Glottometrics 4, 2002, 61-86 *To honor G.K. Zipf*

How many words are there?

András Kornai[1]

Abstract. The commonsensical assumption that any language has only finitely many words is shown to be false by a combination of formal and empirical arguments. Zipf's Law and related formulas are investigated and a more complex model is offered.

Keywords: Vocabulary size, Zipf's law

1. Introduction

Ask the lay person how many different words are there, and you are likely to receive a surprisingly uniform set of answers. English has more words than any other language. There are over a hundred thousand words in unabridged dictionaries. The *OED* has over three hundred thousand, but there are still a few words missing. A five year old knows five thousand words, an adult uses fifteen thousand. Shakespeare used thirty thousand, but Joyce used even more. Some primitive tribes get along with only a few hundred words. Eskimo has seventeen words for snow. Introductory linguistics courses invariably spend a great deal of energy on rebutting these and similar commonly known "facts" about languages in general and English in particular. But the basic fallacy, what we will call the *closed vocabulary* assumption, that there is a fixed number of words S in any given language, is often present even in highly technical work otherwise based on a sophisticated understanding of language. *Open vocabulary*, that words in a language form a denumerably infinite set, is a standard assumption in generative linguistics, where it is justified by pointing at productive morphological processes such as compounding and various kinds of affixation. Yet somehow the existence of such processes generally fails to impress those with more of an engineering mindset, chiefly because the recursive aspect of these processes is weak – the probability of iterated rule application decreases exponentially with the number of iterations.

In this paper we offer a new quantitative argument why vocabulary must be treated as open. We investigate vocabulary size not as an isolated number, but rather as part of the broader task of trying to estimate the frequency of words. The rest of this Introduction establishes the terminology and notation and surveys the literature. Section 2 disposes of some widely used arguments in favor of closed vocabulary by means of counterexamples and introduces the *subgeometric mean property* that will play a crucial role in the subsequent analysis of vocabulary size. Section 3 explores the regions of extremely high and extremely low frequencies, where the basic regularity governing word frequencies, Zipf's Law, is known to fail. Section 4 investigates some widely used alternatives to Zipf's Law, including the beta, lognormal, Waring, and negative binomial distributions, and shows why most of these are inferior to Zipf's Law. We offer our conclusions in Section 5.

[1] Address correspondence to: Andras Kornai, Metacarta Inc. 126 Prospect St, Cambridge, MA 02139 USA. E-mail: andras@kornai.com

1.1. Zipf's Laws

It is far from trivial to define words in spoken or signed language, but in this paper we can steer clear of these difficulties by assuming some conventional orthography or linguistic transcription system that has one to one correspondence between orthographic words (maximum non-white-space non-punctuation strings) and prosodic words. Because a large variety of transcription systems exist, no generality is lost by restricting our attention to text that has already been rendered machine readable. For the sake of concreteness we will assume 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. Using these or similar definitions, counting the number of tokens belonging in the same type becomes a mechanical task. The results of such *word counts* can be used for a variety of purposes, such as the design of more efficient codes, typology, investigations of style, authorship, language development, and statistical language modeling in general.

Given a corpus Q of N word tokens, we find V different types, $V \leq N$. Let us denote the *absolute frequency* (number of tokens) for a type w by $F_Q(w)$, and the *relative frequency* $F_Q(w)/N$ by $f_Q(w)$. Arranging the w in order of decreasing frequency, the rth type (w_r) is said to have *rank* r, and its relative frequency $f_Q(w_r)$ will also be written f_r. As Estoup (1916) and Zipf (1935) noted, 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.

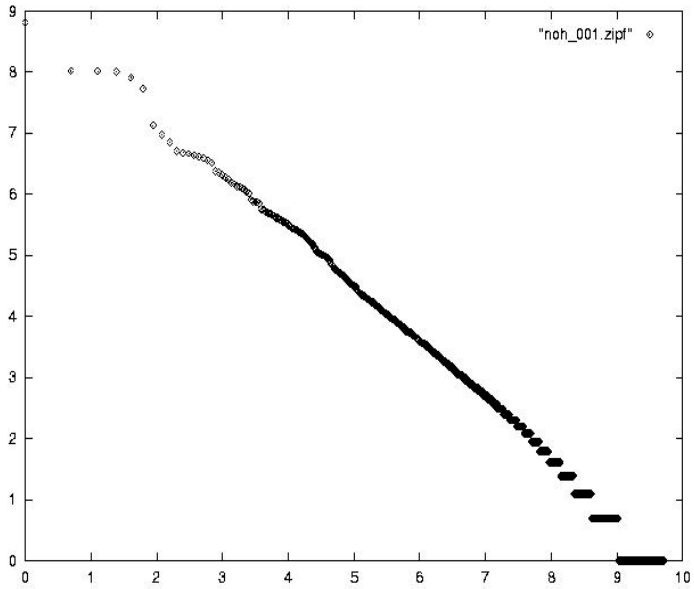

Fig. 1. Plot of log frequency as a function of log rank
for a newspaper issue (150k words)

Denoting the slope of the linear portion by $-B$, B is close to unity, slightly higher on some plots, slightly lower on others. Some authors, like Samuelsson (1996), reserve the term "Zipf's Law" to the case $B = 1$, but in this paper we use more permissive language, since part of our goal is to determine how to formulate this regularity properly. As a first approximation, Zipf's Law can be formulated as

(1) $\log(f_r) = H_N - B_N \log(r)$

where H_N is some constant (possibly dependent on our random sample Q and thus on N, but independent of r). When this formula is used to fit a Zipfian curve to frequency data, with increased corpus size not only the intercept H_N but also the slope B_N will depend on the corpus size N. We reserve the term "Zipfian" to the case where B_N tends to a constant B as N tends to infinity, but do not assume in advance $B = 1$. (1) is closely related to, but not equivalent with, another regularity, often called Zipf's Second Law. Let $V(i,N)$ be the number of types that occur i times: Zipf's Second Law is usually stated as

(2) $\log(i) = K_N - D_N \log(V(i, N))$.

1.2. Background

As readers familiar with the literature will know, the status of Zipf's Law(s) is highly contentious, and the debate surrounding it is often conducted in a spectacularly acrimonious fashion. As an example, we quote here Herdan (1966:88):

> The Zipf law is the supposedly straight line relation between occurrence frequency of words in a language and their rank, if both are plotted logarithmically. Mathematicians believe in it because they think that linguists have established it to be a linguistic law, and linguists believe in it because they, on their part, think that mathematicians have established it to be a mathematical law. [...] Rightly seen, the Zipf law is nothing but the arbitrary arrangement of words in a text sample according to their frequency of occurrence. How could such an arbitrary and rather trivial ordering of words be believed to reveal the most recondite secrets, and the basic laws, of language?

We can divide the literature on the subject in two broad categories: empirical curve fitting and model genesis. The first category is by far the more voluminous, running to several thousand scholarly papers and hundreds of monographs. Here we do not even attempt to survey this literature: *QUALICO* conference volumes and the *Journal of Quantitative Linguistics* offer a good entry point. In mathematical statistics, attempts to discern the underlying mechanism that gives rise to a given distribution are called investigations of *model genesis* a particularly successful example is the explanation why normal distribution appears so often in seemingly unrelated areas, provided by the central limit theorems. Given the sheer bulk of the literature supporting some Zipf-like regularity in domains ranging from linguistic type/token counts to the distribution of wealth, it is natural that statisticians sought, and successfully identified, different mechanisms that can give rise to (1-2) or related laws.

The first results in this direction were obtained by Yule (1924), working on a version of (2) proposed in Willis (1922) to describe the number of species that belong to the same genus. Assuming a single ancestral species, a fixed annual probability s of a mutation that produces a new species, and a smaller probability g of a mutation that produces an entirely new genus, Yule shows that over time the distribution for the number of genera $V(i,N)$ with exactly i species will tend to

(3) $$\frac{c}{i^{1+g/s}}$$

which is the same as (2) with $D_N = s/(s+g)$ and $K_N = -s \log(\zeta(1+g/s))/(s+g)$ independent of N (the Riemann ζ function enters the picture only to keep probabilities summing to one).

This is not to say that words arise from a single undifferentiated ancestor by a process of mutation, but as Zipf already noted, the most frequent words tend to be the historically older ones, which also have the highest degree of polysemy. The essential point of Yule's work is that a simple, uniform process of mutation can give rise, over time, to the characteristically non-uniform 'Zipfian' distribution: merely by being around longer, older genera have more chance to develop more species, even without the benefit of a better than average mutation rate.

The same distribution has been observed in patterns of income by Pareto (1897), and there is again a large body of empirical literature supporting Zipf's Law (known in economics as Pareto's Law). Champernowne (originally in 1936, but not fully published until 1973) offered a model where the uneven distribution emerges from a stochastic process (Champernowne 1952, 1953, 1973, see also Cox and Miller 1965) with a barrier corresponding to minimum wealth.

Zipf himself attempted to search for a genesis in terms of a "principle of least effort", but his work (Zipf 1935, 1949) was never mathematically rigorous, and was cut short by his death. A mathematically more satisfying model specifically aimed at word frequencies was proposed by Simon (1955), who derived (2) from a model of text generation based on two hypotheses: (i) new words are introduced by a small constant probability, and (ii) old words are reused with the same probability that they had in earlier text.

A very different genesis result was obtained by Mandelbrot (1952) in terms of the classic "monkeys and typewriters" scenario. Let us designate an arbitrary symbol on the typewriter as a word boundary, and define "words" as maximum strings that do not contain it. If we assume that new symbols are generated randomly, Zipf's law can be derived for $B > 1$. Remarkably, the result holds true if we move from a simple Bernoulli experiment (zero order Markov process) to higher order Markov processes.

In terms of content, though perhaps not in terms of form, the high point of the Zipfian genesis literature is the Simon-Mandelbrot debate (Mandelbrot 1959, 1961a-c; Simon 1960, 1961a,b). Simon's genesis works equally well irrespective of whether we assume closed ($B < 1$) or open ($B \geq 1$) vocabulary. For Mandelbrot, the apparent flexibility in choosing any number close to 1 is a fatal weakness in Simon's model. While we will argue for open vocabulary, and thus side with Mandelbrot for the most part, we believe his critique of Simon to be too strict in the sense that explaining too much is not as fatal a flaw as explaining nothing. Ultimately, the general acceptance of Mandelbrot's genesis as the linguistically more revealing rests not on his attempted destruction of Simon's model but rather on the fact that we see his model as more assumption-free.

2. Exponential and subexponential decay

Arguments based on counting the frequency of various words and phrases are nothing new: in the 1640s a Swedish sect was deemed heretical (relative to Lutheran orthodoxy) on the basis of larger than expected frequency of forms such as *Christ bleeding, Christ suffering, Christ crucified* found in its Sion Psalmbook. With such a long tradition, predating the foundations of modern probability theory by centuries, it should come as no surprise that a considerable number of those employing word counts still reject the standard statistical view of corpora as samples from some underlying population. In particular, Zipf himself held that collecting more data about word frequency can sometimes distort the picture, and there is an "optimum corpus size". For a modern discussion and critique of this notion see Powers (1998), and for an attempt to recast it in a contemporary statistical framework see Baayen (2001:5.2), who

traces the method to papers published in the eighties by Orlov, Chitashvili, and Khmaladze (non vidi).

The central mathematical method of this paper is to make explicit the dependence of certain model parameters on corpus size N, and let N increase without bounds. Since this method only makes sense if we assume the standard apparatus of mathematical statistics and probability theory, in 2.1 we devote some time to defending the standard view. In 2.2 we introduce a simple normalization technique that makes frequency counts for different values of N directly comparable. In 2.3 we compare the normalized distributions to exponential decay, an unrealistic, but mathematically very tractable model. We introduce the more realistic subgeometric mean property in 2.4, and the empirically observable power law of vocabulary growth in 2.5.

2.1. Corpora as samples

Suppose 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. The issue we wish to address is whether this population is based on a finite (closed) vocabulary, or an infinite (open) one.

It is often argued that synchronic vocabularies are by definition closed, and only in a diachronic sense can vocabulary be considered open. New words enter the language at a perceptible rate, and the Merc shows this effect as well as any other continually growing corpus. But this process is considerably slower than the temporal fluctuations in word frequency occasioned by certain geographic locations, personages, or products getting in the news, and is, at least to some extent, offset by the opposite process of words gradually falling into disuse. What we wish to demonstrate is openness in the synchronic sense, and we will not make any use of the continually growing nature of the Merc corpus in doing so. In fact, we shall argue that even historically closed corpora, such as Joyce's *Ulysses*, offer evidence of being based on an open vocabulary (see 4.1). This of course makes sense only if we are willing to go beyond the view that the words in *Ulysses* comprise the entire statistical population, and view them instead as a sample of Joyce's writing, or even more narrowly, as a sample of Joyce's writing books like *Ulysses*. It is rather unlikely that a manuscript for a sequel to *Ulysses* will some day surface, but the possibility can not be ruled out entirely, and it is only predictions about unseen material that can lend support to any model. The Merc is better for our purposes only because we can be reasonably certain that there will be future issues to check our predictions against.

The empirical foundation of probabilistic arguments is what standard textbooks like Cramér (1955) call the *stability property of frequency ratios:* for any word w, by randomly increasing the sample Q without bounds, $f_Q(w) = F_Q(w)/N$ tends to some $f(w)$ as N tends to infinity. In other words, sample frequencies must 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, the stability of frequency ratios has often been questioned on the basis of the following argument. If vocabulary is not closed, the pie must be cut into more and more slices as sample size is increased, and therefore the relative frequency of a word must, on average, 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 $f(w) = 1/2^r$.

Example 1. If the rth word has probability $p_r = 2^{-r}$, in a random sample Q of size $N = 2^m$ we expect 2^{m-1} tokens of w_1, 2^{m-2} tokens of w_2, ..., 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 $1/N$. Convergence is therefore 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 on the average decline of probabilities could be carried to the point of demonstrating that closed vocabulary is logically necessary. Though not necessary, closed vocabulary is still possible, and it is easy to construct examples, e.g. by using phonemes or syllables instead of words. What we will demonstrate in Theorem 1 is that closed vocabulary is logically incompatible with *observable* properties of word counts.

2.2. Normalization

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 open or closed. 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 scale may well be approximately linear, just as Herdan, quoted above, suggests. 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.

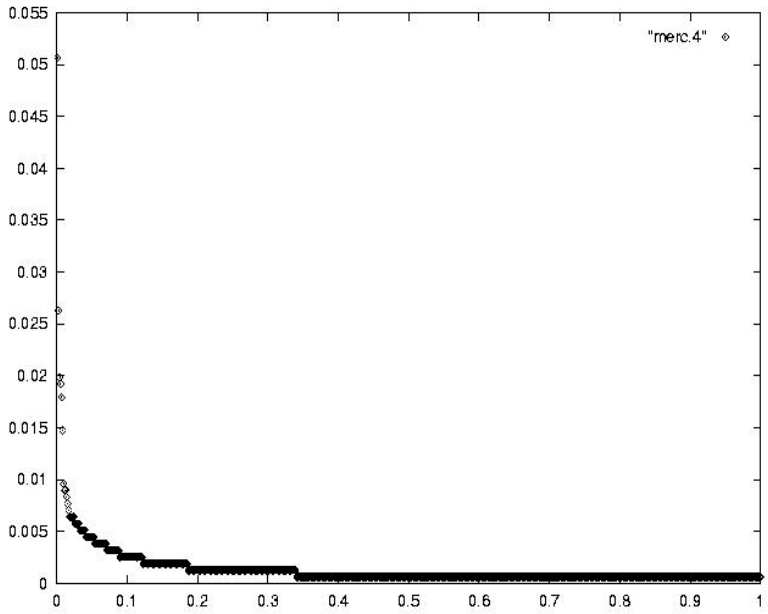

Fig. 2. Frequency as a function of normalized rank (4 articles, 1.5k words)

2.3. Exponential decay

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:

(F1) For any f obtained from a random sample Q of size N, $f(1/V(N))$ tends to p_1, the frequency of the most frequent item, as $N \to \infty$
(F2) For any f obtained from a random sample Q of size N, $f(1) = 1/N$
(F3) Word frequencies decay subexponentially (slower than $\exp(-Dx)$ for any $D > 0$).

Theorem 1. Under conditions (F1-F3) $V(N)$ 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 (F2) we compute $D = \log(N)$, and therefore $V(N) \geq \log(N)(1-1/N)$.

Discussion. Since any theorem is just as good as its premises, let us look at the conditions in some detail. (F1) 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 (see 2.1), there really is no alternative: if frequency ratios can't be expected to

stabilize even for the most frequent word, there is nothing we can hope to accomplish by measuring them. On the surface (F2) 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 (e.g. phoneme or grapheme counts) we would expect every item to appear at least twice as soon as the sample size is twice the inverse probability of the least frequent item. In the final analysis, (F2) rests on the massively supported empirical observation that hapaxes are present in every corpora, no matter how large (see 3.4).

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 *must be* open 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. Baayen (2001: 49) notes that a Turing-Good type estimate (Good 1953) can be used to approximate the rate at which the expected value of $V(N)$ changes by the left derivative $V(1,N)/N$, so that if hapaxes are always present we have $V'(N) \geq 1/N$, and by integrating both sides, $V(N) \geq \log(N)$. While the heuristic force of this simple argument is clear, it is not trivial to turn it into a rigorous proof, inasmuch as Turing-Good estimates are best viewed as Bayesian with a uniform prior over the (finite) set of types, see Nádas (1985).

2.4. The subgeometric mean property

The most novel of our assumptions is (F3), 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: for any $0 \leq p,q < 1$, $p + q = 1$ we find

$$f_{pr+qs} \leq f_r^p f_s^q . \tag{4}$$

In geometrical terms (4) means that $\log(f_r)$ is convex (viewed from below). This may not be strictly true for very frequent items (a concern we will address in 3.1) 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.

2.5. The power law of vocabulary growth

Empirically it seems quite incontestable that $V(N)$ grows with a power of N:

(5) $$V(N) = N^{\rho}$$

where $0 < \rho < 1$ is some constant, dependent on style, authorship, and other factors, but independent of N (Herdan 1964:157 denotes this constant by C). In practice, we almost always have $\rho > 0.4$, but even for smaller positive ρ the empirical power law would still be stronger than the theoretical (logarithmic) lower bound establised in 2.3 above.

In what follows, 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.

In one experiment, we repeatedly doubled the size of our sample to include 1,2,...,128 issues. The samples were selected randomly at each step so as to protect our results against arguments based on diachronic drift, and each sample was kept disjoint from the previous ones. If we plot log vocabulary size against log sample size, this experiment shows a remarkably 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 some Shakespeare plays) by Turner (1997).

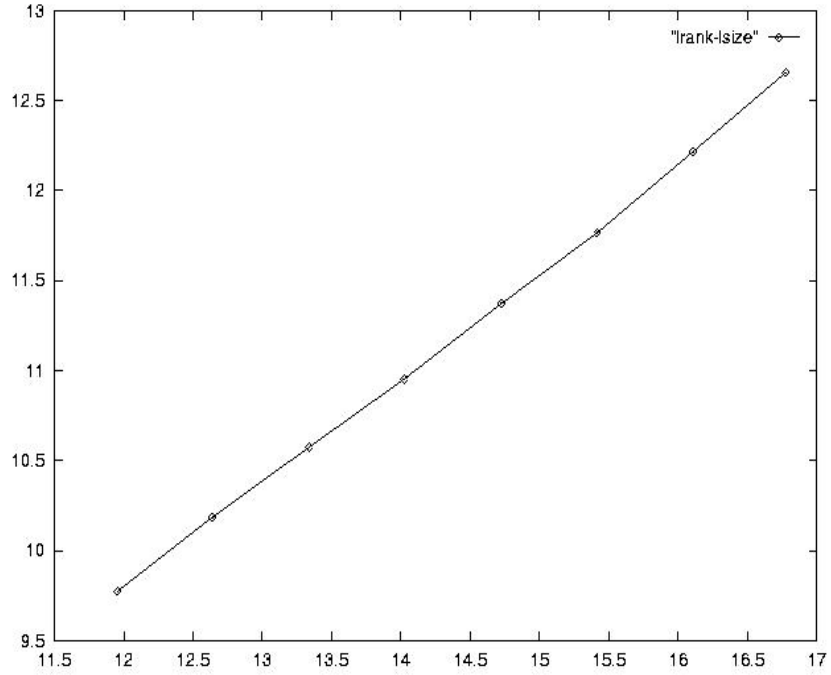

Fig. 3. Growth of vocabulary size $V(N)$ against corpus size N in the Merc on log-log scale

The assumption that a power law relates vocabulary size to sample size goes back at least to Guiraud (1954) (with $\rho = 0.5$) and Herdan (1960). The main novelty in our approach is that we need not postulate (5) as an empirical law, but will derive it as a consequence of Zipf's second law in 3.2. We should add here that (5) is just an approximate empirical law, and some

slower patterns of infinite growth such as $V(N) = N^{D/\log(\log(N))}$ would still look reasonably linear for $N < 10^{10}$ at log-log scale, and would be just as compatible with the observable data.

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 corpus, no matter how large. In fact, hapaxes are not only present, they comprise a significant portion h of the word types, a matter we shall return to in 3.4.

3. Frequency extremes

Implicitly or explicitly, much of the work concerning word frequency assumes a Bernoulli-style experimental 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. The reasons why high frequency items are expected to behave differently are discussed in 3.1, where the relative sizes of the two urns are estimated by a heuristic argument. In 3.2 we argue that there need be no perceptible break between the two urns, and show how the power law (5) can be derived from (1) using only trivial facts about U_C. A more rigorous treatment of low frequency items is given in 3.3, and "ultra-low" frequency items, *hapax legomena* and *dis legomena* are discussed in 3.4.

3.1. Function words vs. content words

Starting with Herdan (1960) it is common to set aside function words in U_F, since their placement is dictated by the rules of syntax rather than by efforts to choose the semantically appropriate term. The same point can be made with respect to other Zipf-like 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. (The issue is more complex for the largest concentrations of individual wealth, because the individuals will often diversify their holdings precisely in order to avoid disproportionate effects of laws and regulations affecting different sectors selectively. If true, this reasoning would suggest that Pareto's Law in economics suffers less from discrepancies at the high end than Zipf's Law in linguistics.) Another reason to set function words aside is that their use is subject to a great deal of idiosyncratic variation, so much so 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:

(U1) exponential decay, $f_N(x) = \exp(-D_N x)$
(U2) left limit, $f_N(1/N)$ is a constant, say $\exp(-c)$
(U3) 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 above. 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 (U1) and (U2) we get $D_N = cN$, which for (U3) gives

$$1/N = \int_{1/2N}^{(V(N)+1/2)/N} \exp(-cNx)dx = \frac{1}{cN}\left[\exp(-c/2) - \exp(-c(V(N)+1/2))\right].$$

Since $V(N) \to \infty$ as $N \to \infty$, the last term can be neglected and we get $c = \exp(-c/2)$. Numerically, 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%.

3.2. High frequency items

From the statistical perspective the tokenization process is arbitrary. We may wish to declare *THE, The*, and *the* to be tokens of the same type, or we may wish to keep them separate. We may declare *a* and *an* to be tokens of the same type, or, if we are so inclined, we may even declare *a, an*, and *the* to be tokens of the same ‘article' type. In French we may have good reasons to tokenize *du* as *de* + *le*, in English we may keep *a* + *priori* together as a single token. Because some 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. Furthermore, it has multiple inflection points, so 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.

In reality, there is no clear-cut boundary between function words and content words based on rank or other observable properties. In fact, many function words like *on* are homographic to content words: for example, 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'. To account for this, 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 lexed 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 3.2 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 $P_k = \sum_{r=1}^{k} p_r$, (to fix ideas, take $0.3 \leq P_k \leq 0.5$) and that Zipf's law is really a statement about U_C. Normalizing for the unit square, again using $V(N)$ as our normalizing factor, sample frequencies are $f(x)$, with $k/V(N) \leq x \leq 1$. The following properties will always hold:

(D1) right limit, $f_N(1) = 1/N$
(D2) left limit, $f_N(k/V(N))$ is a constant
(D3) area under the curve, $\int_{k/V(N)}^{1} f_N(x)dx = (1 - P_k)/V(N)$.

To this we can provisionally add Zipf's law in the form given in (1), or more directly

(6) $f_N(xV(N)) = \exp(H_N - B_N \log(xV(N)))$.

Condition (D1) means $f(1) = \exp(H_N) = 1/N$ therefore $H_N = -\log(N)$. The logarithmic change in H_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 (D1-D2) above hold, and B_N tends to a fixed Zipf's constant B, vocabulary size $V(N)$ must grow with N^{ρ}, $\rho = 1/B$.

Proof. By (D1) 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, $\rho = 1/B$, 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 1961c) 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 (p. 196).

We leave the special case $B = 1$ for 3.3, and conclude our investigation of high frequency items with the following remark. Condition (D3) gives, for $B > 1$, $(1 - P_k)/N^{\rho} = \int_{k/N^{\rho}}^{1} 1/(Nx^B)dx = [1 - (k/N^{\rho})^{1-B}]/N(1 - B)$. Differentiating with respect to $k = xN^{\rho}$ gives $\partial P_k / \partial k = k^{-B}$. Therefore 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.

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

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

The point 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 Table 1 shows, our normalization procedure is consistent with a wide range of choices for k.

3.3. 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 $V(1,N)$, the number of types with exactly 2 tokens by $V(2,N)$ etc. Zipf's second law states that if we plot $\log(i)$ against $\log(V(i,N))$ we get a linear curve with slope close to –1/2. This is illustrated in Fig. 4 below.

Some of the literature (e.g. the web article by Landini (1997)) treats (1) and (2) as separate laws, but really the "second law", $\log(i) = K_N - D_N \log(V(i,N))$, 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. 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 $i \le x^{-B} \le i+1$. Therefore we expect $V(i,N) = V(N)(i^{-\rho} - (i+1)^{-\rho})$. By Rolle's theorem, the second term is $\rho y^{-\rho-1}$ for some $i \le y \le i+1$. Therefore,

$$\log(V(i,N))/(\rho+1) = \log(V(N))/(\rho+1) - \log(\rho)/(\rho+1) - \log(y).$$

Since $\log(\rho)/(\rho+1)$ is a small constant, and $\log(y)$ can differ from $\log(i)$ by no more than $\log(2)$, rearranging the terms we get $\log(i) = \log(V(N))/(\rho+1) - \log(V(i,N))/(\rho+1)$. Since $K_N = \log(V(N))/(1+\rho)$ tends to infinity, we can use it to absorb the constant term bounded by $(\rho-1)/2 + \log(2)$.

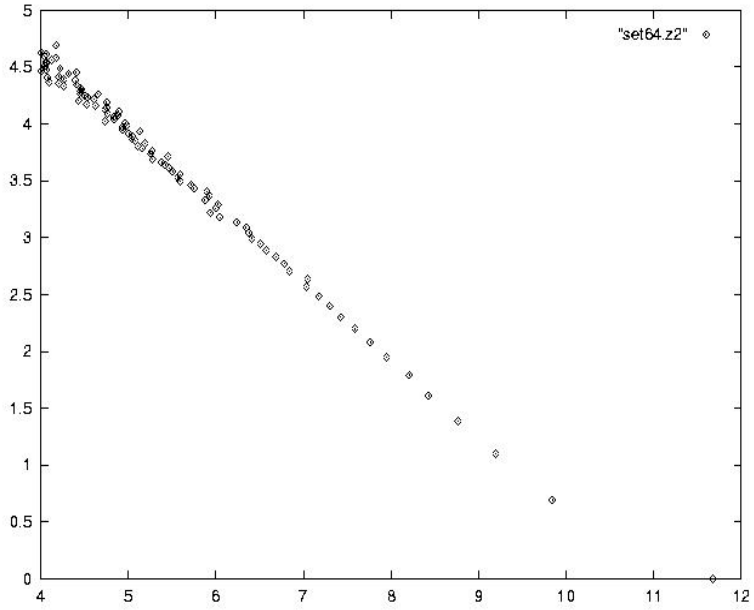

Fig. 4. Number-frequency law on the Merc (10m words)

Discussion. The normalization term K_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 $\log(i) = \log(V(N))/(\rho+1) - \log(V(i,N))/(\rho+1)$ plus some additive constant,

(7) $$V(i,N) = mV(N)/i^{\rho+1}$$

where m is some multiplicative constant. If we wish $\sum_{i=1}^{\infty} V(i,N) = V(N)$ to hold we must choose m to be $1/\zeta(\rho+1)$, which is the reason why Zipfian distributions are sometimes referred to as zeta distributions. Since this argument assumes Zipf's second law to extend well to high frequency items, the case for using $m = 1/\zeta(\rho+1)$ is far from 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 ρ in the vocabulary growth formula (5). 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 $N^{1-\varepsilon}$ for any $\varepsilon > 0$, so B must be 1. The following example sheds some light on the matter.

Example 2. Let $L = \{0,1,...,9)$ 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, $V(i,N) = N/i^2(\log(N) - \log(i))$ and thus $\log(V(N))/2 - \log(V(i,N))/2 = \log(N)/2$

– 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 used in this example 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 is a wealth of other easy to construct examples: as we shall see in Section 4, 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 $B > 1$.

To summarize, Theorem 3 means that distributions that satisfy (1) in the the mid- and the low-frequency range will also satisfy (2) in the low-frequency range. Since the observed fit with (2) is reasonably good, there seems to be no compelling need for a separate urn in the low frequency range. This is in sharp contrast to the high-frequency range, where both theoretical considerations and empirical observations dictate the use of a separate urn.

3.4. Hapax legomena and vocabulary richness

At the extreme low end of the frequency distribution we find *hapax legomena*, types that have only one token. Though misspellings and other errors often end up as hapaxes, it is worth emphasizing that hapaxes are not some accidental contamination of corpora. In the Merc, 46m tokens fall into nearly 600k types, and more than 400k of these (69.8%) are hapaxes. To be sure, over a third of these are numbers (see 5.2), but if we remove numeral expressions from the corpus, we still have 44m tokens, 385k types, of which 218k (56.6%) are hapaxes, consistent with the observation in Baayen (1996) that in large corpora typically more than 50% of the words are hapaxes.

Using $i = 1$ in (7), Zipf's second law predicts that a non-vanishing fraction $mV(N)$ of the vocabulary will be hapaxes, and with $i = 2$ we obtain that roughly a quarter as many will be *dis legomena* (types with exactly two tokens). These predictions have massive support in the quantitative linguistics literature: for example, Herdan (1964:219) only tabulates values of the Waring distribution (see 4.4 below) for the range $0.4 \leq V(1,N)/V(N) \leq 0.6$, because this range covers all values that "are likely to arise in practical work in the area of language".

Baayen (2001:2.4), following Khmaladze (1987, non vidi), defines sequences that have $V(1,N) \rightarrow \infty$ as having a *large number of rare events* (LNRE) if $\lim_{N\rightarrow\infty} V(1,N)/V(N)$ is positive. For a sequence to behave as LNRE it is not necessary for a non-vanishing fraction of *tokens* be rare: in fact, by the power law of vocabulary growth $V(1,N)/N$ will still tend to zero, but a positive fraction h of *types* are rare. As Baayen (2001: 57) notes, word frequency distributions, even when obtained from large samples, are in the LNRE zone. This observation in fact extends to the largest corpora currently available to researchers, web indexes comprising trillions of words, where the ratio of hapaxes is even higher. Assuming $V(1,N) > hV(N)$ we can again use the Turing-Good heuristics (see 2.3 above) for $V'(N) > hV(N)/N$ which, after integration, yields the power law (5) with exponent h.

We can also turn (7) around and use the observed ratio h of hapax legomena to vocabulary size to estimate the theoretical constant m directly, the ratio of dis legomena to vocabulary size to estimate $m/2^{\rho+1}$, and so forth. On the whole we expect better estimates of m from dis legomena than from hapaxes, since the latter also serve as a grab-bag for typos, large numerals, and other marginal phenomena (see 5.2). We can include *tris legomena* and in general use $V(i,N)/V(N)$ to estimate $m/i^{\rho+1}$. Combining the observed numbers of rare words into a single least squares estimate for $2 \leq i \leq 10$, in corpora with at least a few million

words, we can actually obtain better values of the Zipf constant $B = 1/\rho$ than by direct regression of log frequency against log rank.

Clearly, any attempt to model word frequency distributions must take into account the large number of rare words observed, but the large number of hapaxes is only the tip of the iceberg as far as vocabulary growth is concerned. Tweedie and Baayen (1998) survey a range of formulas used to measure vocabulary richness, and argue that many widely used ones, such as the *type-token ratio* $V(N)/N$, fail to define a constant value. In light of the asymptotic considerations used in this paper this comes as no surprise: Guiraud's R, defined as $V(N)/\sqrt{N}$, will tend to zero or infinity if $B < 2$ or $B > 2$ respectively. Dugast's and Rubet's k, defined as $\log(V(N))/\log(\log(N))$, must tend to infinity. Aside from Herdan's C, the main measures of vocabulary richness that can be expected to converge to constant values as sample size increases without bounds are Yule's K, defined as $\sum_{r=1}^{\infty} f_r^2$, entropy, given by $\sum_{r=1}^{\infty} - f_r \log(f_r)$, and in general Good's (1953) spectral measures with $Bt > 1$.

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.

4. Alternatives to Zipf's Law

The most widely used quantitative frequency laws are (1) and (2) as proposed by Zipf. But there are many alternatives, sometimes with easily identifiable champions, but often simply as communities of practice where using a particular model is taken for granted.

4.1. Minor variations

In many cases authors simply express (1-2) using different notation but an algrebrically equivalent formula such as (3). A more interesting case is when the immediate behavior is slightly different, as in Mizutani's Expression:

$$(8) \qquad \sum_{i=1}^{s} V(i,N) = \frac{V(N)s/N}{as/N + bN}$$

with a, b constants (Mizutani 1989). Another kind of correction to (1) was suggested by Mandelbrot (1961b), who introduces an additional parameter $W > 0$ in order to guarantee that the relative frequencies define a proper probability distribution for $B > 1$:

$$(9) \qquad \log(f_r) = \log(B-1) + (B-1)\log(W) - B\log(r+W).$$

With this correction, $\sum_{r=0}^{\infty} f_r \sim (B-1)W^{B-1}\int_W^{\infty} x^{-B}dx = 1$. If W is kept constant, (7) still leaves something to be desired, inasmuch as it assigns a total probability mass of approximately $N^{(1-B)/B}$ to the region of the curve where $r > V(N)$, but at least this error tends to zero as N tends to infinity.

4.2. Beta

Many kinds of minor corrective factors would be compatible with the available empirical evidence, but not all of them show acceptable limit behavior. A case in point is the beta distribution, which Simon (1955) obtained from a model of text generation embodying two assumptions: (i) new words are introduced by a small constant probability, and (ii) old words are reused with the same probability that they had in earlier text. He gives the resulting distribution in the form

(10) $V(i,N) = AB(i,\rho+1)$

where B is the Beta function. The parameter ρ is the same as in Zipf's laws, as can be seen from comparing the estimates for $V(i,N)/V(i+1,N)$ that can be obtained from (7) and (10). In particular, the case $\rho = 1$ corresponds to Zipf's original formulation of the law as $V(i,N)/V(N) = 1/[i(i+1)]$.

But Simon's assumption (i), linear vocabulary growth, is quite problematic empirically. 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 ρ is above 0.82, while in the Merc it was 0.75, but it is still very far from 1. Leaving out the adjacent chapters *Oxen of the Sun* and *Circe* we are left with roughly three-quarters of *Ulysses*, yielding an estimate of $\alpha = 0.116$ or $\rho = 0.825$. Applying these to the two chapters left out, which have 62743 words total, we can compute the number of words in *Ulysses* as whole based on αN, which yields 31122, or based on N^{ρ}, which yields 29804. The actual number of different words is 30014, so the error of the linear estimate, 3.7%, is over five times the 0.7% error of the power law estimate.

The Shakespeare canon provides another example of a "closed" corpus that displays open vocabulary growth. Plotting $\log(n)$ against $\log(V(n,N))$ as in Figure 4 yields Figure 5 (see below). A least squares estimate of ρ at the tail end of the curve yields about 0.52, quite far from unity. If we restrict ourselves to the very tail, we obtain 0.73, and if we use (7) we get 0.76, numbers still very far from what Simon considers the range of interest, "very close to 1". To summarize, the beta distribution is not an adequate model of word frequencies, because it assumes too many words: linear vocabulary growth instead of the power law observable both on dynamically growing corpora such as the Merc and on static ones such as *Ulysses* or the Shakespeare canon (for separate power law counts on *Antony and Cleopatra* and *Richard III* see Turner 1997).

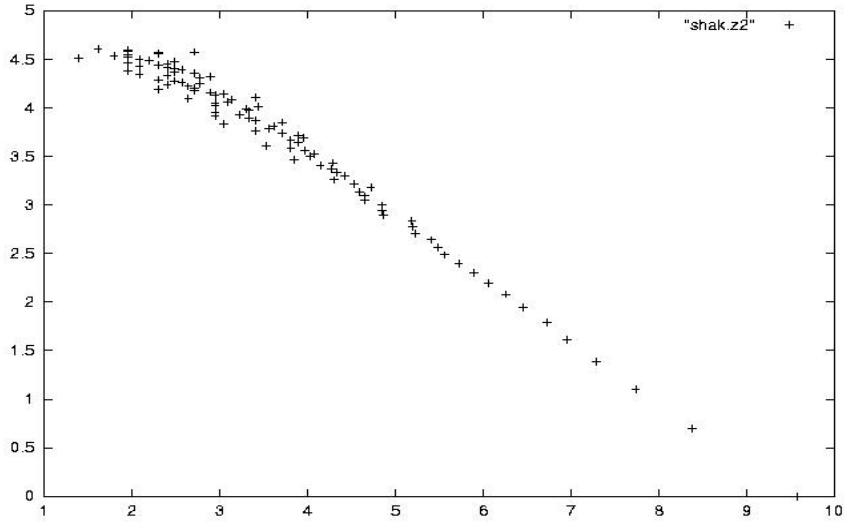

Fig. 5. Number-frequency law on the Shakespeare canon (885k words)

4.3. Lognormal

Another class of distributions that has considerable support in the literature is the *lognormal* family (Herdan 1960). As Champernowne discovered, the same genesis that leads to Pareto's law, when the assumption on minimum wealth is removed, will lead to a lognormal distribution instead. In word frequency counts, the resolution of the corpus presents a minimum barrier (everything that appears must appear at least once), but those in favor of the lognormal hypothesis could argue that this is an artifact of the counting method rather than a structural property of the data.

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 the following 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,

$$(11) \qquad f_N^2(x) = \frac{-f_N'(x)}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\log(f_N(x)) - \mu_N)^2}{\sigma_N^2}\right)$$

contains this term. Using the difference quotient we obtain $p_{k+1} - p_k/V(N)$, and we have $V(N) = N^{\rho}$ for some constant $\rho < 1$. By Zipf's law $\log(f_N(x)) = -\log(N) - B \log(x)$. Using (D2) we get that

$$(11) \qquad \frac{1/V(N)}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(-B\rho \log(N))^2}{\sigma_N^2}\right)$$

is constant, which can hold only if $\rho \log(N) = (1/2)\log(N)^2/\sigma_N^2$ i.e if $\sigma_N^2 = (B/2)\log(N)$.

In other words, the lognormal hypothesis does not lead to a stable limiting distribution: the means drift down with $\log(1/N)$ and the standard deviations open up with $\sqrt{\log(N)}$. This latter divergence, though theoretically more significant than the divergence of the means, is in practice barely noticeable when N grows with the current computational limits of corpus size: the predicted difference between a hundred million word corpus and a ten billion word corpus is less than 12%.

Proponents of the lognormal law could object to the above derivation, pointing out that once we assume that Zipf's first law fits the midrange $N^{\varepsilon} < i < N^{\rho-\varepsilon}$ and Zipf's second law fits the low range, surely the lognormal can not be expected to fit well. (In the high range, introducing a separate urn would benefit the lognormal approach just as much as it benefits the Zipfian.) We should, they might argue, turn the tables and see how well the Zipfian would fit, once we assume the underlying distribution to be lognormal. Because the results would be incompatible with Zipf's laws, the only means of settling the debate is to investigate which of the competing hypotheses fits the data better. Since the fit with lognormal is about as good (or about as bad, depending how we look at it) as with dzeta, there could be no compelling reason to favor one over the other.

While it is true that estimates on the divergence of the lognormal parameters are hard to quantify in the absence of an assumption that the distribution is Zipfian at least to the first approximation, for lower N the divergence is quite strong, and has been noted even by leading proponents of the lognormal hypothesis. Carroll (1967:407) has this to say:

> It will be noted that the mean and the standard deviation vary systematically as a function of sample size; that is, they appear to be biased estimates of the population value. [...] In ordinary sampling theory no bias is expected in the usual measures of central tendency applied to samples of different sizes, and the bias in estimates of the population variance is negligible for large samples and is easily adjusted for by using the number of degrees of freedom as the denominator in the calculation of variance.

Whether this bias is a barely perceptible lack of fit with the data or a gaping hole in the theoretical edifice is a matter of perspective: Herdan (1964:85) repudiates the lognormal largely because "for samples of widely different sizes from the same universe, the conventional statistics, such as the mean and the standard deviation, would not be the same". The main problem with the lognormal is the exact opposite of the problem with beta: beta distribution assumes there to be too many different words, while lognormal requires there to be too few. This problem is quite independent of the numerical details of curve-fitting: a lognormal distribution with fixed means μ and variance σ^2 predicts a fixed $V(N) = \exp(\sigma^2/2 - \mu)$, which Carroll (1967) calls the "theoretical absolute size of the vocabulary". But a fixed upper limit for $V(N)$ is incompatible with the results of our Theorem 1, and more importantly, with the empirically observable power law of vocabulary growth.

4.4. Waring

In a series of influential publications Herdan (1964, 1966) described an alternative to Zipf's law based on the Waring distribution. The basic idea is to model the classes C_i of types that have exactly i tokens: the Waring-Herdan formula asserts that the probability of a type falling in class C_i is given by

$$(12) \qquad \frac{\alpha}{\beta+\alpha}\frac{\beta}{\beta+\alpha+1}\cdots\frac{\beta+i-1}{\beta+\alpha+i}.$$

To account for the case when a token belongs to no previously seen type, the model explicitly contains the class C_0 of words not found in the sample, and assigns probability to it by the $i = 0$ case of the same formula (12). Instead of the number of visible types $V(N)$ we therefore deal with the total number of types $U(N)$ which includes the unseen words as well.

Let us first estimate the appropriate value for α and β. By Zipf's second law, we have $V(i,N)/V(i+1,N) = (1+1/i)^{\rho+1}$, with ρ is a constant < 1. From (12) we get $V(i,N)/V(i+1,N) = (\beta+\alpha+i+1)/(\beta+i)$. Therefore we need

$$1+\frac{\alpha+1}{\beta+i} \sim 1+\frac{\rho+1}{i}$$

which can work well for a range such as $2 \le i \le 200$ only if α is close to ρ and β is close to zero. With this choice, we actually obtain the correct prediction, namely that class C_0 has probability one, meaning that almost all types in the population remain unseen in any sample.

Again, proponents of the Waring distribution could object to the above derivation, especially as it assumes Zipf's second law to provide a reasonable first approximation over a broad range, while in fact the fit in this range is far from perfect. We will therefore use Herdan's own estimators for α and β, which are based on the proportion of hapaxes, $h = V(1,N)/V(N)$ and on the average class size $M = N/V(N)$ as follows:

$$\beta = \frac{1}{\frac{1}{1-h}-\frac{1}{M}-1} \qquad\qquad \beta+\alpha = \frac{\beta}{1-h}.$$

In the larger Merc samples, $1/M$ is already less than 0.01, and if we increase N without bounds, $1/M$ will tend to zero. Thus for very large samples Herdan's estimators yield $\beta = (1-h)/h$ and $\alpha = 1$. Thus we obtain a distribution with a single parameter p, which predicts that in any sample the chances of a random type being manifested as hapax legomena, dis legomena, tris legomena etc. are

$$(13) \qquad \frac{p(1-p)}{1+p}; \qquad \frac{p(1-p)}{(1+p)(1+2p)}; \qquad \frac{p(1-p)}{(1+2p)(1+3p)}; \qquad \ldots$$

These probabilities do not add up to one: the remaining probability, which is p, is assigned to unseen word types. Of course, the whole notion of equiprobable selection of types makes sense only if there is a finite number of types (closed vocabulary): what (13) predicts is that $V(N)/U(N)$, the proportion of visible types among the "theoretical absolute" number of types, is $1-p$.

One way of putting this result is that even if there are infinitely many word types, on the basis of a finite sample, with $V(N)$ manifest types, we can justify no more than $V(N)/(1-p)$ types altogether. This is not dissimilar to the logic underlying the lognormal fit, and in some sense even better, since the lognormal parameters μ and σ diverge as $N \to \infty$, but the Waring parameters α and β appear to converge to stable values 1 and $(1-h)/h$ respectively. Since h, the proportion of hapaxes among the total number of word types seen, is about 1/2, we can conclude that there are roughly as many unseen types as there are visible ones.

Herdan uses h to estimate p, but of course we could apply least squares fit in the entire range of p_i. From (13) we obtain $V(i,N)/V(i+1,N) = 1 + 2p/(1 + ip - p)$, which will be the same as the Zipfian estimate $1 + (\rho + 1)/i$ just in case $\rho = 1/B = 1$. Again, both the Waring and the ζ

distributions fit the observed numbers about equally well for small i, and both leave a lot to be desired for larger i. Therefore, the choice between the two has to be made indirectly, based in this case on the inability of the Waring distribution to support a Zipfian tail for the case of practical interest, $B > 1$.

4.5. Negative binomial

Another influential model originates with the work of Fisher (1943) on species abundance. The main relation can be formulated as

$$\frac{V(i,N)}{V(1,N)} = \frac{\Gamma(i+\alpha)}{i!\Gamma(1+\alpha)}\gamma^{i-1} \tag{14}$$

which is closely related to both (10) and (12). If (12) is to agree with (7) over a broader range, we need

$$1+\frac{\rho+1}{i} \sim \frac{i+1}{\gamma(i+\alpha)}$$

which requires γ to be close to 1 and α to be close to $-\rho$. For example, Efron and Thisted (1976) fits a negative binomial to the Shakespeare data depicted in Figure 5, obtaining $\gamma = 0.9905$, $\alpha = -0.3954$. This translates into an estimate of $\rho \approx 0.4$, much lower than the 0.73 we obtained in 4.2 based on $V(1,N)/V(2,N)$ alone, and still significantly lower than the 0.52 we obtained from the first five $V(i,N)$. Because their method gives nearly uniform weight to the whole range $1 \le i \le 40$ they consider, the discrepancy is quite visible, but for the same range our simpler fit would yield $\rho = 0.36$. Since Theorem 2 indicates that vocabulary growth is determined at the margin, we concentrate at the low end, ignoring values for $i > 5$ entirely.

In general, the negative binomial offers a considerably better fit than Zipf's second law, and in the range of interest, $-1 < \alpha < 0$, yields the same vocabulary growth formula

$$V'(N) = V(1,N)/N \tag{15}$$

as the Turing-Good method, but without assuming a uniform prior on types. As we have seen in 3.2, $V(1,N) = mV(N)$ for some constant $0 < m \le 1$. Combined with (15) we obtain, up to a constant factor, the power law (5) with exponent m. For the Shakespeare canon, $m = 0.46$, roughly halfway between the exponent predicted by (14) and the one computed from $1 \le i \le 5$.

5. Where do the words come from?

The closed vocabulary assumption, that there is a fixed number of words S in any given language, is often couched in terms of rather sophisticated statistical frameworks that assume the existence of unseen words. After discussing this issue in 5.1, we turn to the open vocabulary in 5.2, where we attempt to identify the factors fueling infinite vocabulary growth. We offer our conclusions in 5.3.

5.1. Unseen words

The lognormal and Waring distributions are examples of a broader class of hypotheses that require a distinction between the observed number of types $V(N)$ and the predicted number of types $U(N)$. In some sense, this is a very attractive distinction, for it would surely be absurd to assume that no new sample will ever contain words hitherto unseen. Also, on the basis of such hypotheses, we can obtain quantitative answers to a range of questions:

> What proportion of first names known in small villages are actually in use there? Over 80%, according to the Waring fit used in Schubert and Toma (1983).
>
> How many words could be used in children's reading materials between grades 3 and 9? 609,606, according to the lognormal fit used in Carroll (1971).
>
> How many words did Shakespeare know but never use? At least 35,000, according to the negative binomial fit used by Efron and Thisted (1976).
>
> How many words can appear in Turkish archeological texts given a sample of 7k words? Over half trillion, based on the generalized inverse Gauss-Poisson fit used in Baayen (2001:4.3.1).

It is quite conceivable that in rural Hungary first names were indeed drawn from a rather small closed set, and therefore fitting a Waring or lognormal distribution is appropriate. However, the statistics published in Carroll (1971) give no indication that children's reading materials come from a closed subset of the vocabulary, and extrapolationg by (7) suggest that it would take less than 20 times the current corpus to transcend the 609,606 types predicted by the lognormal fit.

While the third question seems to be about Shakespeare's mental lexicon, in fact Efron and Thisted (1976) posed it in a more conservative fashion: how many new words $\Delta(t)$ do we expect if a new body t times the size of the currently acknowledged Shakespearean canon was discovered? We are certainly in no position to collect 20 times the available Shakespearean corpus (for $t = 0.0004849$ see Thisted and Efron 1987) so let us for the moment pursue the issue based on the Merc.

The hope is that by fitting a Waring distribution, we can describe how many words are known by the journalists at the Merc. In the first sample, $N = 147{,}260$, we estimate $U(N)$ by $V(N)/(1 - V(1,N)/V(N))$, obtaining $U(N) = 35{,}439$. However, in an independent sample of 587k tokens, we find more than this, 38,865 types. For this sample the Waring estimate is $U(N) = 113$k. However, in an independent sample of 4.7m words we find $V(N) = 127$k and $U(N) = 280$k, but we are just as far from the elusive "theoretical absolute size" as we were before, and in an independent sample of 18.3m words indeed we find 310k different words.

What is fueling all this vocabulary growth? If indeed $V(N) \to \infty$ as sample size grows, the answer can not be that the Merc employs, say, a hundred journalists: for the total to come out infinitely large, at least some of them must already know infinitely many words. We are in no position to study the Merc contributors separately, but wherever such studies have been conducted, as on the output of Chaucer, Shakespeare, or Joyce, the same unlimited pattern of vocabulary growth can be seen. In fact, Efron and Thisted (1976) are quite explicit about the fact that their model predicts $\Delta(t) \to \infty$ as $t \to \infty$, they just somehow find this conclusion unpalatable (perhaps because it would make no sense for the case of species abundance that originally gave rise to the model).

From what we have seen so far, not only do writers have infinite vocabularies, they actually provide evidence of this in the (necessarily finite) corpus of their writings. Further,

this conclusion must hold not only for extraordinary literary geniuses but, by the pigeonhole principle, at least for some of the lesser known journalists toiling at the Merc.

5.2. Mixtures

A natural generalization of the Zipfian model is to assume that we have a *mixture* of Zipfians: instead of a single distribution with parameter B we have k independent distributions with parameters $B_1, \cdots, B_k$. In such cases, it follows from Theorem 2 that vocabulary growth will always be dominated by the smallest B_i. One application would be to separate the individual writers of multi-author corpora and inspect their contribution to the overall vocabulary separately. Another application is to look at different classes of words, such as morphologically simple vs. morphologically complex, written using digits vs. letters, etc. For example, we may assume that only 98% of the Merc data are ordinary words, and the remaining 2% are numbers. If we separate the two corpora, indeed we obtain quite different parameters: B_w = 1.65 for words and B_n = 1.31 for numbers. Assuming the power law (5), this means that for corpora with 385m words and beyond, the number of new number types will exceed the number of new word types, even though the number of new number tokens will still be only 2% of the number of new word tokens.

While it is true that the distribution of numbers in the Merc follows the same broad Zipfian patterns as the distribution of words, we certainly do not need an elaborate statistical argument to prove that the number of numbers is infinite, or that writers know infinitely many numbers. To the contrary, our goal here is to protect the conclusion, that vocabulary is open, against the counterargument "yes, but only because it includes numbers". In fact, if we remove numbers, we still find the same Zipfian pattern, and vocabulary growth still obeys the power law (5), though with a smaller exponent.

If we inspect the non-numerical hapaxes more closely, we find that other obviously infinite sources, such as proper names, foreign words, typos and eye-dialect (e.g. *Arrrrrrnnnnnold*) play a significant role. Again we need to assign these to separate mixture components, and argue that the rest still grows without bounds. Remarkably, at this point the bulk of the vocabulary growth is actually provided by productive morphological processes: about 40% of the non-numeric hapaxes are hyphenated, a clear sign of compounding, and over 7% end in the possessive suffix *'s*. Multiply suffixed forms, such as *eclectically* or *ebonizing* provide about 5%, so at 40m words the majority of non-numeric hapaxes are either monomorphemic English words such as *decade* or polymorphemic, but clearly well-formed English. To the extent that numerals freely enter into combination with nouns to form adjectives such as *958-member*, one could even argue that there is no need to treat numbers, proper names, or even foreign words as extragrammatical, but we leave this matter to the side here, as the inclusion of these classes would no doubt weaken our argument in the eyes of many.

5.3. Summary and conclusions

In this paper we answered the question posed in the title by arguing that there are an infinite number of words. We came to this conclusion not on the basis of productive morphological processes, but rather by inspecting the characteristic properties of large corpora, and deriving the open vocabulary result from these properties. Nevertheless, our results support the conclusion that the main grammatical source of infinite vocabulary growth is productive generative morphology, in particular compounding.

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 that the frequency plot is log-covex (subgeometric mean property) and that every corpus has hapaxes. Using these properties and a simple normalization technique we proved in Theorem 1 that vocabulary size $V(N)$ tends to infinity as $N \to \infty$.

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 $V(N) = N^{1/B}$. Theorem 3 established a simple quantitative connection between Zipf's first and second law, suggesting that there is no need to introduce a separate urn for the low range, especially 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. If separate urns are needed at all, they should be used for numerals, typos, eye-dialect, direct quotations form other languages, and other arguably extragrammatical material that can be seen as contaminating the basic vocabulary pattern.

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 approximately with a power $\rho < 1$ of N, and the most widely used competing hypotheses, in particular the beta, lognormal, and Waring distributions, are not well suited for characterizing the observed pattern of word frequencies. The negative binomial, with α negative and γ close to 1, stands out as a realistic alternative to ζ, though a satisfactory genesis, explaining why the cumulative distribution function of the Poisson parameters should follow Γ, is still lacking.

Acknowledgements

The author would like to thank Gabriel Landini, David M.W. Powers, and the anonymous reviewers for valuable suggestions and discussion of earlier drafts of this paper. In particular, the possibility of deriving a logarithmic lower bound for vocabulary size based on the Turing-Good estimates (see the end of 2.3) was called to my attention by reviewer #3.

References

Baayen, R. H. (1996). The effect of lexical specialisation on the growth curve of the vocabulary. *Computational Linguistics 22, 455-480.*

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer Academic Publishers.

Burrows, J. (1987). Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing 2, 61-70.*

Carroll, J. B. (1967). On Sampling from a lognormal model of word-frequency distribution. In: H. Kucera and W. Francis (eds.), *Computational Analysis of Present-Day American English: 406-424.* Providence, RI: Brown University Press.

Carroll, J. B. (1971). *The American Heritage word frequency book*. Boston: Houghton Mifflin.

Champernowne, D. (1952). The graduation of income distributions. *Econometrica 20, 591-615.*

Champernowne, D. (1953). A model of income distribution. *Economic Journal 63, 318-351.*

Champernowne, D. (1973). *The distribution of income*. Cambridge University Press.

Cox, D., Miller, H. (1965). *The theory of stochastic processes*. London: Methuen.

Cramér, H. (1955). *The elements of probability theory*. New York: John Wiley & Sons.

Efron, B., Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika 63, 435-448.*

Estoup, J. (1916). *Gammes Stenographiques*. Paris: Institut Stenographique de France.

Fisher, R., Corbet, A., Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology 12, 42-58.*

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40, 237-264.*

Guiraud, H. (1954). *Les charactères statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-token mathematics*. The Hague: Mouton.

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

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin: Springer.

Khmaladze, E. (1987). The statistical analysis of large number of rare events. *Technical Report MS-R8804*, Dept of Mathematical Statistics, CWI, Amsterdam: Center for Mathematics and Computer Science.

Knuth, D. E. (1971). *The art of computer programming*. Reading MA: Addison-Wesley.

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

Mandelbrot, B. (1952). An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. In: W. Jackson (ed.), *Second Symposium on Information Theory.* London.

Mandelbrot, B. (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.*

Mandelbrot, B. (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.*

Mandelbrot, B. (1961b). On the thory of word frequencies and on related markovian models of discourse. In: R. Jakobson (ed.), *Structure of language and its mathematical aspects: 190-219.* Providence: American Mathematical Society.

Mandelbrot, B. (1961c). Post scriptum to 'final note'. *Information and Control 4, 300-304.*

Mizutani, S. (1989). Ohno's Lexical Law: Its Data Adjustment by Linear Regression. In: S. Mizutani (ed.), *Japanese Quantitative Linguistics: 1-13.* Bochum: Brockmeyer.

Nádas, A. (1985). On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing 33(6), 1414-1416.*

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

Pareto, V. (1897). *Cours d'economie politique*. Lausanne-Paris: Rouge.

Powers, D. M. (1998). Applications and explanations of Zipf's law. In: D. Powers (ed.): *NEMLAP3/CONLL98: New methods in language processing and Computational natural language learning: 151-160.*

Samuelsson, C. (1996). Relating Turing's Formula and Zipf's Law. *Proc. Fourth Workshop on Very Large Corpora*.

Schubert, A., Toma, O. (1984). Estimating the total number of first names based on occurrence frequency (In Hungarian). *Névtani Értesítő 9, 72-80.*

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

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

Simon, H. A. (1961a). Reply to Dr. Mandelbrot's post scriptum. *Information and Control 4, 305-308.*

Simon, H. A. (1961b). Reply to 'final note' by Benoit Mandelbrot. *Information and Control 4, 217-223.*

Thisted, R., Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika 74, 445-455.*

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

Tweedie, F. J., Baayen, R.H. (1998), How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32, 323-352*.

Willis, J. (1922). *Age and area*. Cambridge University Press.

Yule, G. U. (1924). A mathematical theory of evolution. *Philosophical Transactions of the Royal Society B 213, 21ff.*

Zipf, G. K. (1935). *The psycho-biology of language; an introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass: Addison-Wesley.