Large-scale analysis of Zipf's law in English texts

Isabel Moreno-Sánchez, Francesc Font-Clos, Álvaro Corral[1, 2]

[1]*Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*
[2]*Departament de Matemàtiques, Facultat de Ciències,*
Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain
(Dated: September 14, 2015)

Despite being a paradigm of quantitative linguistics, Zipf's law for words suffers from three main problems: its formulation is ambiguous, its validity has not been tested rigorously from a statistical point of view, and it has not been confronted to a representatively large number of texts. So, we can summarize the current support of Zipf's law in texts as anecdotic. We try to solve these issues by studying three different versions of Zipf's law and fitting them to all available English texts in the Project Gutenberg database (consisting of more than 30 000 texts). To do so we use state-of-the art tools in fitting and goodness-of-fit tests, carefully tailored to the peculiarities of text statistics. Remarkably, one of the three versions of Zipf's law, consisting of a pure power-law form in the complementary cumulative distribution function of word frequencies, is able to fit more than 40 % of the texts in the database (at the 0.05 significance level), for the whole domain of frequencies (from 1 to the maximum value) and with only one free parameter (the exponent).

PACS numbers:

INTRODUCTION

Zipf's law constitutes a striking quantitative regularity in the usage of language [1–4]. It states that, for a large enough piece of text, the frequency of use n of any word decreases with its rareness r in the text in an approximately hyperbolic way, i.e., $n \propto 1/r$, where the symbol "$\propto$" denotes proportionality. Technically, r is called the rank, and the most common (i.e., less rare) word is assigned $r = 1$, the second most common, $r = 2$, and so on. A slightly more general formulation includes a parameter in the form of an exponent α; then, the rank-frequency relation takes the form of a power law,

$$n \propto \frac{1}{r^\alpha}. \qquad (1)$$

with the value of α close to one.

This pattern (1) has been found across different languages, literary styles, time periods, and levels of morphological abstraction [2, 5, 6]. More fascinatingly, the same law has been claimed in other codes of communication, as in music [7] or for the timbres of sounds [8], and also in disparate discrete systems where individual units or agents gather into different classes [9], for example, employees into firms [10], believers into religions [11], insects into plants [12], units of mass into animals present in ecosystems [13], visitors or links into web pages [14], telephone calls to users [15], or abundance of proteins (in a single cell) [16]. The attempts to find an explanation have been diverse [3, 15, 17–24], but no solution has raised consensus [20, 25, 26].

Despite its quantitative character, Zipf's law has been usually checked for in a qualitative way, plotting the logarithm of the frequency n versus the logarithm of the rank r and looking for some domain with a roughly linear behavior, with slope more or less close to −1. A more refined approach consists in fitting a straight line to the double-logarithmic plot by linear regression [27]. But several authors have recently pointed out the limitations of this method when applied to probability distributions [11, 28–30], and the advantages of using an asymptotically unbiased and minimum-variance procedure such as maximum likelihood (ML) estimation [30], whose solutions, moreover, are invariant under reparameterizations [31, 32]. One should consider then ML estimation as the most reliable procedure of estimation for parametric models (when a maximum of the likelihood does exist and the number of data is large).

Furthermore, for the particular case of linguistics, the search for Zipf's law has been traditionally performed in very limited sets of texts (less than a dozen in a typical research article [6, 33]). More recently, however, large corpora have been considered –these are representative collections of different texts aggregated together into a single bag, so, instead of many separated texts one deals with one enormous mixed text. When "rare" words are not considered (i.e., for high frequencies), it seems that Zipf's law still holds in these large collections [2, 34–37]. At present, there is agreement that Zipf's law is a rough approximation in lexical statistics, but its range of validity is totally unknown, i.e., we ignore how good Zipf's law is in order to account for the appearance of words, and for which texts it should work –and with which level of precision– and for which texts it should fail.

An extra difficulty emerges when one recognizes the ill-defined nature of Zipf's law. In fact, the law has two formulations, with the first one being Eq. (1), which just counts the frequency of words. For the sake of clarity, the words that are counted are referred to as word types, in order to distinguish them from each repetition, which is called a token. The second formulation of Zipf's law arises when, after counting the frequency of word types,

one performs a second statistics and counts how many values of the frequency are repeated, that is, how many word types have the same frequency. This means that the frequency n is considered the random variable. One can realize that the rank, when normalized by its maximum value in text, is just the empirical estimation of the complementary cumulative distribution function of n, and then, the derivative of the expression for $r(n)$ (the inverse of Eq. (1)) yields a continuous approximation for the probability mass function $f(n)$ of the frequency n. From here one obtains another power law,

$$f(n) \propto \frac{1}{n^\beta}, \tag{2}$$

with the new exponent β fulfilling $\beta = 1 + 1/\alpha$, which yields values of β close to 2. The expression given by Eq. (2) was in fact the first approach followed by Zipf's himself [3], and is usually considered as equivalent to Eq. (1) [3, 9, 14, 15, 33]; however, as it is derived in the continuum limit, both expressions can only be equivalent asymptotically, for large n [38]. Consequently, if one wants to be precise, a natural question follows: which one is the "true" Zipf's law (if any)?

We cannot know a priori which of the two Zipf's laws better describes real texts, but we can argue which of the two representations (that of $n(r)$, Eq.(1), or that of $f(n)$, Eq.(2)) is better for statistical purposes, independently of the functional dependency they provide. It is clear that the rank-frequency representation, given by $n(r)$, presents several difficulties, due to the peculiar nature of the rank variable. First, in Ref. [39], Zipf-like texts were randomly generated following Eq. (1), keeping the original ranks "hidden" (as it happens in the real situation), and it was found that the rank reconstructed from the sample deviated considerably from the original ranks when these were taking large values (which for a power law happens with a high probability). The resulting ML estimations of the exponent α were highly biased and the Kolmogorov-Smirnov test rejected the power-law hypothesis, although the original ranks were power-law indeed.

One might argue that the problem could be escaped by using an upper truncated power-law distribution (introducing then an additional parameter for the truncation), in order to avoid the inconsistency of the rank representation for high values. But a second problem is that the rank is not a true random variable [40], as its values are assigned a posteriori, once the sample (i.e., the text) is analyzed. This means that the rank does not show "enough" statistical fluctuations, that is, if $r_a < r_b$, then the frequency of a is always larger, by construction, than the frequency of b. This does not necessarily happen for a true random variable. The negative correlation between the variable and its frequency of occurrence makes the power-law hypothesis harder to reject. In fact, inflated p-values (not uniformly distributed between 0 and

1) have been found when fitting truncated power laws to simulated power-law rank-frequency representations [39]. This problem could still be avoided by choosing a low enough upper truncation parameter (yielding a very short range of ranks, for which the fluctuations would be very little) but at the expense of disregarding an important part of the data.

A third inconvenience is the impossibility, due to normalization, that a non-truncated power law comprises values of the $\alpha-$exponent smaller than 1. This yields the necessity of introducing a truncation parameter that may be artifactual, i.e., not present in the real system. All this leads to the conclusion that the most reliable method of parameter estimation (ML, in a frequentist framework) cannot be directly applied to the rank-frequency representation. In contrast, the representation in terms of the distribution of frequencies is devoid of these problems [39], as n is a well-defined random variable, and this will be the representation used in this paper for statistical inference. Nevertheless, for alternative arguments, see Ref. [41].

The purpose of this paper is to quantify, at a large, big-data scale, different versions of Zipf's law and their ranges of validity. In the next section, we present and justify the three Zipf-like distributions we are going to fit, and we briefly explain the selected fitting method and the goodness-of-fit test. The corpus of texts under consideration is also detailed. The following section presents the results, with special attention to their statistical significance and their dependence with text length. Finally, we end with the conclusions and some technical appendices.

ZIPF-LIKE DISTRIBUTIONS

As implicit in the introduction, and in contrast with continuous random variables, in the discrete case a power law in the probability mass function $f(n)$ does not lead to a power law in the complementary cumulative distribution or survival function $S(n)$, and vice-versa. Let us specify our definition for both functions, $f(n) = \text{Prob[frequency} = n]$ (as usual), and $S(n) = \text{Prob[frequency} \geq n]$ (changing, for convenience, the usual strict inequality sign by the non-strict inequality). Then, the relation between both is $f(n) = S(n) - S(n+1)$ and $S(n) = \sum_{n'=n}^{\infty} f(n')$.

We consider that the values the random variable takes, given by n, are discrete, starting at the integer value a, taking values then $n = a, a+1, \ldots$ up to infinity. In this study we will fix the parameter a to $a = 1$, in order to fit the whole distribution and not just the tail. Then, although for large n and smooth $S(n)$ we may approximate $f(n) \simeq -dS(n)/dn$, this simplification, which lies in the equivalence between Eqs. (1) and (2), is clearly wrong for small n.

For the first distribution that we consider, the power-law

form is in $f(n)$, then,

$$f_1(n) = \frac{1}{\zeta(\beta, a)n^\beta}. \tag{3}$$

This is just the normalized version of Eq. (2), and then,

$$S_1(n) = \frac{\zeta(\beta, n)}{\zeta(\beta, a)}$$

with $\beta > 1$ and $\zeta(\beta, a) = \sum_{k=0}^\infty (a + k)^{-\beta}$ denotes the Hurwitz zeta function, which ensures normalization of both expressions. A preliminary analysis in terms of this distribution was done in Ref. [42]. In contrast, when the power law is in $S(n)$, this leads to our second case,

$$f_2(n) = \left(\frac{a}{n}\right)^{\beta-1} - \left(\frac{a}{n+1}\right)^{\beta-1} \tag{4}$$

and

$$S_2(n) = \left(\frac{a}{n}\right)^{\beta-1}$$

with $\beta > 1$ again. Note that this corresponds to a power law in the empirical rank-frequency relation.

Finally, it is interesting to consider also the frequency distribution derived by Mandelbrot [38] when ranks are generated independently from a power law in Eq. (1), which is,

$$f_3(n) = \frac{(\beta-1)\Gamma(a)}{\Gamma(a+1-\beta)} \frac{\Gamma(n+1-\beta)}{\Gamma(n+1)} \tag{5}$$

and

$$S_3(n) = \frac{\Gamma(a)\Gamma(n+1-\beta)}{\Gamma(n)\Gamma(a+1-\beta)},$$

with $1 < \beta < 2$ and $\Gamma(\gamma) = \int_0^\infty x^{\gamma-1}e^{-x}dx$ denotes the gamma function [43]. In this case the power law is the underlying theoretical rank-frequency relation $n(r)$. Note that $f_3(n)$ can be written as

$$f_3(n) = \frac{B(n+1-\beta, \beta)}{B(a+1-\beta, \beta-1)}$$

using the beta function [43], $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, with an analogous expression for $S_3(n)$, but do not confuse this distribution with the beta distribution.

In all three cases it is easy to show that we have well-defined, normalized probability distributions, when n takes values $n = a, a+1, \ldots$, with a being a positive integer. Moreover, in the limit $n \to \infty$ all of them yield a power-law tail, $f(n) \propto 1/n^\beta$, so β will be referred to as the power-law exponent. Indeed, it is easy to show that

$$f_2(n) \xrightarrow{n\to\infty} (\beta-1)\frac{a^{\beta-1}}{n^\beta},$$

whereas

$$f_3(n) \xrightarrow{n\to\infty} \frac{(\beta-1)\Gamma(a)}{\Gamma(a+1-\beta)n^\beta}$$

using Stirling's formula [43]. The main difference between the three distributions is in the smaller values of n, taking $f_2(n)$ a convex shape in log-scale (as seen "from above"); $f_3(n)$ a concave one; and $f_1(n)$ being somehow in between, as it is neither concave nor convex.

METHODOLOGY AND DATA

In order to fit these three distributions to the different texts, and test the goodness of such fits, we use maximum likelihood estimation [44] followed by the Kolmogorov-Smirnov (KS) test [45]. The procedure seems similar to the one proposed in Ref. [11], but as a is fixed here, the problems resulting from the search of the optimum a [32, 46] do not arise in this case.

The method of ML estimation proceeds in the following simple way. Given a set of data $\{n_i\}$ with $i = 1, 2, \ldots N$, and a probability mass function parameterized by β, denoted as $f(n; \beta) = f(n)$, the log-likelihood function is obtained as

$$l(\beta) = \sum_{i=1}^N \ln f(n_i; \beta). \tag{6}$$

We are assuming that the data points n_i are independent from each other, in other words, we are calculating the likelihood that the data are generated independently from $f(n; \beta)$. The ML estimation of β is obtained as the value of β which maximizes $l(\beta)$; we undertake this numerically, using Brent's method [45]. In the case of the distribution f_1 the log-likelihood function takes the simple form $l_1(\beta)/N = -\ln(\zeta(\beta, a)) - \beta \ln G$, with G the geometric mean of the set $\{n_i\}$. For the other distributions no closed-form expression is possible and we use Eq. (6) directly.

As mentioned, the goodness-of-fit test is done through the Kolmogorov-Smirnov statistic [11, 45], in the discrete case [47], for which the p-value is calculated from Monte Carlo simulations (due to the fact that, as the value of the exponent is calculated from the same data is going to be tested, the theoretically computed p-value [45] would be positively biased). In this paper we use 100 Monte Carlo simulations for each test. The proper simulation of the 3 distributions is explained in the Appendix. Remember that a small enough p-value leads to the rejection of the fit. Although we perform multiple testing, we do not incorporate any Bonferroni-like correction [48–50], due to the fact that these corrections increase the number of non-rejected null hypothesis (that is, decrease the number of type I errors), inflating the performance of the fits,

in the case of goodness-of-fit tests. Without Bonferroni-like corrections, our acceptance (i.e., non-rejection) of the fits is more strict.

In order to apply this methodology we consider a set of 37 041 texts stored in UTF-8 encoding in the Project Gutenberg database (accessed July 2014 [51]). These texts correspond to different languages, styles, and time periods, although most of them are works of literature from the Western cultural tradition [52]. First of all, parts of text that do not pertain to the piece under consideration (copyright notes, headers,...) are removed by an automatized process. Books that have not been filtered in this step (mainly because they do not have standard delimiters) are discarded. After this, we still keep 97.5 % of the total (i.e., 36 108). To perform our study, we restrict ourselves to the subset of texts in English, which represent the 86 % of these 36 108 (i.e., 31 102).

An important characteristic of each text is its length, L, counting the number of word tokens it contains. It turns out to be that in the database L expands from very small values up to 4 659 068 tokens, with a distribution that is shown in Fig. 1. Observe the roughly uniform distribution up to about $L = 10^5$, and the decay afterwards. We consider only the 31 075 English texts that consist of more than 100 word tokens, as smaller texts would not have statistical value. For each of these texts we select then actual word types (punctuation signs, numbers and any character different from letters are not considered) to count their frequencies n, which will be our primary object of study.

The values of these frequencies, for each text, are available on http://dx.doi.org/10.6084/m9.figshare.1515919, in order to facilitate the reproducibility of our results.

In summary, we apply the above described fitting and goodness-of-fit procedure –using ML estimation and the Kolmogorov-Smirnov test– to a total of 31 075 texts from the English Project Gutenberg database, using three different possibilities for the distribution of frequencies: f_1 (Eq. (3)), f_2 (Eq. (4)), and f_3 (Eq. (5)). This yields a total of $3 \times 31\,075$ fits and associated p-values, which we analyze and interpret in what follows.

RESULTS

Contrary to previous studies where the number of texts considered was, at most, in the order of tens, the large-scale approach taken in this work requires a statistical analysis of the fitting results, as a case-by-case interpretation is out of hand. We first focus on the distribution of p-values, see Fig. 2 and Fig. 3. If *all* texts were truly generated by a mechanism following a given distribution, the corresponding p-values for that distribution would be uniformly distributed between zero and one. As seen in Fig. 2, this is not the case and, furthermore, most texts have rather small p-values for the three fits; nevertheless,
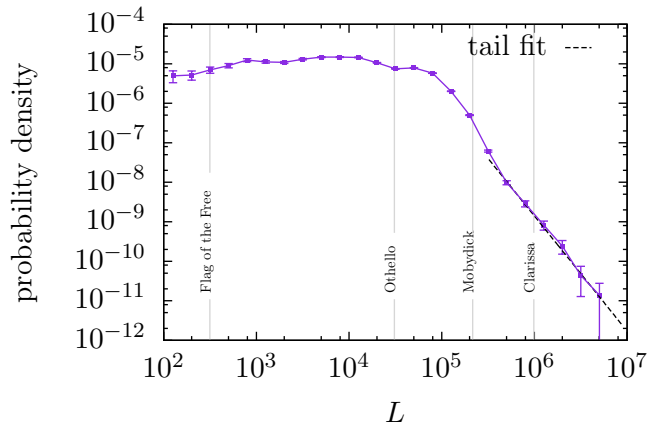
FIG. 1: Estimation of the probability density function of text length L in the English Project Gutenberg database, using logarithmic binning (5 bins per decade). Texts with less than 100 tokens are not considered. A power-law fit of the tail [32] yields an exponent 2.92 ± 0.15.

for distributions f_1 and f_2 there are still many texts that yield high enough p-values. This implies that, although we cannot conclude that the whole database is generated by any of these distributions, these cannot be rejected as good descriptions for large subsets of the database. Regarding distribution f_3, it is clear from the histogram of p-values that it can be discarded as a good description of the distribution of frequencies in any non-negligible subset of texts. So, from now on, we will concentrate on the remaining options, f_1 and f_2, to eventually quantify which of these better describes our corpus. In essence, what we are interested in is which version of Zipf's law, either distribution f_1 or f_2, fits better a reasonable number of texts, and which range of validity these simple one-parameter distributions have.

The outcome is that, independently of the significance level (as long as this is not below our resolution of 0.01 given the number of Monte Carlo simulations), the ratio between the number of texts fitted by distribution f_2 and those fitted by f_1 is nearly constant, taking a value around 2.6. For example, considering significance level (i.e., minimum p-value) equal to 0.05, Fig. 3 shows that distribution f_2 fits about 40 % of all texts, whereas distribution f_1 fits just 15 %. Both percentages include a 2.7 % of texts that are fitted by both distributions simultaneously, although this number does not keep a constant ratio with the other two, rather, it decreases when the significance level is increased (as it is implicit in the values of Fig. 3). Given that the aforementioned ratio of 2.6 is independent of the significance level, it is fair to say that distribution f_2 provides, compared to f_1, a better description of our database. As a visual illustration of the performance of the fits we display in Fig. 4 the word frequency distribution of the longest texts that
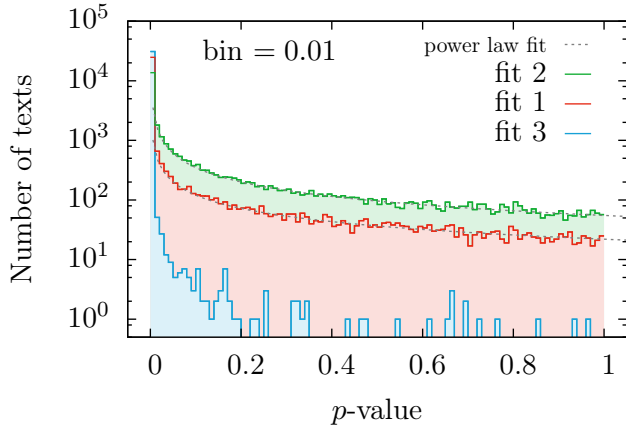
have $p > 0.5$, for distributions f_1, f_2 and f_3.

FIG. 2: Histograms of p-values obtained when the Zipf-like distributions f_1, f_2, and f_3 are fitted to the texts of the English Project Gutenberg. The histograms just count the number of texts in each bin of width 0.01. Note the poor performance of distribution 3 and the best performance of 2. Power-law approximations to the histograms for f_1 and f_2, with respective exponents 0.74 and 0.78, are shown as a guide to the eye.
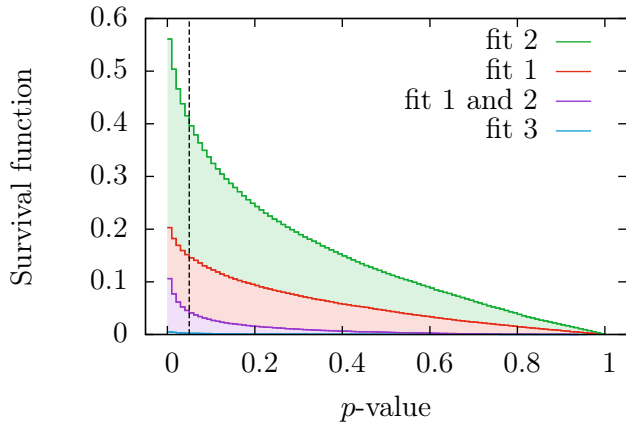

FIG. 3: Complementary cumulative distributions (i.e., survival functions) of p-values obtained when our three distributions are fitted to the texts of the English Project Gutenberg. This corresponds, except for normalization, to the integral of the previous figure, but we have included a fourth curve for the fraction of texts whose p-values for fits 1 and 2 are both higher than the value marked in the abscissa. Note that the values of p can play the role of the significance level. The value for $p = 0$ is not shown, in order to have higher resolution.

The next question we address is the dependence of the performance of fits on text length L. In order to asses this, note that from the shape of the histograms in Fig. 2 we can distinguish two groups of texts: those that lie in
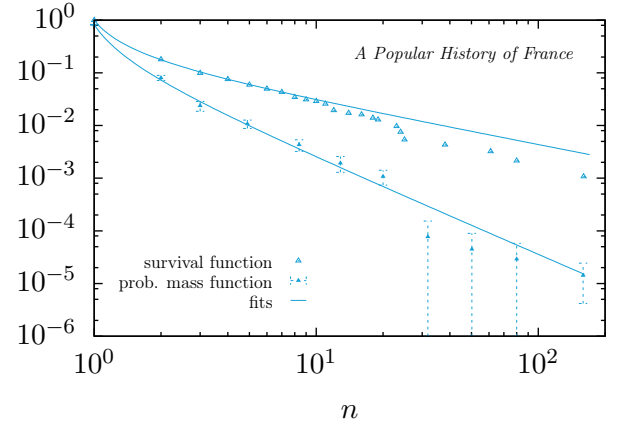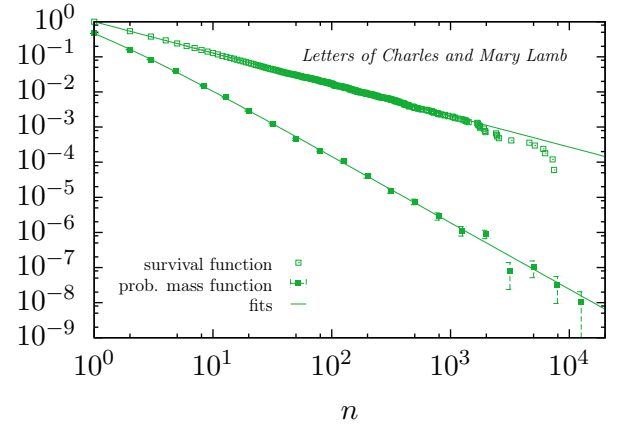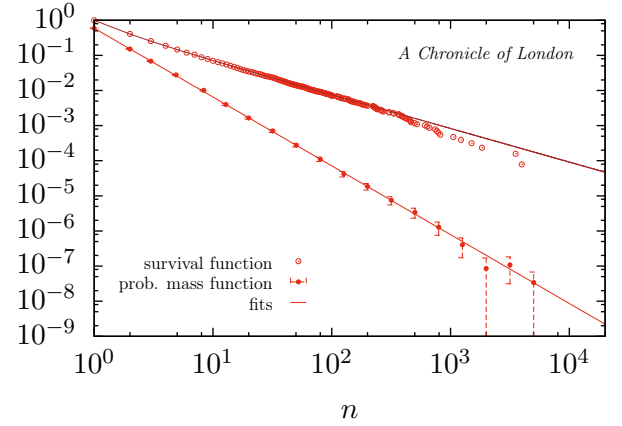

FIG. 4: Complementary cumulative distribution and probability mass function of texts: (a) *A Chronicle of London, from 1089 to 1483*; (b) *The letters of Charles and Mary Lamb*, edited by E. V. Lucas; (c) *A Popular History of France from the Earliest Times, Vol. I*, by F. Guizot. These texts are the ones with the largest length L (83 720, 239 018 and 2 081 respectively) of those that fulfill $p > 1/2$, for fits 1, 2 and 3 respectively. The exponent β takes values 1.96, 1.89, and 1.82, in each case.

the zero bin (whose p-value is strictly less than 0.01), and the rest. Taking the last group, i.e., texts with $p \geq 0.01$, and partitioning it into different subsets according to text length (i.e., looking at the distribution of p conditioned to $p \geq 0.01$ for different ranges of L), it holds that the shape of the resulting distribution of p does not strongly depend on L, as shown in Fig. 5. In contrast, the number of texts that yield p-value near zero certainly varies with L, see Fig. 6. Therefore, in order to compare the performances of f_1 and f_2 as a function of the text length L, it is enough to consider a single value of the significance level (greater than zero) as the results for any other significance level will be the same, in relative terms.

Indeed, Fig. 7(a) shows how distribution f_1 fits some more texts than distribution f_2 for small values of L, up to about 13 000 tokens. But for larger texts, distribution f_2 clearly outperforms distribution f_1, which becomes irrelevant for L beyond 100 000 (at 0.05 significance level), whereas distribution f_2 is able to fit many texts with L larger than 200 000. The figure shows that this is the case no matter if the significance level is 0.05, 0.20, or 0.50; the collapse of the curves in Fig. 7(b) confirms this fact. From Fig. 6 one could infer the same for significance level equal to 0.01. This stability of the performance of the fits for different significance levels arises from the observed fact that the distributions of p-values (conditioned to $p \geq 0.01$) are nearly identical for different L, as shown in Fig. 5.

In order to check the consistency of our fitting procedure, we also perform a direct comparison of models through the likelihood ratio (LR) test [11, 53]. We apply this test to all texts that have been fitted, considering 0.05 as significance level, by at least one of the two distributions f_1 and f_2. Then the log-likelihood-ratio between distributions f_1 and f_2 is

$$R_{1,2} = \sum_{i=1}^{N} \left(\ln f_1(n_i) - \ln f_2(n_i) \right),$$

and, under the null hypothesis that both models are equally good to describe the data, $R_{1,2}$ should be normally distributed with zero mean and a variance that can be estimated as $N\sigma^2$, with σ^2 the variance of the random variable $\ln f_1(n) - \ln f_2(n)$. Large absolute values of $R_{1,2}$ will lead to the rejection of the null hypothesis.

Table I merges the results of the LR test and our previous procedure (based on ML estimation plus the KS test). The total number of texts previously fitted by f_1 or/and f_2 is displayed depending on the sign of the corresponding log-ratio $R_{1,2}$. However we must take into account that the sign of the obtained value of $R_{1,2}$ could be a product of just statistical fluctuations if the true value were zero and thus, the sign of $R_{1,2}$ cannot be trusted in order to discriminate between two models. The probability under the null hypothesis, of obtaining an absolute value of the log-ratio greater than the empirical value
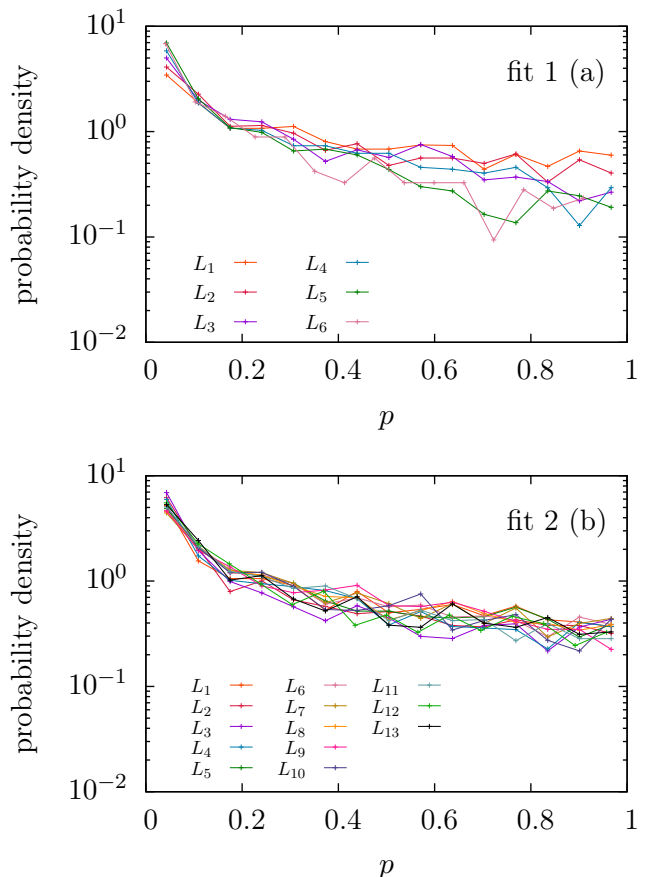

FIG. 5: Estimated probability density functions of p-values conditioned to $p \geq 0.01$ separating for different ranges of text length L. p-values correspond to the fitting of word frequencies to (a) distribution f_1 and (b) distribution f_2. We divide the distribution of text length into 15 intervals of 2 000 texts each. For distribution f_1 only the first seven groups (up to length 34 400) are displayed (beyond this value we do not have enough statistics to see the distribution of p-values greater than 0.01, as displayed in Fig. 6; for distribution 2 this happens only in the last two groups). The intervals L_i range from $L_1 = [115, 5\,291]$ to $L_{13} = [89\,476, 103\,767]$.

$|R_{1,2}|$ is computed through:

$$p_{LR} = \mathrm{erfc}\left(\frac{|R_{1,2}|}{\sqrt{2N\sigma^2}} \right)$$

where erfc is the complementary error function [43]. We will take as statistically significant those $R_{1,2}$ that yield $p_{LR} < 0.05$. Equivalently, at 0.05 significance level, $R_{1,2}$ is significant if its absolute value is greater than $R_c = 1.96\sqrt{N\sigma^2}$. The results are shown in Table II

Note that the LR test cannot conclude if a fit is good or bad, as it only compares the relative performance of two fits; in other words, if the LR test selects a particular distribution, that distribution can still yield a bad fit, in absolute terms. Anyway, there is no mismatch between
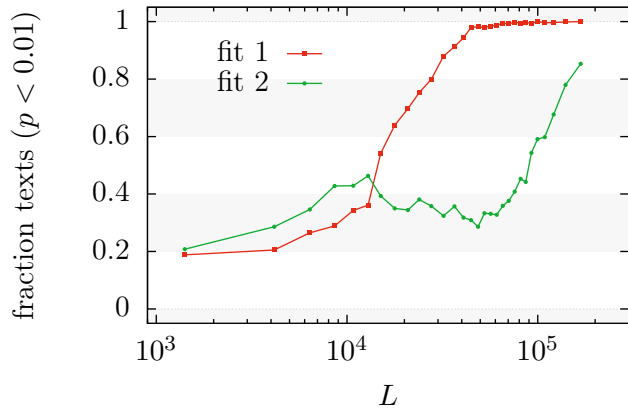
FIG. 6: Number of texts with p-value near zero ($p < 0.01$) in different ranges of L divided by the number of texts in the same ranges, for the fits of distributions f_1 and f_2. Values of L denote the geometric mean of ranges containing 1000 texts each. The higher value for fit 1 (except for L below about 13000 tokens) denotes its worst performance.

the results of both tests: any time the ML-KS method selects one distribution over the other, the LR test either supports the selection or does not give significant results, but it never selects the other option (as shown in Table II).

	$R_{1,2} > 0$	$R_{1,2} < 0$	Total ML-KS
f_1 (exclusively)	3614	81	3695
f_2 (exclusively)	120	11366	11486
f_1 and f_2	431	398	829
Total LR	4165	11845	16010

TABLE I: The number of texts that are fitted by f_1 or f_2 or both at 0.05 significance level of the ML-KS procedure, separated into two columns according to the sign of $R_{1,2}$. Positive $R_{1,2}$ means that the likelihood for f_1 is greater than that for f_2, and conversely for negative $R_{1,2}$. Nevertheless, the sign of $R_{1,2}$ is not an indication of significance, for significant LR tests see Table II.

Taking now those texts whose frequency distributions could be approximated by f_1 or f_2, we draw attention to the distribution of the estimated exponents (i.e., the parameter β). The original formulation of Zipf's law implies $\beta = 2$ and Fig. 8 shows that β is certainly distributed around 2, with a bell-like shape. If we check the effect of the text length L in the distribution of β, we find a decreasing trend of β with L, as can be seen in Fig. 9. We have tested that this observation is not an artifact
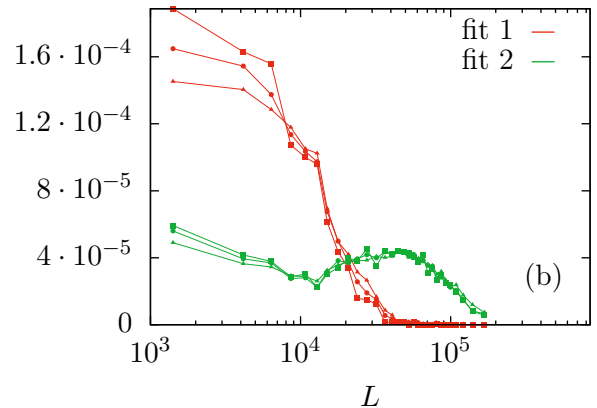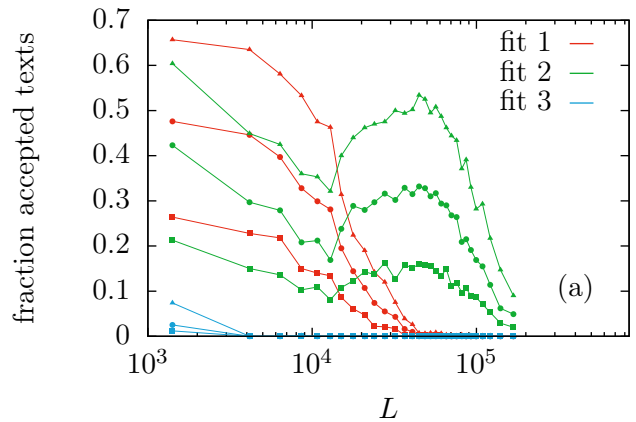

FIG. 7: (a) Histograms showing the fraction of accepted texts by the three distributions as a function of their text length, for three different significance levels p_0: 0.05 (upper curves), 0.20 (middle), 0.50 (lower). To be concrete, for each range of L, the ratio between the number of texts with $p \geq p_0$ and the number of texts in that range is calculated. (b) Same curves (removing those for distribution 3) under rescaling. We rescale the y-axis by the number of $p \geq p_0$, in each case, showing that the relative performance of each fit with regard L is independent on the significance level. Bins are selected to contain 1000 texts each.

	$R_{1,2} > R_c$	$R_{1,2} < -R_c$
f_1 (exclusively)	1666	0
f_2 (exclusively)	0	9423
f_1 and f_2	0	3
Total LR test	1666	9426
None (neither f_1 nor f_2)	510	11431

TABLE II: Number of texts with a significant LR test, at the 0.05 level, either favouring distribution f_1 ($R_{1,2} > R_c$) or distribution f_2 ($R_{1,2} < -R_c$), for different outcomes of the ML-KS procedure (at the 0.05 level also). Note that these cases correspond to a subset of the previous table. An additional row shows the number of texts that are fitted neither by distribution f_1 nor f_2; notice that in this case a significant LR test does not guarantee a good fit.
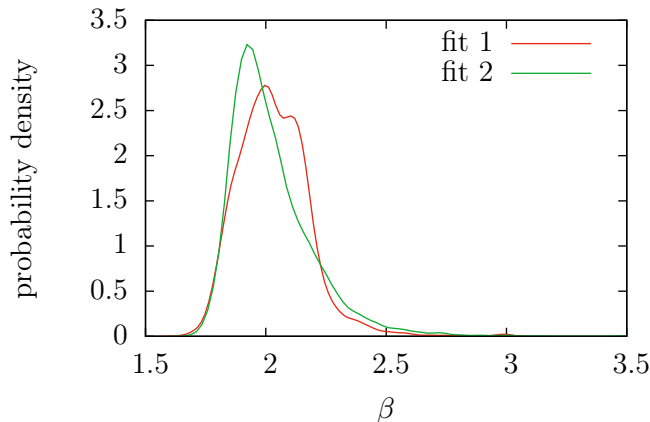
FIG. 8: Estimation of the probability density of the exponent β for texts yielding $p \geq 0.05$ in fit 1 and fit 2. Curves have been calculated from the histograms via normal kernel smoothing method as implemented in MatLab (*ksdensity* function). Estimated mean and standard deviation are 2.03 and 0.15 respectively for fit 1, and 2.02 and 0.17 for fit 2.

of the fitting method, as synthetic texts generated with fixed β do not show this behavior. We have no theoretical explanation for this fact, but notice that this trend is not in disagreement with the claims of Ref. [33], where the stability of the exponent β was demonstrated for a single growing text (i.e., comparing small parts of a text with the whole).

CONCLUSIONS

Zipf's law is probably the most intriguing and at the same time well-studied experimental law of quantitative linguistics, and extremely popular in its wider sense in the science of complex systems. Although the previous literature is vast, as far as we know our work constitutes the first large-scale analysis of Zipf's law in single (non-aggregated) texts. Thus, we are in a position to make a well-grounded statement about the validity of Zipf's law in such texts.

Let us first briefly summarize, however, some key technical points of our study. First, we have analyzed a total of 31 075 English texts from the Project Gutenberg database using rigorous fitting procedures, and have tested how well they are described by three Zipf-like distributions. Our choice of distributions has not been exhaustive; rather, we have limited ourselves to different interpretations of what can be understood as "Zipf's law", in the sense of having a perfect power law either in the probability mass function of word frequencies, or in the complementary cumulative distribution function (whose empirical estimation leads to the rank-frequency relation of the sample), or in the rank-frequency relation of an
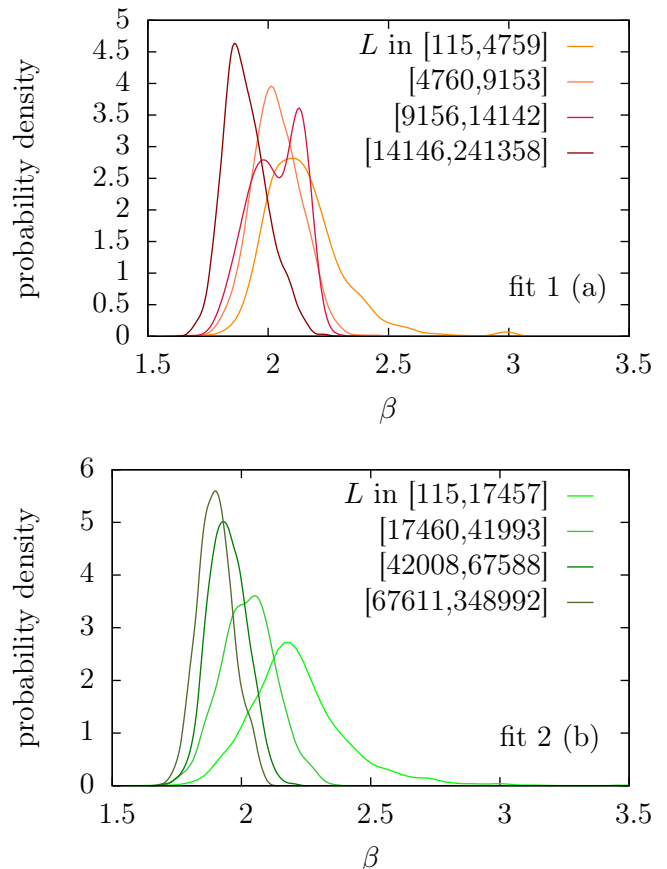
FIG. 9: Estimated probability density of β for fits with $p \geq 0.05$, in different length ranges. We have divided both groups of accepted texts into 4 percentiles according to L. As in the previous figure, the normal kernel smoothing method is applied. (a) For distribution f_1. (b) For distribution f_2.

underlying population. Remarkably, the resulting distributions have a unique parameter, β, which in all cases is the exponent of an asymptotic power law in the probability mass function of the frequency. It is left to explore how other, more complicated extensions of Zipf's law perform on this large corpus, but it is obvious that, by including additional parameters, one might provide good fits to a larger number of texts (although in this case, proper model selection will require to balance number of parameters and parsimony).

Our aim in this paper has not been to fit as many texts as possible, but to test the performance of the simplest Zipf-like distributions within a very strict, conservative framework. Indeed, by requiring the three versions of Zipf's law to hold on the full range of frequencies $n = 1, 2, \ldots$ (and not only on the tail of the distribution) we put ourselves in the strictest range of demands. It is hence remarkable that, e.g., at the standard significance level of 0.05, and for text lengths between 10^4 and 10^5 word to-

kens, more than 40 % of the considered texts are statistically compatible with the pure power law in the complementary cumulative distribution function represented by distribution f_2 (see Fig. 7). So, we can state that, for the corpus under consideration, the most appropriate version of Zipf's law is given by a probability mass function

$$f(n) = \text{Prob[frequency} = n] = \frac{1}{n^{\beta-1}} - \frac{1}{(n+1)^{\beta-1}},$$

or, equivalently, by a complementary cumulative distribution function

$$S(n) = \text{Prob[frequency} \geq n] = \frac{1}{n^{\beta-1}}.$$

Due to the broad coverage of the Project Gutenberg corpus we speculate that this distribution should fit a large fraction of generic (non-technical) English texts. Of course, testing this speculation in front of all possible corpora is an impossible task.

We have also shown that our conclusions regarding the *relative* performance of a pure power law in the probability mass function, given by distribution f_1, versus distribution f_2 are robust with respect to changes in the significance level: about twice as many texts are statistically compatible with distribution f_2 than those compatible with f_1, at any significance level (obviously, in absolute terms, the number of accepted texts varies with the significance level). Hence we can conclude that distribution f_2 gives a better description of English texts than distribution f_1, at least for the corpus considered in this work. We also conclude that distribution f_3, first derived by Mandelbrot [38], is irrelevant for the description of texts in this corpus. Finally, we have corroborated that the exponent β of Zipf's law certainly varies from text to text, as had been previously claimed using other approaches for defining what Zipf's law is [5, 6]. Interestingly, the value $\beta = 2$ originally proposed by Zipf himself is among the most frequent ones.

We believe that our analysis constitutes a major advancement in the understanding of Zipf's law. It is astonishing how good the simplest one-parameter Zipf-like distributions perform on such a large set of texts, particularly with the strict set of requirements we have imposed. This is in sharp contrast for instance with Zipf's law in demography [54] and in the distribution of income [55], where the power law seems to be valid only for the tail corresponding to the largest sizes, as it happens also for the distribution of word frequency in large text corpora, as mentioned above [2, 34–37].

Zipf's law has been subject to much debate, and will probably continue to be so for many years. Indeed, one can always cast doubt on its validity on the basis of some particular examples. Yet it seems clear to us that, in our modern times of big data and large computational capabilities, more efforts should be put towards large-scale analysis of Zipf's law. We hope this paper constitutes a first step in this direction.

APPENDIX: SIMULATION OF DISCRETE ZIPF-LIKE DISTRIBUTIONS

As part of the testing procedure, we need simulated samples from f_1, f_2, and f_3, which are discrete distributions defined for $n = a, a + 1, \dots$. We will give the recipe of simulation for an arbitrary positive integer value of the lower cut-off a. It is simpler to start with f_2, as this is used as an auxiliary distribution in the simulation of the other two.

Simulation of f_2. Fixed a and given the parameter β, we want a set of random numbers whose cumulative distribution function is a discrete power law: $S_2(n) = (a/n)^{\beta-1}$. For that, we first generate a random number u from a uniform distribution in the interval $(0, u_{max})$, with $u_{max} = 1/a^{\beta-1}$. If we take $x = 1/u^{1/(\beta-1)}$, it turns out to be that the values of x yield a continuous power law with $S_2^c(x) = (a/x)^{\beta-1}$, for $x \geq a$, where the superscript c distinguishes the continuous distribution from its discrete analogous one. So, taking n equal to the integer part of x, i.e., $n = \text{int}(x)$, yields a discrete distribution with $S_2(n) = (a/n)^{\beta-1}$, as desired. This is so because, for any X, $\text{int}(X) \geq n$ is equivalent to $X \geq n$ for n integer. In a recipe:

- generate u from a uniform distribution in $(0, 1/a^{\beta-1}]$,

- calculate $x = 1/u^{1/(\beta-1)}$,

- take $n = \text{int}(x)$.

By means of the so-called rejection method [56], simulated integers distributed following f_2 can be used for the simulation of integers following f_1 or f_3. The key point to achieve a high performance in the rejection method is to use a "good" auxiliary function, i.e., one that leads to a low rejection rate. This is certainly the case in our framework, as explained below.

Simulation of f_1. In this case, the steps are:

- generate n from $f_2(n)$,

- generate v from a uniform distribution in the unit interval,

- n is accepted if

$$v \leq \frac{f_1(n)}{f_2(n)C},$$

and rejected otherwise, where C is the rejection constant given by $C = \max_{n \geq a} \{f_1(n)/f_2(n)\}$.

It is easy to check that the maximum of f_1/f_2 is reached at $n = a$ as this is a decreasing function [39]. The acceptance condition above can be simplified by taking $\tau = (1 + n^{-1})^{\beta-1}$, and $b = (1 + a^{-1})^{\beta-1}$, then, the condition becomes:

$$bvn(\tau - 1) \leq a(b - 1)\tau,$$

which is devoid of the calculation of the Hurwitz-zeta function. This is a generalization for $a > 1$ of the method of Ref. [56]. The choice of f_2 as the auxiliary distribution function is justified by the small value that C takes, as this is the expected number of generated values of n until we accept one. For instance, for $\beta = 2$ and $a = 1$ we get $C = 1.2158$.

Simulation of f_3. Proceeding similarly, we get in this case low values of $C = \max_{n \geq a} \{f_3(n)/f_2(n)\}$ as well (we get $C = 2$ in the limit $\beta \to 2$ for $a = 1$). The maximum of $f_3(n)/f_2(n)$ is numerically seen to be reached at $n = a$. In summary, the steps are:

- generate n from $f_2(n)$

- generate v from a uniform distribution in the unit interval

- n is accepted if

$$v f_2(n) \leq a \left(1 - \left(\frac{a}{a+1} \right)^{\beta - 1} \right) \frac{\Gamma(n - (\beta - 1))}{\Gamma(n+1)} \frac{\Gamma(a)}{\Gamma(1 + a - \beta)}$$

and rejected otherwise.

ACKNOWLEDGMENTS

[1] H. Baayen. *Word Frequency Distributions*. Kluwer, Dordrecht, 2001.

[2] M. Baroni. Distributions in text. In A. Lüdeling and M. Kytö, editors, *Corpus linguistics: An international handbook, Volume 2*, pages 803–821. Mouton de Gruyter, Berlin, 2009.

[3] D. Zanette. Statistical patterns in written language. *arXiv*, 1412.3336v1, 2014.

[4] S. T. Piantadosi. Zipf's law in natural language: a critical review and future directions. *Psychon. Bull. Rev.*, 21:1112–1130, 2014.

[5] D. Zanette and M. Montemurro. Dynamics of text generation with realistic Zipf's distribution. *J. Quant. Linguist.*, 12(1):29–40, 2005.

[6] A. Corral, G. Boleda, and R. Ferrer-i-Cancho. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE*, 10(7):e0129031, 2015.

[7] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. Ll. Arcos. Measuring the evolution of contemporary western popular music. *Sci. Rep.*, 2:521, 2012.

[8] M. Haro, J. Serrà, P. Herrera, and A. Corral. Zipf's law in short-time timbral codings of speech, music, and environmental sound signals. *PLoS ONE*, 7:e33993, 2012.

[9] W. Li. Zipf's law everywhere. *Glottom.*, 5:14–21, 2002.

[10] R. L. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.

[11] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.

[12] S. Pueyo and R. Jovani. Comment on "A keystone mutualism drives pattern in a power function". *Science*, 313:1739c–1740c, 2006.

[13] J. Camacho and R. V. Solé. Scaling in ecological size spectra. *Europhys. Lett.*, 55:774–780, 2001.

[14] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143–150, 2002.

[15] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Cont. Phys.*, 46:323 –351, 2005.

[16] C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102, 2003.

[17] H. A. Simon. On a class of skew distribution functions. *Biomet.*, 42:425–440, 1955.

[18] G. A. Miller. Some effects of intermittent silence. *Am. J. Psychol.*, 70(2):311–314, 1957.

[19] R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A.*, 100:788–791, 2003.

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

[21] A. Saichev, Y. Malevergne, and D. Sornette. *Theory of Zipf's Law and of General Power Law Distributions with Gibrat's Law of Proportional Growth*. Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Berlin, 2009.

[22] 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.

[23] J. Peterson, P. D. Dixit, and K. A. Dill. A maximum entropy framework for nonexponential distributions. *Proc. Natl. Acad. Sci. USA*, 110(51):20380–20385, 2013.

[24] B. Corominas-Murtra, R. Hanel, and S. Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA*, 112(17):5348–5353, 2015.

[25] R. Ferrer-i-Cancho and B. Elvevåg. Random texts do not exhibit the real Zipf's law-like rank distribution. *PloS One*, 5(3):e9411, 2010.

[26] R. Dickman, N. R. Moloney, and E. G. Altmann. Analysis of an information-theoretic model for communication. *J. Stat. Mech: Theory Exp.*, page P12022, 2012.

[27] W. Li, P. Miramontes, and G. Cocho. Fitting ranked linguistic data with two-parameter functions. *Entropy*, 12(7):1743–1764, 2010.

[28] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *Eur. Phys. J. B*, 41:255–258, 2004.

[29] H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. *Eur. Phys. J. B*, 58:167–173, 2007.

[30] E. P. White, B. J. Enquist, and J. L. Green. On estimating the exponent of power-law frequency distributions. *Ecol.*, 89:905–912, 2008.

[31] G. Casella and R. L. Berger. *Statistical Inference.*

Duxbury, Pacific Grove CA, 2nd edition, 2002.

[32] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394, 2013.

[33] F. Font-Clos, G. Boleda, and A. Corral. A scaling law beyond Zipf's law and its relation with Heaps' law. *New J. Phys.*, 15:093033, 2013.

[34] R. Ferrer i Cancho and R. V. Solé. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.*, 8(3):165–173, 2001.

[35] A. M. Petersen, J. N. 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.

[36] M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, 2013.

[37] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds. Text mixing shapes the anatomy of rank-frequency distributions: A modern Zipfian mechanics for natural language. *arXiv*, 1409.3870v3, 2015.

[38] B. Mandelbrot. On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI, 1961.

[39] A. Corral and R. Ferrer i Cancho. *(in preparation)*, 2015.

[40] A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Pub. Co., New York, 2nd edition, 1956.

[41] E. G. Altmann and M. Gerlach. Statistical laws in linguistics. *arXiv*, 1502.03296, 2015.

[42] F. Font-Clos and A. Corral. Log-log convexity of type-token growth in Zipf's systems. *Phys. Rev. Lett.*, 114:238701, 2015.

[43] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[44] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford UP, Oxford, 2001.

[45] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 2nd edition, 1992.

[46] A. Corral, F. Font, and J. Camacho. Non-characteristic half-lives in radioactive decay. *Phys. Rev. E*, 83:066103, 2011.

[47] A. Corral, A. Deluca, and R. Ferrer-i-Cancho. A practical recipe to fit discrete power-law distributions. *ArXiv*, 1209:1270, 2012.

[48] H. Abdi. Bonferroni and šidák corrections for multiple comparisons. In N. J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 103–107. Sage, Thousand Oaks, 2007.

[49] J. M. Bland and D. G. Altman. Multiple significance tests: the Bonferroni method. *Brit. Med. J.*, 310:170–170, 1995.

[50] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Roc. B*, 57:289–300, 1995.

[51] Project Gutenberg. http://www.gutenberg.org.

[52] Wikipedia (accessed August 2015). https://en.wikipedia.org/wiki/Project_Gutenberg.

[53] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–33, 1989.

[54] Y. Malevergne, V. Pisarenko, and D. Sornette. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Phys. Rev. E*, 83:036111, 2011.

[55] A. Drăgulescu and V. M. Yakovenko. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A*, 299(1-2):213–221, 2001.

[56] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.