COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13215

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13215

Stochastic Time-Series Analyses Highlight the Day-To-Day Dynamics of Lexical Frequencies

Cameron Holdaway,[a] Steven T. Piantadosi[b]

[a]*Department of Psychology, University of California San Diego*
[b]*Department of Psychology, University of California Berkeley*

Received 15 November 2021; received in revised form 25 August 2022; accepted 9 October 2022

Abstract

Standard models in quantitative linguistics assume that word usage follows a fixed frequency distribution, often Zipf's law or a close relative. This view, however, does not capture the near daily variations in topics of conversation, nor the short-term dynamics of language change. In order to understand the dynamics of human language use, we present a corpus of daily word frequency variation scraped from online news sources every 20 min for more than 2 years. We construct a simple time-varying model with a latent state, which is observed via word frequency counts. We use Bayesian techniques to infer the parameters of this model for 20,000 words, allowing us to convert complex word-frequency trajectories into low-dimensional parameters in word usage. By analyzing the inferred parameters of this model, we quantify the relative mobility and drift of words on a day-to-day basis, while accounting for sampling error. We quantify this variation and show evidence against "rich-get-richer" models of word use, which have been previously hypothesized to explain statistical patterns in language.

Keywords: Bayesian modeling; Corpus methods; Quantitative linguistics

1. Introduction

As languages change, words come and go according to external influence of culture, politics, and technology (Michel et al., 2011). These changes have been examined quantitatively on the timescale of decades (Lieberman, Michel, Jackson, Tang & Nowak, 2007), revealing the growth and death of words over time (Petersen, Tenenbaum, Havlin, & Stanley, 2012; Petersen, Tenenbaum, Havlin, Stanley, & Perc, 2012), and relationships between language use and external factors (Lau, Huang, Ferreira & Vul, 2019). Even on the timescale of days, word

Correspondence should be sent to Cameron Holdaway, Department of Psychology, University of California San Diego, La Jolla, CA 92093, USA. E-mail: choldawa@ucsd.edu

frequencies are *complex*. The use of words in, for instance, the news, will change according to many social, cultural, and political factors, as topics come and go in public discourse—as politicians win or lose elections, as fads come and go, or as interest in specific recent events waxes and wanes. With improvements in data collection and computational power, there has been an increase in corpus-based studies that study language change in large diachronic corpora with improved temporal sensitivity from a variety of sources, such as news and social media, that allow for more nuanced analyses of these complex dynamic processes (Davies, 2013, 2017; Grieve, Nini, & Guo, 2018). For example, Nini et al. applied growth modeling to emerging words in a corpus of North American tweets to show they follow a cubic s-curve (Grieve, Nini & Guo, 2017; Nini, Corradini, Guo, & Grieve, 2017), and other work has shown that the perseverance of new words can be predicted by a natural selection model of lexical evolution (Grieve, 2018). Altmann et al. show that the temporal distribution of word usage exhibits "bursts" and "lulls" that deviate from static frequency distributions (Altmann, Pierrehumbert, & Motter, 2009). Indeed, work suggests that factors beyond simple frequency counts, like local frequency changes (i.e., "momentum"), contribute to language selection in the choice of children's names (Gureckis & Goldstone, 2009). Other work has applied dynamical models to test hypotheses about cultural transmission beyond language, such as novelty biases in baby names (O'dwyer & Kandler, 2017), and random copying as an explanation for innovation and imitation in pop culture (Bentley, Lipo, Herzog & Hahn, 2007). Linguistic changes have also been analyzed through dynamical models with the goal of parsing the variability in word use over time, for example, by controlling for topical fluctuations in corpora (Karjus, Blythe, Kirby & Smith, 2020). These models tease apart random drift from meaningful sociocultural trends by grouping words into topics by weighted co-occurrence, and suggest that a significant portion of the change in word frequencies can be attributed to topical fluctuations.

Perhaps surprisingly, the complex dynamics of language usage give rise to stable statistical "laws" that are the object of study in statistical linguistics (Altmann & Gerlach, 2016; Baayen, 2001). For instance, word frequencies approximately follow a power-law distribution ("Zipf's law") (Piantadosi, 2014; Tsonis, Schultz & Tsonis, 1997; Zipf, 1936, 1949), the growth rate of tokens in a document follows a predictable scaling relationship ("Heaps' law") (Heaps 1978), and the size of a linguistic unit is systematically related to the size of its components ("Menzerath's law") (Altmann & Schwibbe, 1989; Milička, 2014). Many authors have attempted to show how such large-scale patterns can be explained by specific dynamical patterns—for instance, Simon (1955) showed how a stochastic process that preferentially reuses tokens can give rise to power laws, building on earlier work by Yule (1944) (see also Lü, Zhang, and Zhou, 2013; Mitzenmacher, 2004). Preferential reuse implements a "rich-get-richer" scheme where if a word is used frequently, it will tend to be reused even more frequently. Mak & Twitchell (2020) suggest there may be cognitive bases for the emergence of rich-get-richer dynamics showing that preferential attachment can arise in semantic networks where well-connected nodes are better at acquiring new links in paired association tasks. Importantly, however, these kinds of in-principle statistical models have been criticized for making unrealistic assumptions about the processes generating language—people do not choose words at

random—as well as not being empirically tested in any meaningful way (Herdan 1960; Howes 1968; Piantadosi 2014). Indeed, such models arguably target the wrong aspect of language: they derive a coarse pattern like Zipf's law with unrealistic and untested generative assumptions, instead of discovering and formalizing the real underlying dynamics of word usage.

Our goal here is to quantify patterns in real-world frequency use. Indeed, when detailed patterns of word usage are examined, it becomes clear that the overall distribution of frequencies (e.g., Zipf's law) is relatively uninteresting compared to the complex dynamical patterns which even common individual words follow. We use a novel statistical model that is able to infer and summarize dynamical properties of words; this model provides a rich picture of how frequencies in natural language behave, and a way to quantify the influence of specific dynamical properties, including growth/decay, random noise, and momentum. We use this model to answer basic questions about word dynamics: (1) do words that are used more frequently tend to grow in frequency ("rich get richer"), (2) do high-frequency words tend to be the ones that gain (or change) frequency, (3) are the observed changes in frequency well-characterized with random-walk models, and (4) do words really change substantially in frequency, or are apparent changes due to sampling error? For each of these questions, our analyses break the lexicon down by part of speech and word frequency, allowing us to examine whether different syntactic categories or word frequencies exhibit different behavior.

The model we develop is probabilistic and also flexible, meaning that its parameters can encode a variety of qualitatively different behaviors, including simple growth, decay, oscillations, self-reinforcing trends, random drift, and variations thereof as a function of part of speech or word frequency. The key intuition behind our model is that complex dynamics can often be summarized with a few key parameters. A close analog of our model is the dynamics of a physical spring attached to a weight with a damper. In this simple setup, if the spring is stretched some distance x, it experiences a restoring force proportional to a spring constant k times how far it has moved. It may also feel a damping effect c times its velocity which works against movement. Importantly, varying the parameters c and k through positive and negative values leads to very different dynamics, from oscillators that return to rest, to those which speed off to infinity with a small displacement. If one measured the precise dynamics of the spring and fit parameters, one could then discover the trends in the underlying dynamics of what might otherwise be complex-looking dynamics. Reducing the complexity of word dynamics to a few key parameters shares the spirit of dimensionality reduction techniques in machine learning (e.g., Maaten & Hinton, 2008; Tenenbaum, De Silva and Langford, 2000; Torgerson, 1952), but relies on a model-based framework. Specifically, we use recently tools for Bayesian inference to infer posterior distributions on the parameters of a stochastic dynamical system, and treat these parameters as the low-dimensional characterization each word's dynamics.

We apply this analysis to a large, novel collection of news stories gathered from the internet every 20 min for 876 days from November 2015 until May 2018. This period of time has seen significant social changes, including the 2016 presidential election, the #MeToo movement, and various global and national political changes, including the beginning of the Trump

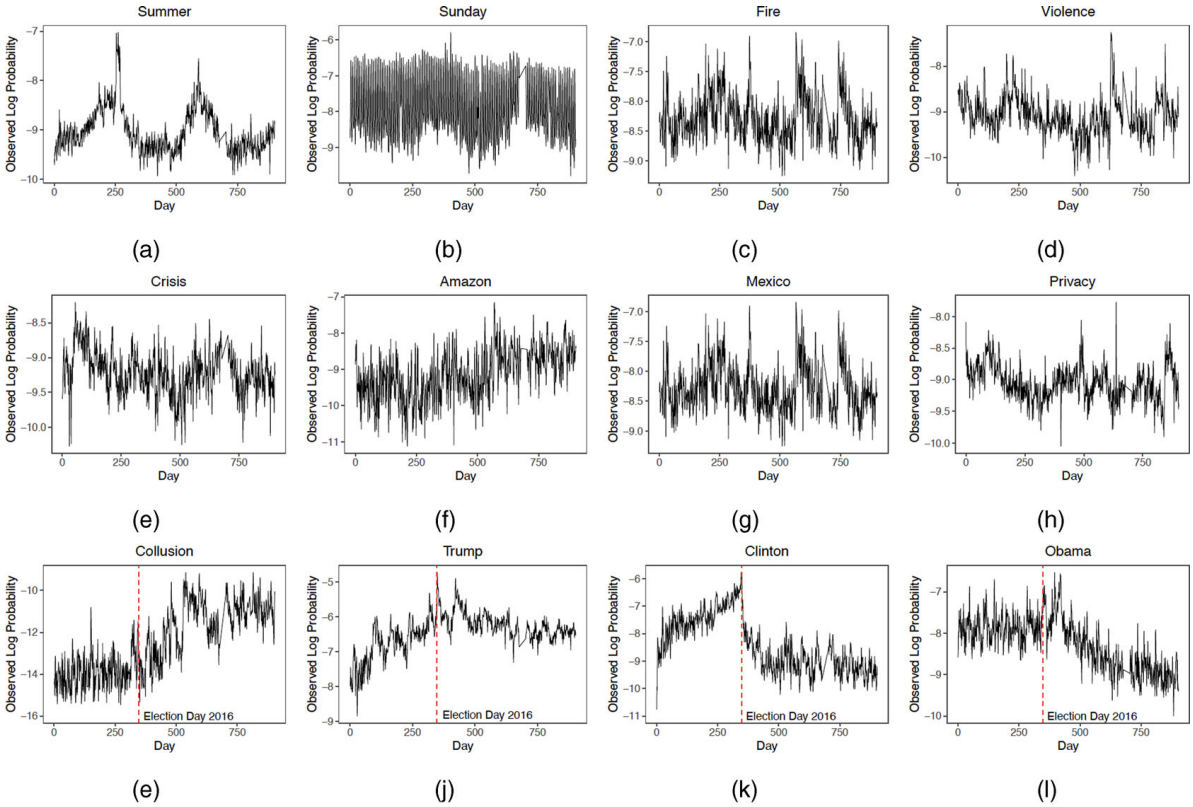

Fig. 1. Observed log frequency counts for 12 words over 876 days. Each word displays qualitatively different dynamics over the time period; some exhibit seasonal spikes on long (a) or short (b) timespans, while others exhibit fluctuations related to topically relevant world events (c–h). Figures (i– l) show politically related words as the time of collection centered around the 2016 Election (denoted by a red vertical line).

presidency. Our corpus is similar in scope to the English portion of the NOW corpus Davies (2017), except that it was gathered independently and is distributed freely. For analysis, we have selected the 20,000 most common words (data and analysis available on Github[1]). Our analysis method notably controls for the volatility in sampling words of different frequencies by inferring not just dynamical parameters, but each word's latent sequence of log probabilities. These latent log probabilities are not fit nor equated with relative token counts, but are inferred in a Bayesian setup where uncertainties in these frequencies are correctly propagated through to inferences on the dynamical parameters. We begin by presenting a qualitative picture of patterns in the word frequency corpus, and then describe the dynamical model and the parameters.

2. Descriptive analysis

Fig. 1 shows the frequency of 12 words over the time span of the corpus. In this figure, the x-axis is the day (starting November 2015) and the y-axis corresponds to the log frequency

of each word, normalized by the total number of word token counts on each day. Upon first inspection, it is clear that there are qualitatively very different trends exhibited across the words. Seasonal trends exist, such as those intuitively exhibited every summer for the word "summer" in Fig. 1a. A cyclical effect on much smaller timescales also exists, such as for days of the week (Fig. 1b). Other words, such as "fire" (Fig. 1c), "violence" (Fig. 1d), "crisis" (Fig. 1e), "Mexico" (Fig. 1g), and "privacy" (Fig. 1h), exhibit sporadic sharp peaks as they become topically relevant to the news cycle, but do not exhibit clear linear trends of increase or decrease over the period of observation. Still different are the dynamics of words like "Amazon" (Fig. 1f) and "collusion" (Fig. 1i), which have linearly increasing trends during the observed time period. These trends can be caused by individual events that spark the word into relevancy, or instead build gradually as they gain prominence. These different paths are shown in the qualitative difference between "Amazon," which does not appear to have a clearly defined event that led to this trend, and "collusion," which began to increase shortly after Election Day 2016 and seems to have stabilized around a new baseline mean frequency that is significantly greater than pre-election levels.

Furthermore, since the data collection window centered around the 2016 US Presidential Election, we can also observe substantially different trends for the words "Trump," "Clinton," and "Obama" before and after the election. In Fig. 1j, the trend for the word "Trump" is shown. The observed log frequency increases up to Election Day 2016, then levels off. An opposite effect is seen with "Clinton," depicted in 1(k), where a precipitous drop in frequency follows the 2016 election. Fig. 1l depicts "Obama" which is steady up to Election Day, then decreases following Inauguration Day 2017. These trends intuitively follow the change in media coverage of these politicians as their careers began or ended.

One difficulty in interpreting these trends is that it is not obvious how to separate meaningful changes from noise. If we assume that the number of times a word is used each day is sampled from an underlying probability distribution over words, then just by chance, a word's observed day-to-day frequencies will vary with time. This makes it difficult to separate such sampling noise from meaningful linguistic changes, or indeed to ask whether some words are actually more volatile than others because such volatility will be dependent on frequency. We address these concerns with a model.

3. Model-based analysis

The central assumption of our model is that each word has a latent probability of use on each day and that this probability may vary in time. We assume that the latent probability evolves according to a drift model with some unknown bias, analog of restoring force, and momentum terms. Formally, we assume that each word has a baseline latent log probability μ which is adjusted via $A \cdot t$ for the t'th day, plus an adjustment l_t. We assume that the l_ts vary in time according to a mixture of drifting from previous values and being resampled from a baseline distribution centered on 0. These correspond, respectively, to slow changes in the topics under discussion in the corpus, or sudden changes due to topic changes that may be

driven by events in the world. Formally,

$$l_t \sim \begin{cases} \text{Normal}(B \cdot l_{t-1} + C \cdot (l_{t-1} - l_{t-2}), 1) & \text{with probability } logit^{-1}(\theta) \\ \text{Normal}(0, \sigma_{jump}) & \text{with probability } 1 - logit^{-1}(\theta). \end{cases} \quad (1)$$

The parameter θ quantifies the relative contribution of drift versus independent sampling. We assume $\theta \sim Normal(0, 3)$. Thus, with probability $logit^{-1}(\theta)$, the word will be determined by the previous days' l_t value via B and C. With probability $1 - logit^{-1}(\theta)$, the deviation l_t will be resampled from centered around 0, corresponding to a "jump" in log frequency over and above the mean frequency μ and linear trend A. An important difference between our analysis and prior analyses of word frequencies is that we treat the l_t as latent variables whose influence is only observed through their influence on frequency counts:

$$c_t \sim \text{Binomial}(exp(\mu + s \cdot l_t + A \cdot t/T),\ N_t) \quad (2)$$

where c_t is how many times the word has been observed on day t, N_t is the total number of words observed on day t, and $s \sim Exponential(1)$ gives the size (scale) of influence of the l_t on the outcome.[2] Here, t is a variable for time, and it is scaled by T, the total number of days, so that A corresponds to changes over the entire duration of the corpus, rather than each day. In this model, if a word happens not to be observed on a given day ($c_t = 0$), the inferred probability is not undefined (or negative infinity) and in fact is inferred from the zero count and the other parameters of the model. Thus, this model is a form of autoregressive hidden-Markov model which treats l_t as potentially autocorrelated deviation from a linear effect $\mu + A \cdot t/T$ on log probability.

By varying the parameters, this model is able to capture many qualitatively different kinds of behavior. To illustrate, it is easiest to first consider the case when $A = C = 0$ and $B \neq 0$. Then, when $B = 1$, this model acts as a simple random walk on log probabilities, where each l_t is centered on l_{t-1}. When $B < 1$, the value of l_t tends to be *less* than l_{t-1}, meaning that deviations from μ tend to be corrected down to zero. In this situation, there is a restoring force to push deviations l_t from μ toward zero, and thus word frequency toward a μ. When $B > 1$, l_t will tend to be *more* extreme than l_{t-1}, corresponding to rich-get-richer dynamics. For any of these, having $A < 0$ will tend to bias log probabilities day-to-day down, and $A > 0$ will tend to bias them up, independent of the previous day. The C term is a kind of "momentum" where when $C > 0$, increases in l_t will be followed with more increases; when $C < 0$, the model will act to "damp" deviations from μ caused by l_p. Overall, then, a rich-get-richer scheme predicts $B > 1$ (the next daily deviation tends to be greater than the last) and/or $C > 0$ (words that are increasing/decreasing tend to do so more in the future). Fig. 2 shows several qualitative patterns that the model can exhibit as A, B, and C are varied. Since these parameters can encode importantly different dynamics, and effectively reduce the dimensionality of the time series, inferring their values from data tests between different hypotheses about how words behave. Moreover, examination of how these parameters relate across words provides a rich space of hypotheses that allow us to determine if words of different frequencies are engaged in similar or different dynamics. We note that the present model differs from prior work like (Nini et al., 2017) in that it focuses on a stochastic model with very simple (linear) underlying

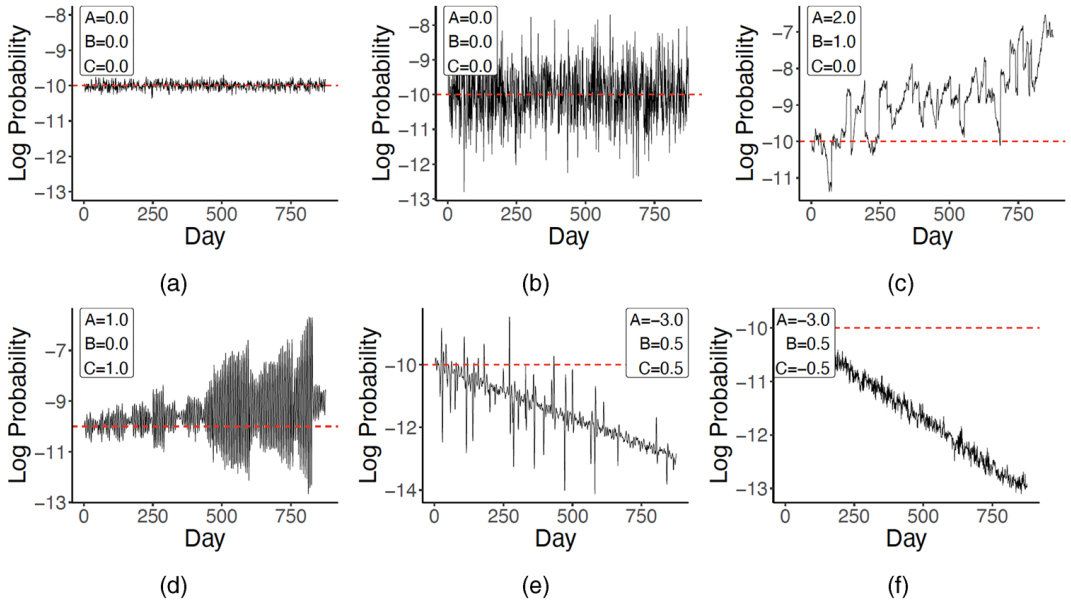

Fig. 2. Simulations of possible dynamics arising from different parameterizations of the model ($\mu = -10$, $\theta = 5$, $\sigma_{jump} = 0.5$, $s = 1$, unless otherwise stated). The "baseline" starting frequency is shown by the red line. (a-b) A pure, random noise model where each day is randomly sampled from a normal distribution centered at the baseline, with small variance $s = 0.1$ (a) and large varaince $s = 0.8$ (b). Two examples of rich-get-richer dynamics in (c) and (d) where recent deviations from baseline increase in subsequent days. And (e-f) decay models where daily variations decrease in conjunction with the overall frequency, where the parameter C determines the "momentum"; the extent to which changes are followed by more changes, or are damped.

dynamics, as opposed to growth curves. This simplification was made intentionally due to the difficulties of large-scale statistical inference (e.g., with hundreds of latent l_t for each word and 20,000 words), but an important direction for future work will be in connecting these modeling approaches by including nonlinear dynamics.

We applied this model to the data using an implementation in RStan Carpenter et al. (2017). Analyses were run on words tagged with part of speech (thus "run" the noun and "run" the verb were considered as different words) (see Methods). We collapsed tagger output to a simplified tag set consisting of nouns (NN), adjectives (JJ), verbs (VB), adverbs (RB), and other function words (FW).

4. Results

Fig. 3 shows each word's μ (baseline frequency) on the x-axis, and inferred model parameters on the y-axis. The levels (colors) represent the density of words in the given category. Subplots are divided by part of speech (columns) and key model parameters (rows). In general, due to the large sample size of 20,000 total words, we focus on the qualitative distribution of parameter values shown in these contours. We discuss each of the patterns these words exhibit in turn.

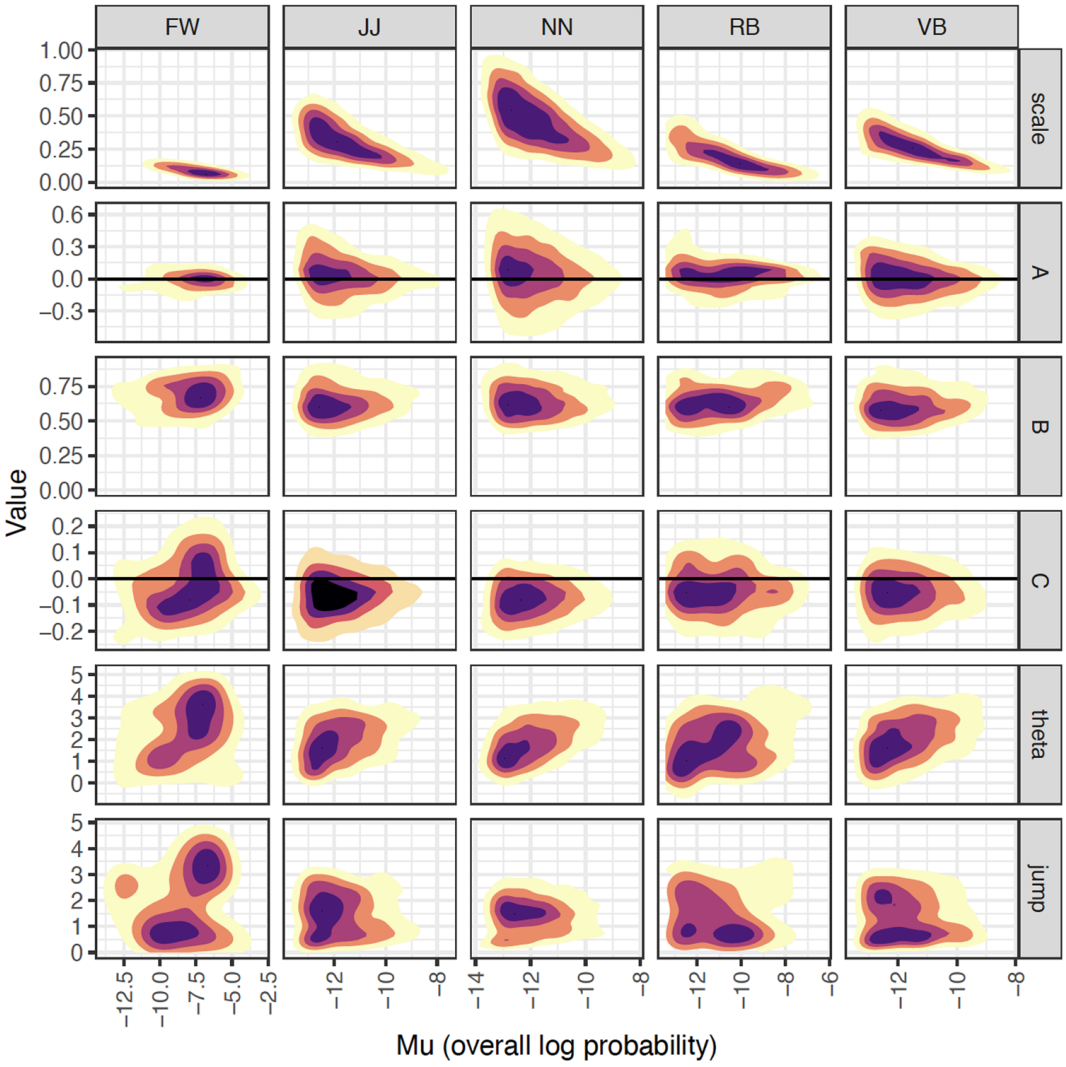

Fig. 3. Hexbin density plots showing the relationship between each key parameter and μ for words in a given part of speech. A reference line at $y = 0$ has been added for the A and C parameters.

4.1. *Rich words do not get richer*

Rich-get-richer models hypothesize that words that are high in frequency, or increasing in frequency will continue to do so. Our model provides evidence that this is *not* how word frequencies change on the timescale of our corpus. The main difference between our results and previous rich-get-richer analyses is that instead of arguing how theoretical distributions might arise from statistical processes (Yule 1944; Simon 1955), we instead parameterize the model is such a way that we can empirically test a variety of rich-get-richer hypotheses. Specifically, our model could potentially have discovered four different versions of rich-get-richer dynamics. If A were positively correlated with μ, then words with high frequency would tend to continue to grow in frequency since A represents the overall increase in

frequency. However, this is not the case: just in raw numbers, A is negatively correlated with μ ($r_\tau = -0.23$, $p < 0.001$). Second, if A tended to be positive for high-frequency words and negative elsewhere, that might support a version of rich-get-richer dynamics. This is shown not to be true, especially for the large open class categories NN and VB: as μ increases, the average A tends toward zero, which reflects no overall change. For the function words FW, A tends to be zero uniformly throughout the range of μ, which should be expected if these are a stable core of the linguistic system.

A third way that rich-get-richer dynamics could arise is if $B > 1$. In this case, daily deviations l_t would tend to be followed by *larger* daily deviations on l_{t+1}. However, for 99.6% of words, $B < 1$, indicating that the average daily deviation is smaller than the previous day, in the drift part of the model. This represents the opposite of rich-get-richer dynamics where deviations from μ tend to be squashed in the future.

Finally, if $C > 0$, then words that increased in frequency previously would tend to increase *more* on the next day. Our results show that C is typically negative, uniformly across word categories, or tending toward 0 for function words. This indicates that language tends to work *against* changes, so that increases in l_p one day will tend to be followed by decreases in l_p the next.

4.2. Words frequencies change, but not all change is random drift

Fig. 3 also shows that there are systematic patterns and trends in word frequencies, as shown by the many words with $A \neq 0$, $scale > 0$, and so on. Moreover, inference of these parameters is done on latent probabilities rather than the actual frequency counts, meaning that these patterns within words are not due to sampling error. Instead, these patterns reflect the best guess about the underlying patterns of change on the latent log probabilities.

Some derivations of Zipf's law derive it from random walks on logarithmic scales (e.g., Kawamura & Hatano, 2002), but our analyses show that there are likely factors at play beyond random walks. There are two random walk models that our results provide strong evidence against. If each word's daily frequency was independently sampled, then we would find $\theta \ll 0$ for all words; instead, we find $\theta > 0$. The distribution of θ values depicted in Fig. 3 shows that the majority of words have a θ value above 0, indicating that there exists meaningful autocorrelation between the l_t. A second simple random walk might be one where each word's value was a random adjustment to the previous day, corresponding to $\theta > 0$ and $A \approx 0, B \approx 1$, $C \approx 0$). Instead, we find that $B < 1$, meaning that each day's l_t tends to be closer to zero than the previous; and in fact, many words have nonzero A and nonzero C, meaning that we find evidence for other factors at play beyond simple random walk dynamics. However, it is important to note that the model is still stochastic—it is a random walk of some sort—but one that formalizes nontrivial influences on usage.

4.3. Part of speech differences in burstiness

While the general distribution of parameters is similar across parts of speech, there are some key differences that demonstrate that word dynamics are not the same for all types of words. Altmann et al. (2009) have previously shown the semantic type of a word can influence

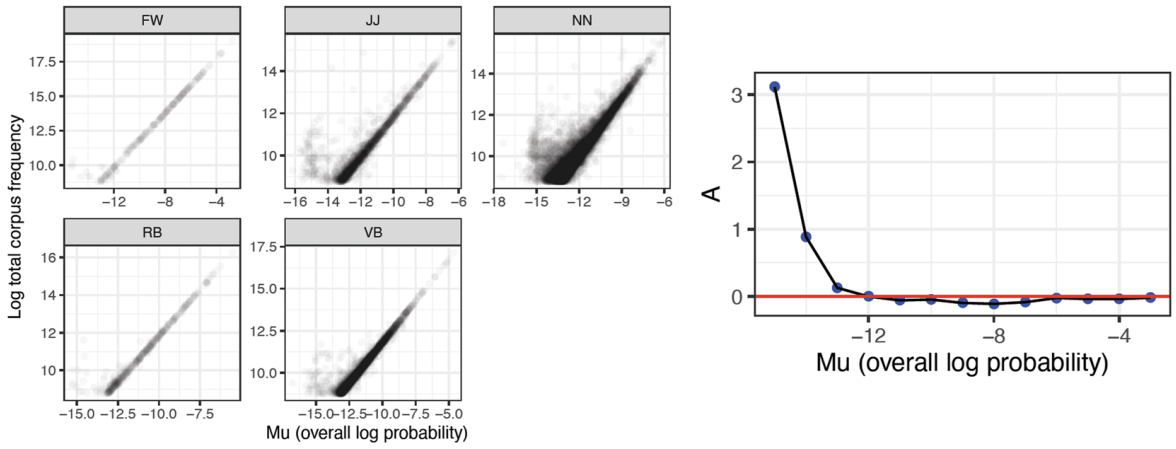

Fig. 4. Relationship between inferred probability μ and total corpus frequency count (left). Average A parameters for words as a function of overall log probability μ. Due to data sparsity, words with $\mu < -15$ have been adjusted to have $\mu = -15$ (right).

the extent to which it exhibits "bursty" recurrence patterns, but their model only looked at the dynamics with respect to the gaps between uses and not any other local frequency changes. In our model, the *scale* parameter, which scales the daily adjustments l_t, can be thought of as capturing the "burstiness" of the word since it quantifies how large of a change the l_t correspond to. This parameter has much higher variance for nouns than function words, which suggests the former are subject to much larger daily adjustments. This would be expected if nouns are the words that primarily vary with topic, and function words tend not to vary daily. Similarly, the A parameter has the largest range for nouns, followed by adjectives and verbs, compared to relatively little variation for function words, which are tightly packed around $A \approx 0$. These suggest that variability is driven by open-class, content words.

4.4. Low-frequency words experience "churn"

Fig. 4 (right) shows the average linear trend in log probability, A, as a function of μ in order to better visualize how frequency *typically* changes. Since A represents a scalar adjustment to the mean of the normal distribution from which a word's daily log frequency is sampled, positive values of A would indicate the word is increasing in frequency.[3] This figure shows that very low μ words tend to have $A > 0$ corresponding to increases in overall frequency. Words with μ between around -12 and -6 tend to have a *negative* A, while the highest μ words tend to have $A \approx 0$. This illustrates again that higher frequency words do not tend to increase in frequency. In fact, among the low-frequency words, the opposite is true: the lowest-frequency words tend to increase, and moderate frequency words tend to decrease. For example, some words that are infrequent (low μ), but are rising in frequency (high A) are: "bitcoin" ($\mu = -14.20$, $A = 3.45$), "timeline" (–14.07, 3.34), "bosnia" (–14.17, 3.03), and "meddling" (–14.02, 3.46), while words that are churning out of the lexicon (higher μ with negative A) are: "miles" ($\mu = -8.80$, $A = -0.40$), "syrian" (–8.97, –0.73), "turkey" (–8.84, –0.98), and "trial" (–8.67, –0.85). This trend may be related to the two "regimes" of

word frequency reported in Ferrer i Cancho and Solé (2001), and also may reflect a kind of regression to the mean for low frequencies, but note that the sampling errors are controlled to some extent because these parameters are estimated over and above l_t. This generally provides support for a kind of global stability in word distributions, despite documented variation: there is a stable, high-frequency core, and a collection of lower-frequency words that come in and out of use.

4.5. Overall probability and μ

Finally, Fig. 4 (left) shows the relationship between inferred μ for each word (x-axis) and the word's total frequency in the corpus (y-axis). The very high correlations here indicate that generally the model's inferences align with the overall frequency, especially for closed-class words like function words. However, the plot for nouns shows that there are a substantial proportion of low-frequency nouns which tend to have slightly different μ than would be expected from their overall frequency. This indicates that the total frequencies observed for these words in a corpus are subject to additional forces like those modeled here. In fact, one does not discover their "true" overall probability without modeling these additional factors, showing that studies of word frequency may need to model dynamical patterns in usage rather than rely on total frequency.

5. Discussion and conclusion

We have shown how a simple probabilistic model can capture some of the patterns governing day-to-day changes in lexical usage. This model was applied to a massive corpus of news articles collected over multiple years and sheds light on the dynamics giving rise to previously observed statistical patterns. By inferring the parameters of this model from data, we were able to test key assumptions of statistical language models. These results show that words vary in frequency, over and above sampling error, with measurable differences across parts of speech and frequency. Our results also show that previously proposed rich-get-richer dynamics do not match patterns in word usage. Instead, these results suggest that there is a relatively stable core of high-frequency words, with some nontrivial dynamics as moderate frequency words tend to become rare, and rare words tend to become frequent. Our results show typically negative momentum terms ($C < 0$) corresponding to word dynamics which do not self-reinforce, but push words back to an average baseline value, and $\theta > 0$ corresponding to strong day-to-day autocorrelation.

The corpus collected for this paper is a powerful new data set for future linguistic research. Its focus on news articles makes it especially informative about local and emerging social trends, and its high temporal resolution highlights temporary fluctuations. Indeed, when it is recognized that languages change on the order of days—if not minutes—it becomes clear that the laws of statistical linguistics are really statements about the average behavior of a very complex dynamical system. Fully understanding the dynamics of language use and the statistical laws that emerge from it will require formalizing and testing *causal* theories of the

underlying dynamics of human communication. Our model, for instance, is consistent with several different possible causal accounts of parameter change: for example, $A > 0$ could reflect changes in communicative need, changes in style, changes in sociolinguistic demographic factors, and so on. The model is agnostic about these causes, and in fact, its results motivate further work to search for and evaluate such factors.

Our general approach has illustrated that one first step in understanding complex dynamics like those found in word frequencies may be to parameterize a space of possible dynamical patterns, and infer the parameters of these dynamics from the data. If the model is appropriately constructed, existing theoretical proposals may map on to particular values of these parameters; inference of the parameters, therefore, effectively works to tests the theories. In this way, model-bases analyses function somewhat like dimensionality-reduction techniques, providing a simpler, low-dimensional characterization of what is sure to be a complex conjunction of social forces.

6. Methods

Data collection: The data were collected using a custom script implemented that scraped online articles every 20 min for 873 days over a 903 day span beginning on November 25, 2015. This script used the $HTML :: ContentExtractor$ Perl library in order to extract text from HTML news articles. Previously loaded URLs were stored in database in order to ensure that news stories were only added to the database once (note this has the possibility of adding the same story twice if it appears at different URLs). For simplicity, the extracted text was converted to lowercase and words were defined as between 1 and 12 consecutive letters which contained a vowel (aeiouy) and no numbers. Word tokens were then summed for each day and stored, along with the day's total token count. These are publicly available on the [ ]. Though our analysis focused on daily variations in unigram frequency, we have made available the full dataset, including bigrams, allowing further fine-grained analysis by other research groups.

This script was left to run on a Linux server for several years, with several interruptions for server maintenance, building maintenance, a large-scale DDOS attack, and network outages. This led to a few gaps (the longest of which was 26 days between day 677 and 703 – This gap is visible in the graphs in Fig. 1). These gaps do not substantially affect the data analysis here because the true log probability for each day is treated as a latent variable, and these gaps are simply days with zero counts, which have imputed probabilities.

Tagging: Tagged was done using the Natural Language Toolkit (NLTK) part-of-speech tagger in Python, but we collapsed to a simplified tag set consisting of nouns (NN), adjectives (JJ), verbs (VB), adverbs (RB), and other function words (FW) (including cardinal numbers, prepositions, particles, personal pronouns, modals, conjunctions, determiners, and predeterminers). Otherwise, tags with fewer than 10 word types were removed.

Statistical inference: We used Stan's implementation of the No-U-Turn Sampler (NUTS) (Carpenter et al. (2017)) which performs similarly to a traditional Hamiltonian Monte Carlo algorithm without user specified step sizes (Hoffman & Gelman, 2014). The Stan script is available on Github,[4] including R scripts for our analysis and visualization. To do this, the

corpus was separated into individual files for each word, and each was run separately using GNU parallel Tange (2011). Each word was run in two separate chains with random starting positions, for 10,000 samples. In the primary analyses, we included all words run regardless of convergence assessment because in general the words showed good convergence properties. For instance, the average (across words) median (within parameters, including l_t, for a single word) $\hat{R}$ (Brooks & Gelman, 1998) was 1.01. The average maximum $\hat{R}$ was 1.45, but this was mostly found in the l_t parameters. One challenge with this model was divergent transitions: we found 13,257 words ran with zero divergent transitions. The Supplementary Information shows that we find similar results when the measured words are restrained to only those with zero divergent transitions (Fig. S1) or very small $\hat{R}$ (Fig. S2).

Notes

1 https://github.com/choldawa/Word-Frequency
2 This parameterization is analogous to choosing Normal($B \cdot l_{t-1} + C \cdot (l_{t-1} - l_{t-2})$, s), but easier for statistical inference.
3 Note that the y scale on this plot must be interpreted with care because these are adjustments to log probability, and so the effect of A on nonlogged probability will depend on μ.
4 https://github.com/choldawa/Word-Frequency

References

Altmann, E. G., & Gerlach, M. (2016). Statistical laws in linguistics. In Degli Esposti, M., Altmann, E., Pachet, F. *Creativity and universality in language*. Lecture Notes in Morphogenesis. Springer, Cham. https://doi.org/10.1007/978-3-319-24403-7_2

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS ONE*, *4*(11), e7678.

Altmann, G., & Schwibbe, M. H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Georg Olms Verlag.

Baayen, R. (2001). *Word frequency distributions*, volume 1. Kluwer Academic Publishers.

Bentley, R. A., Lipo, C. P., Herzog, H. A., & Hahn, M. W. (2007). Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior*, *28*(3), 151–158.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Davies, M. (2013). Corpus of global web-based English: 1.9 Billion words from speakers in 20 countries.

Davies, M. (2017). The new 4.3 billion word now corpus, with 4–5 million words of data added every day. In *The 9th International Corpus Linguistics Conference*.

Ferrer i Cancho, R., & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), 165–173.

Grieve, J. (2018). Natural selection in the modern English lexicon. In *International Conference on Language Evolution* (pp. 153–157).

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online. *English Language & Linguistics*, *21*(1), 99–127.

Grieve, J., Nini, A., & Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics*, *46*(4), 293–319.

Gureckis, T. M., & Goldstone, R. L. (2009). How you named your child: Understanding the relationship between individual decision making and collective outcomes. *Topics in Cognitive Science*, *1*(4), 651–674.

Heaps, H. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc.

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Howes, D. (1968). Zipf's law and Miller's random-monkey model. *American Journal of Psychology*, *81*(2), 269–272.

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, *1*, 1–40.

Kawamura, K., & Hatano, N. (2002). Universality of Zipf's law. *Journal of the Physical Society of Japan*, *71*(5), 1211–1213.

Lau, S. H., Huang, Y., Ferreira, V. S., & Vul, E. (2019). Perceptual features predict word frequency asymmetry across modalities. *Attention, Perception, & Psychophysics*, *81*(4), 1076–1087.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713.

Lü, L., Zhang, Z.-K., & Zhou, T. (2013). Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. *Scientific Reports*, *3*(1), 1–7.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Mak, M. H., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin & Review*, *27*(5), 1059–1069.

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.

Milička, J. (2014). Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, *21*(2), 85–99.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*(2), 226–251.

Nini, A., Corradini, C., Guo, D., & Grieve, J. (2017). The application of growth curve modeling for the analysis of diachronic corpora. *Language Dynamics and Change*, *7*(1), 102–125.

O'dwyer, J. P., & Kandler, A. (2017). Inferring processes of cultural transmission: The critical role of rare variants in distinguishing neutrality from novelty biases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1735), 20160426.

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*.

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, *2*, 943.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425–440.

Tange, O. (2011). GNU parallel - The command-line power tool. *USENIX Magazine*, *36*(1), 42–47.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401–419.

Tsonis, A., Schultz, C., & Tsonis, P. (1997). Zipf's law and the structure and evolution of languages. *Complexity*, *2*(5), 12–13.
Yule, G. U. (1944). *The statistical study of literary vocabulary*. CUP Archive.
Zipf, G. (1936). *The psychobiology of language*. London: Routledge.
Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1: A version of Fig. 3 using only words with zero divergent transitions.

Figure S2:A version of Fig. 3 where maximum $\hat{R} <$ 1.1.