Zipf's word frequency law in natural language: a critical review and future directions

Steven T. Piantadosi

Abstract

The frequency distribution of words has been a key object of study in statistical linguistics for the past 70 years. This distribution approximately follows a simple mathematical form known as *Zipf's law*. This paper first shows that human language has highly complex, reliable structure in the frequency distribution over and above this classic law, though prior data visualization methods obscured this fact. A number of empirical phenomena related to word frequencies are then reviewed. These facts are chosen to be informative about the mechanisms giving rise to Zipf's law, and are then used to evaluate many of the theoretical explanations of Zipf's law in language. No prior account straightforwardly explains all the basic facts, nor is supported with independent evaluation of its underlying assumptions. To make progress at understanding why language obeys Zipf's law, studies must seek evidence beyond the law itself, testing assumptions and evaluating novel predictions with new, independent data.

1 Introduction

One of the most puzzling facts about human language is also one of the most basic: words occur according to a famously systematic frequency distribution such that there are few very high frequency words that account for most of the tokens in text (e.g. "a", "the", "I", etc.), and many low frequency words (e.g. "accordion", "catamaran", "ravioli"). What is striking is that the distribution is mathematically simple, roughly obeying a power law known as *Zipf's law*: the rth most frequent word has a frequency $f(r)$ that scales according to

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

for $\alpha \approx 1$ (Zipf, 1936, 1949)[1]. In this equation, r is called the "frequency rank" of a word, and $f(r)$ is its frequency in a natural corpus. Since the actual observed frequency will depend on the size of the corpus examined, this law states frequencies proportionally: the most frequent word ($r = 1$) has a frequency proportional to 1, the second most frequent word ($r = 2$) has a frequency proportional to $\frac{1}{2^\alpha}$, the third most frequent word has a frequency proportional to $\frac{1}{3^\alpha}$, etc.

Mandelbrot proposed and derived a generalization of this law that more closely fits the frequency distribution in language by "shifting" the rank by an amount β (Mandelbrot, 1962, 1953):

$$f(r) \propto \frac{1}{(r+\beta)^\alpha} \tag{2}$$

for $\alpha \approx 1$ and $\beta \approx 2.7$ (Zipf, 1936, 1949; Mandelbrot, 1962, 1953). This paper will study (2) as the current incarnation of "Zipf's law," although we will use the term "near-Zipfian" more broadly to mean frequency distributions where this law at least approximately holds. Such distributions are observed universally in languages, even in extinct and yet-untranslated languages like Meroitic (R. D. Smith, 2008).

It is worth reflecting on peculiarity of this law. It is certainly a nontrivial property of human language that words vary in frequency at all—it might have been reasonable to expect that all words should be about equally frequent. But given that words do vary in frequency, it is unclear why words should follow such

[1]Note that this distribution is phrased over frequency ranks because the support of the distribution is an unordered, discrete set (i.e. words). This contrasts with, for instance, a Gaussian which is defined over a complete, totally-ordered field ($\mathbb{R}$), and so has a more naturally visualized probability density.

a precise mathematical rule—in particular one that does not reference any aspect of each word's meaning. Speakers generate speech by needing to communicate a meaning in a given world or social context; their utterances obey much more complex systems of syntactic, lexical, and semantic regularity. How could it be that the intricate processes of normal human language production conspire to result in a frequency distribution that is so mathematically simple—perhaps "unreasonably" so (Wigner, 1960)?

This question has been a central concern of statistical language theories for the past 70 years. Derivations of Zipf's law from more basic assumptions are numerous, both in language and in the many other areas of science where this law occurs (for overviews, see Mitzenmacher, 2004; Newman, 2005; Farmer & Geanakoplos, 2006; Saichev, Malevergne, & Sornette, 2010). Explanations for the distribution across the sciences span many formal ideas, frameworks and sets of assumptions. To give a brief picture of the range of explanations that have been worked out, such distributions have been argued to arise from random concatenative processes (Miller, 1957; W. Li, 1992; Conrad & Mitzenmacher, 2004), mixtures of exponential distributions (Farmer & Geanakoplos, 2006), scale-invariance (Chater & Brown, 1999), (bounded) optimization of entropy (Mandelbrot, 1953) or Fisher information (Hernando, Puigdomènech, Villuendas, Vesperinas, & Plastino, 2009), the invariance of such power laws under aggregation (see Farmer & Geanakoplos, 2006), multiplicative stochastic processes (see Mitzenmacher, 2004), preferential re-use (Yule, 1944; Simon, 1955), symbolic descriptions of complex stochastic systems (Corominas-Murtra & Solé, 2010), random walks on logarithmic scales (Kawamura & Hatano, 2002), semantic organization (Guiraud, 1968; D. Manin, 2008), communicative optimization (Zipf, 1936, 1949; Mandelbrot, 1962; Ferrer i Cancho & Solé, 2003; Ferrer i Cancho, 2005a; i Cancho, 2005; Salge, Ay, Polani, & Prokopenko, 2013), random division of elements into groups (Baek, Bernhardsson, & Minnhagen, 2011), first- and second-order approximation of most common (e.g. normal) distributions (Belevitch, 1959), optimized memory search (Parker-Rhodes & Joyce, 1956), among many others.

For language in particular, any such account of the Zipf's law provides a psychological theory about what must be occurring in the minds of language users. Is there a multiplicative stochastic process at play? Communicative optimization? Preferential re-use of certain forms? In the face of such a profusion of theories, the question quickly becomes *which*—if any—of the proposed mechanisms provides a true psychological account of the law. This means an account which is connected to independently testable phenomena and mechanisms, and fits with the psychological processes of word production and language use.

Unfortunately, essentially all of the work in language research has focused on solely deriving the law itself in principle; very little work has attempted to assess the underlying assumptions of the hypothesized explanation, a problem for much work on power laws in science (Stumpf & Porter, 2012)[2]. It should be clear why this is problematic: the law itself can be derived from many starting points. Therefore, the ability of a theory to derive the law provides very weak evidence for that account's cognitive validity. Other evidence is needed.

This paper reviews a wide range of phenomena any theory of word frequency distributions and Zipf's law must be able to handle. The hope is that a review of facts about word frequencies will push theorizing about Zipf's law to address a broader range of empirical phenomena. This review intentionally steers clear from other statistical facts about text (e.g. Heap's law, etc.) because these are throughly reviewed in other work (see Baayen, 2001; Popescu, 2009). Instead, we focus here specifically on facts about word frequencies which are informative about the mechanisms giving rise to Zipf's law[3].

We begin first, however, by pointing out an important feature of the law: it is not as simple as Zipf and other since have suggested. Indeed, some of the simplicity of the relationship between word frequency and frequency rank is the result of a statistical sin that is pervasive in the literature. In particular, the plots which motivate equation (2) almost always have unaddressed, correlated errors, leading them to look simpler than they should. When this is corrected, the complexities of the word frequency distribution become more apparent. This point is important because it means that (2) is at best a good approximation to what is demonstrably a much more complicated distribution of word frequencies. This complication means that

[2]As they write, "Finally, and perhaps most importantly, even if the statistics of a purported power law have been done correctly, there is a theory that underlies its generative process, and there is ample and uncontroversial empirical support for it, a critical question remains: What genuinely new insights have been gained by having found a robust, mechanistically supported, and in-all-other-ways superb power law? We believe that such insights are very rare."

[3]Importantly, however, other statistical properties are also likely informative, as a "full" theory of word frequencies would be able to explain a wide range of empirical phenomena.

detailed statistical analysis of what particular form the word frequency distribution takes (e.g. (1) vs (2) vs lognormal distributions, etc.) will not be fruitful: none is strictly "right."

Following those results, this paper presents and reviews a number of other facts about word frequencies. Each fact about word frequencies is studied because of its relevance to a proposed psychological account of Zipf's law. Most strikingly, Section 3.7 provides experimental evidence that near-Zipfian word frequency distributions occur for novel words in a language production task. Section 4 then reviews a number of formal models seeking to explain Zipf's law in language, and relates each proposed account to the empirical phenomena discussed in Section 3.

2 The word frequency distribution is complex

Quite reasonably, a large body of work has sought to examine what form most precisely fits the word frequency distribution observed in natural language. Zipf's original suggestion of (1) was improved by Mandelbrot to that in (2), but many other forms have been suggested including for instance, a log-normal distribution (Carroll, 1967, 1969), which might be considered a reasonably "null" (e.g. unremarkable) hypothesis.

A superb reference for comparing distributions is Baayen (2001, Chapter 3), who reviews evidence for and against a log-normal distribution (Carroll, 1967, 1969), a generalized inverse Gauss-Poisson model (Sichel, 1975), and a generalized Z-distribution (Orlov & Chitashvili, 1983) for which many other models (due to, e.g., Yule, 1924; Simon, 1955, 1960; Herdan, 1960, 1964; Rouault, 1978; Mandelbrot, 1962) are a special case (see also Montemurro, 2001; Popescu, 2009). Baayen finds with a quantitative model comparison that which model is best depends on which corpus is examined. For instance, the log-normal model is best for the text, *The Hound of the Baskervilles*, but the Yule-Simon model is best for *Alice in Wonderland*. One plausible explanation for this is that none of these simple models—including the Zipf-Mandelbrot law in Equation (2)—is "right,"[4] instead only capturing some aspects of the full distribution of word frequencies.

Indeed, none is right. The apparent simplicity of the distribution is an artifact of how the distribution is plotted. The standard method for visualizing the word frequency distribution is to count how often each word occurs in a corpus, and sort the word frequency counts by decreasing magnitude. The frequency $f(r)$ of the r'th most frequent word is then plotted against the frequency rank r, yielding typically a mostly linear curve on a log-log plot (Zipf, 1936), corresponding to roughly a power law distribution[5]. This approach—though essentially universal since Zipf—commits a serious error of data visualization. In estimating the frequency-rank relationship this way, the frequency $f(r)$ and frequency rank r of a word are estimated on the same corpus, leading to correlated errors between the x-location r and y-location $f(r)$ of points in the plot.

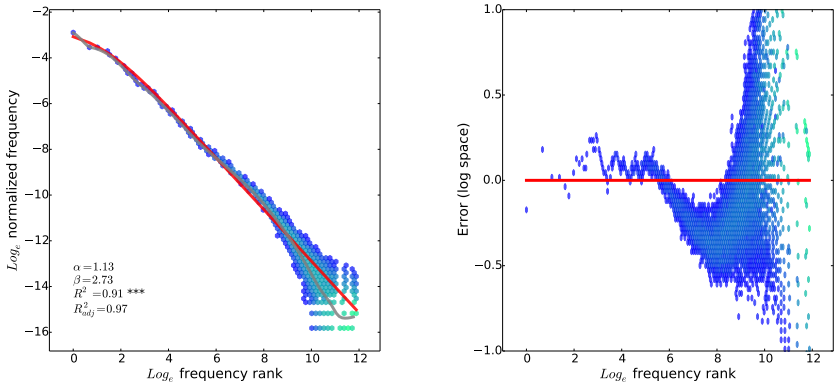
This is problematic because it may suggest spurious regularity[6]. The problem can be best understood by a simple example. Imagine that all words in language were actually equally probable. In any sample (corpus) of words, we will find that some words occur more than others just by chance. When plotted in the standard manner, we will find a strikingly decreasing plot, erroneously suggesting that the true frequency-rank relationship has some interesting structure to be explained. This spurious structure is especially problematic for low frequency words, whose frequencies are measured least precisely. Additionally, in the standard plot, deviations from the Zipfian curve are difficult to interpret due to the correlation of measurement errors: it is hard to tell systematic deviations from noise.

Fortunately, the problem is easily fixed: we may use two independent corpora to estimate the frequency and frequency rank. In the above case where all words are equally probable, use of independent corpora will lead to no apparent structure—just a roughly flat frequency-rank relationship. In general, we need not have two independent corpora from the start: we can imagine splitting our initial corpus into two subcorpora before any text processing takes place. This creates two corpora which are independent bodies of text (conditioned on the general properties of the starting corpus), and so from which we can independently estimate r and $f(r)$. A convenient technique to perform this split is to perform a binomial split on observed frequency of

[4]See Ferrer-i-Cancho and Servedio (2005) for related arguments based on the range of Zipfian exponents.

[5]Since linearity on a log-log plot means that $\log f = a \log r + b$, so $f = e^b r^a \propto r^a$.

[6]Such estimation *also* violates the assumptions of typical algorithms used to fit Zipfian exponents since most fitting algorithms assume that x is known perfectly and only y is measured with error. This concern applies in principle to maximum-likelihood estimation, least squares (on log-log values), and any other technique that places all of measurement error on frequencies, rather than both frequencies and frequency ranks.



(a) (b)

Figure 1: 1(a) shows the relationship between frequency rank (x-axis) and (normalized) frequency (y-axis) for words from the American National Corpus. This is plotted using a two-dimensional hexagonal histogram. Bins are shaded blue to green along a logarithmic scale depending on how many words fall into the bin. The red line shows the fit of (2) to this data. 1(b) shows frequency rank versus the difference (in log space) between a word's frequency and the prediction of (2). This figure shows only a subset of the full y range, cropping some extreme outliers on the right hand side of the plot in order to better visualize this error for the high frequency words.

each word: if we observe a word, say, 100 times, we may sample from a binomial ($N = 100, p = 0.5$) and arrive at a frequency of, say, 62 used to estimate its true frequency, and a frequency of $N - 62 = 38$ to estimate its true frequency rank. This exactly mirrors if we had randomly put tokens of each word into two independent corpora, before any text processing began. The choice of $p = 0.5$ is not necessary, but yields two corpora of approximately the same size. With this method, the deviations from a fit are interpretable and our plotting method no longer introduces any erroneous structure.

Figure 1(a) shows such a plot, giving the frequency/frequency-rank relationship from the American National Corpus (Reppen & Ide, 2004), a freely available collection of written American English. All figures in this paper follow this plotting procedure, unless otherwise noted. The plot shows a two-dimensional histogram of where words fall in frequency/frequency-rank space[7]. The shading of the histogram is done logarithmically with the the number of words falling into each hexagonal bin, and is white for zero-count bins. Because the plot has a logarithmic y-axis, words with zero frequency after the split are not shown. The fit of (2) using a maximum-likelihood method on the separate frequency and frequency rank portions of the corpus is shown in the red solid line. Additionally, a locally-smoothed regression line (LOESS) (Cleveland, Grosse, & Shyu, 1992) is shown in gray. This line corresponds to a local estimate of the mean value of the data, and is presented as a comparison point to see how well the fit of (2) matches the expected value of the points for each frequency rank (x-value). In the corner several key values are reported: the fit α and β, an R^2 measure giving the amount of variance explained by the red line fit, and an adjusted R^2_{adj} capturing the proportion of *explainable* variance captured by the fit, taking the smoothed regression as an estimate of the maximum amount of variance explainable. For simplicity, statistics are computed only on the original R^2, and its significance is shown with standard star notation (three starts means $p < 0.001$).

This plot makes explicit several important properties of the distribution. First, it *is* approximately linear on a log-log plot, meaning the word frequency distribution is approximately power law, and moreover is fit very well by (2) according to the correlation measures. This plot shows higher variability towards the low frequency end, (accurately) indicating that we cannot estimate the curve reliably for low frequency words. While the scatter of points is no longer monotonic, note that the true plot relating frequency to frequency rank *must* be monotonic by definition. Thus, one might imagine estimating the true curve by drawing any monotonic curve through this data. At the low frequency end we have more noise and so greater uncertainty

[7]In these plots, tied ranks are not allowed, so words of the same frequency are arbitrarily ordered.

about the shape of that curve. This plot also shows that equation (2) provides a fairly accurate fit (red) to the overall structure of the frequency-rank relationship across both corpora.

Importantly, because we have estimated r and $f(r)$ in a statistically independent way, deviations from the curve can be interpreted. Figure 1(b) shows a plot of these deviations, corresponding to the residuals of frequency once (2) is fit to the data. Note that if the true generating process were something like (2), the residuals should be only noise, meaning that which are above and below the fit line ($y = 0$ in the residual plot) should be determined entirely by chance. There should be no observable structure to the residual plot. Instead, what Figure 1(b) reveals is that there is considerable structure to the word frequency distribution beyond the fit of the Zipf-Mandelbrot equation, including numerous minima and maxima in the error of this fit. This is most apparent by the "scoop" on the right hand size of the plot, corresponding to mis-estimation higher ranked (lower-frequency) words. This type of deviation has been observed previously with other plotting methods and modeled as a distinct power law exponent by Ferrer i Cancho and Solé (2001), among others.

However, what is more striking, is the systematic deviation observed in the *left* half of this plot, corresponding to low rank (high frequency) words. Even the most frequent words do not exactly follow Zipf's law. Instead, there is a substantial autocorrelation, corresponding to the many local minima and maxima ("wiggles") in the left half of this plot. This indicates that there are further statistical regularities—apparently quite complex—that are not captured by (2). These autocorrelations in the errors are statistically significant using the Ljung-Box Q-test (Ljung & Box, 1978) for residual autocorrelation ($Q = 126810.1, p < 0.001$), even for the most highly-ranked twenty-five ($Q = 5.7, p = 0.02$), fifty ($Q = 16.7, p < 0.001$), or hundred ($Q = 39.8, p < 0.001$) words examined.

Such complex structure should have been expected: *of course* the numerous influences on language production result in a distribution that is complex and structured. However, the complexity is not apparent in standard ways of plotting power laws. Such complexity is probably incompatible with attempts to characterize the distribution with a simple parametric law, since it is unlikely a simple equation could fit all of the minima and maxima observed in this plot. At the same time, almost all of the variance in frequencies is fit very well by a simple law like Zipf's power law or its close relatives. A simple relationship captures a considerable amount about word frequencies, but clearly will not explain everything. The distribution in language is only *near*-Zipfian.

3 Empirical phenomena in word frequencies

Having established that the distribution of word frequencies is more complex than previously supposed, we now review several basic facts about word frequencies which any theory of the Zipfian or near-Zipfian distribution must account for. The plan of this paper is to present these empirical phenomena in this section, and then use them to frame specific model-based accounts of Zipf's law in Section 4. As we will see, the properties of word frequencies reviewed in this section will have much to say about the most plausible accounts of the word frequency distribution in general.

The general method followed in this section is to study relevant subsets of the lexicon and quantify the fit of (2). This approach contrasts somewhat with the vast literature on statistical model comparison to check for power laws (as compared to, e.g., lognormal distributions, etc.). The reason for this is simple: Section 2 provides strong evidence that no simple law can be the full story behind word frequencies because of the complexities of the frequency rank / frequency curve. Therefore, comparisons between simple models will inevitably be between alternatives that are both "wrong".

In general, it is not so important which simple distributional form is a better approximation to human language. What matters more are the general properties of word frequencies that are informative about the *underlying mechanisms* behind the observed distribution. This section tries to bring out those general properties. Do the distributions appear near-Zipfian for systematic subsets of words? Are distributions that look similar to power laws common across word types, or are they restricted to word with certain syntactic or semantic features? Any psychologically-justified theory of the word frequency distribution will depend on appreciating, connecting-to, and explaining these types of high-level features of the lexical frequency distribution.

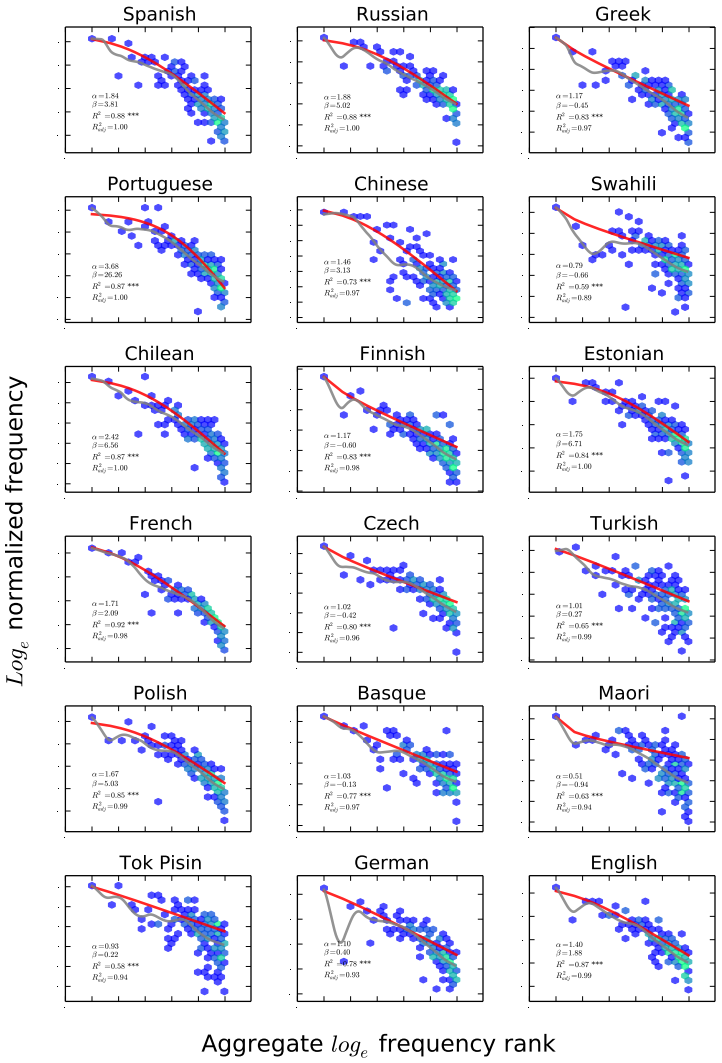
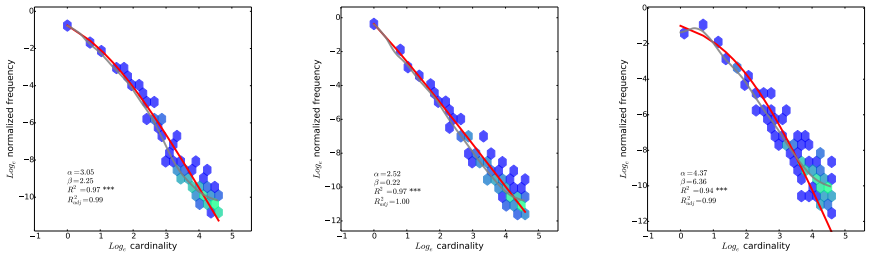


Figure 2: Cross-linguistic word frequency distributions using words from a Swadesh list (data provided by Calude and Pagel, 2011). Here, the x-location of each point (word) is fixed across languages according to the aggregate frequency rank of the word's meaning on an independent set of data. The systematicity here means that the word frequency distribution falls off similarly according to word meaning across languages, and approximately according to a power law like (2) (red).

3.1 Semantics strongly influences word frequency

As a language user, it certainly *seems* like we use words to convey an intended meaning. From this simple point of view, Zipf's law is really a fact about the "need" distribution for how often we need to communicate each meaning. Surprisingly, many accounts of the law make no reference to meaning and semantics (except see 4.3 and some work in 4.4), deriving it from principles independent of the content of language. But this view is incompatible with the fact that even cross-linguistically, meaning *is* systematically related to frequency. Calude and Pagel (2011) examined Swadesh lists from 17 languages representing six language families and compared frequencies of words on the list. Swadesh lists provides translations of simple, frequent words like "mother" across many languages; they are often used to do historical reconstruction. Calude and Pagel (2011) report an average inter-language correlation in log frequency of $R^2 = 0.53$ ($p < 0.0001$) for these



(a) (b) (c)

Figure 3: Power law frequencies for number words ("one", "two", "three", etc.) in English (a), Russian (b) and Italian (c) using data from Google (Lin et al., 2012). Note that here the x-axis is ordered by cardinality, *not* frequency rank, although these two coincide. Additionally, decades ("ten", "twenty", "thirty", etc.) were removed from this analysis due to unusually high frequency from their approximate usage. Here and in all plots the red line is the fit of (2) and the gray line is a LOESS.

common words, indicating that word frequencies are surprisingly robust across languages and predictable from their meanings. Importantly, note that Swadesh words will tend to be high-frequency, so the estimated R^2 is almost certain to be lower for less frequent words. In any case, if meaning has any influence on frequency, a satisfying account of the frequency distribution will have to address it.

We can also see systematic frequency-rank relationship across languages, grouping words by their meaning. Figure 2 shows frequency-rank plots of the Swadesh lists compiled in Calude and Pagel (2011)[8], plotted, like all other plots in the paper, according to the methods in Section 2. However, unlike other plots in this paper, the frequency rank here is fixed across all languages, estimated independently on 25% of data from each language and then collapsed across languages. Thus, the rank ordering—corresponding to the x-location of each meaning on the Swadesh list—does not vary by language and is determined only by aggregate, cross-linguistic frequency (independently estimated from the y-location). We can then compare the frequencies at each rank to see if they follow similar distributions. As these plots reveal, the distributions are extremely similar across languages, and follow a near-Zipfian distribution for the pooled rank ordering.

In this plot, because the rank ordering is fixed across all languages, not only do frequencies fall off like (Equation 2), but they do so as roughly with the same coefficients across cross-linguistically. If frequency was not systematically related to meaning these plots would reveal no such trends.

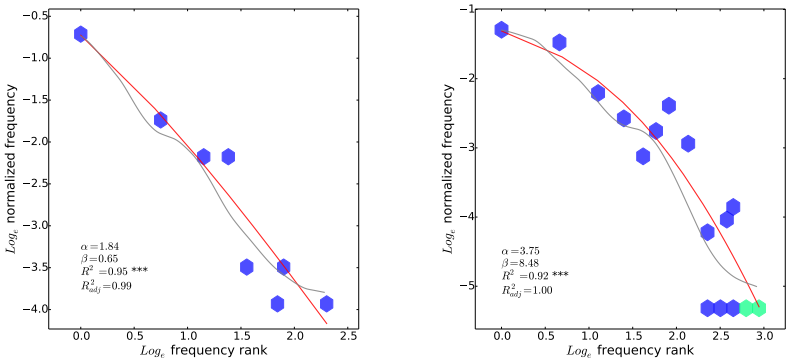
Another domain where the meaning-dependence of frequency is apparent is that of number words (Dehaene & Mehler, 1992; Piantadosi, 2016). Figure 3 shows number word frequencies (e.g. "one", "two", "three", etc.), previously reported in Piantadosi (2016). These plots show *cardinality* vs. frequency in English, Russian, and Italian, using all the data from the Google Books N-gram dataset (Lin et al., 2012). This clearly shows that across languages, number words follow a near-Zipfian distribution according to the magnitude (meaning)—in fact, a very particular one with exponent $\alpha \approx -2$ (the "inverse square law" for number frequency), a finding previously reported by Dehaene and Mehler (1992). Piantadosi (2016) shows that these trends also hold for the decade words, and across historical time. Thus, the frequency of these words is predictable from what cardinality the words refer to, even across languages.

The general point from this section is therefore that word meaning is a substantial determinant of frequency, and it is perhaps intuitively the best causal force in shaping frequency. "Happy" is more frequent than "disillusioned" because the meaning of the former occurs more commonly in topics people like to discuss. A psychologically-justified explanation of Zipf's law in language must be compatible with the powerful influence that meaning has on frequency.

3.2 Near-Zipfian distributions occur for fixed referential content

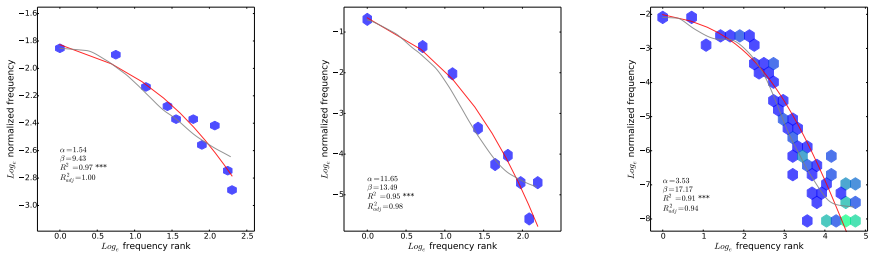
Given that meanings in part determine frequencies, it is important to ask if there are any phenomena which cannot be straightforwardly explained in terms of meaning. One place to look is words which have roughly

[8]We are extremely grateful to the authors for providing this data.



(a) (b)

Figure 4: Distributions for taboo words for (a) sex (gerunds) and (b) feces.



(a) (b) (c)

Figure 5: Frequency distribution in the ANC for words whose scope of meaning has been highly constrained by the natural world (a) months, (b) planets, (c) elements.

the same meaning, at least in terms of referential content. Facts like the *principle of contrast* (Clark, 1987) may mean that true synonyms do not exist in human language. However, *taboo words* provide a class of words which refer at least approximately to the same thing (e.g. "fornicating," "shagging," "fucking," etc.). Figure 4 shows the frequency distribution of several taboo words, gerunds referring to sex 4(a) and synonyms for feces 4(b)[9], plotted using the methods of Section 2 on data from the ANC. Both cases reveal that near-Zipfian word frequency distributions can still be observed for words that have a fixed referential content, meaning that other factors (e.g. formality, social constraints) also play a role in determining word frequency.

3.3 Near-Zipfian distributions occur for naturally constrained meanings

If meanings in part determine word frequencies, it is plausible that the distribution arises from how humans languages segment the observable world into labeled categories (see Section 4.3). For instance, languages are in some sense free to choose the range of referents for each word[10]: should "dog" refer to a specific kind of dog, or a broad class, or to animals in general? Perhaps language evolution's process for choosing the scope of word meanings gives rise to the frequency distribution (for a detailed account, see D. Manin, 2008).

However, the distribution follows a near-Zipfian distribution even in domains where the objects of reference are highly constrained by the natural world. Figure 5(a)-5(c) shows several of these domains[11] chosen

[9]More common taboo words meaning "penis" and "vagina" were not used since many of their euphemisms have salient alternative meanings (e.g. "cock" and "box").

[10]Although there are compelling regularities in at least some semantic domains—see, e.g., Kemp and Regier (2012); Kay and Regier (2003).

[11]In the elements, "lead" and "iron" were excluded due to their ambiguity, and thus frequent use as non-elements. In the months, "May" and "March" were removed for their alternative meanings.

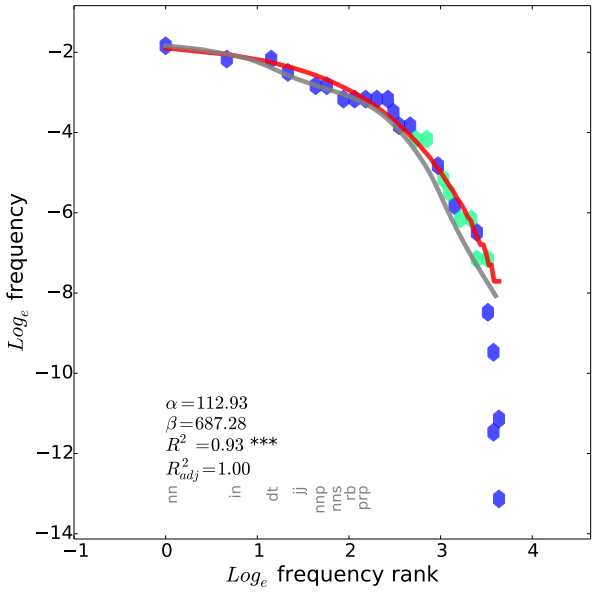


Figure 6: Frequency distribution of syntactic categories from the Penn Treebank.

a priori for their semantic fixedness: months, planets, and element names. Intuitively, in each of these cases, it is likely that the lexicon did not have much freedom in how it labeled the terms in these categories, since the referents of these terms are salient, fixed natural kinds. For instance, our division of the world into 12 months comes from phases of the moon and the seasons, not from a totally free choice that language may easily adapt or optimize. These plots all show close fits by (2), shown in red, and high, reliable correlations.

3.4 The fit of Zipfian distributions vary by category

Zipf's law is stated as a fact about the distribution of words, but it is important to remember that there may not be anything particularly special about analyzing language at the level of words. Indeed, words may not even be a precisely defined psychological class, with many idiomatic phrases stored together by language processing mechanisms, and some wordforms potentially created on the fly by grammatical or morphological mechanisms. It is therefore important to examine the frequency distribution for other levels of analysis.

Figure 6 shows the frequency distribution of various syntactic categories (part of speech tags on individual words) from the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993), using the tagged Brown corpus. This reveals that word categories are also fit nicely by (2)—perhaps even more closely than words—but the shape of the fit (parameters α and β) differs. The quality of fit appears to off for the lowest frequency tags, although it is not clear how much of this effect is due to data sparsity. The general pattern suggests that a full explanation of the word frequency distribution would ideally call on mechanisms general enough to apply to syntactic categories and possibly even other levels of analysis[12].

The same corpus can also be used to examine the fit and parameters *within* syntactic categories. Figure 7(a)-7(c) shows the distribution of words within each of six categories from the treebank: determiners (DT), prepositions / subordinating conjunctions (IN), modals (MD), singular or mass nouns (NN), past participle verbs (VBN), and 3rd person singular present tense verbs (VBZ). None of these were predicted to pattern in any specific way by any particular theory, but were chosen post-hoc as interesting examples of distributions. Determiners, modals, and some verbs appear to have the lowest adjusted correlations when (2) is fit. These figures illustrate that the word types vary substantially in the best-fitting parameters α and β, but show

[12]It is apparently unclear whether N-grams in text follow Zipf's law (see Egghe (1999, 2000); cf. Ha, Sicilia-Garcia, Ming, and Smith (2002); Ha, Hanna, Ming, and Smith (2009)).

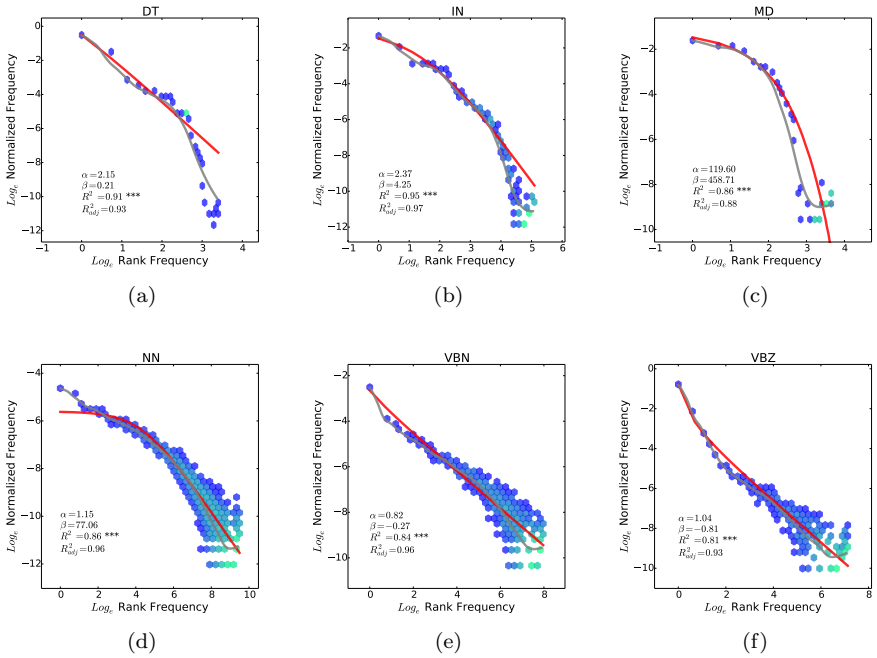


Figure 7: Frequency distribution of words within several syntactic categories from the Penn Treebank: determiners (DT), prepositions or subordinating conjunctions (IN), modals (MD), nouns (NN), past participle verbs (VBN), 3rd person singular present verbs (VBZ). These plots represent a post-hoc-selected subset of all syntactic categories.

in general fairly Zipfian distributions. Additionally, the residual structure (deviation from the red line fit) shows interesting variability between categories. For instance, the verbs (7(f)) show an interesting concavity that is the opposite of that observed in typical Zipfian distributions, bowing to the bottom rather than the top. This concavity is primarily driven by the much larger frequency of the first several words like "is", "has", and "does." These auxiliary verbs may in truth belong in a separate category than other verbs, perhaps changing the shape of this curve. There also appears to be a cluster of low-frequency modals of about all the same frequency. The determiner plot suggests that the rate at which frequency decreases with rank changes through two scaling regimes—a slow fall-off followed by a fast one—which is often argued for the lexicon in general (Ferrer i Cancho & Solé, 2001) and would be inconsistent with the simple fit of (2).

Overall, the variability across part of speech categories suggests that some of the fit of Zipfian distribution arises by collapsing together different parts of speech.

3.5 The distribution of word frequencies is not stationary

An often over-looked factor in the search for explanations of Zipf's law is that word frequencies are not stationary, meaning that the probability of uttering each word changes depending on other factors. This phenomenon occurs at, for instance, a long timescale reflecting the topic of discussion. One is more likely to utter "Dallas" in a discussion about Lyndon Johnson than a discussion about Carl Sagan. The non-stationarity of text is addressed by Baayen (2001, Chapter 5), who notes that the clumpy randomness of real text leads to difficulties estimating vocabulary sizes and distributions. Recently, Altmann, Pierrehumbert, and Motter (2009) showed that word recurrences on a timescale compatible with semantics (not syntax) follow a stretched exponential distribution, with a certain degree of "burstiness." The variability in frequencies is an important method of classification of documents via *topic models* (see Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2007; Blei & Lafferty, 2007, 2009) or *latent semantic analysis* (Landauer, Foltz, & Laham, 1998; Dumais, 2005). Such models work by essentially noting that word frequencies within a document are cues to its semantic topic; one can then work backwards from the frequencies to the topic or set of possible topics. The variability in word frequencies is also useful in information retrieval (Manning & Schütze, 1999, Chapter

15).

The non-stationarity of word frequencies has an important theoretical implication for explanations of Zipf's law. The frequencies of words we observe are actually *averages* over contexts: the probability of uttering a word w is given by

$$P(W = w) = \sum_c P(c)P(W = w|C = c) \tag{3}$$

where $P(W = w|C = w)$ is the probability of w in a particular context c. If the observed frequency is an average over contexts, then our explanation of Zipf's law must respect the fact that it is an average, and not explain it with a model that is incompatible with context-dependent frequencies.

3.6 Word frequency varies according to many forces

Thanks in large part to the recent availability of gigantic, freely-available, longitudinal corpora like Lin et al. (2012), recent studies have also been able to chart changes in word frequencies throughout modern time. These studies generally reveal substantial complexity in the forces that shape word frequencies. Altmann, Pierrehumbert, and Motter (2011) show that a word's *niche*, its characteristic features and the environment in which it is used, strongly influence the word's change in frequency. More specifically, they argue that some of the non-stationarity of word frequencies results from features of individuals like desires to convey information or identify with a particular social group. Petersen, Tenenbaum, Havlin, and Stanley (2012) show that word usage varies according to social, technological, and political pressures. In the simplest case, of course people start saying words like "email" once email is invented; but these trends extend to, for instance, measurable differences in word frequencies and word-birth and death in periods of drastic social and political change. Pagel, Atkinson, and Meade (2007) show that word frequency and language change are closely linked, such that low frequency words tend to evolve the most.

In general, these studies suggest that any theory aiming to explain Zipf's law must connect to the forces that shape frequencies, and with language change in general. How is it that processes affecting how frequencies change and how lexica evolve all yield a relatively conserved distribution across time? How does the nature of—perhaps drastic—language change maintain the distribution? Any theory which is not directly compatible with change must be missing a large part of what determines frequencies.

3.7 Power laws arise from (almost) nothing

A wide range of explanations of Zipf's law make reference to optimization and language change. However, we next show that this cannot be the entire story: a near-Zipfian word frequency distribution occurs even for wholly novel words whose content and use could not have been shaped by any processes of language change.

In a behavioral experiment, twenty five subjects were recruited from Amazon's mechanical turk an online platform that is becoming increasingly popular for experimental psychology (Paolacci, Chandler, & Ipeirotis, 2010; ?, ?; Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2012). Participants were given the following prompt: "An alien space ship crashes in the Nevada desert. Eight creatures emerge, a Wug, a Plit, a Blicket, a Flark, a Warit, a Jupe, a Ralex, and a Timon. In at least 2000 words, describe what happens next." Subjects' relative frequency distribution of each of these eight novel words was then computed on their produced text. Because different subjects may pick a different creatures as their "primary" character in the text, the analysis aggregated statistical by rank across subjects. It used the sampling methods described for Figure 1(a) to determine the estimated frequency $f(r)$ of each subject's r'th most frequent word, and then collapsed this distribution across subjects by rank. Thus, the frequency we report for the r'th most frequent word is the sum (or, scaled, mean) of each individual subject's r'th most frequent word. This aggregation was done to decrease noise since each subject uses each word only a handful of times[13].

The resulting subject-average frequency distribution is shown in Figure 8. This clearly demonstrates near-Zipfian scaling in frequency, despite the fact that all words are in some sense equivalent in the prompt—participants are not told, for instance, that one creature is extra salient, or that they should primarily describe one character. The context was chosen to bias them as little as possible about how much to describe each

[13]However, because we use separate subsets of the sample to estimate r and $f(r)$, this method does not introduce any spurious effects or non-independence errors.

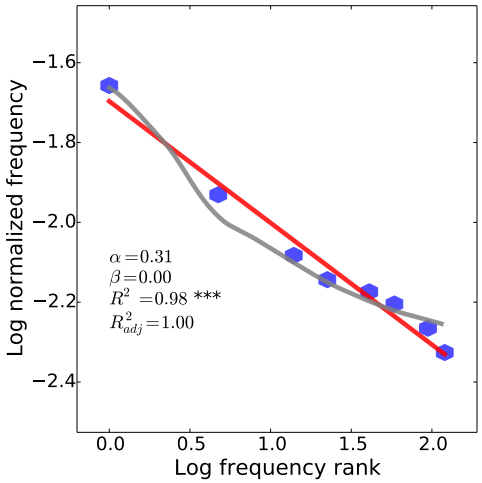


Figure 8: An approximate power law distribution of novel alien names used by subjects in making up a story.

creature and what role it would play in their novel story. Moreover, subjects show this distribution even though they are told almost nothing about the creatures (other than that they crashed from an Alien ship) and are told absolutely nothing about what happens next. Even in this context words *still* approximately follow the power law distribution, although larger-scale studies should be used to check that this effect is seen within individuals and is not the result of averaging together subjects.

In general, these findings suggest that a parsimonious, broad-coverage explanation for near-Zipfian distributions in language—one that can explain this experiment—should be applicable to people speaking about entirely novel, relatively unspecified referents.

3.8 Zipf's law occurs in other human systems

Interestingly, Zipf's law occurs in very many aspects of human society, including communication other than natural language. For instance, Zipfian (or near-Zipfian) frequency distributions occur in music (Manaris et al., 2005; D. H. Zanette, 2006, among others). They are observed in computer systems in the distribution of hardware instructions for programming languages (Shooman & Laemmel, 1977; Chen, 1991; Veldhuizen, 2005; Concas, Marchesi, Pinna, & Serra, 2007, among others), across many levels of abstraction in software (Louridas, Spinellis, & Vlachos, 2008), in n-tuples in computer code (Gan, Wang, & Han, 2009), and in many aspects of the internet (Adamic & Huberman, 2002). These findings complement the general result that Zipfian distributions occur in some form in a striking number of physical and biological systems (W. Li, 2002; Mitzenmacher, 2004; Newman, 2005; Farmer & Geanakoplos, 2006; S. A. Frank, 2009; Saichev et al., 2010). An important question for future work is to determine how broadly the word frequency distribution should be explained—should we seek explanations that unify language with music and perhaps other areas like computer software? Or does the profusion of derivations of Zipf's law mean that we shouldn't place such a strong weight on all-encompassing explanations, as very different mechanisms may give rise to the power law in different domains?

4 Models of Zipf's law

Now that we have reviewed a number of empirical phenomena about word frequencies, we next consider several of the attempts to explain Zipf's law in language, and relate these to the empirical phenomena just reviewed. These include explanations based on very simple statistical models (random typing, preferential re-use), the organization of semantic systems, deep optimization properties of communication, and universal properties of computational systems. As described above, very little of this work has sought independent

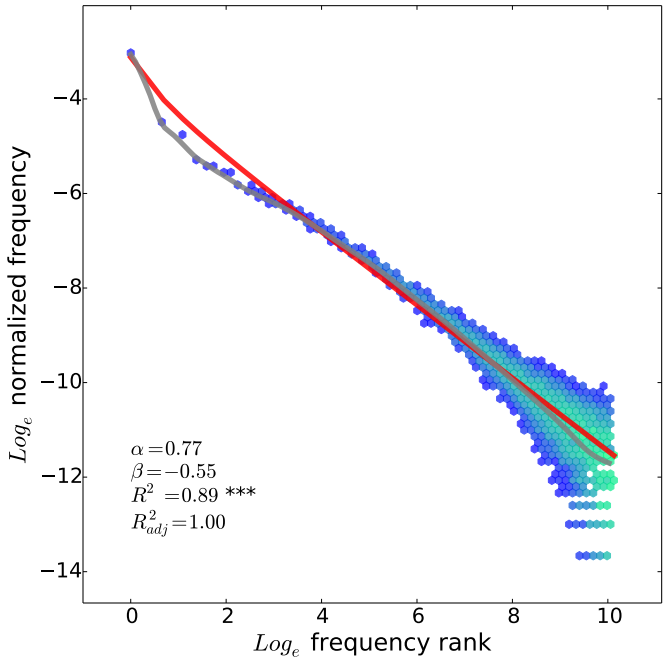


Figure 9: Frequency distribution of the 25,000 most frequent "words" in the ANC, where "e" rather than space (" ") was treated as a word boundary. This exhibits a clear near-Zipfian distribution, with the frequency of these words falling off much like (2).

tests of the key assumptions nor addressed the range of empirical phenomena described above. As we will see, none of the accounts is compellingly adequate alone. However, it may be true that there is no unitary explanation for word frequencies and that multiple causal forces are at play.

4.1 Random typing accounts

Given the ubiquity and robustness of Zipf's law, some have argued that the law is essentially a statistical artifact. This view is even widespread in certain communities and advocated by some prominent linguists like Chomsky (personal communication). The random typing account holds that Zipf's law is uninteresting because it holds even in very trivial statistical systems, like a monkey randomly banging on a typewriter (Miller, 1957; W. Li, 1992; Conrad & Mitzenmacher, 2004). Such a monkey will occasionally hit the space bar, creating a word boundary, and we can then look at the distribution of "word" frequencies. It turns out, that they follow a Zipfian distribution even though words are created entirely at random, one letter at a time. Intuitively, short words will tend to have a high probability, with the probability or frequency of words falling off approximately geometrically in their length. Although this process is clearly not an apt description of how humans generate language (see Howes, 1968; Piantadosi, Tily, & Gibson, 2013), the idea is that it should be treated as a *null hypothesis* about how language may be in the absence of other forces.

Indeed, the theoretical challenge raised by this model can be illustrated by taking a corpus of text and dividing it on a character other than the space (" ") character, treating, for instance, "e" as a word boundary[14]. Doing this robustly recovers a near-Zipfian distribution over these artificial "words," as shown in Figure 9. This shows some interesting deviations from the shape of the curve for natural language, but the general pattern is unmistakably similar to Figure 1(a), with a strong decrease in "word" frequency that falls off like a power law (linear on this plot) with length. So if the distribution occurs for even linguistically nonsensical "word" boundaries (like "e"), perhaps its presence in real language is not in need of explanation.

[14]Such that the string "I ate a enchilada for easter" would be segmented into "words" *I-at*, *-an-*, *nchilada-for-*, *ast*, *r*.

Some work has examined the ways in which the detailed statistics of random typing models look unlike that observed in real human language (Tripp & Feitelson, 1982; Baayen, 2001; Ferrer i Cancho & Solé, 2002; Ferrer i Cancho & Elvevåg, 2010; D. Manin, 2008, 2009). For instance, random typing models predict that the number of word types of a given length should decay exponentially in length; but in real language, this relationship is not even monotonically decreasing (D. Manin, 2009). Indeed, even the particular frequency distribution does not appear well-approximated by simple random typing models (Ferrer i Cancho & Elvevåg, 2010), although in other work Ferrer-i-Cancho is a strong proponent of such models (Ferrer i Cancho & Moscoso del Prado Martín, 2011). Of course, random-typing advocates might point out that tweaking the details of random typing models (e.g. changing letter frequencies, introducing N'th order Markov dependence) might allow them to fit the details of human language (for Zipf's law in Markov processes with random transitions, see Kanter & Kessler, 1995).

As such, a stronger argument than the details of the distribution is to recognize that they do not capture anything like the real causal process and therefore are poor scientific theories (Howes, 1968; Piantadosi et al., 2013). Indeed, once we appreciate that humans know *words* in their entirety and generate them intentionally to convey a meaning, it no longer makes sense to consider null hypotheses based on sub-word processes whose key feature is that a word's frequency is wholly determined by its components (e.g. letters) (Howes, 1968; Ferrer i Cancho & Elvevåg, 2010; Piantadosi et al., 2013). In the real cognitive system, people know whole words and do not emit sub-word components at random, and so clearly such processes cannot explain the cognitive origins of the law—a "deeper" (D. Manin, 2008) explanation is needed.

This counterpoint was articulated early by Howes (1968), but his reply has not been widely appreciated: "If Zipf's law indeed referred to the writings of 'random monkeys,' Miller's [random typing] argument would be unassailable, for the assumptions he bases it upon are appropriate to the behavior of those conjectural creatures. But to justify his conclusion that people also obey Zipf's law for the same reason, Miller must perforce establish that the same assumptions are also appropriate to human language. In fact, as we shall see, they are directly contradicted by well-known and obvious properties of languages." Those facts are, of course, that language is *not* generated at random, by accidentally happening to create a word boundary. The question remains, then, why is it that *real* processes of language generation give rise to this word frequency distribution.

Beyond the theoretical arguments against random typing accounts, such accounts are not compatible with several empirical facts reviewed earlier. The systematicity of word frequencies across meanings (Section 3.1) are particularly problematic for random typing models, since any process that is remotely like random typing will be unable to explain such patterns. One certainly would not be able to explain why cardinal number words also follow a near-Zipfian distribution, ordered precisely by magnitude. Moreover, random typing accounts cannot explain the variability across syntactic categories (Section 3.4)—why would certain word categories appear not to follow the model? Nor can it explain the tendency of subjects to follow the distribution for novel words (Section 3.7), and the simplest forms of random typing models are incompatible with the non-stationarity word frequencies exhibit (Section 3.5).

4.2 Simple stochastic models

One of the oldest approaches to explaining Zipf's law is to posit simple stochastic models of how words tend to be re-used in text. The idea is that preferential re-use will lead to a very skewed frequency distribution since frequent words will tend to get re-used even more. Intuitively, if you say, say, "pineapple" once you are more likely to repeat it later in the text, and such re-use can often be shown under certain assumptions to lead to Zipfian or near-Zipfian distributions. For instance, building on work of Yule (1944), Simon (1955) introduces a stochastic model that assumes (i) preferential re-use of previously frequent words, and (ii) a constant probability of introducing a new word. The stochastic model that Simon describes can be imagined to sequentially generate a text according to these assumptions, giving rise to a particular word frequency distribution over word types. Extensive discussion of this type of model and related ones can be found in Mitzenmacher (2004), Baayen (2001) and Farmer and Geanakoplos (2006), and a sophisticated and recent variant can be found in D. Zanette and Montemurro (2005).

This general class of models occupies an interesting ground between the psychological implausibility of random typing models and psychologically plausible models that capture, for instance, subjects' knowledge of whole words. However, like random typing models, they do not plausibly connect real causal stories

of language generation. As D. Manin (2008) writes, "Simon's model seems to imply that the very fact of some words being frequent and others infrequent is a pure game of chance." Such models only show that *if* language generation behaved like a certain stochastic model, *then* it would give rise to Zipf's law. It fails to establish what exactly it would mean for real human speakers to behave like the model, especially concerning the intentional production of meaningful language.

In this vein, Herdan (1961) wrote of Simon (1955)'s model: "For mathematical models to be of real value it is necessary that (1) the relationship between events of which the mathematical structure is to be a model should be what the mathematician believes it to be; (2) that the assumptions needed for constructing the model should be sensible, i.e. in accordance with how the operations in question take place; and (3) that the formulae derived in this way should fit the observed facts. None of these requirements must be neglected if the model is to fulfill its purpose. It is now a sad fact that model construction in mathematical linguistics seems dogged by the neglect of one or other of these requirements, especially the first, which cannot but have in its wake the neglect of the other two." Human speech is created with a purpose and the explanation for the frequency distribution must take into account this intentionality—why does an intentional process result in the Zipfian distribution? That is the fact that theories should seek to explain.

Further, it is not clear that the randomness of this kind of model can easily be connected to systematic relationships between meaning and frequency (Section 3.1). However, in some situations the simple stochastic model may actually be correct. The near-Zipfian use of novel words (Section 3.7) *may* be explained by these kinds of processes—perhaps in deciding how to continue their story, participants essentially sample from past referents with a probability that scales with recent use. It is useful to consider if this idea might even generalize to all language production: perhaps language is constrained by other factors like syntax, but on a large scale is characterized by stochastic re-use along the lines of Simon's model. Indeed, it is likely that given the non-stationarity of word frequencies (Section 3.5) something like these models must be approximately true. Words really are more likely to be re-used later in discourse. However, the underlying cause of this is much deeper than these models assume. Words are re-used in language (probably) not because of an intrinsic preference for re-use itself, but instead because there is a latent hidden variable, a *topic*, that influences word frequencies.

4.3 Semantic accounts

If the meanings of words in part determine frequency it is useful to consider whether semantic organization itself may give rise to the word frequency distribution. Guiraud (1968) argued that the law could result from basic ternary (true/false/undefined) elements of meaning called *semes* (e.g. animate/inanimate) with each word coding some number of semes. If semes must be communicated in speech this setup can give rise to a Zipfian word frequency distribution. Another hypothesis along the lines of semantics was put forth by D. Manin (2008), who argued that the law could result from labeling of a semantic hierarchy (e.g. Collins & Quillian, 1969; Fellbaum, 1998), combined with a pressure to avoid synonymy. Intuitively, if words label different levels of semantic space and evolve to avoid too much overlap, the lexicon arrives at coverings of semantic space which, he shows via simulation, will result in Zipf's law.

This theory motivated the comparisons in Section 3.3, which examined words whose meanings are strongly constrained by the world. It is unlikely that language had much of a "choice"—or optimizing pressure—in choosing which of the possible ways of labeling months, planets, or elements, since these meanings are highly constrained by the natural world. Yet we see near-Zipfian distributions for even these words. We find similar results for words whose referential content is fixed, like taboo words (Section 3.2). The results on number words (Section 3.1) provide another compelling case where choice of semantic referent by the lexicon is not likely to explain word frequencies which are nonetheless power laws. The behavioral experiment (Section 3.7) additionally indicates even for words which are initially, in some sense, on equal ground and whose specific semantics is not given, people *still* follow a near-Zipfian distribution. All of these results do not indicate that semantic explanations play *no* role in determining word frequencies, but only that they are likely not the entire story[15].

[15]In evaluating theories, one might wonder if these semantic comparisons are essentially just random subsets of words, and that a random subset of a Zipfian distribution may tend to look Zipfian. Therefore, it may not be very strong evidence against theories based on meaning that we still see Zipfian distributions when we control or constrain meaning. However, note that theories based on meaning explain the distribution starting from semantics. They explain patterns across the entire lexicon by

4.4 Communicative accounts

Various authors have also explained the Zipfian distribution according to communicative optimization principles. Zipf (1949) himself derived the law by considering a trade-off between speakers and listener's effort. Mandelbrot (1953) shows how the Zipfian distribution could arise from minimizing information-theoretic notions of cost (Mandelbrot, 1962, 1966), ideas further developed by D. Manin (2009), and Ferrer i Cancho and colleagues (Ferrer i Cancho & Solé, 2003; Ferrer i Cancho, 2005a; i Cancho, 2005) and more recently Salge et al. (2013).

In Ferrer i Cancho and Solé (2003), the authors imagine optimizing a matrix $\mathbf{A} = \{A_{ij}\}$ where A_{ij} is 1 if the i'th word can refer to the j'th meaning. In their framework, speakers pay a cost proportional to the diversity of signals they must convey and listeners pay a cost proportional to the (expected) entropy over referents given a word (for variants and elaborations, see Ferrer i Cancho & Díaz-Guilera, 2007) . There is a single parameter which trades off the cost between speakers and listeners, and the authors show that for a very particular setting of this parameter $\lambda = 0.41$ they recover a Zipfian distribution.

While mathematically sophisticated, their approach makes several undesirable choices. In the implementation, it assumes that meanings are all equally likely to be conveyed, an assumption which is likely far from true even in constrained semantic domains (Figure 5). Later versions of this model (Ferrer i Cancho, 2005b) study variants without this assumption, but it is not clear—for any model—what the psychologically relevant distribution should be for how often each meaning is needed, and how robust this class of models is to that distribution[16], or how such accounts might incorporate other effects like memory latency, frequency effects, or context-based expectations.

Second, the assumption that speakers' difficulty is proportional to the entropy over signals is not justified by data and is not predicted from a priori means—a better a priori choice might have been the entropy over signals *conditioned* on a meaning since this captures the uncertainty for the psychological system. In this vein, none of the assumptions of the model are tested or justified on independent psychological grounds.

Thirdly, this work requires a very specific parameter $\lambda \approx 0.4$ to recover Zipf's law, and the authors show that it no longer does, for $\lambda = 0.5$ or $\lambda = 0.3$. The required specificity of this parameter is undesirable from the perspective of statistical modeling—the so-called "Spearman's principle" (Glymour, Scheines, Spirtes, & Kelly, 1987)—as it suggests non-robustness.

In the context of the corpus analyses provided above, communicative accounts would likely have difficulty explaining near-Zipfian distribution for fixed referential content (Section 3.2) and variability of fits across syntactic categories (Section 3.4). Theories based on communicative optimization like Ferrer i Cancho and Solé (2003) are based on choosing which meanings go with which words—when optimized for communication, this process is supposed to give rise to the law. But we still see it in domains where this mapping is highly constrained (Section 3.3) and for number words (Section 3.1) where it is hard to imagine what such optimization might mean. Therefore, it is unclear on a conceptual level how these accounts might handle such data. It is also not straightforward to see how communicative accounts could accommodate the behavioral results (Section 3.7), since it is hard to imagine in what sense communication of names might be actively optimized by speakers simply telling a story. The intentionality of storytelling—wanting to convey a sequence of events you have just thought of—seems very different than the language-wide optimization of information-theoretic quantities required by communicative accounts.

This is certainly not to say that there is no way a communicative theory could account for the facts or that communicative influences play no role. An adequate theory has just not been formalized or empirically evaluated yet[17].

appealing to semantic properties of single words, and so cannot explain the subsets of words that look Zipfian but don't have the required semantic properties.

[16] A result on a large class of meaning distributions might help that issue.

[17] Moving forward, however, it will be important for communicative accounts to explicitly address *predictability* of words. As Shannon (1948) demonstrated, the predictability (negative log probability) of a word is the measure of the information it conveys. This means that a theory based on communication should be intrinsically linked to theories of what human language comprehenders find predictable (e.g. Demberg & Keller, 2008; Levy, 2008; Levy & Jaeger, 2007; A. Frank & Jaeger, 2008; Jaeger, 2010; Piantadosi, Tily, & Gibson, 2011; N. J. Smith & Levy, in press) and how much information is effectively conveyed for such mechanisms.

4.5 Explanations based on universality

The models described so far explain Zipf's law from psychological or statistical processes. But it is also possible that Zipf's law in language arises from a universal pressure that more generally explains its prevalence throughout the sciences. An analogy is that of the Central Limit Theorem (CLT) and the normal distribution. When a normal distribution is observed in the world (in, e.g., human heights), commonly the CLT is taken to explain *why* that distribution is found, since the theorem shows that normal distributions should be expected in many places—in particular where many independent additive processes are at play[18][19]. It is reasonable to ask if there is a such a theorem for power laws: do they simply arise "naturally" in many domains according to some universal law? Perhaps even the multitude of derivations of Zipf's law indicate that the presence of the law in language is not so surprising or noteworthy.

There are in fact derivations of Zipf's law from very fundamental principles that in principle span fields. Corominas-Murtra and Solé (2010) show that Zipfian distributions of symbol sequences can be derived in the (maximally general) framework of algorithmic information theory (M. Li & Vitányi, 2008), considering symbols to be observations of a system growing in size, but which is constrained to have bounded algorithmic complexity. Their account even explains the exponent $\alpha \approx 1$ observed in language, providing a compelling explanation of Zipf's law in general complex systems. Y. I. Manin (2013) provides a related account deriving Zipf's law from basic facts about Kolmogorov complexity and Levin's probability distribution (see also Veldhuizen, 2005). S. A. Frank (2009) studies entropy maximizing processes, relating power laws to normal distributions and other common laws in the sciences. In general, these accounts say that we should have *expected* Zipf's law to appear in many systems simply due to the intrinsic properties of information, complexity, and computation.

Similarly, there have also been somewhat more deflationary universal explanations. Remarkably, Belevitch (1959), showed how a Zipfian distribution could arise from a first-order approximation to most common distributions; he then showed how the Zipf-Mandelbrot law arose from a second-order approximation. In this kind of account, Zipf's law could essentially be a kind of statistical artifact of using a frequency/frequency-rank plot, when the real underlying distribution of frequencies is any of a large class of distributions.

All of these accounts based on universal a priori notions are interesting because they would explain the surprising scope of Zipf's law across the sciences without requiring many domain-specific assumptions. However, one troubling shortcoming of these theories as explanations is that they have not been used to generate novel predictions; it is hard to know what type of data could falsify them, or how we would know if they are really the "right" explanation as opposed to any of the more psychologically-motivated theories. Do the assumptions they require really hold in human psychology, and how would we know? One interesting test might be for these kinds of explanations to derive predictions for the variance beyond Zipf's law that should be expected in any finite sample, and perhaps in some situations even predict correlated errors like those seen in Figure 1(b). If Zipf's law is universal, we would require additional mechanisms to explain domains where Zipf's law less well or for different parameters (Section 3.4) or how it could also hold given systematic relationships with meaning (Section 3.1). It is unclear if the behavioral experiment (Section 3.7) is compatible with these accounts—what might people be doing psychologically in this experiment, and how does it translate into universal derivations of Zipf's law?

4.6 Other models

We note that there are many other accounts of Zipf's law in language and elsewhere, actually giving rise to a fat tail of theories of the law. For instance, Baek et al. (2011) shows how Zipf's law can be derived from processes that randomly divide elements into groups. Arapov and Shrejder (1978) argue that Zipf's law can be derived by simultaneously maximizing two entropies: the number of different texts creatable by a lexicon and the number of different ways the same text can be created by a lexicon. As argued by D. Manin (2008), this approach compellingly lacks a priori justification and a possible optimizing mechanism. Other

[18] For generalizations of the CLT that are connected to power-laws and similar distributions, see Gnedenko and Kolmogorov (1968) and Roehner and Winiwarter (1985).

[19] In actuality, it may not even be clear for most common situations how the assumptions of the CLT or its generalizations hold (Lyon, 2014). The true reason for the ubiquity of normal distribution may be related to its other properties, such as entropy-maximation (Lyon, 2014), suggesting that maximum-entropy derivations may be most fruitful for explaining Zipf's law broadly (see, e.g. S. A. Frank, 2009).

optimizations of, e.g. Fisher information (Hernando et al., 2009), can also give rise to Zipfian distributions. Popescu (2009, Chapter 9) sketch a simple vocabulary growth model. Parker-Rhodes and Joyce (1956) argue that the distribution arises by a linear search through words in long-term memory ordered by frequency during normal language processing, where the time required to scan a word is proportional to the number of words scanned. To date, there is no evidence for this kind of process in normal language use. In general, it is not clear that any of these kinds of accounts could handle the gamut of empirical phenomena reviewed above, and to our knowledge none have proposed and evaluated independent tests of their assumptions.

5 Conclusion and forward directions

Word frequencies are extremely interesting. They are one of the most basic properties of humans' communicative system and play a critical role in language processing and acquisition[20]. It is, in short, remarkable that they can be well-characterized by a simple mathematical law. With good cause, many have attempted to derive this law from more basic principles. Notably, theories of language production or discourse do not explain the law.

This review has highlighted several limitations in this vast literature. First, the method of plotting word frequency distributions has obscured an important fact: word frequencies are not actually so simple. They show statistically-reliable structure beyond Zipf's law that likely will not be captured with any simple model. At the same time, the large-scale structure is robustly Zipfian.

Second, essentially all of the prior literature has focused very narrowly on deriving the frequency/frequency-rank power law, while ignoring these types of broader features of word frequencies. This in some sense represents a misapplication of effort towards explaining an effect—the Zipfian distribution—instead of uncovering the causal forces driving word frequencies in the first place. This is what makes so many derivations of Zipf's law unsatisfying: they do not account for any psychological processes of word production, especially the intentionality of choosing words in order to convey a desired meaning. A focus on explaining what words are needed at each point in a normal conversation would begin to explain *why* word frequencies look like they do. Until then, a deep mystery remains: why should language generation mechanisms follow such a precise mathematical law, even in cases of constrained meanings and totally novel words, but apparently not identically for all syntactic categories?

It should be clear that this question will only be addressable by broadly studying properties of word frequencies beyond the frequency distribution itself. The empirical phenomena reviewed here (Section 3) have aimed to encourage more comprehensive evaluation of theories of the Zipfian distribution that is observed. This review has revealed that likely none of the previous accounts are sufficient alone and that the facts surrounding word frequencies are complex and subtle. A sticking point for many theories will be the behavioral results showing Zipf's law for novel words (Section 3.7). These results likely have to do with properties of human memory since it is hard to think of other pressures in this experiment that would lead people into power-law use of words. Indeed, human memory has independently been characterized as following powers laws (see Wickelgren, 1974, 1977; Wixted & Ebbesen, 1991, 1997; Wixted, 2004a, 2004b). Such scaling relationships are broadly observed elsewhere in cognition (Kello et al., 2010). If these properties of memory are the underlying cause of near-Zipfian laws in language, it could provide a parsimonious and general explanation, able to unify word frequencies with memory, while also explaining the occurrence of related laws in other systems humans use like computer software and music (Section 3.8).

Interestingly, if human memory is the underlying cause of Zipf's law in language, we are left to ask why memory has the form that it does. A plausible hypothesis advocated by Anderson and Schooler (1991) is that memory is well-adapted to environmental stimuli, meaning that Zipfian structures in the real world might ultimately create the observed form of word frequencies distributions. Of course, any such theory of word frequencies would require substantial elaboration in order to address the complexities of how well

[20] While it rarely enters into discussions of the origins of Zipf's law, it's important to point out that people really do appear to *know* word frequencies. Evidence for this is apparent in both detailed, controlled (e.g. Dahan, Magnuson, & Tanenhaus, 2001) and broad-coverage (e.g. Demberg & Keller, 2008) analyses of language processing (see Ellis, 2002, for a review). Similarly, frequency effects are observed in language production (Oldfield & Wingfield, 1965; Jescheniak & Levelt, 1994; Levelt, 1999). These effects show that speakers know something about the frequencies with which words occur in their input, and that this type of knowledge is used in online processing.

Zipfian distributions fit different types of words, the residual deviations from the distribution observed in language (Section 2), and interactions with semantics (Section 3.1, 3.2).

In general, the absence of novel predictions from authors attempting to explain Zipf's law has led to a very peculiar situation in the cognitive sciences, where we have a profusion of theories to explain an empirical phenomenon yet very little attempt to distinguish those theories using scientific methods. This is problematic precisely because there *are* so many ways to derive Zipf's law that the ability to do so is extremely weak evidence for any theory. An upside of this state of the field is that it is ripe for empirical research. The downside is that because proposals of theories have not been based on incremental empirical discoveries, many can be easily shown to be inadequate using only minimal new data. The key will be for explanations of Zipf's law to generate novel predictions and to test their underlying assumptions with more data than the law itself. Until then, the prior literature on Zipf's law has mainly demonstrated that there are many ways to derive Zipf's law. It has not provided any means to determine which explanation, if any, is on the right track.

6 Acknowledgments

I'm very grateful to Leon Bergen, Ev Fedorenko, and Kyle Mahowald for providing detailed comments on this paper. Andreea Simona Calude James generously shared the data visualized in Figure 2. I am highly appreciative of Dmitrii Manin, Bob McMurray and an anonymous reviewer for providing extremely helpful comments on this work. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number F32HD070544. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics*, *3*(1), 143–150.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS ONE*, *4*(11), e7678.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLOS ONE*, *6*(5), e19009.

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396.

Arapov, M., & Shrejder, Y. (1978). Zakon cipfa i princip dissimmetrii sistem [Zipf's law and system dissymmetry principle]. *Semiotics and Informatics*, *10*, 74–95.

Baayen, R. (2001). *Word frequency distributions* (Vol. 1). Kluwer Academic Publishers.

Baek, S. K., Bernhardsson, S., & Minnhagen, P. (2011). Zipf's law unzipped. *New Journal of Physics*, *13*(4), 043004.

Belevitch, V. (1959). On the statistical laws of linguistic distributions. *Annales de la Societe Scientifique de Bruxelles*, *73*(3), 301–326.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 17–35.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, *10*, 71.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

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.

Carroll, J. B. (1967). On sampling from a lognormal model of word frequency distribution. *Computational analysis of present-day American English*, 406–424.

Carroll, J. B. (1969). *A rationale for an asymptotic lognormal form of word-frequency distributions.*

Chater, N., & Brown, G. D. (1999). Scale-invariance as a unifying psychological principle. *Cognition*, *69*(3), B17–B24.

Chen, Y.-S. (1991). Zipf's law in natural languages, programming languages, and command languages: the Simon-Yule approach. *International journal of systems science*, *22*(11), 2299–2312.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum.*

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. *Statistical models in S*, 309–376.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Concas, G., Marchesi, M., Pinna, S., & Serra, N. (2007). Power-laws in a large object-oriented software system. *Software Engineering, IEEE Transactions on*, *33*(10), 687–708.

Conrad, B., & Mitzenmacher, M. (2004). Power laws for monkeys typing randomly: the case of unequal probabilities. *Information Theory, IEEE Transactions on*, *50*(7), 1403–1414.

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of zipf's law. *Physical Review E*, *82*(1), 011102.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive psychology*, *42*(4), 317–367.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230.

Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-world phrases.

Egghe, L. (2000). The distribution of N-grams. *Scientometrics*, *47*(2), 237–252.

Ellis, N. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(2), 143–188.

Farmer, J. D., & Geanakoplos, J. (2006). *Power laws in economics and elsewhere* (Tech. Rep.). Santa Fe Institute Tech Report.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferrer i Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. *Physica A: Statistical Mechanics and its Applications*, *345*(1), 275–284.

Ferrer i Cancho, R. (2005b). Zipf's law from a communicative phase transition. *The European Physical Journal B-Condensed Matter and Complex Systems*, *47*(3), 449–457.

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(06), P06009.

Ferrer i Cancho, R., & Elvevåg, B. (2010). Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLOS ONE*, *5*(3).

Ferrer i Cancho, R., & Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, *2011*, L12002.

Ferrer-i-Cancho, R., & Servedio, V. D. (2005). Can simple models explain zipf's law in all cases? *Glottometrics*, *11*, 1-8. Retrieved from `http://groups.lis.illinois.edu/amag/langev/paper/ferrer05zipfLawSimpleModels.html`

Ferrer i Cancho, R., & Solé, R. (2002). Zipf's law and random texts. *Advances in Complex Systems*, *5*(1), 1–6.

Ferrer i Cancho, R., & Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788.

Ferrer i Cancho, R., & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), 165–173.

Frank, A., & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Cognitive Science Society*.

Frank, S. A. (2009). The common patterns of nature. *Journal of evolutionary biology*, *22*(8), 1563–1585.

Gan, X., Wang, D., & Han, Z. (2009). N-tuple Zipf Analysis and Modeling for Language, Computer Program and DNA. *arXiv preprint arXiv:0908.0500*.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Gnedenko, B. V., & Kolmogorov, A. (1968). *Limit distributions for sums of independent random variables* (Vol. 233). Addison-Wesley Reading.

Guiraud, P. (1968). The semic matrices of meaning. *Social Science Information*, *7*(2), 131–139.

Ha, L. Q., Hanna, P., Ming, J., & Smith, F. (2009). Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, *32*(1), 101–113.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the 19th international conference on computational linguistics-volume 1* (pp. 1–6).

Herdan, G. (1960). *Type-token mathematics* (Vol. 4). Mouton.

Herdan, G. (1961). A critical examination of simon's model of certain distribution functions in linguistics. *Applied Statistics*, 65–76.

Herdan, G. (1964). *Quantitative linguistics*. Butterworths London.

Hernando, A., Puigdomènech, D., Villuendas, D., Vesperinas, C., & Plastino, A. (2009). Zipf's law from a fisher variational-principle. *Physics Letters A*, *374*(1), 18–21.

Howes, D. (1968). Zipf's Law and Miller's Random-Monkey Model. *The American Journal of Psychology*, *81*(2), 269–272.

i Cancho, R. F. (2005). Hidden communication aspects inside the exponent of zipf's law. , *11*, 98–119.

Jaeger, F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824.

Kanter, I., & Kessler, D. (1995). Markov processes: linguistics and zipf's law. *Physical review letters*, *74*(22), 4559–4562.

Kawamura, K., & Hatano, N. (2002). Universality of zipf's law. *arXiv preprint cond-mat/0203455*.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*(15), 9085–9089.

Kello, C. T., Brown, G. D., Ferrer-i Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223–232.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*, 849–856.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag.

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

Li, W. (2002). Zipf's law everywhere. *Glottometrics*, *5*, 14–21.

Lin, Y., Michel, J., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Louridas, P., Spinellis, D., & Vlachos, V. (2008). Power laws in software. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *18*(1), 2.

Lyon, A. (2014). Why are Normal Distributions Normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649.

Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., & Davis, R. B. (2005). Zipf's law, music classification, and aesthetics. *Computer Music Journal*, *29*(1), 55–69.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 486–502.

Mandelbrot, B. (1962). On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 190–219.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P. Lazarsfield & N. Henry (Eds.), *Readings in Mathematical Social Sciences*. Cambridge, MA: MIT Press.

Manin, D. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*(7), 1075–1098.

Manin, D. (2009). Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics*, *16*(3), 274–285.

Manin, Y. I. (2013). Zipf's law and L. Levin's probability distributions. *arXiv preprint arXiv:1301.0427*.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.

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.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, *44*(1), 1–23.

Miller, G. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 311–314.

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

Montemurro, M. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, *300*(3), 567–578.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, *46*(5), 323–351.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Orlov, J., & Chitashvili, R. (1983). Generalized Z-distribution generating the well-known rank-distributions. *Bulletin of the Academy of Sciences, Georgia*, *110*, 269–272.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*(7163), 717–720.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Parker-Rhodes, A., & Joyce, T. (1956). A theory of word-frequency distribution. *Nature*, *178*, 1308.

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific reports*, *2*.

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychonomic Bulletin and Review*, 1–10. Retrieved from `http://colala.berkeley.edu/papers/piantadosi2016rational.pdf` doi: 10.3758/s13423-015-0963-8

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526. Retrieved from `http://colala.berkeley.edu/papers/PNAS-2011-Piantadosi-1012551108.pdf`

Piantadosi, S. T., Tily, H., & Gibson, E. (2013). Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin [arXiv:1209.1751]. *ArXiv e-prints*. Retrieved from `https://arxiv.org/pdf/1307.6726`

Popescu, I.-I. (2009). *Word frequency studies* (Vol. 64). Walter de Gruyter.

Reppen, R., & Ide, N. (2004). The American National Corpus overall goals and the first release. *Journal of English Linguistics*, *32*(2), 105–113.

Roehner, B., & Winiwarter, P. (1985). Aggregation of independent paretian random variables. *Advances in applied probability*, 465–469.

Rouault, A. (1978). Lois de Zipf et sources Markoviennes. In *Annales de l'institut h. poincare.*

Saichev, A., Malevergne, Y., & Sornette, D. (2010). *Theory of Zipf's law and beyond* (Vol. 632). Springer.

Salge, C., Ay, N., Polani, D., & Prokopenko, M. (2013). *Zipf's Law: Balancing Signal Usage Cost and Communication Efficiency* (Tech. Rep.). Santa Fe Institute Working Paper #13-10-033.

Shannon, C. (1948). *The Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press.

Shooman, M., & Laemmel, A. (1977). Statistical theory of computer programs information content and complexity. In *Compcon fall'77* (pp. 341–347).

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, *70*(351a), 542–547.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 425–440.

Simon, H. A. (1960). Some further notes on a class of skew distribution functions. *Information and Control*, *3*(1), 80–88.

Smith, N. J., & Levy, R. (in press). The effect of word predictability on reading time is logarithmic. *Cognition*.

Smith, R. D. (2008). Investigation of the zipf-plot of the extinct meroitic language. *arXiv preprint arXiv:0808.2904*.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Stumpf, M. P., & Porter, M. A. (2012). Critical truths about power laws. *Science*, *335*(6069), 665–666.

Tripp, O., & Feitelson, D. (1982). Zipf's law re-visited. *Studies on Zipf's law*, 1–28.

Veldhuizen, T. L. (2005). Software libraries and their reuse: Entropy, kolmogorov complexity, and zipf's law. *arXiv preprint cs/0508023*.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, *2*(4), 775–780.

Wickelgren, W. A. (1977). *Learning and memory.* Prentice-Hall Englewood Cliffs, NJ.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications on pure and applied mathematics*, *13*(1), 1–14.

Wixted, J. T. (2004a). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological review*, *111*(4), 864–879.

Wixted, J. T. (2004b). The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.*, *55*, 235–269.
Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological science*, *2*(6), 409–415.
Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*(5), 731–739.
Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *213*, 21–87.
Yule, G. U. (1944). *The statistical study of literary vocabulary*. CUP Archive.
Zanette, D., & Montemurro, M. (2005). Dynamics of text generation with realistic zipf's distribution. *Journal of Quantitative Linguistics*, *12*(1), 29–40.
Zanette, D. H. (2006). Zipf's law and the creation of musical context. *Musicae Scientiae*, *10*(1), 3–18.
Zipf, G. (1936). *The Psychobiology of Language*. London: Routledge.
Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.