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



2013 New J. Phys. 15 093033

(http://iopscience.iop.org/1367-2630/15/9/093033)

New Journal of Physics
The open access journal for physics

# A scaling law beyond Zipf's law and its relation to Heaps' law

**Francesc Font-Clos**[1,2,4], **Gemma Boleda**[3] **and Álvaro Corral**[1]

[1] Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Bellaterra, Barcelona, Spain
[2] Department de Matemàtiques, Universitat Autònoma de Barcelona, Edifici C, E-08193 Bellaterra, Barcelona, Spain
[3] Department of Linguistics, The University of Texas at Austin, 1 University Station B5100, Austin, TX, USA
E-mail: fontclos@crm.cat

*New Journal of Physics* **15** (2013) 093033 (16pp)
Received 4 March 2013
Published 23 September 2013
Online at http://www.njp.org/
doi:10.1088/1367-2630/15/9/093033

**Abstract.** The dependence on text length of the statistical properties of word occurrences has long been considered a severe limitation on the usefulness of quantitative linguistics. We propose a simple scaling form for the distribution of absolute word frequencies that brings to light 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 that 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. In the latter case, the distribution of frequencies is well approximated 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 a generic alternative to Heaps' law, which turns out to be intimately connected to the distribution of frequencies, thanks to its scaling behavior.

[4] Author to whom any correspondence should be addressed.

## Contents



## 1. Introduction

Zipf's law is perhaps one of the best pieces of evidence about the existence of universal physical-like laws in cognitive science and the social sciences. Classic examples where it applies include the population of cities, company income and the frequency of words in texts or speech [1]. In 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. When a power-law relation

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

holds for a large enough range, with the exponent $\beta$ more or less close to 1, Zipf's law is considered to be fulfilled (with $\propto$ denoting proportionality). An equivalent formulation of the law is obtained in terms of the probability distribution of the frequency $n$, such that it plays 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 $D(n)$ as the probability mass function of $n$ (or the probability density of $n$, in a continuous approximation) [2–6]. Note that this formulation implies performing double statistics (i.e. doing statistics twice), first counting words to get frequencies and then counting repetition of frequencies to get the distribution of frequencies.

The criteria for the validity of Zipf's law are arguably rather vague (long enough text, large enough range, exponent $\beta$ 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 $\beta$ to 1 translates into an interval (0.7,1.2), or even beyond that [6–8]. Moreover, no rigorous methods have been usually required for the fitting of the power-law distribution. Linear regression in double-logarithmic scale is the most common method, either for $n(r)$ or for $D(n)$, despite the fact that it is well known that this procedure suffers from severe drawbacks and can lead to flawed results [9, 10]. Nevertheless, once these limitations are assumed, the fulfillment of Zipf's law in linguistics is astonishing, being valid no matter the author, style or language [1, 6, 7]. So, the law is universal, at least in a qualitative sense.

At a theoretical level, many different competing explanations of Zipf's law have been proposed [6], such as random (monkey) typing [11, 12], preferential repetitions or proportional growth [13–15], the principle of least effort [1, 16–18], and, beyond linguistics, Boltzmann-type approaches [19] or even avalanche dynamics in a critical system [20]; most of these options have generated considerable controversy [21–23]. 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 more 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 [24]. 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 [25, 26]), 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 the exponent $\alpha$ smaller than one. However, even simple log–log plots of $V_L$ versus $L$ do not show a convincing linear behavior [27] and therefore, the evidence for this law is somewhat weak (for a notable exception see [5]). Nevertheless, a number of works have derived the relationship $\beta = 1/\alpha$ between Zipf's and Heaps' exponents [2, 5, 28], at least in the infinite-system limit [29, 30], using different assumptions.

Despite the relevance of Zipf's law, and its possible relations with criticality, few systematic studies about the dependence of the law on system size (i.e. text length) have been carried out. It was Zipf himself [1, pp. 144] who first observed a variation in the exponent $\beta$ 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 $\beta$ with $L$ has been confirmed later, see [25, 31], leading to the conclusion that the practical usefulness of Zipf's law is rather limited [25].

More recently, using rather large collections of books from single authors, Bernhardsson *et al* [32] find a decrease of the exponents $\gamma$ and $\alpha$ with text length, in correspondence with the increase in $\beta$ 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 however that this is not an assumption in itself, just notation, and it 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

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

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, $\langle n(L)\rangle \propto L^{2-\gamma(L)}$ [24], which comparing to assumption (i) leads to

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

so, $0 < \alpha(L) < 1$. This relationship between $\alpha$ and $\gamma$ is in agreement with previous results if $L$ is fixed [2, 29, 30]. It was claimed in [32] that $\alpha(L)$ decreases from 1 to 0 for increasing $L$ and therefore $\gamma(L)$ decreases from 2 to 1. The resulting functional form

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

is in fact the same functional form 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 [24, 33–35]. Note that this implies that the tail of the frequency distribution is not a power law but an exponential one, 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 modeled by a power law [9, 36]. However, what is most uncommon about this functional form is the fact that it has a 'critical' exponent that depends on system size. The values of exponents should not be influenced by external scales. So, here we look for an alternative picture that is more in 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 of $L$ and $V_L$, and only the *scale* of $D_L(n)$ changes with these variables. This implies that the shape parameters of $D_L(n)$ (in particular, any exponent) do not change with $L$; only one scale parameter changes with $L$, increasing linearly. This is explained in section 2, while section 3 one is devoted to the validation of our scaling form in real texts, using both plain words and their corresponding lemma forms; in the latter case an alternative to Zipf's law can be proposed, consisting of a double power-law distribution (which is a distribution with two power-law regimes that have different exponents). Our findings for words and lemmas suggest that the previous observation that the exponent in Zipf's law depends on text length [25, 31, 32], might be an artifact of the increasing weight of a second regime in the distribution of frequencies beyond a certain text length. Section 4 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 deemed more appropriate from a statistical point of view [36].

## 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 type 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 assumptions introduced by Bernhardsson *et al* we just need one assumption, which is the independence of the function $G_L$ with respect to $L$; so

$$r = G(n/L). \tag{1}$$

This turns out to be a scaling law, with $G(x)$ a scaling function. It means that if in the first 10 000 tokens of a book there are five types 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 tokens, and for the first 100 000 and for the whole book. These types need not necessarily be the same ones, although in some cases they might be. In fact, instead of assuming as in [32] that the frequency of the most used type scales linearly with $L$, what we assume is just that this is true for all types, at least on average. Notice that this is not a straightforward assumption, as, for instance [5], considers instead that $n$ is just a (particular) function of $r/V_L$.

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 $S_L(n) = \text{Prob}[\text{frequency} \geqslant n]$. 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 $S_L(n)$:

$$D_L(n) = -\frac{\partial S_L(n)}{\partial n} = \frac{g(n/L)}{LV_L}, \tag{2}$$

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, $D_L(n) = g(n/L)/L^{1+\alpha}$, which fulfills invariance under a scaling transformation, or, equivalently, fulfills the definition of a generalized homogeneous function [24, 37]

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

where $\lambda_L$, $\lambda_n$ and $\lambda_D$ are the scale factors, related in this case through

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

and

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

**Table 1.** Total text length and vocabulary before ($L_{\rm tot}$, $V_{\rm tot}$) and after ($L^{(\rm l)}_{\rm tot}$, $V^{(\rm l)}_{\rm tot}$) 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.

| Title | Author | Language | Year | $L_{\rm tot}$ | $V_{\rm tot}$ | $L^{(\rm l)}_{\rm tot}$ | $V^{(\rm l)}_{\rm tot}$ |
|---|---|---|---|---|---|---|---|
| Artamène | Scudéry siblings | French | 1649 | 2 078 437 | 25 161 | 1 737 556 | 5008 |
| Clarissa | Samuel Richardson | English | 1748 | 971 294 | 20 490 | 940 967 | 9041 |
| Don Quijote | Miguel de Cervantes | Spanish | 1605–1615 | 390 436 | 21 180 | 378 664 | 7432 |
| La Regenta | L Alas 'Clarín' | Spanish | 1884 | 316 358 | 21 870 | 309 861 | 9900 |
| Le Vicomte de 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 | 9141 |
| Ulysses | James Joyce | English | 1918 | 268 144 | 29 448 | 242 367 | 12 469 |

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 $\lambda_D$, which is

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

So, $D_L(n)$ is not a generalized homogeneous function, but presents an even more general form. In any case, the validity of the proposed scaling law, equation (1), can be checked by performing a very simple rescaled plot, displaying $LV_L D_L(n)$ versus $n/L$. A resulting data collapse support the independence of the scaling function with respect to $L$. This is undertaken in section 3.

## 3. Data analysis results

To test the validity of our predictions, summarized in equation (2), 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 statistics of homogeneous texts as possible). In addition to the statistics of the words in the texts, we consider the statistics of lemmas (roughly speaking, the stem forms of the word; for instance, *dog* for *dogs*). In the lemmatized version of each text, each word is substituted by its corresponding lemma, and the statistics are collected in the same way as they are collected for word forms. Appendix A provides detailed information on the lemmatization procedure, and table 1 summarizes the most relevant characteristics of the analyzed books.

First, we plot the distributions of word frequencies, $D_L(n)$ versus $n$, for each book, considering either the whole book or the first $L/L_{\rm tot}$ fraction, where $L_{\rm tot}$ is the real, complete text length (i.e. if $L = L_{\rm tot}/2$ 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 figure 1. The pattern described by Bernhardsson *et al* (equivalent to Zipf's findings for the change of the exponent $\beta$) seems to hold, as the absolute value of the slope in log–log scale (i.e. the apparent power-law exponent $\gamma$) decreases with increasing text length.

However, a scaling analysis reveals an alternative picture. As suggested by equation (2), plotting $LV_L D_L(n)$ against $n/L$ for different values of $L$ yields a collapse of all the curves onto

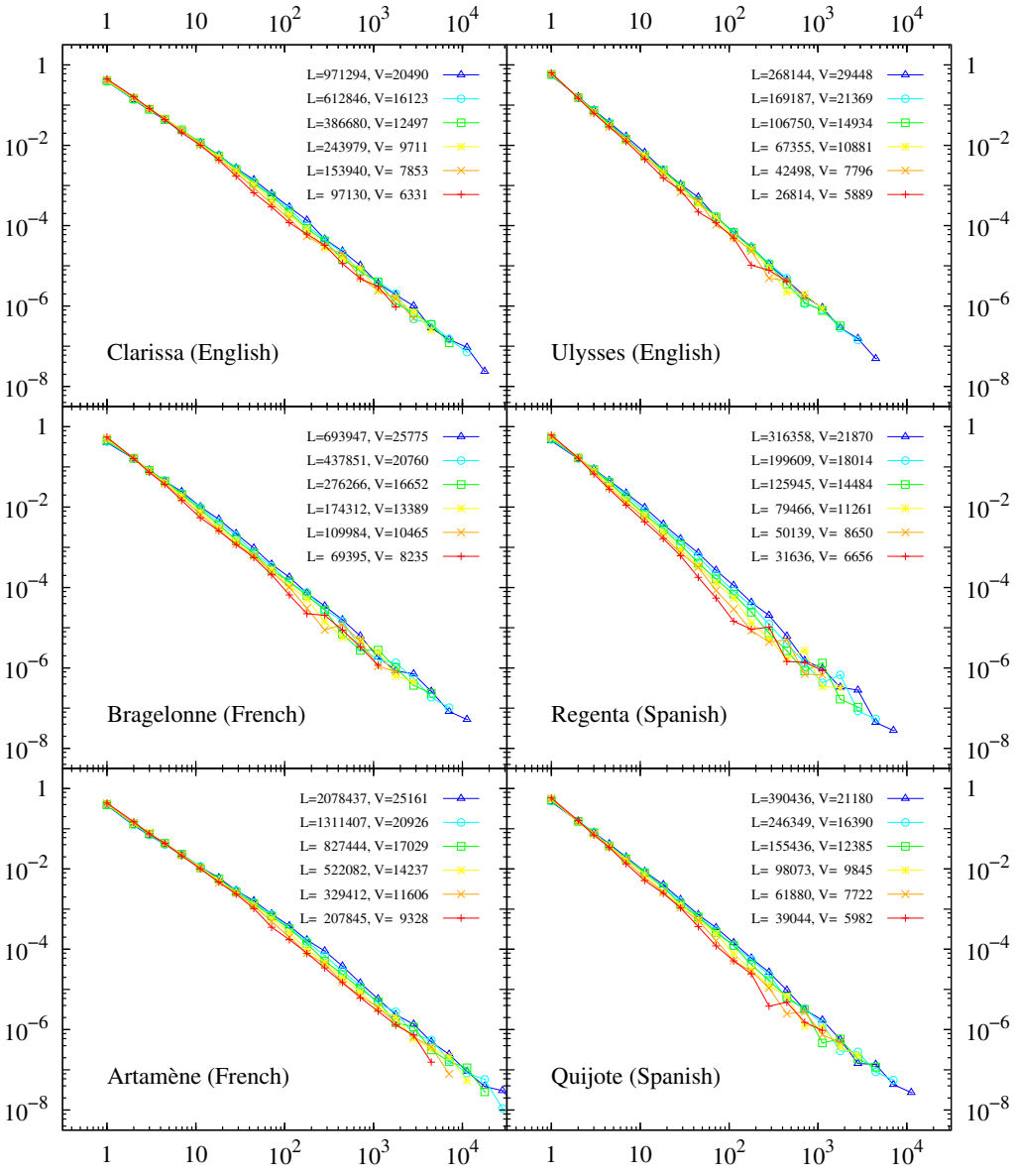


**Figure 1.** Density of word frequencies $D_L(n)$ ($y$-axis) against absolute frequency $n$ ($x$-axis), for six different books, taking text length $L = L_{\text{tot}}/10$, $L_{\text{tot}}/10^{4/5}$, $L_{\text{tot}}/10^{3/5}$, ..., $L_{\text{tot}}$. The slope seems to decrease with text length.

a unique $L$-independent function for each book, which represents the scaling function $g(x)$. Figure 2 shows this for the same books and parts of the books as in figure 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 $n \geqslant 6$. So, our scaling hypothesis is validated, independently of the particular shape that $g(x)$ takes. Note that $g(x)$ is independent of $L$ but not 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. A double power law, a stretched exponential, a Weibull

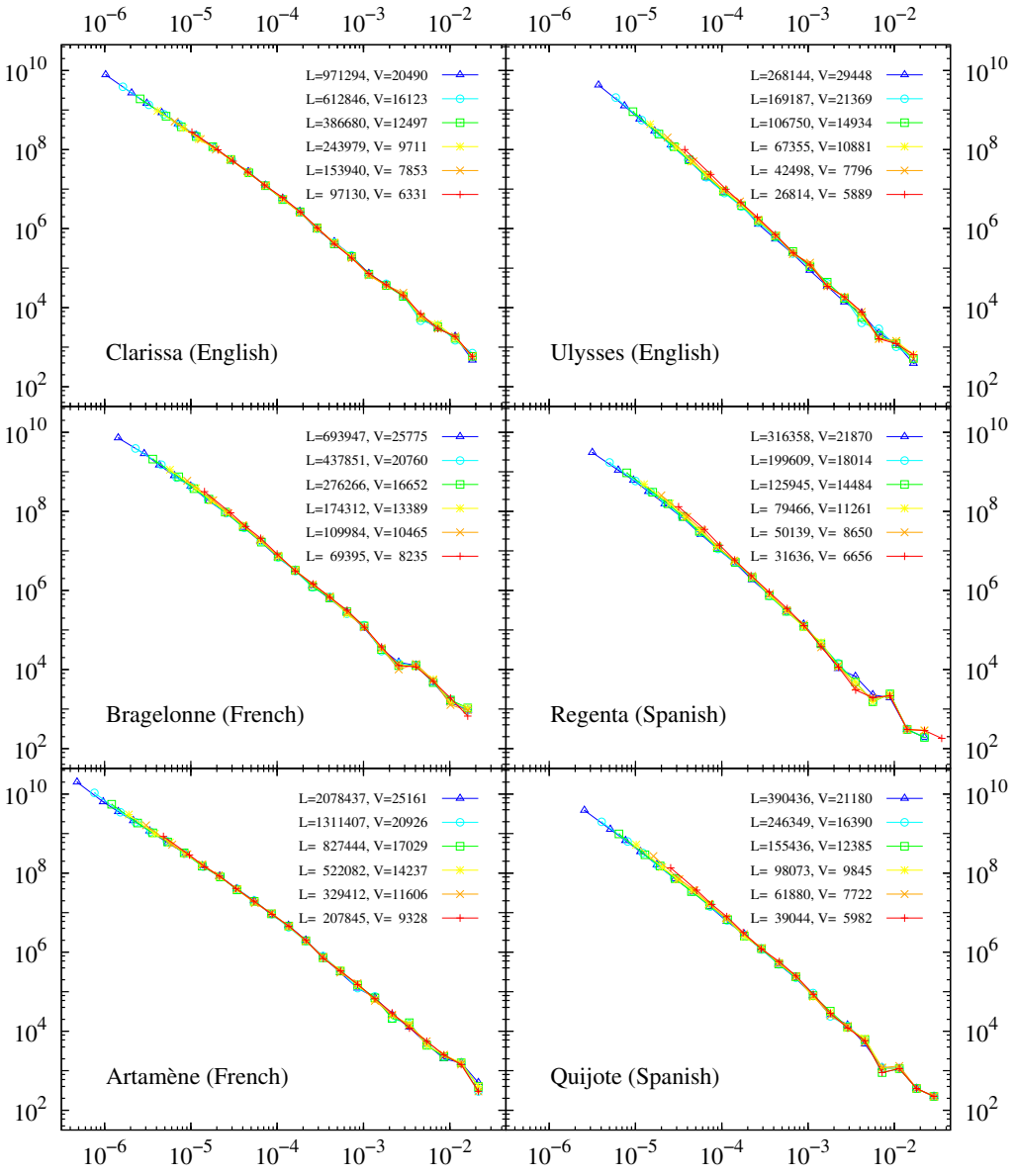


**Figure 2.** Rescaled densities $LV_L D_L(n)$ ($y$-axis) against relative frequency $n/L$ ($x$-axis), for the same books and fractions of text as in figure 1. The rescaled densities collapse onto a single function, independently of the value of $L$, validating our proposed scaling form for $D_L(n)$ (equation (2)) and making it 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.

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 two orders of magnitude [36].

Things turn out to be somewhat different after the lemmatization process. The scaling ansatz is still clearly valid for the frequency distributions, see figure 3, but with a different kind of scaling function $g(x)$, with a more defined characteristic shape, due to a more pronounced

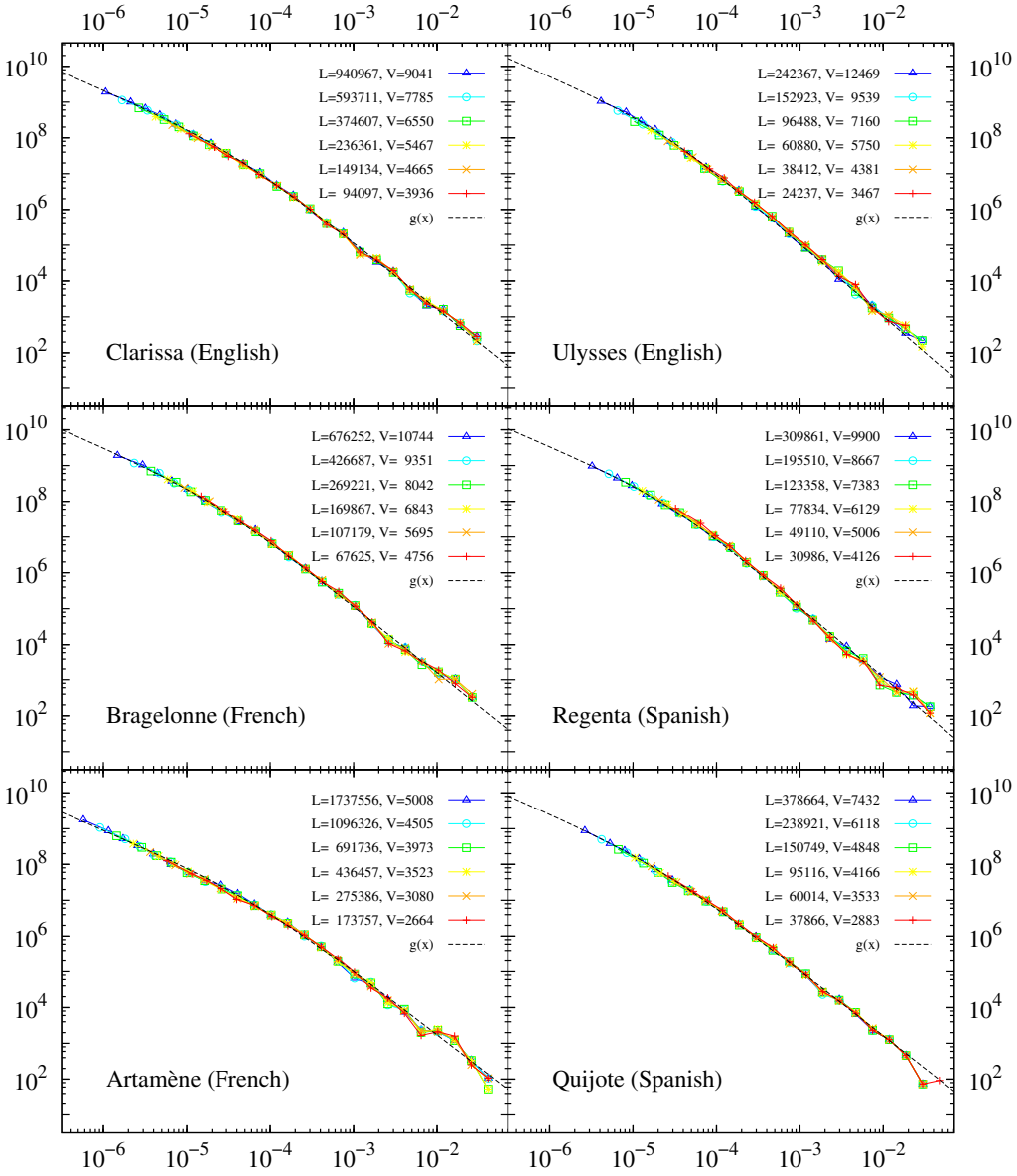


**Figure 3.** Same rescaled distributions as in previous figure ($LV_L D_L(n)$ versus $n/L$), but for the frequencies of lemmas. The data collapse guarantees the fulfillment of the scaling law also in this case. The fit resulting from the double power-law distribution, equation (3), is also included.

log–log curvature or convexity. In fact, close examination of the data leads us 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.

**Table 2.** Values of the parameters $n_a$, $\gamma$ and $a$ for the lemmatized versions (indicated with the superscript (l)) of the seven complete books. The fits are performed numerically through MLE, while the standard deviations come from Monte Carlo simulations, see appendix B.

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

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

$$g(x) = \frac{k}{x(a + x^{\gamma-1})}. \tag{3}$$

This function has two free parameters, $a$ and $\gamma$ (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 $a$ (more precisely, by $a^{\frac{1}{\gamma-1}}$), 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: firstly, in order to explore an alternative to the power law in the $V_L$ versus $L$ relation (which is not clearly supported by the data, see next section); secondly, to allow for a better comparison of our results and those of [32]; thirdly, to keep the number of parameters minimum. Thus, we do not look for the most accurate fit but for the simplest description of the data.

Then, defining $n_a = a^{\frac{1}{\gamma-1}} L$, the corresponding word-frequency density (or, more properly, lemma-frequency density or type-frequency density) turns out to be

$$D_L(n) \propto \frac{1}{n\left(1 + (n/n_a)^{\gamma-1}\right)} \tag{4}$$

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

The data collapse in figure 3 and the good fit imply that the Zipf-like exponent $\gamma$ 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 (MLE) of the frequencies of the whole books, as well as Monte Carlo estimates of their uncertainties. We have confirmed the stability of $\gamma$ fitting only a power-law tail from a fixed common relative frequency, for different values of $L$ [36].

Regarding the low-frequency exponent, one could find a better fit if the exponent was not fixed to be one; however, our data does not allow this value to be well constrained. 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.

Although double power laws have been previously fit to rank-frequency plots for unlemmatized multi-author corpora [27, 38, 39], the resulting exponents for large ranks (low frequencies) are different than the ones obtained for our lemmatized single-author texts. Note that [27] also proposed that the crossover between both power laws happened for a constant number of types, around 7900, independently of corpus size. This corresponds indeed to $r = 7900$ and therefore, from equation (1), to a fixed relative frequency. This is certainly in agreement with our results, supporting the hypothesis that rank-frequency plots and frequency distributions are stable in terms of relative frequency.

## 4. An asymptotic approximation of Heaps' law

Coming back to our scaling ansatz, equation (2), 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

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

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,

$$V_L = G\left(\frac{1}{L}\right). \tag{5}$$

This just means, compared to equation (1), that the number of types 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 research [2, 5, 29, 30, 32].

However, this power-law growth of $V_L$ with $L$ is not what is observed in texts, in general. Due to the accurate fit that we can achieve for lemmatized texts, we can explicitly derive an

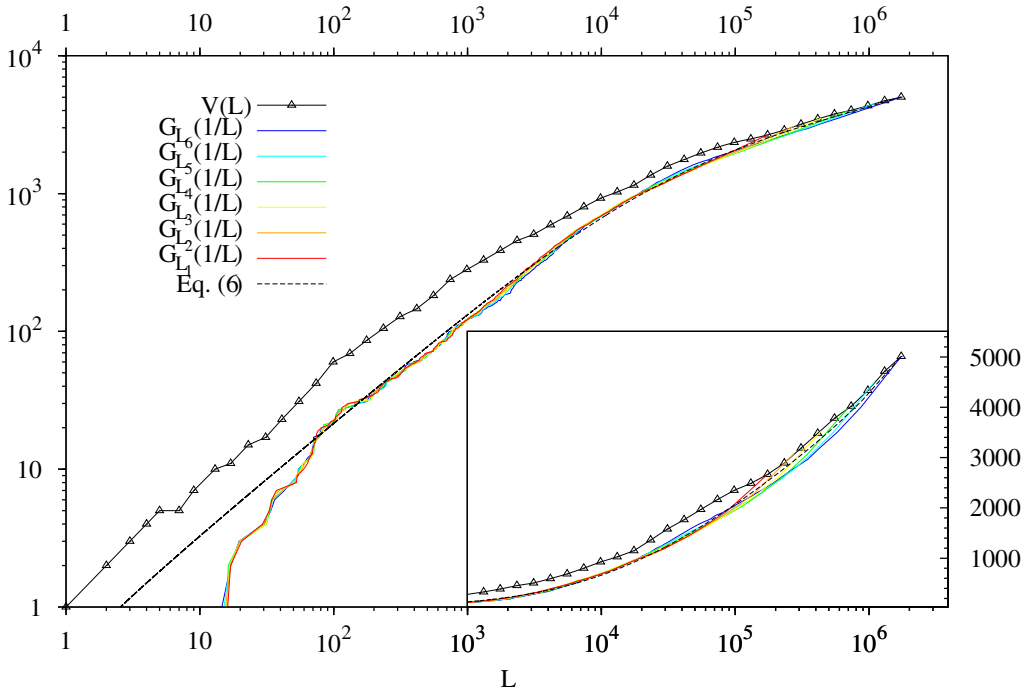


**Figure 4.** The actual curve $V_L$ (solid black with triangles) for the lemmatized version of the book *Artamène*, together with the curves $V_L = G(1/L)$ obtained by using the empirical inverse of the rank-frequency plot, $r = G(n/L)$, with $L_i = L_{\text{tot}}/10^{(6-i)/5}$ (colors), and the analytical expression (7) with parameters determined from the fit of $D_{L_{\text{tot}}}(n)$, equation (6) (dashed black).

asymptotic expression for $V_L$ given our proposal for $g(x)$. As we have just shown, $g(x)$ is not normalized to one, rather, $\int_{1/L}^{\infty} g(x)\,\mathrm{d}x = V_L$. Hence, substituting $g(x)$ from equation (2) and integrating

$$V_L = \int_{1/L}^{\infty} \frac{k}{x(a + x^{\gamma-1})}\mathrm{d}x = \frac{k}{a}\int_{1/L}^{\infty} \frac{ax^{-\gamma}}{ax^{1-\gamma}+1}\mathrm{d}x =$$
$$= \frac{k}{a(1-\gamma)}\ln(ax^{1-\gamma}+1)\Big|_{1/L}^{\infty} = \frac{k}{a(\gamma-1)}\ln(aL^{\gamma-1}+1). \qquad (6)$$

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 figure 4 as an example. Nevertheless, the result from equation (6) is not a good fit either, due to a wrong proportionality constant. This is caused by the continuous approximation in equation (6).

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 (6) must

be rewritten as

$$V_L = G\left(\frac{1}{L}\right) = G\left(\frac{L_{\rm tot}/L}{L_{\rm tot}}\right) = \sum_{n \geqslant L_{\rm tot}/L} \frac{g(n/L_{\rm tot})}{L_{\rm tot}}$$
$$= \frac{1}{L_{\rm tot}} \sum_{n \geqslant L_{\rm tot}/L} \frac{k}{\left(\frac{n}{L_{\rm tot}}\right)\left(a + \left(\frac{n}{L_{\rm tot}}\right)^{\gamma-1}\right)}, \tag{7}$$

where we have used the fact that $S_{L_{\rm tot}}(n') = \sum_{n \geqslant n'} D_{L_{\rm tot}}(n)$, with $n' = L_{\rm tot}/L$ (notice that in the discrete case, $g(x) \neq -G'(x)$). This is consistent with the fact that, indeed, the maximum likelihood parameters $\gamma$ 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 $\gamma$ and $a$.

## 5. Conclusions

In summary, we have shown that, contrary to claims in previous research [25, 31, 32], 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 determines 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 lengths. Moreover, we find that in the case of lemmatized texts the distribution can be well described by a double power law, with a large-frequency exponent $\gamma$ 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 ones reported in [27, 38] for non-lemmatized corpora. 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 [40]. 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.

## Acknowledgments

We appreciate a collaboration with R Ferrer-i-Cancho, who also put AC in contact with GB. Financial support is acknowledged from grants FIS2009-09508 from the Ministerio de Ciencia y Tecnología, FIS2012-31324 from the Ministerio de Economía y Competitividad and 2009-SGR-164 from Generalitat de Catalunya, which also supported FF-C through grant 2012FI_B 00422 and GB through AGAUR grant 2010BP-A00070.

**Table A.1.** 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).

| 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 |

## Appendix A. Lemmatization

To analyze the distribution of frequencies of lemmas, the texts needed to be 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*[5] for Spanish and English and *TreeTagger* [41] 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), so 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.1). The exceptions are *Ulysses*, because

[5] FreeLing (http://nlp.lsi.upc.edu/freeling).

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 MLE. This well-known procedure consists firstly in computing the log-likelihood function $\mathcal{L}$, which in our case reads

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

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

$$K = \left[\sum_{n=1}^{n_{\max}} \frac{1}{n(b+n^{\gamma-1})}\right]^{-1}.$$

Note that we have reparameterized the distribution compared to the main text, introducing $b = n_a^{\gamma-1} = aL^{\gamma-1}$. Then, $\mathcal{L}$ is maximized with respect to the parameters $\gamma$ and $b$; this has been done numerically using the simplex method [42]. The error terms $\sigma_\gamma$ and $\sigma_b$, representing the standard deviation of each estimator, are computed from Monte Carlo simulations. From the resulting maximum-likelihood parameters $\gamma^*$ and $b^*$, synthetic data samples are simulated, and the MLE parameters of these samples are calculated in the same way; their fluctuations yield $\sigma_\gamma$ and $\sigma_b$. 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 [36, 43] for details for a similar case). In a summarized recipe, the procedure simply is:

1. numerically compute the MLE parameters, $\gamma^*$ 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 $\gamma^m$, $b^m$;
4. compute the standard deviations $\sigma_\gamma$ and $\sigma_b$ 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 to $b$ and $\gamma$.

## References

[1] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[2] Mandelbrot B B 1961 *Structures of Language and its Mathematical Aspects* ed R Jacobsen (New York: American Mathematical Society) pp 214–7
[3] Ferrer i Cancho R and Hernández-Fernández A 2008 Power laws and the golden number *Problems of General, Germanic and Slavic Linguistics* ed G Altmann, I Zadorozhna and Y Matskulyak (Chernivtsi: Books–XXI) pp 518–23
[4] Adamic L A and Huberman B A 2002 *Glottometrics* **3** 143
[5] Kornai A 2002 *Glottometrics* **4** 61
[6] Zanette D 2012 *Statistical Patterns in Written Language*
[7] Zanette D and Montemurro M 2005 *J. Quantum Linguist.* **12** 29

[8] Ferrer i Cancho R 2005 *Eur. Phys. J.* B **44** 249
[9] Clauset A, Shalizi C R and Newman M E J 2009 *SIAM Rev.* **51** 661
[10] Corral A, Font F and Camacho J 2011 *Phys. Rev.* E **83** 066103
[11] Miller G A 1957 *Am. J. Psychol.* **70** 311
[12] Li W 1992 *IEEE Trans. Inform. Theory* **38** 1842
[13] Simon H A 1955 *Biometrica* **42** 425
[14] Newman M E J 2005 *Contemp Phys.* **46** 323
[15] Saichev A, Malevergne Y and Sornette D 2010 *Theory of Zipf's Law and Beyond* (*Lecture Notes in Economics and Mathematical Systems* vol 632) (Berlin: Springer)
[16] Ferrer i Cancho R and Solé R V 2003 *Proc. Natl Acad. Sci. USA* **100** 788
[17] Corominas-Murtra B, Fortuny J and Solé R V 2011 *Phys. Rev.* E **83** 036115
[18] Ferrer i Cancho R 2005 *Eur. Phys. J.* B **47** 449
[19] Düring B, Matthes D and Toscani G 2009 *Riv. Mat. Univ. Parma* **1** 199
[20] Bak P 1996 *How Nature Works: The Science of Self-Organized Criticality* (New York: Copernicus)
[21] Mitzenmacher M 2004 *Internet Math.* **1** 226
[22] Ferrer i Cancho R and Elvevåg B 2010 *PLoS ONE* **5** e9411
[23] Dickman R, Moloney N R and Altmann E G 2012 *J. Stat. Mech.* **2012** P12022
[24] Christensen K and Moloney N R 2005 *Complexity and Criticality* (London: Imperial College Press)
[25] Baayen H 2001 *Word Frequency Distributions* (Dordrecht: Kluwer)
[26] Herdan G 1964 *Quantitative Linguistics* (London: Butterworths)
[27] Gerlach M and Altmann E G 2013 *Phys. Rev.* X **3** 021006
[28] van Leijenhorst D and van der Weide T 2005 *Inform. Sci.* **170** 263
[29] Serrano M A, Flammini A and Menczer F 2009 *PLoS ONE* **4** e5372
[30] Lü L, Zhang Z-K and Zhou T 2010 *PLoS ONE* **5** e14139
[31] Powers D M W 1998 *NeMLaP3/CoNLL '98*: *Proc. of the Joint Conf. on New Methods in Language Processing and Computational Natural Language Learning* (Stroudsburg, PA: Association for Computational Linguistics) pp 151–60
[32] Bernhardsson S, da Rocha L E C and Minnhagen P 2009 *New J. Phys.* **11** 123015
[33] Stauffer D and Aharony A 1994 *Introduction To Percolation Theory* 2nd edn (Boca Raton, FL: CRC)
[34] Zapperi S, Lauritsen K B and Stanley H E 1995 *Phys. Rev. Lett.* **75** 4071
[35] Corral A and Font-Clos F 2013 *Self-Organized Critical Phenomena* ed M Aschwanden (Berlin: Open Academic Press) pp 183–228
[36] Corral A, Boleda G and Ferrer-i-Cancho R 2013 in preparation
[37] Hankey A and Stanley H E 1972 *Phys. Rev.* B **6** 3515
[38] Ferrer i Cancho R and Solé R V 2001 *J. Quantum Linguist.* **8** 165
[39] Petersen A M, Tenenbaum J N, Havlin S, Stanley H E and Perc M 2012 *Sci. Rep.* **2** 943
[40] Baixeries J, Elvevåg B and Ferrer-i R 2013 Cancho *PLoS ONE* **8** e53227
[41] Schmid H 1994 *Proc. Int. Conf. on New Methods in Language Processing* vol 12 (Manchester: Citeseer) pp 44–9
[42] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes in FORTRAN* 2nd edn (Cambridge: Cambridge University Press)
[43] Devroye L 1986 *Non-Uniform Random Variate Generation* (New York: Springer)