39 Distributions in text

Marco Baroni

June 30, 2006

1 Introduction

The frequency of words and other linguistic units plays a central role in all branches of corpus linguistics. Indeed, the use of frequency information is what distinguishes corpus-based methodologies from other approaches to language. Thus, not surprisingly, the distribution of frequencies of words and combinations of words in corpora has played a central role in the debate between proponents and detractors of the corpus-based approach (see, e.g., Abney 1996). One would then expect that the study of word frequency distributions plays a central role in the corpus linguistics curriculum. This is not the case. The standard introductions to the field (e.g., Biber/Conrad/Reppen 1998; McEnery/Wilson 2001) do not discuss the topic at all, and even an introduction explicitly geared towards the statistical aspects of the discipline, such as Oakes (1998), mentions Zipf's law (see section 3 below) only in passing (pp. 54-55).

This state of affairs may be due to the fact that the study of word frequency distributions originated outside mainstream linguistics. George Kingsley Zipf, undoubtedly the father of *lexical statistics* (the study of word frequency distributions), was trained as a philologist and considered himself a "human ecologist". Other important pioneers of the field were the psychologist George Miller, the mathematician Benoit Mandelbrot (of Mandelbrot set fame) and the Nobel Prize winning economist Herbert Simon. Thus, the argumentations and terminology found in the early literature often sound rather exotic to linguists (e.g., Mandelbrot's "temperature of discourse" approach). Still today, most articles about lexical statistics appear in relatively obscure journals and they are often rooted in traditions, in particular that of the former Soviet Union, that are not well known in the English-centered world of corpus linguistics (Sampson 2002). The heavy involvement of non-linguists in the study of lexical statistics continues to this day. Judging from the affiliations of the authors of the recent *Glottometrics* volumes in honor of Zipf, word frequency distributions are more of interest to theoretical physicists than to theoretical linguists. The relatively recent publication of Baayen (2001), a thorough introduction to lexical statistics that summarizes

much of the earlier work, but recasts problems and solutions in the perspective of modern corpus/computational linguistics, will probably contribute to give more prominence to the domain.

This article introduces some of the empirical phenomena pertaining to word frequency distributions and the classic models that have been proposed to capture them. In particular, section 2 introduces the basic analytical tools and discusses the patterns typically encountered in corpora/texts. Section 3 presents Zipf-Mandelbrot's law, the most famous model proposed to account for frequency distributions. Section 4 shortly reviews some of the practical consequences and applications of frequency distribution modeling. Section 5 concludes by suggesting some directions for further study.

2 Distributions

2.1 Counting tokens and types

In order to study word frequency distribution, we must first of all count all the instances (*tokens*) of all distinct words (*types*) that occur in the corpus of interest (I use the term corpus in the most general way, to refer to any text or collection of texts that is the object of a linguistic study). Neither deciding what must be counted as a token, nor mapping tokens to types are trivial tasks. Consider the following mini-corpus:

> The woman went to Long Beach and to Anaheim on bus number
> 234. However, the man didn't go.

First, we will have to decide whether punctuation marks are tokens or not and whether to keep or remove strings containing digits. Both choices affect the shape of frequency distributions (punctuation marks are few and very frequent, numbers are many and rare). Next, we face a number of token segmentation problems. For example, we must decide whether we should split *didn't* into two words (and if we do, where do we split it). Moreover, *Long Beach* should perhaps be counted as a single word. Again, these choices will affect our counts in a systematic way. Having decided which strings to ignore, and how to segment the remaining text, we can count the tokens in the corpus. For example, if we decide to ignore punctuation and numbers, to treat *Long Beach* as two words and *didn't* as a single word, the mini-corpus above will have 17 tokens: *The, woman, went, to, Long, Beach, and, to, Anaheim, on, bus, number, However, the, man, didn't, go.*

Now, we must map each word token to a word type. In order to do this, we have to decide whether our counts should be sensitive to the distinction between upper and lower case or not: intuitively, *The* and *the* in the mini-corpus above should be counted as instances of the same word, but it would be wrong to treat the parts of the name *Long Beach* as instances

of the adjective *long* and noun *beach*, respectively. In English, ignoring the distinction between upper and lower case will have distorting effects on proper name counts, but by preserving case distinctions we will duplicate word types that occur both in sentence-initial position and elsewhere. If we distinguish between upper and lower case, the mini-corpus tokenized as above will contain 16 types, one of them (*to*) represented by two tokens.

If we have the relevant resources (most importantly, a list of word-form/lemma correspondences), we can map tokens to lemma types. In the mini-corpus above, *went* and *go* would be treated as tokens of the same lemma type. On the one hand, more sophisticated tokenization/type mapping steps are likely to lead to cleaner counts. On the other, the errors and imprecisions inherent in any form of automated pre-processing can have a serious distorting effect on the data. For example, if all the words that are not recognized by our lemmatizer are mapped to a type *unknown*, we will transform many low frequency items into a single artificial high frequency type.

In the corpora analyzed in this article, unless stated otherwise, punctuation marks, strings containing digits and strings made entirely of non-alphabetic characters are not counted as tokens; all other white-space or punctuation-delimited strings constitute separate tokens (in English, some special strings are split into multiple tokens – e.g., *wouldn't* is tokenized as *would*, *n't*); upper- and lower-case types and not merged; lemmatization is not performed. The token and type counts I report are based on this tokenization/type mapping scheme. Issues related to corpus pre-processing, tokenization and lemmatization are discussed in Articles 25 and 26 of this handbook.

2.2 The basic tools of lexical statistics

Once we have tokenized a corpus and mapped each token to a type, we can count the number of tokens in the corpus, or *corpus size* (N), and the number of types, or *vocabulary size* (V). For example, in the mini-corpus above, given the tokenization and type mapping rules I adopted, N is 14 and V is 13.

The starting point for any further analysis will be a *frequency list*, i.e., a list that reports the number of instances (tokens) of each word (type) that we encountered in the corpus. Consider for example the toy frequency list in table 1.

The data in a frequency list can be re-organized in two ways that are particularly useful to study word frequency distributions: as *rank/frequency profiles* and as *frequency spectra*. To obtain a rank/frequency profile, we simply replace the types in the frequency list with their frequency-based ranks, by assigning rank 1 to the most frequent type, rank 2 to the second most frequent word, etc. In the example of table 1, *barks* would be assigned

type	f	type	f
again	2	he	1
and	3	her	1
another	1	that	2
bark	1	this	1
barks	6	will	1
dog	3	with	1
friends	1		

Table 1: A toy frequency list

rank 1, *and* and *dog* would be assigned rank 2 and 3 (ranking of words with the same frequency is arbitrary), etc. This produces the rank/frequency profile in table 2.

r	f	r	f
1	6	8	1
2	3	9	1
3	3	10	1
4	2	11	1
5	2	12	1
6	1	13	1
7	1		

Table 2: A toy rank/frequency profile

A frequency spectrum is a list reporting how many types in a frequency list have a certain frequency. The spectrum corresponding to the frequency information in table 1 is presented in table 3.

f	V(f)
1	8
2	2
3	2
6	1

Table 3: A toy frequency spectrum

The first row of table 3 tells us that there are 8 words with frequency 1 ($V(1) = 8$; *another, bark, friends, he, her, this, will, with*). The second row tells us that there are 2 words with frequency 2 ($V(2) = 2$; *again, that*), etc.

A rank/frequency profile and the corresponding frequency spectrum contain the same information, and it is thus possible to derive one from the other. However, as we will see, rank/frequency profiles are particularly useful to study the properties of high frequency items and frequency spectra are useful to study the properties of low frequency items.

2.3 Typical frequency patterns

Table 4 shows the top and bottom ranks and corresponding frequencies in the Brown corpus of American English (see Appendix).

top frequencies			bottom frequencies		
rank	fq	word	rank range	fq	randomly selected examples
1	62642	the	7967-8522	10	recordings undergone privileges
2	35971	of	8523-9236	9	Leonard indulge creativity
3	27831	and	9237-10042	8	unnatural Lolotte authenticity
4	25608	to	10043-11185	7	diffraction Augusta postpone
5	21883	a	11186-12510	6	uniformly throttle agglutinin
6	19474	in	12511-14369	5	Bud Councilman immoral
7	10292	that	14370-16938	4	verification gleamed groin
8	10026	is	16939-21076	3	Princes nonspecifically Arger
9	9887	was	21077-28701	2	blitz pertinence arson
10	8811	for	28702-53076	1	Salaries Evensen parentheses

Table 4: Top and bottom of the Brown frequency list

The top ranks are occupied by function words such as *the*, *of* and *and*. Frequency decreases quite rapidly: the most frequent word is almost twice as frequent as the second most frequent word. The difference in frequency becomes less dramatic as we go down the list, but the ranks are still spread across a wide frequency range. Because of their very high frequencies, the 10 top-ranked word types alone account for about 23% of the total token count in the Brown (232,425 occurrences over 996,883 tokens in total). This is to say that in the Brown more than one word in five comes from the set *the, of, and, to, a, in, that, is, was, for.*

The picture is very different at the bottom of the list, where there are massive frequency ties, and more ties as the frequency decreases: for example, there are 4,137 words with frequency 3 (ranks from 16939 to 21076), 7,624 words with frequency 2 (ranks from 21077 to 28701), 24,374 words with frequency 1 (ranks from 28702 to 53076). Since the Brown corpus contains 53,076 distinct types in total, the words occurring once constitute almost half of its vocabulary. The words occurring 3 times or less constitute almost 70% of the vocabulary. At the same time, this 70% of types account for only about 5% of the overall Brown token count (52,033 tokens over 996,883 total tokens). The lowest frequency elements are of course content words. As the random examples reported in the table show, not all the lowest frequency words are neologisms, new derivations or exotic forms. For example, words such as *pertinence* and *parentheses* are probably not going to strike the average English speaker as new or unusual.

The dichotomy between the extremely high token frequency of the most frequent types and the large number of low frequency types affects the classic summary statistics in peculiar ways. The average frequency of word types in the Brown is of 19 tokens. However, this value is inflated by the very high frequencies of the most common words: more than 90% of the types in

the Brown corpus have frequency lower than the average. The median value is 2 (i.e., 50% types have frequency greater than or equal to 2, and 50% types have frequency less than or equal to 2). The mode (the most common value), of course, is 1.
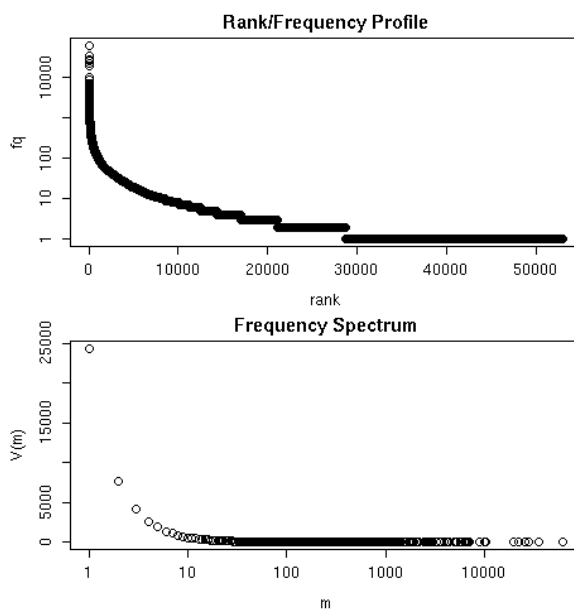

Figure 1: Rank/frequency profile and frequency spectrum of the Brown corpus.

The upper panel of figure 1 illustrates the rank/frequency profile of the Brown corpus. Frequency (on the y axis) is plotted on a logarithmic scale, because the frequency of the most frequent words is so much higher than the frequency of the long tail of rare words that a figure of this size without a logarithmic transformation would look like the letter L. The plot illustrates very clearly what we already observed: the frequency curve decreases very steeply from the extremely high values corresponding to the most frequent words, and it becomes progressively flatter, until it reaches a very wide plateau in correspondence to the ranks assigned to the tail of words occurring once (increasingly narrower plateaus corresponding to words occurring 2, 3, 4 times etc. are also visible). The lower panel of figure 1 plots the frequency spectrum of the Brown (again, token frequency – this time on the x axis – is on a logarithmic scale). The lowest frequency classes are represented by a very large (and rapidly decreasing) number of types (the types that occur once, the types that occur twice, etc.), and there is a long tail of high frequency classes represented by only 1 or 0 types.

The frequency distribution of the Brown is not specific to this corpus, but typical of natural language texts, independently of tokenization/type

mapping method, size, language, textual typology, etc. To illustrate this, let us consider the British National Corpus (BNC – see Appendix), which differs from the Brown in that it represents British rather than American English, it is based on more recent texts, it includes a spoken language section and, perhaps most importantly, it is much larger. The Brown contains about one million tokens, whereas the written section of the BNC contains 86,480,906 tokens, and the spoken section contains 10,423,654 tokens.
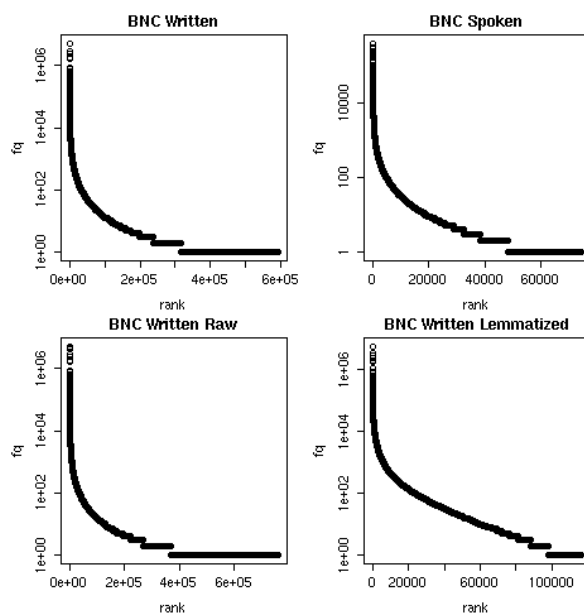

Figure 2: Rank/frequency profiles of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right).

Figures 2 and 3 present rank/frequency profiles and frequency spectra for the BNC. The top two panels of figure 2 show the rank/frequency profiles of the BNC written and spoken sections, respectively. The top two panels of figure 3 show the corresponding spectra. The overall pattern is very similar to the one we observed in the Brown: few very frequent words, many low frequency words. This second fact is perhaps surprising: one could reasonably expect that in a very large sample of a language the words that are encountered only once become a minority. This is obviously not the case: in the written section of the BNC, the words occurring only once account for 46% of all the types, and the proportion of words occurring 3 times or less is of 66%. In the spoken section, these proportions are smaller (perhaps suggesting less lexical variety in speech?) but still very significant: 35% of the types occur only once and 56% occur 3 times or less. The mean token frequency of types in the written BNC is of about 146 tokens but
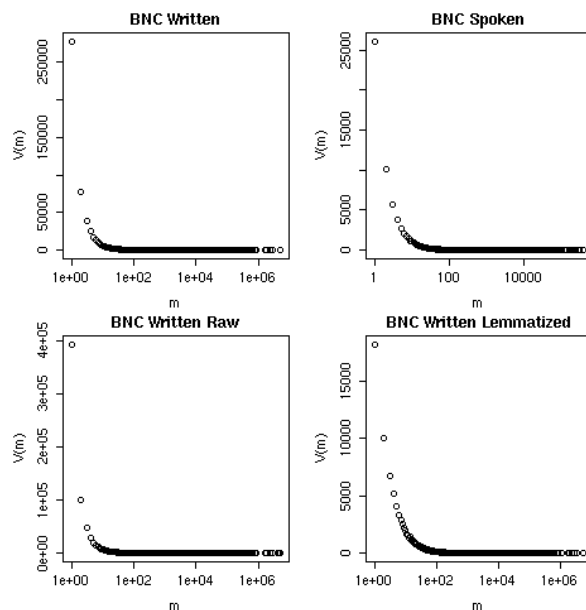
7

Figure 3: Frequency spectra of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right).

more than 95% of the types have a frequency below this value. Like in the Brown, the median is 2 and the mode is 1. Corpus after corpus, we find that the mean is a value much higher than the median (and, as is intuitive, it increases in function of corpus size), the median is 2 or 1 and the mode is 1. Thus, the mean is not a meaningful indicator of central tendency, whereas the median and the mode are not very interesting since they tend to have the same values in all corpora. The third panels of figures 2 and 3 show the rank/frequency profile and frequency spectrum in a version of the written BNC in which strings containing digits and other non-alphabetic symbols were counted as regular words. Again, we encounter a very similar pattern. Not surprisingly, the portion of the distribution taken by words occurring only once is even more prominent. The bottom right panels of figures 2 and 3 report the rank/frequency profile and frequency spectrum of the lemmas in the written BNC. Although the number of very low frequency forms is lower than in the non-lemmatized counterpart (top left panels), the overall pattern is essentially the same, which shows that such pattern cannot be simply explained in terms of the presence of inflected forms in non-lemmatized corpora.

Figures 4 and 5 display rank/frequency profiles and frequency spectra for four more texts/corpora of very different kinds. The top left panels present data from *The War of the Worlds*, the famous H. G. Wells novel from 1898,
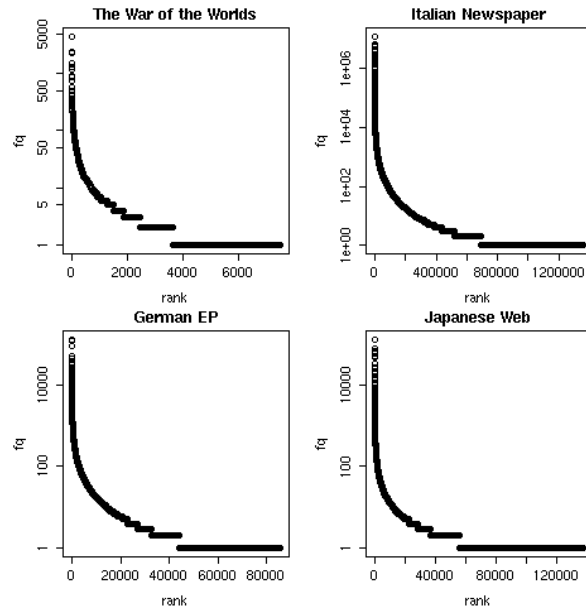
8

Figure 4: Rank/frequency profiles of *The War of the Worlds* (top left), the Italian *la Repubblica* corpus (top right), a section of the German *EuroParl* corpus (bottom left) and a corpus of Japanese web-pages (bottom right).
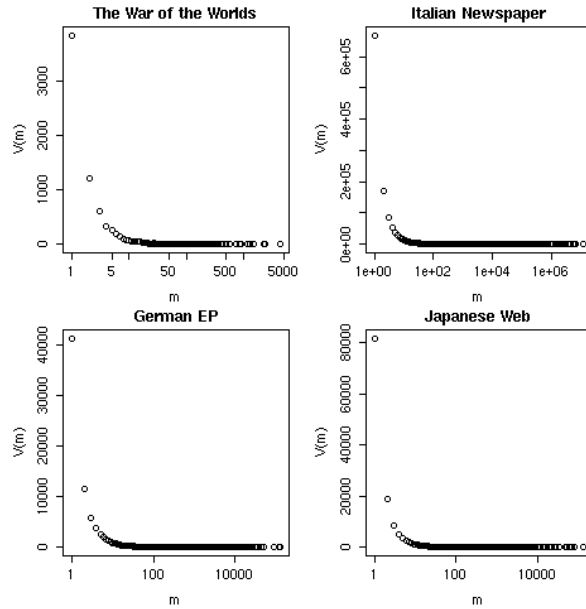

Figure 5: Frequency spectra of *The War of the Worlds* (top left), the Italian *la Repubblica* corpus (top right), a section of the German *EuroParl* corpus (bottom left) and a corpus of Japanese web-pages (bottom right).

9

which, unlike the Brown or the BNC, is a "corpus" made of a single, coherent text. Moreover, compared to the other corpora analyzed here, it is very small, being comprised of 60,160 tokens. The top right panels present data from the *la Repubblica* corpus, containing (at the time in which the data used here were extracted) 325,290,035 tokens of Italian newspaper text (see Appendix). The bottom left panels present data from the year-2002 section of the German EuroParl corpus (see Appendix), collecting transcriptions of European Parliament proceedings. This corpus contains 3,090,142 tokens. Finally, the bottom right panels present data from a corpus of Japanese web pages collected in 2004 with the method described in Ueyama/Baroni (2006) and tokenized with the ChaSen system (Matsumoto/Kitauchi/Yamashita/et al. 2000). It contains 2,175,736 tokens. Despite the obvious differences among these corpora, the rank/frequency profiles and the frequency spectra reveal strikingly similar overall patterns, in turn resembling those that we encountered in the Brown and BNC: few very high frequency types, and long tails of very rare words.

The same skewed shape also emerges if instead of looking at words we look at sequences of words, or *ngrams*, such as *bigrams* or *trigrams* (sequences of two and three words, respectively) . Such distributions are even more skewed than those of words (given that the potential vocabulary of possible ngrams is much higher). This is illustrated for the Brown corpus in figure 6. Among the trigrams, the types with frequency 1 constitute 92% of the vocabulary!
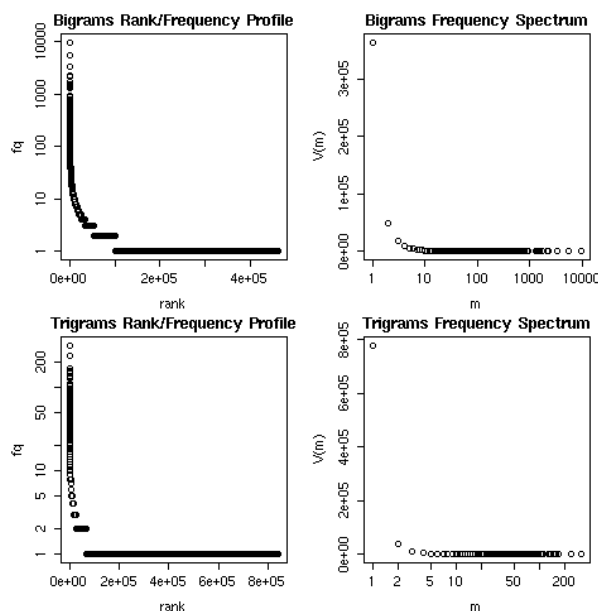

Figure 6: Rank/frequency profiles and frequency spectra of the bigrams (top) and trigrams (bottom) in the Brown corpus.

The distribution of word and ngram frequencies is rather different from the typical count distributions that are studied in introductory statistics classes. For example, if we divide the male students of a certain high-school into classes based on their height, we expect that most students will fall into the medium class, fewer students will be classified as tall or short, and very few students will turn out to be extremely tall or extremely short. The distribution of words is akin to finding a population made of few giants, rather few people in the medium height range and an army of dwarves.

3 Zipf(-Mandelbrot)'s law

The typical skewed structure of word frequency distributions was first systematically studied by Zipf (1949, 1965), who observed in various data-sets that frequency is a non-linearly decreasing function of rank (decreasing more sharply among high ranks than among low ranks), and proposed the following model, which became known as *Zipf's law*, to predict the frequency of a word given its rank:

$$f(w) = \frac{C}{r(w)^a} \tag{1}$$

In this formula, $f(w)$ and $r(w)$ stand for frequency and rank of word w, respectively. C and a are constants to be determined on the basis of the available data. To understand why this is a plausible model, assume for now that $a = 1$ (but the same point could be illustrated with other values of this parameter), so that equation (1) can be simplified to $f(w) = \frac{C}{r(w)}$. Then, the most frequent word in the corpus, having rank 1, must have frequency C. Suppose that in a certain corpus we find that the most frequent word has frequency 60,000 and thus we set $C = 60,000$. The second most frequent word is predicted to have frequency $C/2 = 30,000$, half the frequency of the first word. The third most frequent word will have frequency $C/3 = 20,000$, one third of the first word. On the other hand, the 100th most frequent word (the word with rank 100) will have frequency $C/100 = 600$. The 101st most frequent word will have frequency $C/101 = 594.06$, i.e. about 99% of the frequency of the 100th word. The 102nd most frequent word will have frequency $C/102 = 588.23$, about 98% of the frequency of the 100th word. Thus, the model predicts a very rapid decrease in frequency among the most frequent words, which becomes slower as the rank grows, leaving very long tails of words with similar low frequencies. Zipf's law does not predict frequency ties, since there are no ties among ranks, but it approximates the empirically attested plateaus by predicting a very large number of words with very similar non-integer frequencies. For example, the model above with a set to 1 and C set to 60,000 predicts that about 80,000 words will have frequencies between 1.5 and 0.5!

Zipf's law is an inverse power function, i.e., frequency is proportional to a negative power ($-a$) of rank. That frequency decreases when rank increases is obvious, given that ranks are based on frequency. However, compared to other distributions commonly used to model decay in natural and artificial phenomena, such as the exponential distribution, a power law distribution decreases more slowly, leaving a long tail of low frequency items. Zipfian distributions are not limited to word frequencies, but are also encountered in completely unrelated phenomena such as city populations, incomes (in economics, a variant of Zipf's law is known as *Pareto's law*), frequency of citations of scientific papers and visits to web-sites. It should be clear that these are all distributions of the few-giants/many-dwarves type. For a short survey of attested Zipfian distributions, see Li (2002).

Mathematically, Zipf's law has the convenient property that, if we take the logarithm of both sides, we obtain a linear function (recall that the logarithm of a fraction equals the difference of the logarithms of its numerator and denominator, and that $\log x^k$ equals $k \log x$):

$$\log f(w) = \log C - a \log r(w) \qquad (2)$$

This is the equation of a straight line with intercept $\log C$ and slope $-a$. Thus, Zipf's law predicts that rank/frequency profiles will appear as straight lines in double logarithmic space (i.e., plotting log frequency as a function of log rank). The values of the intercept and the slope (and thus of Zipf's law's parameters C and a) can be easily estimated using the standard method of least squares, implemented in most statistical packages (see, e.g., Dalgaard 2002 for the R implementation). Figure 7 presents some of the rank/frequency profiles we already saw plotted in double logarithmic space. As it can be seen, fitting a straight line to the log-log curves is not unreasonable (indeed, Zipf probably came up with his formula by looking at plots of this sort), although the fit is far from perfect, especially at the edges.

At the right edge of the curves, among the highest ranks (lowest frequencies), we notice a "bell-bottom" pattern due to the increasingly wider horizontal lines corresponding to the rare words that are assigned different ranks but have the same frequency. This is expected, since we are fitting a model predicting no ties (but many words with very near continuous frequencies) to an empirical curve that for high ranks is essentially a discrete step function. More worryingly, for the two largest corpora (BNC and *la Repubblica*) we observe a curvature suggesting that frequency is dropping more rapidly than what is predicted by Zipf's law. This tendency is already noticeable, to a lesser extent, in the Brown and Japanese web corpus curves. Zipf and other early scholars had no access to large corpora where the phenomenon is clear (we do not observe this curvature in *The War of the Worlds*). The BNC and *la Repubblica* plots suggest that we should perhaps
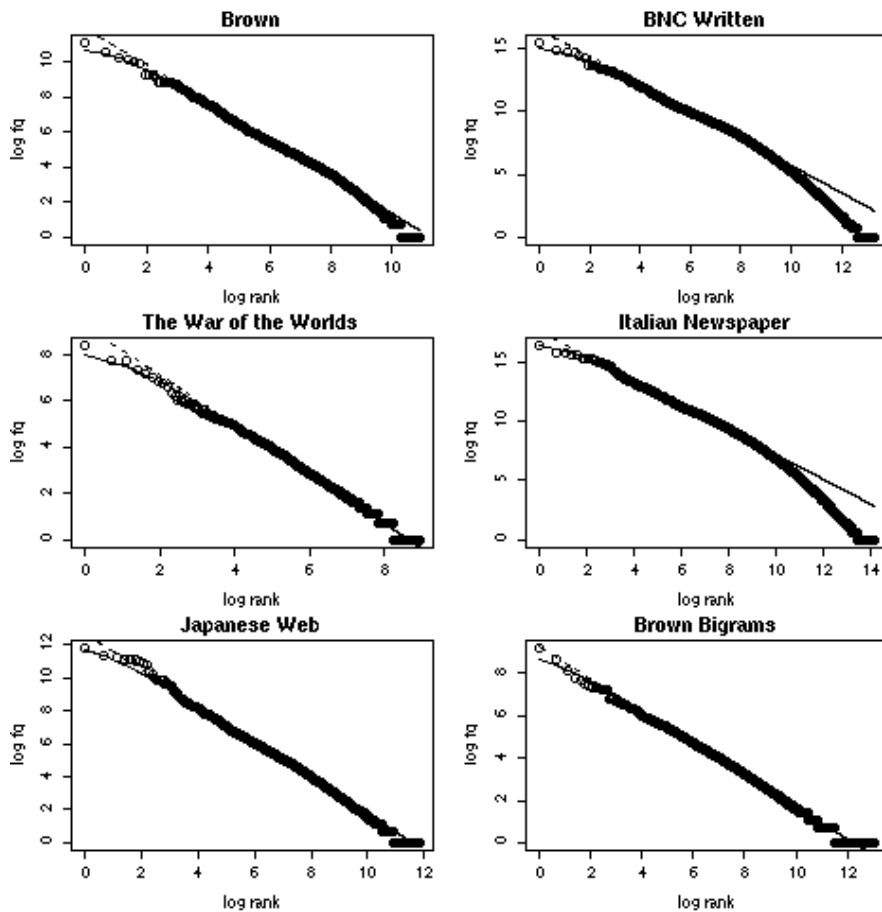
Figure 7: Log rank/log frequency plots with Zipf and Zipf-Mandelbrot fits for the Brown (top left), written BNC (top right), *The War of the Worlds* (middle left), *la Repubblica* (middle right), the Japanese web-page corpus (bottom left), the Brown bigrams (bottom right).

be fitting two straight lines to the data: one for the top ranks and one, with a steeper slope, for the bottom ranks. Indeed, Ha/Stewart/Hanna/et al. (2006) obtain a good fit to a large English corpus with two lines, one for the top 5000 ranks and another (with a slope twice as steep) for the remaining ranks, and present similar results for other languages.

At the other end of the plot (low ranks, high frequencies) we observe, again, a downward curvature of the empirical profile, i.e., the attested high frequencies tend to be lower than what would be predicted by their rank according to Zipf's law. The pattern was observed early on, and Mandelbrot (1953) added a parameter to Zipf's law to take care of this downward curvature:

$$f(w) = \frac{C}{(r(w) + b)^a} \tag{3}$$

Zipf's original law is a special case of Zipf-Mandelbrot's law with $b = 0$. A reasonably small value of b will lower the frequency of the first few ranks in a significant manner but it will hardly affect higher ranks. For example, if we assume like above that $C = 60,000$ and $a = 1$, and furthermore that $b = 1$, then for the most frequent word Zipf's law (equation 1) predicts a frequency of $60,000/1 = 60,000$ whereas the Zipf-Mandelbrot's formula (equation 3) predicts half this frequency: $60,000/(1 + 1) = 30,000$. On the other hand, for the word with rank 1000 the difference in predicted frequency between the two formulas is minimal ($60,000/1,000 = 60$ with Zipf's formula, and $60,000/1,001 = 59.94$ with Mandelbrot's variant). Mandelbrot's formula no longer corresponds to a straight line in double logarithmic space:

$$\log f(w) = \log C - a \log(r(w) - b) \tag{4}$$

This makes sense empirically since we just saw that the log rank/log frequency profiles are not quite straight lines, but it complicates the math since we can no longer use a simple least squares linear fit model as with Zipf's original equation. In my experience, reasonable fits can be obtained by first setting b to 0 and calculating $\log C$ and a with the least squares method, and then increasing b in small steps until the goodness of fit of equation (4) applied to the first few ranks (those that will be considerably below the predicted straight line) stops improving. Figure 7 presents Zipf and Zipf-Mandelbrot fits to the empirical frequency rank profiles (as dashed and continuous lines, respectively). The Zipf parameters were found with the least squares method applied to the first 10,000 ranks. The extra Zipf-Mandelbrot parameter b was calculated with the method I just described, applied to the top 20 ranks (top 2 ranks in the Japanese corpus). As expected, in all plots the difference between the Zipf and Zipf-Mandelbrot curves is noticeable only for the lowest ranks (top left).
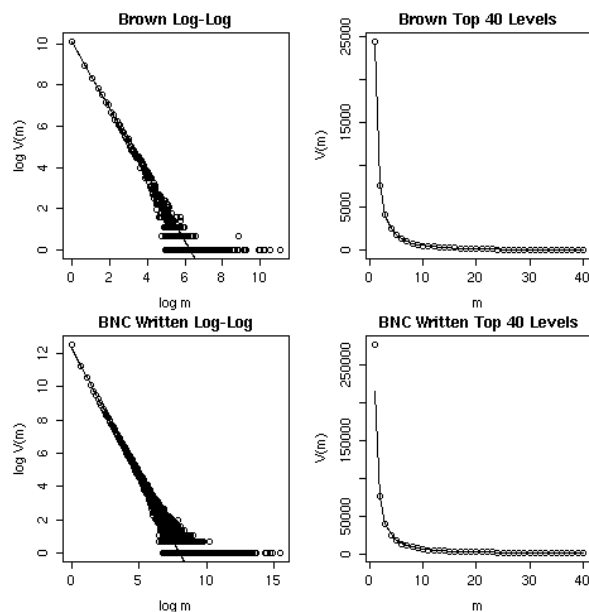
Figure 8: Frequency spectra in log-log space (left panels) and first 40 classes of the frequency spectrum (right panels) in the Brown (top row) and written BNC (bottom row), with Zipfian fits based on equation (5).

The a parameter is close to 1 for all the word frequency curves, ranging from 1.04 (*la Repubblica*) to 1.09 (*BNC* and Japanese web corpus). The tendency of a to be close to 1 is well known, and it justifies the simplified version of Zipf's law sometimes found in the literature, in which the formula is reduced to $f = C/r$, by assuming $a = 1$.

In the Brown bigram rank/frequency profile, the estimated a value is 0.76, well below the values typical of single word curves. Also, the plot suggests that for bigrams there is no need for the extra parameter b. The bigram frequencies look most decidedly like a straight line, without clear signs of downward curvatures at the top or bottom. Zipf's law may provide a better fit to ngram distributions than to single words (Ha/Sicilia-Garcia/Ming/et al. 2002).

3.1 Zipf's law for frequency spectra

Zipf (1965, p. 40 and ff.) also analyzed the frequency spectrum in terms of a power law of the form:

$$V(f) = \frac{C}{f^a} \tag{5}$$

Again, the parameters can be estimated with a simple linear least squares fit in double logarithmic space. Figure 8 shows that Zipf's power law for

15

frequency spectra provides reasonable fits to the Brown and BNC corpora (parameters estimated with the least squares method using the top 50 frequency classes).

Observing how Zipf(-Mandelbrot)'s law for rank/frequency profiles fits high frequency words better and how the frequency spectrum law fits low frequency words better, Naranan/Balasubrahmanyan (1998) propose to use (a variation of) the former to model function words and (a variation of) the latter to model content words.

3.2   Explanations of Zipf's law

Language after language, we find that Zipf(-Mandelbrot)'s law fits the data reasonably well. This has prompted many scholars to seek an explanation for this pattern. Zipf famously proposed to interpret it in terms of a "least effort" principle: the tension between the goal of the speaker to minimize production efforts by using only few words very frequently and the goal of the listener to minimize perceptual confusion by having a large vocabulary of distinct words would lead to the compromise distribution predicted by Zipf's law, with few high frequency types and many low frequency types. Mandelbrot derived his version of the law from information-theoretic notions, as the optimal solution to the problem of minimizing the average cost per unit of information in a text. Taking a different approach, other scholars (most famously, Simon 1955), observing how widespread Zipf's law is across phenomena that are clearly not related (such as word frequency distributions, city sizes and income distributions), studied under which general conditions such a distribution might arise.

Interestingly, texts constructed by generating characters (including white space) in random order also exhibit a Zipfian distribution (Miller 1957; Li 1992). Intuitively, when combining characters randomly, short words will be few but much more likely to occur by chance, whereas long words will be many but each of them will be extremely unlikely. Thus, in the output of the random generation process we will observe the by-now-familiar few-giants/many-dwarves pattern. Some authors (e.g., Miller 1957) take the fact that random text has a Zipfian distribution as evidence against "deep" explanations of Zipf's law in terms of principles of language or communication. However, unlike in random text generation, the frequency with which a speaker selects a word will not depend on the length of the characters that compose it (the effect, as already observed by Zipf, is likely to go in the other direction, with a tendency for more frequently used words to be shortened). Thus, the random text experiments are not "explaining" Zipf's law in natural language in any psychologically plausible sense.

4   Practical consequences

Although most of the literature on word frequency distributions is highly theoretical, the basic patterns of frequency in corpora have important consequences in practical work. First and most importantly, the Zipfian nature of word frequency distributions causes data sparseness problems. No matter how large a corpus is, most of the words occurring in it have very low frequency and a small set of frequent words constitutes the large majority of the tokens in the corpus. The distribution of bigrams and linguistic units larger than the word is even more skewed. Anybody working with corpora should be aware of these facts.

For example, according to the guidelines in Sinclair (2005), a trained lexicographer will need to inspect at least 20 instances of an unambiguous word to get an idea of its behavior. Even in a large corpus such as the (written) BNC, a lexicographer will find that less than 14% of the words have a frequency of 20 or higher. In a completely different area, Möbius (2003) observes that speech synthesis researchers often accept poor modeling of rare words (and other relevant units) in virtue of the fact that they are rare. However, Möbius observes that, because of the Zipfian nature of linguistic data, although each rare unit has a very low probability to occur, the overall probability that at least one rare unit will occur in a sentence approaches certainty.

Another facet of the data sparseness problem is that even large corpora do not sample the whole vocabulary of the language they represent: as sample increases, the number of types (vocabulary size) keeps increasing. This is illustrated for the Brown corpus in figure 9, where I plotted the overall number of types (V) and the number of words occurring once (V(1)) found in the first 100K, 200K, etc. tokens, up to the full corpus size. Even at full corpus size the vocabulary is still growing.
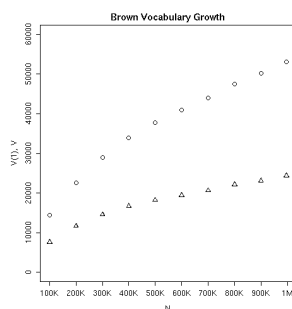

Figure 9: The Brown corpus vocabulary growth curve: number of types (circles) and hapax legomena (triangles) for 10 increasingly larger token samples (N).

Baayen (2001, pp. 49-50) shows that the growth rate of the vocabulary,

the rate at which the vocabulary size increases as sample size increases, can be estimated as follows:

$$G = \frac{V(1)}{N} \qquad (6)$$

In equation (6), $V(1)$ is the number of words occurring once (*hapax legomena*, Ancient Greek for "said once") in a sample of size N. The formula should make intuitive sense: the proportion of hapax legomena that we encountered up to the Nth token is a reasonable estimate of how likely it is that word $N + 1$ will be a hapax legomenon, i.e., a word that we have not seen before and that will consequently increase vocabulary size. In the Brown corpus, $G = 24375/996883 = .024$, indicating that the vocabulary size is still growing at a relatively fast pace. The vocabulary is still growing (although at a slower pace) in much larger corpora, such as the written section of the BNC ($G = .003$) and the *la Repubblica* corpus ($G = .002$).

An important consequence of the fact that even large corpora are not sampling the full vocabulary of types they are drawn from is that the standard method of estimating the probability of occurrence of a word (or ngram) by its relative frequency in a corpus is very inaccurate. On the one hand, the word types that are not in the corpus are wrongly assigned 0 probability. On the other hand, the probability of the words that do occur in the corpus is overestimated, since they take up probability mass that should have been assigned to unseen words. Indeed, much work in corpus-based computational linguistics (see, e.g., Manning/Schütze 1999) focuses on ways to solve problems deriving from data sparseness, e.g., by assigning some probability mass to unseen words/ngrams with heuristic methods, or by clustering words into classes to obtain more robust statistics, or by using massive data collections, such as the web.

Another consequence of the fact that V keeps growing with corpus size is that we cannot use it as a measure of lexical richness when comparing corpora of different sizes: larger corpora will tend, trivially, to have more types. The fact that V increases with N (in ways that are not captured by simple functional relations) also affects nearly all the "constants" that have been proposed in the literature as measures of lexical richness (Tweedie/Baayen 1998), which turn out to vary with corpus size, and thus are not true constants. Statistical models of word frequency distributions (such as those introduced in Baayen 2001) provide formulas for the expectation (mean) and variance of quantities such as vocabulary size at arbitrary sample sizes. Thus, they allow the comparison of corpora of different sizes (we can compute, e.g., the expected number of types we would see in a smaller corpus X if we could "stretch" it to the length of a larger corpus Y), and, under certain assumptions, to assess whether the vocabulary size difference between the corpora is statistically significant. Word frequencies require specialized statistical models, since some crucial aspects of standard methods, such as

the assumption that the central limit theorem guarantees the normality of sample averages for reasonably large samples, are not appropriate for the extremely skewed word/ngram frequency data. Lexical-statistical models have been applied most extensively in stylometry and in the study of morphological productivity (see Articles 52 and 43 of this handbook, respectively), but also in terminology (Kageura 1998) and collocation mining (Evert 2004). Unfortunately, the experiments reported in Evert/Baroni (2006) indicate that the prediction quality of current lexical-statistical models is not very high, probably because typical corpus data severely violate the randomness assumption that lies at the core of statistical modeling.

The Zipfian distribution of word frequencies is not only "bad news". The fact that we can expect words in pretty much any natural language text to have this distribution (and the coefficient a to be close to 1) has found many applications, ranging from index compression (Baldi/Frasconi/Smyth 2003, section 4.1.2) to term weighting in information retrieval (Witten/Moffat/Bell 1999, section 4.4), to cryptography (Landini/Zandbergen 1998), to Bayesian modeling of morpheme frequencies (Creutz 2003).

5 Conclusion

This article introduced the typical patterns of frequency distribution encountered in corpora/texts, it proposed Zipf-Mandelbrot's law as a descriptive model that captures such patterns and it illustrated some of the consequences of these patterns for corpus-based work. Of course, I only scratched the surface of the large body of studies on lexical statistics.

The interested reader should proceed to Baayen (2001), a very thorough (and mathematically challenging) introduction to word frequency distributions with an emphasis on statistical modeling. I am not aware of contemporary introductions to lexical statistics at a less advanced level. Muller (1977) is an introduction in French to the basic concepts of word frequency analysis.

The *Journal of Quantitative Linguistics* and *Glottometrics* often feature articles on relevant topics. In 2002, the latter published three special issues in honor of George Kingsley Zipf. The recent HSK handbook on quantitative linguistics (Köhler/Altmann/Piotrowski 2005) features several articles on various aspects and applications of lexical statistics.

For those interested in the hands-on approach, the LEXSTATS software developed by Harald Baayen for the statistical analysis of word frequency distributions is freely available from his site:

`http://www.mpi.nl/world/persons/private/baayen`

The free zipfR package developed by Stefan Evert and myself provides

similar functionalities, and it is integrated in the popular open source statistical package R. For information, visit the site:

http://purl.org/stefan.evert/zipfr

Finally, Wentian Li maintains a very up-to-date Internet bibliography on Zipf's law and related principles at:

http://www.nslij-genetics.org/wli/zipf

Acknowledgments

References

Abney, Steven (1996), Statistical methods and linguistics. In: Klavans, J. & Resnik, P. (eds) *The balancing act: Combining symbolic and statistical approaches to language.* Cambridge MA: MIT Press, 1-23.

Baayen, Harald (2001), *Word frequency distributions.* Dordrecht: Kluwer.

Baldi, Pierre/Frasconi, Paolo/Smyth, Padhraic (2003), *Modeling the Internet and the web.* Chichester: Wiley.

Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus linguistics.* Cambridge: Cambridge University Press.

Creutz, Mathias (2003), Unsupervised segmentation of words using prior distributions of morph length and frequency. In: *Proceedings of ACL 03*, 280-287.

Dalgaard, Peter (2002), *Introductory statistics with R.* New York: Springer.

Evert, Stefan (2004), *The statistics of word cooccurrences: Word pairs and collocations.* PhD thesis, University of Stuttgart/IMS.

Evert, Stefan/Baroni, Marco (2006), Testing the extrapolation quality of word frequency models. In: *Proceedings of Corpus Linguistics 2005*, available from http://www.corpus.bham.ac.uk/PCLC

Ha, Le Quan/Sicilia-Garcia, Elvira/Ming, Ji/Smith, Francis (2002), Extension of Zipf's law to words and phrases. In: *Proceedings of COLING 2002*, 315-320.

Ha, Le Quan/Stewart, Darryl/Hanna, Philip/Smith, Francis (2006), Zipf and Type-Token rules for the English, Spanish, Irish and Latin languages. In: *Web Journal of Formal, Computational & Cognitive Linguistics*, 8, available from `http://fccl.ksu.ru/`

Kageura, Kyo (1998), A Statistical Analysis of Morphemes in Japanese Terminology. In: *Proceedings of COLING-ACL 98*, 638-645.

Köhler, Reinhard/Altmann, Gabriel/Piotrowski, Rajmund (eds) (2005). *Quantitative linguistics: An international handbook*. Berlin: Mouton de Gruyter.

Landini, Gabriel/Zandbergen, René (1998), A well-kept secret of mediaeval science: The Voynich manuscript. In: *Aesculapius* 1998, available from `http://www.voynich.nu/extra/aes.html`

Li, Wentian (1992), Random texts exhibit Zipf's-law-like word frequency distribution. In: *IEEE Transactions on Information Theory* 38, 1842-1845.

Li, Wentian (2002), Zipf's Law everywhere. In: *Glottometrics* 5, 14-21.

Mandelbrot, Benoit (1953), An informational theory of the statistical structure of languages", in Jackson, W. (ed) *Communication theory*. London: Butterworth, 486-502.

Manning, Christopher/Schütze, Hinrich (1999), *Foundations of statistical natural language processing*. Cambridge MA: MIT Press.

Matsumoto, Yuji/Kitauchi, Akira/Yamashita, Tatsuo/Hirano, Yoshitaka/Matsuda, Hiroshi/Takaoka, Kazuma/Asahara, Masayuki (2000), *Morphological analysis system ChaSen version 2.2.1 manual*. NIST Technical Report.

McEnery, Tony and Andrew Wilson (2001), *Corpus linguistics, 2nd edition*. Edinburgh: Edinburgh University Press.

Miller, George (1957), Some effects of intermittent silence. In: *American Journal of Psychology* 52, 311-314.

Möbius, Bernd (2003), Rare events and closed domains: Two delicate concepts in speech synthesis. In: *International Journal of Speech Technology* 6, 57-71.

Muller, Charles (1977), *Principes et méthodes de statistique lexicale.* Paris: Hachette.

Naranan, S./Balasubrahmanyan V. K. (1998), Models for power law relations in linguistics and information science. In: *Journal of Quantitative Linguistics* 5, 35-61.

Oakes, Michael (1998), *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Sampson, Geoffrey (2002), Review of Harald Baayen: Word Frequency Distributions. In: *Computational Linguistics* 28, 565-569.

Simon, Herbert (1955), On a class of skew distribution functions. In: *Biometrika* 42, 425-440.

Sinclair, John (2005), Corpus and text: Basic principles. In Wynne, M. (ed) *Guide to good practice in developing linguistic corpora*, available from `http://ahds.ac.uk/litlangling/linguistics/index.html`

Ueyama, Motoko/Baroni, Marco (2006), Automated construction and evaluation of a Japanese web-based reference corpus. In: *Proceedings of Corpus Linguistics 2005*, available from `http://www.corpus.bham.ac.uk/PCLC`

Tweedie, Fiona/Baayen, Harald (1998), How variable may a constant be? Measures of lexical richness in perspective. In: *Computers and the Humanities* 32, 323-352.

Witten, Ian/Moffat, Alistair/Bell, Timothy (1999), *Managing gigabytes, 2nd edition.* San Francisco: Morgan Kaufmann.

Zipf, George Kingsley (1949), *Human behavior and the principle of least effort.* Cambridge MA: Addison-Wesley.

Zipf, George Kingsley (1965), *The psycho-biology of language.* Cambridge MA: MIT Press.