Zipf's law outside the middle range

András Kornai
Belmont Research
andras@kornai.com

Abstract

Zipf (1949) already noted that the linear relationship that he observed between log frequency and log rank is strongest in the middle range: both very high and very low frequency items tend to deviate from the log-log regression line. In this paper the causes for such deviations are investigated and a more detailed statistical model is offered. The *subgeometric mean property* of frequency counts is introduced and used in proving that the size of the vocabulary tends to infinity as sample size is increased without bounds.

0 Introduction

In spite of its venerable history (starting with Pareto 1897, Estoup 1916, Willis 1922, Yule, 1924) and considerable empirical support, Zipf's law remains one of the least understood phenomena in mathematical linguistics. Given a corpus of N word tokens, arranging word types in order of descending token frequency (called "rank" and denoted by r in what follows), the plot of log frequencies against log ranks shows, at least in the middle range, a reasonably linear relation. Fig. 1 shows this for a single issue of an American newspaper, the *San Jose Mercury News*, or *Merc* for short.

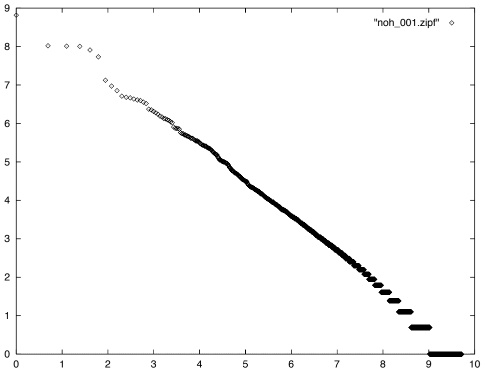

Figure 1: Log-log plot for a newspaper issue (150k words)

Denoting the slope of the linear portion by $-B$, B is close to unity, slightly higher on some plots, slightly lower on others. As Mandelbrot repeatedly stressed in the Simon-Mandelbrot debate and elsewhere, the apparent flexibility in choosing any number close to 1 is actually cause for serious concern, inasmuch as for $B \leq 1$, $\log(p_r) \sim r^{-B}$ would imply that $\sum_{r=1}^{\infty} p_r$ diverges. Some authors like Samuelsson (1996) in fact reserve the term "Zipf's law" to the case $B = 1$ and observe, quite correctly, that this formulation of the law implies a finite vocabulary. While the narrower terminology would be historically more faithful (for a discussion and critique of Zipf's original notion of an optimum corpus size see Powers 1998) in this paper we will use the term "Zipf's law" in the broader sense. Since our very first Theorem states that under very general conditions vocabulary size can *not* be expected to be finite, it is worth discussing this matter in some detail.

Let us assume for the moment that the primary focus of our interest is the journalistic/nonfiction-literary style exemplified by the Merc, or that even more narrowly, our focus is just the Merc and we have no intention of generalizing our results to other newspapers, let alone other stylistic ranges. While the Merc is a finite corpus, growing currently at a rate of 60M words/year, our goal is not an exhaustive characterization of past issues, but rather predicting word frequencies for future issues as well. Therefore, the *population* we care about is an infinite one, comprising all *potential* issues written in "Merc style" and each issue is but a finite *sample* from this infinite population. Though undoubtedly this population shows a certain amount of diachronic drift as new words enter the language and old ones fall into disuse, here we take an "adiabatic" perspective on increasing sample size, and treat our population (and by means of randomizing article order, also our samples) essentially as a synchronic slice.

Given some sample S of articles and some word w, it is a simple matter to obtain a *sample count* $F_S(w)$ of w and divide it by the *sample size* N to obtain the *relative sample frequency* $f_S(w) = F_S(w)/N$. Standard textbooks like Cramér (1955) take it as axiomatic that by randomly increasing S without bounds, $f_S(w) \to f(w)$ as $N \to \infty$, i.e. that for every word, sample frequencies will converge to a fixed constant $0 \leq f(w) \leq 1$ that is the *probability* (population frequency) of the word. In the context of using ever-increasing corpora as samples this *stability property of frequency ratios* has often been questioned, both on grounds of diachronic drift, which we can safely disregard here, and on the basis of the following argument: if the lexicon is not closed, then the true probability of a word should, on average, decay as sample size is increased. While it is certainly true that the average will tend to zero, the probabilities of individual words need not tend to zero.

Section 1 of this paper provides a fully worked out example of this phenomenon, compares Zipf's law to a simple *exponential decay* model and derives a lower bound on vocabulary growth. The case of high frequency items is discussed in Section 2, and we turn to the case of low frequency items in Section 3. Unless noted otherwise, we illustrate our main points with a corpus of some 300 issues of the *Merc* totaling some 43M words. While this is not a large corpus by contemporary standards, it is still an order of magnitude larger than the classic Brown and LOB corpora on which so much of our current ideas about word frequencies was first developed and tested, and empirical regularities observed on a corpus this size can not be dismissed lightly.

As a practical matter, we need a definition of when a token belongs to a type that is capable of handling the issues of `lexing` (punctuation, capitalization, etc.) that inevitably arise in the course of any corpus-based work. This will have little effect on our conclusions, but for the sake of concreteness we will assume here that all characters are lowercased and all special characters, except for hyphen and apostrophe, are mapped on whitespace. The terminal symbols or *letters* of our alphabet are therefore $L = \{a, b, ...z, 0, 1, ...9,', -\}$ and all word types are strings in L^*, though word tokens are strings over a larger alphabet including capital letters, punctuation, and special characters.

1 Exponential decay

Since word frequencies span many orders of magnitude, it is difficult to get a good feel for their rate of convergence just by looking at frequency counts. The log-log scale used in Zipf plots is already an indication of the fact that to get any kind of visible convergence exponentially growing corpora need to be considered. Much of traditional quantitative linguistic work stays close to the Zipfian optimum corpus size of $10^4 - 10^5$ words simply because it is based on a closed corpus such as a single book or even a short story or essay. But as soon as we go beyond the first few thousand words, relative frequencies are already in the 10^{-6} range. Such words of course rarely show up in smaller corpora, even though they are often perfectly ordinary words such as *uniform* that are familiar to all adult speakers of English. Let us therefore begin by considering an artificial example, in which samples are drawn from an underlying geometrical distribution $f(w_r) = 1/2^r$.

Example 1. If the rth word has probability $p_r = 2^{-r}$, in a random sample S of size $N = 2^m$ we expect 2^{m-1} tokens of w_1, 2^{m-2} tokens of $w_2, \ldots, 2$ tokens of w_{m-1}, 1 token of w_m and one other token, most likely another copy of w_1. If this expectation is fulfilled, the frequency ratio based estimate $f_S(w_r)$ of each probability $p_r = f(w_r)$ is correct within $1/N$ i.e. convergence is limited only by the resolution offered by corpus size N, yet the number of types $V(N)$ observed in a sample of N tokens still tends to infinity with $\log_2(N)$.

Discussion. Needless to say, in an actual experiment we could hardly expect to get results this precise, just as in $2N$ tosses of a fair coin the actual value of heads is unlikely to be exactly N. Nevertheless, the mathematical expectations are as predicted, and the example shows that no argument based based on the average decline of probabilities could be carried to the point of demonstrating that a closed/finite vocabulary is logically necessary. Though not necessary, finite vocabulary is still possible: what we will demonstrate in Theorem 1 is that this possibility is logically incompatible with *observable* properties of corpora, as long as these are treated as random samples from an underlying probability distribution rather than as objects fully defining a probability distribution.

In short, we assume that population frequencies give a probability distribution over L^*, but for now remain neutral on the issue of whether the underlying vocabulary is finite (closed) or infinite (open). We also remain neutral on the rate of convergence of frequency ratios, but note that it can be seen to be rather slow, and not necessarily uniform. If rates of convergence were fast to moderate, we would expect empirical rankings based on absolute frequencies to approximate the perfect ranking based on population frequencies at a comparable rate. For example one could hope that any word that has over twice the average sample frequency $1/V(N)$ is already "rank stabilized" in the sense that increasing the sample size will not change its rank. Such hopes are, alas, not met by empirical reality: doubling the sample size can easily affect the ranking of the first 25 items even at the current computational limits of N, 10^9-10^{10} words. For example, moving from a 10M corpus of the Merc to a 20M corpus already affects the rankings of the first *four* items, changing *the, of, a, to* to *the, of, to, a*.

Since sample rank is an unreliable estimate of population rank, it is not at all obvious what Zipf's law really means: after all, if we take any set of numbers and plot them in decreasing order, the results charted on log-log scales may well be approximately linear. As a first step, we will *normalize* the data, replacing absolute rank r by relative rank $x = r/V(N)$. This way, the familiar Zipf-style plots, which were not scale invariant, are replaced by plots of function values $f(x)$ restricted to the unit square. $f(1/V(N)) = f(w_1)$ is the probability of the most frequent item, $f(1) = f(V(N)/V(N)) = 1/N$ is the probability of the least frequent item, and for technical reasons we define the values of f between $r/(V(N))$ and $(r+1)/V(N)$ to be $p(w_{r+1})$. A small sample (four articles) is plotted in this style in Fig. 2. Since the area under the curve is $1/V(N)$, by increasing the sample size plots of this kind get increasingly concentrated around the origin.

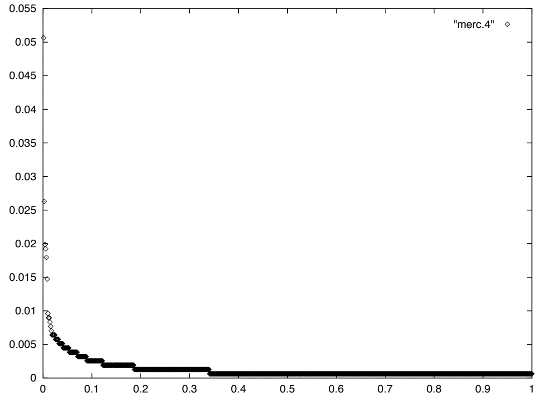

Figure 2: Bounded plot of 4 articles (1.5k words)

In approximating such a curve an obvious choice would be to try *exponential decay* i.e. $f(x) \sim Ce^{-Dx}$ with some constants $C, D > 0$. However, for reasons that will shortly become apparent, no such curve provides a very good fit, and we merely use the exponential model as a tool to derive *from first principles* a lower bound for $V(N)$. We will use the following facts:

(1) For any f obtained from a random sample S of size N, $f(1/V(N))$ tends to p_1, the frequency of the most frequent item, as $N \to \infty$

(2) For any f obtained from a random sample S of size N, $f(1) = 1/N$

(3) Word frequencies decay subexponentially (slower than $\exp(-Dx)$ for any $D > 0$).

Theorem 1. Under conditions (1-3) $V(N)$ grows at least as fast as $\log(N)(1 - 1/N)$.

Proof: $1/V(N) = \sum_{r=1}^{V(N)} f(r/V(N))/V(N)$ is a rectangular sum approximating $\int_0^1 f(x)dx$. Since $f(x)$ is subexponential, for any $g(x) = \exp(-Dx)$ that satisfies $g(1/V(N)) \geq p_1$ and $g(1) \geq 1/N$, we have $g(x) \geq f(x)$ everywhere else in the interval $[1/V(N), 1]$, and therefore $1/V(N) < \int_0^1 \exp(-Dx)dx = (1 - \exp(-D))/D$. Using (2) we compute $D = \log(N)$, $V(N) \geq \log(N)(1 - 1/N)$.

Discussion. Since any theorem is just as good as its premises, let us look at the conditions in some detail. (1) is simply the axiom that sample frequencies for the single most frequent item will tend to its population frequency. Though this is not an entirely uncontroversial assumption, we believe that the preceding discussion provides sufficient grounds for adopting it. On the surface (2) may look more dubious: there is no *a priori* reason for the least frequent word in a sample to appear only once. For example, in closed vocabulary Bernoulli experiments we would expect every word to appear at least twice as soon as the sample size is twice the inverse probability of the least frequent word. In the final analysis, (2) rests on the massively supported empirical observation that hapaxes are present in every corpora, no matter how large.

It may therefore be claimed that the premises of the theorem in some sense include what we set out to prove (which is of course true of every theorem) and certainly in this light the conclusion that vocabulary size *must be* infinite is less surprising. In fact a weaker bound can already be derived from $g(1/V(N)) \geq p_1$, knowing $g(x) = \exp(-Dx)$ and $D = \log(N)$. Since $\exp(-\log(N)/V(N)) \geq p_1$ we have $V(N) \geq \log(N)/\log(1/p_1)$, an estimate that is weakest for small p_1.

The most novel of our assumptions is (3), and it is also the empirically richest one. For any exponent D, exponentially decaying frequencies would satisfy the following *geometric mean property:*

> if r and s are arbitrary ranks, and their (weighted) arithmetic mean is t, the frequency at t is the (weighted) geometric mean of the frequencies at r and s.

What we find in frequency count data is the *subgeometric mean property,* namely that frequency observed at the arithmetic mean of ranks is systematically *lower* than frequency computed as the geometric mean, i.e. that decay is *slower* than exponential.

This may not be strictly true for very frequent items (a concern we will address in Section 2) and will of necessity fail at some points in the low frequency range, where effects stemming from the resolution of the corpus (i.e. that the smallest gap between frequency ratios cannot be smaller than $1/N$) become noticeable: if the rth word has i tokens but the $(r+1)$th word has only $i - 1$ tokens, we can be virtually certain that their theoretical probabilities (as opposed to the observed frequency ratios) differ less than by $1/N$. At such "steps" in the curve, we cannot expect the geometric mean property to hold: the observed frequency of the rth word, i/N, is actually higher than the frequency computed as the geometric mean of the frequency of e.g. the $(r-1)$th and $(r+1)$th words, which will be $\sqrt{i(i-1)}/N$. To protect our Theorem 1 from this effect, we could estimate the area under the curve by segregating the steps up to $\log(\log(N))$ from the rest of the curve by two-sided intervals of length N^{ϵ}, but we will not present the details here because $\log(N)$ is only a lower bound on vocabulary size, and as a practical matter, not a very good one.

In fact, if we repeatedly double the size of our Merc corpus to include 1,2,...,128 issues, and plot log vocabulary size against log sample size we get a very good linear relationship (see Fig. 3), indicating that $V(N) \sim N^q$, with $q \approx 0.75$. A similar "power law" relationship has been observed in closed corpora (including several Shakespeare plays) by Turner (1997). The assumption that a power law relates vocabulary size to sample size goes back at least to Guiraud (1954) (with $q = 0.5$) and Herdan (1960) – the only novelty in our approach is that we will derive this power law as a consequence of Zipf's second law in Theorem 2 later.

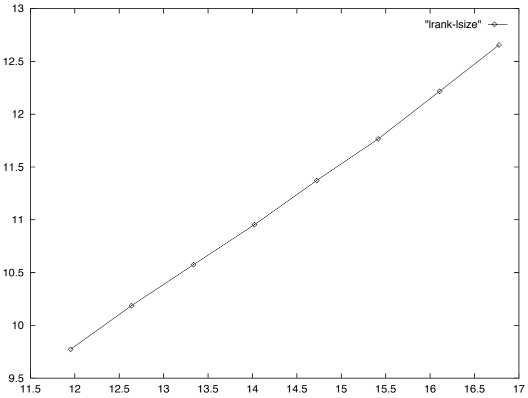

Figure 3: Growth of vocabulary size *V(N)* against corpus size *N* in the Merc on log-log scale

The lesson that we would like to take away from Theorem 1 is not the quantitative form of the relationship $V(N) \geq \log(N)(1 - 1/N)$, since this is a rather weak lower bound, but the qualitative fact that vocabulary size grows in an unbounded fashion when sample size is increased. Less than logarithmic growth is logically inconsistent with the characteristic properties of corpora, namely their subexponential decay and that singletons (hapaxes) are present in every sample, no matter how large.

2 High frequency items

Implicitly or explicitly, much of the work concerning word frequency assumes a Bernoulli-style setup, in which words (tokens) are randomly drawn (with replacement) from a large urn containing all word types in fixed proportions. Though clearly not intended as a psychologically realistic model of speech or writing, it is nevertheless a very useful model, and rather than abandoning it entirely, our goal here is to refine it to fit the facts better. In particular, we follow Mandelbrot's (1961c) lead in assuming that there are *two* urns, a small one U_F for function words, and a larger one U_C for content words.

Inasmuch as the placement of function words is dictated by the rules of syntax rather than by efforts to choose the semantically appropriate term,[1] it seems appropriate to set aside function words in U_F. Also, the use of function words is subject to so much individual variation that principal component analysis on the function word counts is effective in separating different authors (Burrows, 1987).

Our first task is to estimate the relative sizes of the two urns. Let $f_N(x)$ be a family of $[0,1] \rightarrow [0,1]$ functions with the following properties:

(4) exponential decay, $f_N(x) = \exp(-D_N x)$

(5) left limit, $f_N(1/N)$ is a constant, say $\exp(-c)$

(6) linear area law, $\int_{1/2N}^{(V(N)+1/2)/N} f_N(x)dx = 1/N$

To fix ideas, the f_N should be thought of as normalized frequency distributions, but the x axis is scaled by N rather than $V(N)$ as before: values of f_N for $x > V(N)/N$ are simply 0. Also, we think of the values $f_N(r/N)$ as providing the ordinate for trapezoidal sums approximating the integrals, rather than the rectangular sums used in Section 1. Since the width of the trapezoids is $1/N$ and their height sums to 1, the trapezoidal sum is $1/N$ rather than $1/V(N)$ as before.

From (4) and (5) we get $D_N = cN$, which for (6) gives $1/N = \int_{1/2N}^{(V(N)+1/2)/N} \exp(-cNx)dx = (1/cN)[\exp(-c/2) - \exp(-c(V(N) + 1/2))]$. Since $V(N) \rightarrow \infty$ as $N \rightarrow \infty$, the last term can be neglected and we get $c = \exp(-c/2)$. Numerically,

[1] The same point can be made with respect to other Pareto-Zipf laws. For example, in the case of city sizes it stands to reason that the growth of a big city like New York is primarily affected by local zoning laws and ordinances, the pattern of local, state, and federal taxes, demographic and economic trends in the region, and immigration patterns: the zoning laws etc. that affect Bombay are almost entirely irrelevant to the growth of New York. But once we move to mid-sized and small population centers, the general spatial patterns of human settlement can be expected to assert themselves over the special circumstances relevant to big cities.

this yields $c = 0.7035$ meaning the frequency of the most frequent item is 49.4866%.

While our argument is clearly heuristic, it strongly suggests that nearly half of the tokens may come from function words i.e. the two urns are roughly the same size. An alternative to using two separate urns may be to tokenize every function word as an instance of a catchall 'functionword' type. The standard list in Vol 3 of Knuth (1971) contains 31 words said to cover 36% of English text, the 150 most frequent used in Unix covers approximately 40% of newspaper text, and to reach 49.5% coverage on the Merc we need less than 200 words. By grouping the appropriate number of function words together we can have the probability of the dominant type approximate 49.5%.

Because tokenization is from the statistical perspective arbitrary, but reorganization of the data in the tokenization step (using only finite resources) is often desirable, it should be emphasized that at the high end we cannot in general expect Zipf-like regularity, or any other regularity. For example, Fig. 1 completely fails to show the linear pattern predicted by Zipf's law, and because of the multiple inflection points, fitting other smooth curves is also problematic at the high end. The geometric mean property is also likely to fail for very high frequency items, but this does not affect our conclusions, since the proof can be carried through on U_C alone, either by segregating function words in U_F or by collecting them in a single functionword type that is added to U_C.

Since some function words like *on* are homographic to content words (e.g. in *The cat is on the mat* we see the locative meaning of *on* rather than the purely prepositional one as in *go on* 'continue') ideally U_C should also contain some function word homographs, albeit with different probabilities. It would require sophisticated sense disambiguation to reinterpret the frequency counts this way, and we make no further efforts in this direction here, but note that because of this phenomenon the use of two separate urns need not result in a perceptible break in the plots even if the functional wordsenses are governed by laws totally different from the laws governing the contentful wordsenses.

It will be evident from Table 1 below that in the Merc no such break is found, and as long as markup strings are `lex`ed out just as punctuation, the same is true of most machine readable material. Several explanations have been put forth, including the notion that elements of a vocabulary "collaborate", but we believe that the smooth interpenetration of functional and contentful wordsenses, familiar to all practicing linguists and lexicographers, is sufficient to explain the phenomenon. Be it as it may, in the rest of this Section we assume the existence of some rank boundary $k, (30 < k < 200)$ such that all words in $1 \leq r \leq k$ are function words and all words with $r > k$ are content words. As we shall show shortly, the actual choice of k does not affect our argument in a material way.

We assume that the function words have a total probability mass $P_k = \sum_{r=1}^{k} p_r (0.3 \leq P_k \leq 0.5)$ and that Zipf's law is really a statement about U_C. Normalizing for the unit square, now using $V(N)$ as our normalizing factor, sample frequencies are $f(x)(k/V(N) \leq x \leq 1)$. The following properties will always hold:

(7) right limit, $f_N(1) = 1/N$

(8) left limit, $f_N(k/V(N))$ is a constant

(9) area power law, $\int_{k/V(N)}^{1} f_N(x)dx = (1 - P_k)/V(N)$

To this we can provisionally add Zipf's law, $\log(f_N(xV(N))) = C_N - B_N \log(xV(N))$ or more directly

$$f_N(xV(N)) = \exp(C_N - B_N \log(xV(N))) \qquad (10)$$

Condition (7) means $f(1) = \exp(C_N) = 1/N$ therefore $C_N = -\log(N)$. The logarithmic change in C_N corresponds to the fact that as corpus size grows, unnormalized Zipf plots shift further to the right – notice that this is independent of any assumption about the rate of vocabulary growth. In fact, if we use Zipf's law as a premise, we can state that vocabulary grows with a power of corpus size as

Theorem 2. If corpora satisfy Zipf's law, grow such that assumptions (7-8) above hold, and B_N tends to a fixed Zipf's constant B, vocabulary size $V(N)$ must grow with $N^q, q = 1/B$.

Proof. By (7) we have Zipf's law in the form $f_N(x) = 1/Nx^{B_N}$. If $f_N(k/V(N))$ is to stay constant as N grows, $N(k/V(N))^{B_N}$ must be constant. Since k (the number of function words) is assumed to be constant, we get $\log(N) + B_N \log(k) - B_N \log(V(N))$ constant, and as B_N converges to B, $\log(N) \sim B \log(V(N))$. Therefore, $N = V(N)^B$ within a constant factor.

In our notation, $q = 1/B$, and as $V(N) \leq N$, we obtained as a side result that frequency distributions with $B < 1$ are sampling artifacts in the sense that larger samples from the same population will, of necessity, have a B parameter ≥ 1. Thus we find Mandelbrot (1961b) to be completely vindicated when he writes

> [...] Zipf's values for B are grossly underestimated, as compared with values obtained when the first few most frequent words are disregarded. As a result, Zipf finds that the observed values of B are close to 1 or even less than 1, while we find that the values of B are not less than 1 [...] (p196)

We leave the special case $B = 1$ for Section 3, and conclude our investigation of high frequency items with the following remark. Equation (9), what we called the "power law" for area under the curve, gives, for $B > 1, (1 - P_k)/N^q = \int_{k/N^q}^{1} 1/(Nx^B)dx = [1 - (k/N^q)^{1-B}]/N(1 - B)$. Differentiating with respect to $k = xN^q$ gives $\partial P_k/\partial k = k^{-B}$ meaning that at the boundary between content words and function words we expect $p_k \sim 1/k^B$. Looking at four function words in the Merc in the range where we would like to place the boundary, Table 1 summarizes the results.

Word	Rank	Frequency	B
be	30	0.0035	1.66
had	75	0.0019	1.45
other	140	0.0012	1.36
me	220	0.00051	1.41

Table 1: $B = -\log(p_k)/log(k)$ (estimates)

Our goal here is not to compute B on the basis of estimated ranks and frequencies of a few function words, but rather to show that a smooth fit can be made at the function word boundary k. The proper procedure is to compute B on the basis of fitting the mid- (and possibly the low-) frequency data, and select a k such that the transition is smooth. As the chart shows, our normalization procedure is consistent with a wide range of choices for k.

3 Low frequency items

The fundamental empirical observation about low frequency items is also due to Zipf – it is sometimes referred to as his "second law" or the *number-frequency law*. Let us denote the number of singletons in a sample by c_1, the number of types with exactly 2 tokens by c_2 etc. Zipf's second low states that if we plot $\log(n)$ against $\log(c_n)$ we get a linear curve with slope close to -1/2. This is illustrated in Fig. 4 below:

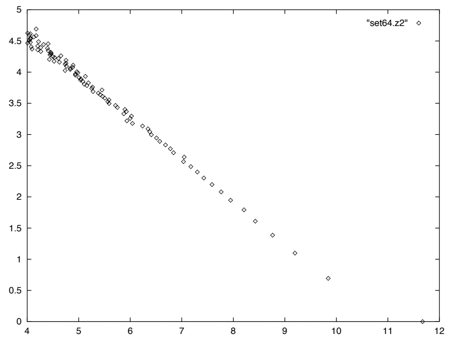

Figure 4: Num-freq law on the Merc (10M words)

Some of the literature (e.g. the excellent web article by Landini (1997)) treats these as separate laws, but really the "second law", $\log(i) = H_N - D_N \log(c_i)$, is a straightforward consequence of the first, as Zipf already argued more heuristically.

Theorem 3. If a distribution obeys Zipf's first law with slope parameter B it will obey Zipf's second law with slope parameter $D = B/(1 + B)$.

Proof. Using the notation established above, for sample size N we have $f_N(x) = 1/Nx^B$, so the probability that an item is between i/N and $(i + 1)/N$ if $i \leq x^{-B} \leq i + 1$. Therefore we expect $c_i = V(N)(i^{-q} - (i + 1)^{-q})$. By Rolle's theorem, the second term is qy^{-q-1} for some $i \leq y \leq i + 1$. Therefore, $\log(c_i)/(q + 1) = \log(V(N))/(q + 1) - \log(q)/(q + 1) - \log(y)$. Since $\log(q)/(q + 1)$ is a small constant, and $\log(y)$ can differ from $\log(i)$ by no more than $\log(2)$, rearranging the terms we get $\log(i) = \log(V(N))/(q + 1) - \log(c_i)/(q + 1)$. Since $H_N = \log(V(N))/(1 + q)$ tends to infinity, we can use it to absorb the constant term bounded by $(q - 1)/2 + \log(2)$.

Discussion. The normalization term H_N is necessitated by the fact that "second law" plots would otherwise show the same drift as "first law" plots. Using this term we can state the second law in a much more useful format. Since $\log(i) = \log(V(N))/(q + 1) - \log(c_i)/(q + 1)$ plus some additive constant,

$$c_i = V(N)/i^{q+1}$$

times some multiplicative constant m. If we wish $\sum_{i=1}^{\infty} c_i = V(N)$ to hold we must choose m to be $1/\zeta(q + 1)$. Since this argument assumes Zipf's second law to extend well to high frequency items, the case for using $m = 1/\zeta(q + 1)$ is not totally compelling, but it is reassuring to see that for $B \geq 1$ we always find a bound constant ($6/\pi^2$ for $B = 1$) that will make the distribution consistent.

Therefore we find Mandelbrot's (1961c) criticism of $B = 1$ to be somewhat less compelling than the case he made against $B < 1$. Recall from the preceding that B is the reciprocal of the exponent q in the vocabulary growth formula $V(N) = N^q$. If we choose a very "rich" corpus, e.g. a table of logarithms, virtually every word will be unique, and $V(N)$ will grow faster than $N^{1-\epsilon}$ for any $\epsilon > 0$, so B must be 1. The following example sheds some light on the matter.

Example 2. Let $L = \{0, 1, \ldots, 9\}$ and our word tokens be the integers (in standard decimal notation). Further, let two tokens share the same type if their smallest prime factors are the same. Our size N corpus is constructed by N drawings from the exponential distribution that assigns frequency 2^{-i} to the number i. It is easy to see that the token frequency will be $1/(2^p - 1)$ for p prime, 0 otherwise.

Therefore, our corpora will not satisfy Zipf's law, since the rank of the ith prime is i, but from the prime number theorem $p_i \sim i\log(i)$ and thus its log frequency $\sim -i\log(i)\log(2)$. However, the corpora will satisfy Zipf's second law, since, again from the prime number theorem, $c_i = N/i^2(\log(N) - \log(i))$ and thus $\log(V(N))/2 - \log(c_i)/2 = \log(N)/2 - \log(\log(N))/2 - \log(N)/2 + \log(i) + \log(\log(N) - \log(i))/2$ which is indeed $\log(i)$ within $1/\log(N)$.

Example 2 shows that Theorem 3 can not be reversed without additional conditions (such as $B > 1$). A purist might object that the definition of token/type relation is weird. However, it is just an artifact of the Arabic system of numerals that the smallest prime in a number is not evident: if we used the canonical form of numbers, everything after the first prime could simply be discarded as mere punctuation. More importantly, there are several standard families of distributions that can, when conditions are set up right, satisfy the second law but not the first one with any $B > 1$.

Yule (1922) and Simon (1955) explored variants of the beta distribution in this context. This is not the place to give a full appraisal of the Simon-Mandelbrot debate (Mandelbrot 1959 1961a 1961b, Simon 1960 1961a 1961b), but it seems clear to us that the use of such models can hardly be faulted on grounds of mathematical rigor. It can, however, be faulted on grounds of empirical validity. One example used in Simon (1955) and subsequent work is Joyce's *Ulysses.* The general claim of $V(N) = \alpha N$ is made for *Ulysses* with $\alpha \approx 0.115$. However, instead of linear vocabulary growth, in Ulysses we find the same power law that we have seen in the Merc (cf. Fig. 3 above). To be sure, the exponent q is closer to 0.9, while in the Merc it was 0.75, but it is still very far from 1. Leaving out the two longest chapters *Oxen of the Sun* and *Circe* we are left with roughly two-thirds of Ulysses, yielding an estimate of $\alpha = 0.111$ or $q = 0.822$. Applying these to the two chapters left out, which have 96268 words total, we can compute the number of different words based on αN, which yields 10723, or based on N^q, which yields 12422. The actual number of words used in these two chapters is 13448, so the error of the power law estimate is 8.5% versus the 25.5% error of the linear estimate.

Tweedie and Baayen (1998) survey a range of formulas relating $V(N)$ to N, and identify our q as Herdan's C. If we are satisfied with Zipf's Law at least as a first approximation to the empirically observable frequency distribution, clearly $C = q = 1/B$, where $B > 1$ is the Zipfian parameter. In light of our results so far, type token ratio *must* tend to zero since $V(N)/N \sim N^{q-1}$ and $q = 1/B < 1$. Guiraud's R will tend to zero or infinity if $B < 2$ or $B > 2$ respectively. Dugast's and Rubet's k, defined by Tweedie and Baayen as $\log(V(N))/\log(\log(N))$, must tend to infinity. From the second law, the ratio of hapax legomena to vocabulary size $c_1/V(N)$ is a constant m, the ratio of dis legomena to vocabulary size is a different constant $c_2/V(N) = m/2^{q+1}$, and in general $c_i/V(N)$ is m/i^{q+1}. On the whole we expect better estimates of m from dis legomena than from hapaxes, since the latter serve as a grab-bag for typos, large numerals, and other marginal phenomena. In light of these simple asymptotic considerations it comes as no surprise that most of the "lexical richness" measures discussed by Tweedie and Baayen are not constant. From the Zipfian vantage point it is also clear that Yule's K, which is essentially $\sum_{i=1}^{\infty} p_i^2$, entropy, given by $\sum_{i=1}^{\infty} -p_i\log(p_i)$, and indeed all of Good's (1953) spectral measures with $Bt > 1$ are converging to constant values, as sample size increases without bounds.

Our results therefore cast those of Tweedie and Baayen in a slightly different light: some of the measures they investigate are truly useless (divergent or converging to the same constant independent of the Zipfian parameter B) while others are at least in principle useful, though in practice estimating them from small samples may be highly problematic. In many cases, the relationship between a purely Zipfian distribution with parameter B and a proposed measure of lexical richness such as K is given by a rather complex analytic relation (in this case, $K = \zeta(2B)/\zeta(B)$) and even this relation can be completely obscured if effects of the high-frequency function words are not controlled carefully. This important methodological point, made very explicitly in Mandelbrot's early work, is worth reiterating, especially as there are still a large number of papers (see Naranan and Balasubrahmanyan (1993) for a recent example) which treat the closed and the open vocabulary cases as analogous.

Finally, let us consider another class of distributions that has considerable support in the literature, the *lognormal* family (see e.g. Carroll 1967). Here the problem is in the opposite direction: while the beta distribution assumes there to be too many different words, lognormal would require there to be too few. Theorem 1 proves that under reasonably broad conditions $V(N) \to \infty$, meaning that the average frequency, $1/V(N)$, will tend to zero as sample size increases. But if average frequency tends to zero, average log frequency will diverge. In fact, using Zipf's second law we can estimate it to be $-\log(N)$ within an additive constant R. As a simple "left limit" argument shows, the variance of log frequencies also diverges with $\sqrt{B\log(N)/2}$. To see this, we need to first estimate $f_N^{'}(k/V(N))$, because the functional equation for lognormal distribution,

$$f_N^2(x) = \frac{-f_N^{'}(x)}{\sqrt{2\pi}} \exp(\frac{-1}{2}\frac{(\log(f_N(x)) - \mu_N)^2}{\sigma_N^2})$$

contains this term. Using the difference quotient we obtain $p_{k+1} - p_k/V(N)$, and we have $V(N) = N^q$ for some constant $q < 1$. By Zipf's law $\log(f_N(x)) = -\log(N) - B\log(x)$. Using (8) we get that

$$\frac{1/V(N)}{\sqrt{2\pi}} \exp(\frac{-1}{2}\frac{(-Bq\log(N))^2}{\sigma_N^2})$$

is constant, which can hold only if $q\log(N) = (1/2)\log(N)^2/\sigma_N^2$ i.e if $(B/2)\log(N) = \sigma_N^2$. In other words, the lognormal hypothesis does not lead to a stable limiting distribution: the means drift down with $\log(1/N)$ and, more importantly, the variances open up with $\sqrt{\log(N)}$. Another way of putting our result is that a lognormal that fits the Zipfian midrange $N^\epsilon < i < N^{q-\epsilon}$ well can never fit the low range well, if the latter satisfies Zipf's second law.

4 Summary and conclusions

In this paper we inspected Zipf's law separately for the high-, mid-, and low-frequency ranges. For the high-frequency range we proposed that a separate urn, containing only a few dozen to a few hundred function words, be used, and argued that this urn will contain somewhere between 30% and 50% of the total probability mass.

For the mid- and low-frequency range we noted the following *subgeometric mean property*: for ranks r and s, the observed frequency $f(\frac{r+s}{2})$ is less than the geometric mean of $f(r)$ and $f(s)$. Using this property and a simple normalization technique we proved in Theorem 1 that vocabulary size $V(N)$ tends to infinity as $N \rightarrow \infty$.

It is in the middle range that Zipf's law appears strongest, and here estimates of the Zipf constant B clearly give $B > 1$ which corresponds, as we have shown in Theorem 2, to a vocabulary growth rate $V(N) = N^{\frac{1}{B}}$. We could use this theorem to give a simple proof that a mixture of Zipfian distributions with constants $B_1, B_2, \ldots B_t$ will always be dominated by the smallest B_i, independent of the size of the mixture weights.

Because distributions that satisfy Zipf's first law in the the mid- and the low-frequency range will also satisfy "Zipf's second law" in the low-frequency range (Theorem 3), there seems to be no compelling need for a separate urn in the low frequency range, and we have not endeavored to introduce one, particularly as the B of this urn, were it lower than the B of the mid-frequency urn, would dominate the whole distribution for large N.

Altogether, there appears to be considerable empirical support for the classical Zipfian distribution with $B > 1$, both in the Merc and in standard closed corpora such as *Ulysses.* There seems to be no way, empirical or theoretical, to avoid the conclusion that vocabulary size grows with a power $q < 1$ of N, and competing hypotheses, in particular the lognormal, are not well suited for characterizing distributions that satisfy this power law.

Acknowledgement

The author would like to thank Gabriel Landini and David M.W. Powers for valuable suggestions and discussion of earlier drafts of this paper.

References

J.F Burrows. 1987. Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2:61–70.

John B. Carroll. 1967. On sampling from a lognormal model of word-frequency distribution. In H. Kucera and W.N. Francis, editors, *Computational Analysis of Present-Day American English*, pages 406–424. Brown University Press, Providence, RI.

Harald Cramér. 1955. *The elements of probability theory.* John Wiley & Sons, New York.

J.R. Estoup. 1916. *Gammes Stenographiques.* Institut Stenographique de France, Paris.

I.J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

H. Guiraud. 1954. *Les charactères statistiques du vocabulaire.* Presses Universitaires de France.

Gustav Herdan. 1960. *Type-Token Mathematics.* Mouton.

Donald E. Knuth. 1971. *The Art of Computer Programming.* Addison-Wesley.

Gabriel Landini. 1997. Zipf's laws in the Voynich Manuscript. *http://web.bham.ac.uk /G.Landini/evmt/zipf.htm.*

Benoit Mandelbrot. 1959. A note on a class of skew distribution functions. analysis and critique of a paper by H.A. Simon. *Information and Control*, 2:90–99.

Benoit Mandelbrot. 1961a. Final note on a class of skew distribution functions: analysis and critique of a model due to Herbert A. Simon. *Information and Control*, 4:198–216.

Benoit Mandelbrot. 1961b. On the thory of word frequencies and on related markovian models of discourse. In Roman Jakobson, editor, *Structure of language and its mathematical aspects*, pages 190–219. American Mathematical Society.

Benoit Mandelbrot. 1961c. Post scriptum to 'final note'. *Information and Control*, 4:300–304.

S. Naranan and V.K. Balasubrahmanyan. 1993. Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, 52:728–738.

Vilfredo Pareto. 1897. *Cours d'economie politique.* Rouge.

David M.W. Powers. 1998. Applications and explanations of Zipf's law. In D.M.W. Powers, editor, *NEMLAP3/CONLL98: New methods in language processing and Computational natural langyuage learning*, pages 151–160. ACL.

Christer Samuelsson. 1996. Relating Turing's formula and Zipf's law. *Proc. Fourth Workshop on Very Large Corpora.*

Herbert A. Simon. 1955. On a class of skew distribution functions. *Biometrika*, 42:425–440.

Herbert A. Simon. 1960. Some further notes on a class of skew distribution functions. *Information and Control*, 3:80–88.

Herbert A. Simon. 1961a. Reply to dr. Mandelbrot's post scriptum. *Information and Control*, 4:305–308.

Herbert A. Simon. 1961b. Reply to 'final note' by benoit Mandelbrot. *Information and Control*, 4:217–223.

Geoffry R. Turner. 1997. Relationship between vocabulary, text length and Zipf's law. *http://www.btinternet.com/ g.r.turner/ZipfDoc.htm.*

Fiona J. Tweedie and R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352.

J.C. Willis. 1922. *Age and area.* Cambridge University Press.

G. Udny Yule. 1922. A mathematical theory of evolution. *Phil. Trans. Roy. Soc.*, B213:21ff.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley.