Stochastic Model for the Vocabulary Growth in Natural Languages

Martin Gerlach and Eduardo G. Altmann

Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany

(Received 4 December 2012; revised manuscript received 20 March 2013; published 14 May 2013)

We propose a stochastic model for the number of different words in a given database which incorporates the dependence on the database size and historical changes. The main feature of our model is the existence of two different classes of words: (i) a finite number of core words, which have higher frequency and do not affect the probability of a new word to be used, and (ii) the remaining virtually infinite number of noncore words, which have lower frequency and, once used, reduce the probability of a new word to be used in the future. Our model relies on a careful analysis of the Google Ngram database of books published in the last centuries, and its main consequence is the generalization of Zipf's and Heaps' law to two-scaling regimes. We confirm that these generalizations yield the best simple description of the data among generic descriptive models and that the two free parameters depend only on the language but not on the database. From the point of view of our model, the main change on historical time scales is the composition of the specific words included in the finite list of core words, which we observe to decay exponentially in time with a rate of approximately 30 words per year for English.

DOI: 10.1103/PhysRevX.3.021006 Subject Areas: Complex Systems, Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

Even in our time of big data [1–3], there is no indication of a saturation of the vocabulary size (total number of different words) with increasing database size. In order to clarify whether it is meaningful to estimate a vocabulary size in the limit of infinitely large databases, it is essential to understand not only the birth and death of words [4–6], but also the process governing the usage of new words and its dependence on database size. The interest in this problem is motivated by fundamental linguistic studies [7,8] as well as by applications in search engines, which require an estimation of the number of different words in a given database [9–11].

The scaling between the number of different words, N, and the size of the database in words, M, as $N \sim M^{\lambda}$, is known as Heaps' law [12] and has been studied in different linguistic [13–15] and nonlinguistic [16,17] contexts. The universality and interest of this empirical scaling is surpassed only by Zipf's law [18], which states that the frequency $F(r)$ of the rth most frequent word in a database decays as $F(r) \sim 1/r$. The relation between Heaps' law and Zipf's law has been the subject of great recent interest [19–21]. Furthermore, it is well known that deviations of the Heaps' and Zipf's laws are observed in the tails of Heaps' and Zipf's plots (i.e., for large N and r, respectively) [22–24]. Similar deviations of fat-tailed distributions appear in a variety of social and physical systems [25,26] and are crucial when extrapolating to the limit of large databases.

In this paper, we propose a stochastic growth model whose predictions go beyond the simpler scalings of Heaps' and Zipf's laws and are compatible with actual observations in the tail of the corresponding distributions. Our model is in the same spirit of, but differs from, the simpler versions of Yule's, Simon's, Gibrat's, and preferential attachment growth models [26–29] because it contains two categories of words and leads to two scaling regimes in the Heaps' and Zipf's plots. These findings are supported by a statistical analysis of the Google Ngram database, indicating that the only two free parameters needed in the description of these scalings remain unchanged over centuries and depend only on the language, and that there is a slow change of words belonging to each category. The latter adds to the recent interest in language dynamics as a complex system [30,31].

The paper is organized as follows: In Sec. II, we present a statistical analysis of the Google Ngram database in terms of word frequencies, as well as the growth of the vocabulary. This will then lead us to the formulation of our stochastic model for the vocabulary growth in Sec. III. In Sec. IV, we investigate dynamical aspects on historical time scales within the framework of our model.

II. DATA ANALYSIS

A. Data

The main motivation for our model comes from empirical observations. As databases, we use the Google Ngram corpus [1] for English, German, French, Spanish, and Russian, which provides data of the word frequencies (occurring in printed books), with a yearly resolution for

a period of several hundred years (1520–2000). Our main interest in this database stems from its large size (several millions of books with $>10^{11}$ words) and from the long time span it covers (thus enabling us to trace historical changes in the usage of language). We consider as words only the 1-grams consisting uniquely of letters present in the alphabet of the corresponding language. This pragmatic definition reduces the effect of symbolic sequences, foreign words, numbers, or scanning problems in our observations and should be taken into account when interpreting our findings about the vocabulary. For each language, we use two different partitions of the database: (i) yearly (y), in which case $y(t)$ corresponds to the database of the year t; and (ii) cumulative (Y), in which case $Y(t) = \sum_{t'=to}^{t} y(t')$. We consider only words which appear at least $n = 41$ times in order to avoid biases due to the filtering mechanism used in the Google Ngram database; see Ref. [32], Sec. I, for further details. Here, we show our detailed analysis for the largest database (English, $t_0 = 1520$, $t \in [1805, 2000]$). For the other four languages, we report the main findings and leave the details for Ref. [32].

B. Zipf's analysis

Our first empirical analysis focuses on the distribution of word frequencies. In his seminal work, Zipf proposed that the frequency of the rth most frequent word in a given text is given by $F(r) = F(1)/r$ [18]. It is easy to see that this scaling has to break for large r: Because of the divergence of the harmonic series, for sufficiently large databases, one arrives at $\sum_{r=1}^{N} F(r) > 1$ (sum of frequencies larger than text size). In English, $F(1) \approx 0.07$ (the frequency of "the") and $\sum_{r=1}^{N} F(r) > 1$ for $N \approx 10^6$, meaning that $F(r)$ has to decay faster than $1/r$ for $r \gtrsim 10^6$. This well-known expectation, which is clearly seen in our data shown in Fig. 1(a), motivated numerous different generalizations of Zipf's proposal [33–35]. While many of these generalizations were shown to provide a better account of particular databases, they remain, to a great extent, unsatisfactory because they lack the simplicity and universality of Zipf's original proposal (e.g., the parameters vary depending on the size, topic, or date of publication of the analyzed texts [36,37]). Motivated by the new magnitude of our large database, we apply rigorous statistical tests to determine which of the previously proposed distributions provide a better account of the data. We select seven of the most popular previously proposed heavy-tailed distributions with at most two free parameters [8,23,24]: power law, two power laws, shifted power law, log normal, Weibull, and power laws with exponential cutoffs (in the tail and beginning, respectively). The parameters for each distribution were obtained numerically by means of maximum likelihood (ML) estimation [38]. In addition, we (i) calculate the probability that the data were generated by that model (χ^2 p-value [39,40]) and (ii) compare which model is more likely to describe the data (relative likelihood [41,42]) for each fit (for details see Ref. [32], Sec. IIA).

The results show that it is extremely unlikely ($p < 10^{-15}$) that the data were drawn exactly from any of the proposed distributions, a consequence of the large databases which makes any small (true) deviation incompatible with these simple fits. On the other hand, the results show unequivocally that for English the distribution with two power laws is the best fit ($1 - p < 10^{-15}$) for all databases with a size larger than 10^9 words. We confirm that the double power law is also the best fit for the English Wikipedia, a strong indication of the validity of this result

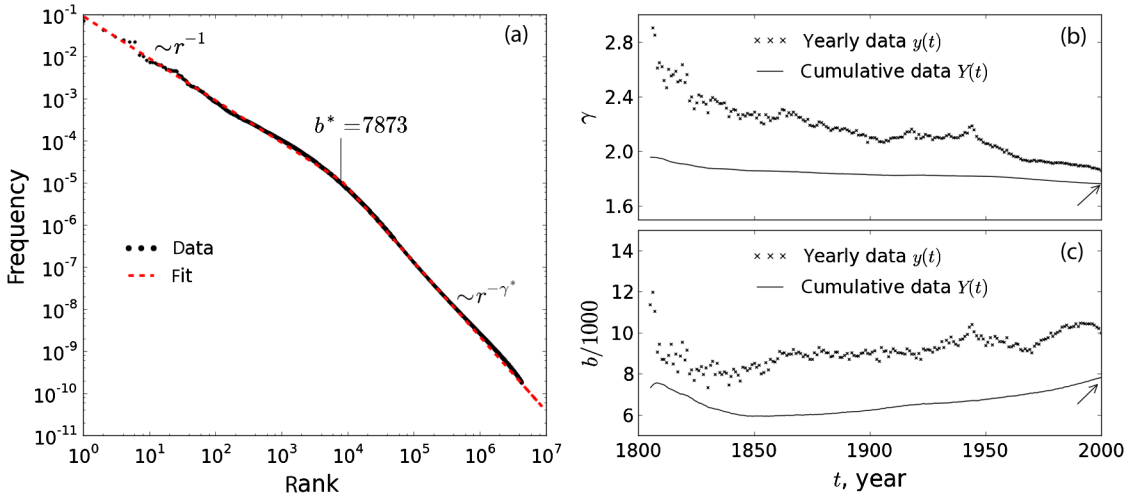

FIG. 1. The rank-frequency distribution shows double-scaling behavior (Zipf's plot). (a) Rank-frequency distribution for the English database $Y(2000)$ (solid line) and a ML fit of Eq. (1) (dashed line). (b,c) parameters γ and b obtained from ML fits of Eq. (1) to yearly $y(t)$ (x symbols) and accumulated $Y(t)$ (solid line) database. Arrows indicate the values of the parameters γ^* and b^* obtained for the fit in (a). Results are shown for the time range $t \in [1805, 200]$ in which data are most reliable; accumulation starts in $t_0 = 1520$.

in databases of different origin (see Ref. [32], Sec. IIB, for the detailed analysis on both databases, which uses methods reported in Refs. [43]).

We now discuss in detail the best two-parameter model that we identify from our data:

$$F_{\mathrm{dp}}(r;\gamma,b) = C\begin{cases} r^{-1} & r \le b \\ b^{\gamma-1}r^{-\gamma} & r > b, \end{cases} \quad (1)$$

characterizing a double power law (dp), where b and γ are free parameters, and $C = C(\gamma, b)$ is the normalization constant [44]. The effect of the threshold n applied to the frequency of words is that, in practice, data of $F(r)$ are limited to $F(r) \ge n/M$ (M is the observed number of words). Zipf's original law is recovered for high-frequency words, and a critical rank $r = b$ determines a transition to a power law with exponent γ. Double power laws were proposed as a generalization of Zipf's law in Ref. [45] and further investigated in Refs. [46,47]. These insightful works used distributions with two power-law exponents, γ_1, γ_2, and were motivated by the visual inspection of double logarithmic plots. Our improved statistical analysis confirms and extends these observations for the simpler distribution, Eq. (1). Besides the likelihood analysis and visual inspection given in Fig. 1, a third strong evidence in favor of distribution (1) comes from the comparison of the estimated parameters of different corpora shown in Figs. 1(b) and 1(c). Very similar values, $b \in [7 \times 10^3, 12 \times 10^3]$ and $\gamma \in [1.8, 2.5]$, were obtained for nonoverlapping databases (for the English Wikipedia, $b = 7830$, $\gamma = 1.68$), and the fluctuations become smaller for increasing database size. These observations strongly suggest that the same fixed parameters provide a good description of all English texts [e.g., $y(1900)$ and $y(2000)$]. Therefore, hereafter we do not consider individual fits for each database and instead assume that Eq. (1) is valid with $b = b^* = 7873$ and $\gamma = \gamma^* = 1.77$, values obtained for our largest database $Y(2000)$.

Similar findings also apply to the other languages. In Table I we summarize the parameters γ^* and b^* obtained from a ML fit of the largest database $Y(2000)$ of the respective language to Eq. (1). French and Spanish are also best described by Eq. (1) for databases exceeding a particular size, and they yield values for γ^* and b^* similar to English. For German and Russian, Eq. (1) constitutes only the second-best model. However, we have strong indications that Eq. (1) provides a better account of the tails ($r \gg b^*$), and, therefore, we expect that even larger databases will reveal the double power law as the best fit also in these languages (see Ref. [32], Sec. IIB, for details). Apart from being the smallest databases among the investigated languages, another feature affecting the fitting in German and, especially, in Russian is the higher degree of inflection in the morphology of these languages. We recall that no lemmatization was applied in our definition of words, and, therefore, inflected words (obtained, e.g., by adding a suffix) are counted as distinct words. This reasoning explains the higher measured values of b^* (vocabulary in the r^{-1} regime). From the fitting perspective, however, the large values of b^* in German and Russian require even larger databases to characterize the deviations from the r^{-1} regime for $r \gg b^*$.

TABLE I. Parameters b^*, γ^*, and $C^* = C(\gamma^*, b^*)$ obtained from the ML fit of Eq. (1) for the largest database $Y(2000)$ for all considered languages.

Language	b^*	γ^*	$C^* = C(\gamma^*, b^*)$
English	7873	1.77	0.0922
French	8208	1.78	0.0920
Spanish	8757	1.78	0.0915
German	19 863	1.62	0.0828
Russian	62 238	1.94	0.0789

C. Heaps' analysis

We now turn to our second empirical analysis: the dependence of the number of different words, N, on the size of the database, i.e., total number of words, M. The classical result for this relation is the empirical Heaps' law [12], which states that $N \sim M^{\lambda}$ with $\lambda \in [0, 1]$ ($A \sim B$ indicates that $A/B =$ constant for large B). We start searching for the implications of our finding of a generalized Zipf's law to the Heaps' analysis of vocabulary growth. A simple and powerful approach is the so-called Zipfian ensemble (ZE) [21], which can be traced [47] back to Mandelbrot [48]. This approach assumes that the occurrence of every possible word is governed by a Poisson process with an intensity proportional to its frequency (see Ref. [32], Sec. IIIA). It was shown that, under this or similar assumptions (e.g., stochastic processes with fixed frequencies for words), Heaps' law can be interpreted as a direct consequence of a Zipfian rank-frequency distribution $F(r) \sim r^{-\gamma}$ [9,13,14,19,21] and vice versa [20,49,50], where $\gamma = 1/\lambda$ [48]. Here, we want to draw attention to the fact that these observations are not restricted to Zipf's and Heaps' laws; i.e., assuming a stochastic model, the relationship between $F(r)$ and $N(M)$ can always be established. The expectation of the ZE of Eq. (1) with a threshold $n \gg 1$ is (see Ref. [32], Sec. IIIB)

$$N_{\mathrm{dp}}(M;\gamma,b) = C_n\begin{cases} M & M \ll M_b \\ M_b^{1-1/\gamma}M^{1/\gamma} & M \gg M_b, \end{cases} \quad (2)$$

where M_b is the number of words such that $N(M_b) = b$ and the scaling constant $C_n = C/n$ [$C \approx F(1)$ being the frequency of the most common word, as can be seen from Eq. (1)]. Thus, the effect of the threshold n applied to the growth curve of the vocabulary simply amounts to rescaling the constant C. While the expected (average) number of distinct words over many realizations of the stochastic process leads to a sharp transition between the two

regimes, the values of $N_{\mathrm{dp}}(M \approx M_b)$ might depend more strongly on the particular realization.

In Fig. 2, we show that the data in the Google Ngram database obey the scalings of Eq. (2). In Fig. 2(a), we present the $N(M)$ curve for English. While for the yearly database $y(t)$ we obtain a set of points for each t, the cumulative database $Y(t)$ builds a curve of vocabulary growth for increasing t. Despite the differences in these databases, all the data lie in a relatively narrow region of the plot which resembles a single curve compatible with the double scaling of Eq. (2). This curve is well described by the $N(M)$ curve obtained from the combination of the double power-law distribution Eq. (1) with fixed parameters (γ^*, b^*) and the assumption of Poisson usage of words, in the spirit of the ZE. Similar observations apply to all considered languages, as shown in Fig. 2(b). On closer inspection [see Fig. 2(c)], the fine details of the $N(M)$ curve are not compatible with the fluctuations expected from the strongly simplifying assumptions of the ZE. Nevertheless, it is remarkable that the agreement between the model and data remains within 50% for different databases and over 9 orders of magnitude in size.

Here, it is worth revisiting the question about the finitude of the vocabulary. Even after more than 10^6 different words, the $N(M)$ data in Fig. 2 do not seem to saturate. To further investigate this point, we perform the ZE with the same rank-frequency distribution from Eq. (1) (fixed b^*, γ^*) but varying the maximum possible number of different words $N_{\mathrm{ZE}}^{\max}$, e.g., 1, 2, 5, 10, and 100 times the observed number of distinct words in our largest database $Y(2000)$. It can be seen in Fig. 2(d) that the differences for the predicted growth curves for such different hypothetical vocabulary sizes are negligible compared to the fluctuations of the real data. From this, we conclude that, given the data accessible so far, the possible vocabulary can be regarded, for all practical purposes, to be infinite (although bounded by combinatorial arguments due to a finite alphabet and word length). The fact that the same distribution, Eq. (1), with fixed parameters accounts for the observation across all years shows that the observation of different numbers of

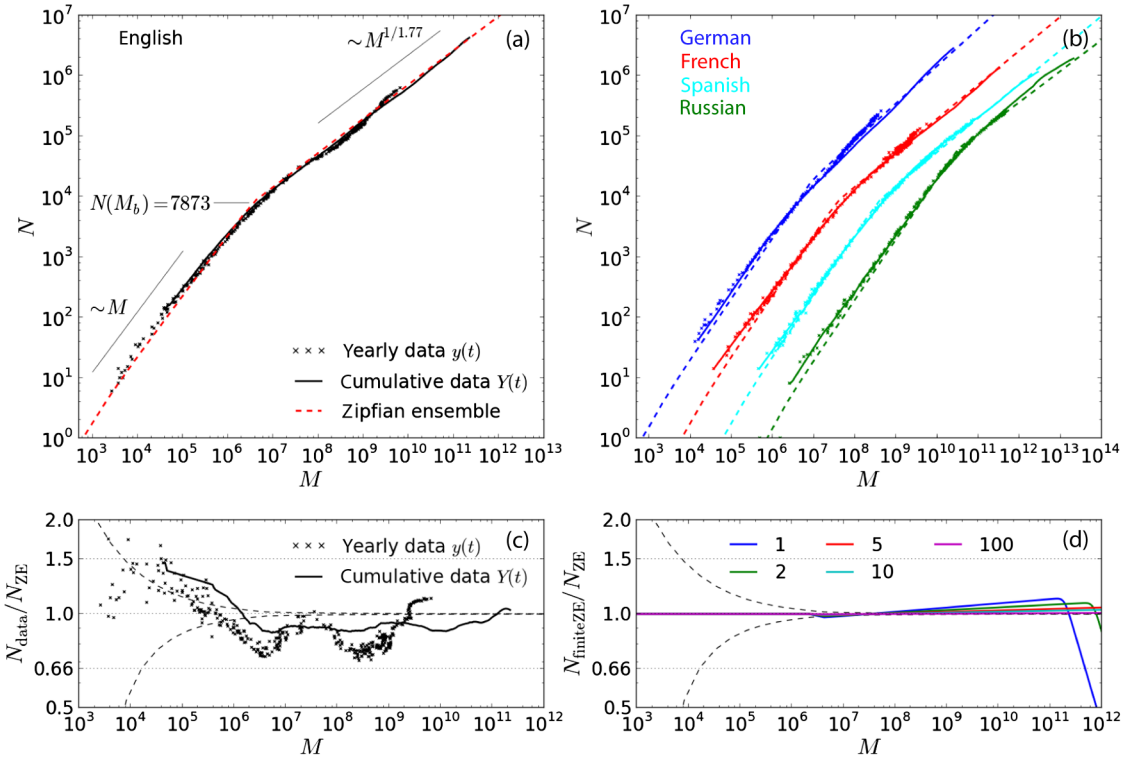

FIG. 2. Vocabulary N as a function of database size M (Heaps' plot). (a) Number of distinct words as a function of the number of words for yearly $y(t)$ (x symbols) database, cumulative $Y(t)$ (solid line) database, and the Zipfian ensemble (dashed line), assuming $n = 41$ and the rank-frequency distribution Eq. (1) with $b^* = 7873$ and $\gamma^* = 1.77$. (b) Same curves as in (a) but for different languages showing the same scaling behavior. In order to increase visibility, the curves for French, Spanish, and Russian were shifted, respectively, by one, two, and three decades with respect to their x values. (c) Difference of the curves in (a): deviation of the data $y(t)$ and $Y(t)$ (N_{data}) from the ZE growth curve (N_{ZE}). The dashed lines show the 95%-confidence interval of the ZE. (d) Deviation of a ZE growth curve with a hypothetically finite vocabulary ($N_{\mathrm{finite\,ZE}}$) from the ZE growth curve with infinite vocabulary (N_{ZE}) assuming rank-frequency distribution, Eq. (1). The possible size of the total vocabulary is given in units k of the number of observed distinct words in $Y(2000)$, such that $N_{\mathrm{ZE}}^{\max} = k \times 4\,263\,717$ with $k = 1, 2, 5, 10, 100$. Note that, for $M \to \infty$, $N_{\mathrm{finiteZE}}(M) \to N_{\mathrm{ZE}}^{\max}$, and therefore the deviation for $k = 1$ becomes visible.

words is driven mainly by the different database size and not by a change in vocabulary richness over time.

III. MODEL

In this section, we propose a simple generative model which recovers and allows for an improved interpretation of the double scalings in our empirical findings—Eqs. (1) and (2).

Our approach is different from Zipf's original explanation based on a principle of least effort between speakers and listeners [18,51] but, instead, is in line with the tradition of Yule-type stochastic growth models explaining fat-tailed distributions [26–29]. The main novelty in our model is that it contains two classes of word types: a core vocabulary and a noncore vocabulary [46]. At each step, a word (i.e., word token) is drawn ($M \mapsto M+1$) and attributed to one of the distinct words (i.e., word type) depending on probabilities specified below; see Fig. 3 for a sketch of the model. The total number of word types is given by $N = N_c + N_{\bar{c}}$, where ($N_{\bar{c}}$) N_c is the number of (non)core words. The new word token can either be a new word type ($N \mapsto N+1$) with a probability p_{new} or an already existing word type ($N \mapsto N$) with a probability $1 - p_{\text{new}}$. In the latter case, a (previously used) word type is attributed to the word token at random with a probability proportional to the number of times this word type has occurred before. In the former case, the new word type can either originate from a finite set of $N_c^{\max}$ core words ($N_c \mapsto N_c + 1$) with a probability p_c, or it can come from a potentially infinite set of noncore words ($N_{\bar{c}} \mapsto N_{\bar{c}} + 1$). In our simplest model, we consider p_c to be a constant, i.e., $p_c^0 \lesssim 1$, which becomes zero only if all core words were drawn ($N_c = N_c^{\max}$):

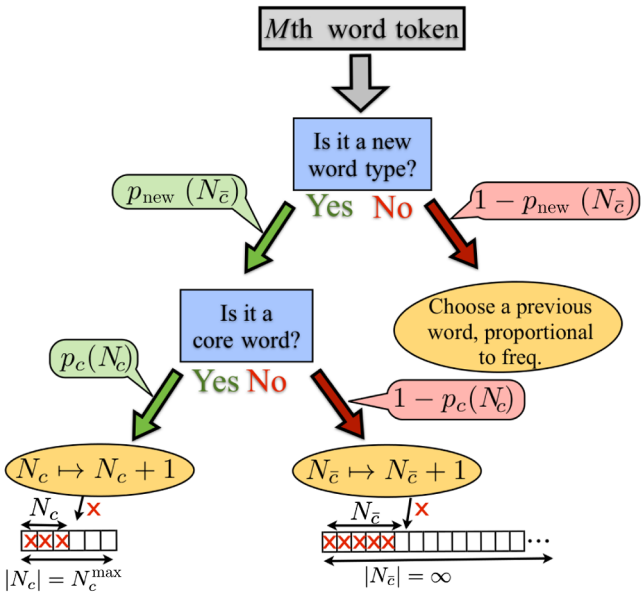

FIG. 3. Illustration of our generative model for the usage of new words.

$$p_c(N_c) = \begin{cases} p_c^0 & \text{if } N_c < N_c^{\max} \\ 0 & \text{if } N_c = N_c^{\max}. \end{cases} \tag{3}$$

The final element of our model, which establishes the distinguishing aspect of core words, is the dependence of p_{new} on N. We choose p_{new} (and p_c) to depend on N and not on M because an increase in N necessarily reflects that fewer undiscovered words exist, while an increase in M is strongly affected by repetitions of frequently used words. By definition, we think of core words as necessary in the creation of any text, and, therefore, the usage of a new core word in a particular text should be expected and thus not affect the probability of using a new (noncore) word type in the future, i.e., $p_{\text{new}} = p_{\text{new}}(N_{\bar{c}})$. On the other hand, if a noncore word is used for the first time ($N_{\bar{c}} \mapsto N_{\bar{c}} + 1$), the combination of this word and the previously used (core and noncore) words leads to a combinatorial increase in possibilities of expression of new ideas with the already used vocabulary and thus to a decrease in the marginal need for additional new words [47]. In our model, this argument suggests that p_{new} should decrease with $N_{\bar{c}}$. Taking these factors into account, we propose as an update rule for p_{new} after each occurrence of a new noncore word

$$p_{\text{new}} \mapsto p_{\text{new}}\left(1 - \frac{\alpha}{N_{\bar{c}} + s}\right), \tag{4}$$

with the decay rate $\alpha > 0$ and the constant $s \gg 1$, which is introduced simply in order to damp the reduction of p_{new} for small $N_{\bar{c}}$ (for simplicity, we use $s = N_c^{\max}$). The main justification for the exact functional form in Eq. (4) is that it allows us to recover the empirical observations reported in Sec. II, as shown below. An alternative *a posteriori* justification will be given at the end of this section, and it shows that Eq. (4) can be interpreted as a direct consequence of an unlimited noncore vocabulary.

We now show how this model recovers Eqs. (1) and (2). We require that $1 - p_c^0 \ll 1$, which simply means that it is much more likely to draw core words than noncore words initially. In this case, we can obtain approximately exact solutions for $N(M)$ in the two limiting cases considered in Eq. (2). When $N \ll N_c^{\max}$, which implies $N_c, N_{\bar{c}} \ll N_c^{\max}$, it follows from Eqs. (3) and (4) that $p_{\text{new}} \approx \text{const}$, and therefore we trivially obtain that $N \sim M^1$. This case resembles the very beginning of the vocabulary growth, when most new word types belong to the set of core words. In the case $N \gg N_c^{\max}$, $p_c = 0$ and $N \approx N_{\bar{c}}$, so Eq. (4) becomes, in the continuum limit,

$$\frac{\mathrm{d}}{\mathrm{d}N} p_{\text{new}}(N) = -\alpha \frac{p_{\text{new}}(N)}{N}, \tag{5}$$

from which it follows that $p_{\text{new}} \sim N^{-\alpha}$.

We now obtain the expected growth curve $N(M)$. Notice that our model can be considered a biased random walk in N, which, as an approximation, can be mapped onto a

binomial random walk by the coordinate transformation $N(M)$ such that $p_{\text{new}}(N) = p_{\text{new}}[N(M)]$. The resulting Poisson-binomial process [52] can be treated analytically; e.g., the transformation $N(M)$ is then given by the average of the vocabulary growth:

$$N(M) = \int_0^M \mathrm{d}M' p_{\text{new}}(M') = \int_{N(0)}^{N(M)} \mathrm{d}N' \left| \frac{\mathrm{d}M'}{\mathrm{d}N'} \right| p_{\text{new}}(N'). \quad (6)$$

Using $p_{\text{new}} \sim N^{-\alpha}$, this equation holds (self-consistently) by assuming a sublinear growth for the vocabulary $N \sim M^{\lambda}$, where the relation $\lambda = (1+\alpha)^{-1}$ is established (for details see Ref. [32], Sec. IV). In accordance with Eq. (2), we identify the following relation between the parameters: $N_c^{\max} = b$ and $\alpha = \gamma - 1$. The fitting parameters of Eq. (1) can thus be interpreted as follows: b is the size of the core vocabulary, and γ controls the sensitivity of the probability of using a new word to the number of already-used words in Eq. (5).

Since the probability of usage for already-used word types is assumed to be proportional to the number of times it occurred before, we guarantee that Eq. (2) implies (1) [20], meaning that the double scaling in the Zipf plot is also recovered from our generative model. While the previous arguments show that the correct scalings are obtained by our model, in order to obtain an agreement with the data, it is essential to (i) use the normalization constant C in order to determine the initial probability of finding a new word in Eq. (4), (ii) rescale the distribution using the threshold n as M/n, and (iii) account for the disproportionally large weight of the first word types (in the Zipf plot). Taking these points into account, direct simulations of the model in Fig. 3 with the traditional parameters $b = b^*$ and $\gamma = \gamma^*$ lead to Zipf's and Heaps' curves, which resemble the original fits. See Ref. [32], Sec. V, for all details.

It is worth comparing the generative model with the model of random usage of words with fixed frequency, the ZE model discussed in the previous section. While the ZE model allowed us to obtain Heaps' curves from Zipf's distributions (and vice versa), in the generative model, we simultaneously obtain the double-scaling regime in both cases. It is important to stress that individual texts or single databases should not be considered as the output of single realizations of our generative model. Instead, we consider that not only texts but also all databases have a negligible size when compared to the language as a whole and therefore should be thought of as a small subsample ($M_{\text{database}} \ll M$) of the output of our generative model, retrieved after the model achieved its stationary state ($M \to \infty$). In this case, changes in word frequencies become negligible (at the scale of M) during the creation of the database (at the scale of M_{database}). Therefore, the vocabulary growth of the created database is well approximated by the ZE model with $F_{\text{dp}}(r)$.

Finally, we use our previous calculations and provide an *a posteriori* justification of the key assumption of our model, Eq. (4). Our starting point is the observation—see Fig. 2(d)—that vocabulary is, for all practical purposes, infinite. We therefore postulate that

$$N(M) \overset{M\to\infty}{\longrightarrow} \infty, \quad (7)$$

and, by following (in reverse order) the previous calculations, we naturally arrive at Eq. (4). From the first line of Eq. (6), we see that in order to fulfill our postulate (7), p_{new} has to decay at least as slow as $p_{\text{new}}(M) \sim M^{-\delta}$, with $\delta \leq 1$ for $M \to \infty$. In a minimal model, it is reasonable to assume such a power-law decay, in which case the first line of Eq. (6) implies that $N(M) \sim M^{\lambda}$, with $\lambda = 1 - \delta$. Making a transformation of variables from M to N, we obtain

$$p_{\text{new}}(N) = p_{\text{new}}(M(N)) \sim N^{-1+(1/\lambda)} = N^{-\alpha}. \quad (8)$$

In turn, this is equivalent to Eq. (5), from which we recover Eq. (4) as a discretized version. Thus, we see that Eq. (4) is a minimal assumption for an unbounded vocabulary.

IV. HISTORICAL CHANGES

The model described so far has been shown to give a good account of all databases and all years with the same two fixed parameters, $N_c^{\max} = b^* = 7,873$ and $\alpha = \gamma^* - 1 = 0.77$ in the case of English. A natural question is, therefore, what actually changes in historical time scales? Considering two different databases (say, two different years), our model does not consider any differences in the actual composition of the core vocabulary. Even if the value of $N_c^{\max}$ remains constant, this does not mean that the *same* words are observed for all years. From the point of view of our model, the main change a word can experience is to enter or to leave the group of core words. For instance, comparing the decades 1891–1900 and 1991–2000, the most frequent words that left the core vocabulary were *majesty*, *doubtless*, *furnished*, *monsieur*, *Napoleon*, and *hitherto*, while the ones that entered were *cultural*, *context*, *technology*, *programs*, *environmental*, and *computer* [53].

In order to quantify this effect, we investigate the replacement of words from the core vocabulary in the yearly databases $y(t)$ in the time $t \in [1805, 2000]$ in Fig. 4. We calculate the fraction $f(t, \Delta t)$ of core words (i.e., with rank $r < b^* = 7873$, fixed for all t) from $y(t)$ that remain in the set of core words in $y(t + \Delta t)$. Figure 4(a) shows that all curves can be qualitatively described by an exponential decay

$$f(t, \Delta t) = f_0 e^{-\kappa|\Delta t|}, \quad (9)$$

independent of whether forward ($\Delta t > 0$) or backward time ($\Delta t < 0$) was considered. This is further supported in Figs. 4(b) and 4(c), where the parameters f_0 and κ

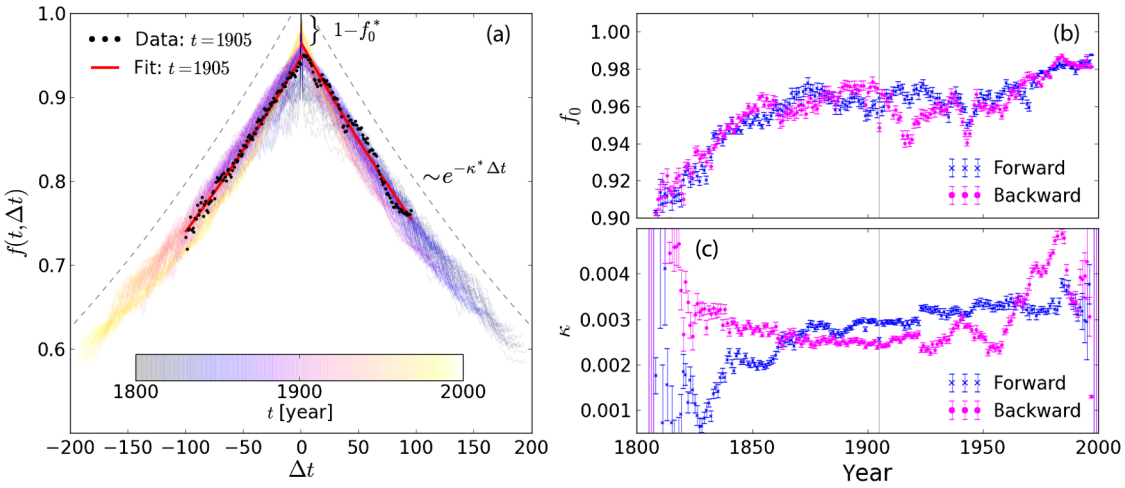

FIG. 4. Historical change in the composition of core words in the English vocabulary. (a) Fraction $f(t, \Delta t)$ of core words in $y(t)$ that remain in the set of core words in $y(t + \Delta t)$ for $t \in [1805, 2000]$ (pale colors) and, in particular, for $t = 1905$ (black dots) with the corresponding exponential fit (red line). (b,c) Parameters f_0 and κ in the exponential decay Eq. (9) of the curves in (a) obtained through least-square fits. Forward (backward) decay refers to $\Delta t > 0$ ($\Delta t < 0$).

obtained numerically from a least-square fit [38] of Eq. (9) for all curves $f(t, \Delta t)$ with $t \in [1805, 2000]$ are presented. In order to avoid biases due to different number of points in the fit, for each t we performed a fit with the same number of points $\min\{2000 - t, t - 1805\}$ forwards and backwards in time. On closer inspection, two features connected to the interpretation of the parameters f_0 and κ deserve a more careful discussion. The parameter $f_0 < 1$ represents the discontinuous change of core words in two subsequent years. It strongly depends on the different selection of books in the construction of the respective databases and can be attributed to the finite size of the database, which leads to a wrong estimation of the "true" core words. Consistently with this interpretation, Fig. 4(b) shows that f_0 grows over time, due to the fact that database size increases, leading to a better sampling of words. Nevertheless, a value of $f_0 \approx 0.98$ indicates that this is still far from being negligible (e.g., for $N_c^{\max} = 7,873$, this means that around 150 words of the set of core words will be different because of finite sampling). In contrast, the decay rate κ describes the continuous replacement of core words over time, with a rate of $\kappa N_c^{\max} \approx 30$ words per year. The most intriguing observation in Fig. 4(c) is that this change experiences an acceleration over time, as κ grows by more than 50% from 1805 to 2000.

Finally, it is worth discussing the implications of these findings on our generative model. The characteristic time scale of the core-vocabulary replacement ($\approx 1/\kappa$) is on the order of centuries. This means that, on the scale of a few decades, our generative model holds with the assumption of a constant core vocabulary. On longer time scales, our model has to be refined in order to include (i) a probability of replacement of the words belonging to the core vocabulary and (ii) a finite memory or a distinction between core and noncore words in the preferential attachment part of our model.

V. DISCUSSION

In summary, we have shown that the rank-frequency distribution and the vocabulary growth of languages can be best described by simple two-scaling functions. The only two free parameters of the functions are related to each other and remain almost unchanged over centuries, as well as databases, and depend only on the considered language. We have also shown that these empirical findings can be interpreted as the result of a finite number of words belonging to a core vocabulary, which have different properties from the remaining virtually unlimited number of words, as summarized in Table II. This conclusion was achieved based on a simple generative stochastic model for the vocabulary growth. Finally, we found that in English, the composition of the core vocabulary experiences an exponential decay with a rate of 30 words per year, which has been, remarkably, steadily accelerating in the past decades.

It is worth comparing these findings in view of previous results. As far as we are aware, our analysis provides the first rigorous statistical confirmation of similar previous

TABLE II. Properties of core (c) and noncore ($\bar{c}$) words in our model.

	Core words	Noncore words
Number	Finite: $N_c^{\max} \in [10^3, 10^4]$	Infinite: $N_{\bar{c}} \to \infty$
Frequency	Larger ($r > b^*$)	Smaller ($r < b^*$)
Effect on p_{new}	None	Reduction

proposals [45–47] of the double-scaling generalizations of Zipf's law—Eq. (1). The consequence of this on vocabulary growth and Heaps' law (see also [47]), which we drew based on a Poisson usage of words [21], is that the rate of introduction of new words decays but never vanishes with increasing database size. This is in contrast to recent claims that reported a convergence to a maximum vocabulary size [14]. We note that this previous analysis was based on single books, and therefore the database sizes were close to our transition point $N_c^{\max}$, which we believe could have been misinterpreted as a systematic decay. A generalization of a Yule-type process to obtain double-scaling degree distribution in a network of words was introduced in Ref. [54]. Two crucial differences from our model are that it yields fixed exponents and cannot be understood as a generative model of texts (word by word). Interestingly, in Ref. [6] an analysis of the network constructed from the thesaurus also showed the existence of a set of core words of almost the same size as ours.

Our simple model and expression for the vocabulary growth as a function of database size have important practical consequences. Simply knowing the database size (in number of words, M, or potentially in bits), and using the language-dependent parameters (C, $N_c^{\max} = b^*$, $\alpha = \gamma^* - 1$) reported above, from Eq. (2) one can immediately estimate the expected number of different words, N, appearing more than n times. This is crucial for search engines and data-mining programs because it allows for an estimation of the memory to be allocated prior to the scanning of an unknown database, e.g., in the construction of the inverted index [9–11]. Even the fluctuations around this expectation can be easily computed through our generative model or through the Poisson assumption of word usage. Of course, this strong assumption ignores correlations and typically underestimates the expected fluctuations, so our model should be considered as the simplest null model. The existence of a transition between two scalings (which is within reach of even single, large books) shows that simple estimations based only on the traditional Zipf's law have to be generalized. For instance, a commonly used index of vocabulary richness of a text is Herdan's coefficient given by the ratio $\log N/\log M$ [8]. In view of our results, the coefficient is highly dependent on which of the two scaling regimes is reached with the given size of the text.

We now compare our observations of change on historical time scales to other historical changes in language usage. For the whole vocabulary, we obtain that the vocabulary size is mainly driven by the available database size. This is in contrast to previous conclusions based on the same Google Ngram database that detected a growth of vocabulary in time [1]. Here, it is important to note that this previous analysis included a substantially different filtering of the listed 1-grams to achieve valid words in the vocabulary, including a frequency criterion and manual classification. Still, our results show that, also in this case, a more careful analysis of the role of the database size is needed. For the core vocabulary, we observe a fairly constant number of constituents over centuries. The number of words common to core vocabularies of different databases was found to decay exponentially with the time between publication of the databases; e.g., for English, the decay rate is approximately 30 words per year and the half-life of the core vocabulary is approximately 200 years. It is worth comparing these numbers with recent studies that reported half-lifes for (i) the regularization of verbs (750 to 10 000 years) [5] and (ii) a fundamental vocabulary of 200 words (300 to 38 000 years) [4]. Perhaps our most intriguing finding is the approximately linear increase of the rate in time, which eventually confirms the overall acceleration of language change and society in general, as propagated in Ref. [1].

Our results can be extended in many directions and open new possibilities of studies of vocabulary change. Directly related to our observations and model, the specific value of the parameter $\gamma^* \approx 1.77$ still needs to be explained, which is intriguingly similar across different languages. Another important point is to assess the limitations of our estimations due to the role of correlations inside real texts and databases, and to determine how this could be introduced into our model. Furthermore, it remains to be shown whether the transition between two scalings due to the existence of a core vocabulary can be related to the phenomenon of phase transitions in ranking stability of complex systems recently reported in Ref. [55]. Finally, we believe that our model provides the correct null model for normalizations due to database sizes and that, therefore, future investigations of historical effects on the vocabulary should take this into account.

ACKNOWLEDGMENTS

We are indebted to J. Miotto and R. Guimerà for valuable discussions about the data analysis. We thank F. Ghanbarnejad and J. Leitão for the careful reading of the manuscript.

[1] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, *Quantitative Analysis of Culture Using Millions of Digitized Books*, Science **331**, 176 (2011).

[2] J. Gao, J. Hu, X. Mao, and M. Perc, *Culturomics Meets Random Fractal Theory: Insights into Long-Range Correlations of Social and Natural Phenomena over the Past Two Centuries*, J. R. Soc. Interface **9**, 1956 (2012).

[3] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, *Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death*, Sci. Rep. **2**, 313 (2012).

[4] M. Pagel, Q. D. Atkinson, and A. Meade, *Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History*, Nature (London) **449**, 717 (2007).

[5] E. Lieberman, J.-P. Michel, J. Jackson, T. Tang, and M. A. Nowak, *Quantifying the Evolutionary Dynamics of Language*, Nature (London) **449**, 713 (2007).

[6] D. Levary, J.-P. Eckmann, E. Moses, and T. Tlusty, *Loops and Self-Reference in the Construction of Dictionaries*, Phys. Rev. X **2**, 031018 (2012).

[7] G. Wimmer and G. Altmann, *Review Article: On Vocabulary Richness*, J. Quant. Linguist. **6**, 1 (1999).

[8] R. H. Baayen, *Word Frequency Distributions* (Kluwer Academic Publishers, Dordrecht, Netherlands, 2001).

[9] R. Baeza-Yates and G. Navarro, *Block Addressing Indices for Approximate Text Retrieval*, J. Am. Soc. Inf. Sci. **51**, 69 (2000).

[10] H. E. Williams and J. Zobel, *Searchable Words on the Web*, Int. J. Digit. Libr. **5**, 99 (2005).

[11] B. Croft, D. Metzler, and T. Strohmann, *Search Engines: Information Retrieval in Practice* (Addison-Wesley, Boston, MA, 2009).

[12] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, New York, 1978).

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

[14] S. Bernhardsson, L. E. Correa da Rocha, and P. Minnhagen, *The Meta Book and Size-Dependent Properties of Written Language*, New J. Phys. **11**, 123015 (2009).

[15] Y. Sano, H. Takayasu, and M. Takayasu, *Zipf's Law and Heaps' Law Can Predict the Size of Potential Words*, Prog. Theor. Phys. Suppl. **194**, 202 (2012).

[16] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, *Collective Dynamics of Social Annotation*, Proc. Natl. Acad. Sci. U.S.A. **106**, 10511 (2009).

[17] R. W. Benz, S. J. Swamidass, and P. Baldi, *Discovery of Power-Laws in Chemical Space*, J. Chem. Inf. Model. **48**, 1138 (2008).

[18] G. K. Zipf, *The Psycho-Biology of Language* (Routledge, London, England, 1936).

[19] D. C. van Leijenhorst and T. P. van der Weide, *A Formal Derivation of Heaps' Law*, Information Sciences (NY) **170**, 263 (2005).

[20] D. H. Zanette and M. A. Montemurro, *Dynamics of Text Generation with Realistic Zipf's Distribution*, J. Quant. Linguist. **12**, 29 (2005).

[21] I. Eliazar, *The Growth Statistics of Zipfian Ensembles: Beyond Heaps' Law*, Physica (Amsterdam) **390**, 3189 (2011).

[22] M. A. Montemurro, *Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics*, Physica (Amsterdam) **300**, 567 (2001).

[23] W. Li, P. Miramontes, and G. Cocho, *Fitting Ranked Linguistic Data with Two-Parameter Functions*, Entropy **12**, 1743 (2010).

[24] G. Jäger, *Power Laws and Other Heavy-Tailed Distributions in Linguistic Typology*, Adv. Compl. Syst. **15**, 1150019 (2012).

[25] M. P. H. Stumpf and M. A. Porter, *Critical Truths About Power Laws*, Science **335**, 665 (2012).

[26] M. E. J. Newman, *Power Laws, Pareto Distributions and Zipf's law*, Contemp. Phys. **46**, 323 (2005).

[27] G. U. Yule, *A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.*, Phil. Trans. R. Soc. B **213**, 21 (1925).

[28] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Log-normal Distributions*, Internet Math. **1**, 226 (2004).

[29] M. V. Simkin and V. P. Roychowdhury, *Re-inventing Willis*, Phys. Rep. **502**, 1 (2011).

[30] C. Castellano, S. Fortunato, and V. Loreto, *Statistical Physics of Social Dynamics*, Rev. Mod. Phys. **81**, 591 (2009).

[31] A. Baronchelli, V. Loreto, and F. Tria, *Language Dynamics*, Adv. Compl. Syst. **15**, 1203002 (2012).

[32] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.3.021006 for further details.

[33] B. Mandelbrot, An Informational Theory of the Statistical Structure of Language, *Communication Theory* (Butterworth, Woburn, MA, 1953), p. 486.

[34] J. Tuldava, *The Frequency Spectrum of Text and Vocabulary*, J. Quant. Linguist. **3**, 38 (1996).

[35] S. K. Baek, S. Bernhardsson, and P. Minnhagen, *Zipf's Law Unzipped*, New J. Phys. **13**, 043004 (2011).

[36] A. Cohen, R. N. Mantegna, and S. Havlin, *Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts*, Fractals **05**, 95 (1997).

[37] R. Ferrer i Cancho, *The Variation of Zipf's Law in Human Language*, Eur. Phys. J. B **44**, 249 (2005).

[38] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes* (Cambridge University Press, Cambridge, England, 2007), 3rd ed.

[39] R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit-Techniques* (Marcel Dekker, New York, 1986).

[40] J. R. Taylor, *An Introduction to Error Analysis* (University Science Books, Sausalito, CA, 1997).

[41] H. Akaike, *A New Look at the Statistical Model Identification*, IEEE Trans. Autom. Control **19**, 716 (1974).

[42] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, New York, 2002), 2nd ed.

[43] F. Johansson *et al.*, mpmath: A Python Library for Arbitrary-Precision Floating-Point Arithmetic (Version 0.14), 2010, http://code.google.com/p/mpmath/; E. Jones *et al.*, SciPy: Open Source Scientific Tools for Python, 2001, http://www.scipy.org/; Wikimedia: Dump of the English Wikipedia on 02/06/2012, http://dumps.wikimedia.org/enwiki/20120601/; University of Pisa Multimedia Lab: Wikipedia Extractor, http://medialab.di.unipi.it/wiki/Wikipedia_Extractor. See Ref. [32] for access dates.

[44] The sharp transition between the two regimes in Eq. (1) might seem artificial. We believe that alternative distributions which interpolate between the two scalings could provide a similarly good account of the data. The advantage of the distribution Eq. (1) is that the transition point $r = b$ appears explicitly as a free parameter and can be independently estimated from data.

[45] S. Naranan and V. Balasubrahmanyan, *Models for Power Law Relations in Linguistics and Information Science*, J. Quant. Linguist. **5**, 35 (1998).

[46] R. Ferrer i Cancho and R. V. Solé, *Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited*, J. Quant. Linguist. **8**, 165 (2001).

[47] A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, *Languages Cool as They Expand: Allometric Scaling and the Decreasing Need for New Words*, Sci. Rep. **2**, 943 (2012).

[48] B. Mandelbrot, *On the Theory of Word Frequencies and on Related Markovian Models of Discourse*, Structure of Language and Its Mathematical Aspects: Proceedings of Symposia in Applied Mathematics Vol. XII (American Mathematical Society, Providence, RI, 1961).

[49] H. A. Simon, *On a Class of Skew Distribution Functions*, Biometrika **42**, 425 (1955).

[50] A. P. Masucci, A. Kalampokis, V. M. Eguíluz, and E. Hernández-García, *Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space*, PLoS ONE **6**, e17333 (2011).

[51] B. Corominas-Murtra, J. Fortuny, and R. V. Solé, *Emergence of Zipf's Law in the Evolution of Communication*, Phys. Rev. E **83**, 036115 (2011).

[52] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1968), 3rd ed., Vol. I.

[53] These examples are the six most frequent words which belong to the core vocabulary (i.e., $r < b^* = 7873$) in every single year in one decade and in none of the years in the other decade (ordered by the average frequency in the decade in which they belonged to the core vocabulary).

[54] S. N. Dorogovtsev and J. F. F. Mendes, *Language as an Evolving Word Web*, Proc. R. Soc. Lond. B Biol. Sci. **268**, 2603 (2001).

[55] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, and A.-L. Barabási, *Dynamics of Ranking Processes in Complex Systems*, Phys. Rev. Lett. **109**, 128701 (2012).