Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death

Alexander M. Petersen,[1,2] Joel Tenenbaum,[2] Shlomo Havlin,[3] and H. Eugene Stanley[2]

[1] *IMT Lucca Institute for Advanced Studies, Lucca 55100, Italy*

[2] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*

[3] *Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

(Dated: July 19, 2011)

How often a given word is used, relative to other words, can convey information about the word's linguistic utility. Using *Google* word data for 3 languages over the 209-year period 1800–2008, we found by analyzing word use an anomalous recent change in the birth and death rates of words, which indicates a shift towards increased levels of competition between words as a result of new standardization technology. We demonstrate unexpected analogies between the growth dynamics of word use and the growth dynamics of economic institutions. Our results support the intriguing concept that a language's lexicon is a generic arena for competition which evolves according to selection laws that are related to social, technological, and political trends. Specifically, the aggregate properties of language show pronounced differences during periods of world conflict, e.g. World War II.

In a monumental effort, *Google Inc.* has recently unveiled a database of words, in seven languages, after having scanned approximately 4% of the world's books [1]. The massive project [2] allows for a novel view into the growth dynamics of word use and the birth and death processes of words in accordance with evolutionary selection laws [3]. In this paper, we analyze the complex dynamics of the number of uses $u_i(t)$ of word i in year t, which we regard as a proxy for the linguistic utility of a given word.

Do words, in all their breadth and diversity, display any common patterns that are consistent with fundamental classes of competition dynamics? To address this question, we analyze the growth rates of word use, defined here as the relative change in $u_i(t)$ over the time interval $\Delta t \equiv 1$ year. We focus on the 209-year time period 1800 − 2008 for the English, Spanish, and Hebrew text corpuses, which together comprise over 1×10^7 distinct words. Since the

---

[1] Corresponding author: Alexander M. Petersen
E-mail: petersen.xander@gmail.com or alexander.petersen@imtlucca.it

number of books and the number of distinct words are both growing exponentially in time (see Supporting Online Material (SOM) Fig. S1), we define the relative word use $f_i(t)$ as the fraction of uses of word i out of all word uses in the same year, $f_i(t) \equiv u_i(t)/N_u(t)$, where $N_u(t) \equiv \sum_{i=1}^{N_w(t)} u_i(t)$ is the total number of indistinct word uses digitized from books printed in year t, and $N_w(t)$ is the total number of distinct words digitized from books printed in year t. Hence, we focus our analysis on the growth rate of relative word use,

$$r_i(t) \;\equiv\; \ln f_i(t+\Delta t) - \ln f_i(t) = \ln\left(\frac{f_i(t+\Delta t)}{f_i(t)}\right) . \tag{1}$$

The relative use of a word depends on the intrinsic grammatical utility of the word (related to the number of "proper" sentences that can be constructed using the word), the semantic utility of the word (related to the number of meanings a given word can convey), and the context of the word's use. Because of the comprehensive extent of the *Google* database, we are able for the first time to analyze the statistical properties of *all* words over several hundred years in an entire *corpus*, whereas previous studies were able only to analyze individual texts [4, 5], collections of topical texts [6], and a relatively small snapshot of the British corpus [7].

We distinguish words with equivalent meanings but with different spellings (e.g. color versus colour), since we view the competition between synonyms and misspelled counterparts in the linguistic arena as a key ingredient in complex evolutionary dynamics [3]. A prime example for the evolutionary arena governed by word utility is the case of irregular and regular verb use in English. By analyzing the regularization rate of irregular verbs through the history of the English language, Lieberman et al. [8] show that the irregular verbs that are used more frequently are less likely to be overcome by their regular verb counterparts. Specifically, they find that the irregular verb death rate scales as the inverse square root of the word's relative use.

We use $f_i(t)$ as a proxy for the "fitness" of a given word, which determines the survival capacity of the word in relation to its competitors. With the advent of spell-checkers in the digital era, the fitness of a "correctly" spelled word is now larger than the fitness of related "incorrectly" spelled words. But not only "defective" words can die, even significantly used words can die. Fig. 1 shows an example of three once-significant words, "Radiogram", "Roentgenogram" and "Xray," which competed in the linguistic arena for the majority share of nouns referring to what is now commonly known as an "Xray." The word"Roentgenogram" has since become extinct, even though it was the most common term for several decades in the 20th century.

Fig. 2 illustrates the current era of heightened word competition, demonstrated through an anomalous increase in the *death rate* of existing words and an anomalous decrease in the *birth rate* of new words. In the past 10–20 years, the total number of distinct words has significantly decreased, which we find is due largely to the extinction of both misspelled words and nonsensical print errors, and simultaneously, the decreased birth rate of new misspelled variations. This observation is consistent with both the decreasing marginal need for new words and also the broad adoption of automatic spell-checkers and corresponds to an increased efficiency in modern written language. Figs. S2 and S3 show that the birth rate is largely comprised of words with relatively large median f_c while the death rate is almost entirely comprised of words with relatively small median f_c. Sources of error in the reported birth and death rates could be explained by OCR (optical character recognition) errors in the digitization process, which could be responsible for a certain fraction of the misspelled words. Also, the digitization of many books in the computer era does not require OCR transfer, since the manuscripts are themselves digital, and so there may be a bias resulting from this recent paradigm shift. Nevertheless, many of the trends we observe are consistent with the trajectories that extend back several hundred years.

Complementary to the death of old words is the birth of new words, which are commonly associated with new social and technological trends. Such topical words in modern media can display long-term persistence patterns analogous to earthquake shocks [9], and can result in a new word having larger fitness than related "out-of-date" words (e.g. log vs. blog, memo vs. email). Here we show that a comparison of the growth dynamics between different languages can also illustrate the local cultural factors (e.g. national crises) that influence different regions of the world. Fig. S4 shows how international crisis can lead to globalization of language through common media attention. Notably, such global factors can perturb the participating languages (here considered as arenas for word competition), while minimally affecting the nonparticipating regions, e.g. the Spanish speaking countries during WWII, see Fig. S4(A). Furthermore, we note that the English corpus and the Spanish corpus are the collections of literature from several nations, whereas the Hebrew corpus is more localized.

Between birth and death, one contends with the interesting question of how the use of words evolve when they are "alive". We focus our efforts toward quantifying the relative change in word use over time, both over the word lifetime and throughout the course of history. In order to analyze separately these two time frames, we select two sets of words: (i) relatively new words with "birth year" $t_{0,i}$ later than 1800, so that the relative age $\tau \equiv t - t_{0,i}$ of word i is the number of years after the word's first occurrence in the database, and (ii) relatively common words, typically with $t_{0,i}$ prior to 1800. We analyze dataset #(i) words, summarized in Table S1, so that we can control for properties of the growth dynamics that are related to the various stages of a word's life trajectory (e.g. an "infant" phase, an "adolescent" phase, and a "mature" phase). For comparison, we also analyze dataset #(ii) words, summarized in Table S2, which are typically in a stable mature phase. We select the relatively common words using the criterion $\langle f_i \rangle \geq f_c$, where $\langle f_i \rangle$ is the average relative use of the word i over the word's lifetime T_i, and f_c is a cutoff threshold which we list in Table S2. In Table S3 we summarize the entire data for the 209-year period 1800–2008 for each of the four *Google* language sets analyzed.

Modern words typically are born in relation to technological or cultural events, such as "Antibiotics." We ask if there exists a characteristic time for a word's general acceptance. In order to search for patterns in the growth rates as a function of relative word age, for each new word i at its age τ, we analyze the "use trajectory" $f_i(\tau)$ and the "growth rate trajectory" $r_i(\tau)$. So that we may combine the individual trajectories of words of varying prevalence, we normalize each $f_i(\tau)$ by its average $\langle f_i \rangle = \sum_{\tau=1}^{T_i} f_i(\tau)/T_i$ over the word's entire lifetime, obtaining a normalized use trajectory $f'_i(\tau) \equiv f_i(\tau)/\langle f_i \rangle$. We perform the analogous normalization procedure for each $r_i(\tau)$, normalizing instead by the growth rate standard deviation $\sigma[r_i]$, so that $r'_i(\tau) \equiv r_i(\tau)/\sigma[r_i]$ (see SOM).

Since some words will die and other words will increase in use as a result of the standardization of language, we hypothesize that the average growth rate trajectory will show large fluctuations around the time scale for the transition of a word into regular use. In order to quantify this transition time scale, we create a subset $\{i\,|T_c\}$ of word trajectories i by combining words that meets an age criteria $T_i \geq T_c$. Thus, T_c is a threshold to distinguish words that were born in different historical eras and which have varying longevity. For the values $T_c = 25, 50, 100$, and 200 years, we select all words that have a lifetime longer than T_c and calculate the average and standard deviation for each set of growth rate trajectories as a function of word age τ. In Fig. 3 we plot $\sigma[r'_i(\tau|T_c)]$ which shows a broad peak around $\tau_c \approx$ 30–50 years for each T_c subset. Since we weight the average according to $\langle f_i \rangle$, the time scale τ_c is likely associated with the characteristic time for a new word to reach sufficiently wide acceptance that the word is included in a typical dictionary.

How much do the growth rates vary from word to word? The answer to this question can help distinguish between candidate models for the evolution of word utility. Hence, we analyze the probability density function (pdf) for the normalized growth rates $R \equiv r'_i(\tau)/\sigma[r'(\tau|T_c)]$ so that we can combine the growth rates of words of varying ages. The empirical pdf $P(R)$ shown in Fig. 4 is remarkably symmetric and is centered around $R \approx 0$, just as is found for the growth rates of institutions governed by economic forces [10–13]. Since the R values are normalized and detrended according to the age-dependent standard deviation $\sigma[r'(\tau|T_c)]$, the standard deviation by construction is $\sigma(R) = 1$.

A candidate model for the growth rates of word use is the Gibrat proportional growth process [12], which predicts a Gaussian distribution for $P(R)$. However, we observe the "tent-shaped" pdf $P(R)$ which is a double-exponential or Laplace distribution, defined as

$$P(R) \equiv \frac{1}{\sqrt{2}\sigma(R)} \exp[-\sqrt{2}|R - \langle R \rangle|/\sigma(R)] \ . \tag{2}$$

Here the average growth rate $\langle R \rangle$ has two properties: (a) $\langle R \rangle \approx 0$ and (b) $\langle R \rangle \ll \sigma(R)$. Property (a) arises from the fact that the growth rate of distinct words is quite small on the annual basis ($\gamma_w \approx 0.011$ shown in Fig. S1) and property (b) arises from the fact that R is defined in units of standard deviation. The Laplace distribution predicts a pronounced excess number of very large events compared to the standard Gaussian distribution. For example, comparing the likelihood of events above the 3σ event threshold, the Laplace distribution displays a five-fold excess in the probability $P(|R - \langle R \rangle| > 3\sigma)$, where $P(|R - \langle R \rangle| > 3\sigma) = \exp[-3\sqrt{2}] \approx 0.014$ for the Laplace distribution, whereas $P(|R - \langle R \rangle| > 3\sigma) = \text{Erfc}[3/\sqrt{2}] \approx 0.0027$ for the Gaussian distribution. The large R values correspond to periods of rapid growth and decline in the utility of words during the crucial "infant" and "adolescent" lifetime phases. In Fig. S5 we also show that the growth rate distribution $P(r')$ for the relatively common words comprising dataset #(ii) is also well-described by the Laplace distribution.

For hierarchical systems consisting of units each with complex internal structure [14] (e.g. a given country consists of industries, each of which consists of companies, each of which consists of internal subunits), a non-trivial scaling relation between the standard deviation of growth rates $\sigma(r|S)$ and the system size S has the form $\sigma(r|S) \sim S^{-\beta}$. The theoretical prediction in [14, 15] that $\beta \in [0, 1/2]$ has been verified for several economic systems, with empirical β values typically in the range $0.1 < \beta < 0.3$ [15, 16].

Since different words have varying lifetime trajectories as well as varying relative utilities, we now quantify how the standard deviation $\sigma(r|S_i)$ of growth rates r depends on the cumulative word frequency $S_i \equiv \sum_{t'=0}^{t} f_i(t)$. We observe a size-variance relation $\sigma(r|S_i) \sim S_i^{-\beta}$ for the growth rates of words with large S_i. The emergent scaling is surprising, given the fact that words do not have internal structure, yet still display the analogous growth patterns of larger economically-driven institutions that do have complex internal structure. To explain this within our framework of words as analogs of economic entities, we hypothesize that the analog to the subunits of word use are the books in which the word appears. Hence, S_i is proportional to the number of books in which word i appears. As a result, we find β values that are consistent with nontrivial correlations in word use between books. This phenomenon may be related to the fact that books are topical [6], and that book topics are correlated with cultural trends. Fig. S6B shows scaling behavior for large S_i, with $\beta \approx 0.10 - 0.21$ depending on the corpus, where a positive β value means that words with larger cumulative word frequency have smaller annual growth rate fluctuations.

Recent theoretical work [17] shows that there is a fundamental relation between the size-variance exponent β and the Hurst exponent H which quantifies the auto-correlations in a stochastic time series. The unexpected relation $\langle H \rangle = 1 - \beta > 1/2$ (corresponding to $\beta < 1/2$) indicates that the temporal long-term persistence, whereby on average large values are followed immediately by large values and smaller values followed by smaller values, can manifest in non-trivial β values (i.e. $\beta \neq 0$ and $\beta \neq 0.5$). Thus, the $f_i(\tau)$ of common words with large S_i display strong positive correlations and have β values that cannot be explained by a either a Gibrat proportional growth, which predicts $\beta = 0$, or a Yule-Simon Urn model, which predicts $\beta = 0.5$.

To test this connection between memory ($H \neq 1/2$) and size-variance scaling ($\beta < 1/2$), we calculate the Hurst exponent H_i for each time series belonging to the more relatively common words analyzed in dataset (ii) using detrended fluctuation analysis (DFA) [17, 18]. We plot the relative use time series $f_i(t)$ for the words "polyphony," "Americanism," "Repatriation," and "Antibiotics" in Fig. S7A, along with DFA curves (see SOM section) from which H is derived in Fig. S7B. The H_i values for these four words are all significantly greater than $H_r = 0.5$, which is the expected Hurst exponent for a stochastic time series with no temporal correlations. In Fig. S8 we plot the distribution of H_i values for the English fiction corpus and the Spanish corpus. Our results are consistent with the theoretical prediction $\langle H \rangle = 1 - \beta$ [17] relating the variance of growth rates to the underlying temporal correlations in each $f_i(t)$. This relation shows that the complex evolutionary dynamics we observe for words use growth is fundamentally related to the dynamics of cultural topic bursting [19, 20].

This study is motivated by analogies with other complex dynamic systems, such as the growth rates of economic institutions, e.g. companies [10–12] and countries [11, 13], and the growth rates of animal populations [21]. We find a striking analogy between the relative use of a word, which can quantitatively represent the intrinsic value of the word, and the value of a company (e.g. measured by its market capitalization or sales). This suggests a common underlying mechanism: just as firms compete for market share leading to business opportunities, and animals compete for food and shelter leading to reproduction opportunities, words are competing for use among the books that constitute a corpus.

Acknowledgements: We thank Will Brockman, Fabio Pammolli and Massimo Riccaboni for critical comments and insightful discussions and DTRA for financial support.

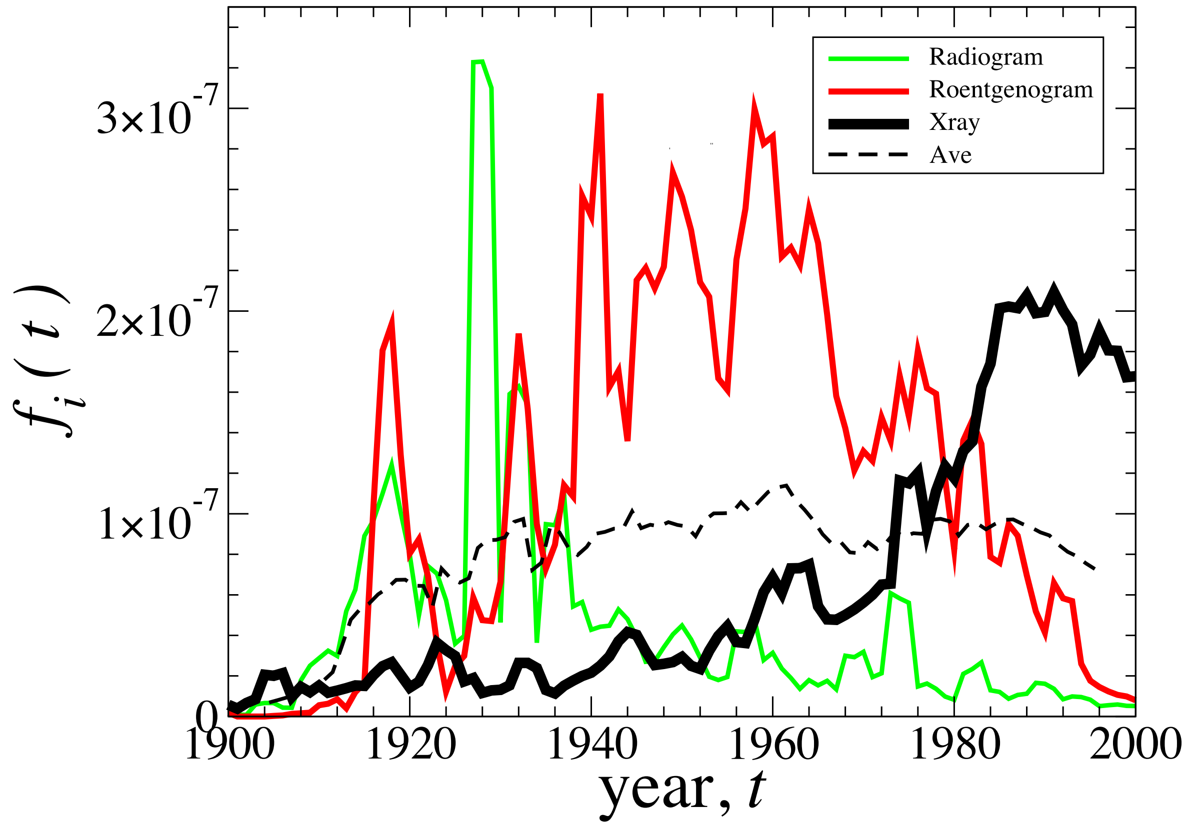

FIG. 1: The extinction of the English word "Roentgenogram" as a result of word competition with two competitors, "Xray" and "Radiogram." The average of the three $f_i(t)$ is relatively constant over the 80-year period 1920–2000, indicating that these 3 words were competing for limited linguistic "market share." We conjecture that the higher fitness of "Xray" is due to the efficiency arising from its shorter word length and also due to the fact that English has become the base language for scientific publication.

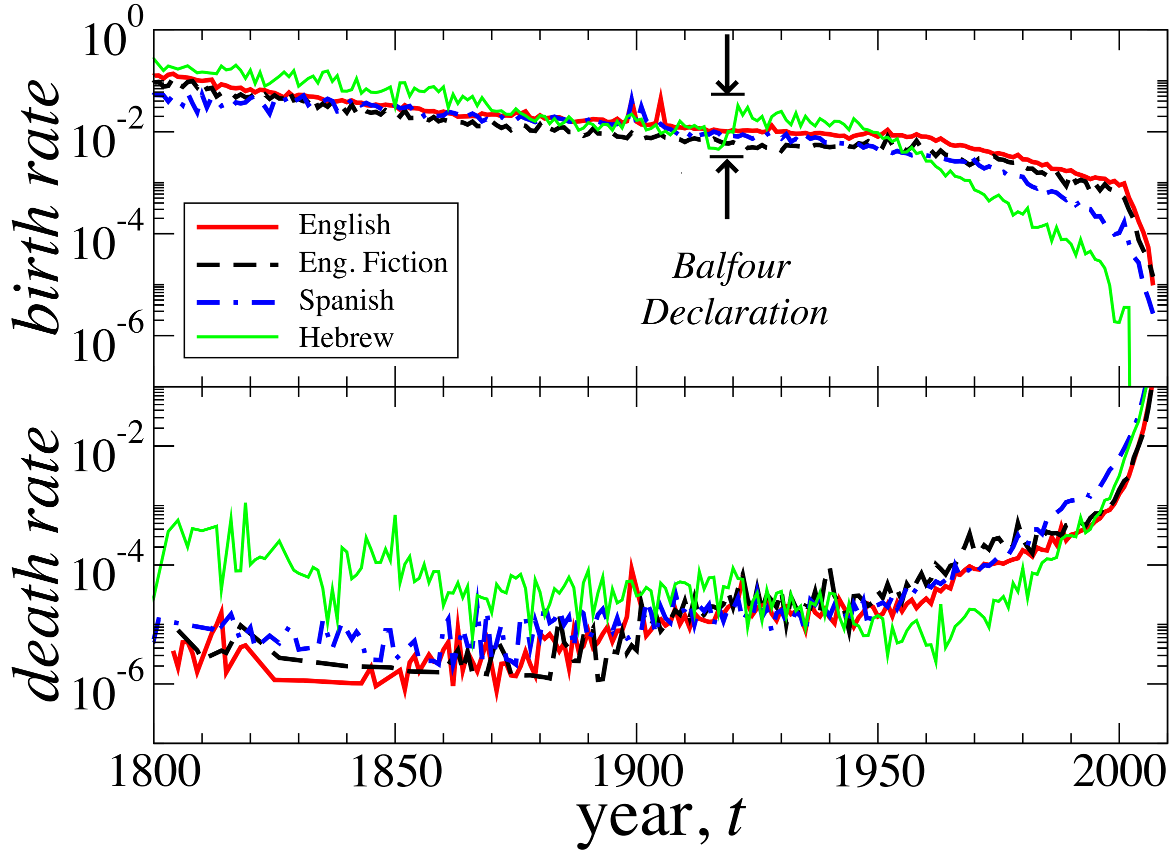

FIG. 2: The birth rate r_b and the death rate r_d of words demonstrate the inherent time dependence of the competition level between words in each of 4 corpora analyzed. The modern print era shows a marked increase in the death rate of words (e.g. low fitness, misspelled and outdated words). There is also a simultaneous decrease in the birth rate of new words, consistent with the decreasing marginal need for new words. This fact is also reflected by the sub-linear Heaps' law exponent $b \approx 0.5$ (see Fig. S13 and the SOM discussion). Note the impact of the Balfour Declaration in 1917, the circumstances surrounding which effectively rejuvenated Hebrew as a national language, resulting in a significant increase in the birth rate of Hebrew words.

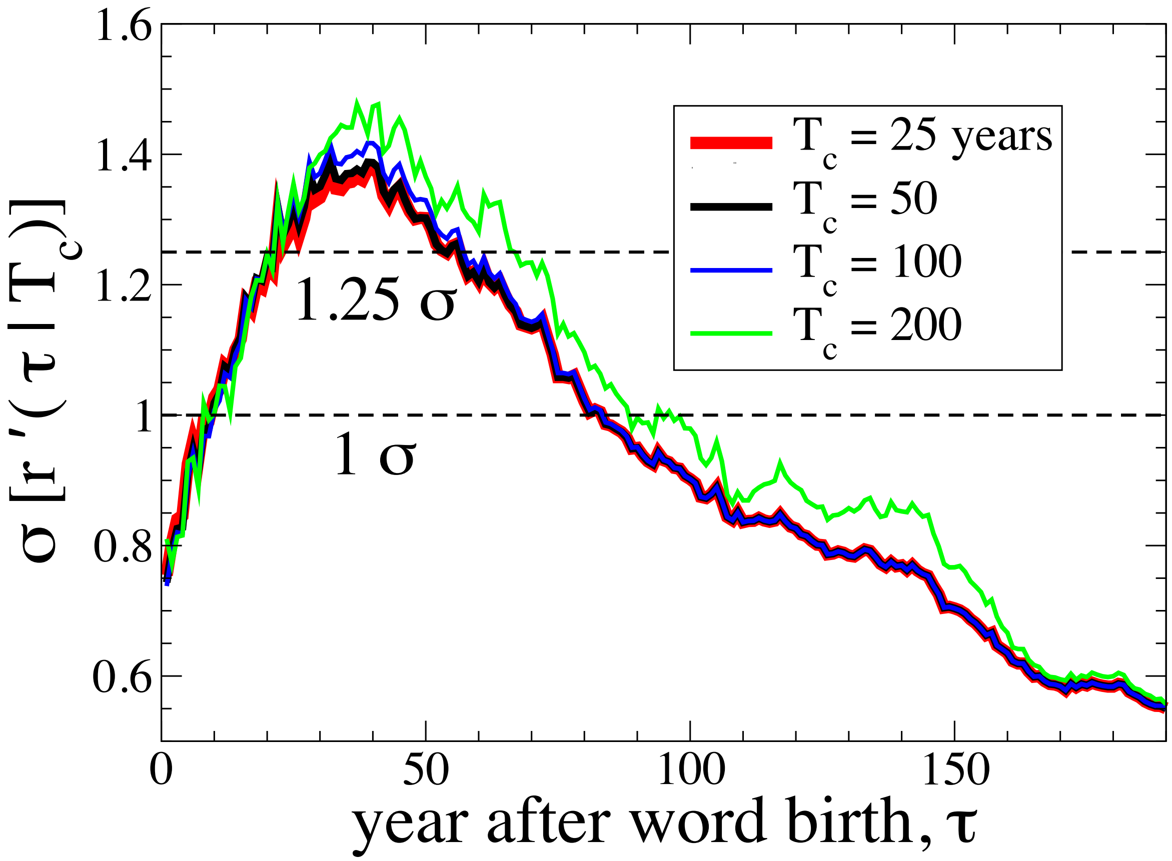

FIG. 3: Quantifying the tipping point for word use. The maximum in the standard deviation σ of growth rates during the "adolescent" period $\tau \approx 30$–50 indicates the characteristic time scale for words being incorporated into the standard lexicon, i.e. inclusion in popular dictionaries. In Fig. S9 we plot the average growth rate trajectory $\langle r'(\tau|T_c)\rangle$ which also shows relatively large positive growth rates during approximately the same period $\tau \approx 30$–50 years.

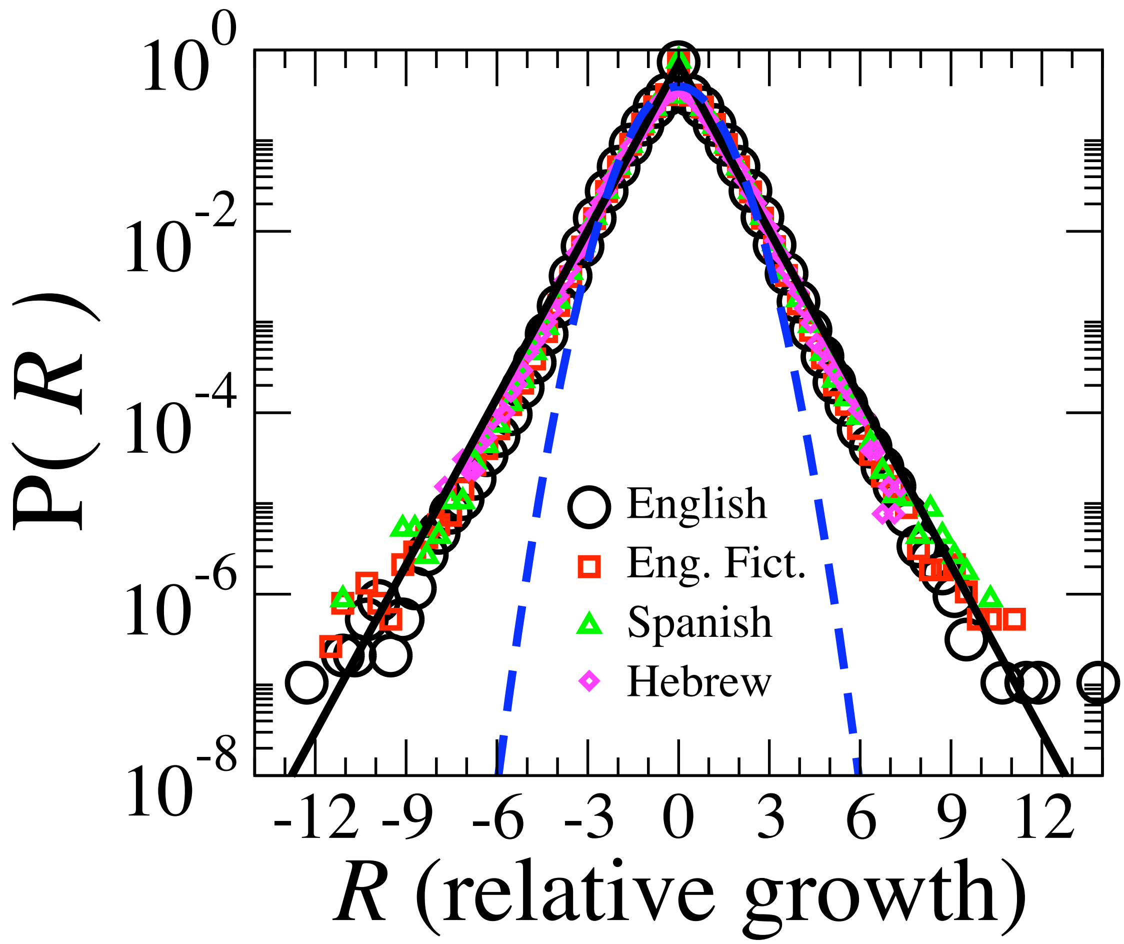

FIG. 4: Evidence for complex evolution of word use. We find Laplace distributions, defined in Eq. (2), for the annual word use growth rates for relatively new words, as well as for relatively common words (shown in Fig. S5) for English, Spanish and Hebrew. These Laplace distributions, which are symmetric and centered around $R \approx 0$, exhibit an excess number of large positive and negative values when compared with the Gaussian distribution. The Gaussian distribution is the predicted distribution for the Gibrat growth model, which is a candidate null-model for the growth dynamics of word use [11]. These large growth rates illustrate the possibility that words, many of which correspond to cultural and technological events, can have large variations even over the course of a year. For comparison, we plot a Gaussian distribution (dashed blue curve) with unit variance, which displays fast parabolic decay on the semi-logarithmic plot. The data agree remarkably well over the entire range $-12\sigma \leq R \leq 12\sigma$ with the Laplace distribution (solid black line). We analyze word use data over the time period 1800-2008 for new words i with lifetimes $T_i \geq T_c$, where in these panels we choose $T_c = 100$ years (see SOM methods section and Table S1 for a detailed description).

[1] M. Jean-Baptiste, et al., Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **331**, 176 (2011).

[2] *Google* n-gram project.
`http://ngrams.googlelabs.com/datasets`

[3] M. A. Nowak, *Evolutionary Dynamics: exploring the equations of life* (BelknapHarvard, Cambridge, MA 2006).

[4] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA 1949).

[5] A. A. Tsonis, C. Schultz, P. A. Tsonis, Zipf's law and the structure and evolution of languages. *Complexity* **3**, 12 (1997).

[6] M.Á. Serrano, A. Flammini, F. Menczer, Modeling Statistical Properties of Written Text. *PLoS ONE* **4(4)**, e5372 (2009).

[7] R. Ferrer i Cancho, R. V. Solé, Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics* **8**, 165 (2001).

[8] E. Lieberman, et al., Quantifying the evolutionary dynamics of language. *Nature* **449**, 713 (2007).

[9] P. Klimek, W. Bayer, S. Thurner, The blogosphere as an excitable social medium: Richter's and Omori's Law in media coverage. (under review) arXiv:1102.2091v1 [physics.soc-ph].

[10] L. A. N. Amaral, et al., Scaling Behavior in Economics: I. Empirical Results for Company Growth. *J. Phys. I France* **7**, 621 (1997).

[11] D. Fu., et al., The growth of business firms: Theoretical framework and empirical evidence.' *Proc. Natl. Acad. Sci.* **102**, 18801 (2005).

[12] M. H. R. Stanley, et al., Scaling behaviour in the growth of companies. *Nature* **379**, 804 (1996).

[13] D. Canning, et al., Scaling the volatility of gdp growth rates. *Economic Letters* **60**, 335 (1998).

[14] L. A. N. Amaral, et al., Power Law Scaling for a System of Interacting Units with Complex Internal Structure. *Phys. Rev. Lett.* **80**, 1385 (1998).

[15] M. Riccaboni, et al., The size variance relationship of business firm growth rates. *Proc. Natl. Acad. Sci.* **105**, 19595 (2008).

[16] V. Plerou, et al., Similarities between the growth dynamics of university research and of competitive economic activities. *Nature* **400**, 433 (1999).

[17] D. Rybski, et al., Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci. USA* **106**, 12640 (2009).

[18] C. K. Peng, et al., Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685 (1994).

[19] A. L. Barabási, The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207 (2005).

[20] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.* **105**, 15649 (2008).

[21] T. H. Keitt, H. E. Stanley, Dynamics of North American breeding bird populations *Nature* **393**, 257 (1998).

[22] R. Ferrer i Cancho, The variation of Zipf's law in human language. *Eur. Phys. J. B* **44**, 249 (2005).

[23] X. Gabaix, Zipf's law for cities: An explanation. *Quarterly Journal of Economics* **114**, 739 (1999).

[24] R. Ferrer i Cancho, R. V. Solé, Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **100**, 788 (2003).

[25] S. V. Buldryev, et al. Behavior in Economics: II. Modeling of Company Growth. *J. Phys. I France* **7**, 635 (1997).

[26] Y. Lee, et al., Universal Features in the Growth Dynamics of Complex Organizations. *Phys. Rev. Lett.* **81**, 3275 (1998).

[27] K. Matia, et al., Statistical properties of business firms structure and growth. *Europhys. Lett.* **67**, 498 (2004).

[28] S. V. Buldyrev, et al., The growth of business firms: Facts and theory. *J. Eur. Econ. Assoc.* **5**, 574 (2007).

[29] J. Growiec, et al., *Econ. Lett.* **98**, 207 (2008).

[30] B. Podobnik, et al., Common scaling behavior in finance and macroeconomics. *Eur. Phys. J. B* **76**, 487 (2010).

[31] Y. Liu, et al., The Statistical Properties of the Volatility of Price Fluctuations. *Phys. Rev. E* **60**, 1390 (1999).

[32] S. Picoli Jr., et al., Scaling behavior in the dynamics of citations to scientific journals. *Europhys. Lett.* **75**, 673 (2006).

[33] S. Picoli Jr., R. S. Mendes, Universal features in the growth dynamics of religious activities. *Phys. Rev. E* **77**, 036105 (2008).

[34] Y. Schwarzkopf, R. L. Axtell, J. D. Farmer, The cause of universality in growth fluctuations. *Arxiv e-print.* physics.soc-ph: 1004.5397. Under Review.

[35] B. Podobnik, et al., Size-dependent standard deviation for growth rates: Empirical results and theoretical modeling. *Phys. Rev. E* **77**, 056102 (2008).

[36] K. Hu, et al., Effect of Trends on Detrended Fluctuation Analysis. *Phys. Rev. E* **64**, 011114 (2001).

Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death

Alexander M. Petersen,[1,2] J. Tenenbaum,[2] S. Havlin,[3] H. Eugene Stanley[2]

[1]IMT Lucca Institute for Advanced Studies, Lucca 55100, Italy

[2]Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA

[3]Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

(2011)

Supporting Online Material

I. QUANTIFYING THE BIRTH RATE AND THE DEATH RATE OF WORDS

Just as a new species can be born into an environment, a word can emerge in a language. Evolutionary selection laws can apply pressure on the sustainability of new words since there are limited resources (here books) for the use of words. Along the same lines, old words can be driven to extinction when cultural and technological factors limit the use of a word, in analogy to the environmental factors that can limit the survival capacity of a species by altering the ability of the species to obtain food in order to survive and reproduce.

We define the birth year $y_{0,i}$ as the year t corresponding to the first instance of $f_i(t) \geq 0.05 f_i^m$, where f_i^m is median word use $f_i^m = Median\{u_i(t)\}$ of a given word over its recorded lifetime in the *Google* database. Similarly, we define the death year $y_{f,i}$ as the last year t during which the word use satisfies $f_i(t) \geq 0.05 f_i^m$. We use the relative word use threshold $0.05 f_i^m$ in order to avoid anomalies arising from extreme fluctuations in $f_i(t)$ over the lifetime of the word.

The significance of word births $\Delta_b(t)$ and word deaths $\Delta_d(t)$ for each year t is related to the size of a language. We define the birth rate r_b and death rate r_d by normalizing the number of births and deaths in a given year t to the total number of distinct words $N_w(t)$ recorded in the same year t, so that

$$r_b \equiv \Delta_b(t)/N_w(t) \ , \tag{S1}$$

$$r_d \equiv \Delta_d(t)/N_w(t) \ . \tag{S2}$$

This definition yields a proxy for the rate of emergence and disappearance of words with respect to their individual lifetime use. We restrict our analysis to words with lifetime $T_i \geq 2$ years and words with a year of first recorded use $t_{0,i}$ that satisfies the criteria $t_{0,i} \geq 1700$, which biases for relatively new words in the history of a language.

Fig. 2 is a log-linear plot of the relative birth and death rates for the 208-year period 1800–2007. The modern era of publishing, which is characterized by more strict editing procedures at publishing houses and very recently computerized word editing with spell-checking technology, shows a drastic increase in the death rate of words, along with a recent decrease in the birth rate of new words. This phenomenon reflects the decreasing marginal need for new words, consistent with the sub-linear Heaps law exponent $b \approx 0.5$ in Fig. S13.

II. ZIPF ANALYSIS OF WORD USE

The well known Zipf law quantifies the relative use of words, and has been verified by many analyses of single bodies of literature [4], online texts such as *Wikipedia* [6], and a large dataset of english texts [7]. Interesting deviations from the traditional Zipf scaling have been observed in select cases of human language, such as in the schizophrenic lexicon [22]. We find statistical regularity in the distribution of relative word use for four large datasets each comprising more than a half million distinct words taken from millions of books [1], suggesting that there are fundamental relations between the most common and the least common words in a given language.

Fig. S10(A) shows the probability density function (pdf) of the relative word use f_i and is characterized by a striking two-regime scaling [7],

$$P(f) \sim f^{-\alpha} = \begin{cases} P(f) \sim \; f^{-1.7} \quad , \quad f \lesssim f_c \;\; [\textit{“unlimited lexicon”}] \\ P(f) \sim \; f^{-2} \quad\;\; , \quad f \gtrsim f_c \;\; [\textit{“kernel lexicon”}] \; . \end{cases} \tag{S3}$$

The high-use regime $f > f_c$ is in agreement with the Zipf scaling law. These two regimes, coined the “kernel lexicon” and the “unlimited lexicon” by Ferrer i Cancho and Solé [7] are thought to be related to the cognitive constraints on the brain’s vocabulary. The Zipf scaling law quantifies the rank-frequency of the r-ranked word by the relation

$$f(r) \sim r^{-\zeta} \tag{S4}$$

with a scaling exponent $\zeta \approx 1$ [4]. Eq. (S4) has been shown to also quantify other socio-economic systems such as city size [23]. We find similar scaling for the pdf of word uses u_i in Fig. S11. Interestingly, it has been shown that The Zipf law emerges as the result of the “principle of least effort” which minimizes the communication noise between speakers (writers) and hearers (readers) [24].

The two scaling exponents α and ζ are related asymptotically by [7]

$$\alpha \approx 1 + 1/\zeta \; . \tag{S5}$$

In Fig. S12 we plot α values calculated for each year t using Hill’s maximum likelihood estimator (MLE). We characterize the two scaling regimes using the threshold $f_c = 10^{-5}$ so that the two regimes are: (A) $10^{-8} \leq f \leq 10^{-6}$ and (B) $10^{-4} \leq f \leq 10^{-1}$. For the “kernel lexicon” regime $f \gtrsim f_c$ we verify Zipf’s scaling law $\zeta \approx 1$ (corresponding to $\alpha \approx 2$) for the English corpus and the Spanish corpus for years $t > 1900$. For the “unlimited lexicon” regime $f \lesssim f_c$

we calculate $\zeta \approx 1.43$ corresponding to $\alpha \approx 1.7$. The Hebrew corpus, however, varies significantly, possibly as a result of the language's unique history and the revival of the language in the 19th century.

III. HEAPS LAW AND THE INCREASING MARGINAL UTILITY OF NEW WORDS

The utility of a new word is established as it becomes disseminated throughout the literature of a given language. As a proxy for this utility, one may study how often new words are invoked in lieu of preexisting competitors. Specifically, one can study the linguistic value of new words and ideas by analyzing the relation between N_u, the total number of words printed in a body of text, and N_w, the number of these which are distinct (i.e. the vocabulary size). The marginal utility of new words, $\partial N_u / \partial N_w$, therefore addresses the question: how much additional literature may follow from adding one more word to the vocabulary of a corpus?

For individual books, the relation between N_u and N_w empirically observed as the scaling relation

$$N_w \sim (N_u)^b \ , \tag{S6}$$

with $b < 1$, where Eq. (S6) is known as Heaps' law [6]. Using a stochastic model for the growth of book vocabulary size as a function of book size, Serrano et al. [6] proposed that $b = 1/\alpha$, where α is the scaling exponent in the pdf of relative word use as defined in Eq. (S3). Thus, the marginal utility of new words is directly related to the distribution of relative word use by

$$\frac{\partial N_u}{\partial N_w} \sim (N_w)^{\alpha-1} \ . \tag{S7}$$

To determine the marginal utility of words for an entire language, we test the relation in Fig. S13

$$N_w(t) \sim [N_u(t)]^b \approx [N_u(t)]^{1/\alpha} \ . \tag{S8}$$

Using linear regression of the variables plotted log-log axes, we find that $b \approx 0.5$ for each corpus, in good agreement with the $\alpha \approx 2$ values calculated in Fig. S12(B). As a result, we note that there is an *increasing marginal utility* for additional words since $\alpha > 1$. Alternatively, this relation indicates that there is a *decreasing marginal need* for additional words, since $\partial N_w / \partial N_u \sim (N_w)^{1-\alpha}$.

Since it is possible to estimate $\alpha = \alpha(t)$ on an annual basis using the pdf of relative word use, one can monitor the marginal utility of new words over time. Interestingly, Fig. S12(B) shows a marked decrease in α over the last 10 years of the study, indicating that, while new words are still being added, we do not obtain as much "mileage" out of new words as we obtained in the past. Consequently, we show that it is possible that the marginal utility of new words can provide insight into the underlying social and technological progress of a society.

IV. QUANTIFYING THE WORD USE TRAJECTORY

Next we ask how word use evolves through the various stages of its lifetime. Since words appear to compete for use in a word-space that is based on utility, we seek to quantify the average lifetime trajectory of word use. The lifetime trajectories of different words will vary, since each trajectory depends not only on the intrinsic utility of word i, but also on the "birth-year" $t_{0,i}$ of word i.

Here we define the age or trajectory year $\tau = t - t_{0,i}$ as the number of years after the word's first appearance in the database. In order to compare word use trajectories across time and across varying utility, we normalize the trajectories for each word i by the average use

$$\langle f_i \rangle \equiv \frac{1}{T_i} \sum_{t=t_{0,i}}^{t_{f,i}} f_i(t) \tag{S9}$$

over the lifetime $T_i \equiv t_{f,i} - t_{0,i} + 1$ of the word, leading to the normalized trajectory,

$$f'_i(\tau) = f'_i(t - t_{i,0} | t_{i,0}, T_i) \equiv f_i(t - t_{i,0}) / \langle f_i \rangle \ . \tag{S10}$$

By analogy, in order to compare various growth trajectories, we normalize the relative growth rate trajectory $r'_i(t)$ by the standard deviation over the entire lifetime,

$$\sigma[r_i] \equiv \sqrt{\frac{1}{T_i} \sum_{t=t_{0,i}}^{t_{f,i}} [r_i(t) - \langle r_i \rangle]^2} \ . \tag{S11}$$

Hence, the normalized relative growth trajectory is

$$r'_i(\tau) = r'_i(t - t_{i,0} | t_{i,0}, T_i) \equiv r_i(t - t_{i,0}) / \sigma[r_i] \ . \tag{S12}$$

Using these normalized trajectories, Fig. S9 shows the weighted averages $\langle f'(\tau|T_c) \rangle$ and $\langle r'(\tau|T_c) \rangle$ and the weighted standard deviations $\sigma[f'(\tau|T_c)]$ and $\sigma[r'(\tau|T_c)]$. We compute $\langle \cdots \rangle$ and $\sigma[\cdots]$ for each trajectory year τ using all N_t trajectories (Table S1) and using all words that satisfy the criteria $T_i \geq T_c$ and $t_{i,0} \geq 1800$. We compute the weighted average and the weighted standard deviation using $\langle f_i \rangle$ as the weight value for word i, so that $\langle \cdots \rangle$ and $\sigma[\cdots]$ reflect the lifetime trajectories of the more common words that are "new" to each corpus.

We analyze the relative growth of word use in a fashion parallel to the economic growth of financial institutions, and show in Fig. S11(B) that the pdf $P(r')$ for the relative growth rates is not only centered around zero change corresponding to $r \approx 0$ but is also symmetric around this average. Hence, for every word that is declining, there is another word that is gaining by the same relative amount. Since there is an intrinsic word maturity $\sigma[r'(\tau|T_c)]$ that is not accounted for in the quantity $r'_i(\tau)$, we further define the detrended relative growth

$$R \equiv r'_i(\tau)/\sigma[r'(\tau|T_c)] \tag{S13}$$

which allows us to compare the growth factors for new words at various life stages. The result of this normalization is to rescale the standard deviations for a given trajectory year τ to unity for all values of $r'_i(\tau)$. Figs. 4 and S10(B) show common growth patterns $P(R)$, independent of corpus. Moreover, we find that the Laplace distributions $P(R)$ found for the growth rates of word use are surprisingly similar to the distributions of growth rates for economic institutions of varying size, such as scientific journals, small and large companies, universities, religious institutions, entire countries and even bird populations [10–13, 16, 21, 25–35].

V. QUANTIFYING THE LONG-TERM SOCIAL MEMORY

In order to gain understanding of the overall dynamics of word use, we have focused much of our analysis on the distributions of f_i and r_i. However, distributions of single observation values discard information about temporal ordering. Hence, in this section we also examine the temporal correlations in each time series $f_i(t)$ to uncover memory patterns in the word use dynamics. To this end, we compare the autocorrelation properties of each $f_i(t)$ to the well-known properties of the time series corresponding to a 1-dimensional random walk.

In a time interval δt, a time series $Y(t)$ deviates from the previous value $Y(t-\delta t)$ by an amount $\delta Y(t) \equiv Y(t) - Y(t-\delta t)$. A powerful result of the central limit theorem, also known as Fick's law of diffusion, is that if the displacements are independent (uncorrelated corresponding to a simple Markov process), then the total displacement $\Delta Y(t) = Y(t) - Y(0)$ from the initial location $Y(0) \equiv 0$ scales according to the total time t as

$$\Delta Y(t) \equiv Y(t) \sim t^{1/2} \ . \tag{S14}$$

However, if there are long-term correlations in the time series $Y(t)$, then the relation is generalized to

$$\Delta Y(t) \sim t^{H} \ , \tag{S15}$$

where H is the Hurst exponent which corresponds to positive correlations for $H > 1/2$ and negative correlations for $H < 1/2$.

Since there may be underlying social/political/technological trends that influence each time series $f_i(t)$, we use the detrended fluctuation analysis (DFA) method [18, 36] to analyze the residual fluctuations $\Delta f_i(t)$ after we remove the local linear trends using time windows of varying length Δt. The time series $\tilde{f}_i(t|\Delta t)$ corresponds to the locally detrended time series using window size Δt. Hence, we calculate the Hurst exponent H using the relation between the root-mean-square displacement $F(\Delta t)$ and the window size Δt [17, 18],

$$F(\Delta t) = \sqrt{\langle \Delta \tilde{f}_i(t|\Delta t)^2 \rangle} = \Delta t^H \ . \tag{S16}$$

Here $\Delta \tilde{f}_i(t|\Delta t)$ is the local deviation from the average trend, analogous to $\Delta Y(t)$ defined above.

Fig. S7 shows 4 different $f_i(t)$ in panel (A), and plots the corresponding $F_i(\Delta t)$ in panel (B). The calculated H_i values for these 4 words are all significantly greater than the uncorrelated $H = 0.5$ value, indicating strong positive long-term correlations in the use of these words, even after we have removed the local trends. In these cases, the trends are related to political events such as war in the cases of "Americanism" and "Repatriation", or the bursting associated with new technology in the case of "Antibiotics," or new musical trends in the case of "polyphony."

In Fig. S8 we plot the pdf of H_i values calculated for the relatively common words analyzed in Fig. S5. We also plot the pdf of H_i values calculated from shuffled time series, and these values are centered around $\langle H \rangle \approx 0.5$ as expected from the removal of the intrinsic temporal ordering. Thus, using this method, we are able to quantify the social memory characterized by the Hurst exponent which is related to the bursting properties of linguistic trends, and in general, to bursting phenomena in human dynamics [9, 19, 20].

VI. DATA METHODS FOR ANALYZING NEW WORDS

In order to analyze the lifetime trajectories of relatively new words, we select data according to several criteria (i) We analyze words of length 20 or less in order to diminish the effects of insignificant strings of characters which can appear, e.g. bababadalgharaghtakamminarronnkonnbronntonnerronntuonnthunntrovarrhounawnskawntoohoohoordenenthurnuk. (ii) We only consider words that have their first appearance $y_{0,i}$ after year $Y_0 \equiv 1800$. We then compare words of varying longevity, where the lifetime of word i is $T_i \equiv t_{f,i} - t_{0,i} + 1$, and group the words into 4 career trajectory sets according to the four different thresholds $T_i \geq T_c \equiv \{25, 50, 100, 200\}$ years. For the calculation of the average word-use trajectory $\langle u(\tau) \rangle$ and growth rate trajectory $\langle r'(\tau) \rangle$, we use a sparsity threshold $s_c \equiv 0.2$ so that we consider words that have at most $s_c \cdot T_i$ years with no recorded use, corresponding to $u_i(t) = 0$ for year t. In the case that $u_i(t) = 0$, we use the approximation $f_i(t) \equiv f_i(t-1)$. Using these data cuts, we still have a significant number N_t of unique word trajectories to analyze, as shown in the data summary in column 6 of Table S1.

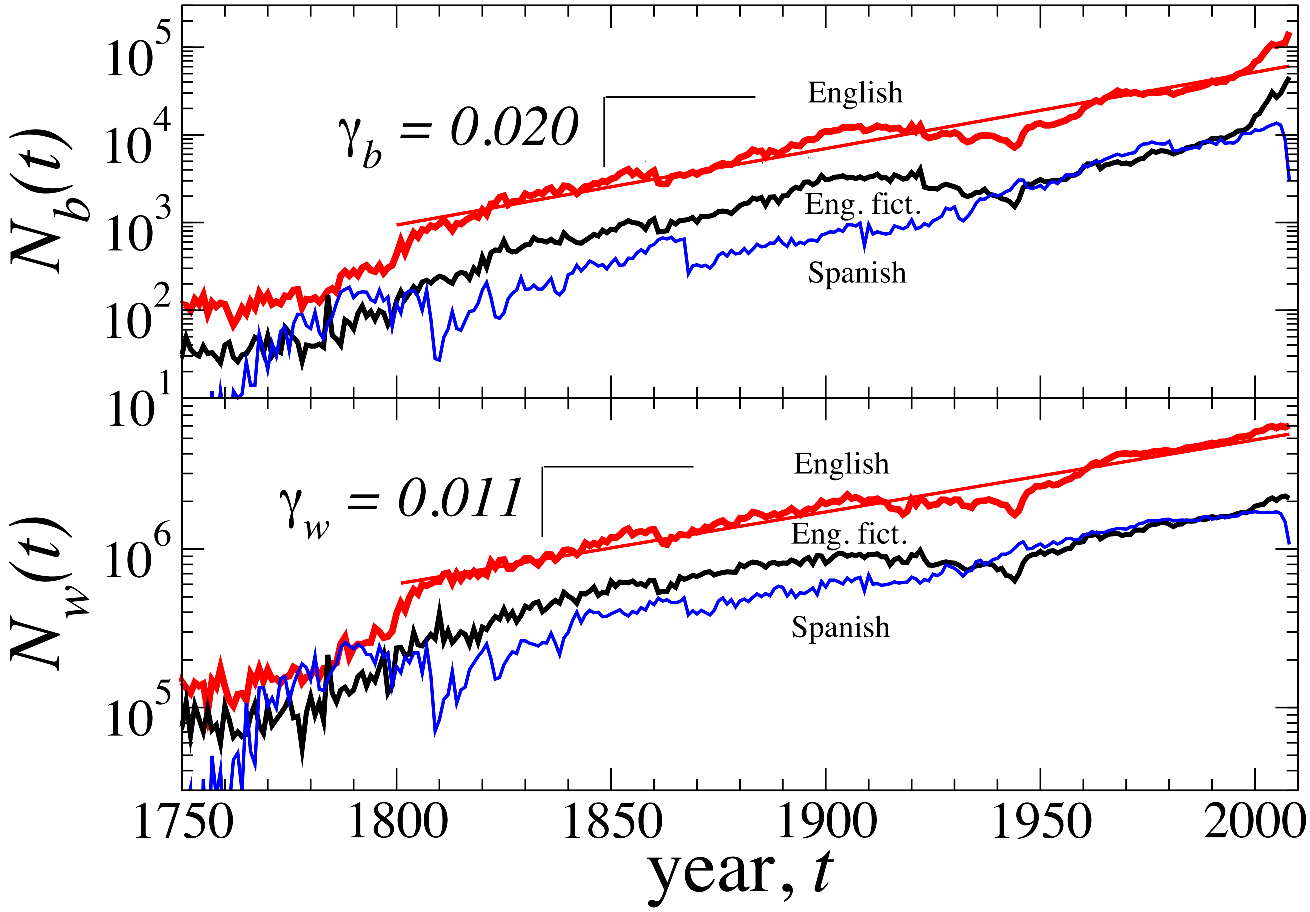

FIG. S1: The exponential growth in the number of books $N_b(t)$ and the number of distinct words $N_w(t)$ by year for three selected literature sets over the 259-year period 1750–2008. We calculate $N_b(t)$ by analyzing the most common words in each corpus: "and" (English and English fiction) and "el" (Spanish). We set $N_b(t)$ equal to the number of books digitized that have at least one occurrence of the most common word. Each case shows an increasing exponential trend; but marked deviations occur around periods of national/international war during which national productive capacity was most likely diverted towards the war effort. If the growth of the vernacular (words used) is exponential, $N_u(t) \propto \exp[\gamma t]$, then the logarithmic change in the growth rate of total words used, $\ln N_u(t) - \ln N_u(t-\Delta t) = \gamma$. Thus, for exponential growth of vocabulary, the growth rate of relative word use is simply $r_i(t) \equiv \ln f_i(t) - \ln f_i(t-\Delta t) = \ln[u_i(t)/u_i(t-\Delta t)] - \ln[N_u(t)/N_u(t-\Delta t)] = \ln[u_i(t)/u_i(t-\Delta t)] - \gamma$. We approximate the exponential trend for the English corpus over the years 1800–2008 and calculate the growth rate $\gamma_b = 0.0201 \pm 0.0004$ books/year and $\gamma_w = 0.0105 \pm 0.0002$ words/year. The rate parameter γ is related to the entry rate of new ideas and events, and is also related to the "crowding out" of old ideas and events, whereby newer topics typically have a higher probability of being discussed in literature.

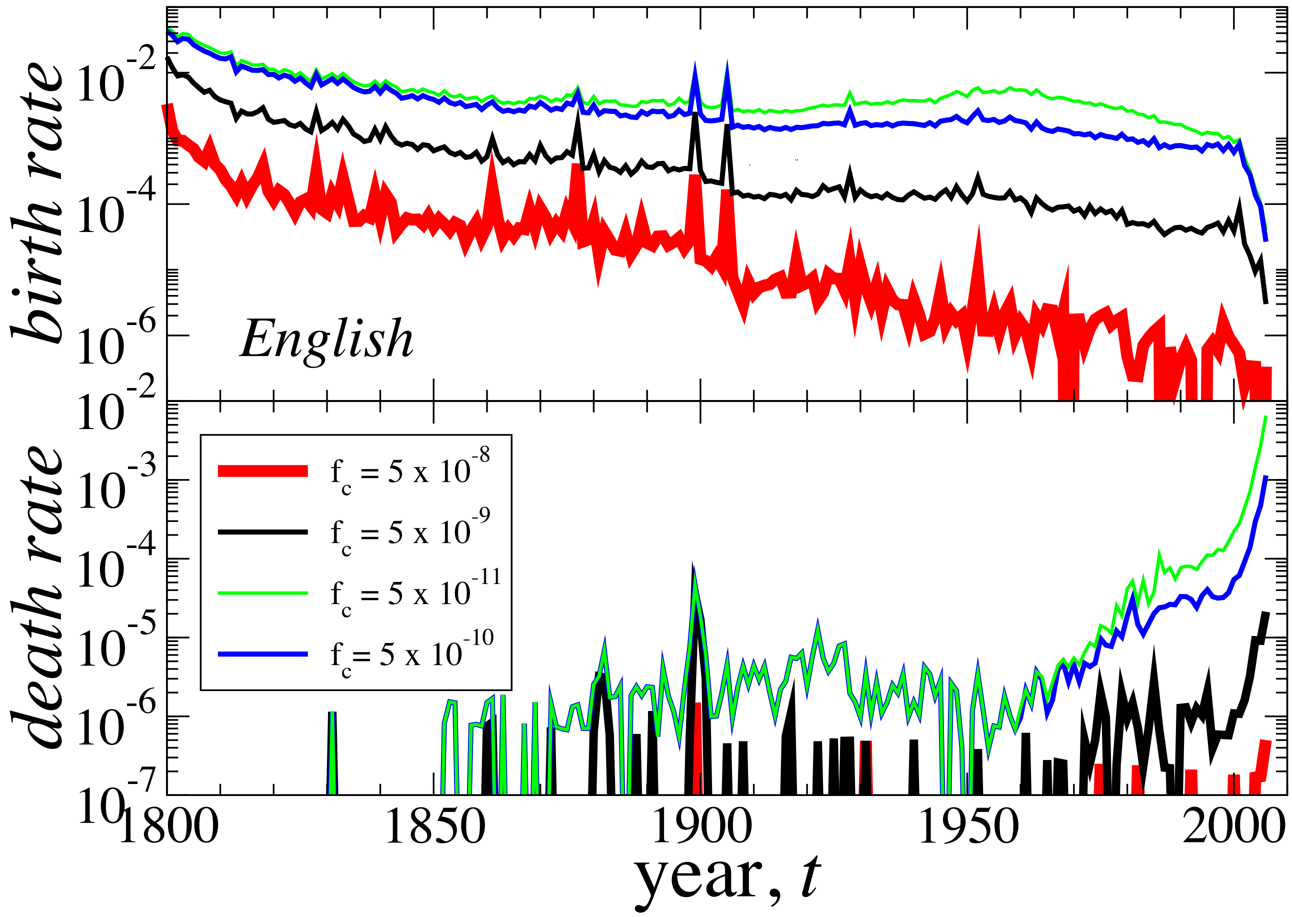

FIG. S2: The birth and death rates of a word depends on the relative use of the word. For the English corpus, we calculate the birth and death rates for words with median lifetime relative use $\mathrm{Med}(f_i)$ satisfying $\mathrm{Med}(f_i) > f_c$. The difference in the birth rate curves corresponds to the contribution to the birth rate of words in between the two f_c thresholds, and so the small difference in the curves for small f_c indicates that the birth rate is largely comprised of words with relatively large $\mathrm{Med}(f_i)$. Consistent with this finding, the largest contribution to the death rate is from words with relatively low $\mathrm{Med}(f_i)$. By visually inspecting the lists of dying words, we confirm that words with large relative use rarely become completely extinct (see Fig. 1 for a counterexample word "Roentgenogram" which was once a frequently used word, but has since been eliminated due to competitive forces with other high-fitness competitors).

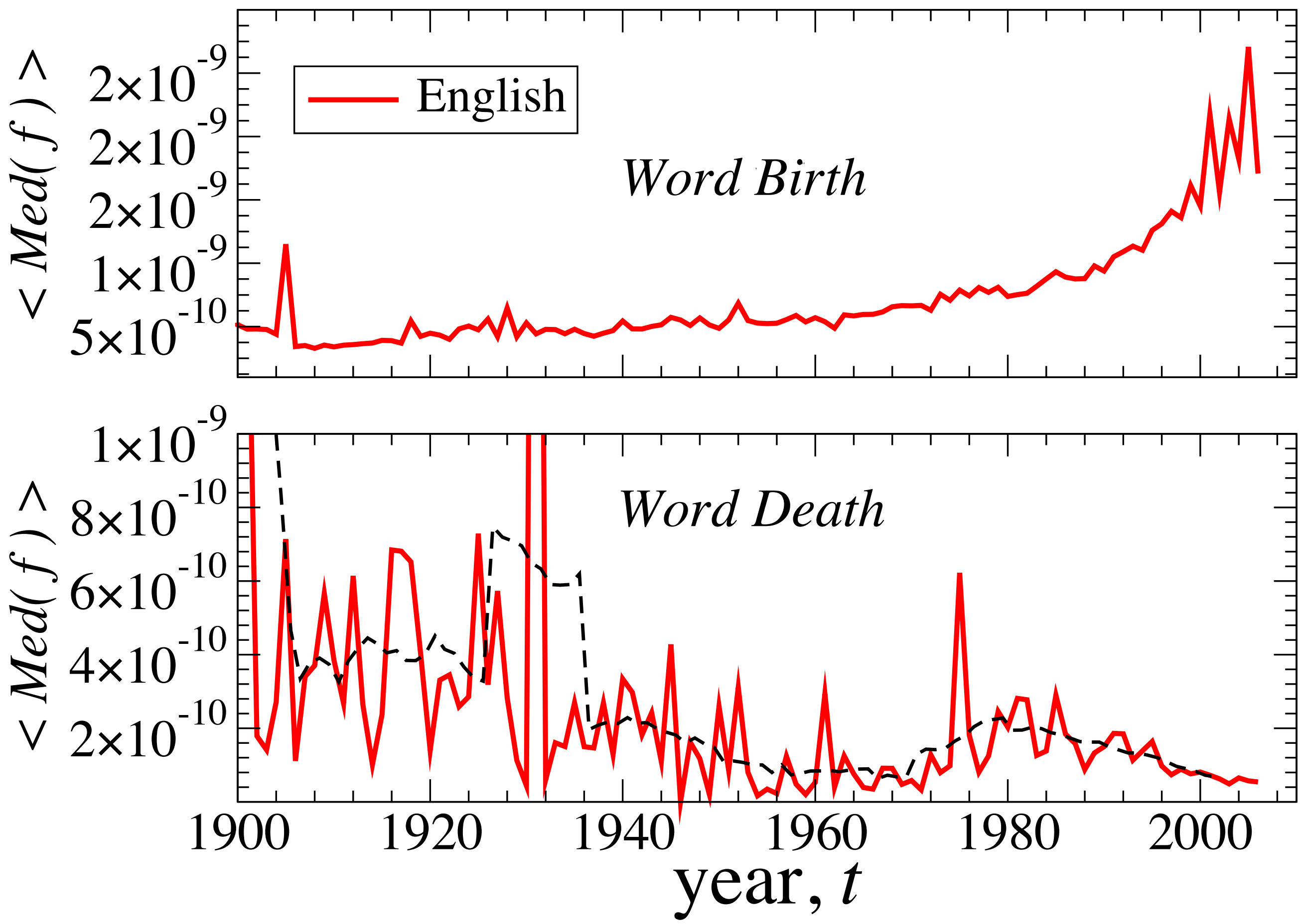

FIG. S3: Trends in the relative uses of words that either were born or died in a given year show that the degree of competition between words is time dependent. For the English corpus, we calculate the average median lifetime relative use $\langle \mathrm{Med}(f_i) \rangle$ for all words i born in year t (top panel) and for all words i that died in year t (bottom panel), which also includes a 5-year moving average (dashed black line). The relative use ("utility") of words that are born shows a dramatic increase in the last 20–30 years, as many new technical terms, which are necessary for the communication of modern devices and ideas, are born with relatively high intrinsic fitness. Conversely, with higher editorial standards and the recent use of word processors which include spelling standardization technology, the words that are dying are those words with low relative use, which we also confirm by visual inspection of the lists of dying words to be misspelled and nonsensical words.

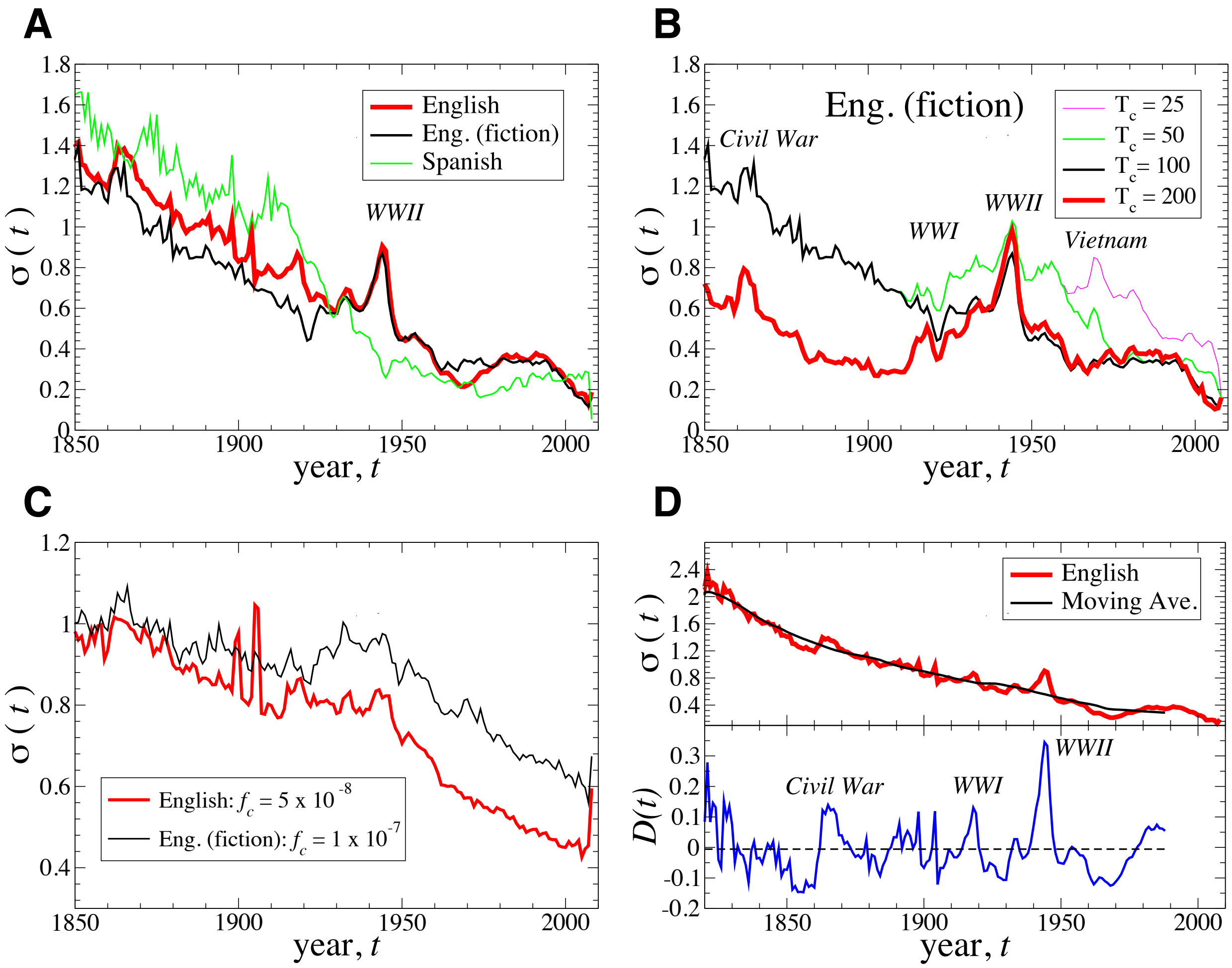

FIG. S4: Historical factors influence the evolution of word utility. The variation $\sigma(t) \equiv \sigma(r|t)$ in the growth rate $r_i(t)$ of relative word use defined in Eq. (1) demonstrates the increased variation in growth rates during periods of international crisis (e.g. World War II). The increase in $\sigma(t)$ during the World War II, despite the overall decreasing trend in $\sigma(t)$ over the 159 year period 1850–2008, demonstrates a"globalization" effect, whereby societies are brought together by a common event and a unified media. Such contact between relatively isolated systems necessarily leads to information flow, much as in the case of thermodynamic heat flow between two systems that are at different temperatures and are brought into contact. (A) The variation $\sigma(t)$ calculated for the relatively new words with $T_c = 100$. The Spanish corpus does not show an increase in $\sigma(t)$ during World War II, indicative of the relative isolation of South America and Spain from the European conflict. (B) $\sigma(t)$ for four sets of relatively new words i that meet the criteria $T_i \geq T_c$ and $t_{i,0} \geq 1800$. The oldest "new" words, corresponding to $T_c = 200$, demonstrate the strong increase in $\sigma(t)$ during World War II, with a peak around 1945. (C) The standard deviation $\sigma(t)$ in the growth rates $r_i(t)$ for the most common words that meet the criterion that the average relative use $\langle f_i \rangle > f_c$ over the entire lifetime. For this set of words, $\sigma(t)$ are also decreasing with time, consistent with a "crowding out" effect. (D) We compare the variation $\sigma(t)$ for common words with the 20-year moving average over the time period 1820–1988, which also demonstrates an increasing $\sigma(t)$ during times of national/international crisis, such as the American Civil War (1861–1865), World War I (1914–1918) and World War II (1939–1945), and recently during the 1980s and 1990s, possibly as a result of new digital media (e.g. the internet) which offer new environments for the evolutionary dynamics of word use. $D(t)$ is the difference between the moving average and $\sigma(t)$.

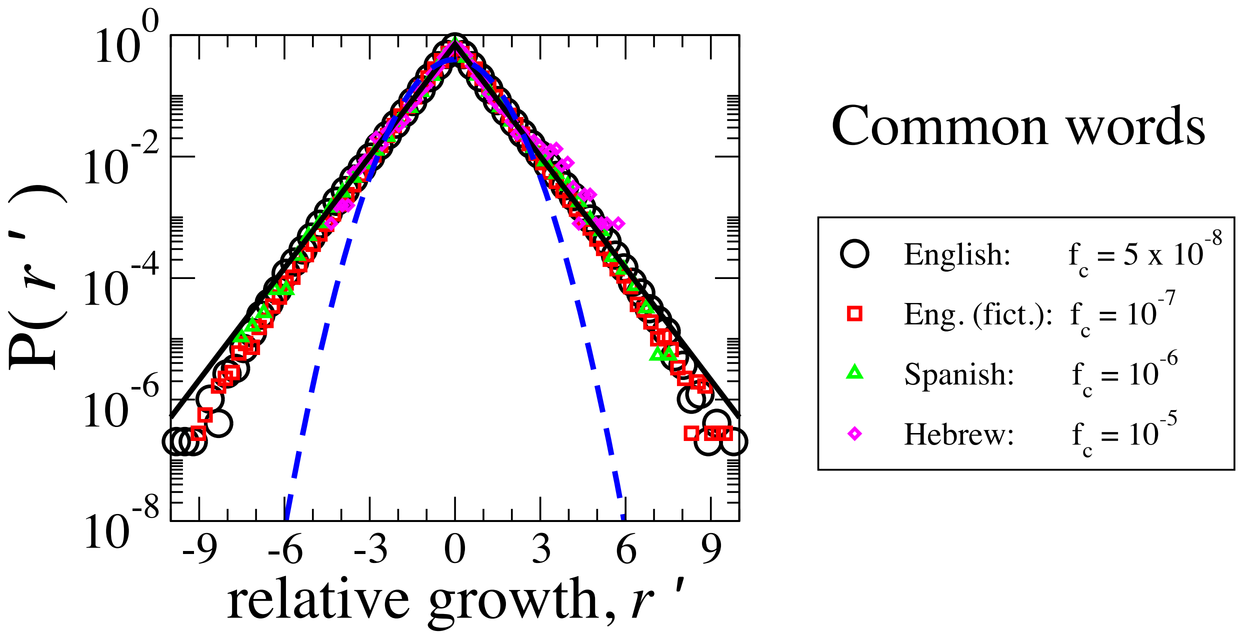

FIG. S5: PDF of the annual relative growth rate r' for dataset #ii words which have average relative use $\langle f_i \rangle \geq f_c$. Hence, these select words correspond to relatively common words. We find that there is a common distribution for both relatively new words (compare with corresponding panels in Fig. S11) as well as notable words, which are used relatively frequently. We plot a Laplace distribution with unit variance (solid black lines) and the Gaussian distribution with unit variance (dashed blue curve) for reference. In order to select relatively frequently used words, we use the following criteria: $T_i \geq 10$ years, $1800 \leq t \leq 2008$, and $\langle f_i \rangle \geq f_c$. There is no need to account for the age-dependent trajectory $\sigma[r'(\tau|T_c)]$, as in the normalized growth defined in Eq. (S13), for these relatively common words since they are all most likely in the mature phase of their lifetime trajectory.

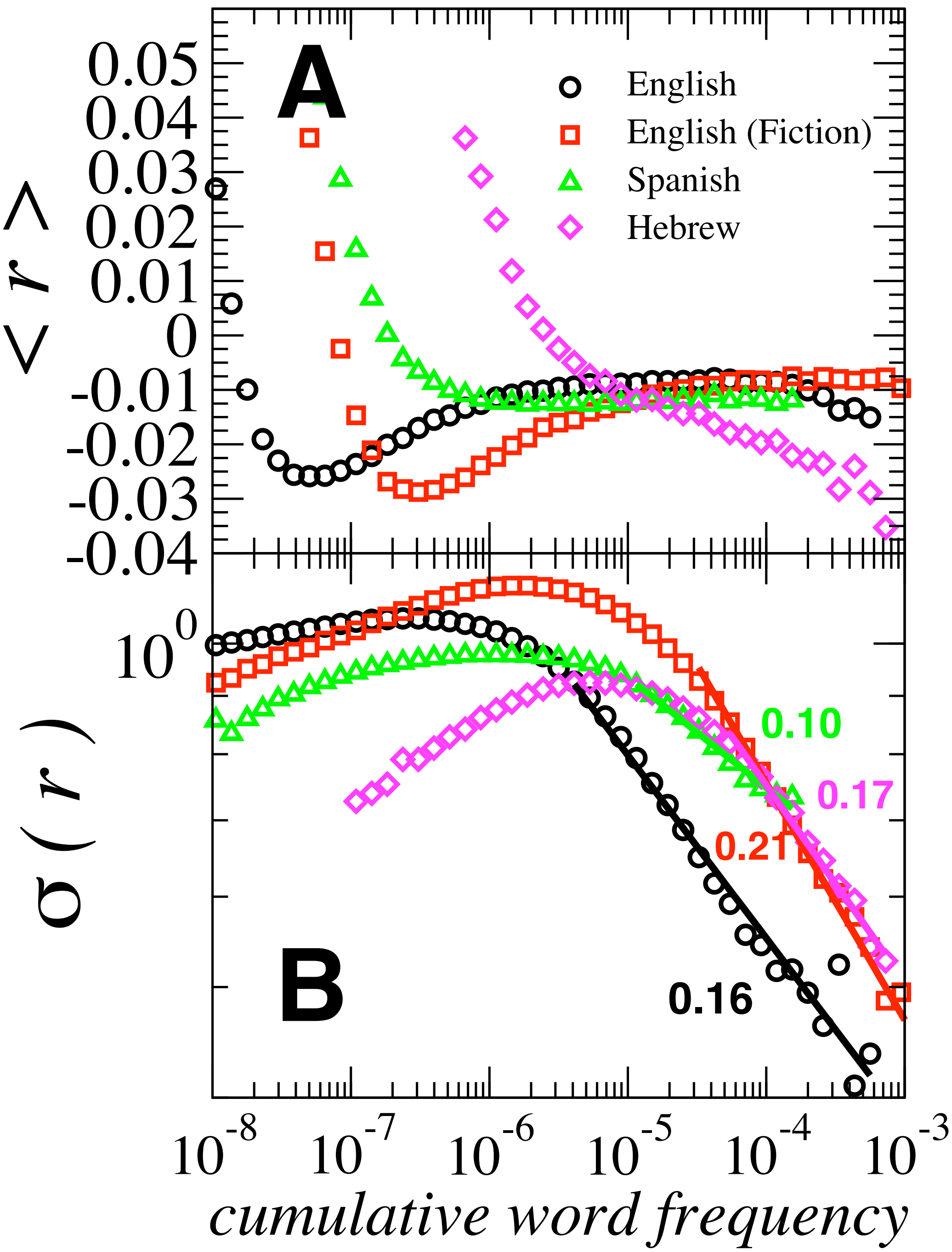

FIG. S6: The dependence of growth rates on the cumulative word frequency $S_i \equiv \sum_{t'=0}^{t} f_i(t)$ calculated for a combination of new [dataset (i)] and common [dataset (ii)] words that satisfy the criteria $T_i \geq 10$ years (we verify similar results for threshold values $T_c = 50, 100$, and 200 years). (A) Average growth rate $\langle r \rangle$ saturates at relatively constant (negative) values for large S. (B) Scaling in the standard deviation of growth rates $\sigma(r|S) \sim S^{-\beta}$ for words with large S. This scaling relation is also observed for the growth rates of large economic institutions, ranging in size from companies to entire countries [13, 15]. Here this size-variance relation corresponds to scaling exponent values $0.10 < \beta < 0.21$, which are related to the non-trivial bursting patterns and non-trivial correlation patterns in literature topicality. We calculate $\beta_{Eng.} \approx 0.16 \pm 0.01$, $\beta_{Eng.fict} \approx 0.21 \pm 0.01$, $\beta_{Spa.} \approx 0.10 \pm 0.01$ and $\beta_{Heb.} \approx 0.17 \pm 0.01$.

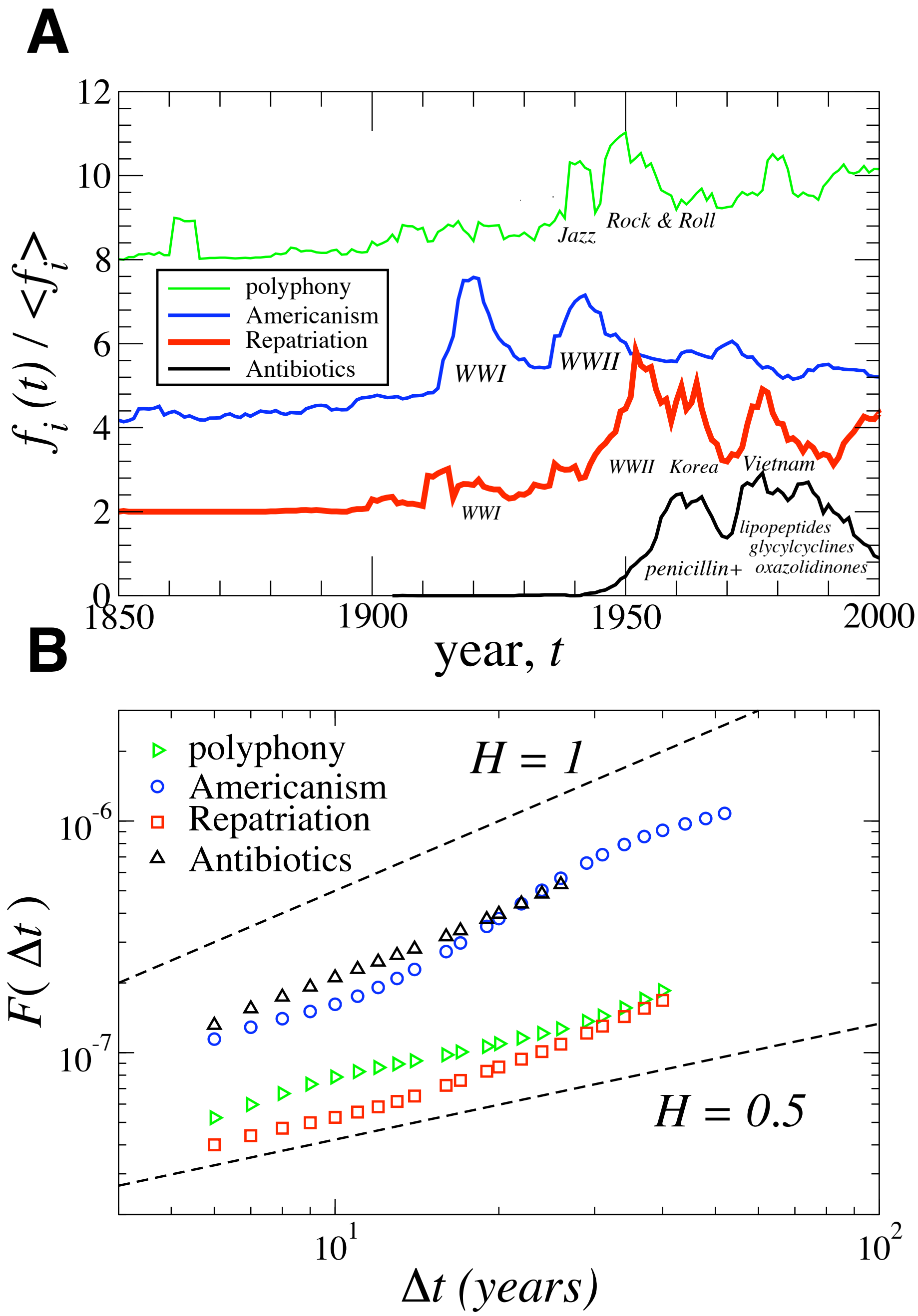

FIG. S7: Quantifying the strong positive correlations in social memory using the relative word use $f_i(t)$. **(A)** Four example $f_i(t)$, given in units of the average use $\langle f_i \rangle$, show bursting of use as a result of social and political "shock" events. We choose these four examples based on their relatively large $H_i > 0.5$ values. The use of "polyphony" in the English corpus shows peaks during the eras of jazz and rock and roll. The use of "Americanism" shows bursting during times of war, and the use of "Repatriation" shows an approximate 10-year lag in the bursting after WWII and the Vietnam War. The use of the word "Antibiotics" is related to technological advancement. The top 3 curves are vertically displaced by a constant from the value $f_i(1800) \approx 0$ so that the curves can be distinguished. **(B)** We use detrended fluctuation analysis (DFA) to calculate the Hurst exponent H_i for each word in order to quantify the long-term correlations ("memory") in each $f_i(t)$ time series. Fig. S8 shows the probability density function $P(H)$ of H_i values calculated for the relatively common words found in English fiction and Spanish, summarized in Table S2.

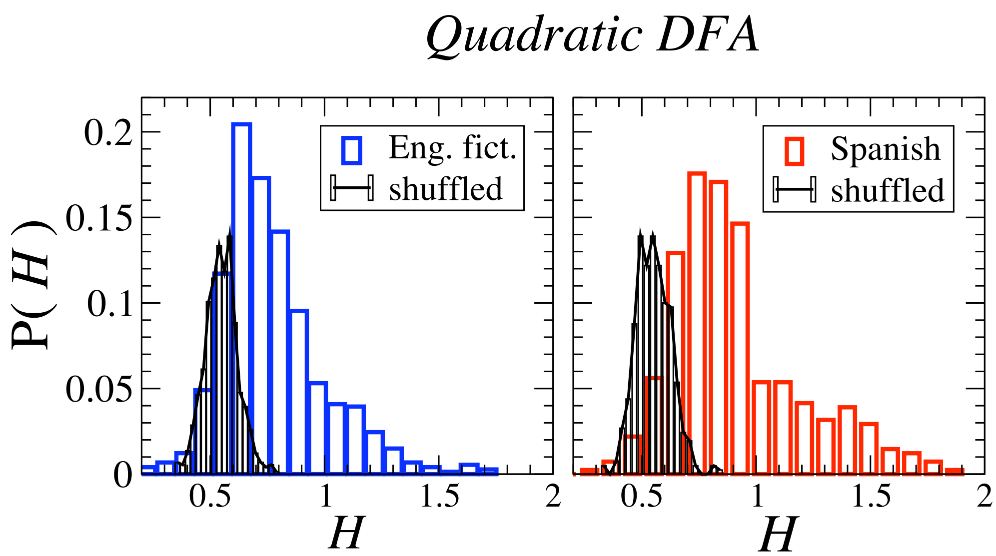

FIG. S8: Results of detrended fluctuation analysis (DFA)[17, 18] on the words analyzed in Fig. S5 show strong long-term memory with positive correlations ($H > 0.5$), indicating strong correlated bursting in the dynamics of word use, possibly corresponding to historical, social, or technological events. We calculate $\langle H_i \rangle \pm \sigma = 0.77 \pm 0.23$ (Eng. fiction) and $\langle H_i \rangle \pm \sigma = 0.90 \pm 0.29$ (Spanish). The size-variance β values calculated from the data in Fig. S6 confirm the theoretical prediction $\langle H \rangle = 1 - \beta$. Fig. S6 shows that $\beta_{Eng.fict} \approx 0.21 \pm 0.01$ and $\beta_{Spa.} \approx 0.10 \pm 0.01$. For the shuffled time series, we calculate $\langle H_i \rangle \pm \sigma = 0.55 \pm 0.07$ (Eng. fiction) and $\langle H_i \rangle \pm \sigma = 0.55 \pm 0.08$ (Spanish), which are consistent with time series that lack temporal ordering (memory).

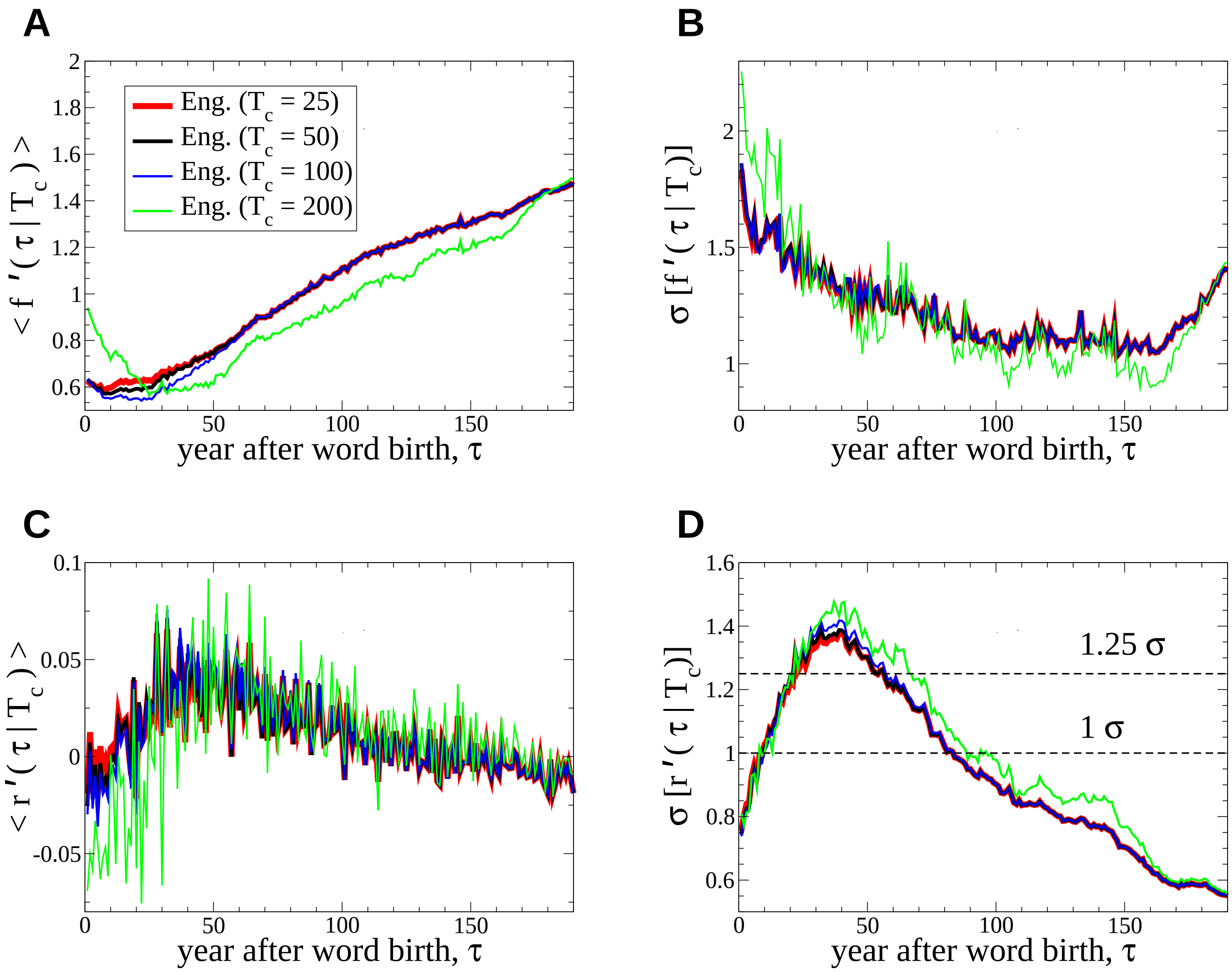

FIG. S9: Characteristics of the time-dependent word trajectory show the time scales over which a typical word becomes relevant or fades. For 4 values of T_c, we show the word trajectories for dataset (i) words in the English corpus, although the same qualitative results hold for the other languages analyzed. Recall that T_c refers to the subset of timeseries with lifetime $T_i \geq T_c$, so that two trajectories calculated using different thresholds $T_c^{(1)}$ and $T_c^{(2)}$ only vary for $\tau < Max[T_c^{(1)}, T_c^{(2)}]$. We show weighted average and standard deviations, using $\langle f_i \rangle$ as the weight for word i contributing to the calculation of each time series in year τ. (A) The relative use increases with time, consistent with the definition of the weighted average which biases towards words with large $\langle f_i \rangle$. For words with large T_i, the trajectory has a minimum which begins to reverse around $\tau \approx 40$ years, possibly reflecting the amount of time it takes to reach a critical utility threshold that corresponds to a relatively high fitness value for the word in relation to its competitors. (B) The variations in $\langle f(\tau|T_c) \rangle$ decrease with time reflecting the transition from the insecure "infant" phase to the more secure "adult" phase in the lifetime trajectory. (C) The average growth trajectory is qualitatively related to the logarithmic derivative of the curve in panel (A), and confirms that the region of largest positive growth is $\tau \approx$ 30–50 years. (D) The variations in the average trajectory are larger than 1.25 σ for $30 \lesssim \tau \lesssim 50$ years and are larger than 1.0 σ for $10 \lesssim \tau \lesssim 80$ years. This regime of large fluctuations in the growth rates conceivably corresponds to the time period over which a successful word is accepted into the standard lexicon, e.g. a word included in an official dictionary or an idea/event recorded in an encyclopedia or review.

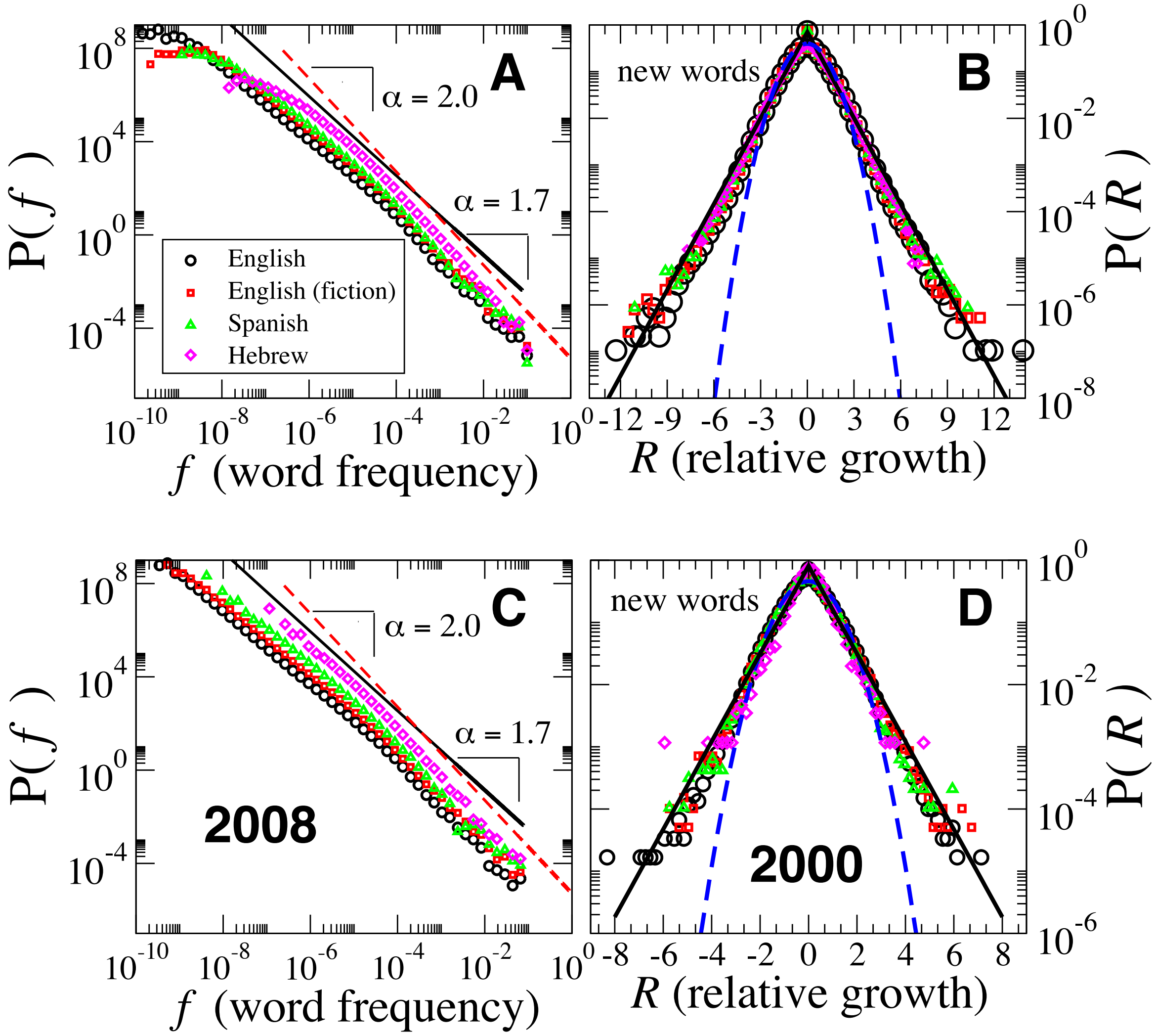

FIG. S10: (A) The distribution of relative word use f for all words aggregated over the time period 1800-2008 shows a crossover in the Zipf scaling from $\alpha \approx 1.7$ to $\alpha \approx 2$, where the latter value is in agreement with the Zipf law [7]. (B) The Laplace distribution of annual relative change in word use for the relatively new words corresponding to dataset #1 (see Table S1 for data summary), using word data over the time period 1800-2008 for words with a lifetime $T_i \geq T_c \equiv 100$ years and a sparsity threshold $s_c \equiv 0.2$. For comparison we plot a Gaussian distribution with unit variance (dashed blue curve), which displays rapid parabolic decay on the semi-logarithmic axis. The data agree over the entire range with the Laplace distribution (solid black line) defined in Eq. (2) with $\sigma \equiv 1$. The pdf $P(f)$ in panel (C) for year 2008 data only and the pdf $P(R)$ in panel (D) for year 2000 data only show that the distributions are also stable for word data aggregated over only a single year.

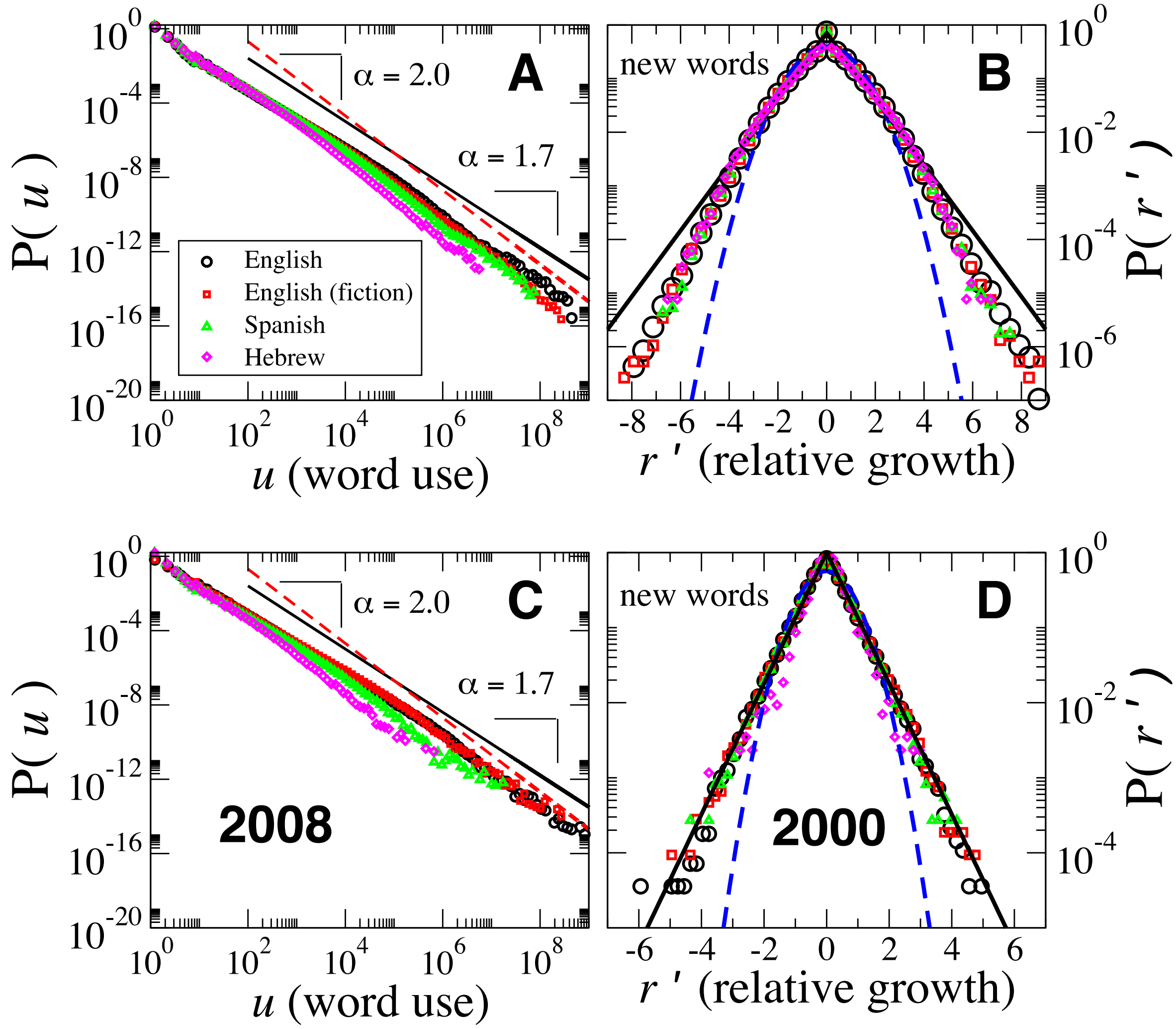

FIG. S11: Distributions of word use u and relative change r' to be compared with Fig. S10. (A) Distribution of word use u for all words aggregated over the time period 1800-2008 also shows a crossover in the Zipf scaling from $\alpha \approx 1.7$ to $\alpha \approx 2$, where the latter value is in agreement with the Zipf law [7]. (B) Distribution of annual relative change in word use r' for the relatively new words corresponding to dataset #1 (see Table S1 for data summary), using word data over the time period 1800-2008 for words with a lifetime $T_i \geq T_c \equiv 100$ years and a sparsity threshold $s_c \equiv 0.2$. For comparison we plot a Gaussian distribution with unit variance (dashed blue curve), which displays fast parabolic decay on the semi-logarithmic axis. The pdf $P(r')$ is more heavy in the tails than the Gaussian distribution but less-heavy than the Laplace distribution with unit variance (solid black line). The pdf $P(u)$ in panel (C) for only year 2008 data and the pdf $P(r')$ in panel (D) for year 2000 data only show that the distributions are also stable on an annual basis. The pdfs $P(r')$ have a standard deviation $\sigma(r'|t)$ that is year dependent (e.g. $\sigma(r'|t = 2008) \approx 0.7$). This observation explains why $P(r')$ in panel (B) is not fit well by the Laplace distribution in the tails, since this $P(r')$ is actually a mixture of distributions with varying widths $\sigma(r'|t)$. To account for this variation, we use the growth factor R defined in Eq. (S13) to better quantify the growth rates of word use by accounting for word maturity effects.

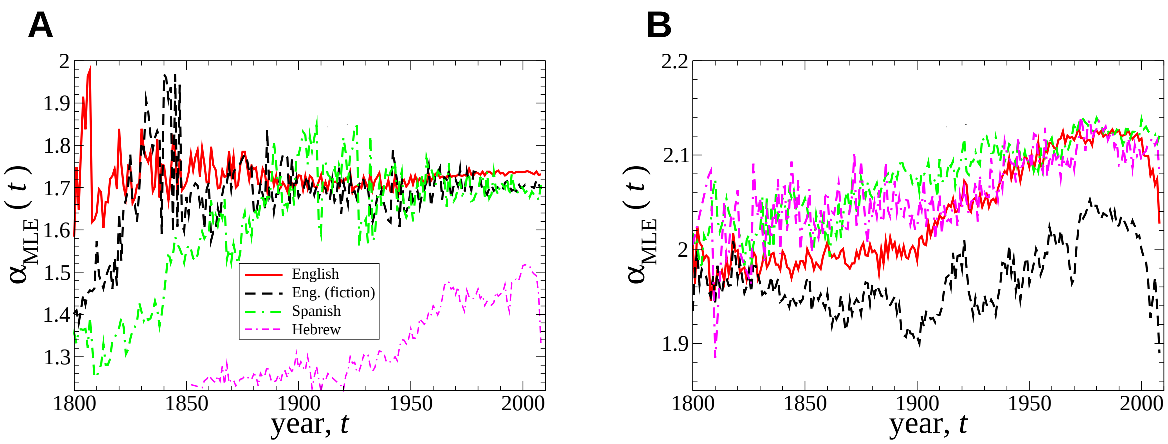

FIG. S12: Annual scaling exponent which quantifies the relative use of words according to the power-law distribution $P(f(t)) \sim f^{-\alpha(t)}$, which is related to the rank-frequency Zipf law exponent ζ by the relation $\alpha \approx 1 + 1/\zeta$. We observe a crossover in the exponent α in Fig. S10(A), and so we use the maximum likelihood estimation (MLE) method to calculate the exponent for both low word use ("unlimited lexicon") and high word use ("kernel lexicon") regimes [7]. (A) For the low word use regime $10^{-8} < f < 10^{-6}$, we calculate the exponent $\alpha \approx 1.7$ which is smaller than the value $\alpha \approx 2$ predicted by the Zipf law. (B) We confirm Zipf scaling corresponding to $\alpha \approx 2$ for the tail of $P(f)$ using the range $f > 10^{-4}$ for each corpus.

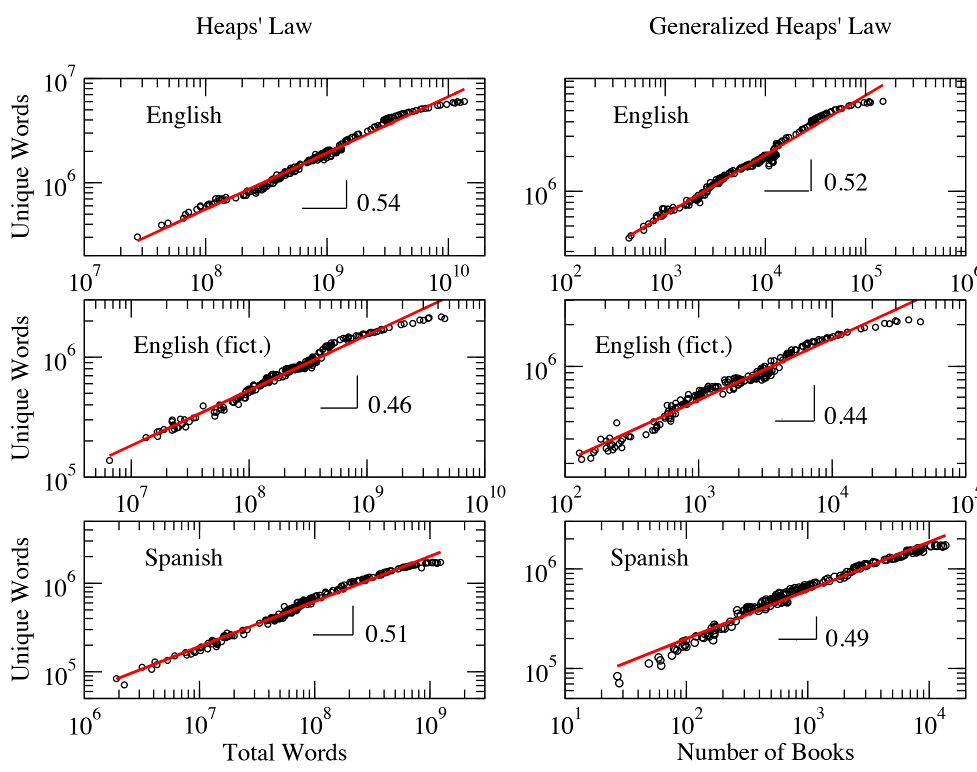

FIG. S13: Heaps' law quantifies the marginal utility of adding a new word to a vocabulary, here in the context of an entire corpus. In the panels on the left, we show scatter plots of $N_u(t)$, the total number of words published vs $N_w(t)$, the total number of unique words used (vocabulary size). Each data set shows a strong scaling relation $N_w(t) \sim [N_u(t)]^b$ with $b \approx 0.5$ over several orders of magnitude for each corpus analyzed. In the panels on the right, for comparison we also show plots of $N_b(t)$, the total number of books published vs $N_w(t)$, the total number of unique words used in each year. Estimates for the scaling exponent b values are listed in each panel. A simple model of topicality in a text [6] shows that $b \approx 1/\alpha$, where α is the pdf scaling exponent defined in Eq. (S3). We verify this theoretical prediction, which further demonstrates an increasing marginal utility of new words, meaning that each additional word added to a vocabulary superlinearly increases the total number of words written in a corpus, $N_u \sim (N_w)^{\alpha}$.

TABLE S1: Summary of annual growth trajectory data for varying threshold T_c, and $s_c = 0.2$, $Y_0 \equiv 1800$ and $Y_f \equiv 2008$.

Corpus,	Annual growth $R(t)$ data					
(1-grams)	$T_c(years)$	$N_t(words)$	% (of all words)	$N_R(values)$	$\langle R \rangle$	$\sigma[R]$
English	25	302,957	4.1	31,544,800	2.4×10^{-3}	1.00
English fiction	25	99,547	3.8	11,725,984	-3.0×10^{-3}	1.00
Spanish	25	48,473	2.2	4,442,073	1.8×10^{-3}	1.00
Hebrew	25	29,825	4.6	2,424,912	-3.6×10^{-3}	1.00
English	50	204,969	2.8	28,071,528	-1.7×10^{-3}	1.00
English fiction	50	72,888	2.8	10,802,289	-1.7×10^{-3}	1.00
Spanish	50	33,236	1.5	3,892,745	-9.3×10^{-4}	1.00
Hebrew	50	27,918	4.3	2,347,839	-5.2×10^{-3}	1.00
English	100	141,073	1.9	23,928,600	1.0×10^{-4}	1.00
English fiction	100	53,847	2.1	9,535,037	-8.5×10^{-4}	1.00
Spanish	100	18,665	0.84	2,888,763	-2.2×10^{-3}	1.00
Hebrew	100	4,333	0.67	657,345	-9.7×10^{-3}	1.00
English	200	46,562	0.63	9,536,204	-3.8×10^{-3}	1.00
English fiction	200	21,322	0.82	4,365,194	-3.5×10^{-3}	1.00
Spanish	200	2,131	0.10	435,325	-3.1×10^{-3}	1.00
Hebrew	200	364	0.06	74,493	-1.4×10^{-2}	1.00

TABLE S2: Summary of data for the relatively common words that meet the criterion that their average word use $\langle f_i \rangle$ over the entire word history is larger than a threshold f_c, defined for each corpus. In order to select relatively frequently used words, we use the following three criteria: the word lifetime $T_i \geq 10$ years, $1800 \leq t \leq 2008$, and $\langle f_i \rangle \geq f_c$.

Corpus,	Data summary for relatively common words					
(1-grams)	f_c	$N_t(words)$	% (of all words)	$N_{r'}(values)$	$\langle r' \rangle$	$\sigma[r']$
English	5×10^{-8}	106,732	1.45	16,568,726	1.19×10^{-2}	0.98
English fiction	1×10^{-7}	98,601	3.77	15,085,368	5.64×10^{-3}	0.97
Spanish	1×10^{-6}	2,763	0.124	473,302	9.00×10^{-3}	0.96
Hebrew	1×10^{-5}	70	0.011	6,395	3.49×10^{-2}	1.00

TABLE S3: Summary of *Google* corpus data. Annual growth rates correspond to data in the 209-year period 1800–2008.

Corpus,	Annual use $u_i(t)$ 1-gram data					Annual growth $r(t)$ data		
(1-grams)	$N_u(uses)$	Y_i	Y_f	$N_w(words)$	$Max[u_i(t)]$	$N_r(values)$	$\langle r \rangle$	$\sigma[r]$
English	3.60×10^{11}	1520	2008	7,380,256	824,591,289	310,987,181	2.21×10^{-2}	0.98
English fiction	8.91×10^{10}	1592	2009	2,612,490	271,039,542	122,304,632	2.32×10^{-2}	1.03
Spanish	4.51×10^{10}	1532	2008	2,233,564	74,053,477	111,333,992	7.51×10^{-3}	0.91
Hebrew	2.85×10^{9}	1539	2008	645,262	5,587,042	32,387,825	9.11×10^{-3}	0.90