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.

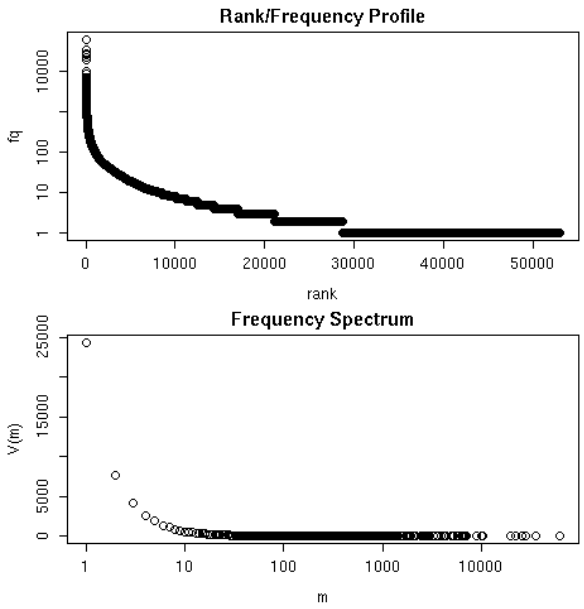

Figure 1: Rank/frequency profile and frequency spectrum of the Brown corpus.

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.

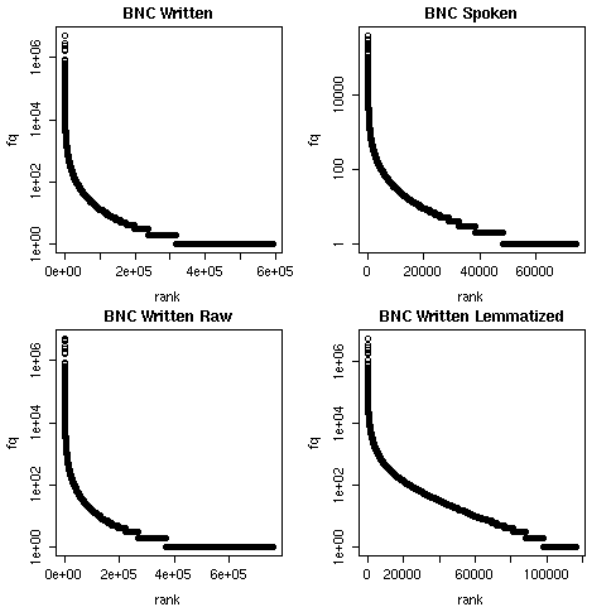

Figure 2: Rank/frequency profiles of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right).

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

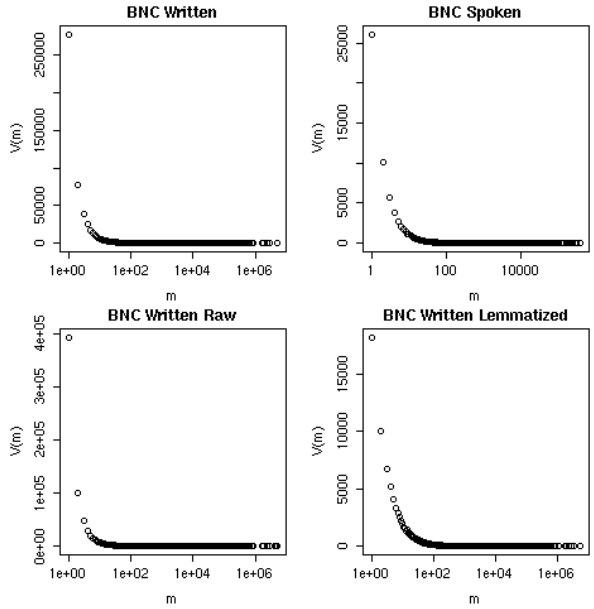

Figure 3: Frequency spectra of the written (top left) and spoken (top right) sections of the BNC, of the written BNC with minimal pre-processing (bottom left) and of the lemmatized written BNC (bottom right).

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,