OPEN ACCESS Freely available online

PLoS one

Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution

Ramon Ferrer-i-Cancho[1]*, Brita Elvevåg[2]

1 Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain, **2** Clinical Brain Disorders Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract

Background: Zipf's law states that the relationship between the frequency of a word in a text and its rank (the most frequent word has rank 1, the 2nd most frequent word has rank 2,…) is approximately linear when plotted on a double logarithmic scale. It has been argued that the law is not a relevant or useful property of language because simple random texts - constructed by concatenating random characters including blanks behaving as word delimiters - exhibit a Zipf's law-like word rank distribution.

Methodology/Principal Findings: In this article, we examine the flaws of such putative good fits of random texts. We demonstrate - by means of three different statistical tests - that ranks derived from random texts and ranks derived from real texts are statistically inconsistent with the parameters employed to argue for such a good fit, even when the parameters are inferred from the target real text. Our findings are valid for both the simplest random texts composed of equally likely characters as well as more elaborate and realistic versions where character probabilities are borrowed from a real text.

Conclusions/Significance: The good fit of random texts to real Zipf's law-like rank distributions has not yet been established. Therefore, we suggest that Zipf's law might in fact be a fundamental law in natural languages.

Citation: Ferrer-i-Cancho R, Elvevåg B (2010) Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. PLoS ONE 5(3): e9411. doi:10.1371/journal.pone.0009411

Editor: Enrico Scalas, University of East Piedmont, Italy

Received September 23, 2009; **Accepted** January 11, 2010; **Published** March 9, 2010

Funding: This work was partially supported by the project Secuencias Simbólicas: Análisis, Aprendizaje, Minería y Evolución, Barcelona (SESAAME-BAR) (TIN2008-06582-C03-01) of the Spanish Ministry of Science and Innovation, http://web.micinn.es/ (RFC); and the National Institute of Mental Health Intramural Research Program (NIMH-IRP), http://intramural.nimh.nih.gov/ (BE). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rferrericancho@lsi.upc.edu

Introduction

Imagine that one takes a text, counts the frequency of every word and assigns a rank to each word in a decreasing order of frequency. This would result in the most frequent word having a rank of 1, the second most frequent word having a rank of 2 and so on. The histogram of such word ranks is said to conform to Zipf's law for word frequencies [1]. In its simplest form, the law states that $f(r)$, the frequency of a word or rank r obeys

$$f(r) \sim r^{-\alpha}, \qquad (1)$$

where α is a constant, the so-called exponent of the law (typically $\alpha \approx 1$ [1]). In other words, Eq. 1 indicates that frequency decays linearly as the rank increases on double logarithmic scale. Although the law was originally thought to reveal principles of language functioning [1], many have argued against its relevance [2–7]. Their major claim is that the statistics of simple random sequences of characters - including a special one that behaves as a word delimiter - reproduces Zipf's law for word frequencies [2,4,5]. Henceforth, we refer to this special character as a space or a blank. For instance, the random text

wbqcrw h q rorjleabeyxkrlpqkpchnesguliwkb mrltn q a rss vfs w a h rlzpxxtxbkqetfwfpqudgwaorqwgqmo wyngwtbseuodboxaw x rldua eucx mmard xgqzv uu pueuerc pkizuauyrwi bllhjddv bp anud xbxvjyymioymvzebc tdtsecdijntssyepqdubcvxjd evavybwvejp w z uvspufvdvuzyf t nllifznwatic

has been generated using English letters ranging from 'a' to 'z' (the separation between words in our example is arbitrary and due to automatic formatting).

The idea that random sequences of characters reproduce Zipf's law stems from the seminal work of Mandelbrot [8] and was reformulated in various works [2,4,5,9]. We refer to a random sequence of characters of the type listed above as a random text so as to be consistent with [2] although more appropriate names have been discussed [10]. The simplest version of a random text is based upon the assumption that all characters are equally likely [2,7]. We define N as the number of regular characters of the random text and p_b as the probability of a blank. The above example was generated with $N = 26$ and $p_b = 0.18$, which was deemed suitable for English [4,5]. It is noteworthy that when constructing the example above, we assumed that all characters are independent, that all letters from 'a' to 'z' are equally likely and two or more blanks in a row are not permitted. If two blanks in a row are not allowed then words with no characters (i.e. empty words) cannot be generated.

There have been many arguments against the meaningfulness or relevance of Zipf's law [2,4,6,7]. However, there are also reasons that such arguments might be flawed:

- *Problem 1*
  The studies that question the relevance to natural language of Zipf's law argue for the matching between Eq. 1 and random texts. However, Eq. 1 is only an approximation for the real rank histogram. The best candidate for the actual rank distribution remains an open question [11–14] for two reasons: first, the goodness of the fit provided by Eq. 1 in a statistically rigorous sense is questionable and, second, the best function may not be unique [15]. If it turned out that when using statistically rigorous methods that real texts do not usually fit Eq. 1, then the arguments against the relevance of Zipf's law would be seriously challenged.
- *Problem 2*
  As far as we know, in none of the popular articles that argue against the meaningfulness of Zipf's law [2–7] is there an accurate enough derivation of Zipf's law (Eq. 1) from random texts. This is of crucial importance because real texts and random texts may seem to have consistent rank distributions if not regarded with enough precision simply because two distinct tiny objects may look similar if our lens is not powerful enough. Notice that in [2–7] an exact derivation of Zipf's law from the assumptions of a random text is absent. Instead, only equations that are valid for the ensemble of words of a certain length are provided. For instance, Li [2] defines $p_w(L)$ as the probability of any particular word of length L and proves that

$$p_w(L) < \frac{C}{(r(L)+B)^{\alpha'}} \leq p_w(L-1), \tag{2}$$

  where B, C and α' are constants and $r(L)$ is the rank of any word of length L (a similar derivation can be found in [7]). Miller & Chomsky [4] showed that the probability of any word of length L obeys

$$p_w(L) \sim (r'(L)+B')^{-\alpha''}, \tag{3}$$

  where B' and α'' are constants and $r'(L)$ is now the mean rank of all the possible words of length L. In contrast, notice that Zipf's law (Eq. 1) is a law of individual ranks, not a law of a rank chosen to represent all words of the same length (e.g., the average rank or words of the same length). Recently, it has been proven that $p(r)$, the probability of observing a word of rank r in a random text, obeys [16], for sufficiently large r,

$$c_1 r^{-\alpha} \leq p(r) \leq c_2 r^{-\alpha}, \tag{4}$$

  where c_1 and c_2 are two positive constants. Although the derivation of Eq. 4 in [16] for a general class of random texts is a milestone in the history of random texts, notice that Eq. 4 is weaker than the definition of Zipf's law in Eq. 1.
- *Problem 3*
  Eqs. 2, 3 and 4 are derived in the context of a very long text. It is not known *a priori* if the parameters of the underlying exact distribution of ranks depend upon the text length or if the distribution that is obtained in the context of a very long text is the same as that of a random text of the size of the order of real texts.
- *Problem 4*
  As far as we know, in none of the popular articles that question the meaningfulness to natural language of Zipf's law [2–7] is there any comparison between the rank histograms of actual texts and those of random texts. Rather it is simply taken for granted that an approximate agreement with Eq. 1 is sufficient. To the best of our knowledge, in none of these cases is either a visual comparison between the rank histogram of a real text and that of a random text provided (e.g., by plotting both histograms together), nor are more convenient rigorous tests of the goodness of fit of random texts for real texts performed. In some exceptional cases, a visual comparison between a real text and an equation similar to Eq. 3 has been made [17] but the comparison implies the misuse of an equation that was originally derived for the mean rank of words of the same length to the individual ranks of actual Zipf's law-like rank distributions. Although Mandelbrot did not show simultaneously real and artificial rank distributions, arguably he inappropriately used equations that had been derived for individual ranks (e.g., Fig. 1 of [18] and Fig. 2 of [8]).

To address Problem 1, we evaluate the goodness of fits of random texts to real texts *directly* by means of samples of ranks produced by the real process and not *indirectly* through Eq. 1. To address Problem 2 we study the consistency between rank samples from a random text and rank samples from a real text using three rigorous statistical tests. We skip the mathematical challenge of obtaining the missing exact rank distribution for individual ranks. To address Problem 3, we compare real texts with random texts of the same length. In this way, we can establish that putative differences cannot be attributed to simply differences in the text length. To address Problem 4, we compare visually the rank histogram of random texts with those of real texts so as to provide an estimate of the enormous differences between both and then we perform rigorous statistical tests to show that the real word rank histograms are inconsistent with those of random texts.

We exclude from our analysis a variant of the random text that generates empty words. Empty words are obtained when producing two blanks in a row, which is allowed in [4–7,16] but not permitted in Li's version [2] (see Text S1). In other cases, it is not clear if empty words are allowed, e.g., [19]. Excluding empty words in our study is justified by the fact that the goal of this article is to evaluate the fit of random texts for real Zipf's law-like word rank distributions. As far as we know, in none of Zipf's pioneering works [1,20] and in the many studies that followed, have empty words been included or even considered in real rank histograms. Indeed, their existence in real texts is very questionable.

Many authors have discussed the explanatory adequacy of random texts for real Zipf's law-like word rank distributions indirectly from inconsistencies between random texts and real texts beyond the distribution of ranks [19,21–24]. One of the most typical and recurrent examples is the claim that real word lengths are not geometrically distributed as expected from a random text experiment [21–24]. However, the question that we seek to address here is: do random texts really fit the real Zipf's law-like distribution accurately as suggested by many [2,4–7,25]?

To our knowledge, only a few studies have addressed this question [19,26,27] but in a qualitative manner and only for certain versions of the random text model. In this article, we go a step forward by bringing rigorous statistical tests into the debate and considering all the variants of the random text model that have been considered in the literature. In particular, we compare visually some rank histograms from English texts with those of different versions of the random text model and test rigorously the goodness of fit of random texts on actual histograms in a set of ten texts. We demonstrate that - contrary to what has previously been suggested - random texts fail to fit actual texts even visually.

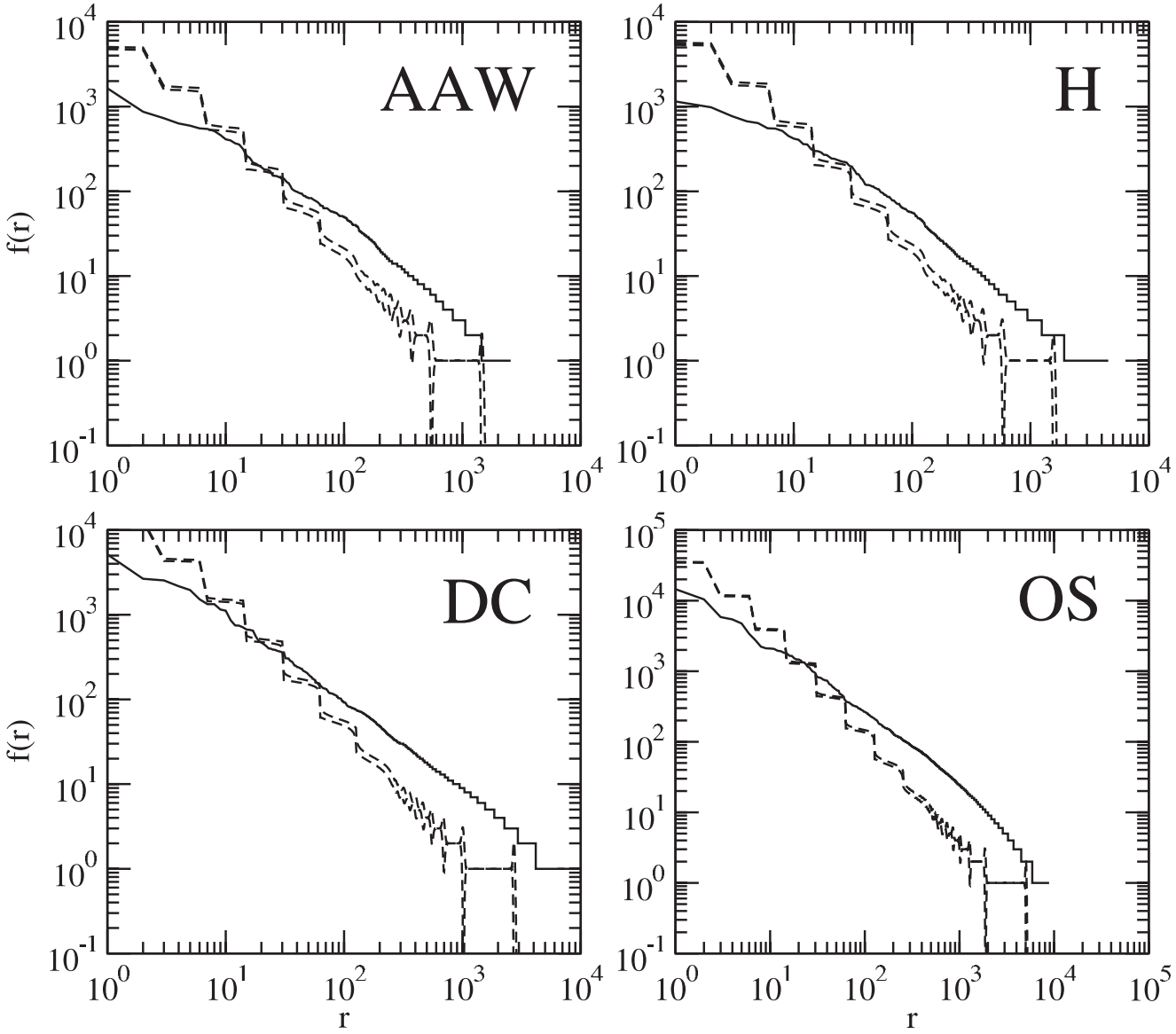

Figure 1. The rank histograms of English texts versus that of random texts (RT_1). A comparison of the real rank histogram (thin black line) and two control curves with the 3σ upper and lower bounds of the expected histogram of a random text of the same length in words (dashed lines) involving four English texts. $f(r)$ is the frequency of the word of rank r. For the random text we use the model RT_1 with alphabet size $N=2$. The expected histogram of the random text is estimated averaging over the rank histograms of 10^4 random texts. For ease of presentation, the expected histogram is cut off at expected frequencies below 0.1. AAW: *Alice's adventures in wonderland*. H: *Hamlet*. DC: *David Crockett*. OS: *The origin of species*.
doi:10.1371/journal.pone.0009411.g001

Finally, we shed light on the failure of random texts to fit actual texts from the perspective of cognitive science and discuss the implications of our negative results for the meaningfulness to natural language of Zipf's law.

Results

In this article we employ a set of ten English texts (eight novels and two essays) to evaluate the goodness of fit of random texts in Table 1. A summary of their statistical properties is shown in Table 2.

The Versions of the Random Text

We consider three different versions of the random text (RT) model without empty words that have been considered in the literature. All the versions generate a random sequence of independent characters. These three version are (the subindex indicates the number of parameters of the version of the random text):

- RT_1
  All characters, including the blank are equally likely. This model is specified with a single parameter: N, the number of characters other than space. $N \in \{2,4,6,26\}$ was used in [2]. $N=5$ was used by [7] allowing empty words. An example of RT_1 with $N=2$ is

  uu kuuuuk k kkk uu u kkuuukuuk uk kukukuuu u ukku kukkk uku uku ku u kuk kukk uuuk k kk kku uuu u kuukkuk u kku kuukuu u uukk ku uuk kukk u ukkkkuuu k ukku kuku kuk k k uku k uuku uu kuukukuukk kukku k uk u

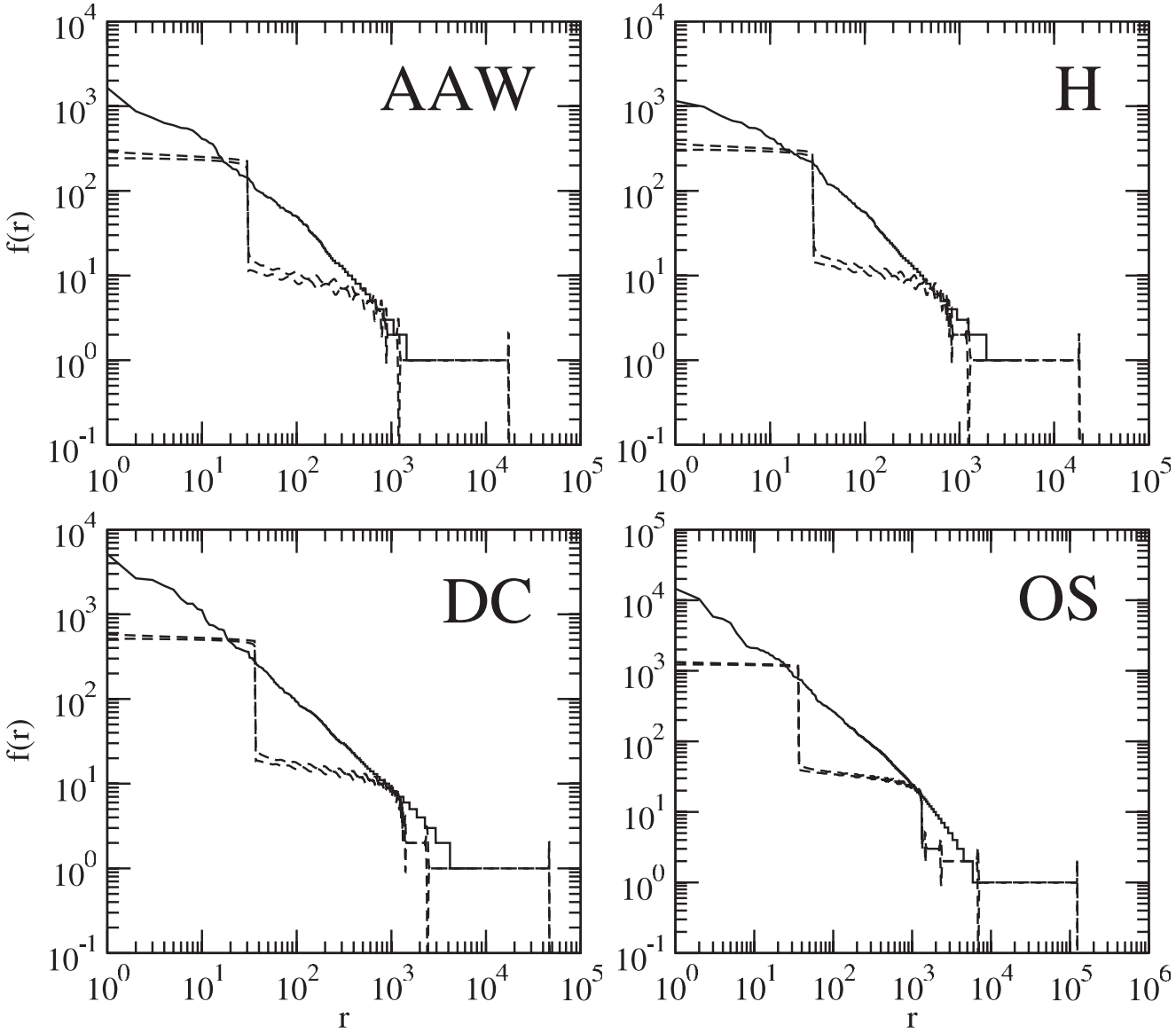

Figure 2. The rank histograms of English texts versus that of random texts (RT_2). The same as Fig. 1 for the model RT_2 with alphabet size N and probability of blank p_b obtained from the real text.
doi:10.1371/journal.pone.0009411.g002

- RT_2
  All characters except the blank are equally likely. This model is specified with two parameters, N as in RT_1 plus p_b, the probability of blank for the 2nd and following characters of a word (notice that in our case, the probability that the current word under construction has no character when the blank is produced is zero). Allowing empty words, $N=26$ and probability of blank $p_b=0.18$ was argued to be suitable for English [4,5] without explaining how p_b was estimated. Here we obtain N and p_b from real normalized texts (see Materials and Methods for details about our text normalization). N is obtained from the number of different characters of the text (except the blank). p_b is computed from the formula

$$p_b = \frac{N_b}{N_c - N_b}, \qquad (5)$$

  where N_b is the number of blanks and N_c is the total number or characters (including blanks). In our text normalization, N_b is equivalent to the number of different words (i.e. the maximum rank). p_b is the proportion of blanks after excluding the first character of each word, which cannot be a blank in our versions of the random text model. An example of RT_2 with N and p_b borrowed from *Alice's adventures in wonderland* is

  i 0xbple f h gxadchrdcty hz trsykj o b axurvg qfu k kg3kx vwzsj3 xw0t3f nq ryb uhibb nqhtqb zfgnfk v gdq p30ajh 30 c p k3cgozfe3vt hdmzc k0q bw fs c kgu lm0tx bh av eu v cmbosjbis a3aks mucjtefrtvko t uyprnz eyti h3do hm0mx w0kbecyd ti v qoyowzcfiykv3wb

- RT_{N+1}
  All characters can take any probability. This model is specified with $N+1$ parameters (i.e. the $N+1$ probabilities of each of

Table 1. Summary of English texts employed.

Title	Abbreviation	Author
Alice's adventures in wonderland	AAW	Lewis Carroll (1832–1898)
The adventures of Tom Sawyer	ATS	Mark Twain (1835–1910)
A Christmas carol	CC	Charles Dickens (1812–1870)
David Crockett	DC	John S. C. Abbott (1805–1877)
An enquiry concerning human understanding	ECHU	David Hume (1711–1776)
Hamlet	H	William Shakespeare (1564–1616)
The hound of the Baskervilles	HB	Sir Arthur Conan Doyle (1859–1930)
Moby-Dick: or, the whale	MB	Herman Melville (1819–1891)
The origin of species by means of natural selection	OS	Charles Darwin (1809–1882)
Ulysses	U	James Joyce (1882–1941)

The data set of English texts employed in our study.
doi:10.1371/journal.pone.0009411.t001

the characters). Three parameter settings have been considered in the literature:

- L_1
  $N=2$ with the probability of the two characters other than space being 0.47 and 0.2 and the probability of space is 0.33 [2].
- L_2
  $N=4$ with the probability of the four characters other than space being 0.5, 0.13, 0.1 and 0.07 and the probability of space being 0.2 [2].

Table 2. Statistics of the English texts.

Abbreviation	T (in words)	N (in chars.)	p_b	$\max(r)$	$\mu(r)$	$\sigma(r)$
AAW	27342	28	0.254	2574	254.05	466.60
CC	29253	30	0.240	4263	463.31	887.22
H	32839	28	0.253	4582	474.39	932.44
ECHU	57958	36	0.212	4912	433.91	861.35
HB	59967	39	0.244	5568	472.87	990.44
ATS	73523	31	0.248	7169	612.45	1298.53
DC	78819	36	0.228	7385	668.60	1346.19
OS	209176	36	0.207	8955	589.94	1274.53
MB	218522	36	0.229	17190	1291.67	2909.44
U	269589	36	0.228	29213	2425.63	5444.95

Statistical properties of the English texts. See Table 1 for the meaning of each abbreviation. Texts are sorted by increasing length. T is the text length in words. N is the number of different characters excluding the blank. p_b is the estimated probability of blank. $\max(r)$ is the maximum rank or the observed vocabulary size. $\mu(r)$ and $\sigma(r)$ are, respectively, the mean and the standard deviation of the rank.
doi:10.1371/journal.pone.0009411.t002

- *Real*
  Real character probabilities extracted from the target writing as in [19].

An example of RT_{N+1} with real character probabilities borrowed from *Alice's adventures in wonderland* is

tel g shs oo fagl t ersu fa r esnrlaod k ni ihe a o e sh foie r do aorhdaev aiot t oseldtiyie wq t thsynt w e sptsnsm heooeat utdgeco a iyeb sniemt ehdoy t thruw twaame eatendeisidle mc nlhtt ih a utfd anulbgleta nlh ohe gt eehitofnet

Visual Fitting

Here we aim to compare rank histograms from real texts and expected rank histograms from random texts. If random texts really reproduce the rank histogram of real texts, then the histogram of real texts and those of the random texts should completely overlap. We will see that this is not the case.

Here our emphasis is on providing a fair visual comparison. We use the term fair in two senses. First, we consider real and artificial texts of the same length in words. Notice that the equations that have been derived so far for the rank distribution of RT_1 and RT_2 texts are derived in the limit of a very large text in which all words of the same length must have the same frequency of occurrence because they are *a priori* equally likely [2,4,5,7]. If the text is not long enough, the frequency of words of the same length may differ noticeably. Here we aim to equate the text size of both the model and the real text. Second, we do not misuse a theoretical equation that is not valid for individual ranks as in [17]. The theoretical rank distribution or even the theoretical expected rank histogram of random texts are not available, even in their most simples versions. Therefore, we work on the expected rank histogram of random texts, which can be easily estimated by simulating the process and averaging the rank histogram over a sufficiently large number of artificial texts. Third, we do not use binning as in [26] which could shadow the differences between actual texts and random texts.

In the interest of being concise, for visual fits, we chose four works representing different genres and covering the whole range of text lengths in the sample. Fig. 1 shows the rank histogram of the four selected English texts versus the expected rank histogram of a RT_1. From visual inspection, it is obvious that the agreement between the random text and the real text is poor. The histograms of random texts are clearly above the corresponding real histograms for small ranks and clearly below for larger ranks. Additionally, the curves of real histograms are smoother as compared to the pronounced staircase decrease of random texts, especially for small ranks. Fig. 2 shows that RT_2 with N and p_b taken from the real text does not improve the quality of the visual fit. The staircase decrease of the histogram of random texts becomes more radical and the plateaus are huge as compared to those of Fig. 1. One may infer from Figs. 1 and 2 that RT_1 gives a better fit than RT_2 in general, but the difference in fitting is mainly due to the small value of N employed in Fig. 1, which produces smaller plateaus with regard to Fig. 2.

It is well known that if characters other than the blank have unequal probabilities then the rank histogram smoothes [2,25,28]. The point is: would this apparently dramatic improvement be enough to achieve a perfect fit? Fig. 3 shows that these random texts (RT_{N+1} model) still deviate from the real texts from which they borrow the character probabilities. For instance, Fig. 3 shows that random texts display pronounced humps for high frequencies and wider plateaus in the low frequency domain with regard to the

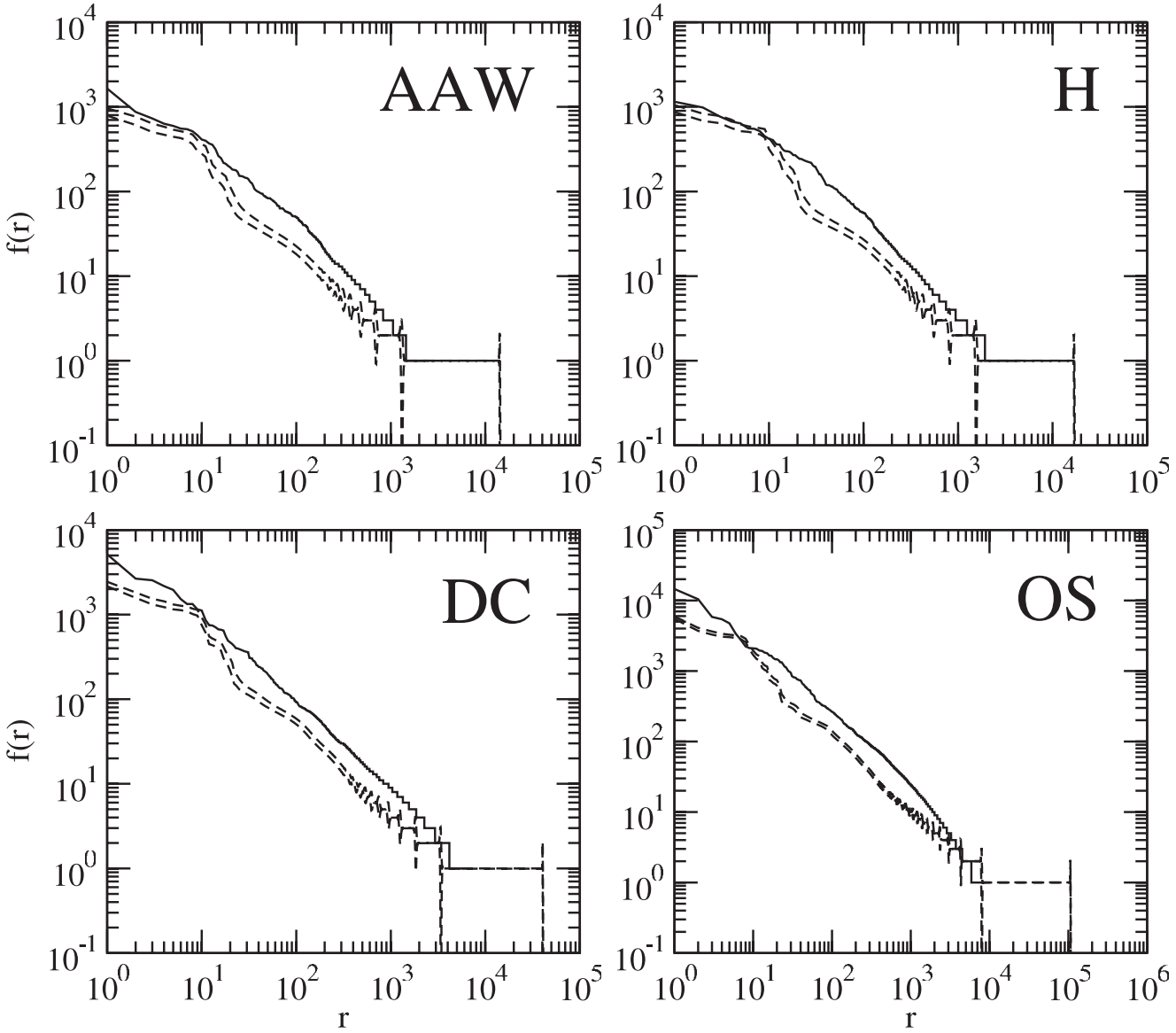

Figure 3. The rank histograms of English texts versus that of random texts (RT_{N+1}). The same as Fig. 1 for the model RT_{N+1} with alphabet size N and character probabilities obtained from the real text.
doi:10.1371/journal.pone.0009411.g003

real text. Besides, the rank histogram of random texts is clearly longer than that of real texts in Figs. 2 and 3. The fact that the plateaus at the low frequency region are much broader for RT_{N+1} texts than for real texts is well-known [19]. In the next section we show that the differences between real texts and random texts with non-commensurate character probabilities are statistically significant as well as for all the parameters suggested in the literature for RT_1 and RT_2.

In the next section, we employ rigorous statistical fitting, not because we think that it is strictly necessary when large visual differences between random and real texts are found (e.g., Figs. 1 and 2), but so as to provide a foundation for a more mathematically precise understanding of the differences between real texts and random texts and to extend, in a concise way, the analysis to the texts and parameters settings not considered in the figures. Notice that the poor visual fit of random texts shown in Figs. 1, 2 and 3 also applies to the real texts in Table 1 not visually examined in these figures so as to conserve space.

Rigorous Statistical Fitting

We detailed in the introduction that we did not seek to evaluate the goodness of fits of random texts for actual rank histograms through Zipf's law because this implies the risk that the target equation, i.e. Eq. 1, is not accurate enough [11,14]. A typical way of testing the fit of a certain model to real data is from the exact distribution that characterizes the model [12]. However, as mentioned in the introduction, this is impossible in the current situation because the exact rank distribution of random texts is unknown. To our knowledge, only approximations have been derived. Furthermore, the intention of our article was not to derive this equation *per se*. In light of the absence of such an exact distribution, we evaluate the consistency of ranks from a real text with those of a random text of the same length

(in words) through three different statistics of the rank r in a text of length T:

- $\max(r)$, the maximum rank. $\max(r)$ is the observed vocabulary size and measures the width of the rank histogram). Notice that the actual vocabulary of a random text model is infinite (*a priori*, any string of at least one letter can be formed), contrary to the actual vocabulary of a writer, which although large is finite. Support for $\max(r)$ comes from previous work suggesting that the pattern of observed vocabulary growth is useful for distinguishing between natural and RT_{N+1} texts [19]. We find that $\max(r)$ is indeed useful for any kind of random text and that simply its value is enough to distinguish random texts from real texts for the versions and parameter settings considered in this article.
- $\mu(r)$, the mean rank.
- $\sigma(r)$, the standard deviation of the rank.

To our knowledge, the expectation of these statistics for a text of a certain finite length has not previously been reported. If the rank distribution of the real texts and that of the random texts are the same, statistically significant differences between the value of the above statistics in real texts and those of random texts should not be found or be exceptional. Here we consider the whole set of ten English texts including the four works we examined in detail in the previous section (Table 1).

For each real text, we estimate the expectation and standard deviation of these statistics by generating 10^4 independent random texts for all the versions and parameters of the random text reviewed above. Notice that the length in words of the random texts is the same as that of the real text. Then we calculate k, the distance to the mean (measured in units of the standard deviation) between the value of real value of the statistic in the target text and that of a random text of a certain version and parameter setting. The three rank statistics yield three distances, i.e.,

$$k_{\max(r)} = \frac{\max(r) - \mu(\max(r))}{\sigma(\max(r))} \tag{6}$$

$$k_{\mu(r)} = \frac{\mu(r) - \mu(\mu(r))}{\sigma(\mu(r))} \tag{7}$$

$$k_{\sigma(r)} = \frac{\sigma(r) - \mu(\sigma(r))}{\sigma(\sigma(r))}, \tag{8}$$

The sign of the distance indicates whether the actual value is smaller than the expected ($k<0$) or larger than expected ($k>0$) for the hypothesis of a random text. Table 3 shows a summary of these signed distances for the texts in our data set.

How can we determine the significance of these distances? The Chebyshev inequality provides us with an upper bound of the, p-value, the probability that the value of the distance is due to mere chance for any kind of distribution. This upper bound is $1/|k|^2$, where $|...|$ is the absolute value operator [29]. Henceforth we use the term absolute distance to refer to $|k|$. We estimate the mean (μ) and standard deviation (σ) that are needed to compute the distances (Eqs. 6, 7 and 8) by simulating the version of the random text with the parameter setting under consideration a certain number of times (10^4 in our case). Table 3 shows, that all absolute distances (for any novel, any version of the random text and any parameter setting) are above 36.8. The minimum absolute distance is achieved by RT_{N+1} for CC with the parameters setting L_2 and the statistic $\mu(r)$. This means that the distance p-values, in all cases do not exceed $1/36^2 \approx 8 \cdot 10^{-4}$. Next we examine some concrete examples of the huge distance between a real text and a certain random text model and parameter setting using the results in Table 3. The minimum absolute distance achieved by any statistic for:

- the fair die rolling experiment considered in [7] (RT_1 with $N=5$) is 76.8 standard deviations, which is achieved by the text CC (*A Christmas carol*, by Dickens). This means that the p-value of the differences for any statistic and for all texts does not exceed $1/76.8^2 \approx 2 \cdot 10^{-4}$. In our version of the model, we do not allow for empty words to make the model more realistic.
- the variant of the random text model considered in [4] (RT_2) is 93.1 standard deviations, which is achieved by the text AAW (*Alice's adventures in wonderland*, by Carrol). This means that the p-value of the differences for any statistic and for all the texts does not exceed $1/93.1^2 \approx 10^{-4}$. In our version of the model, we do not allow for empty words to make the model realistic and estimate the parameters from the real text.
- the random text with unequal letter probabilities (RT_{N+1} with the three different parameters settings) is 36 standard deviations, which is achieved by CC with the parameter setting L_2 (this is the minimum distance for all versions of the random texts and parameter settings). Thus, the p-value of the differences for all statistics and for all the tests does not exceed $1/36^2 \approx 8 \cdot 10^{-4}$. This is striking, since it has been claimed that unequal letter probabilities improve the fit of random texts to the rank distribution of real texts dramatically [25]. In contrast, we show that the hypothesis of a random text is still rejected with unequal letter probabilities.

Next we focus on the sign of the distances in order to shed light on the nature of the disagreement between real and random texts. The sign of the distance indicates whether the actual value is too small ($k<0$) or too large ($k>0$) for the hypothesis of a random text. In all the cases shown in Table 3, the sign of this new distance is negative (the real values of the statistic are too small) except for RT_1 with $N=2$ and RT_{N+1} with the parameters setting L_1, where that distance is positive (the real values of the statistic are too large in these cases). A further statistical test confirming the results obtained thus far is presented in Text S1.

Discussion

We have seen that three different rank statistics are able to show, independently, that ten English texts and random texts with different versions and parameters settings are statistically inconsistent in all cases. We have seen that for the majority of the parameter settings considered, the nature of the disagreement is that the real rank statistic is smaller than that expected for a random text.

Although we have shown the poor fits of random texts by means of rigorous statistical tests, our limited exploration of the parameter space cannot exclude the possibility that random texts provide good fits for actual rank histograms with parameter values not considered here. Notice that random texts fail both with arbitrarily chosen parameters, e.g., the fair die rolling experiment [7] with $N=5$ and $p_b=1/6$ (model RT_1), and with parameters inferred from the target text, which would seem *a priori* more likely to yield a good fit. Despite our limited exploration of the parameter space, in the absence of concrete parameter values for

Table 3. Distance to the mean in standard deviations.

				RT_1			RT_2		RT_{N+1}	
Abbrv.	k	$N=2$	$N=4$	$N=5$	$N=6$	$N=26$	-	L_1	L_2	*Real*
AAW	$\max(r)$	42.6	−97.5	−133.2	−163.4	−573.4	−160.6	54.0	−74.3	−147.1
	$\mu(r)$	130.5	−59.7	−78.6	−94.2	−312.9	−93.1	173.0	−46.3	−85.7
	$\sigma(r)$	56.3	−83.4	−119.5	−156.8	−2033.6	−153.1	74.1	−63.0	−135.5
CC	$\max(r)$	99.1	−80.7	−120.0	−151.1	−555.4	−158.5	116.8	−53.7	−139.3
	$\mu(r)$	267.3	−54.5	−76.8	−93.6	−317.8	−98.0	347.2	−36.8	−87.1
	$\sigma(r)$	136.5	−72.8	−111.9	−149.5	−1969.6	−159.2	169.9	−48.7	−134.7
H	$\max(r)$	103.5	−86.6	−127.6	−158.4	−581.5	−157.7	121.5	−58.6	−142.3
	$\mu(r)$	277.8	−58.4	−81.4	−97.6	−331.3	−97.5	361.9	−40.5	−89.1
	$\sigma(r)$	142.1	−77.8	−118.3	−155.8	−2017.9	−154.6	176.8	−53.1	−135.4
ECHU	$\max(r)$	75.6	−133.8	−184.6	−226.0	−795.6	−275.9	93.7	−98.9	−240.5
	$\mu(r)$	247.4	−81.9	−108.5	−129.7	−431.3	−155.9	328.9	−61.6	−137.4
	$\sigma(r)$	106.3	−112.7	−161.7	−210.2	−2494.6	−278.7	138.1	−82.9	−227.2
HB	$\max(r)$	92.0	−131.4	−182.2	−225.5	−791.3	−246.2	112.8	−93.8	−207.3
	$\mu(r)$	272.8	−82.6	−109.2	−131.3	−432.7	−142.6	366.7	−60.5	−121.9
	$\sigma(r)$	127.8	−112.0	−161.0	−211.0	−2482.7	−238.3	165.5	−79.9	−189.0
ATS	$\max(r)$	120.7	−137.9	−195.9	−242.1	−854.7	−253.9	143.7	−97.7	−219.7
	$\mu(r)$	369.8	−87.6	−118.6	−142.1	−469.4	−148.1	488.6	−63.6	−130.6
	$\sigma(r)$	173.0	−118.1	−173.1	−226.1	−2620.6	−241.3	218.5	−83.9	−199.6
DC	$\max(r)$	119.2	−143.6	−201.8	−250.1	−882.0	−294.2	143.9	−102.1	−246.5
	$\mu(r)$	404.6	−89.5	−120.6	−145.5	−482.0	−168.9	540.4	−64.0	−143.9
	$\sigma(r)$	175.7	−121.8	−177.0	−232.0	−2678.1	−288.0	224.7	−86.6	−226.7
OS	$\max(r)$	72.9	−258.2	−341.1	−419.4	−1446.2	−539.9	100.0	−205.0	−443.3
	$\mu(r)$	349.3	−148.4	−189.5	−228.6	−754.5	−289.1	486.6	−119.9	−240.5
	$\sigma(r)$	117.7	−203.8	−279.5	−362.9	−3939.7	−514.2	164.6	−160.9	−390.7
MB	$\max(r)$	222.1	−221.6	−311.1	−392.2	−1418.5	−470.9	266.4	−155.8	−382.7
	$\mu(r)$	849.5	−137.8	−184.8	−226.4	−765.8	−266.8	1152.2	−98.0	−221.3
	$\sigma(r)$	352.8	−184.8	−265.5	−350.9	−3908.0	−444.5	452.7	−130.7	−339.9
U	$\max(r)$	404.3	−200.7	−303.2	−398.6	−1491.0	−481.1	466.6	−120.8	−388.7
	$\mu(r)$	1672.5	−133.7	−190.8	−241.7	−828.3	−285.2	2206.4	−78.8	−235.9
	$\sigma(r)$	693.6	−175.0	−266.7	−364.3	−4068.4	−462.1	862.0	−107.3	−354.5

Summary of k, the distance to the mean (in standard deviations), between real values and those of random texts for three different rank statistics: $max(r)$ (the maximum rank), $\mu(r)$ (the mean rank) and $\sigma(r)$ (the standard deviation of the rank). The first column contains the abbreviation of the text (see Table 1 for the meaning of each abbreviation). Texts are sorted by increasing length. The columns after the first column correspond to different versions of the random text model and different parameter settings. For each text and parameter setting, we show $k_{\max(r)}$, $k_{\mu(r)}$ and $k_{\sigma(r)}$, the distances from each of the three rank statistics. N is the number of characters other than space. L_1 and L_2 are two parameter settings borrowed from [2]. *Real* indicates that all character probabilities are obtained from the original text. Distances are computed from the estimated mean and standard deviation of the rank of a certain random text through 10^4 independently generated replicas. The random texts have the same length in words as the target real text.

doi:10.1371/journal.pone.0009411.t003

which random texts fit real rank histograms accurately, the meaningfulness for natural languages of Zipf's law-like word distributions remains viable.

We believe that the quest for parameters that provide a good fit of random texts on real texts is a tough challenge for detractors of the meaningfulness of Zipf's law, because real writers do not produce words by concatenating independent events under a certain termination probability. Real writers extract words from a mental lexicon that provides almost 'ready to use' words [30]. Our main point here is that generally the lexicon provides root word forms that can be completed with affixes. The valid root forms are basically determined *a priori*. Although writing can be a very creative exercise, real writers do not construct words 'on the fly' as in the random texts that have previously been presented as an argument against the utility of probability distributions in language. Although some writers do invent many words, their creativity is limited by the need to be understood by their readers. Indeed, the meaning of invented words has to be guessed from the surrounding words. If the context words are also invented, then the reader is likely to get completely lost. Considered from this perspective, random texts are a case of maximum word creativity, and are not limited by a need to be understood (recall the meaningless examples of random texts in the Introduction and the Results section) but only constrained by the prior character probabilities.

There are still many models of Zipf's law for which the goodness of fit to real texts has not been studied rigorously (e.g., [31,32]). A remarkable exception is [21]. Further research is necessary in

order to establish which models provide the best fit in a statistically rigorous sense. Indeed, this is yet another reason to conclude that two fundamental research problems about Zipf's law in natural languages, namely its meaningfulness and a realistic explanation of it, remain open.

Materials and Methods

Materials

To simplify the analysis, we normalize the English texts in Table 1 by removing all marks, lower casing all letters, converting all spaces into blanks and leaving only one blank after each word. In this way, we obtain a sequence of words whose length is at least one character and separated by a single blank. A similar normalization procedure is used in [19] although this study does not provide enough details to determine if its normalization procedure is exactly the same as ours.

After text normalization, there is a small fraction of word characters that are not letters in the English alphabet. Most of these characters are digits or accents. To make sure that our results are not a consequence of these infrequent characters we repeated the fitting tests excluding words not made exclusively of English lowercase letters from 'a' to 'z' after text normalization. We found that the results were qualitatively identical: each of the three rank statistics is able to reject the hypothesis of a random text in all cases.

Computational Methods

Here we aim to provide some guidelines to perform the computer calculations presented in this article for easy replication of our results. In what follows we consider the computational efficiency of three issues: (i) the generation of random words; (ii) counting the frequency of random words; (iii) and sorting.

Random word generation. Here we explain how to generate a random word efficiently. We start with the simplest (or naïve) algorithm of random word generation (we assume that the space delimiting words does not belong to the word):

1. Start with an empty string of characters s.
2. Generate a random character c and add it to s.
3. Generate a uniform random deviate $x \sim U(0,1)$.
4. While $x \geq p_b$ do
5. Generate a random character c and add it to s.
6. Generate a uniform random deviate $x \sim U(0,1)$.

Generating a uniformly distributed random letter (steps 2 and 5) for the models RT_1 and RT_2 with a standard random uniform given the alphabet size N is straightforward. Generating a random letter for RT_{N+1} where probabilities come from a real text can also be easily performed in $\theta(1)$ time using a table look-up method [33]. If character probabilities come from a real text (as in the parameter setting *Real* (Results section), we can place all the characters other than space from that text in a table and then choose one uniformly in $\theta(1)$ time. This needs space $\theta(N_c - N_b)$, where N_c is the number of characters of the text after text normalization (including blanks) and N_b is the number of blanks. If the character probabilities are given *a priori* but are rational numbers (as in the parameter settings L_1 and L_2 borrowed from [2]), then we can generate a table where the relative frequency of each character is the same as the desired probability. Alternatively, for the case of RT_{N+1} in general, one can use an inversion method with a guided table [33]. The fact that normally $N \ll N_c - N_b$ in large enough texts implies that this inversion method requires less space than the table look-up method while keeping the $\theta(1)$ time for generating a random letter.

Imagine that a random word has length L (we assume $L \geq 1$ in our random texts). The naïve algorithm above needs invoking a random uniform deviate generator $2L$ times, i.e. L times for generating each of the L random characters (steps 2 and 5) and L times for determining if more characters have to be added or not (steps 3 and 6). We can reduce the number of random uniform deviates that need to be generated using the following algorithm:

1. Generate a random geometric deviate $L \sim G(p_b)$.
2. Generate a random word w of length L,

where is Step 2 is performed through the following algorithm

1. Start with an empty string of characters s.
2. Repeat L times
3. Generate a random character c and add it to s.

Of key importance is the generation of the random geometric deviate in $\theta(1)$ time. It is possible to generate a random geometric deviate L with parameter p_b $(L \geq 1)$ from a random uniform deviate x through the formula [33,34]

$$L = 1 + \left\lfloor \frac{\log x}{\lambda} \right\rfloor, \tag{9}$$

where $\lambda = \log(1 - p_b)$. To save computation time, the constant λ is calculated only once. The second version of the algorithm needs to generate only $L+1$ uniform deviates (one for generating the geometric deviate and L for each of the L characters comprising the word) whereas the naïve first version required $2L$ uniform deviates.

It is still possible to generate a random word of length L with only L uniform deviates. The idea is to allow the blank to be among the characters that can be generated once the first non-blank character has been placed. The algorithm is

1. Start with an empty string of characters s.
2. Generate a random character c (c cannot be a blank) and add it to s.
3. Generate a random character c (possibly a blank).
4. While c is different than blank do
5. Add c to s.
6. Generate a random character c (possibly a blank).

Word frequency counting. We define T as the length of a text in words. By ignoring the length of a word, the frequency of a word efficiently can be counted in $\theta(1)$ time and $\theta(T)$ space using a hashing table [35] of character strings.

With simultaneous random word generation and counting, the time efficiency can be improved by employing more memory for the case of RT_1 and RT_2. The idea is to keep the hashing table only for counting the frequency of the words of lengths greater than L_{max} and using a matrix $F = \{f_{ij}\}$ for counting the frequency of each of the N^i words of length i such that $1 \leq i \leq L_{max}$. f_{ij} is the frequency of the $j-th$ word of length i with $1 \leq j \leq N^i$. In this way, a random word of length L_{max} or smaller can be simultaneously generated and counted involving only two random deviates with the following simple algorithm:

1. Generate a random geometric deviate $i \sim G(p_b)$
2. If $i \leq L_{max}$ then

3. Generate a random uniform number $j \sim U[1,N^i]$.
4. Increase f_{ij} by one.
5. else
6. Generate a random word w of length i by means of the algorithm above.
7. Increase the frequency of w by updating the hashing table of character strings.

The extra memory needed for the table of words of length not exceeding L_{max} is

$$S(L_{max}) = \sum_{L=1}^{L_{max}} N^L \tag{10}$$

$$= \begin{cases} \frac{N(N^{L_{max}}-1)}{N-1} & \text{if } N \geq 2 \\ L_{max} & \text{if } N = 1 \end{cases} \tag{11}$$

Sorting. Sorting natural numbers efficiently is needed to calculate ranks. Obtaining the ranks of a certain text (real or random) requires sorting the word frequencies from the random text in decreasing order. All the above techniques may not contribute to increase significantly the speed of the computer calculations if the sorting takes more than $\theta(T)$ time, where T is the length in words of the text. In our case, we can take advantage of the fact that frequencies lie within the interval $[1,T]$ and then we can use counting sort [35], which allows one to sort elements in linear time.

Supporting Information

Text S1
Found at: doi:10.1371/journal.pone.0009411.s001 (0.24 MB PDF)

Acknowledgments

We are grateful to R. Gavaldà, B. McCowan, J. Mačutek and E. Wheeler for helpful comments. We also thank M. A. Serrano, A. Arenas, S. Caldeira for helpful discussions and J. Cortadella for computing facilities.

Author Contributions

Conceived and designed the experiments: RFiC BE. Performed the experiments: RFiC. Analyzed the data: RFiC. Contributed reagents/materials/analysis tools: RFiC. Wrote the paper: RFiC BE.

References

1. Zipf GK (1972) Human behaviour and the principle of least effort. An introduction to human ecology. New York: Hafner reprint. 1st edition: Cambridge, MA: Addison-Wesley, 1949.
2. Li W (1992) Random texts exhibit Zipf's-law-like word frequency distribution. IEEE T Inform Theory 38: 1842–1845.
3. Rapoport A (1982) Zipf's law re-visited. Quantitative Linguistics 16: 1–28.
4. Miller GA, Chomsky N (1963) Finitary models of language users. In: Luce RD, Bush R, Galanter E, eds. Handbook of Mathematical Psychology. New York: Wiley, volume 2. pp 419–491.
5. Miller GA (1957) Some effects of intermittent silence. Am J Psychol 70: 311–314.
6. Mitzenmacher M (2003) A brief history of generative models for power law and lognormal distributions. Internet Mathematics 1: 226–251.
7. Suzuki R, Tyack PL, Buck J (2005) The use of Zipf's law in animal communication analysis. Anim Behav 69: 9–17.
8. Mandelbrot B (1953) An informational theory of the statistical structure of language. In: Jackson W, ed. Communication theory. London: Butterworths. 486 p.
9. Nicolis JS (1991) Chaos and information processing. Singapore: World Scientific.
10. Ferrer-i-Cancho R, Gavaldà R (2009) The frequency spectrum of finite samples from the intermittent silence process. Journal of the American Association for Information Science and Technology 60: 837–843.
11. Herdan G (1964) Quantitative linguistics. London: Butterworths.
12. Miller DW (1995) Fitting frequency distributions: philosophy and practice. Volume I: discrete distributions. New York: Book Resource.
13. Miller DW (1995) Fitting frequency distributions: philosophy and practice. Volume I: continous distributions. New York: Book Resource.
14. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703.
15. Popescu II, Altmann G, Köhler R (2009) Zipf's law – another view. Quality and Quantity: in press.
16. Conrad B, Mitzenmacher M (2004) Power laws for monkeys typing randomly: the case of unequal probabilities. IEEE Transactions on Information Theory 50: 1403–1414.
17. Manning CD, Schütze H (1999) Foundations of statistical natural language processing. CambridgeMA: MIT Press, chapter Introduction.
18. Mandelbrot B (1966) Information theory and psycholinguistics: a theory of word frequencies. In: Lazarsfield PF, Henry NW, eds. Readings in mathematical social sciences. Cambridge: MIT Press. pp 151–168.
19. Cohen A, Mantegna RN, Havlin S (1997) Numerical analysis of word frequencies in artificial and natural language texts. Fractals 5: 95–104.
20. Zipf GK (1935) The psycho-biology of language. Boston: Houghton Mifflin.
21. Balasubrahmanyan VK, Naranan S (1996) Quantitative linguistics and complex system studies. J Quantitative Linguistics 3: 177–228.
22. Ferrer i Cancho R (2007) On the universality of Zipf's law for word frequencies. In: Grzybek P, Köhler R, eds. Exact methods in the study of language and text. To honor Gabriel Altmann. Berlin: Gruyter. pp 131–140.
23. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46: 323–351.
24. Manin DY (2008) Zipf's law and avoidance of excessive synonymy. Cognitive Science 32: 1075–1098.
25. Wolfram S (2002) A new kind of science. Champaign: Wolfram Media.
26. Ferrer i Cancho R, Solé RV (2002) Zipf's law and random texts. Advances in Complex Systems 5: 1–6.
27. Ferrer i Cancho R (2005) Zipf's law from a communicative phase transition. European Physical Journal B 47: 449–457.
28. Bell TC, Cleary JG, Witten IH (1990) Text Compression. Englewood Cliffs, NJ: Prentice-Hall.
29. DeGroot MH (1989) Probability and statistics. Reading, MA: Addison-Wesley. 2nd edition.
30. Levelt WJM (2001) Spoken word production: a theory of lexical access. Proc Natl Acad Sci USA 98: 13464–13471.
31. Ferrer i Cancho R, Solé RV (2003) Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences USA 100: 788–791.
32. Zanette D, Montemurro MA (2005) Dynamics of text generation with realistic Zipf's distribution. J Quantitative Linguistics 12: 29–40.
33. Devroye L (1986) Non-uniform random variate generation. New York: Springer-Verlag.
34. Dagpunar JS (1988) Principles of random variate generation. Oxford: Clarendon Oxford Science Publications.
35. Cormen TH, Leiserson CE, Rivest RL (1990) Introduction to algorithms. Cambridge, MA: The MIT Press.