Journal of Theoretical Biology 405 (2016) 140–149

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Culturomics as a data playground for tests of selection: Mathematical approaches to detecting selection in word use

Suzanne S. Sindi [a,*], Rick Dale [b,1]

[a] Applied Mathematics, University of California, Merced, USA
[b] Cognitive and Information Sciences, University of California, Merced, USA

HIGHLIGHTS

- We develop a neutral model of the evolution of word frequencies.
- We compare the frequency of stable words in American and British English as annotated by Google Ngram.
- Empirical word frequencies deviate from neutral simulations indicating selective processes.
- Word frequencies vary in concert with one another as influenced by cultural forces.

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form
1 December 2015
Accepted 28 December 2015
Available online 21 January 2016

Keywords:
Evolution
Language
Word frequencies
Drift
Selection
Causal state

ABSTRACT

In biological evolution traits may rise and fall in frequency due to genetic drift, where variant frequencies change by chance, or by selection where advantageous variants will rise in frequency. The neutral model of evolution, first developed by Kimura in the 1960s, has become the standard against which selection is detected. While the balance between these two important forces – drift and selection – has been well established in biology there are other domains where the contribution of these processes is still coming together. Although the idea of natural selection has been applied to the cultural domain since the time of Darwin, it has proven more challenging to positively identify cultural traits under selection both because of a lack of established tests for selection and a lack of large cultural data sets. However, in recent years with the accumulation of large cultural data sets many cultural features from pre-history pottery to modern baby names have been shown to evolve according to the neutral theory. But there is accumulating empirical evidence from cultural processes suggesting that the neutral theory alone cannot account for all features of the data. As such, there has been a renewed interest in determining whether there is selection amidst drift. Here we analyze a subset English word frequencies, and determine whether frequency change reveals processes of selection.

Inspired by the Moran and Wright–Fisher models in population genetics, we developed a neutral model of word frequency variation to assess when linguistic data appears to depart from neutral evolution. As such, our model represents a possible "test for selection" in the linguistic domain. We explore how the distribution of word use has changed for sets of words in English for more than 100 years (1901–2008) as expressed in vocabulary usage in published books, made available by Google Ngram. When comparing empirical word frequency changes to our neutral model we find pervasive and systematic departures from neutrality.

1. Introduction

Evolution refers to the process by which a system changes over time. While most commonly thought of in the biological context of genetic heritability, the concept of evolution applies far more broadly. In particular, cultural phenomena have long been considered in the context of evolutionary processes. Indeed, not long after Darwin's "Origin of the Species" (Darwin, 1859), Schleicher (1869) wrote "What Darwin lays down of the animal creation in general, can be equally said of the organisms of speech." Although the evolutionary process is quite different in these two domains, two forces are thought to account for much of the observed variation – selection and drift.

In many cases, cultural change seems governed by the neutral theory, derived from concepts of genetic drift (Boyd and Richerson, 1988; Crow and Kimura, 1970; Neiman, 1995). This powerful account

* Corresponding author. Fax: +1 209 228 4060.
E-mail addresses: ssindi@ucmerced.edu (S.S. Sindi), rdale@ucmerced.edu (R. Dale).
[1] Fax: +1 209 228 4007.

http://dx.doi.org/10.1016/j.jtbi.2015.12.012
0022-5193/

of evolutionary dynamics – that variational change is purely frequency dependent – has been applied to everything from pre-history pottery (Bentley et al., 2004) to modern baby names (Hahn and Bentley, 2003) and much more besides (see Bentley et al., 2011, for a review). A chief goal of those using the neutral theory is to demonstrate that aggregate statistical properties of cultural variants exhibit structure that is elegantly predicted by the theory. In this paper, we consider the case of vocabulary change across a century of published books. Indeed, in this linguistic domain, the neutral theory has already been successfully applied. It shows the power-law relationship between word frequency and cumulative probability (Reali and Griffiths, 2009), patterns of color word use (Acerbi and Bentley, 2014) and other cultural products related to language e.g., turnover in academic terminology: (Bentley, 2008).

Processes of selection have also become influential in the cultural domain, but there is continued debate about the nature of the selective mechanisms. For example, Acerbi and Mesoudi (2015) note that most researchers in this domain are now cultural Darwinists, yet there is still debate about the nature of the selection itself.[2] Processes of selection may be driven mostly by individual cognitive agents, making copies of variants present in the cultural pool. Statistics at the population level are driven by individual, local decisions. Another quite different perspective is that there are constraints inherent in the cultural variants transmitted that may dictate their fitness – which may reflect mechanisms outside the individual. For example, certain cultural or behavioral patterns may be more or less memorable, so that a cognitive agent may be more or less constrained by the structure already present in the cultural milieu. As Acerbi and Mesoudi (2015) note, it is likely that both of these processes are relevant to cultural evolution and change – we need new tools to explore these mechanisms and identify signatures of selection.

With the recent accumulation of massive cultural datasets, there has been a renewed interest in identifying selective elements amidst the cultural drift.[3] For example, in the domain of baby names, Gureckis and Goldstone (2009) demonstrated that name choices are biased by potential cultural factors, such as the popularity of a name. They do this by demonstrating that the patterns of change across years, for various names, are best captured by models that factor in such choice biases. Acerbi and Bentley (2014) demonstrate that processes of selection may be present in various cultural domains, such as popular songs and baby names (see also Bentley, 2008). They use patterns in "turnover rates" for variants – the extent to which a song or baby name pops out of top lists. Their measure of turnover exhibits a distribution that is not as well predicted by the neutral theory as a model that includes biases, such as selection by conformity (for similar discussions in other domains see Kandler and Shennan, 2013; Steele et al., 2010).

There has been considerable discussion about these and other tests of selection, and their effectiveness in determining the presence of selection in natural systems (see Linnen and Hoekstra, 2009, for review). For example, recently, (Zhai et al., 2009) conducted a study of the statistical power of various tests of selection in simulated biological evolution. One finding they describe is that if selective sweeps over variants are relatively rare, then population genetic tests will have little statistical power to detect them. These kinds of tests have been elegantly applied in the cultural domain, as well. For example, Rogers and Ehrlich (2008) compared functional vs. symbolic features of canoes from cultures in Polynesia. Using statistics of the sort in population genetics with silent vs. non-silent variants, they find that functional features have stabilized, while symbolic features tend to change more quickly.

Given these previous studies, there is ample room for developing and exploring new ideas about selection, especially in the cultural context. The recent advent of "big data" in social and cultural sciences offers new "data playgrounds" to explore the change in cultural variants. These new sources of data may reveal new techniques and principles in various cultural domains, such as the analysis of historical word-frequency changes. If the principles underlying this change are non-trivially similar to those in the biological cases, then development in one domain will be of use to the other. We note that detecting selection in biological data sets is a comparatively well-studied problem. Adapting selective tests to cultural evolution is complicated by significant differences between cultural and biological evolution. For example, the concept of "alleles" – heritable elements which selection is thought to operate on – and "inheritance" are far more rigid in the biological domain than in the linguistic or cultural domain. But while the relationship between biological and cultural evolution is imperfect, commonalities between the two domains mean that similar evolutionary models and statistical tests may be shared (Mesoudi, 2007). Throughout our analysis and discussion, we discuss commonalities and differences between evolution in the linguistic and biological domain.

In this study, we take three approaches to exploring the domain of vocabulary change. First, we develop a simple mathematical model representing neutral evolution of word frequencies. Of course, the precise nature of the neutral model is crucial for drawing any inferences (Blythe, 2012), and we do not mean to propose a definitive neutral model that accomplishes the best contrast with a system that shows selection. Instead, it offers a tractable basis for developing expected statistical properties under the fundamental "random copying" conditions that have been proposed in other cultural domains (e.g. Hahn and Bentley, 2003).

Second, we implement this neutral model in a series of simulations that have properties of expected word frequency distributions (Zipf, 1949). This allows us to demonstrate the effectiveness of the mathematical formalism in precisely describing neutral numeric simulations. The simulation also offers a numerical baseline with which to compare the observed Google Ngram data.

Third, finally, we conduct an analysis of word frequency change in English from 1901 to 2008 through Google Ngram (Michel et al., 2011), calculating the measures that are used in the mathematical and simulated models. The "playground" we chose for this paper was intended to be very simple to facilitate derivation and discussion. First, we ignore for the present paper the possibility of words to "turnover" – instead, we focus on a set of "stable variants," those which occur in all of the years from 1901 to 2008 ($N=6489$) in at least 1000 books each year. This means our analyses are on the variation over time within this closed set of words – words may grow or diminish in prominence, but they do not vanish from the "population." We note that focusing on "stable" linguistic variants has a strong analogy to examining biological evolution by examining changes in a "core" set of genes and not the entire genome (Daubin et al., 2002).

The tests of selection we propose may provide a basis for determining the presence of "selective sweeps" (Zhai et al., 2009), and mining the data for clusters of words that are undergoing selection, and at what time. Previous work on Google Ngram uses intuitive and well-known historical events to find these patterns (Michel et al., 2011). In our case, we assume that we do not know that interesting patterns of change are occurring, and use the tests of selection to (i) determine that change exceeds random copying and (ii) find the clusters in which the change is occurring.

[2] In language, for example, adaptationism has become a prominent way discussing language origins and change especially due to Pinker and Bloom (1990). There is extensive ongoing investigation about the locus of selection – whether genetic, individual, cultural, and in what combination (Christiansen and Kirby, 2003; D'Andrade, 2002; Hurford et al., 1998).

[3] This also occurred, to some extent, sometime ago when the neutral theory was proposed in population genetics models; considerable evidence has since been adduced that selection is a relatively frequent phenomenon (Gillespie, 2004).

We design statistical signatures that are suitable for data that contain historical change over a probability distribution of variants. Akin to Gureckis and Goldstone (2009), we argue that patterns of *frequency change* in historical data are crucial ingredients to assessing neutrality or selection. Statistical signatures derived from these historical data show that there is pervasive and systemic departure from neutrality. In particular, we focus on three statistical signatures present in the Google Ngram data.

First, we note deviations in the *kurtosis of frequency changes*. Selection amplifies or dampens the change in word frequencies, by operating at the tails of variational drift, producing increased kurtosis in the distribution of even first-order frequency changes. Second, we note *non-independent changes in word frequencies*. Words are part of a broader cultural system and so their frequency changes are not statistically independent of each other; a dimensionality reduction over frequency changes demonstrates latent structure underlying vocabulary change. Finally, we provide statistical evidence that *changes in word frequencies are complex*. The frequency change of words seems to be generated by complex and often biased models. That is, word frequencies exhibit fluctuations which are not indicative of random unbiased copying.

Though the neutral model does a very good job of accounting for broad statistical characteristics of the observed Google Ngram data, we find distinct departures from neutrality. The measures we explore offer new ideas in the cultural domain for tests of selection, and suggest that both neutrality and bias are part of the story of this closed set of English words. In the General Discussion, we revisit concerns about cultural selective mechanisms, as described in Acerbi and Mesoudi (2015). The statistical signatures we introduce may help us to identify cultural constraints at the population level, highlighting the role of shared structure across word-use patterns. For example, a global event such as a world war presents obvious strong external structure that shapes word usage. The presence of such an effect is obvious, but as we demonstrate, if we did not know about such cultural events, it would be possible to identify these "selective sweeps" over words using the analyses we describe.

2. Methods

2.1. Google Ngram data

The cultural domain we explore in this paper is how the distribution of word use has changed for the set of words that occur in both American and British English from 1901 to 2008. There are impressive data available through Google Ngram for this purpose, offering a precise distribution of word occurrence across these years. Although this data set represents only a limited view of cultural evolution, computational and quantitative exploration of vocabulary change is in and of itself a highly active area of research (e.g. Dale and Lupyan, 2012; Michel et al., 2011; Reali and Griffiths, 2009; Lieberman et al., 2007; Petersen et al., 2012b, 2012a), and indeed vocabulary is a chief source of data regarding cultural change itself (e.g., recently, Anthony, 2010).

As mentioned in the introduction, we first identify the subset of words in the Google Ngram data set which correspond to "stable variants". These are words which have been a stable part of the English language during the entire time span in question. In our analysis, we restricted to the set of words that occur in at least 1000 books per year in each of the years under consideration (1901–2008). There were over 7000 books annotated for each year considered (see Fig. 1) books, peaking at 149,373 books in 2008. In total there are 6489 words present in the current data set, representing the stable vocabulary set for English across 108 years.

Although restricting to only the stable variants in the Google Ngram data set removes the process of loss and gain of words, it is analogous to a research approach in the biological case. By restricting the data set to this subset we are performing the analogous study to examining evolution of an organism by focusing on fluctuation in the "core genome" (Hsiang and Baillie, 2005; Richter and King, 2013; Waterhouse, 2015; Tatusov et al., 2003).

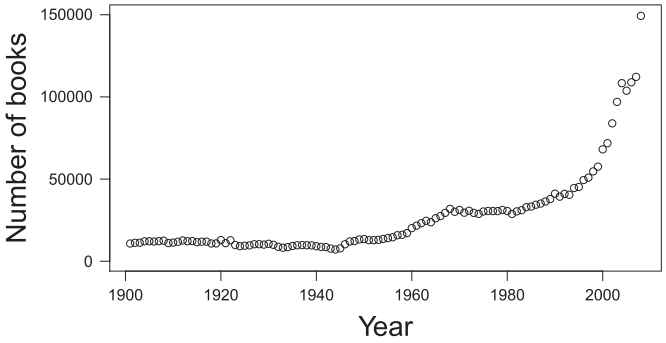

Fig. 1. *Books per year in Google Ngram*. Our analysis considered only the set of "stable variants," words which occurred each year from 1901 to 2008 and in at least 1000 books per year. Since each year had at least 7000 annotated books, our restriction required that words in consideration, at most, occurred in 14% of the books in each year.

To compare Google Ngram data to our neutral model, we normalize the empirical word frequencies. First, we convert all word frequencies to normalized proportions for each year from 1901–2008. Next, we convert the frequency of each word to its z-score. That is, if $f_{w,t}$ is the frequency of word w at time t, and Y is the number of years, we determine the mean and variance of the empirical distributions:

$$\overline{f}_w = \left(\frac{1}{Y}\right)\sum_t f_{w,t}, \tag{1}$$

and

$$(\sigma_{f_w})^2 = \left(\frac{1}{Y}\right)\sum_t \left(f_{w,t} - \overline{f}_w\right)^2. \tag{2}$$

Then, we convert the empirical distributions to their normalized frequencies (z-scores):

$$p_{w,t} = (f_{w,t} - \overline{f}_w)/\sigma_{f_w}. \tag{3}$$

We note that if the data were generated independently according to the neutral model, we expect $p_{w,t} \sim N(0,1)$. In the analysis below, we also consider the change in normalized frequency:

$$\Delta p_{w,t} = p_{w,t} - p_{w,t-1} = (f_{w,t} - f_{w,t-1})/\sigma_{f_w}. \tag{4}$$

Under the assumptions of neutrality, we expect $\Delta p_{w,t} \sim N(0,2)$.

2.2. Neutral model of word frequency evolution

We take as inspiration for our work the neutral theory of molecular evolution first proposed by Kimura (1985) and the Wright–Fisher model (Ewens, 2004). In the simplest formulation this neutral model considers a haploid asexually reproducing population of N individuals. We map the Wright–Fisher model from biological to linguist evolution by considering each word as an allele; our trait of interest is the frequency of an allele at a particular locus, denoted as black, white and gray circles in Fig. 2. In the neutral model, all alleles have equal fitness and so individuals are chosen from the previous generation at random. Thus, while in expectation allele frequencies remain constant, they will vary as an unbiased random walk.

We adapt the neutral Wright–Fisher model to the linguistic domain in a straightforward manner. We consider the population in a given year as the concatenated text of all books from that year. The frequency of a word is simply the number of instances in that word divided by the total length of the concatenated text. We

generate the text for the following year by randomly sampling words from the previous year, just as in the Wright–Fisher model.

Finally, to model the word frequency data from Google Ngram we add some additional detail. First, as we were interested in the frequencies of stable linguistic variants, we track only a subset of the full collection of words and we require that each word is sampled at least once in the next generation. Second, as the size of written texts increases throughout the term of our study, we considered size of the text at each successive generation to increase in accordance with the empirical data from Google Ngram. Third, to more accurately model the frequencies of words, we begin with the underlying assumption that the initial frequencies of words sampled in the population follow a Zipf distribution by seeding the simulation with the empirical distribution of our stable variants at the year 1901.

Although word frequencies generated by this neutral model are dependent, analysis of simulated data indicated that these dependencies were weak. Thus, for computational simplicity, in the results below we generate word frequency trajectories according to the neutral model by sampling each word independently.

As for the Google Ngram data, distributions of our neutral model are converted into their corresponding z-scores. As mentioned above, according to the neutral model, we expect each word trajectory to behaving as a normal random variable with $\mu = 0$ and $\sigma = 1$. Indeed as shown in Fig. 3, computational realizations of our neutral process do generate frequencies with the expected distributions of normality $N(0, 1)$ and as expected kurtosis $K=0$. In what follows, we systematically examine differences between simulations of our neutral evolutionary process and the Google Ngram data.

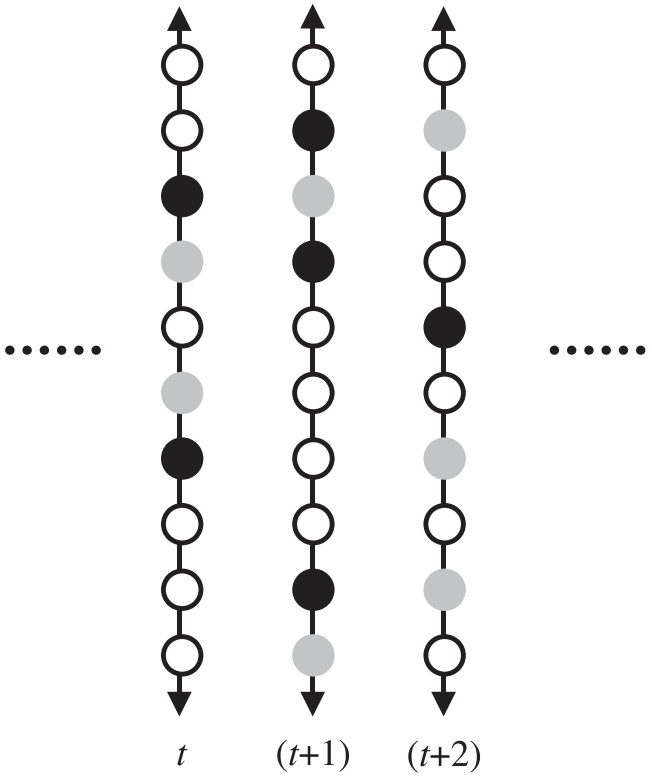

Fig. 2. *Wright–Fisher model.* In the Wright–Fisher model of evolution individuals in generation $(t+1)$ are chosen from the previous generation t by random. This *frequency dependent selection* creates allele frequencies which vary as an unbiased random walk.

3. Results

3.1. Frequency changes exhibit excessive kurtosis

We first consider the overall stability of word frequencies by examining the patterns in the change of frequency. Under the neutral model, we expect changes in word frequencies to follow a normal distribution and, as such, would have excess kurtosis (K) near 0. Our core set of words deviate with significantly high K. The higher the excessive kurtosis, the more likely that future changes in frequency will be either extremely small or extremely large (Rachev et al., 2011). As such, kurtosis is sometimes termed the "volatility of volatility."

Words in the observed data can show much greater kurtosis in the frequency distribution over the century of data. Each word can be given a K value from its own century worth of data (the distribution of its changes from 1901 to 2008). In Fig. 4 we show the distribution of K values for each word on a log scale, because some words exhibit very high K values. In the comparative neutral model, K is significantly greater than 0, but the general distribution of observed words has a much longer tail and by a Kolmogorov–Smirnov (K–S) test shows a difference in distribution, $D = 0.19, p < 10^{-10}$, and a t-test finds a different in mean, Welch's $t(7635.51) = 21.8, p < 10^{-10}$. These results remain strongly significant in log-transformed data, $p's < 10^{-10}$.

Positive kurtosis in a distribution can be commonly attributed to two factors: the presence of outliers in the distribution or that the underlying distribution itself is non-normal. For example, in the inset of Fig. 4 we observe that the frequency in the word "landing" increases substantially during the 1940s.

We continue to assess the presence of outliers in our data by examining the year-to-year variation in frequencies. We find even more substantial differences in the tendency for frequency change to take place *within* a year. By taking the year-to-year difference in frequency, and assigning a K value to the distribution of change within each year, we see substantial divergence from neutrality. Observed data involved considerably higher entropy, suggesting that words tend to change much more erratically from year to year than what would be expected from the neutral model. Again, a K–S test shows substantial differences in the log-transformed data, $D = 1, p < 10^{-10}$, and the means of these distributions are, of course, significantly different, Welch's $t(108.15) = 18.7, p < 10^{-10}$. In Fig. 5, observed data are in green, and neutral data in red. The correlated behavior of word frequency changes per year is indicative of the influence of non-neutral forces acting on the temporal dynamics.

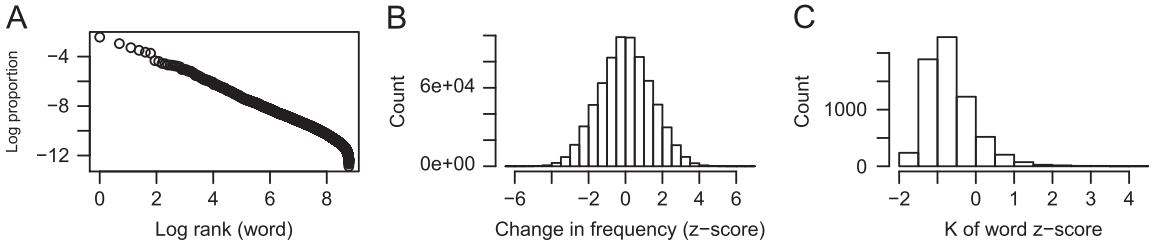

Fig. 3. *Simulations of neutral model.* Computational simulations of the neutral model demonstrate expected behavior in the z-scores and kurtosis (K). (A) Zipfian distribution holds between log word rank and word frequency. (B) Word frequency changes are normal. (C) Word frequency changes have K near approximately 0 (normal).

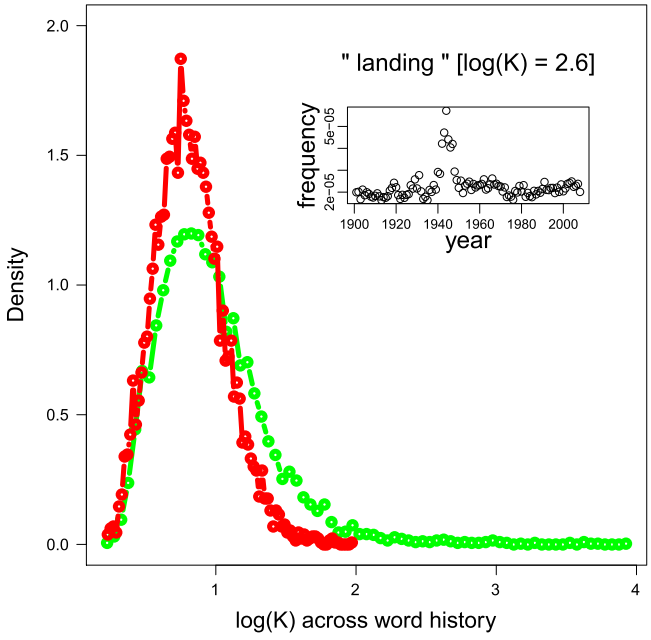

Fig. 4. *Kurtosis of empirical word frequencies.* We examine the kurtosis in the change of word frequencies. By analyzing the center of changes in word frequencies for each word, we find that the observed Google Ngram data (green) has substantially higher kurtosis that the our neutral model (red). In order to plot on log scale, we use K where normality = 3 (i.e., non-excess K). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

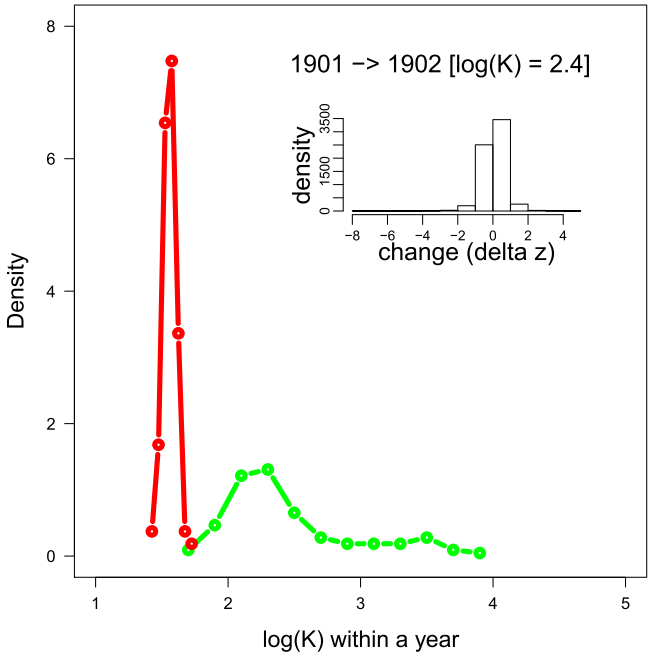

Fig. 5. *Kurtosis by year.* We take the 107 ($Y-1$) distributions of frequency changes and plot the density function of the K for these within-year fluctuations (green). There is distinct departure from neutrality (red), indicating that the year-to-year frequency changes are non-normal. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

3.2. Correlated word frequency changes

Our interpretation of the non-neutral behavior is that words change in their frequency due to the influences of social and cultural forces. If this is the case then, unlike the neutral model where word frequencies would change independently, we expect the observed word frequencies to undergo changes in concert. Previous studies such as Petersen et al. (2012a) and Michel et al. (2011) have observed cultural forces influencing word frequencies and we see similar behavior. For example, we observe a similar distribution of word frequencies for "war" and "Germany" during this 108-year period in a manner that reflects the temporal impact of World War I and World War II (see Fig. 6).

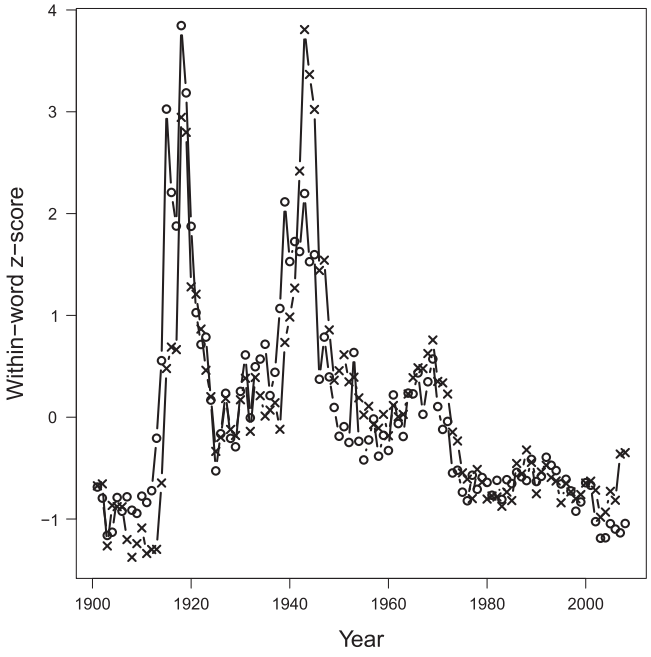

Fig. 6. *Cultural effects on word frequencies ("war" and "Germany").* Word frequencies change visibly due to the impact of culture. In this case we note that the frequencies for "war" (x) and "Germany" (o) change in a highly correlated fashion likely due to the response from World War I and World War II.

We sought to determine the extent to which word frequencies vary together. A simple way to examine large data sets for the signature of dependence is through a principal components analysis (PCA) of the data set. PCA is a computational technique to explain the spread variance in a (usually) high dimensional data set (Pearson, 1901). In our data, we have a total of 6489 words each of which we can treat as a point in either the 108-dimensional space of normalized frequencies ($p_{w,t}$) or 107-dimensional space of frequency changes ($\Delta p_{w,t}$).

PCA is closely related to the singular value decomposition (SVD) and is an attempt to uncover lower dimensional structures. Mathematically, it is often interpreted as fitting an ellipsoid of minimal radii around a cloud of points. If all axes of the ellipsoid have similar length, this would suggest the data are unstructured. If a few directions – principal components – have substantially longer axes than others then this suggests significant correlation and that the high-dimensional data can be well described by lower-dimensional projection along the extremal principal components.

As evident in Fig. 8, we observe significant deviation from neutrality for the empirical data. We observe the very similar PCA structure for the normalized frequencies between the empirical and neutrally simulated data, though there is still a significant difference in the drop in component score for the observed data: by K–S test $D = .32, p < 10^{-4}$, though mean PC score cannot be inferred to be different, Welch's $t(213.9) = .23, p = .82$. For the frequency change patterns, we find much greater different in the observed principal-component scores, with the observed data having much more lower-dimensional structure, K–S test $D = .39, p < 10^{-6}$, and means are different, Welch's $t(124.9) = 5.3, p < 10^{-6}$. In particular this observation shows that our "core" words undergo dependent changes in frequency which are likely the response to cultural phenomena.

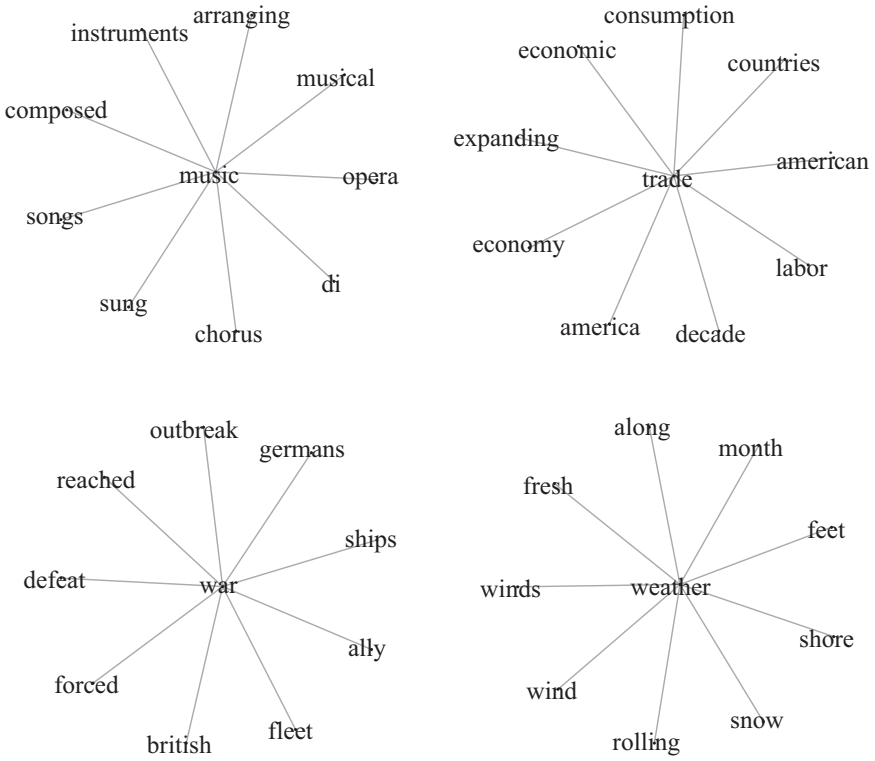

Fig. 7. *Words changing together in time.* PCA results from frequency changes show that words move together in ways that are intuitive when we look for most-closely similar words in component space. Here, we show the words most correlated in their frequency changes with 'music,' 'trade,' and 'war,' and 'weather.'

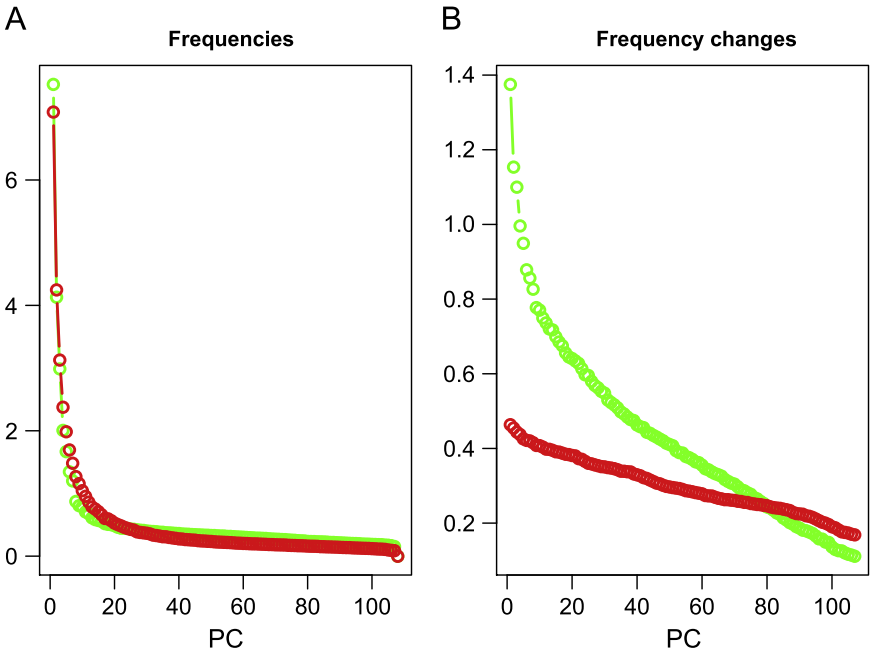

Fig. 8. *Principal component analysis.* We examine both our empirical data as well as simulated neutral data through principal component analysis (PCA). (A) We note that there is no difference in variances associated with the principal components for the word frequencies. However, when we examine the change in word frequencies. (B) We observe significant differences in the empirical Google Ngram data.

Although the cultural impact of global war has been previously noted (Petersen et al., 2012a), the use of PCA provides a historically agnostic way to identify forces and patterns of correlated change. We identify relevant clusters of correlated words by projecting them onto the lower-dimensional space characterized by the 10 largest principal components. We consider two words to be "highly correlated" if their Pearson correlation coefficient exceeds 0.9. As expected, the empirical data exhibits significantly more highly correlated pairs of words (58,181 pairs with 0.9 correlations or greater) than data generated under the neutral model (14,635). We can now choose words of interest and find which words lie in their "historical cluster" of nonindependent change (see Fig. 7).

It is worth noting, as we will revisit below, that these non-independent changes in word frequencies mirror the genetic

hitchhiking effect where loci undergoing positive selection will impact the frequencies of nearby loci (Gillespie, 2004; Maynard and Haigh, 2007).

3.3. Generative models of word frequency changes

Lastly, we utilize Causal State Modeling (CSM) to demonstrate that word frequencies are generated by processes exhibiting significant deviations from a random-unbiased walk. CSM emerged as a topology-agnostic method for determining the underlying "discrete causal states" and the transition probabilities in a discrete state Markov process (Crutchfield, 1994; Shalizi and Shalizi, 2004). In fact, an outcome of CSM is the inference of a hidden Markov model (HMMs) but without needing to assume an underlying topology of hidden states. The underlying theoretical principles were developed for doubly-infinite sequences from a discrete symbol space but can be adapted for finite sets of sequences (Shalizi and Shalizi, 2004). In what follows we briefly describe CSM as applied to our word frequency data and then demonstrate that word frequency changes do not follow a neutral model (applied introductions can be found in diverse domains, including: Boschetti, 2008; Dale and Vinson, 2013; Kelly et al., 2012).

CSM relies on a discrete sequence space, and so we first convert observed word frequency changes to discrete states. While there are several natural ways to do this, we simply convert each frequency change to its sign by assigning a symbol 1 if the word frequency increases and -1 if the word frequency decreases (see Fig. 9). While we lose information in the magnitude of the change, we can more clearly discern overall trends in the nature of the sign changes. In particular, we interpret the sign of frequency as being emitted from the outcome of a Bernoulli trial where we choose to increase with probability p and decrease with probability $(1-p)$. However, the parameter p associated with this Bernoulli trial may in fact change according to a hidden (causal) state.

The goal of CSM is to identify the set of equivalence classes (hidden causal states) which are associated with a specific set of emission probabilities and to uncover the rates of transitions between classes. In the infinite sequence case, the equivalence classes consist of sets of infinite prefixes; in the finite case a word length L is used. For example, consider discrete sequence x_n generated by a series of independent coin flips where we associate 1 with heads and -1 with tails. Suppose that $L=3$ and we observe that for all strings of length L are equally likely to be followed by 1 or -1:

$$P(x_{n+3}=1\,|\,x_n,x_{n+1},x_{n+2})=P(x_{n+3}=-1\,|\,x_n,x_{n+1},x_{n+2})=0.5.$$

From this we might reasonably conclude that the series of symbols is generated by the outcomes of a single unbiased coin. As such, all length L strings would belong to the same causal state (Fig. 10(A)). However, if the emitted symbol depends on the current state then there may be more than one causal state. In Fig. 10(B) we illustrate a system with two causal states, each of which can be interpreted as a biased coin. Coin s_1 is more likely to emit a $+1$ and remain in the same state than a -1 and change states and similarly for s_{-1}. We again note that since we do not observe the sequence of causal states directly, they represent "hidden" states.

We utilized a previously published method for CSM (Kelly et al., 2012) on both the empirical and neutral data ($L=3, p<0.005$). In comparison to the neutral data, the Google Ngram data differed in two substantial ways. First, the Ngram data had a significantly different distribution of causal states. For neutrally generated data, only 5/6489 words required more than a single causal state. For the Google Ngram data 431/6489 ($\approx 7\%$) required at least two causal states, $\chi^2(1)=428.7, p<10^{-10}$. Larger numbers of causal states indicate words with more complicated dynamics. For example, the fairly innocuous sounding word "burden" was predicted to have 5 causal states (see Fig. 11).

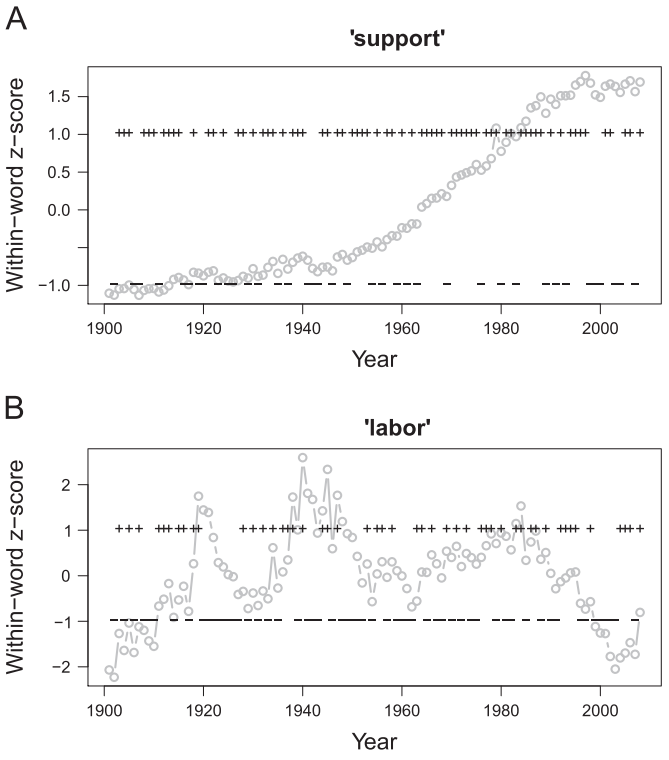

Fig. 9. *Sign of frequency change.* In order to employ causal state modeling (CSM) we translate our word frequencies to a discrete state space by examining the sign of the change in frequency with each year. Here we show two representative examples (A) 'support' and (B) 'labor'. In gray are the observed frequencies, and these are converted into signed changes, labeled as '+' (+1) for increasing frequency, and '−' (−1) for decreasing frequency, from year to year. Note that 'support' shows distinct regimes of consistent upward frequency change, while 'labor' shows an ebb and flow where change is stable for a time until it shifts.

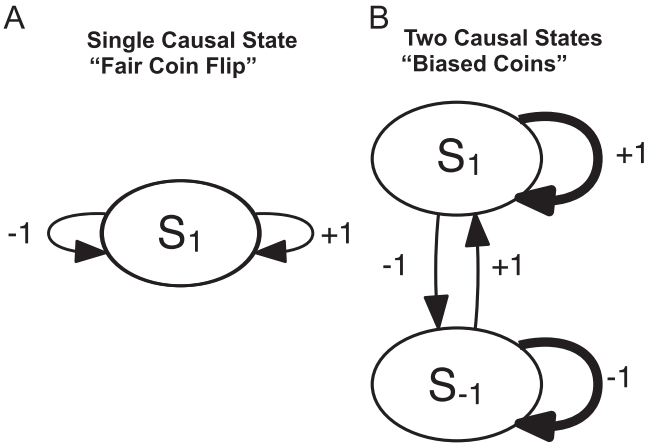

Fig. 10. *Causal state example.* (A) We show a simple model with a single causal state generating sequences of $+1, -1$ with each term generated independently and each symbol equally likely. (B) We demonstrate a simple model with two causal states. The causal state s_1 is more likely (thicker arrow) to emit a $+1$ and remain in the same state than to emit a -1 and transition to state s_{-1}.

Second, in comparison with the neutral model, words with a single causal state obeyed biased behavior. That is, the transition probability $+1/-1$ deviated from 0.50 more in the observed (4.1%) than neutral case (3.8%), a small but reliable difference, Welch's $t(11,982.13)=5.04, p<10^{-5}$. These subtly different means are primarily due to many random and observed words essentially approximating a coin flip. However, the distributions between the observed and neutral case are different as well, with more bias seen in the observed case – higher divergences from the

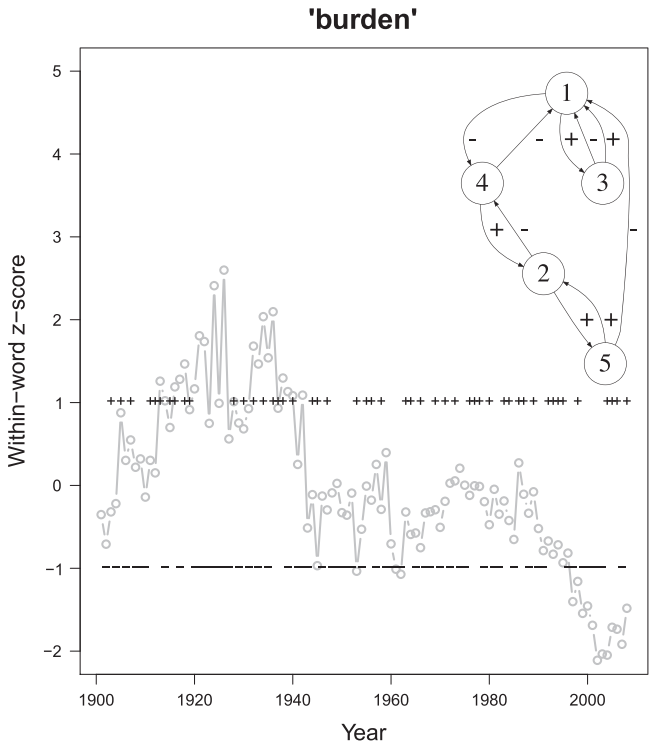

Fig. 11. *Causal state analysis*: *"burden"*. Words which undergo complex behavior in the patterns of increasing and decreasing (main figure) may be characterized by more complicated causal state models (inset). In this example, the word "burden" is predicted to have 5 causal states.

coin flip, even with only a single causal state, K–S test $D = .08, p < 10^{-10}$. This indicates that even for words whose frequency change dynamics are consistent with a single causal state, they were likely to be either increasing or decreasing during the entire duration. Since we have focused on stable variants, one possibility for a word decreasing in frequency is that it is on its way to being lost. This is perhaps consistent with previous analyses of word frequency dynamics which indicate patterns of novelty (birth) and loss (death) of words are significant contributors to the statistical properties of overall word frequency dynamics (Petersen et al., 2012a).

4. General discussion

We note that in biological evolution there are several forces which are commonly responsible for departures from neutrality: drift, gene flow, mutations and selection (Gillespie, 2004). We have showcased signatures in observed linguistic data that signal these processes. By comparing to a stochastic neutral model of evolution we have attempted to account for fluctuations in frequencies that could be attributed to drift. Through focusing only on stable variants and re-normalizing the data, we hoped to mediate the influences of gene flow and mutations. Thus, after attempting to account for these other forces impacting word frequencies we are left to conclude that the "non-neutral behavior" we observe in word frequencies is consistent with a biased selective process.

As we described in the introduction, the mechanism underlying such cultural fluctuation is still under debate (Acerbi and Bentley, 2014). One position we could take on the PCA results is that they are consistent with the idea that cultural forces – such as war – shape the way language is used. In this case, it is not merely the choices of individuals aggregating at the population level, but also constraints present already in the cultural environment that render particular words more or less useful under relevant conditions. The procedures described in this paper may offer a means of identifying these "selective sweeps" in historical linguistic data. We would agree with Acerbi and Bentley (2014) that this cannot account for all of the data. The general "kick" that seems to be present for various words from year to year – expressed quantitatively in kurtosis – may reflect a kind of biased copying at the individual level that is amplified in a socially interacting community. One candidate for the form for bias at the cultural level is suggested by evolutionary biology; Turelli and Barton (1994) note that disruptive selection – a type of natural selection that favors extreme over intermediate values of a trait – can also generate high kurtosis. At present, we cannot identify with any certainty the mechanism of these effects. However the development of further "big data" approaches in cultural data (e.g., Acerbi and Bentley, 2014) may provide incremental operationalization of them, such as our PCA analysis here.

Besides this inevitable difficulty in isolating selective mechanism, there are other limitations to the current study that future work may explore. As described at the outset of this paper, we ignore "turnover" rate, and thus neglect vocabulary loss or, equally important, the introduction of new variants into the vocabulary. It is certainly possible that the frequency of words we observe in our stable set will be influenced by words which occur during some subset of the years in our study. Nevertheless, the terms we explore here are among the most frequent words in the English language, occurring consistently for over 100 years. The fact that even this core vocabulary shows an ebb and flow that may be attributed to amplification through cultural selection is perhaps intuitive but, we would argue, nevertheless surprising given the stability of these terms. The most intuitive account of change for such variants would have been a neutral model, since one might expect these stable variants to be fixed and subject only to subtle frequency-dependent copying.

Our neutral model could be enhanced, and indeed a more complex modeling framework itself may be a useful addition to this kind of corpus analysis. A modeling framework could help isolate the role of selection, and help us to investigate the role of different mechanisms in generating the patterns of results we observe in the historical data. For example, by exploring various social factors, Nunn et al. (2009) use agent-based modeling along with general-linear models to demonstrate the power of selection, but also the potential role of different social variables, such as prestige or consensus. This may be especially valuable in view of these large historical data, because they present quantifiable "gold standard" for what is observed in language change patterns. Agent-based results may be compared to them directly. This is, of course, a strategy that has been used elsewhere, such as in English verb change (Hare and Elman, 1995), and is represented in wider literature on language evolution (e.g., Hurford et al., 1998).

An additional limitation to our present study is that some words are subject to sudden "pulses" or changes that take place in just one year, and this can, by itself, increase *K*. *K* may therefore be useful for identifying potential divergences, but perhaps cannot be used alone to determine whether sweeps are taking place in some coherent way. We would propose that coupling the *K* statistic with PCA would indicate that these pulses are not simply noise in any observed data. The two together would suggest divergence from neutrality in a manner that suggests lower-dimensional – "culturally functional" – patterns of variation. Further, it is possible that a linguistic analog of "genetic hitchhiking" is occurring whereby frequencies words will fluctuate not through direct selection but their association with words under selection (Barton, 2000). Coupling our statistical approaches with the study of bi-grams, also available through Google Ngram, may help to disambiguate these processes.

Though we have related our analysis of stable variants with that of a "core genome", we note that there are other linguistic data sets which may be appropriate to consider. In 1955, Swadesh established

a list of words deemed common across all human languages (Swadesh, 1955) that has since been revised several times. We compared our list of stable variants with the 100 Swadesh words.[4] We found that 96 of the Swadesh words occurred in our list of stable variants (the 4 missing words were louse, fingernail, belly and liver). While we might expect that, given their essential nature, words on the Swadesh list would be characterized by a single unbiased causal state. However, we found about an equivalent fraction of the Swadesh words had more than 1 causal state (8/96).

Another interesting potential extension is to explore words that are *more* stable than what would be anticipated by chance. While we have focused on change and explored potential sources of selection – like major cultural events – it may also be linguistically interesting to quantify more or less stability (resistance to change) relative to our various baselines. We leave this to future exploration, but note that the "skeletal" features of language in the form of grammatical word classes may show this pattern, and exhibit slower change, though there is ongoing discussion about the quantification of this idea (e.g., Greenhill et al., 2010).

Finally, we have attempted to draw commonalities between cultural and biological evolution. We note that rather than simply an application of insights from biological evolution, cultural evolution offers many opportunities for discovering new and relevant dynamics that are currently intractable in the realm of traditional evolutionary biology. Due to social media, Google, and other online services, we are now in the presence of a multitude of massive cultural data sets. In contrast to the traditional biological realm, many of these cultural data sets, such as the one we examine, demonstrate highly frequent observations over a lengthy temporal scale. This offers us the ability to examine time-varying selective processes and directly consider flow of cultural units between historically disparate populations. This may also provide a "playground," as one might call it, for developing further tests of selection that may be newly applicable in the biological case. As such, we believe that the cultural domain represents an important source and opportunity for quantitative modeling and analysis in ways that will likely complement and spur new ideas in other scientific disciplines (e.g., Bentley et al., 2011).

5. Conclusion

In our analysis, we have examined a large set of curated word frequency data from Google Ngram representing words used in English books over the past 100 years. To target our analysis, we focused on "stable variants" – commonly used words throughout the past 100 years. Surprisingly, we observed that even these stable parts of the English language undergo substantial deviations under what we expect from a neutral model.

Acknowledgments

This project was supported as part of NSF INSPIRE Track 1, BCS-1344279 to both authors of this paper.

References

Acerbi, A., Bentley, R.A., 2014. Biases in cultural transmission shape the turnover of popular traits. Evol. Hum. Behav. 35 (3), 228–236.
Acerbi, A., Mesoudi, A., 2015. If we are all cultural darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. Biol. Philos. 30 (4), 481–503.
Anthony, D.W., 2010. The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton University Press, Princeton, New Jersey.
Barton, N.H., 2000. Genetic hitchhiking. Philos. Trans. R. Soc. B: Biol. Sci. 355 (1403), 1553–1562.
Bentley, R.A., 2008. Random drift versus selection in academic vocabulary: an evolutionary analysis of published keywords. PLoS One 3 (8), e3057.
Bentley, R.A., Earls, M., O'Brien, M.J., 2011. I'll have what she's having: mapping social behavior, MIT Press.
Bentley, R.A., Hahn, M.W., Shennan, S.J., 2004. Random drift and culture change. Proc. R. Soc. Lond. Ser. B: Biol. Sci. 271 (1547), 1443–1450.
Blythe, R.A., 2012. Neutral evolution: a null model for language dynamics. Adv. Complex Syst. 15 (03–04).
Boschetti, F., 2008. Mapping the complexity of ecological models. Ecol. Complex. 5 (1), 37–47.
Boyd, R., Richerson, P.J., 1988. Culture and the Evolutionary Process. University of Chicago Press, Chicago and London.
Christiansen, M.H., Kirby, S., 2003. Language evolution: the hardest problem in science?. Stud. Evol. Lang. 3, 1–15.
Crow, J.F., Kimura, M., 1970. An introduction to population genetics theory.
Crutchfield, J.P., 1994. The calculi of emergence: computation, dynamics and induction. Phys. D: Nonlinear Phenom. 75 (1), 11–54.
Dale, R., Lupyan, G., 2012. Understanding the origins of morphological diversity: the linguistic niche hypothesis. Adv. Complex Syst. 15 (03n04), 1150017.
Dale, R., Vinson, D.W., 2013. The observer's observer's paradox. J. Exp. Theor. Artif. Intell. 25 (3), 303–322.
D'Andrade, R., 2002. Cultural darwinism and language. Am. Anthropol. 104 (1), 223–232.
Darwin, C., 1859. On the origin of the species by natural selection.
Daubin, V., Gouy, M., Perriere, G., 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12 (7), 1080–1090.
Ewens, W.J., 2004. Mathematical Population Genetics 1: I. Theoretical Introduction, vol. 27. Springer Science & Business Media, New York.
Gillespie, J.H., 2004. Population Genetics: A Concise Guide. JHU Press, Baltimore and London.
Greenhill, S.J., Atkinson, Q.D., Meade, A., Gray, R.D., 2010. The shape and tempo of language evolution. Proc. R. Soc. Lond. B: Biol. Sci. 277 (1693), 2443–2450.
Gureckis, T.M., Goldstone, R.L., 2009. How you named your child: understanding the relationship between individual decision making and collective outcomes. Top. Cogn. Sci. 1 (4), 651–674.
Hahn, M.W., Bentley, R.A., 2003. Drift as a mechanism for cultural change: an example from baby names. Proc. R. Soc. Lond. B: Biol. Sci. 270 (Suppl 1), S120–S123.
Hare, M., Elman, J.L., 1995. Learning and morphological change. Cognition 56 (1), 61–98.
Hsiang, T., Baillie, D.L., 2005. Comparison of the yeast proteome to other fungal genomes to find core fungal genes. J. Mol. Evol. 60 (4), 475–483.
Hurford, J.R., Studdert-Kennedy, M., Knight, C., 1998. Approaches to the Evolution of Language: Social and Cognitive Bases. Cambridge University Press, Cambridge.
Kandler, A., Shennan, S., 2013. A non-equilibrium neutral model for analysing cultural change. J. Theor. Biol. 330, 18–25.
Kelly, D., Dillingham, M., Hudson, A., Wiesner, K., 2012. A new method for inferring hidden Markov models from noisy time sequences. PLoS One 7 (e29703), 1.
Kimura, M., 1985. The Neutral Theory of Molecular Evolution. Cambridge University Press, New York.
Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., Nowak, M.A., 2007. Quantifying the evolutionary dynamics of language. Nature 449 (7163), 713–716.
Linnen, C.R., Hoekstra, H.E., 2009. Measuring natural selection on genotypes and phenotypes in the wild. In: Cold Spring Harbor Symposia on Quantitative Biology, vol. 74. Cold Spring Harbor Laboratory Press, pp. 155–168.
Maynard, J., Haigh, J., 2007. The hitch-hiking effect of a favourable gene. Genet. Res. 89 (5–6), 391–403.
Mesoudi, A., 2007. Biological and cultural evolution: similar but different. Biol. Theory 2 (2), 119.
Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182.
Neiman, F.D., 1995. Stylistic variation in evolutionary perspective: inferences from decorative diversity and interassemblage distance in illinois woodland ceramic assemblages. Am. Antiq., 7–36.
Nunn, C.L., Thrall, P.H., Bartz, K., Dasgupta, T., Boesch, C., 2009. Do transmission mechanisms or social systems drive cultural dynamics in socially structured populations? Anim. Behav. 77 (6), 1515–1524.
Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11), 559–572.
Petersen, A.M., Tenenbaum, J., Havlin, S., Stanley, H.E., 2012a. Statistical laws governing fluctuations in word use from word birth to word death. Sci. Rep. 2.
Petersen, A.M., Tenenbaum, J.N., Havlin, S., Stanley, H.E., Perc, M., 2012b. Languages cool as they expand: allometric scaling and the decreasing need for new words. Sci. Rep. 2.
Pinker, S., Bloom, P., 1990. Natural language and natural selection. Behav. Brain Sci. 13 (04), 707–727.
Rachev, S.T., Kim, Y.S., Bianchi, M.L., Fabozzi, F.J., 2011. Financial Models with Lévy Processes and Volatility Clustering. vol. 187. John Wiley & Sons, Hoboken, New Jersey.

[4] http://ielex.mpi.nl./wordlist/Swadesh100/

Reali, F., Griffiths, T.L., 2009. Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. Proc. R. Soc. Lond. B: Biol. Sci., rspb20091513.
Richter, D.J., King, N., 2013. The genomic and cellular foundations of animal origins. Ann. Rev. Genet. 47, 509–537.
Rogers, D.S., Ehrlich, P.R., 2008. Natural selection and cultural rates of change. Proc. Natl. Acad. Sci. 105 (9), 3416–3420.
Schleicher, A., 1869. Darwinism Tested by the Science of Language, vol. 41. London.
Shalizi, C.R., Shalizi, K.L., 2004. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 504–511.
Steele, J., Glatz, C., Kandler, A., 2010. Ceramic diversity, random copying, and tests for selectivity in ceramic production. J. Archaeol. Sci. 37 (6), 1348–1358.
Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. Int. J. Am. Linguist., 121–137.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al., 2003. The cog database: an updated version includes eukaryotes. BMC Bioinform. 4 (1), 41.
Turelli, M., Barton, N., 1994. Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? Genetics 138 (3), 913–941.
Waterhouse, R.M., 2015. A maturing understanding of the composition of the insect gene repertoire. Curr. Opin. Insect Sci. 7, 15–23.
Zhai, W., Nielsen, R., Slatkin, M., 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol. Biol. Evol. 26 (2), 273–283.
Zipf, G.K., 1949. Human behavior and the principle of least effort.