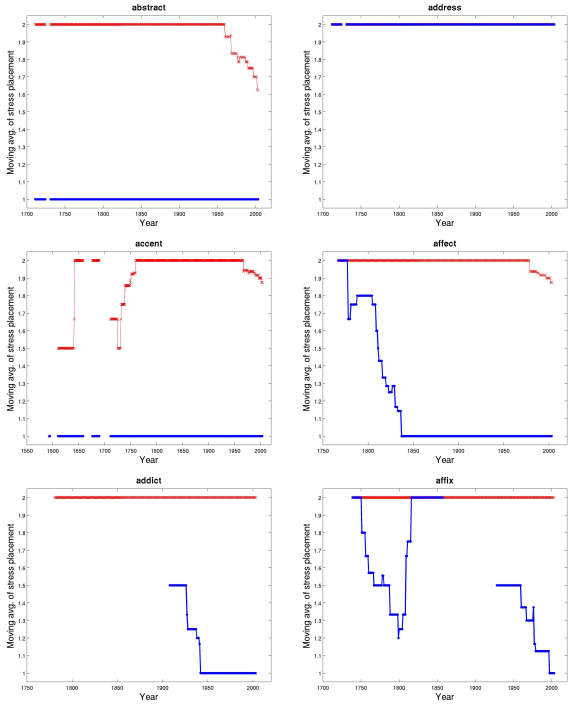

survey
Moving average of stress placement
Year

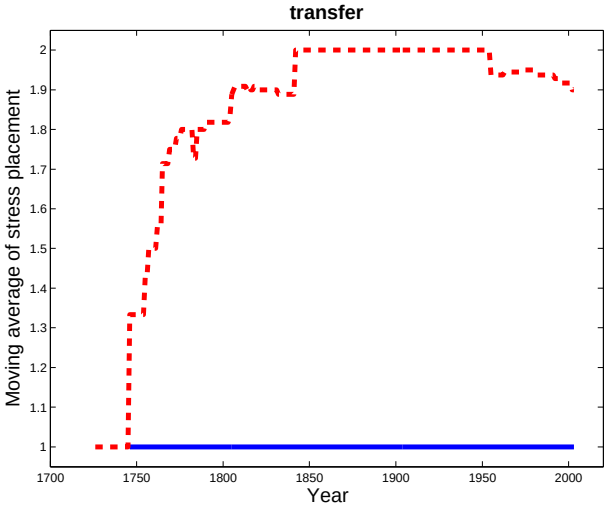

sublease
Moving average of stress placement
Year

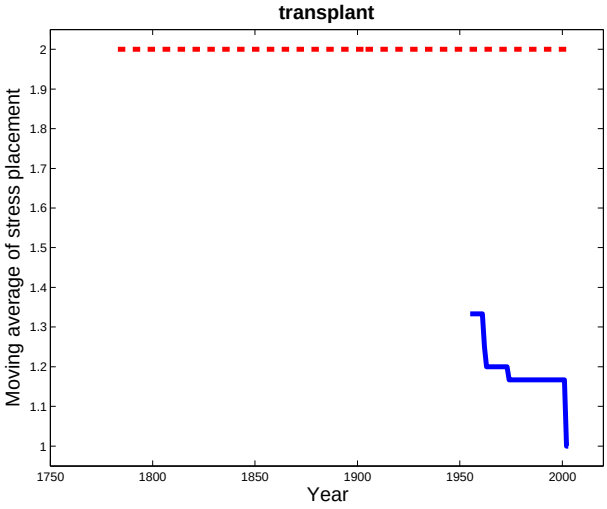

suspect
Moving average of stress placement
Year

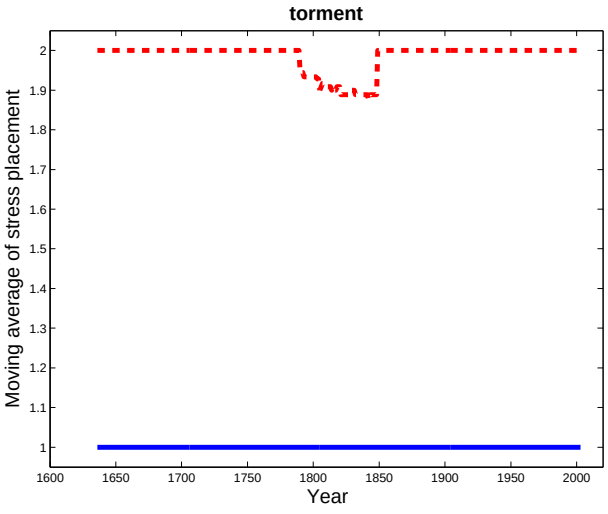

surcharge
Moving average of stress placement
Year

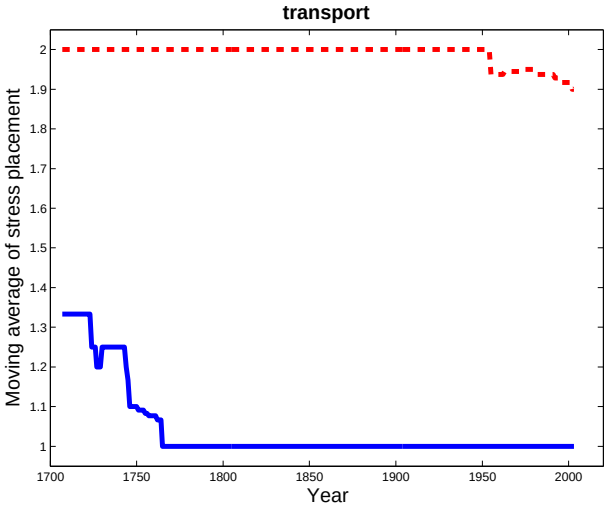
Noun trajectories are blue, verb trajectories are red. A point was indcluded at time t for the N form of a N/V pair if 2 or more dictionaries in the window $(t-25, t+25)$ listed it (and similarly for the V form)

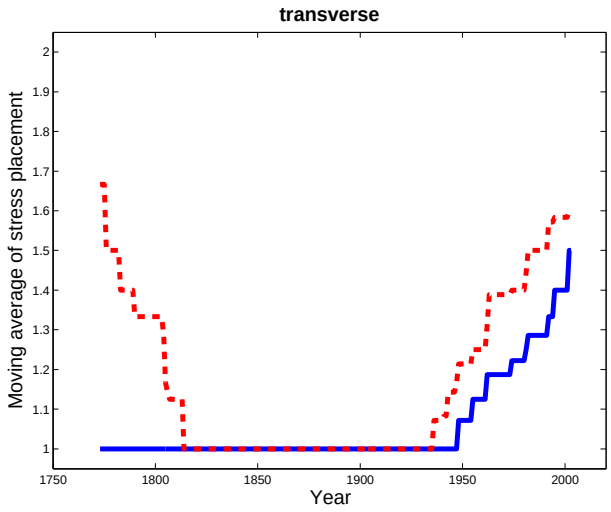


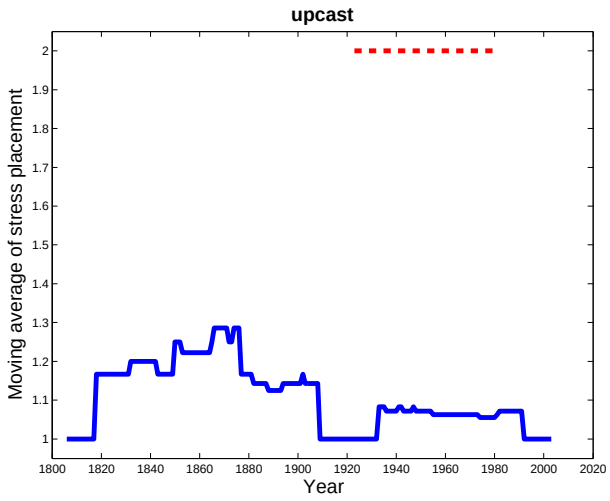


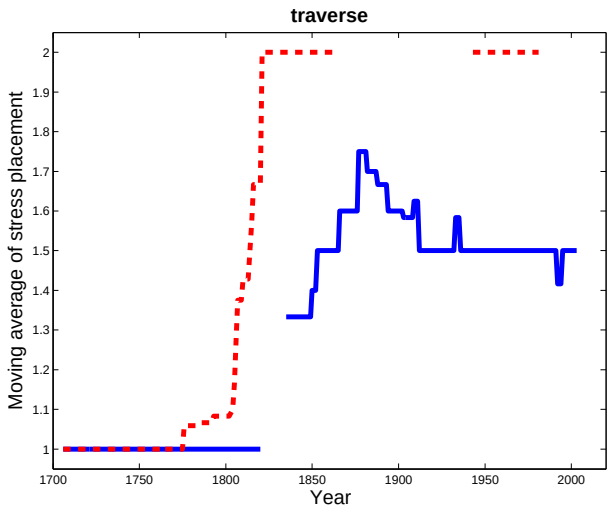


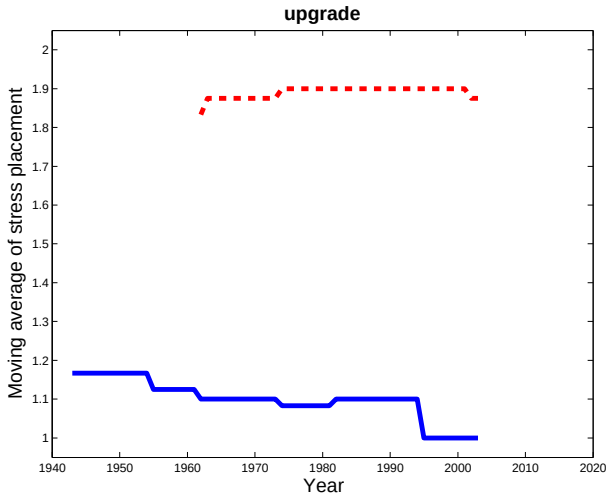


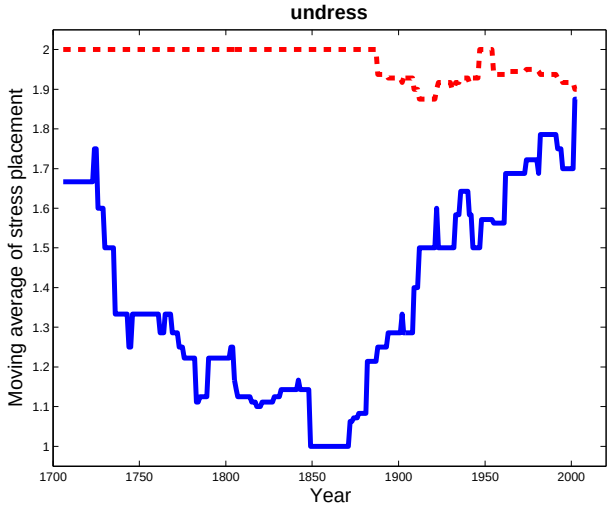


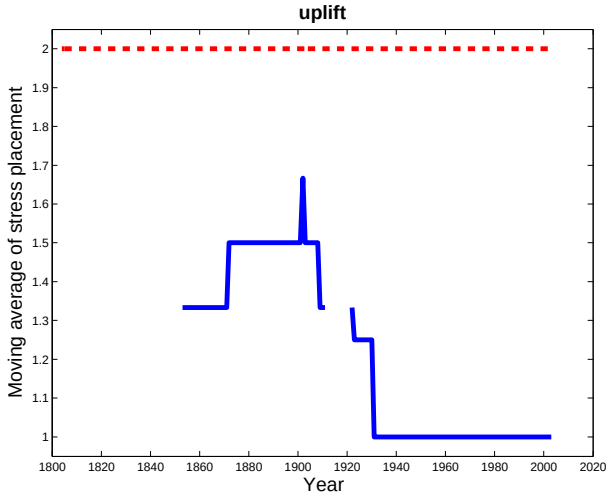

transverse
Moving average of stress placement
Year

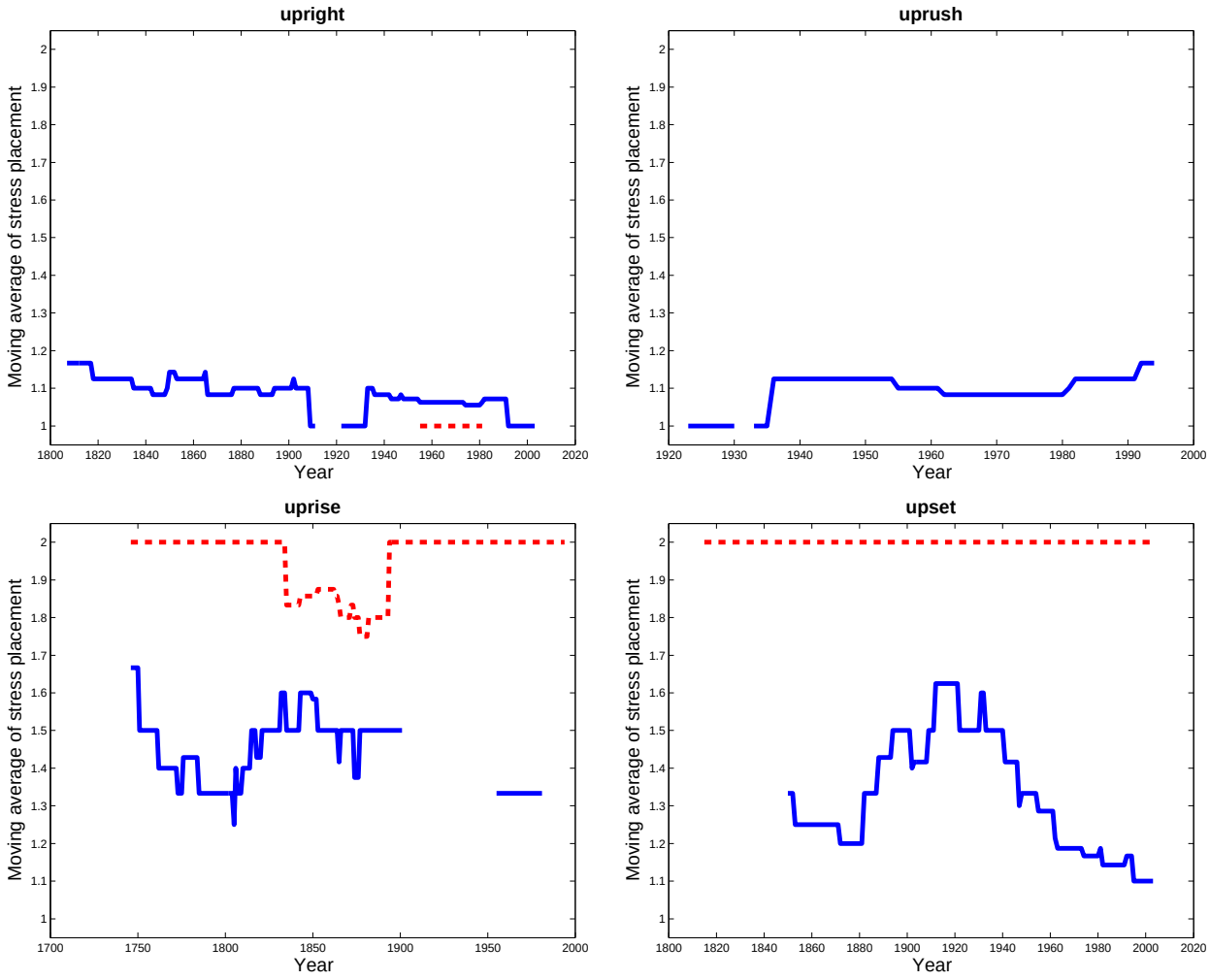

upcast
Moving average of stress placement
Year


traverse
Moving average of stress placement
Year


upgrade
Moving average of stress placement
Year


undress
Moving average of stress placement
Year


uplift
Moving average of stress placement
Year


upright
uprush
uprise
upset
Moving average of stress placement
Year