CRM^R

CENTRE DE RECERCA MATEMÀTICA

A scaling law beyond Zipf's law and its relation with Heaps' law

F. Font-Clos, G. Boleda, Á. Corral

A SCALING LAW BEYOND ZIPF'S LAW AND ITS RELATION WITH HEAPS' LAW

FRANCESC FONT-CLOS, GEMMA BOLEDA, AND ÁLVARO CORRAL

ABSTRACT. The dependence with text length of the statistical properties of word occurrences has long been considered a severe limitation for the usefulness of quantitative linguistics. We propose a simple scaling form for the distribution of absolute word frequencies which uncovers the robustness of this distribution as text grows. In this way, the shape of the distribution is always the same and it is only a scale parameter which increases linearly with text length. By analyzing very long novels we show that this behavior holds both for raw, unlemmatized texts and for lemmatized texts. For the latter case, the word-frequency distribution is well fit by a double power law, maintaining the Zipf's exponent value $\gamma \simeq 2$ for large frequencies but yielding a smaller exponent in the low frequency regime. The growth of the distribution with text length allows us to estimate the size of the vocabulary at each step and to propose an alternative to Heaps' law, which turns out to be intimately connected to Zipf's law, thanks to the scaling behavior.

1. INTRODUCTION

Zipf's law is perhaps one of the best evidences of the existence of universal physical-like laws in cognitive science and in the social sciences. Classic examples of it include the population of cities, the assets of companies, and the frequency of words in texts or speech [41]. Taking the latter case, the law is obtained directly by counting the number of repetitions, i.e., the absolute frequency n, of all words in a long enough text, and assigning increasing ranks, $r = 1, 2, \ldots$, to decreasing frequencies. If a power-law relation

$$n \propto \frac{1}{r^{\beta}}$$

holds for a large enough range, with exponent β more or less close to 1, then it is considered that Zipf's law is fulfilled (with $\propto$ denoting proportionality). An equivalent formulation of the law is obtained in terms of the probability distribution of the frequency n, making it to play the role of a random variable, for which a power-law distribution

$$D(n) \propto \frac{1}{n^{\gamma}},$$

should hold, with $\gamma = 1 + 1/\beta$ (taking values close to 2) and $D(n)$ the probability mass function of n (or the probability density of n, in a continuous approximation)

1

[26, 38, 16, 1]. Note that this picture implies to perform a double statistics, first counting words to get frequencies and then counting repetition of frequencies to get the distribution of frequencies.

We can recognize that the criteria for the validity of Zipf's law are rather vague (long enough text, large enough range, exponent β more or less close to 1). Generally, a long enough text means a book, a large range can be a bit more than an order of magnitude, and the proximity of the exponent β to 1 translates into an interval (0.7,1.2), or even beyond that [39, 38, 14]. Moreover, no rigorous methods have been usually required for the fitting of the power-law distribution, with linear regression in double-logarithmic scale being the most common one, either for $n(r)$ or for $D(n)$. But it is well known that this procedure has severe drawbacks and can lead to flawed results [7]. Nevertheless, once these limitations have been assumed, the fulfillment of Zipf's law in linguistics is astonishing, being valid no matter the author, style, or language [41, 39, 38]. So, the law is universal, at least in a qualitative sense.

At a theoretical level, many different competing explanations of Zipf's law have arisen [38], as random (monkey) typing [27, 24], preferential repetitions or proportional growth [35, 29, 32], the least effort principle [41, 18, 8, 15], and, beyond linguistics, Boltzmann-type approaches [13], or even avalanche dynamics in a critical system [4]; remarkably, most of these options have generated considerable controversy [28, 19, 12]. In any case, the power-law behavior is the hallmark of scale invariance, i.e., the impossibility to define a characteristic scale, either for frequencies or for ranks. Although power laws are sometimes also referred to as scaling laws, we will make a precise distinction here. In short, a scaling law is any function invariant under a scale transformation (which is a linear dilation or contraction of the axes). In one dimension the only scaling law is the power law, but this is not true with more than one variable [6]. Note that in text statistics, other variables to consider in addition to frequency are the text length L (the total number of words, or tokens) and the size of the vocabulary V_L (i.e., the number of different words, or types).

Somehow related to Zipf's law is Heaps' law (also called Herdan's law [2, 23]), which states that the vocabulary V_L grows as a function of the text length L as a power law,

$$V_L \propto L^{\alpha},$$

with exponent α smaller than one. However, even simple log-log plots of V_L versus L do not show a convincing linear behavior and therefore, the evidence for this law is somewhat weak. Nevertheless, a number of works have derived the relationship $\beta = 1/\alpha$ between Zipf's and Heaps' exponents [37], at least in the infinite-system limit [34, 25], using different assumptions.

Despite the relevance of Zipf's law, and its possible relations with criticality, few systematic studies of the dependence of the law on system size (i.e., text length) have been performed. But it was Zipf himself [41, pp. 144] who first

observed a variation in the exponent β when the system size was varied. In particular, "small" samples would give $\beta < 1$, while "big" ones yielded $\beta > 1$. However, that was attributed to "undersampling" and "oversampling", as Zipf believed that there was an optimum system size under which all words occurred in proportion to their theoretical frequencies, i.e., those given by the exponent $\beta = 1$. This increase of β with L has been confirmed later, see Refs. [30, 2], leading to the conclusion that the practical usefulness of Zipf's law is rather limited [2]. More recently, using rather large collections of books from single authors, Bernhardsson et al. [5] find a decrease of the exponents γ and α with text length, in correspondence with the increase in β found by Zipf and others. They propose a size-dependent word-frequency distribution based on three main assumptions:

 (i) The vocabulary scales with text length as $V_L \propto L^{\alpha(L)}$, where the exponent $\alpha(L)$ itself depends on the text length. Note that this is not an assumption in itself, just notation, and is also equivalent to writing the average frequency $\langle n \rangle = L/V_L$ as $\langle n(L) \rangle \propto L^{1-\alpha(L)}$.

 (ii) The maximum frequency is proportional to the text length, i.e. $n_{\max} = n(r = 1) \propto L$.

(iii) The functional form of the word frequency distribution $D_L(n)$ is that of a power law with an exponential tail, with both the scale parameter $c(L)$ and the power-law exponent $\gamma(L)$ depending on the text length L. That is,

$$D_L(n) = A \frac{e^{-n/c(L)}}{n^{\gamma(L)}},$$

with $1 < \gamma(L) < 2$.

Taking $c(L) = c_0 L$ guarantees that $n_{\max} \propto L$; moreover, the form of $D_L(n)$ implies that, asymptotically, $\langle n(L) \rangle \propto L^{2-\gamma(L)}$ [6], which, comparing with condition (i) leads to

$$\alpha(L) = \gamma(L) - 1,$$

so, $0 < \alpha(L) < 1$. This relation between α and γ is in agreement with previous results if L is fixed [26, 25, 34]. It was claimed in Ref. [5] that $\alpha(L)$ seems to decrease from 1 to 0 for increasing L and therefore $\gamma(L)$ decreases from 2 to 1. The resulting functional form,

$$D_L(n) = A \frac{e^{-n/(c_0 L)}}{n^{1+\alpha(L)}},$$

is in fact the same appearing in many critical phenomena, where the power-law term is limited by a characteristic value of the variable, $c_0 L$, arising from a deviation from criticality or from finite-size effects [6, 36, 40, 10]. Note that this implies that the tail of the frequency distribution is not a power law but an exponential, and therefore the frequency of most common words is not power-law distributed. This is in contrast with recent studies that have clearly established that the tail of $D_L(n)$ is well modelled by a power law [7, 9]. But what is most uncommon

about this functional form is to have a "critical" exponent that depends on the system size, rather, the values of exponents should not be influenced by external scales. So, we seek for an alternative picture that is in more agreement with typical scaling phenomena.

Our proposal is that, although the word-frequency distribution $D_L(n)$ changes with system size L, the shape of the distribution is independent on L and V_L, and it is only the scale of $D_L(n)$ which changes with these variables. This implies that the shape parameters of $D_L(n)$ (in particular, any exponent) do not change with L and only one scale parameter changes with L, increasing linearly. This is explained in the next section, while the third one is devoted to the validation of our scaling form in real texts, using both plain words or their corresponding lemma forms; in the latter case an alternative to Zipf's law can be proposed, consisting in a double power-law distribution. Our findings for words and lemmas suggest that the previous observation that the Zipf's exponent depends on the text length [30, 2, 5], might be an artifact of the increasing weight of a second regime in the distribution of frequencies beyond a certain text length. The fourth section investigates the implications of our scaling approach for Heaps' law. Although the scaling ansatz we propose has a counterpart in the rank-frequency representation, we prefer to illustrate it in terms of the distribution of frequencies, as this approach has been found more appropriate from a statistical point of view [9].

2. The scaling form of the word-frequency distribution

Let us come back to the rank-frequency relation, in which the absolute frequency n of each word is a function of its rank r. Defining the relative frequency as $x \equiv n/L$ and inverting the relationship, we can write

$$r = G_L(x).$$

Note that here we are not assuming a power-law relationship between r and x, just a generic function G_L, which may depend on the text length L. Instead of the three hypothesis introduced by Bernhardsson et al. we just need one hypothesis, which is the independence of the function G_L on L; so

$$r = G(n/L),$$

This turns out to be a scaling law, with $G(x)$ a scaling function. It means that if in the first 10,000 words of a book there are 5 words with relative frequency larger than or equal to 2%, that is, $G(0.02) = 5$, then this will still be true for the first 20,000 words, and for the first 100,000 and for the whole book. These words need not necessarily be the same ones, although in some cases they might be. In fact, instead of assuming as in Ref. [5] that the frequency of the most used word scales linearly with L, what we assume is just that this is true for all words, at least on average.

Now let us introduce the survivor function or complementary cumulative distribution function $S_L(n)$ of the absolute frequency, defined in a text of length L as $S_L(n) = \text{Prob}[\text{frequency} \geq n]$. Note that, estimating from empirical data, $S_L(n)$ turns out to be essentially the rank, but divided by the total number of ranks, V_L, i.e., $S_L(n) = r/V_L$. Therefore, using our ansatz for r we get

$$S_L(n) = \frac{G(n/L)}{V_L}.$$

Within a continuous approximation the probability mass function of n, $D_L(n) = \text{Prob}[\text{frequency} = n]$, can be obtained from the derivative of $S_L(n)$,

$$(1) \qquad D_L(n) = -\frac{\partial S_L(n)}{\partial n} = \frac{g(n/L)}{LV_L},$$

where g is minus the derivative of G, i.e., $g(x) = -G'(x)$. If one does not trust the continuous approximation, one can write $D_L(n) = S_L(n) - S_L(n+1)$ and perform a Taylor expansion, for which the result is the same, but with $g(x) \simeq -G'(x)$. In this way, we obtain simple forms for $S_L(n)$ and $D_L(n)$, which are analogous to standard scaling laws, except for the fact that we have not specified how V_L changes with L. If Heaps' law holds, $V_L \propto L^\alpha$, we recover a standard scaling law, $D_L(n) = g(x/L)/L^{1+\alpha}$, which fulfills invariance under a scaling transformation, or, equivalently, fulfills the definition of a generalized homogeneous function [6, 22],

$$D_{\lambda_L L}(\lambda_n n) = \lambda_D D_L(n),$$

where λ_L, λ_n, and λ_D are the scale factors, related in this case through

$$\lambda_n = \lambda_L \equiv \lambda$$

and

$$\lambda_D = \frac{1}{\lambda^{1+\alpha}}.$$

However, in general (if Heaps' law does not hold), the distribution $D_L(n)$ still is invariant under a scale transformation but with a different relation for λ_D, which is

$$\lambda_D = \frac{V_L}{\lambda V_{\lambda L}}.$$

So, $D_L(n)$ is a not a generalized homogeneous function, but presents an even more general form. In any case, the validity of the proposed scaling law, Eq. (1), can be checked by performing a very simple rescaled plot, displaying $LV_L D_L(n)$ versus n/L. A resulting data collapse would be in favor of the independence of the scaling function on L. This is undertaken in the next section.

3. Data analysis results

To test the validity of our predictions, summarized in Eq. (1), we analyze a corpus of literary texts, comprised by seven large books in English, Spanish, and French (among them, some of the longest novels ever written, in order to have as much as possible statistics of homogeneous texts). In addition to performing the statistics of the words in the texts, we consider the lemmatized version of each text, where each word is substituted by its corresponding lemma (roughly speaking, the stem form of the word), and we consider the statistics of lemmas as well. Appendix A provides detailed information on the lemmatization procedure, and Table 1 summarizes the most relevant characteristics of each of the books.

Title	author	language	year	L_{tot}	V_{tot}	$L_{\text{tot}}^{(l)}$	$V_{\text{tot}}^{(l)}$
Artamène	Scudéry siblings	French	1649	2,078,437	25,161	1,737,556	5,008
Clarissa	Samuel Richardson	English	1748	971,294	20,490	940,967	9,041
Don Quijote	Miguel de Cervantes	Spanish	1605-1615	390,436	21,180	378,664	7,432
La Regenta	L. Alas "Clarín"	Spanish	1884	316,358	21,870	309,861	9,900
Bragelonne	A. Dumas (father)	French	1847	693,947	25,775	676,252	10,744
Moby-Dick	Herman Melville	English	1851	215,522	18,516	204,094	9,141
Ulysses	James Joyce	English	1918	268,144	29,448	242,367	12,469

TABLE 1. Total text length and vocabulary before $(L_{\text{tot}}, V_{\text{tot}})$ and after $(L_{\text{tot}}^{(l)}, V_{\text{tot}}^{(l)})$ the lemmatization process, for all the books considered (including also their author, language, and publication year). The text length for lemmas is shorter than for words because for a number of word tokens their corresponding lemma type could not be determined, and they were ignored.

First, we plot the distributions of word frequencies, $D_L(n)$ versus n, for each book, considering either the whole book or the first L/L_{tot} fraction, where L_{tot} is the real, complete text length (i.e., if $L = L_{\text{tot}}/2$ we consider just the first half of the book, no average is performed over parts of size L). For a fixed book, we observe that different L leads to distributions with small but clear differences, see Fig. 1. The pattern described by Bernhardsson et al. (equivalent to Zipf's findings for the change of the exponent β) seems to hold, as the absolute value of the slope in log-log scale (i.e., the apparent power-law exponent γ) decreases with increasing text length.
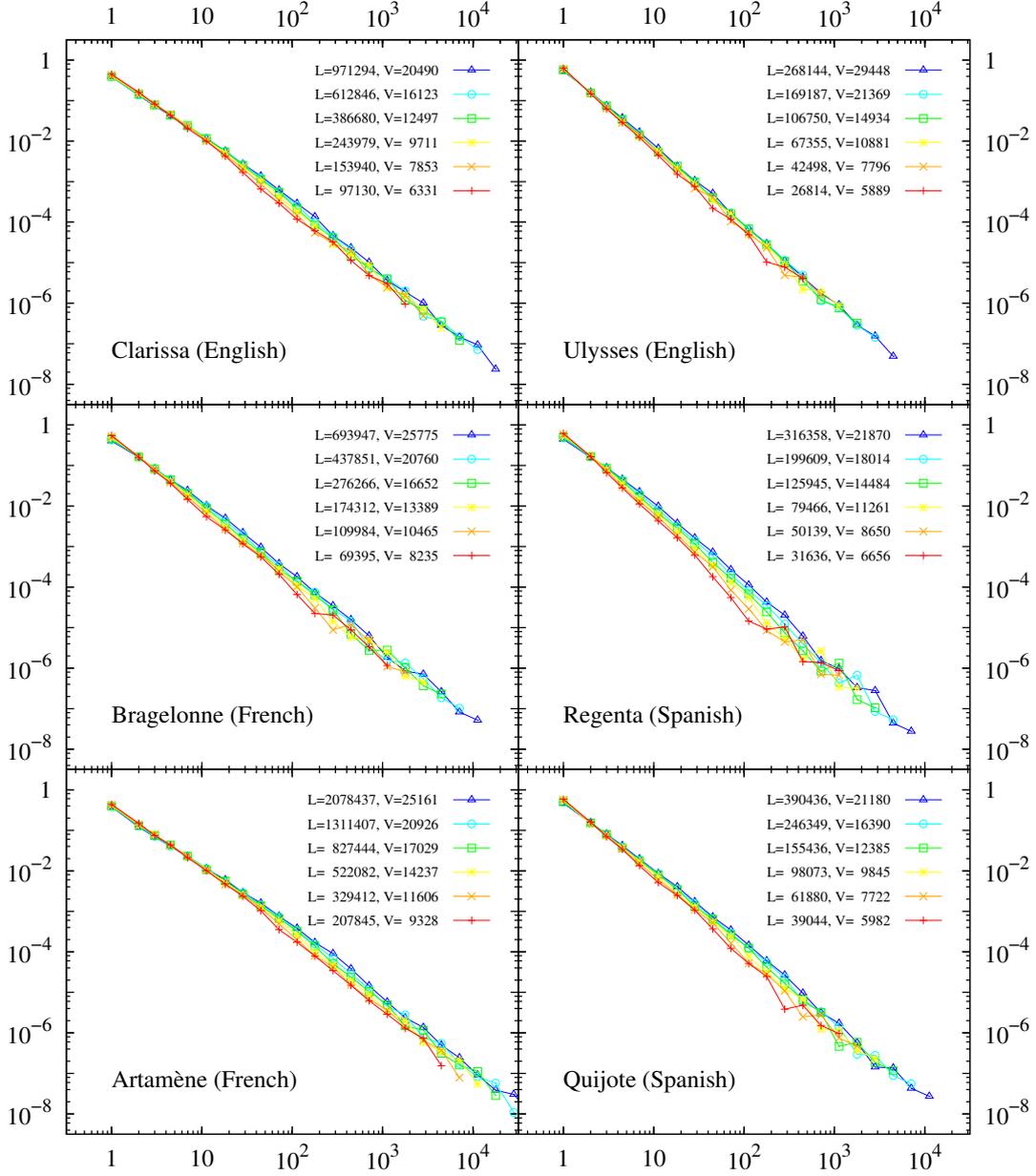
FIGURE 1.   Density of word frequencies $D_L(n)$ (y-axis) against absolute frequency n (x-axis), for six different books, taking text length $L = L_{tot}/10$, $L_{tot}/10^{4/5}$, $L_{tot}/10^{3/5}$, $\ldots$, L_{tot}. The slope seems to decrease with text length.

However, a scaling analysis reveals an alternative picture. As suggested by Eq. (1), plotting $LV_LD_L(n)$ against n/L for different values of L yields a collapse of all the curves onto a unique $L-$independent function, which represents the scaling function $g(x)$. Figure 2 shows this for the same books and parts as in Fig. 1. The data collapse can be considered excellent, except for the smallest frequencies. For the largest L the collapse is valid up to $n \simeq 3$ if we exclude *La Regenta*, which only collapses for about $n \geq 6$. So, our scaling hypothesis is validated, independently of the particular shape that $g(x)$ takes. Note that $g(x)$ is independent on L but not on the book, i.e., each book has its own $g(x)$, different from the rest. In any case, we observe a slightly convex shape in log-log space, which leads to the rejection of the power-law hypothesis for the whole range of frequencies. Nevertheless, the data does not show any clear parametric functional form. Any of a double power law, a stretched exponential, a Weibull, or a lognormal tail could be fit to the distributions. This is not incompatible with the fact that the large n tail can be well fit by a power law (the Zipf's law), for more than 2 orders of magnitude [9].

Things turn out to be somewhat different after the lemmatization process. The scaling ansatz is still clearly valid for the frequency distributions, see Fig. 3, but with a different kind of scaling function $g(x)$, with a more defined characteristic shape, due to a more pronounced log-log curvature or convexity. In fact, close comparison of the data leads to conclude that the lemmatization process enhances the goodness of the scaling approximation, specially in the low frequency zone. It could be reasoned that, as lemmatized texts have a significantly reduced vocabulary compared to the original ones, but the total length remains essentially the same, they are somehow equivalent to much longer texts, if one considers the length-to-vocabulary ratio. Although this matter needs to be further investigated, it supports the idea that our main hypothesis, the scale-invariance of the distribution of frequencies, holds more strongly for longer texts.

Due to the clear curvature of $g(x)$ in the lemmatized case, we go one step further and propose a concrete function to fit these data, namely,

$$(2) \qquad g(x) = \frac{k}{x(a + x^{\gamma-1})}.$$

This function has two free parameters, a and γ (with $\gamma > 1$ and $a > 0$), and behaves as a double power law, that is, for large x, $g(x) \sim x^{-\gamma}$ (we still have Zipf's law), while for small x, $g(x) \sim x^{-1}$. The transition point between both power-law tails is determined by a, and k is fixed by normalization. But an important issue is that it is not $g(x)$ which is normalized to one but $D_L(n)$. We select a power-law with exponent one for small x for three reasons: first, in order to explore an alternative to the power law in the V_L versus L relation (which is not clearly supported by data, see next section); second, to compare better our results with those of Ref. [5]; and third, to keep the number of parameters
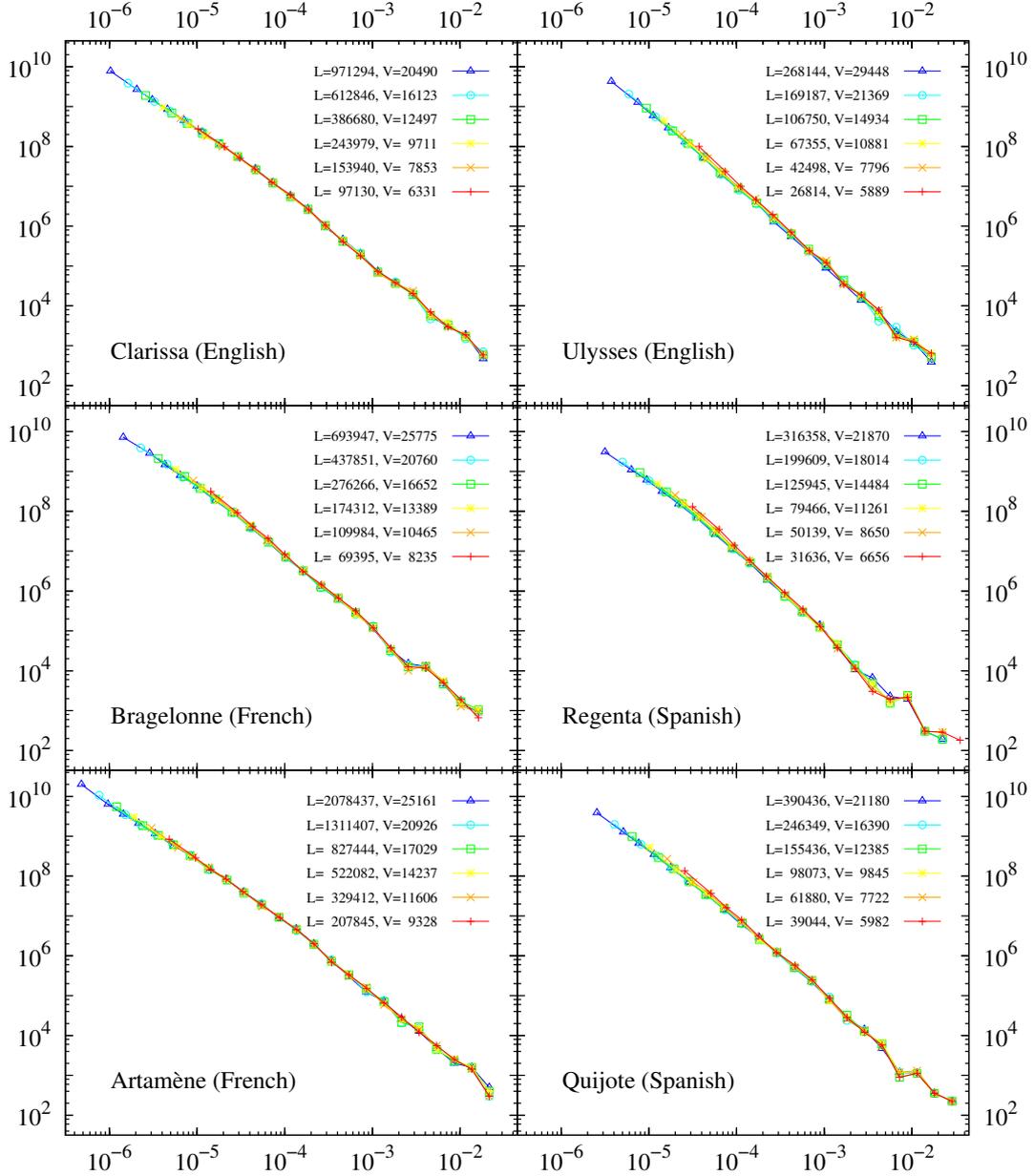
FIGURE 2.   Rescaled densities $LV_L D_L(n)$ (y-axis) against relative frequency n/L (x-axis), for the same books and fractions of text as in Fig. 1. As it is apparent, the rescaled densities collapse onto a single function, independently of the value of L, validating our proposed scaling form for $D_L(n)$ [Eq.(1)] and making clear that the decrease of the log-log slope with L is not a consequence of a genuine change in the scaling properties of the distribution.

minimum. Thus, we do not look for the most accurate fit but for the simplest description of the data. Although double power laws have been previously fit to rank-frequency plots for unlemmatized corpora [17, 21], the resulting exponents for large ranks (low frequencies) are different than for our lemmatized texts.

Then, defining $n_a = a^{\frac{1}{\gamma-1}}L$, the corresponding word-frequency density turns out to be

$$(3) \qquad D_L(n) \propto \frac{1}{n\left(1 + (n/n_a)^{\gamma-1}\right)},$$

with n_a the scale parameter (the scale parameter of $g(x)$ was $a^{\frac{1}{\gamma-1}}$).

The data collapse in Fig. 3 and the good fit imply that the Zipf-like exponent γ does not depend on L, but the transition point between both power laws, n_a, obviously does. Hence, as L grows the transition to the $\sim n^{-\gamma}$ regime occurs at higher absolute frequencies, given by n_a, but fixed relative frequencies, given by $a^{\frac{1}{\gamma-1}}$. In Table 2 we report the fitted parameters for all seven books, obtained by maximum likelihood estimation for the frequencies of the whole books, as well as Monte Carlo estimates of their uncertainties. We have confirmed the stability of γ fitting only a power-law tail from a fixed common relative frequency, for different values of L [9].

title	$n_a \pm \sigma_{n_a}$	$\gamma \pm \sigma_\gamma$	$a \pm \sigma_a$
Artamène[l]	129.7 ± 12.6	1.807 ± 0.026	$(4.65 \pm 0.91) \cdot 10^{-4}$
Clarissa[l]	32.70 ± 2.17	1.864 ± 0.021	$(1.40 \pm 0.24) \cdot 10^{-4}$
Don Quijote[l]	7.91 ± 0.75	1.827 ± 0.020	$(1.35 \pm 0.22) \cdot 10^{-4}$
La Regenta[l]	9.45 ± 0.66	1.983 ± 0.021	$(3.68 \pm 0.62) \cdot 10^{-5}$
Bragelonne[l]	14.56 ± 1.23	1.866 ± 0.018	$(9.10 \pm 1.37) \cdot 10^{-5}$
Moby-Dick[l]	8.21 ± 0.53	2.050 ± 0.024	$(2.42 \pm 0.47) \cdot 10^{-5}$
Ulysses[l]	5.38 ± 0.31	2.020 ± 0.017	$(1.79 \pm 0.28) \cdot 10^{-5}$

TABLE 2. Values of the parameters n_a, γ, and a for the lemmatized versions (indicated with the superscript l) of the seven complete books. The fits are performed numerically through maximum likelihood estimation, while the standard deviations come from Monte Carlo simulations, see Appendix B.

Regarding the low-frequency exponent, one could find a better fit if the exponent were not fixed to be one; however, our data does not allow to constrain well
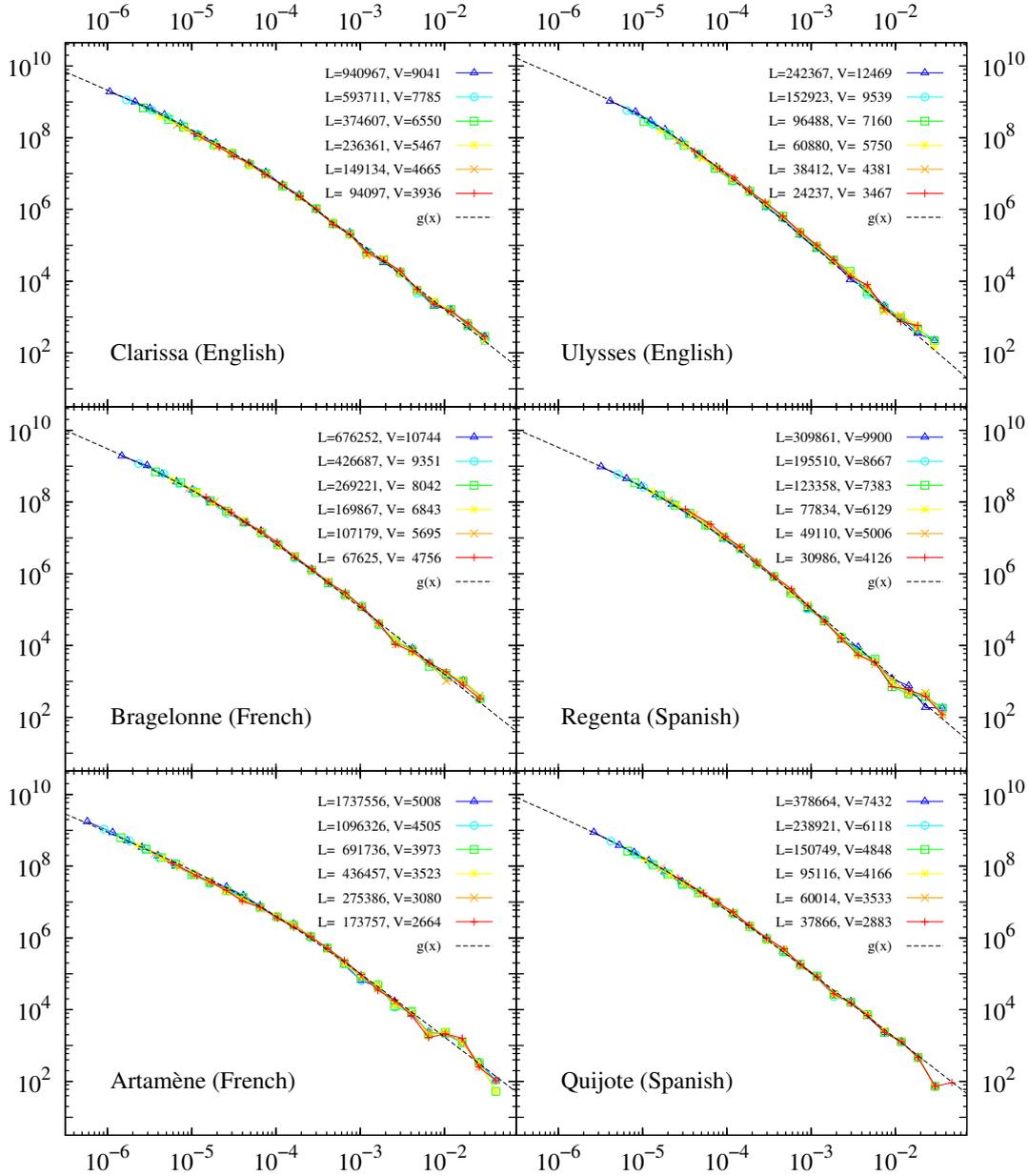
FIGURE 3. Same rescaled distributions as in previous figure ($LV_L D_L(n)$ versus n/L), but for the frequencies of lemmas. The data collapse guarantees the fulfillment of the scaling law also in this case. The fit resulting from the double power-law distribution, Eq. (2), is also included.

this value. A more important point is the influence of lemmatization errors in the characteristics of the low-frequency regime. Although the tools we use are rather accurate, rare words are likely to be assigned a wrong lemma. This limitation is intrinsic to current computational tools and has to be considered as a part of the lemmatization process. Nevertheless, the fact that the behavior at low frequencies is robust in front of a large variation in the percentage of lemmatization errors implies that our result is a genuine consequence of the lemmatization. See Appendix A for more details.

4. An asymptotic approximation of Heaps' law

Coming back to our scaling ansatz, Eq. (1), we notice that the normalization of $D_L(n)$ will allow us to establish a relationship between the word-frequency distribution and the growth of the vocabulary with text length. In the continuous approximation,

$$1 = \int_1^\infty D_L(n)dn = \frac{1}{V_L}\int_1^\infty g(n/L)\frac{dn}{L} = \frac{1}{V_L}\int_{1/L}^\infty g(x)dx = \frac{1}{V_L}G\left(\frac{1}{L}\right),$$

where we have used the previous relation $g(x) = -G'(x)$, and have additionally imposed $G(\infty) \equiv 0$, for which it is necessary that $g(x)$ decays faster than a power law with exponent one. So,

$$(4) \qquad\qquad V_L = G\left(\frac{1}{L}\right).$$

This just means that the number of words with relative frequency greater or equal than $1/L$ is the vocabulary size V_L, as this is the largest rank for a text of length L. It is important to notice the difference between saying that $G_L(1/L) = V_L$, which is a trivial statement, and stating that $G(1/L) = V_L$, which provides a link between Zipf's and Heaps' law, or more generally, between the distribution of frequencies and the vocabulary growth, by approximating the latter by the former. The quality of such an approximation will depend, of course, on the goodness of the scale-invariance approximation. In the usual case of a power-law distribution of frequencies extending to the lowest values, $g(x) \propto 1/x^\gamma$, with $\gamma > 1$, then $G(x) \propto 1/x^{\gamma-1}$, which turns into Heaps' law, $V_L \propto L^\alpha$, with $\alpha = \gamma-1$, in agreement with previous works.

However, this power-law growth of V_L with L is not what is observed in texts. Due to the accurate fit that we can achieve for lemmatized texts, we can explicitly derive an asymptotic expression for V_L given our proposal for $g(x)$. As we have just shown, $g(x)$ is not normalized to one, rather, $\int_{1/L}^\infty g(x)dx = V_L$. Hence, substituting $g(x)$ from Eq. (2) and integrating,

$$V_L = \int_{1/L}^\infty \frac{k}{x(a + x^{\gamma-1})}dx = \frac{k}{a}\int_{1/L}^\infty \frac{ax^{-\gamma}}{ax^{1-\gamma}+1}dx =$$
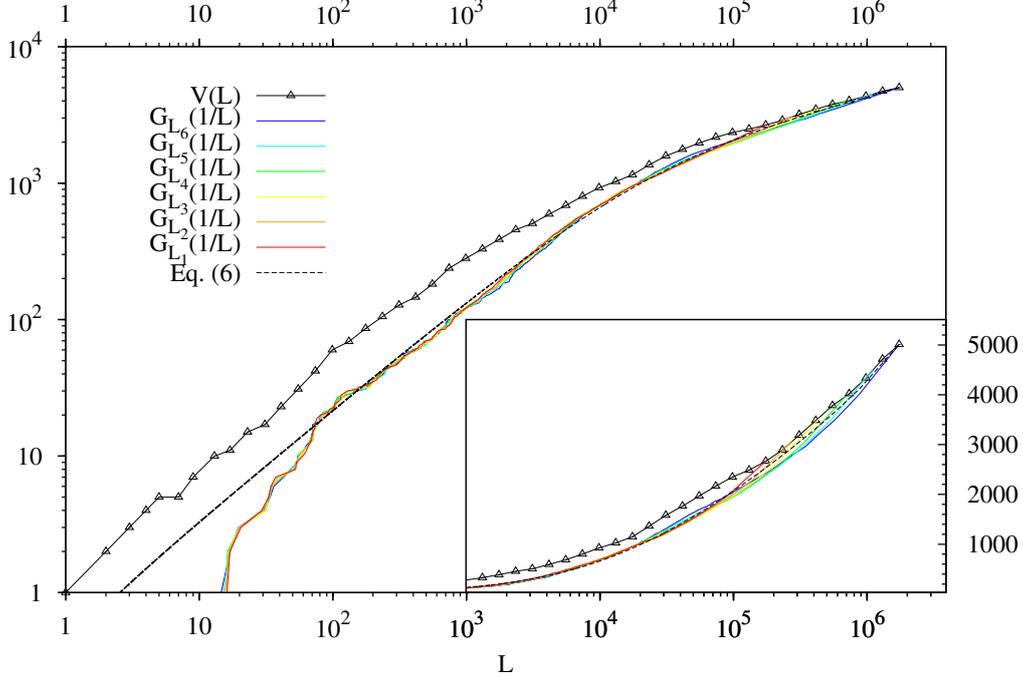
FIGURE 4.   The actual curve V_L (solid black with triangles) for the lemmatized version of the book *Artamène*, together with the curves $V_L = G(1/L)$ obtained by using the empirical inverse of the rank-frequency plot, $r = G(n/L)$, with $L_i = L_{\text{tot}}/10^{(6-i)/5}$ (colors), and the analytical expression Eq. (6) with parameters determined from the fit of $D_{L_{\text{tot}}}(n)$, Eq. (5) (dashed black).

$$(5) \qquad = \frac{k}{a(1-\gamma)} \ln(ax^{1-\gamma}+1)\Big|_{1/L}^{\infty} = \frac{k}{a(\gamma-1)} \ln(aL^{\gamma-1}+1).$$

In this case V_L is not a power law, and behaves asymptotically as $\propto \ln L$. This is a direct consequence of our choice for the exponent 1 in the left-tail of $g(x)$. Indeed, it seems clear that the vocabulary growth curve greatly deviates from a straight line in log-log space, for it displays a prominent convexity, see Fig. 4 as an example. Nevertheless, the result from Eq. (5) is not a good fit either, due to a wrong proportionality constant. This is caused by the continuous approximation in Eq. (5).

For an accurate calculation of V_L we must treat our variables as discrete and compute discrete sums rather than integrals. In the exact, discrete treatment of

$D_L(n)$, equation (5) must be rewritten as

$$V_L = G\left(\frac{1}{L}\right) = G\left(\frac{L_{\text{tot}}/L}{L_{\text{tot}}}\right) = \sum_{n \geq L_{\text{tot}}/L} \frac{g(n/L_{\text{tot}})}{L_{\text{tot}}} =$$

(6)
$$= \frac{1}{L_{\text{tot}}} \sum_{n \geq L_{\text{tot}}/L} \frac{k}{\left(\frac{n}{L_{\text{tot}}}\right)\left(a + \left(\frac{n}{L_{\text{tot}}}\right)^{\gamma-1}\right)}$$

where we have used that $S_{L_{\text{tot}}}(n') = \sum_{n \geq n'} D_{L_{\text{tot}}}(n)$ (notice that in the discrete case, $g(x) \neq -G'(x)$). This is consistent with the fact that, indeed, the maximum likelihood parameters γ and a have been computed assuming a discrete probability function (see Appendix B), and so has the normalization constant. We would like to stress that no fit is performed in Figure 4, that is, the constant k in $g(x)$ is directly derived from the normalizing constant of $D_L(n)$, and depends only on γ and a.

5. Conclusions

In summary, we demonstrate that, contrary to what is claimed in previous works [2, 30, 5], Zipf's law in linguistics is extraordinarily stable under changes in the size of the analyzed text. A scaling function $g(x)$ provides a constant shape for the distribution of frequencies of each text, $D_L(n)$, no matter its length L, which only enters into the distribution as a scale parameter and determining the size of the vocabulary V_L. The apparent size-dependent exponent found previously seems to be an artifact of the slight convexity of $g(x)$ in a log-log plot, which is more clearly observed for very small values of x, accessible only for the largest text lenghts. Moreover, we find that in the case of lemmatized texts the distribution can be well described by a double power-law behavior with a large-frequency exponent γ that does not depend on L, and a transition point n_a that scales linearly with L. The small-frequency exponent is different than the one reported in Ref. [17] for a non-lemmatized corpus. Further, the stability of the shape of the frequency distribution allows one to predict the growth of vocabulary size with text length, resulting in a generalization of the popular Heaps' law. The robustness of Zipf-like parameters under changes in system size opens the way to more practical applications of word statistics. In particular, we provide a consistent way to compare statistical properties of texts with different lengths [3]. Another interesting issue would be the application of the same scaling methods to other fields in which Zipf's law has been proposed to hold, as economics and demography, for instance.

Appendix A. Lemmatization

To analyze the distribution of frequencies of lemmas, we needed to have the texts lemmatized. To manually lemmatize the words would have exceeded the possibilities of this project, so we proceeded to automatic processing with standard computational tools: *FreeLing* [20] for Spanish and English and *TreeTagger* [33] for French. The tools carry out the following steps:

(1) Tokenization: Segmentation of the texts into sentences and sentences into words (tokens).
(2) Morphological analysis: Assignment of one or more lemmas and morphological information (tag) to each token. For instance, *found* in English can correspond to the past tense of the verb *find* or to the base form of the verb *found*. At this stage, both are assigned whenever the word form *found* is encountered.
(3) Morphological disambiguation: An automatic tagger assigns the single most probable lemma and tag to each word form, depending on the context. For instance, in *I found the keys* the tagger would assign the lemma *find* to the word *found*, while in *He promised to found a hospital*, the lemma *found* would be preferred.

All these steps are automatic, such that errors are introduced at each step. However, the accuracy of the tools is quite high (e.g., around 95-97% at the token level for morphological disambiguation), such that a quantitative analysis based on the results of the automatic process can be carried out. Also note that step 2 is based on a pre-existing dictionary (of words, not of lemmas, also called a lexicon): only the words that are in the dictionary are assigned a reliable set of morphological tags and lemmas. Although most of the tools used heuristically assign tag and/or lemma information to words that are not in the dictionary, we only count tokens of lemmas for which the corresponding word types are found in the dictionary, so as to minimize the amount of error introduced by the automatic processing. This comes at the expense of losing some data. However, the dictionaries have quite a good coverage of the vocabulary, particularly at the token level, but also at the type level (see Table A). The exceptions are *Ulysses*, because of the stream of consciousness prose, which uses many non-standard word forms, and *Artamène*, because 17th century French contains many word forms that a dictionary of modern French does not include.

Appendix B. Maximum likelihood fitting

The fitted values of Table 2 have been obtained by maximum-likelihood estimation (MLE). This well-known procedure consists firstly in computing the log-likelihood function $\mathcal{L}$, which in our case reads,

title	types	tokens
Clarissa	68.0 %	96.9 %
Moby-Dick	70.8 %	94.7 %
Ulysses	58.6 %	90.4 %
Don Quijote	81.3 %	97.0 %
La Regenta	89.5 %	97.9 %
Artamène	43.6 %	83.6 %
Bragelonne	89.8 %	97.5 %
Seitsemän v.	89.8 %	95.4 %
Kevät ja t.	96.2 %	98.3 %
Vanhempieni r.	96.5 %	98.5 %
average	78.4 %	95.0 %

TABLE 3. Coverage of the vocabulary by the dictionary in each language, both at the type and at the token level. Remember that we distinguish between a word *type* (corresponding to its orthographic form) and its *tokens* (actual occurrences in text).

$$\mathcal{L} = \frac{1}{V_L} \sum_{i=1}^{V_L} \ln D_L(n_i) = \ln K - \frac{1}{V_L} \sum_{i=1}^{V_L} \ln \left(n_i(b + n_i^{\gamma-1}) \right)$$

with n_i the V_L values of the frequency and the normalization constant K in the discrete case equal to

$$K = \left[\sum_{i=1}^{V_L} \frac{1}{n_i(b + n_i^{\gamma-1})} \right]^{-1}.$$

Note that we have reparameterized the distribution with regard the main text, introducing $b = n_a^{\gamma-1} = aL^{\gamma-1}$. Then, $\mathcal{L}$ is maximized with respect to the parameters γ and b; this has been done numerically using the simplex method [31]. The error terms σ_γ and σ_b, representing the standard deviation of each estimator, are computed from Monte Carlo simulations: from the resulting maximum-likelihood parameters γ^* and b^*, synthetic data samples are simulated, and the MLE parameters of these samples are calculated in the same way; their fluctuations yield σ_γ and σ_b. We stress that no continuous approximation has been made, that is, the simulated data follows the discrete probability function $D_L(n)$ (this is done

using the rejection method, see Ref. [11, 9] for details for a similar case). In a summarized recipe, the procedure simply is:

(1) Numerically compute the MLE parameters, γ^* and b^*.
(2) Draw M datasets, each of size V_L, from the discrete probability function $D_L(n; \gamma^*, b^*)$.
(3) For each dataset $m = 1 \ldots M$, compute the MLE parameters γ^m, b^m.
(4) Compute the standard deviations σ_γ and σ_b of the sets $\{\gamma^m\}_{m=1}^M$ and $\{b^m\}_{m=1}^M$.

The standard deviations of n_a and a are computed in the same way using their relationship with b and γ.

References

[1] L. A. Adamic and B. A. Huberman. Zipfs law and the Internet. *Glottometrics*, 3:143–150, 2002.
[2] H. Baayen. *Word Frequency Distributions*. Kluwer, Dordrecht, 2001.
[3] J Baixeres, B. Elvevåg, and R. Ferrer i Cancho. The evolution of the exponent of zipf's law in language ontogeny. *PLoS ONE, in press*, 2013.
[4] P. Bak. *How Nature Works: The Science of Self-Organized Criticality*. Copernicus, New York, 1996.
[5] Sebastian Bernhardsson, Luis Enrique Correa da Rocha, and Petter Minnhagen. The meta book and size-dependent properties of written language. *New J. Phys.*, 11(12):123015, 2009.
[6] K. Christensen and N. R. Moloney. *Complexity and Criticality*. Imperial College Press, London, 2005.
[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.
[8] Bernat Corominas-Murtra, Jordi Fortuny, and Ricard V. Solé. Emergence of Zipf's Law in the Evolution of Communication. *Phys. Rev. E*, 83:036115, Mar 2011.
[9] A. Corral, G. Boleda, and R. Ferrer-i-Cancho. *in preparation*, 2013.
[10] A. Corral and F. Font-Clos. Criticality and self-organization in branching processes: application to natural hazards. In M. Aschwanden, editor, *Self-Organized Critical Phenomena*, 2013.
[11] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.
[12] Ronald Dickman, Nicholas R. Moloney, and Eduardo G. Altmann. Analysis of an information-theoretic model for communication. *J. Stat. Mech: Theory Exp.*, 2012(12): P12022+, December 2012.

[13] Bertram Dring, Daniel Matthes, and Giuseppe Toscani. A boltzmann-type approach to the formation of wealth distribution curves. *Notes of the Porto Ercole School*, 2008.

[14] R. Ferrer i Cancho. The variation of Zipf's law in human language. *Eur. Phys. J. B*, 44:249–257, 2005.

[15] R. Ferrer i Cancho. Zipf's law from a communicative phase transition. *Eur. Phys. J. B*, 47:449–457, 2005.

[16] R. Ferrer i Cancho and A Hernández-Fernández. Power laws and the golden number. In G. Altmann, I. Zadorozhna, and Y Matskulyak, editors, *Problems of general, germanic and slavic linguistics*, number Books - XII, pages 518–523. Chernivtsi, 2008.

[17] 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.

[18] 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.

[19] Ramon Ferrer i Cancho and Brita Elvevåg. Random texts do not exhibit the real zipf's law-like rank distribution. *PLoS ONE*, 5(3):e9411, 03 2010.

[20] FreeLing.

[21] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages, December 2012.

[22] Alex Hankey and H. Eugene Stanley. Systematic application of generalized homogeneous functions to static scaling, dynamic scaling, and universality. *Phys. Rev. B*, 6:3515–3542, Nov 1972.

[23] G. Herdan. *Quantitative Linguistics*. Butterworths, 1964.

[24] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on*, 38(6):1842 –1845, nov 1992.

[25] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.

[26] B. B. Mandelbrot. On the theory of word frequencies and on related Markovian models of discourse. In R. Jacobsen, editor, *Structures of Language and its Mathematical Aspects*. American Mathematical Society, New York, 1961.

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

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

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

[30] David M. W. Powers. Applications and explanations of zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, pages 151–160, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

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

[32] 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.

[33] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Citeseer, Manchester, 1994.

[34] M. A. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE*, 4(4):e5372, 04 2009.

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

[36] Dietrich Stauffer and Ammon Aharony. *Introduction To Percolation Theory*. CRC Press, 2 edition, July 1994.

[37] D.C. van Leijenhorst and Th.P. van der Weide. A formal derivation of heaps' law. *Inform. Sciences*, 170(24):263 – 272, 2005.

[38] D. Zanette. *Statistical Patterns in Written Language*. 2012.

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

[40] S. Zapperi, K. B. Lauritsen, and H. E. Stanley. Self-organized branching processes: Mean-field theory for avalanches. *Phys. Rev. Lett.*, 75:4071–4074, 1995.

[41] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

Francesc Font-Clos
Centre de Recerca Matemàtica
Edifici C, Campus Bellaterra
E-08193 Bellaterra (Barcelona), Spain
and
Departament de Matemàtiques
Universitat Autònoma de Barcelona
Edifici C
E-08193 Bellaterra (Barcelona), Spain

Gemma Boleda
Department of Linguistics
The University of Texas at Austin
University Station
B5100, Austin, TX, USA

Álvaro Corral
Centre de Recerca Matemàtica
Edifici C, Campus Bellaterra
E-08193 Bellaterra (Barcelona), Spain