NSURJ
NEVADA STATE UNDERGRADUATE RESEARCH JOURNAL

Changes in the Coefficients of Zipf's Law for English Corpora of Different Contexts

Robert A. Arnold and Deena R. Schmidt

University of Nevada, Reno. Department of Mathematics and Statistics. Reno, NV

DOI: http://dx.doi.org/10.15629/6.7.8.7.5_3-1_S-2017-1

Abstract: The statistical behavior of languages is of great interest to linguists and has been since the mid-20th century. The frequency distribution of words is often modeled with a discrete probability distribution called Zipf's Law (Zipf, 1936, 1949). Research involving this modeling approach has undergone increasingly intense scrutiny over the last decade. It turns out that there is an interesting gap in the research of Zipf's Law, as pointed out by Steven Piantadosi (2014) – namely, language is not static and changes from context to context, and there is comparatively little examination of these fluctuations using Zipf's Law. This paper is a first step toward examining how language changes between different time periods and different types of written and spoken media via comparing the parameters of best fit for the data using Zipf's Law.

1 Introduction

In the parlance of linguistics, and as defined for this paper, a corpus (plural, corpora) is any ordered collection of words where repeated elements are allowed. This may simply be a list of random words, a cohesive body of work, or a compilation of many books and other media forms. The meaning of "repeated elements are allowed" is that the quotations (a) "the dogs are good dogs" and (b) "the dogs are good" may be considered as two different corpora, with (a) having five words and (b) having four words, even though both corpora draw from the same set of four tokens: "the", "dogs", "are", and "good".

This relates to two additional important definitions for this paper, which are the count (or number of occurrences of a word) and frequency (or proportion) of a word in a given corpus. In the previous example, the word "dogs" has a count of 2 and frequency of 2/5 in quotation (a) and a count of 1 and frequency of 1/4 in (b).

Finally, the frequency rank of a word is given by the index of a list which orders the words appearing in a corpus by frequency, from most frequent to least frequent. For example, in quotation (a) "dogs" has frequency rank 1 due to appearing more frequently than any other word, whereas the tokens "the", "are", and "good" each tie for frequency rank 2.

1.1 Zipf's Law

Given a corpus of grammatical natural language text (not necessarily English), the frequency of a word is approximately inversely proportional to its frequency rank in the frequency table. This empirical observation of natural languages is referred to as Zipf's Law (Zipf, 1936, 1949). Though he was not the first person to notice that particular behavior of natural languages, the law is named after George Kingsley Zipf, a linguist who investigated the law and promoted its use in multiple disciplines. Formally stated, Zipf's Law is described by the following relation:

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{1}$$

Here, r is the frequency rank of a given word within a text, f(r) is the frequency, and $\alpha>0$ is a parameter which is determined empirically. In practice, many corpora yield $\alpha \approx 1$. When measuring the frequency of words in a corpus, the most frequent word has rank 1 (r = 1), the second most frequent word (with r = 2) will tend to be about half as frequent as the most frequent word, the third most frequent word (with r = 3) will be a third as frequent as the most frequent word, and so on.

Zipf's Law is a special case of a power law distribution. In particular, it is a discrete distribution with probability mass function given by

$$f(r) = Cr^{-\alpha}, \qquad r = 1, 2, \ldots \tag{2}$$

where r is a positive integer, $\alpha > 1$, and $C > 0$ is a normalization constant. Zipf's Law is also known as a discrete Pareto distribution, or a zeta distribution (Newman, 2005). This distribution is typically shown graphically on a log-log plot as a straight line with slope equal to $-\alpha$. Note that $\alpha > 1$ is required for the mean to be finite. If we restrict the sample space from the set of positive integers to a bounded set of integers $\{x_{min}, x_{min+1}, \ldots, x_{max}\}$, then we allow for $\alpha > 0$.

The significance of Zipf's Law is that it identifies a persistent pattern that appears to be universally present in all natural languages (Baroni, 2009). Understanding the structure of natural language is of material consequence because, for instance, natural language translation software requires algorithms which take a string of words in one language and return some other, different string of words in another language. These algorithms encode instructions which presuppose that a pattern in one language is equivalent to a pattern in another language. Translation software is used daily by millions of people in order to facilitate commerce, cultural exchanges, and travel. Therefore, any advancement in the understanding of patterns which appear in all languages is beneficial for improving communication around the world.

Modern investigations of Zipf's Law have focused on two broad research areas: 1. Using Zipf's Law to compare different languages (Gelbukh and Sidorov, 2001), and 2. Using Zipf's Law on small corpora (such as The Hound of the Baskervilles and Alice in Wonderland (Baayen, 2001)) which also includes restricted categories of words (such as "nouns" and "verbs") within the same language (Marcus et al. 1993). These areas naturally have some overlap in the research. For example, a list of very common words (such as "water", "father", "mother") that is translated into a sample of world languages (such as English, French, German, Swedish, etc.) is called a Swadesh List (Calude and Pagel, 2011). Such a list counts as several comparable corpora taken from world languages, but also counts as a category of words that is relatively restricted to content that is universally present in every human society. The key observation is that these avenues of research – different languages and small corpora – neatly miss the very 'non-staticness' of language that Piantadosi (2014) points out, by not examining or taking into consideration change in language over an extended time period or change in language between contexts. For example, the goal of a newspaper is to inform a general audience, while the goal of an academic paper is to inform a much more highly educated general audience or specified experts. These different contexts may result in changes in the word frequencies.

These hidden variables could distort the results of other experiments. Is the time period of a corpus a confounding variable that a researcher comparing two or more corpora needs to control for? This gap in the research needs to be addressed.

1.2 Research Question

Given the aforementioned context surrounding research on Zipf's Law, the main question we investigate in this paper is: Are there substantial differences between large corpora from the same language but from different contexts? Specifically, what would result in applying Zipf's Law to large corpora organized by time period (for example, a corpus of text taken from the 19th century compared with a corpus of text taken from 1900 – 1950) and by media form (for example, a corpus of text taken solely from newspapers compared with a corpus of text taken solely from works of fiction)?

2 Background

2.1 Using Zipf's Law to Compare Different Languages

Zipf's Law has often been used to study differences among world languages. The questions answered by such studies ask how Zipf's Law changes depending on what language is used. If the grammatical structure of the language is radically different, does that significantly change the expected results from fitting a power curve like Zipf's Law?

Gelbukh and Sidorov showed that there exists a more than 1% change in the power law exponent for Zipf's Law (2001) – this refers to the α parameter in Equation (1). The best-fit coefficient was calculated using a linear regression method employing maximum likelihood estimates (Larsen and Marx, 2006) for a corpus of approximately 2.5 million words of English text versus a corpus of 2 million words of Russian text. Each word was taken from a wide range of genres including children's books and science fiction. This breadth and depth of the corpora show that there is a measurable difference in the parameter when comparing two different languages and that this difference is not incidental – that is, a 1% difference is very large when taking into account the number of words and different genres of the corpora with which Gelbukh and Sidorov were working.

More recently, researchers have gone on to explore languages outside of the Indo-European family, such as Chinese, Japanese, and Korean (Lu et al. 2013), noting that thus far, the bulk of research into Zipf's Law and its relationship with natural language have focused on Indo-European languages, as opposed to other language families, such as the Sino-Tibetan family (Chinese) or the Altaic family (Korean). Lu Linyuan's study found that characters taken from these languages exhibited behavior that significantly deviated from the expected power-law pattern following Zipf's Law built from the study of words taken from European languages – notably, that Chinese, Korean, and Japanese characters did not follow Zipf's Law. However, Lu's study did not correctly implement Zipf's Law; the debilitating flaw in that paper lies not in its mathematics, but in the assumption of semantic equality between a word taken from a European language and a Chinese character. Many characters in Chinese form compound words with other characters, behaving much more like morphemes than words. For example, 毕业生 (bì yè shēng, English: graduate) is composed of three characters, each having a meaning (in order: "complete", "industry", and "green") which is unrelated to the whole word. In other words, Chinese characters and words cannot be considered comparable to each other.

These studies show that there is no shortage of attention being given to the applications of Zipf's Law to corpora taken from world languages. However, in this paper, we do not examine the tendency for language to change over time or for the language to change depending on the context.

2.2 Using Zipf's Law to Compare Small Corpora

We discuss studies comparing texts written solely in the same language, specifically English. What is available to the researcher interested in comparing texts using Zipf's Law when the texts are written in the same language? The set of rules describing the usage of the language suggests itself.

Comparisons have been made between the Zipf's Law parameter α in Equation (1) for categories of words within a language, such as parts of speech (Marcus et al. 1993). In their paper, they compared grammatical categories such as determiners, nouns, and verbs in the third person present form (in English), and found that these forms followed Zipf's Law closely, though there were interesting variations. For example, plotting rank versus frequency on a logarithmic scale shows that the curve tends to curve concave down. The presence of a pattern can aid in the task of highlighting abnormal text selections. For example, one deviation from the concave-down behavior observed in Marcus et al. (1993) – in the case of using verbs to construct a Zipf's Law curve of best fit—is that the curve is concave up instead of down, as noted by Piantadosi (2014).

That language changes according to context misses the heart of the matter because the corpora described in the previous paragraph are very small or highly restricted, and not representative of an entire language. The coefficients of best fit for a single book may significantly differ from a corpus that draws from a large pool of text.

The research reviewed in this paper is recent, but there is a simple explanation for why there are an abundance of experiments being done now when looking at the historical context. There are now huge, organized databases of texts and computers capable of quickly sorting through and producing tables of these data. Thus, there is a proliferation of research into the differences between world languages and larger corpora, which only recently became possible due to technological advancements (Moreno-Sanchez et al. 2016). Current research is concerned with the features of language in the present, with little concentration on the effect of language evolution over time. Piantadosi (2014) discusses social forces changing or adding new words (such as "email"), but nowhere does he mention a systematic study that examines word frequencies over a long period of time. To our knowledge, no studies have tested the way Zipf's Law changes over time and between contexts, which led us to our main question.

3 Methods

To investigate our main question, we formulate the following hypotheses. The null hypothesis is that the parameters of the Zipf's Law fit between the corpora organized by time period and those organized by media form will not be significantly different. Then the alternative hypothesis is that the parameters of the Zipf's Law fit will have significant differences. The methods described below are used to determine whether or not we can reject the null hypothesis for the data we consider.

3.1 Data and Software

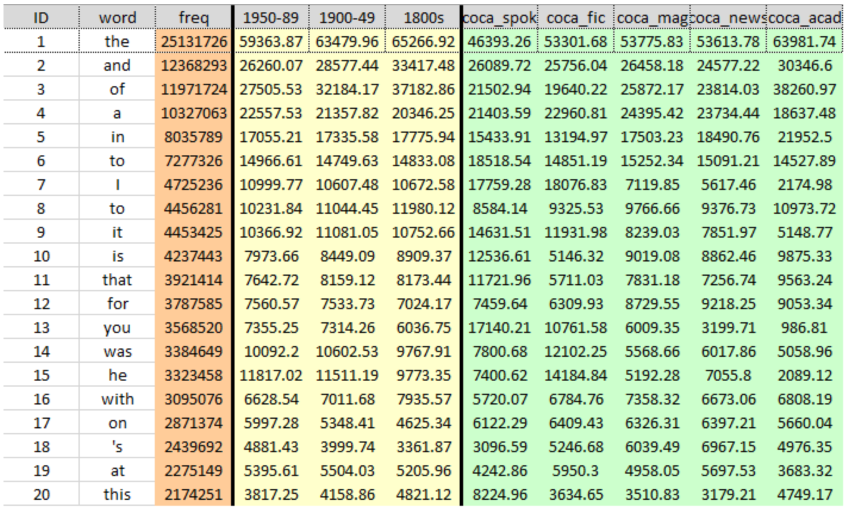
We purchased data from Brigham Young University's CORPORA, specifically the word frequency data for the top 100,000 words from the Corpus of Contemporary American English (COCA) (Davies 2008) and the Corpus of Historical American English (COHA) (Davies 2010). From the COCA website: "[The] COCA includes 440 million words taken from 190,000 texts during 1990-2012, evenly divided (~88 million words each) into spoken, fiction, magazine, newspaper, academic." From the COHA website: "The COHA data includes 385 million words of text in 116,000 different texts from the 1810s-2000s, in fiction, popular magazines, newspapers, and non-fiction (books)." Note that this is truncated data with very large sample sizes which sets up the need for using a truncated version of Zipf's law to analyze the data (see Methods). For this paper, we did not investigate how the data was collected or question the integrity of the data. Since it is one of the most extensive databases available to date, many graduate students and professors use the corpus, a list of which can be seen on the corpus website at (http://corpus.byu.edu/researchers.asp).

These data are organized into eight sub-corpora which we have grouped into two main categories: time-period data and media-form data. The time-period data consist of words taken from texts from the 1800s, 1900-1949, and 1950-1989, each category without regard for different media forms. The media-form data consist of words taken from newspapers, magazines, fiction literature, academia, and spoken (speech from television or radio), each category specifically taken from between 1990-2012. These data are analyzed using the R statistical language

(R Core Team, 2016), with the eight sub-corpora represented in R as separate "data-frame" objects, each with two columns: the frequency ranks of given words (sorted from most frequent to least frequent for each sub-corpora), and the frequencies of the corresponding words.

For each sub-corpus, any word which had zero tokens (zero appearances in the corpus) was removed from the list. The rationale for not including zero-frequency words is simple: if a word doesn't appear in the corpus, then creating a model which notes its absence is a strange thing to do. For example, "e-mail" is, of course, a word that doesn't appear in the 1800s. When building a model for words taken from texts from the 1800s, why have a data point that says "e-mail" appeared zero times? To be fair, there would then be a need to have such a data point for every English word that is not present in the 1800s corpus. For this reason, we decided to discard any word which doesn't appear. See Table 1 for a screenshot of the spreadsheet from which the data is drawn and used in the model (Davies 2008, 2010).

Table 1: Data screenshot: Word frequency data for the top 100,000 words from the Corpus of Contemporary American English (COCA) (Davies 2008) and the Corpus of Historical American English (COHA) (Davies 2010). The leftmost column (ID) is the frequency rank of the word in the second column (word). The orange column (freq) is the total number of occurrences (or counts) of the word in the COCA. The yellow columns correspond to the time-period frequency data and the green columns correspond to the media-form frequency data (reported in per-million word frequencies).



ID	word	freq	1950-89	1900-49	1800s	coca_spok	coca_fic	coca_mag	coca_news	coca_acad
1	the	25131726	59363.87	63479.96	65266.92	46393.26	53301.68	53775.83	53613.78	63981.74
2	and	12368293	26260.07	28577.44	33417.48	26089.72	25756.04	26458.18	24577.22	30346.6
3	of	11971724	27505.53	32184.17	37182.86	21502.94	19640.22	25872.17	23814.03	38260.97
4	a	10327063	22557.53	21357.82	20346.25	21403.59	22960.81	24395.42	23734.44	18637.48
5	in	8035789	17055.21	17335.58	17775.94	15433.91	13194.97	17503.23	18490.76	21952.5
6	to	7277326	14966.61	14749.63	14833.08	18518.54	14851.19	15252.34	15091.21	14527.89
7	I	4725236	10999.77	10607.48	10672.58	17759.28	18076.83	7119.85	5617.46	2174.98
8	to	4456281	10231.84	11044.45	11980.12	8584.14	9325.53	9766.66	9376.73	10973.72
9	it	4453425	10366.92	11081.05	10752.66	14631.51	11931.98	8239.03	7851.97	5148.77
10	is	4237443	7973.66	8449.09	8909.37	12536.61	5146.32	9019.08	8862.46	9875.33
11	that	3921414	7642.72	8159.12	8173.44	11721.96	5711.03	7831.18	7256.74	9563.24
12	for	3787585	7560.57	7533.73	7024.17	7459.64	6309.93	8729.55	9218.25	9053.34
13	you	3568520	7355.25	7314.26	6036.75	17140.21	10761.58	6009.35	3199.71	986.81
14	was	3384649	10092.2	10602.53	9767.91	7800.68	12102.25	5568.66	6017.86	5058.96
15	he	3323458	11817.02	11511.19	9773.35	7400.62	14184.84	5192.28	7055.8	2089.12
16	with	3095076	6628.54	7011.68	7935.57	5720.07	6784.76	7358.32	6673.06	6808.19
17	on	2871374	5997.28	5348.41	4625.34	6122.29	6409.43	6326.31	6397.21	5660.04
18	's	2439692	4881.43	3999.74	3361.87	3096.59	5246.68	6039.49	6967.15	4976.35
19	at	2275149	5395.61	5504.03	5205.96	4242.86	5950.3	4958.05	5697.53	3683.32
20	this	2174251	3817.25	4158.86	4821.12	8224.96	3634.65	3510.83	3179.21	4749.17

The leftmost column in Table 1 is the frequency rank of the word given in the second column. The orange highlighted column is the corresponding overall raw frequency (the number of occurrences) of the word in the COCA. The three yellow columns with the respective headings 1950-89, 1900-49, and 1800s are the per-million word frequencies for those time-period categories taken from the COHA. The five green columns that are headed respectively with coca_spok, coca_fic, coca_mag, coca_news, and coca_acad are the per-million word frequencies for those media-form categories taken from the COCA.

Note that the word “the” is the most frequent word in all cases considered (r = 1). The number of occurrences of “the” in the 1800s corpus, for instance, is 65,266.92 per million words. The word “of” has an overall frequency rank of 3 which means that it is the third most frequent word overall, but it has rank 2 (instead of rank 3) in four out of the eight sub-corpora we considered. Since each sub-corpus is sorted from the most frequent word to the least frequent word before frequency ranks are assigned, only the rank and the frequency are used in the analysis, not the words themselves. In the case that two or more words have the same frequency, they are arbitrarily ranked.

Lastly, note that these data sets are very large since the word frequencies are reported as counts per-million words. Thus, the sample size (n) in our model will be very large, and we discuss the modeling issues that result from very large n in the next section.

3.2 Statistical Analysis

Many real-world examples violate Zipf’s law because the upper and/or lower tails of the data do not follow a power law, i.e. those tails are not linear on a log-log plot of rank versus frequency (Newman 2005, Langlois et al. 2014, see Figure 1 below). Typically, the plots may be curvy at first for small rank but then straighten out into a linear region where the power law is a good fit, and curve downward in the tail. Our word frequency data show this pattern which led us to truncate the data to the linear intermediate region and then fit the model described below to that region. We follow the method for bounded discrete power laws outlined in Langlois et al. (2014).

The model used to fit the data is a discrete bounded power law distribution (or discrete truncated Pareto distribution) with a given lower bound ($x_{\min}$) and upper bound ($x_{\max}$). The probability mass function is given by

$$f(x) = Cx^{-\alpha}, x = 1,2,\ldots \qquad (3)$$

$$\text{where} \sum_{x=x_{\min}}^{x_{\max}} Cx^{-\alpha} = 1$$

This is a truncated version of the original Zipf probability distribution (where C > 0 is a normalization constant), and a more accurate model to fit to word frequency data than other popular methods, such as the method described in Power-Law Distributions in Empirical Data (Clauset et al. 2009). Clauset’s method (2009) does not allow for both an upper and lower bound which is clearly needed in our case.

3.2.1 Parameter Estimation

Here we outline the steps needed to estimate the power law exponent (slope parameter α) in Equation (3).

- The best fit parameter is given by the maximum likelihood estimator (MLE) of α which is obtained by maximizing the log-likelihood function LL(α) defined in Equation (5) below (Bauke 2007, Langlois et al. 2014). In practice, it is more computationally effective to minimize the negative log-likelihood function which has an equivalent optimum:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}[-LL(\alpha)] \qquad (4)$$

$$\text{where} \quad LL(\alpha) =$$

$$-\alpha\left(\sum_{i=1}^{n} \ln x_i\right) - n\ln\zeta(\alpha, x_{\min}, x_{\max}). \qquad (5)$$

The function $\zeta(\alpha, x_{\min}, x_{\max})$ is given by

$$\zeta(\alpha, x_{\min}, x_{\max}) = \zeta(\alpha, x_{\min}) - \zeta(\alpha, x_{\max}) \qquad (6)$$

where $\zeta(\alpha, x)$, the generalized or Hurwitz zeta function, is given by

$$\zeta(\alpha, x) = \sum_{i=1}^{\infty} \frac{1}{(i+x)^{\alpha}}. \qquad (7)$$

Note that when x = 1, ζ(α,1)=ζ(α) which is the Riemann zeta function.

- Computational results in simulated discrete truncated power law data showed no bias in $\hat{\alpha}$, the MLE for α, for the parameter range (α between 1 and 6) and sample sizes considered here.
- It is important to point out that $\hat{\alpha}$ depends on the values for x_{min} and x_{max}. We have to choose these carefully, and this is not straightforward in our case. Even formal ways of doing this (see Bauke 2007) involve a bit of subjective input when truncating the data. This issue will be discussed further in the next section where we describe our approach to uncertainty quantification in our estimates.

3.2.2 Uncertainty Quantification

The standard approach to uncertainty quantification for the α estimates would be to use an asymptotically normal approximation where the uncertainty comes from having a small sample size *n* (i.e. the uncertainty depends on *n*). Following Equation 3.6 from Clauset et al. (2009), the standard error of $\hat{\alpha}$, σ, is derived from the width of the likelihood maximum, which is given by:

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n) \qquad (8)$$

In our case, we have very large sample sizes (ranging from 868,761,000,000 (Newspaper corpus) to 940,050,340,000 (1800s corpus)) so the standard error of the estimates is essentially zero. Instead, the main source of uncertainty in $\hat{\alpha}$ comes from how the data are truncated, i.e. uncertainty in $x_{\min}$ and $x_{\max}$.

Techniques for estimating the lower and upper bounds ($x_{\min}$ and $x_{\max}$) from raw data are discussed in Bauke (2007), and we follow the approach described below:

- Choose conservative (broad) intervals for candidate $x_{\min}$ and $x_{\max}$ bounds by visually inspecting the graphs of rank versus frequency for each of the eight corpora. The intervals we used are listed in Table 2, and they are shown as vertical lines on each plot in Figure 1.
- The intervals are chosen so that the most linear intermediate region lies between the upper $x_{\min}$ range and the lower $x_{\max}$ range (i.e. the inner 2 vertical lines). The lower $x_{\min}$ and upper x_{max} cutoffs (i.e. the outer 2 vertical lines) are chosen to exclude excessive curviness in the tails. See Figure 1 for details.
- Bootstrap α estimates by sampling $x_{\min}$ and $x_{\max}$ uniformly from the intervals described above. In each iteration, use those values to truncate the data and then compute $\hat{\alpha}$. We compile a distribution of α estimates from 5000 bootstrap iterations for each corpus, and those distributions are shown in Figure 2.
- The bootstrapped distributions of $\hat{\alpha}$ are used to test the hypothesis that the parameter estimates do not differ significantly from one another, as shown in Figures 2 and 3.

3.2.3 Accommodations for Count Data

Methods such as the one by Clauset *et al.* (2009), assume that the data being used are a random sample. However, here, the data provided are not a random sample, but are *count data*.

To clarify the difference between a random sample and count data, consider a random sample from a discrete power law distribution, which consists of a set of integers: many 1s and other small integers and few big integers. One can count the occurrence of each integer in the sample to get count data. We define this count of 1s to be c_1, and so on. If we do this for all integers *i* in the sample from 1 to $x_{\max}$, we get a set of count data $\{c_i\}$. Note that we can reconstruct the sample from the count data.

To use the estimation and uncertainty quantification methods described above, we must compute the log-likelihood value from these counts c_i instead of sample values x_i as follows:

$$LL(\alpha) = -\alpha\left(\sum_{i=1}^{x_{\max}} c_i(\ln i)\right) - n\ln\zeta(\alpha, x_{\min}, x_{\max}), \quad (9)$$

$$\text{where } n = \sum_{i=1}^{x_{\max}} c_i. \qquad (10)$$

4 Results

The results of the Zipf's Law parameter (α) estimation and bootstrapped uncertainty analysis are given in Table 2 and Figures 1, 2, and 3. Table 2 summarizes the bootstrap results following the method outlined above for estimating α in the eight different corpora considered. For each bootstrap iteration, the upper and lower cutoff values ($x_{\min}$ and $x_{\max}$) were chosen by uniformly sampling from the corresponding intervals shown in the right two columns, and the estimation of α was done via MLE (see

Table 2: Summary of bootstrap results for estimating the exponent α in the eight different corpora. For each bootstrap iteration, the upper and lower cutoff values ($x_{\min}$ and $x_{\max}$) were chosen by uniformly sampling from the corresponding intervals shown in the right two columns, and the estimation of α was done via MLE (see Methods for details). The left two columns show the median and mean point estimates for α from 5000 bootstrap iterations for each case.

Corpus	Median $\hat{\alpha}$	Mean $\hat{\alpha}$	Range of $x_{\min}$	Range of $x_{\max}$
1800s	1.054	1.064	[10,150]	[1400,2200]
1900-49	1.048	1.056	[10,200]	[1400,2200]
1950-89	1.012	1.020	[30,250]	[1200,2400]
Spoken	1.155	1.159	[32,200]	[1800,2800]
Fiction	1.110	1.109	[20,200]	[1600,3000]
Magazine	0.9319	0.9364	[30,150]	[1600,2400]
Newspaper	0.9286	0.9433	[40,200]	[1400,2200]
Academic	0.8594	0.8586	[32,100]	[7200,2200]

Methods for details). The left two columns show the median and mean point estimates for α from 5000 bootstrap iterations for each case. The median estimates range from 0.8594 (Academic corpus) to 1.155 (Spoken corpus), and the mean estimates are very similar. Note that the median α ˆare also reported in Figures 1 and 2.

Figure 1 shows the Zipf's Law fit for each of the eight corpora using the median α estimate from 5000 bootstrap iterations. In each plot, the dark gray circles represent individual data points and the red line is the model fit. The graphs are all plotted on a log-log scale with rank on the x-axis and word frequency on the y-axis. The vertical lines in each graph show the bounds for the $x_{\min}$ and $x_{\max}$ intervals used in the analysis of each corpus (see Table 2 for exact values and see Methods for details on the analysis). The median $\hat{\alpha}$ value is reported at the top of each graph. We compute both the median $\hat{\alpha}$ and the mean $\hat{\alpha}$ (see Table 2), but we focus on the median because it is more robust to outliers (in the bootstrapped distributions) than the mean.

The degree of similarity (or difference) between parameter estimates for different corpora can be inferred from the degree of overlap among bootstrap distributions of $\hat{\alpha}$. A standard method for comparison would be to compute confidence intervals (CIs) and look for overlapping CIs across corpora, but looking for overlap in the bootstrap distributions plotted on the same graph does exactly this. We did interpret the degree of overlap as follows: if two distributions do not overlap at all, we infer that the α values are significantly different; if two distributions overlap, we visually inspect the degree of overlap to infer whether or not the parameter estimates differ.

Figure 2 summarizes the results of the bootstrap distributions of $\hat{\alpha}$ generated via 5000 bootstrap iterations for each of the eight corpora. The solid black vertical line is the median α estimate and the dashed line is the mean α estimate. As in Figure 1, the median $\hat{\alpha}$ value is reported at the top of each figure. The top three graphs on the left-hand-side correspond to the time-period corpora, and we see that the median $\hat{\alpha}$ values are very close, and further, that these distributions overlap significantly. Figure 3 (top panel) shows a comparison plot among the time-period corpora with the distributions plotted on the same graph. The significant overlap suggests that these corpora do not differ significantly in their parameter estimates.

The remaining graphs in Figure 2 correspond to the media-form corpora, and in this case, we see a clear distinction between the media-form corpora. In particular, the bootstrap distributions for Fiction, Academic, and Spoken essentially do not overlap at all with any other media-form distribution, but Newspaper and Magazine look very similar and almost completely overlap. See Figure 3 (middle panel) for a comparison plot among the media-form corpora with the distributions plotted on the same graph. These results suggest that the α estimates for Fiction, Academic, and Spoken are significantly different from one another and from the Newspaper/Magazine pair, and that the estimates for Newspaper and Magazine are not significantly different.

To further illustrate the relationship among these $\hat{\alpha}$ distributions, Figure 3 (bottom panel) is a summary plot that shows the bootstrap distributions for all eight corpora on the same graph. Here, we see that the three time-period corpora overlap with Fiction and both Newspaper and Magazines, whereas Academic and Spoken appear almost as outliers on opposite ends of the range of α estimates shown.

Lastly, we note that using the truncated version of Zipf's Law is necessary to correctly estimate α in each of the corpora. Including the most frequent words has the effect of underestimating α due to the non-linear curve of those top ranked data points (we ran the above analysis by setting $x_{\min} = 1$ to verify that this is the case; results not shown). Including the tail of the distribution by setting $x_{\max}$ to the largest value in the sample has the effect of overestimating α due to the downward curve in the tail (see Figure 1).

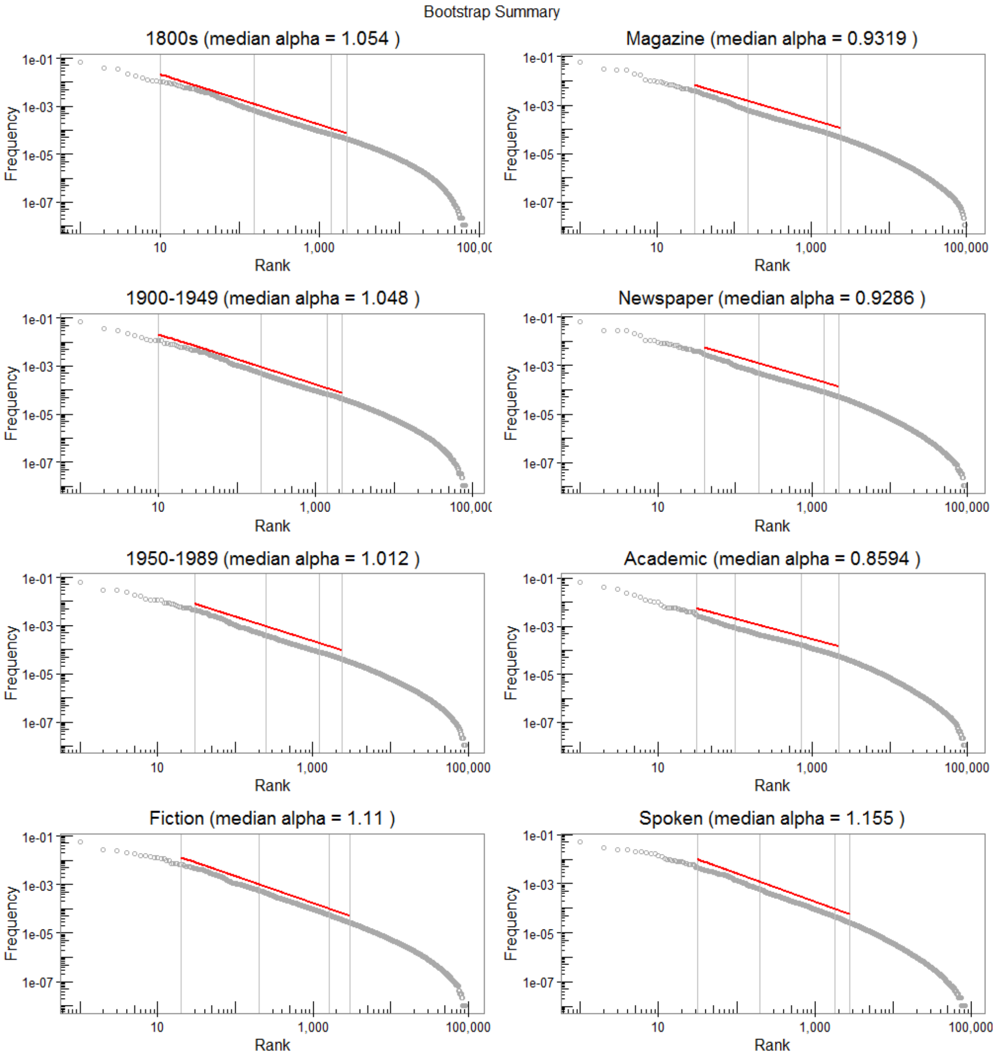


Figure 1: Bootstrap summary of Zipf's Law fits for the eight corpora. Dark gray circles represent individual data points. Red line is the fitted model with slope equal to the median α estimate from 5000 bootstrap iterations (also reported at the top of each graph). The graphs are all plotted on a log-log scale with rank on the x-axis and word frequency (counts of words divided by total sum of word counts) on the y-axis. Vertical lines show the bounds for the $x_{\min}$ and $x_{\max}$ intervals used in the analysis of each corpus (see Table 2 for exact values).

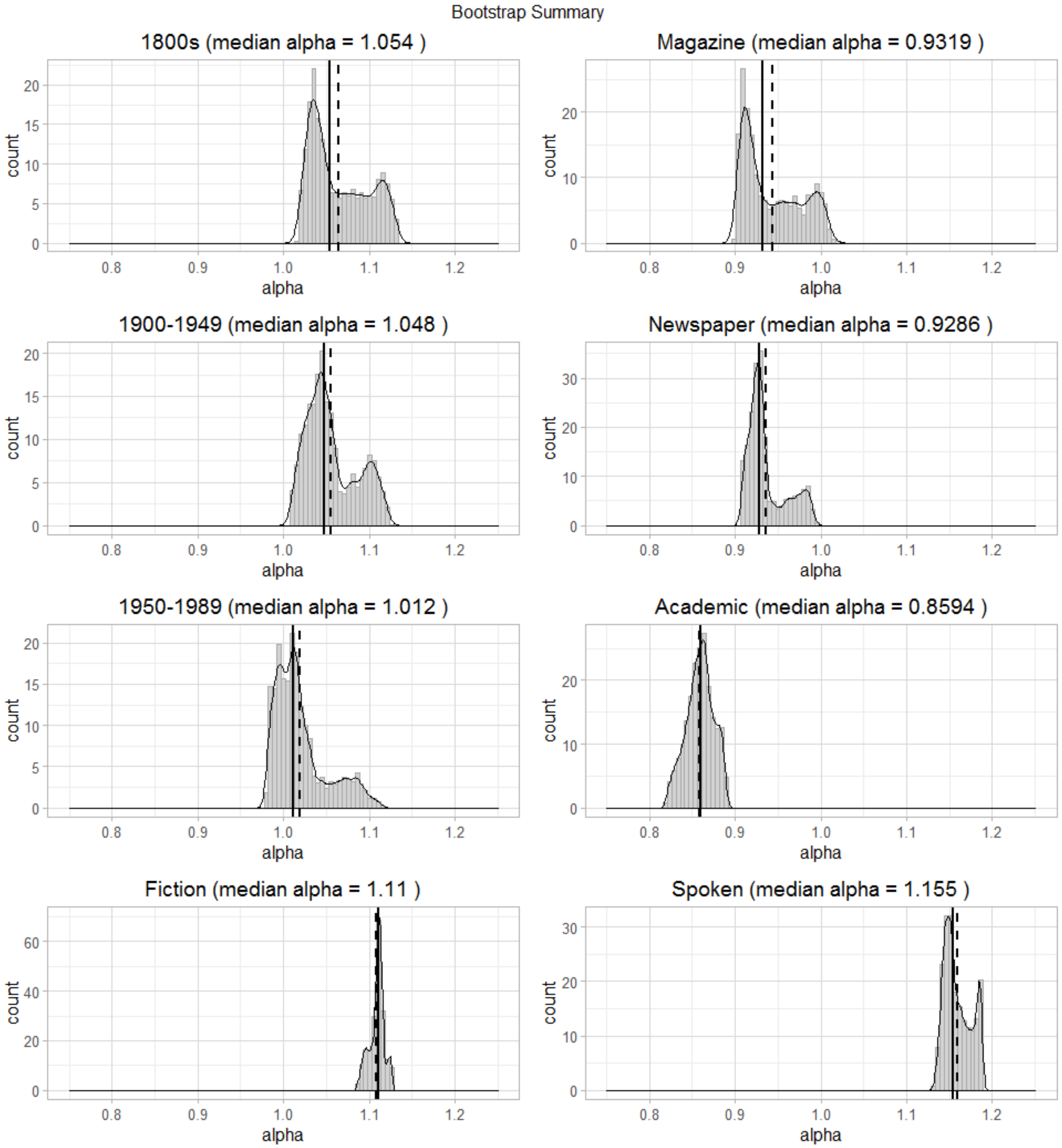


Figure 2: Summary of bootstrap distributions generated from 5000 bootstrap iterations for the each of the eight corpora. The solid black vertical line is the median α estimate and the dashed line is the mean α estimate. The median α estimate is reported at the top of each figure. The black line tracing the histogram bars is the smoothed density function (scaled by the sample size) used to compare to the histogram counts. The time-period corpora are shown in the top three graphs in the left panel. The remaining graphs correspond to the media-form corpora. Note that the distributions for the Fiction, Academic, and Spoken corpora do not overlap significantly with any other distribution. Newspaper and Magazine look very similar, as do the three time-period distributions.

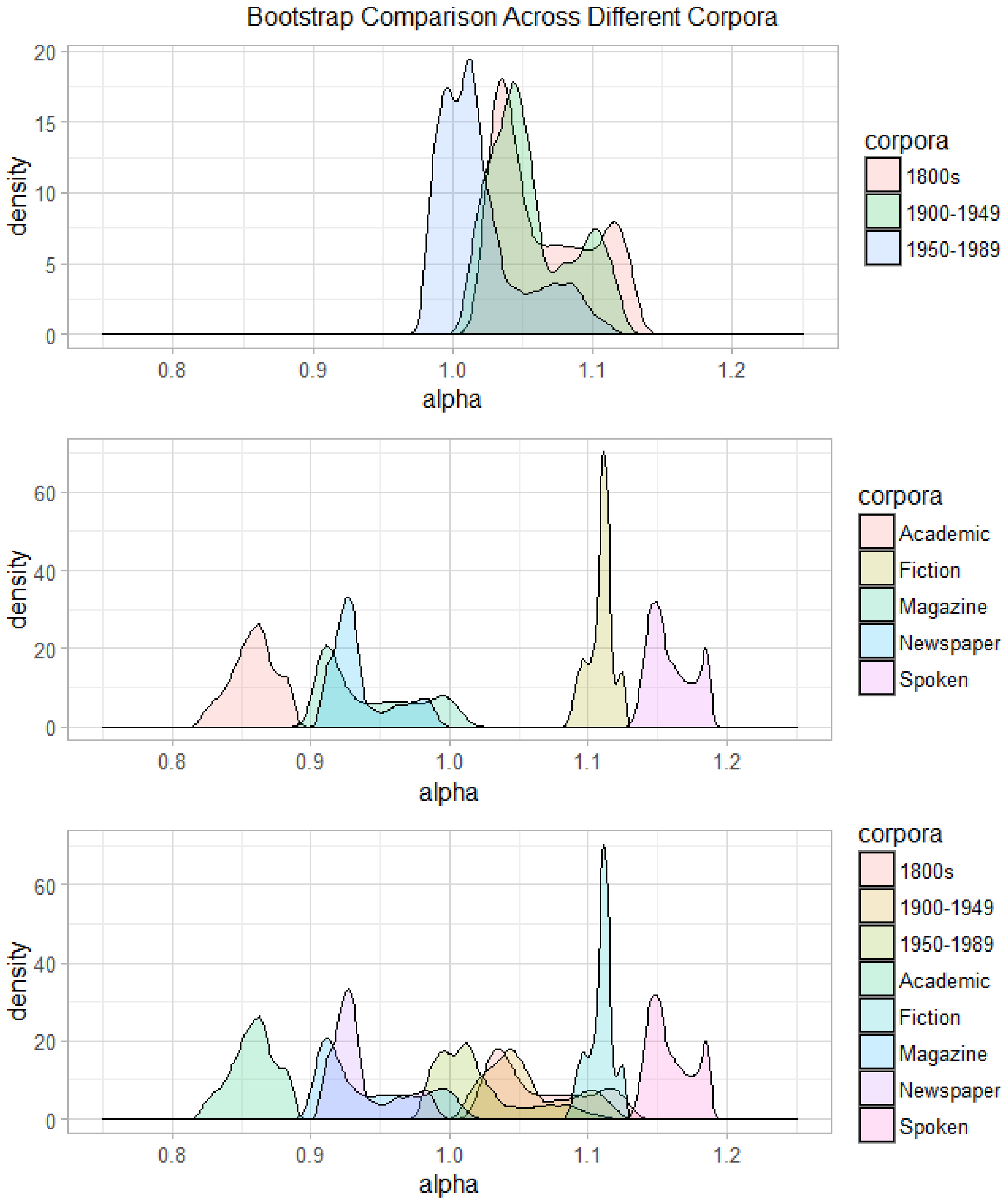


Figure 3: Comparison of bootstrap distributions for α estimates across different corpora (same distributions as shown in Figure 2 but organized by category to illustrate how much the distributions overlap). Top panel: time-period data. Middle panel: media-form data. Bottom panel: all eight corpora. The time-period distributions overlap substantially, whereas the media-form distributions are spread out in comparison.

Discussion

The results of our analysis show that there are substantial differences between large corpora from the same language but from different contexts, and therefore, we conclude that the null hypothesis should be rejected. We find that there are significant differences between the Zipf's Law parameter estimates for the media-form data (but not for the time-period data). This is illustrated nicely in Figure 3, showing the largely non-overlapping spread of the bootstrap distributions for α estimates among the time-period corpora. For three corpora (Academic, Fiction, and Spoken), the corresponding bootstrap distributions do not overlap with any other media-form distribution. Fiction, and to a small extent Spoken, overlap with the three time-period corpora bootstrap distributions. However, by comparison, the time-period data overlap significantly with each other which suggests that their parameter estimates do not differ significantly within their group.

There are several possible directions for future research from here. First, it is clear from looking at any of the log-log plots of the data that Zipf's Law – which, on a log-log scale, is a straight line – is only useful for modelling the behavior of the intermediate frequency words of these corpora (i.e. the truly linear region of the rank-frequency plot). Since it was necessary to truncate the data to get useful inference, a further examination of how to choose upper and lower bounds would be one direction for future study.

Another future direction is to determine a model for the remainder of the data (i.e. data in the upper and lower tails of the distribution representing the most and least frequent words) using something other than Zipf's Law. Putting such a model together with the Zipf's Law fit in a piecewise graph would form a more accurate, complete model for the data.

Additionally, the manner in which the COCA and the COHA data was collected should be further investigated. Being able to source which texts the frequency data was taken from is important to the integrity of the model. For example, in the time-period data for the 1800s, suppose that there is a word which incidentally is only found in newspapers or non-fiction books from that era, but not magazines or fiction novels. Then there is a need to account for how the word frequency is affected by the media-form variable or not.

Lastly, the goodness of fit of the model to the data should be more closely examined. Information such as a Kolmogorov-Smirnoff statistic (Chakravarti 1967) for each corpus would be relevant because it would offer inference on how well the model fits the data. That could also be used to compare the quality of model fits for each corpus to each other.

Acknowledgements

The authors would like to thank two anonymous reviewers who helped to improve the paper, as well as Paul Hurtado for helpful discussions about the method implemented in the paper.

References

Baayen, R. (2001). *Word Frequency Distributions* (Vol. 1). Berlin: Springer.

Baroni, M. (2009). Distributions in Text. In: Ludeling, A. & Kyto, M., editors. Corpus Linguistics: An International Handbook, Vol 2. Mouton de Gruyter, Berlin, 803–821.

Bauke, H. (2007). *Parameter Estimation for Power-Law Distributions by Maximum Likelihood Methods.* European Physical Journal B, 58, 167–173.

Calude, A. S., & Pagel, M. (2011). *How do we Use Language? Shared Patterns in the Frequency of Word Use across 17 World Languages.* Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), 1101–1107.

Chakravarti, I., Laha, R., & Roy, J. (1967). Handbook of Methods of Applied Statistics, volume 1. John Wiley and Sons.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009) *Power-law distributions in empirical data*. SIAM Review, 51(4): 661–703.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Davies, M. (2010). *The Corpus of Historical American English: 400 million words, 1810-2009.* Available online at http://corpus.byu.edu/coha/.

Fries, C. C. (1952). *The Structure of English*. New York: Harcourt Brace.

Gelbukh, A. & Sidorov, G. (2001). Zipf and Heaps Laws' Coefficients Depend on Language. In D. Hutchison (eds.), *Lecture Notes in Computer Science* (vol. 2004) (332-335). Berlin: Springer.

Gillespie, C. S. (2015). *Fitting Heavy Tailed Distributions: The poweRlaw Package*. Journal of Statistical Software, 64(2), 1-16. URL http://www.jstatsoft.org/v64/i02/.

Langlois, D., Cousineau, D. & Thivierge, J. P. (2014). *Maximum likelihood estimators for truncated and censored power-law distributions show how neuronal avalanches may be misevaluated.* Physical Review E, 89, 012709.

Larsen, R. J. & Marx, M. L. (2006). *An Introduction to Mathematical Statistics and Its Applications* (4th ed.). Upper Saddle River: Pearson Prentice Hall.

Lu, L., Zhang, Z. & Zhou, T. (2013). Deviation of Zipf's and Heaps' Laws in Human Languages with Limited Dictionary Sizes. Sci. Rep. 3, 1082; DOI: 10.1038/srep01082.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational linguistics, 19(2), 313–330.

Moreno-Sanchez, I., Font-Clos, F. & Corral, A. (2016). *Large-Scale Analysis of Zipf's Law in English Texts.* PLoS ONE, 11(1): e0147073.

Newman, M. E. J. (2005). *Power laws, Pareto distributions, and Zipf's law.* Contemporary Physics, 46, 323–351.

Piantadosi, S.T. (2014). *Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions*. Psychometric Bulletin and Review, 21(5), 1112–1130.

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Zipf, G. (1936). *The Psychobiology of Language*. London: Routledge.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.