entropy

MDPI

Communication

The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies

Álvaro Corral [1,2,3,4,*] and Isabel Serra [1,5]

[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
[3] Barcelona Graduate School of Mathematics, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain
[4] Complexity Science Hub Vienna, Josefstädter Straβe 39, 1080 Vienna, Austria
[5] Barcelona Supercomputing Center-Centro Nacional de Supercomputación, Jordi Girona 29, E-08034 Barcelona, Spain
* Correspondence: acorral@crm.es

Received: 31 December 2019; Accepted: 12 February 2020; Published: 17 February 2020

Abstract: An important body of quantitative linguistics is constituted by a series of statistical laws about language usage. Despite the importance of these linguistic laws, some of them are poorly formulated, and, more importantly, there is no unified framework that encompasses all them. This paper presents a new perspective to establish a connection between different statistical linguistic laws. Characterizing each word type by two random variables—length (in number of characters) and absolute frequency—we show that the corresponding bivariate joint probability distribution shows a rich and precise phenomenology, with the type-length and the type-frequency distributions as its two marginals, and the conditional distribution of frequency at fixed length providing a clear formulation for the brevity-frequency phenomenon. The type-length distribution turns out to be well fitted by a gamma distribution (much better than with the previously proposed lognormal), and the conditional frequency distributions at fixed length display power-law-decay behavior with a fixed exponent $\alpha \simeq 1.4$ and a characteristic-frequency crossover that scales as an inverse power $\delta \simeq 2.8$ of length, which implies the fulfillment of a scaling law analogous to those found in the thermodynamics of critical phenomena. As a by-product, we find a possible model-free explanation for the origin of Zipf's law, which should arise as a mixture of conditional frequency distributions governed by the crossover length-dependent frequency.

Keywords: quantitative linguistics; brevity law; abbreviation law; power laws; scaling; Zipf's law

1. Introduction

The usage of language, both in its written and oral expressions (texts and speech), follows very strong statistical regularities. One of the goals of quantitative linguistics is to unveil, analyze, explain, and exploit those linguistic statistical laws. Perhaps the clearest example of a statistical law in language usage is Zipf's law, which quantifies the frequency of occurrence of words in such written and oral forms [1–6], establishing that there is no unarbitrary way to distinguish between rare and common words (due to the absence of a characteristic scale in "rarity"). Surprisingly, Zipf's law is not only a linguistic law, but seems to be a rather common phenomenon in complex systems where discrete units self-organize into groups, or types (persons into cities, money into persons, etc. [7]).

Zipf's law can be considered as the "tip of the iceberg" of text statistics. Another well-known pattern of this sort is Herdan's law, also called Heaps' law [2,8,9], which states that the growth of vocabulary with text length is sublinear (however, the precise mathematical dependence has been debated [10]). Herdan's law has been related to Zipf's law, sometimes with too simple arguments, although rigorous connections have been established as well [8,10]. The authors of [11] provide another example of relations between linguistic laws, but, in general, no general framework encompassing all laws exists.

Two other laws—the law of word length and the so-called Zipf's law of abbreviation or brevity law— are of particular interest in this work. As far as we know, and in contrast to the Zipf's law of word frequency, these two laws do not have non-linguistic counterparts. The law of word length finds that the length of words (measured in number of letter tokens, for instance) is lognormally distributed [12,13], whereas the brevity law determines that more frequent words tend to be shorter, and rarer words tend to be longer. This is usually quantified between a negative correlation between word frequency and word length [14].

Very recently, Torre et al. [13] parameterized the dependence between mean frequency and length, obtaining (using a speech corpus) that the frequency averaged for fixed length decays exponentially with length. This is in contrast with a result suggested by Herdan (to the best of our knowledge not directly supported by empirical analysis), who proposed a power-law decay, with exponent between 2 and 3 [12]. This result probably arose from an analogy with the word-frequency distribution derived by Simon [15], with an exponential tail that was neglected.

The purpose of our paper is to put these three important linguistic laws (Zipf's law of word frequency, the word-length law, and the brevity law) into a broader context. By means of considering word frequency and word length as two random variables associated to word types, we will see how the bivariate distribution of those two variables is the appropriate framework to describe the brevity-frequency phenomenon. This leads us to several findings: (i) a gamma law for the word-length distribution, in contrast to the previously proposed lognormal shape; (ii) a well-defined functional form for the word-frequency distributions conditioned to fixed length, where a power-law decay with exponent α for the bulk frequencies becomes dominant; (iii) a scaling law for those distributions, apparent as a collapse of data under rescaling; (iv) an approximate power-law decay of the characteristic scale of frequency as a function of length, with exponent δ; and (v) a possible explanation for Zipf's law of word frequency as arising from the mixture of conditional distributions of frequency at different lengths, where Zipf's exponent is determined by the exponents α and δ.

2. Preliminary Considerations

Given a sample of natural language (a text, a fragment of speech, or a corpus, in general), any word type (i.e., each unique word) has an associated word length, which we measure in number of characters (as we deal with a written corpus), and an associated word absolute frequency, which is the number of occurrences of the word type on the corpus under consideration (i.e., the number of tokens of the type). We denote these two random variables as ℓ and n, respectively.

Zipf's law of word frequency is written as a power-law relation between $f(n)$ and n [6], i.e.,

$$f(n) \propto \frac{1}{n^\beta} \text{ for } n \geq c,$$

where $f(n)$ is the empirical probability mass function of the word frequency n, the symbol $\propto$ denotes proportionality, β is the power-law exponent, and c is a lower cut-off below which the law losses its validity (so, Zipf's law is a high-frequency phenomenon). The exponent β takes values typically close to 2. When very large corpora are analyzed (made from many different texts an authors) another (additional) power-law regime appears at smaller frequencies [16,17],

$$f(n) \propto \frac{1}{n^\alpha} \text{ for } a \leq n \leq b,$$

with α a new power law exponent smaller than β, and a and b lower and upper cut-offs, respectively (with $a < b < c$). This second power law is not identified with Zipf's law.

On the other hand, the law of word lengths [12] proposes a lognormal distribution for the empirical probability mass function of word lengths, that is,

$$f(\ell) \sim \mathrm{LN}(\mu, \sigma^2),$$

where LN denotes a lognormal distribution, whose associated normal distribution has mean μ and variance σ^2 (note that with the lognormal assumption it would seem that one is taking a continuous approximation for $f(\ell)$; nevertheless, discreteness of $f(\ell)$ is still possible just redefining the normalization constant). The present paper challenges the lognormal law for $f(\ell)$. Finally, the brevity law [14] can be summarized as

$$\mathrm{corr}(\ell, n) < 0,$$

where $\mathrm{corr}(\ell, n)$ is a correlation measure between ℓ and n, as, for instance, Pearson correlation, Spearman correlation, or Kendall correlation.

We claim that a more complete approach to the relationship between word length and word frequency can be obtained from the joint probability distribution $f(\ell, n)$ of both variables, together with the associated conditional distributions $f(n|\ell)$. To be more precise, $f(\ell, n)$ is the joint probability mass function of type length and frequency, and $f(n|\ell)$ is the probability mass function of type frequency conditioned to fixed length. Naturally, the word-frequency distribution $f(n)$ and the word-length distribution $f(\ell)$ are just the two marginal distributions of $f(\ell, n)$.

The relationships between these quantities are

$$f(\ell) = \sum_{n=1}^{\infty} f(\ell, n),$$

$$f(n) = \sum_{\ell=1}^{\infty} f(\ell, n),$$

$$f(\ell, n) = f(n|\ell) f(\ell).$$

Note that we will not use in this paper the equivalent relation $f(\ell, n) = f(\ell|n) f(n)$, for sampling reasons ($n$ takes many more different values than ℓ; so, for fixed values of n one may find there is not enough statistics to obtain $f(\ell|n)$). Obviously, all probability mass functions fulfil normalization,

$$\sum_{\ell=1}^{\infty} \sum_{n=1}^{\infty} f(\ell, n) = \sum_{n=1}^{\infty} f(n|\ell) = \sum_{\ell=1}^{\infty} f(\ell) = \sum_{n=1}^{\infty} f(n) = 1.$$

We stress that, in our framework, each type yields one instance of the bivariate random variable (ℓ, n), in contrast to another equivalent approach for which it is each token that gives one instance of the (perhaps-different) random variables, see [7]. The use of each approach has important consequences for the formulation of Zipf's law, as it is well known [7], and for the formulation of the word-length law (as it is not so well known [12]). Moreover, our bivariate framework is certainly different to the that in [18], where the frequency was understood as a four-variate distribution with the random variables taking 26 values from a to z, and also to the generalization in [19].

3. Corpus and Statistical Methods

We investigate the joint probability distribution of word-type length and frequency empirically, using all English books in the recently presented Standardized Project Gutenberg Corpus [20], which comprises more than 40,000 books in English, with a total number of tokens equal to 2,016,391,406 and a total number of types of 2,268,043. We disregard types with $n < 10$ (relative frequency below 5×10^{-9}) and also those not composed exclusively by the 26 usual letters from a to z (previously, capital letters were transformed to lower-case). This sub-corpus is further reduced by the elimination of types with length above 20 characters; to avoid typos and "spurious" words (among the eliminated types with $n \geq 10$ we only find three true English words: *incomprehensibilities*, *crystalloluminescence*, and *nitrosodimethylaniline*). This reduces the numbers of tokens and types, respectively, to 2,010,440,020 and 391,529. Thus, all we need for our study is the list of all types (a dictionary) including their absolute frequencies n and their lengths ℓ (measured in terms of number of characters).

Power-law distributions are fitted to the empirical data by using the version for discrete random variables of the method for continuous distributions outlined in [21] and developed in Refs. [22,23], which is based on maximum-likelihood estimation and the Kolmogorov–Smirnov goodness-of-fit test. Acceptable (i.e., non-rejectable) fits require p-values not below 0.20, which are computed with 1000 Monte Carlo simulations. Complete details in the discrete case are available in Refs. [6,24]. This method is similar in spirit to the one by Clauset et al. [25], but avoiding some of the important problems that the latter presents [26,27]. Histograms are drawn to provide visual intuition for the shape of the empirical probability mass functions and the adequacy of fits; in the case of $f(n|\ell)$ and $f(n)$, we use logarithmic binning [22,28]. Nevertheless, the computation of the fits does not make use of the graphical representation of the distributions.

On the other side, the theory of scaling analysis, following the authors of [21,29], allows us to compare the shape of the conditional distributions $f(n|\ell)$ for different values of ℓ. This theory has revealed a very powerful tool in quantitative linguistics, allowing in previous research to show that the shape of the word-frequency distribution does not change as a text increases its length [30,31].

4. Results

First, let us examine the raw data, looking at the scatter plot between frequency and length in Figure 1, where each point is a word type represented by an associated value of n and an associated value of ℓ (note that several or many types can overlap at the same point, if they share their values of ℓ and n, as these are discrete variables). >From the tendency of decreasing maximum n with increasing ℓ, clearly visible in the plot, one could arrive to an erroneous version of the brevity law. Naturally, brevity would be apparent if the scatter plot were homogenously populated (i.e., if $f(\ell, n)$ would be uniform in the domain occupied by the points). However, of course, this is not the case, as we will quantify later. On the contrary, if $f(\ell, m)$ were the product of two independent exponentials, with $m = \ln n$, the scatter plot would be rather similar to the real one (Figure 1), but the brevity law would not hold (because of the independence of ℓ and m, that is, of ℓ and n). We will see that exponentials distributions play an important role here, but not in this way.

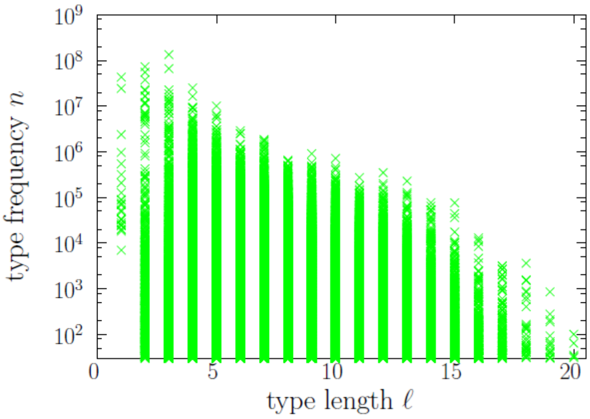

Figure 1. Illustration of the dataset by means of the scatter plot between word-type frequency and length. Frequencies below 30 are not shown.

A more acceptable approach to the brevity-frequency phenomenon is to calculate the correlation between ℓ and n. For the Pearson correlation, our dataset yields $\text{corr}(\ell, n) = -0.023$, which, despite looking very small, is significantly different from zero, with a p-value below 0.01 for 100 reshufflings of the frequency (all the values obtained after reshuffling the frequencies keeping the lengths fixed are between −0.004 and 0.006). If, instead, we calculate the Pearson correlation between ℓ and the logarithm m of the frequency we get $\text{corr}(\ell, m) = -0.083$, again with a p-value below 0.01. Nevertheless, as neither the underlying joint distributions $f(\ell, n)$ or $f(\ell, m)$ resemble a Gaussian at all, nor the correlation seems to be linear (see Figure 1), the meaning of the Pearson correlation is difficult to interpret. We will see below that the analysis of the conditional distributions $f(n|\ell)$ provides more useful information.

4.1. Marginal Distributions

Let us now study the word-length distribution, $f(\ell)$, shown in Figure 2. The distribution is clearly unimodal (with its maximum at $\ell = 7$), and although it has been previously modeled as a lognormal [12], we get a nearly perfect fit using a gamma distribution,

$$f(\ell) = \frac{\lambda}{\Gamma(\gamma)} (\lambda \ell)^{\gamma-1} e^{-\lambda \ell}, \tag{1}$$

with shape parameter $\gamma = 11.10 \pm 0.02$ and inverted scale parameter $\lambda = 1.439 \pm 0.003$ (where the uncertainty corresponds to one standard deviation, and $\Gamma(\gamma)$ denotes the gamma function). Notice then that, for large lengths, we would get an exponential decay (asymptotically, strictly speaking). However, there is an important difference between the lognormal distribution proposed in [13] and the gamma distribution found here, which is that the former case refers to the length of tokens, whereas in our case we deal with the length of types (of course, length of tokens and length of types is the same length, but the relative number of tokens and types is different, depending on length). This was already distinguished by Herdan [12], who used the terms occurrence distribution and dictionary distribution, and proposed that both of them were lognormal. In the caption of Figure 2 we provide the log-likelihoods of both the gamma and lognormal fits, concluding that the gamma distribution yields a better fit for the "dictionary distribution" of word lengths. The fit is specially good in the range $\ell > 2$.

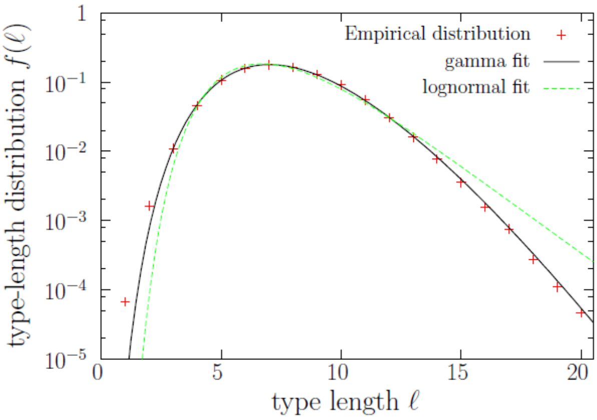

Figure 2. Probability mass function $f(\ell)$ of type length, together with gamma and lognormal fits. Note that the majority of types are those with lengths between 4 and 13, and that $f(\ell)$ is roughly constant between 5 and 10. The superiority of the gamma fit is visually apparent, and this is confirmed by log-likelihood equal to −872,175.2 in front of the value −876,535.1 for the lognormal (a discrete gamma distribution slightly improves the fit, but the simple continuous case here is enough for our purposes). The parameters resulting for the gamma fit are given in the text, and those for the lognormal are $\mu = 1.9970 \pm 0.0005$ and $\sigma = 0.3081 \pm 0.0003$.

Regarding the other marginal distribution, which is the word-frequency distribution $f(n)$ represented in Figure 3, we get that, as expected, Zipf's law is fulfilled with $\beta = 1.94 \pm 0.03$ for

$n \geq 1.9 \times 10^5$ (this is almost three orders of magnitude), see Table 1. Another power-law regime in the bulk, as in [16], is found to hold for one order of magnitude and a half (only), from $a \simeq 400$ to $b \simeq 14{,}000$, with exponent $\alpha = 1.41 \pm 0.005$, see Table 2. Note that although the truncated power law for the bulk part of the distribution is much shorter than the one for the tail (1.5 orders of magnitude in front of almost 3), the former contains many more data (50,000 in front of ~1000), see Tables 1 and 2 for the precise figures. Note also that the two power-law regimes for the frequency translate into two exponential regimes for m (the logarithm of n).

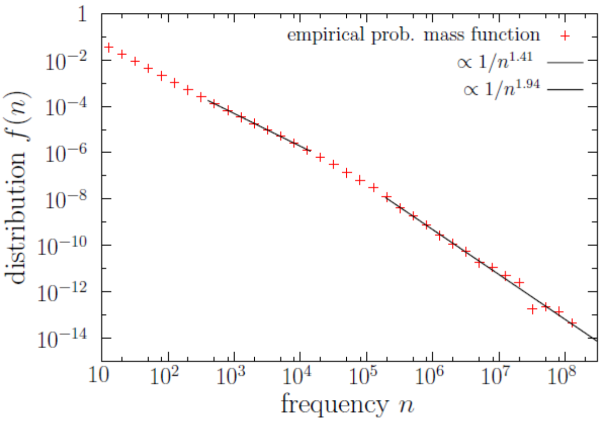

Figure 3. Probability mass function $f(n)$ of type frequency (this is a marginal distribution with respect $f(\ell, n)$). The results of the power-law fits are also shown. The fit of a truncated continuous power law, maximizing number of data, yields $\alpha = 1.41$; the fit of a untruncated discrete power law yields $\beta = 1.94$.

Table 1. Results of the fitting of an discrete untruncated power law to the conditional distributions $f(n|\ell)$, denoted by a fixed ℓ, and to the marginal distribution $f(n)$, denoted by the range $1 \leq \ell \leq 20$. N is total number of types, n_{max} is the frequency of the most frequent type, c is the lower cuf-off of the fit, ordermag is $\log_{10}(n_{max}/c)$, N_c is the number of types with $n \geq c$, β is the resulting fitting exponent, σ_β is its standard deviation, and p is the p-value of the fit. For the conditional distributions, the possible fits are restricted to the range $n > \langle n^2|\ell\rangle / \langle n|\ell\rangle$. The fit proceeds by sweeping 50 values of c per order of magnitude and using 1000 Monte Carlo simulations for the calculation of p. Of all the fits with $p \geq 0.20$ for a given ℓ, the one with smaller c is selected. Outside the range $5 \leq \ell \leq 14$, the number of types in the tail (below 10) is too low to yield a meaningful fit.

ℓ	N	n_{max} ($\times 10^5$)	c ($\times 10^5$)	Ordermag	N_c	$\beta \pm \sigma_\beta$	p
5	41,773	101	15.8	0.80	19	2.75 ± 0.46	0.97
6	62,277	29.0	3.80	0.88	60	2.79 ± 0.24	0.31
7	69,653	18.6	2.88	0.81	55	2.51 ± 0.21	0.32
8	63,574	6.55	1.10	0.78	133	2.82 ± 0.17	0.25
9	50,595	9.12	1.10	0.92	79	2.82 ± 0.21	0.25
10	35,679	7.16	0.83	0.93	69	2.90 ± 0.24	0.75
11	21,536	2.73	0.40	0.84	83	3.03 ± 0.23	0.58
12	11,973	3.49	0.46	0.88	34	2.78 ± 0.33	0.65
13	6240	2.28	0.44	0.72	13	2.57 ± 0.52	0.27
14	3035	0.77	0.24	0.51	12	2.67 ± 0.56	0.22
≤20	391,529	1341	1.91	2.85	927	1.94 ± 0.03	0.44

Table 2. Results of the fitting of a truncated power law to the conditional distributions $f(n|\ell)$, denoted by a fixed ℓ, and to the marginal distribution $f(n)$, denoted by the range $1 \leq \ell \leq 20$. N is total number of types; a and b are the lower and upper cut-offs of the fit, respectively; N_{ab} is the number of types with $a \leq n \leq b$; α is the resulting fitting exponent; σ_α is its standard deviation; and p is the p-value of the fit. The fit of a continuous power law is attempted in the range $n < 0.1\langle n^2|\ell\rangle/\langle n|\ell\rangle$, sweeping 20 values of a and b per order of magnitude and using 1000 Monte Carlo simulations for the calculation of p. Of all the fits with $p \geq 0.20$, for a given ℓ, the one with larger b/a is selected, except for $f(n)$, where the largest N_{ab} is used.

ℓ	N	a ($\times 10^2$)	b ($\times 10^3$)	**Ordermag**	N_{ab}	$\alpha \pm \sigma_\alpha$	p
1	26	126	2510	2.30	23	1.391 ± 0.155	0.24
2	636	20	2510	3.10	188	1.486 ± 0.045	0.24
3	4282	7.94	4470	3.75	1171	1.428 ± 0.016	0.30
4	17,790	0.40	398	4.00	10,618	1.402 ± 0.005	0.20
5	41,773	5.62	178	2.50	5747	1.426 ± 0.009	0.37
6	62,277	3.98	39.8	2.00	8681	1.421 ± 0.009	0.27
7	69,653	2.00	28.2	2.15	13,392	1.449 ± 0.007	0.25
8	63,574	2.51	11.2	1.65	9849	1.417 ± 0.010	0.41
9	50,595	2.00	10.0	1.70	8850	1.400 ± 0.010	0.25
10	35,679	1.12	8.91	1.90	8454	1.428 ± 0.010	0.21
11	21,536	0.56	1.41	1.40	6227	1.469 ± 0.015	0.22
12	11,973	0.63	5.01	1.90	3866	1.411 ± 0.013	0.51
13	6240	0.56	3.98	1.85	2144	1.396 ± 0.019	0.90
14	3035	0.25	2.24	1.95	1567	1.496 ± 0.022	0.27
15	1384	0.22	2.00	1.95	777	1.488 ± 0.031	0.59
16	612	0.28	0.45	1.20	256	1.569 ± 0.082	0.22
17	296	0.13	0.14	1.05	205	1.784 ± 0.110	0.24
18	107	0.11	0.16	1.15	79	2.008 ± 0.172	0.28
≤20	391,529	3.98	14.1	1.55	51,972	1.413 ± 0.005	0.21

4.2. Power Laws and Scaling Law for the Conditional Distributions

As mentioned, the conditional word-frequency distributions $f(n|\ell)$ are of substantial relevance. In Figure 4, we display some of those functions, and it turns out that n is broadly distributed for each value of ℓ (roughly in the same qualitative way it happens without conditioning to the value of ℓ). Remarkably, the results of a scaling analysis [21,29], depicted in Figure 5, show that all the different $f(n|\ell)$ (for $3 \leq \ell \leq 14$) share a common shape, with a scale determined by a scale parameter in frequency. Indeed, rescaling n as $n\langle n|\ell\rangle/\langle n^2|\ell\rangle$ and $f(n|\ell)$ as $f(n|\ell)\langle n^2|\ell\rangle^2/\langle n|\ell\rangle^3$, where the first and second empirical moments, $\langle n|\ell\rangle$ and $\langle n^2|\ell\rangle$, are also conditioned to the value of ℓ, we obtain an impressive data collapse, valid for ~7 orders of magnitude in n, which allows us to write the scaling law

$$f(n|\ell) \simeq \frac{\langle n|\ell\rangle^3}{\langle n^2|\ell\rangle^2} g\left(\frac{n\langle n|\ell\rangle}{\langle n^2|\ell\rangle}\right) \text{ for } 3 \leq \ell \leq 14,$$

where the key point is that the scaling function $g(...)$ is the same function for any value of ℓ. For $\ell > 14$ the statistics is low and the fulfilment of the scaling law becomes uncertain. Defining the scale parameter $\theta(\ell) = \langle n^2|\ell\rangle/\langle n|\ell\rangle$, we get alternative expressions for the same scaling law,

$$f(n|\ell) \simeq \frac{\langle n|\ell\rangle}{\theta^2(\ell)} g\left(\frac{n}{\theta(\ell)}\right) \propto \frac{1}{\theta^\alpha(\ell)} g\left(\frac{n}{\theta(\ell)}\right) \text{ for } 3 \leq \ell \leq 14,$$

where constants of proportionality have been reabsorbed into g, and the scale parameter has to be understood as proportional to a characteristic scale of the conditional distributions (i.e., θ is the characteristic scale, up to a constant factor; it is the relative change of θ what will be important for us). The reason for the fulfillment of these relations is the power-law dependence between the moments and the scale parameter when a scaling law holds, this power-law dependence is $\langle n|\ell\rangle \propto \theta^{2-\alpha}$ and $\langle n^2|\ell\rangle \propto \theta^{3-\alpha}$ for $1 < \alpha < 2$, see [21,29].

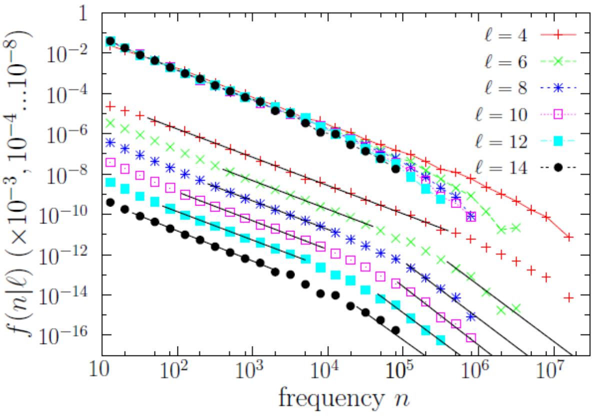

Figure 4. Probability mass functions $f(n|\ell)$ of frequency n conditioned to fixed value of length ℓ, for several values of ℓ. Distributions are shown twice: all together and individually, displaced in the vertical axis by diverse factors 10^{-3}, 10^{-4} up to 10^{-8}, for clarity sake of the power-law fits, represented by dark continuous lines.

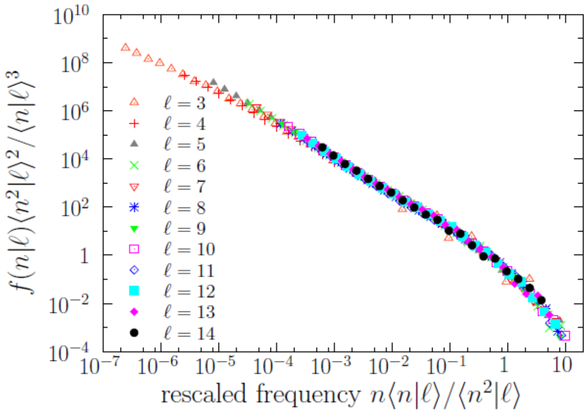

Figure 5. Word frequency probability mass functions $f(n|\ell)$ conditioned to fixed value of length rescaled by the ratio of powers of moments, as a function as rescaled frequency, for all values of length from 3 to 14. The data collapse guarantees the fulfilment of a scaling law.

The data collapse also unveils more clearly the functional form of the scaling function g, allowing to fit its power-law shape in two different ranges. The scaling function turns out to be compatible with a double power-law distribution, i.e., a (long) power law for $n/\theta < 0.1$ with exponent α at ~1.4 and another (short) power law for $n/\theta > 1$ with exponent β at ~2.75; in one formula,

$$g(y) \propto \begin{cases} 1/y^{1.4} & \text{for } y \ll 1, \\ 1/y^{2.75} & \text{for } y > 1, \end{cases} \tag{2}$$

for $y = n/\theta$. In other words, there is a (smooth) change of exponent (a change of log-log slope) at a value of $n \simeq C\theta(\ell)$, with the proportionality constant C taking some value in between 0.1 and 1 (as the transition from one regime to the other is smooth there is not a well defined value of C that separates both). Fitting power laws to those ranges we get the results shown in Tables 1 and 2. Note that $C\theta(\ell)$ can be understood as the characteristic scale of $f(n|\ell)$ mentioned before, and can be also called a frequency crossover.

Nevertheless, although the power-law regime for intermediate frequencies ($n < 0.1\theta$) is very clear, the validity of the other power law (the one for large frequencies) is questionable, in the sense that the power law provides an "acceptable" fit but other distributions could do the same good job, due to the limited range spanned by the tail (less than one order of magnitude). Our main reason to fit a power law to the large-frequency regime is the comparison with Zipf's law ($\beta \simeq 2$), and, as we see, the resulting value of β for $f(n|\ell)$ turns out to be rather large (the results of β for all $f(n|\ell)$ turn out to be statistically compatible with $\beta = 2.75$). In addition, we will show in the next subsection that the high-frequency behavior of the conditional distributions (power law or not) has nothing to do with Zipf's law.

4.3. Brevity Law and Possible Origin of Zipf's Law

Coming back to the scaling law, its fulfillment has an important consequence: it is the scale parameter $\theta(\ell)$ and not the conditional mean $\langle n|\ell\rangle$ what sets the scale of the conditional distributions $f(n|\ell)$. Figure 6 represents the brevity law in terms of the scale parameter as a function of ℓ (the conditional mean value is also shown, for comparison, overimposed to maps of $f(n,\ell)$ and $f(n|\ell)$). Note that the authors of [13] dealt with the conditional mean, finding an exponential decay $\langle n|\ell\rangle \propto 26^{-0.6\ell}$. Using our corpus (which is certainly different), we find that such an exponential decay for the mean is valid in a range of ℓ between 1 and 5, approximately. In contrast, the scale parameter θ shows an approximate power-law decay from about $\ell = 6$ to 15, with an exponent δ around 3 (or 2.8, to be more precise), i.e.,

$$\theta(\ell) \propto \frac{1}{\ell^{\delta}}$$

(note that Herdan assumed this exponent to be 2.4, with no clear empirical support [12]). Beyond $\ell = 15$, the decay of $\theta(\ell)$ is much faster. Nevertheless, these results are somewhat qualitative.

With these limitations, we could write a new version of the scaling law as

$$f(n|\ell) \simeq \ell^{\delta\alpha} g\left(\ell^{\delta} n\right) \tag{3}$$

where the proportionality constant between θ and ℓ^{δ} has been reabsorbed in the scaling function g. The corresponding data collapse is shown in Figure 7, for $5 \le \ell \le 14$. Despite the rough approximation provided by the power-law decay of $\theta(\ell)$, the data collapse in terms of scaling law (3) is nearly excellent for $\delta = 2.8$. This version of the scaling law provides a clean formulation of the brevity law: the characteristic scale of the distribution of n conditioned to the value of ℓ decays with increasing ℓ as $1/\ell^{\delta}$; i.e., the larger ℓ, the shorter the conditional distribution $f(n|\ell)$, quantified by the exponent δ.

However, in addition to a new understanding of the brevity law, the scaling law in terms of ℓ provides, as a by-product, an empirical explanation of the origin of Zipf's law. In the regime of ℓ in which the scaling law is approximately valid, i.e., for $\ell_1 \le \ell \le \ell_2$, we can obtain the distribution of frequency as a mixture of conditional distributions (by the law of total probability),

$$f(n) = \int_{\ell_1}^{\ell_2} f(n|\ell) f(\ell) d\ell$$

(where we take a continuous approximation, replacing sum over ℓ by integration; this is essentially a mathematical rephrasing). Substituting the scaling law and introducing the change of variables $x = \ell^{\delta} n$ we get

$$f(n) = \int_{\ell_1}^{\ell_2} \ell^{\delta\alpha} g\left(\ell^{\delta} n\right) f(\ell) d\ell \propto \int_{\ell_1^{\delta} n}^{\ell_2^{\delta} n} \left(\frac{x}{n}\right)^{\alpha} g(x) \frac{x^{-1+1/\delta}}{n^{1/\delta}} dx$$

$$= \frac{1}{n^{\alpha+1/\delta}} \int_{\ell_1^{\delta} n}^{\ell_2^{\delta} n} x^{\alpha-1+1/\delta} g(x) dx$$

where we also have taken advantage of the fact that, in the region of interest, $f(\ell)$ can be considered (in a rough approximation) as constant.

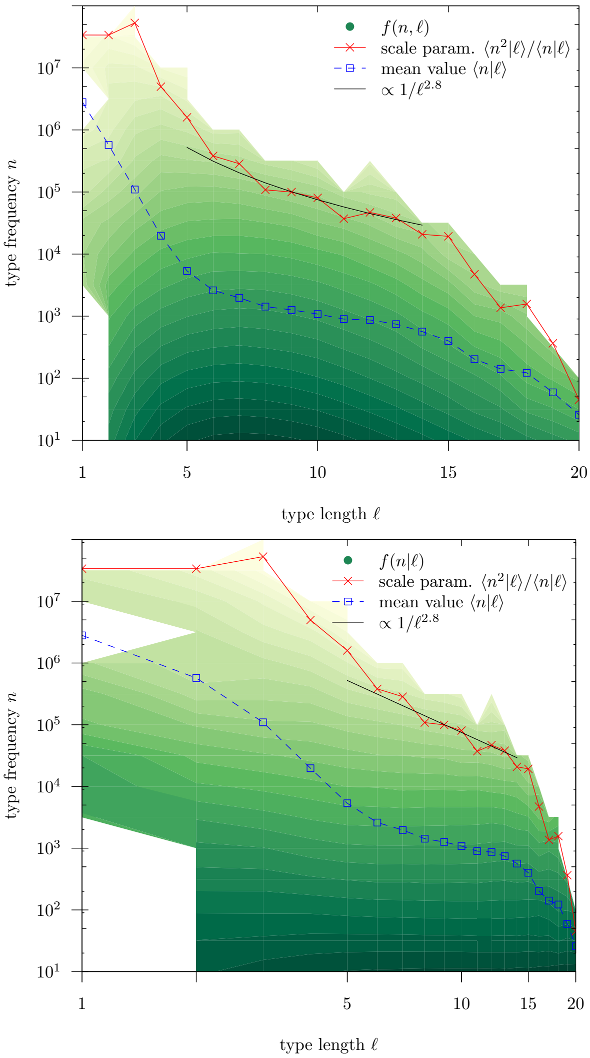

Figure 6. Estimated value of the scale parameter θ of the frequency conditional distributions ($\theta = \langle n^2|\ell\rangle / \langle n|\ell\rangle$) as a function of type length ℓ, together with conditional mean value $\langle n|\ell\rangle$. A decaying power law with exponent 2.8, shown as a guide to the eye, is close to the values of the scale parameter for $6 \leq \ell \leq 13$. The curves are overimposed to the values of the joint distribution $f(n,\ell)$ in the (**top panel**) and to the conditional distribution $f(n|\ell)$ in the (**bottom panel**). Notice that in the last case both axes are logarithmic. The shadower the green color, the higher the value of $f(n,\ell)$ and $f(n|\ell)$.

From here, we can see that in the case where the frequency is small ($n \ll \theta(\ell_2)$), the integration limits are also small, and then the last integral scales with n as $n^{1/\delta}$ (because we have that $g(x) \propto 1/x^{\alpha}$), which implies that we recover a power law with exponent α for $f(n)$, i.e., $f(n) \propto 1/n^{\alpha}$. However, for larger frequencies (n above $\theta(\ell_2)$ but below $\theta(\ell_1)$), the integral does not scale with n but can be considered instead as constant and then we get Zipf's law as

$$f(n) \propto n^{-\left(\alpha+\frac{1}{\delta}\right)}.$$

This means that Zipf's exponent can be obtained from the values of the intermediate-frequency power-law conditional exponent α and the brevity exponent δ as

$$\beta_z = \alpha + \frac{1}{\delta},$$

where we have introduced a subscript z in β to stress that this is the β exponent appearing in Zipf's law, corresponding to the marginal distribution $f(n)$, and to distinguish it from the one of the conditional distributions, that we may call β_c. Note then that β_c plays no role in the determination of β_z, and, in fact, the scaling function does not need to have a power-law tail to obtain Zipf's law. This sort of argument is similar to the one used in statistical seismology [32], but in that case the scaling law was elementary (i.e., $\theta = \langle n|\ell\rangle$).

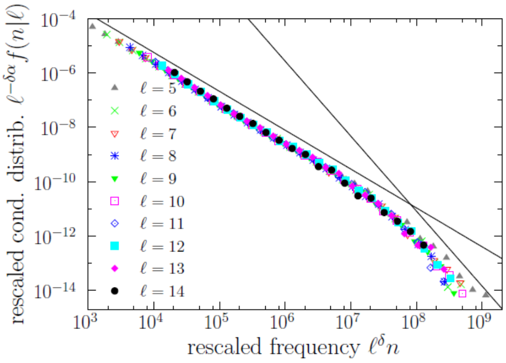

Figure 7. Same as Figure 5, from $\ell = 5$ to 14, changing the scale factors from combination of powers of moments ($\langle n|\ell\rangle$ and $\langle n^2|\ell\rangle$) to powers of length (in concrete, $\ell^{-\delta}$ and $\ell^{\delta\alpha}$). The collapse signals the fulfilment of a scaling law. Two decreasing power laws with exponents 1.43 and 2.76 are shown as straight lines, for comparison.

We can check the previous exponent relation using the empirical values of the exponent. We do not have a unique measure of α, but from Table 2, we see that its value for the different $f(n|\ell)$ is quite well defined. Taking the harmonic mean between the values $4 \leq \ell \leq 14$ we get $\bar{\alpha} = 1.43$, which together with $\delta = 2.8$ leads to $\beta_z \simeq 1.79$, not far from the ideal Zipf's value $\beta_z = 2$ and closer to the empirical value $\beta_z = 1.94$. The reason to calculate the harmonic mean of the exponents comes from the fact that it is the maximum-likelihood outcome when untruncated power-law datasets are put together [33]; when the power laws are truncated, the result is closer to the untruncated case when the range b/a is large.

5. Conclusions

Using a large corpus of English texts, we have seen how three important laws of quantitative linguistics, which are the type-length law, Zipf's law of word frequency, and the brevity law, can be put into a unified framework just considering the joint distribution of length and frequency.

Straightforwardly, the marginals of the joint distribution provide both the type-length distribution and the word-frequency distribution. We reformulate the type-length law, finding that the gamma distribution provides an excellent fit of type lengths for values larger than 2, in contrast to the previously proposed lognormal distribution [12] (although some previous research was dealing not with type length but with token length [13]). For the distribution of word frequency, we confirm the well-known Zipf's law, with an exponent $\beta_z = 1.94$; we also confirm the second intermediate power-law regime that emerges in large corpora [16], with an exponent $\alpha = 1.4$.

The advantages of the perspective provided by considering the length-frequency joint distribution become apparent when dealing with the brevity phenomenon. In concrete, this property arises very clearly when looking at the distributions of frequency conditioned to fixed length. These show a well-defined shape, characterized by a power-law decay for intermediate frequencies followed by a faster decay, which is well modeled by a second power law, for larger frequencies. The exponent α for the intermediate regime turns out to be the same as the one for the usual (marginal) distribution of frequency, $\alpha \simeq 1.4$. However, the exponent for higher frequencies β_c turns out to be larger than 2 and unrelated to Zipf's law.

At this point, scaling analysis reveals as a very powerful tool to explore and formulate the brevity law. We observe that the conditional frequency distributions show scaling for different values of length, i.e., when the distributions are rescaled by a scale parameter (proportional to the characteristic scale of each distribution), these distributions collapse into a unique curve, showing that they share a common shape (although at different scales). The characteristic scale of the distributions turns out to be well described by the scale parameter (given by the ratio of moments $\langle n^2|\ell\rangle / \langle n|\ell\rangle$), instead than by the mean value ($\langle n|\ell\rangle$). This is the usual case when the distributions involved have a power-law shape (with exponent $\alpha > 1$) close to the origin [29]. This also highlights the importance of looking at the whole distribution and not to mean values when one is dealing with complex phenomena.

Going further, we obtain that the characteristic scale of the conditional frequency distributions decays, approximately, as a power law of the type length, with exponent δ, which allows us to rewrite the scaling law in a form that is reminiscent to the one used in the theory of phase transitions and critical phenomena. Despite that the power-law behavior for the characteristic scale of frequency is rather rough, the derived scaling law shows an excellent agreement with the data. Note that taking together the marginal length distribution, Equation (1), and the scaling law for the conditional frequency distribution, Equation (3), we can write for the joint distribution

$$f(\ell, n) \propto \lambda^{\gamma} \ell^{\delta\alpha+\gamma-1} g(\ell^{\delta} n) e^{-\lambda\ell},$$

with the scaling function $g(x)$ given by Equation (2), up to proportionality factors.

Finally, the fulfilment of a scaling law of this form allows us to obtain a phenomenological (model free) explanation of Zipf's law as a mixture of the conditional distributions of frequencies. In contrast to some accepted explanations of Zipf's law, which put the origin of the law outside the linguistic realm (such as Simon's model [15], where only the reinforced growth of the different types counts; other explanations are in [19,34]), our approach indicates that the origin of Zipf's law can be fully linguistic, as it depends crucially on the length of the words (and the length is a purely linguistic attribute). Thus, at fixed length, each (conditional) frequency distribution shows a scale-free (power-law) behavior, up to a characteristic frequency where the power law (with exponent α) breaks down. This breaking-down frequency depends on length through the exponent δ. The mixture of different power laws, with exponent α and cut at a scale governed by the exponent δ, yields a Zipf's exponent $\beta_z = \alpha + \delta^{-1}$. Strictly speaking, our explanation of Zipf's law does not fully explain Zipf's law, but transfers the explanation to the existence of a power law with a smaller exponent ($\alpha \simeq 1.4$) as well as to the crossover frequency that depends on length as $\ell^{-\delta}$. Clearly, more research is necessary to explain the shape of the conditional distributions. It is noteworthy that a similar phenomenology for Zipf's law (in general) was proposed in [34], using the concept of "underlying unobserved variables", which in the case of word frequencies were associated (without quantification) to part of speech (grammatical categories). From our point of view, the "underlying unobserved variables" in the case of word frequencies would be instead word (type) lengths.

Although our results are obtained using a unique English corpus, we believe they are fully representative of this language, at least when large corpora are used. Naturally, further investigations are needed to confirm the generality of our results. Of course, a necessary extension of our work is the use of corpora on other languages, to establish the universality of our results, as done, e.g., in [14]. The length of words is simply measured in number of characters, but nothing precludes the use of number of phonemes or mean time duration of types (in speech, as in [13]). At the end, the goal of this kind of research is to pursue a unified theory of linguistic laws, as proposed in [35]. The line of research shown in this paper seems to be a promising one.

Author Contributions: Methodology, Á.C. and I.S.; writing, Á.C.; visualization, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish MINECO, through grants FIS2012-31324, FIS2015-71851-P, PGC-FIS2018-099629-B-I00, and MDM-2014-0445 (María de Maeztu Program). I.S. was funded through the Collaborative Mathematics Project from La Caixa Foundation.

Acknowledgments: We are indebted to Francesc Font-Clos for providing the valuable corpus released in [20]. Our interest in the brevity law arose from our interaction with Ramon Ferrer-i-Cancho, in particular from the reading of [35,36].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Boston, MA, USA, 1949.
2. Baayen, R.H. *Word Frequency Distributions*; Kluwer: Dordrecht, The Netherlands, 2001.
3. Baroni, M. Distributions in text. In *Corpus linguistics: An International Handbook*; Lüdeling, A., Kytö, M., Eds.; Mouton de Gruyter: Berlin, Germany, 2009; Volume 2, pp. 803–821.
4. Zanette, D. Statistical patterns in written language. *arXiv* **2014**, arXiv:1412.3336v1.
5. Piantadosi, S.T. Zipf's law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef] [PubMed]
6. Moreno-Sánchez, I.; Font-Clos, F.; Corral, A. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* **2016**, *11*, e0147073. [CrossRef] [PubMed]
7. Corral, A.; Serra, I.; Ferrer-i-Cancho, R. The distinct flavors of Zipf's law in the rank-size and in the size-distribution representations, and its maximum-likelihood fitting. *arXiv* **2019**, arXiv:1908.01398.
8. Mandelbrot, B. On the theory of word frequencies and on related Markovian models of discourse. In *Structure of Language and its Mathematical Aspects*; Jakobson, R., Ed.; American Mathematical Society: Providence, RI, USA, 1961; pp. 190–219.
9. Heaps, H.S. *Information retrieval: Computational and Theoretical Aspects*; Academic Press: Cambridge, MA, USA, 1978.
10. Font-Clos, F.; Corral, A. Log-log convexity of type-token growth in Zipf's systems. *Phys. Rev. Lett.* **2015**, *114*, 238701. [CrossRef]
11. Altmann, E.G.; Gerlach, M. Statistical laws in linguistics. In *Creativity and Universality in Language. Lecture Notes in Morphogenesis*; Esposti, M.D., Altmann, E.G., Pachet, F., Eds.; Springer: Berlin/Heidelberger, Germany, 2016.
12. Herdan, G. The Relation Between the Dictionary Distribution and the Occurrence Distribution of Word Length and its Importance for the Study of Quantitative Linguistics. *Biometrika* **1958**, *45*, 222–228. [CrossRef]
13. Torre, I.G.; Luque, B.; Lacasa, L.; Kello, C.T.; Hernández-Fernández, A. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. Open Sci.* **2019**, *6*, 191023. [CrossRef]
14. Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*; Bentz, C., Jäger, G., Yanovich, I., Eds.; University of Tübingen: Tübingen, Germany, 2016.
15. Simon, H.A. On a class of skew distribution functions. *Biometrika* **1955**, *42*, 425–440. [CrossRef]
16. Ferrer i Cancho, R.; Solé, R.V. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.* **2001**, *8*, 165–173. [CrossRef]
17. Williams, J.R.; Bagrow, J.P.; Danforth, C.M.; Dodds, P.S. Text mixing shapes the anatomy of rank-frequency distributions. *Phys. Rev. E* **2015**, *91*, 052811. [CrossRef]
18. Stephens, G.J.; Bialek, W. Statistical mechanics of letters in words. *Phys. Rev. E* **2010**, *81*, 066119. [CrossRef] [PubMed]
19. Corral, A.; García del Muro, M. From Boltzmann to Zipf through Shannon and Jaynes. *Entropy* **2020**, *22*, 179. [CrossRef]
20. Gerlach, M.; Font-Clos, F. A standardized Project Gutenberg Corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* **2020**, *22*, 126. [CrossRef]
21. Peters, O.; Deluca, A.; Corral, A.; Neelin, J.D.; Holloway, C.E. Universality of rain event size distributions. *J. Stat. Mech.* **2010**, *11*, P11030. [CrossRef]
22. Deluca, A.; Corral, A. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.* **2013**, *61*, 1351–1394. [CrossRef]

23. Corral, A.; González, A. Power law distributions in geoscience revisited. *Earth Space Sci.* **2019**, *6*, 673–697. [CrossRef]
24. Corral, A.; Boleda, G.; Ferrer-i-Cancho, R. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef]
25. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703. [CrossRef]
26. Corral, A.; Font, F.; Camacho, J. Non-characteristic half-lives in radioactive decay. *Phys. Rev. E* **2011**, *83*, 066103. [CrossRef]
27. Voitalov, I.; van der Hoorn, P.; van der Hofstad, R.; Krioukov, D. Scale-free networks well done. *Phys. Rev. Res.* **2019**, *1*, 033034. [CrossRef]
28. Deluca, A.; Corral, A. Scale invariant events and dry spells for medium-resolution local rain data. *Nonlinear Proc. Geophys.* **2014**, *21*, 555–567. [CrossRef]
29. Corral, A. Scaling in the timing of extreme events. *Chaos Solitons Fract.* **2015**, *74*, 99–112. [CrossRef]
30. Font-Clos, F.; Boleda, G.; Corral, A. A scaling law beyond Zipf's law and its relation to Heaps' law. *New J. Phys.* **2013**, *15*, 093033. [CrossRef]
31. Corral, A.; Font-Clos, F. Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Phys. Rev. E* **2017**, *96*, 022318. [CrossRef]
32. Corral, A. Statistical features of earthquake temporal occurrence. In *Modelling Critical and Catastrophic Phenomena in Geoscience*; Bhattacharyya, P., Chakrabarti, B.K., Eds.; Springer: Berlin/Heidelberger, Germany, 2007.
33. Navas-Portella, V.; Serra, I.; Corral, A.; Vives, E. Increasing power-law range in avalanche amplitude and energy distributions. *Phys. Rev. E* **2018**, *97*, 022134. [CrossRef]
34. Aitchison, L.; Corradi, N.; Latham, P.E. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Comput. Biol.* **2016**, *12*, e1005110. [CrossRef]
35. Ferrer-i-Cancho, R. Compression and the origins of Zipf's law for word frequencies. *Complexity* **2016**, *21*, 409–411. [CrossRef]
36. Ferrer-i-Cancho, R.; Bentz, C.; Seguin, C. Compression and the origins of Zipf's law of abbreviation. *arXiv* **2015**, arXiv:1504.04884.