Zipf's Law: What and Why?

Łukasz Dębowski
Institute of Computer Science, Polish Academy of Sciences
ldebowsk@ipipan.waw.pl, http://www.ipipan.waw.pl/~ldebowsk/

December 20, 2000

Abstract

These notes form an explanatory introduction to Zipf's law and resume some of the existing literature.

1 Introduction

As the basic statistical properties of language texts, one could consider some elementary formulas about the quantitative entities which could be easily defined and which would be obeyed for all human languages. At the first glance, one might be skeptical about the existence of such simple formulas. Human languages are very diverse and qualitatively complex. Why should there be any simple quantitative description of them? Such descriptions, however, exist. The most famous quantitative law of language is Zipf's law and it deals with the frequency distribution of words.

Before formulating this quantitative law, one should introduce some quantities it can deal with. Token t_i of a text will be meant as an occurrence of a word on the i-th position in the text, where $i = 1, \ldots, N$ and N is the number of tokens. Word w_j of a text will be meant as a word listed once on the j-th position in a lexicon collected out of the text, where $j = 1, \ldots, A$ and A is the number of words. For each position i there is exactly one position j such that $t_i = w_j$. For each position j there is one or more positions i such that $t_i = w_j$. For any texts representative in a given language, the following quantities can be introduced:

- $f = f(w)$ – frequency of word w,
- $F = F(f)$ – frequency of frequency f,
- $r = r(w)$ – rank of word w.

Frequency $f(w)$ is the number of tokens being word w. [1] Since frequency $f(w)$ is discrete-valued, one can easily define frequency of frequency $F(f)$ as the number of words w for which $f(w) = f$. Having sorted the list of words against the frequencies descendently, one can define rank $r(w)$ as the position of word w on the sorted list where the positions are numbered starting at 1. In order to determine the value of $r(w)$ for

[1] Strictly speaking, $f(w)$ should be called "count" but name "frequency" is preferred in literature.

$F(f(w)) > 1$, it is assumed here that the sorting procedure chooses at random one of the possible monotonic lists, so each word has a different rank.

By the definition of rank, frequency $f(w)$ is a monotonically decreasing function of rank $r(w)$. The definition of rank does not state, however, what kind of decreasing function of $r(w)$ frequency $f(w)$ is. The problem whether there is some specific correspondence between these quantities was investigated by Zipf. Strictly speaking, it was not Zipf who explored the problem first. More information about the history can be found in [MS99, Li, Sam72]. Nevertheless, it was Zipf's works [Zip35, Zip49] that became famous. In his works, Zipf surveyed a formula which he observed for long texts in English and which reads: [2]

$$f(w) \propto \frac{1}{r(w)}. \tag{1}$$

Besides this formula, which deals with almost no qualitative concepts, Zipf discovered some other quantitative laws dealing with semantics, namely statistics of homonymy and polysemy. The subject will be omitted here.

It was checked that formula (1) holds also for other human languages, as well as for the artificial ones (e.g. programming languages). Some examples are given in [Sam72], [MS99]. Formula (1), known simply as Zipf's law, is observed not only in linguistics, but also in ecology, sociology, economics, and physics. The law is universal for many statistics in which dependencies between ranks and frequencies are investigated.

There is a rich and variously mathematically advanced literature about Zipf's law, which is scattered over multiple journals in many domains. Some interdisciplinary references can be found on-line at site [Li]. A book on specific investigations by quantitative linguists is [GA82]. These notes have been compiled after some consultation with [MS99], [Li92], [Sam72], [Sam88], and [GLSW96].

Why is law (1) so universal? One might think that it must be caused by some principle which is universal as well. In fact, the kind of responsible principle may seem disappointing. Zipf's law can be derived as a consequence of very simple chaos. The key point is the observation that texts consisting of randomly generated letters and spaces also obey the law. The fact was already recognized by Miller in an introduction to the Zipf's book in 1965 [Mil65] but some simple mathematical explanation at high-school level was given by Li as late as in 1992 [Li92].

In the following sections, detailed calculations are presented for several aspects of Zipf's law.

- In section 2, the necessity of Mandelbrot's correction of formula (1) is explained. Furthermore, a very simple estimate of absolute frequencies $f(w)$ is given for the most common words.

- In section 3, Zipf's law is derived for random texts using slightly more general model than in [Li92]. The differences and similarities in parameters for random texts and texts in meaningful languages are sketched, as well.

- In section 4, the difference between Zipf's law for a set of texts representing the whole language and Zipf's law for a finite text is discussed. An effect of ranking words for finite texts calculated in [GLSW96] is cited. Extrapolating the quoted formula, some idea is presented how frequencies of frequencies $F(f)$ and number of tokens N may change depending on frequency f and number of words A.

[2] Symbol $\propto$ stands for a proportionality relation.

- In section 5, a model inspired by [Orl82] is presented which extends Mandelbrot's correction onto the ensembles of finite texts generated by the same source. For such ensembles, the number of free parameters is reduced in comparison to Mandelbrot's correction for a single text.

Section 6 summarizes the notes briefly.

2 Mandelbrot's correction

Zipf's law has fascinated Benoit M. Mandelbrot, the same person who later made his famous research in fractal geometry. In 1954, he published an article [Man54] in which he discussed Zipf's law (1). He corrected it into the form:

$$f(w) \propto \frac{1}{[r(w)+\rho]^{1+\epsilon}}, \tag{2}$$

which fitted the language data better. The correction contained two new constants: $0 < \epsilon \ll 1$ and $\rho \gg 1$,

Mandelbrot made some more remarks on law (2) in his book on fractals and scaling power laws [Man83]. He wrote there that it was Zipf's law that among other observations had inspired him to discover fractal geometry. Mandelbrot, however, followed some Zipf's teleological view present in [Zip35, Zip49], and tried to justify power laws (1), (2) as an optimization effect of competition in lowering effort between the speaker and the hearer.

Having learned about the work of Li [Li92], showing the obeyance of Zipf's law by random texts, we got reluctant to perceive the high-rank behavior of this law as an optimization effect. We would like to underline here an opposite view. The $1/r(w)^{1+\epsilon}$ behavior for large ranks $r(w)$ can be fully explained, including some estimate for the value of ϵ, by mere random letter effects.

In order to foresee the analogy, one may observe that the number of different words grows exponentially with respect to their length. It is an effect of word building which constructs the vast majority of lexicon using a finite number of short elements. If one represents the text in terms of these elements rather than letters, one will obtain a more chaotic text, much more resembling Li's text of random letters and spaces.

More interesting in Mandelbrot's correction (2) is a possible cause of its low-rank behavior. Because $\rho \gg 1$ for natural languages, the frequencies for the lowest ranks do not decrease so rapidly with respect to the ranks. For texts of random letters $\rho \approx 1$, as it is shown in section 4. The difference between the random texts and the natural languages may be due to the existence of some grammatical structure in the latter ones. The structure makes some functional words appear often, regularly and independently of style, following some approximately constant number of substantive tokens. Usually these functional words occupy the lowest ranks. From the point of view of language learning, it is nice that such a simple rank statistics can filter out so well the most important functional words. An effective learning algorithm for grammar acquisition from texts should use the rank information.

In order that a language could have some grammatical structure, it needs not only ρ to be much bigger than 1 but firstly, ϵ must be greater than 0. The role of ρ has been just explained. What is the role of $\epsilon > 0$? One can try to find probability $p(w)$ of a word. It can be estimated as $p(w) = f(w)/N$, where N—the number of tokens—is the

normalization constant. If one had exactly (1), one could compute the normalization summing by ranks:

$$p(w) = \frac{f(w)}{N} = \frac{f(w)}{\sum_{w'} f(w')} = \frac{1/r(w)}{\sum_{r=1}^{A} 1/r}, \tag{3}$$

where A, the number of words, is the highest rank. Asymptotically:

$$\sum_{r=1}^{A} \frac{1}{r} \approx \log A, \tag{4}$$

so one would get:

$$p(w) \approx \frac{1/r(w)}{\log A}. \tag{5}$$

If formula (5) were true, the larger one had the text the smaller all probabilities would be. In the limit of infinite A, all probabilities decrease to 0 since also $\log A \to \infty$.

It was claimed, however, that the natural language has a grammatical structure. If one considers such a language, there must be some functional words which appear regularly and for which $p(w)$ must be well-defined as non-zero values. Thus, ϵ might not be less or equal to 0. Of course, inequality $\epsilon > 0$ holds also for any language in which probabilities of words are well-defined, including texts of random letters which do not have any grammatical structure.

One can approximate the normalization in case of $\epsilon > 0$ as:

$$\sum_{r=1}^{\infty} \frac{1}{(r+\rho)^{1+\epsilon}} \approx \int_0^{\infty} \frac{1}{(r+\rho)^{1+\epsilon}} dr = \frac{1}{\epsilon \rho^{\epsilon}}. \tag{6}$$

Now, proportionality (2) can be enriched with the approximated normalization:

$$f(w) \approx \frac{\epsilon \rho^{\epsilon}}{[r(w)+\rho]^{1+\epsilon}} \cdot N. \tag{7}$$

For the most frequent words, their ranks are small in comparison with ρ and one may approximate:

$$\max_{w'} f(w') \approx \frac{\epsilon}{\rho} \cdot N. \tag{8}$$

In the real natural languages, the parameters are: $\epsilon \approx 10^{-1}$, $\rho \in [10^1, 10^2]$ so the most frequent words usually have their frequencies in range of $[0.001N, 0.01N]$.

3 Random texts

This section presents an argumentation that also the random texts exhibit Zipf's law. A following probabilistic model is adopted for generating the consecutive letters and spaces of a random text. It is assumed that the characters are tossed independently and the parameters are:

- p – probability of generating any chosen letter,
- a – length of the alphabet,
- $(1 - ap)$ – probability of generating a space.

The introduced zeroth order Markov model is slightly more complex than the model presented in [Li92], which fixes $(1-ap)=p$, or $p=1/(a+1)$. Let $p(l)$ be the probability of any chosen l-letter-long word, let $n(l)$ be the number of l-letter-long words, and let $N(l)$ be the number of words that are not longer than l. It is obvious that:

$$p(0)=(1-ap), \quad p(l+1)=p(l)\cdot p \quad \Rightarrow \quad p(l)=p^l(1-ap), \tag{9}$$

$$n(0)=1, \quad n(l+1)=n(l)\cdot a \quad \Rightarrow \quad n(l)=a^l, \tag{10}$$

$$N(0)=1, \quad N(l+1)=N(l)\cdot a+1 \quad \Rightarrow \quad N(l)=\begin{cases} l & \text{if} \quad a=1, \\ (a^{l+1}-1)/(a-1) & \text{if} \quad a\neq 1. \end{cases} \tag{11}$$

Of course, the normalization holds:

$$\sum_{l=0}^{\infty} p(l)n(l)=1. \tag{12}$$

From (9), one can observe that the frequency of a l-letter-long word decays exponentially with respect to word length l: $f(l)/N \approx p(l)=(1-ap)\exp\left[l\log p\right]$. Straightforwardly, one gets:

$$l \approx \frac{\log\left[f(l)/N\right]}{\log p}-\frac{\log\left[1-ap\right]}{\log p}. \tag{13}$$

In the model, the longer the word is the less frequent it is. It is reasonable to assume that the maximal rank $r(l)$ of a l-letter-long word is estimated by $N(l)-1$, viz. [Li92]. In the estimator, 1 was subtracted from $N(l)$ because only the non-empty words are counted for ranking. Using this approximation, one gets an exponential decay for $a=1$:

$$r(l) \approx N(l)-1=l-1 \quad \Longrightarrow \quad l \approx r(l)+1, \tag{14}$$

$$p(l)=p^l(1-ap) \approx p^{r(l)+1}(1-ap), \tag{15}$$

but for $a\neq 1$, one obtains a power law:

$$r(l) \approx N(l)-1=\frac{a^{l+1}}{a-1}-\frac{a}{a-1} \quad \Longrightarrow \quad l \approx \frac{\log\left[r(l)+\frac{a}{a-1}\right]-\log\left[\frac{a}{a-1}\right]}{\log a}, \tag{16}$$

$$p(l)=p^l(1-ap) \approx \exp\left[\frac{\log p}{\log a}\log\left[\frac{r(l)+\rho}{\rho}\right]\right](1-ap)=\left[\frac{\rho}{r(l)+\rho}\right]^{1+\epsilon}(1-ap), \tag{17}$$

where:

$$\rho=\frac{a}{a-1}, \quad \epsilon=-\frac{\log p}{\log a}-1. \tag{18}$$

Power law (17) is exactly Zipf's law. Some estimates have been obtained for parameters ρ and ϵ, as well. The longer the alphabet is the closer ρ is to 1. Condition $\rho \approx 1$ need not hold so well for the natural languages since it leaves quite little room for the low-rank functional words. On the other hand, estimate (18) works well for ϵ also in the natural languages. According to [Li92], using simply $p=1/(a+1)$ and Latin alphabet value $a=26$ leads to $\epsilon \approx 0.12$, which is close to the value for English. If one adopted a first order Markov model, in which characters are tossed along the unigram letter distribution instead of constant probability p, probably one could estimate ϵ even better.

4 Hapax legomena

Up till now, no effects have been considered which are caused by the finiteness of language text for which Zipf's law is formulated. Actually, it was only argued what should be the law like if the text were so large and representative that all fluctuations of interesting functions in it could be neglected. Such case is called in statistical physics thermodynamic limit. In fact, language is such a complex system of very long-distance correlations, that even collecting huge amounts of texts, one is unable to approach the thermodynamic limit so efficiently as it is possible in natural systems explored in physics. Because of the complexity of the correlations, the obtainable linguistic data are mostly sparse. The problem how this sparseness scales with respect to the data size is very interesting both theoretically and practically. Some fundamental calculations on empirical ranking that can be used to estimate the scaling of sparseness were presented in the article [GLSW96]. Their results will be reported on shortly.

A system that is not in its thermodynamic limit will be called briefly a finite one. An infinite system will stand for a system which is large enough to be practically in its thermodynamic limit. If one restricts oneself to a finite text, one can no longer make the assumption that there exists a strict proportionality of the observed frequency to the theoretical probability: $f(w) \propto p(w)$. The proportionality is the empirical definition of probability $p(w)$ but only in the infinite text. In a finite text of N tokens, observed word frequency $f(w)$ must be treated as a random variable. The mean of $f(w)$ is approximately equal to $Np(w)$ but the standard deviation may not be left out of account. Furthermore, it is rather $p(w)$ than $f(w)$ itself which strictly obeys Zipf's law (2). To deal with the empirical ranking, one should also introduce finite-text ranks $R(w)$, which are counted using sorted $f(w)$ values while infinite-text ranks $r(w)$ are counted using sorted $p(w)$ values. $R(w)$ need not be equal to $r(w)$.

Another problem is the discreteness. Frequency $f(w)$ may be only a natural number while $p(w)$ can be real. Given word w, the longer the available text is, the better $f(w)/N$ can fit any value of $p(w)$. For any finite text there will be, however, numerous rare words which appear just several times and no reasonable value of $p(w)$ could be estimated for them from the data. This problem is one of the major in machine learning of natural language since learning needs some capability of generalization and it is impossible to generalize well out of too sparse data. As it was mentioned, it may be useful to know how large the scale of the sparseness is on average and how it changes with respect to the data size.

Günther and others have proposed in [GLSW96] some ecologically inspired Markov model of text growth for finite texts. The model leads to the following geometrical probability distribution of $f(w)$ for words that appear at least once:

$$P(f(w) = f) = \frac{1}{Np(w)} \left[1 - \frac{1}{Np(w)}\right]^{f-1}, \quad f \geq 1, \quad Np(w) > 1, \tag{19}$$

and:

$$p(w) \propto \frac{1}{r(w)^{1+\epsilon}}. \tag{20}$$

The mean and the standard deviation are:

$$E(f(w)) = Np(w), \tag{21}$$

$$\sqrt{D^2(f(w))} = Np(w)\sqrt{1 - \frac{1}{Np(w)}}. \tag{22}$$

In the model, the standard deviation has the order of the mean and this property does not vary with respect to the text size. In majority of statistical phenomena in nature, standard deviation scales as square root of mean: $\sqrt{D^2(X)} \propto \sqrt{E(X)}$, if the mean itself scales as the system size N. Here, $\sqrt{D^2(X)} \propto E(X)$ was introduced arbitrarily to strongly differentiate finite-text ranks $R(w)$ from theoretical ranks $r(w)$. It is not obvious if such a scaling is artificial for natural language or not. The distribution of functional words is largely style independent. On the other hand, many substantive words can rank very low in certain texts of language while ranking very high in the others. There are some attempts to classify words by their cross-style behavior but still there is no sufficient mathematical model. To learn more, consult [Sam72].

For distribution (19), one cannot see well the Zipf's proportionality in a simple plot of $f(w)$ versus $r(w)$ because the standard deviation is very large. It is interesting to figure out how the plot of $f(w)$ versus $R(w)$ may look like. In order to do this, one needs to find conditional distribution $P(f(w) = f|R(w) = r)$. Using distribution (19) for $\epsilon = 0$, Günther and others have calculated in [GLSW96] that the conditional mean obeys a stepwise power law: [3]

$$E(f(w)|R(w) = r) \approx \lfloor \frac{A}{r} \rfloor, \tag{23}$$

where A is the number of words. A is kept constant so that distribution (20) be normalizable. Conditional variance $D^2(f(w)|R(w) = r)$ is much smaller than $D^2(f(w))$. One may think of the frequency as a function of the finite-text rank while numerically meaning $f(r) = E(f(w)|R(w) = r)$.

Ecological model (19) using simply $\epsilon = 0$ for very large texts is not appropriate for natural language, as it was argued in section 3. Nevertheless, formula (23) is convenient to find some simple estimates for frequencies of frequencies $F(f)$. Let $M(f)$ be the maximal rank in the set of words having frequency f. Then, $F(f) = M(f) - M(f + 1)$. For (23), one obtains $M(f) \approx A/f$ and:

$$F(f) \approx \frac{A}{f} - \frac{A}{f+1} = \frac{A}{f(f+1)}. \tag{24}$$

In this way, one could estimate the number of *hapax legomena*, i.e. the words that appear just once. It is simply $F(1) = A/2$ – as much as half the total number of words A (which is different to number of tokens N). It is a lot. In fact, even being less severe, the problem of *hapax legomena* challenges machine learning of natural language.

The formula (24) is correct only for $F(f) > 1$. Let frequency $\tilde{f} = f(\tilde{r})$ be the minimal one such that $F(\tilde{f}) = 1$. For (23), both frequency $\tilde{f}$ and its rank $\tilde{r}$ are square roots of A:

$$\tilde{f} \approx \sqrt{A}, \quad \tilde{r} \approx \sqrt{A}. \tag{25}$$

Values $\tilde{f}$ and $\tilde{r}$ will be called the middle frequency and rank. The words rarer than $\tilde{f}$ will be called simply rare words, and the words more frequent than $\tilde{f}$ will be called frequent words. A convenient way to count number of tokens N as a function of A is to sum the frequencies by the ranks up to $\tilde{r}$ and the frequencies of frequencies by the frequencies up to $\tilde{f}$:

$$N = \sum_{r=1}^{\tilde{r}} f(r) + \sum_{f=1}^{\tilde{f}} f \cdot F(f). \tag{26}$$

[3] The value of floor function $\lfloor x \rfloor$ is the biggest integer smaller than x.

Using (23) and (24), one obtains two symmetrical sums:

$$N \approx \sum_{r=1}^{\sqrt{A}} \frac{A}{r} + \sum_{f=1}^{\sqrt{A}} \frac{A}{f+1} \approx 2A(\log\sqrt{A} + \gamma) - A = A(\log A + 2\gamma - 1), \tag{27}$$

where Euler's gamma is: $\gamma \approx 0.577$. One can observe that for $\epsilon = 0$, number of tokens N grows slightly quicker than number of words A, which is a statistically positive phenomenon. In principle, the larger text one had the bigger the word frequencies would be and the better one could estimate $p(w)$. Unfortunately, not only the percentage of *hapax legomena* in the text decreases very slowly to zero—i.e. inverse-logarithmically—but so does the percentage of tokens being the most frequent word.

It is interesting to check the difference in the asymptotic behavior for the real case of $\epsilon > 0$ in order to learn if the problem of *hapax legomena* is less severe. By a quadruple analogy to (1), (2), and (23), one might assume that in the case of $\epsilon > 0$ and $\rho \approx 0$ there is:

$$E(f(w)|R(w) = r) \approx \lfloor \frac{A^{1+\epsilon}}{r^{1+\epsilon}} \rfloor, \tag{28}$$

with the usual meaning of the symbols. It is assumed that ϵ is small and only the linear expressions in ϵ are preserved. Then, the frequency of frequency is:

$$F(f) \approx \frac{A}{f^{1-\epsilon}} - \frac{A}{(f+1)^{1-\epsilon}} = (1-\epsilon) \cdot \frac{A}{f^{2-\epsilon}} - \frac{(1-\epsilon)(2-\epsilon)}{2!} \cdot \frac{A}{f^{3-\epsilon}} + \ldots, \tag{29}$$

where the expansion has its limit for $f > 1$.

For computing the middle frequency and rank, one may cut off $\mathcal{O}\left(\frac{1}{f^{3-\epsilon}}\right)$ in (29) and show that they are:

$$\tilde{f} = A^{(1+\epsilon)/2} \cdot \frac{1+\epsilon/2}{A^{\epsilon/4}}, \quad \tilde{r} = A^{(1+\epsilon)/2} \cdot \frac{A^{\epsilon/4}}{1+\epsilon/2}. \tag{30}$$

Approximating the sum in (26) by integrals, one obtains:

$$N \approx \int_1^{\tilde{r}} \frac{A^{1+\epsilon}}{r^{1+\epsilon}} dr + \int_1^{\tilde{f}} \frac{(1-\epsilon)A}{f^{1-\epsilon}} df = -\frac{A^{1+\epsilon}}{\epsilon r^{\epsilon}}\Big|_1^{\tilde{r}} + \frac{(1-\epsilon)A}{\epsilon f^{-\epsilon}}\Big|_1^{\tilde{f}}. \tag{31}$$

Performing the differences, one gets explicitly:

$$N \approx \frac{A}{\epsilon}\left[A^{\epsilon/2} - 1\right] A^{\epsilon/2} + \frac{A}{\epsilon}\left[A^{\epsilon/2} - 1\right](1-\epsilon) = \frac{A}{\epsilon}\left[A^{\epsilon/2} - 1\right]\left[A^{\epsilon/2} + (1-\epsilon)\right], \tag{32}$$

where the first component of the sum is the contribution of the frequent words, and the second one is the contribution of the rare words. One can see again that the number of tokens grows quicker than the number of words, as in the case of $\epsilon = 0$. The case of $\epsilon > 0$ is even more statistically positive since N/A scales as A^{ϵ}, which grows quicker than $\log A$ in the former case.

In the case of $\epsilon > 0$, one has the warranty that the percentage of tokens being the most frequent word will not decrease to zero while enlarging the text. Scaling as A^{ϵ}, however, is still not a lot. One is not able to improve drastically one's estimates of $p(w)$ simply enlarging the dataset of texts by a feasible factor. What is even worse, in the model, A still grows with respect to the text size. All frequencies of frequencies scale like $A \propto N^{1-\epsilon}$, the absolute number of *hapax legomena* being still about $A/2$ and growing as well.

5 Comparing various finite texts

The existence of grammar makes some functional words to obtain the lowest ranks. These words appear very regularly and their frequencies do not decrease versus ranks so quickly. One can model this modification of Zipf's law very robustly by introducing constant ρ as in (2). Having combined this effect with stepwise distribution of frequencies (28) for finite texts, one obtains the formula:

$$f(r) \approx \lfloor \left[\frac{A+\rho}{r+\rho} \right]^{1+\epsilon} \rfloor. \tag{33}$$

Formula (33) fits the whole range of finite-text data consistently better than simply (1) or even (2). Furthermore, it was researched and theoretically modeled by Orlov that there is some dependence between size of text N in a given language and the numerical values not only of A, but also of ρ and ϵ. Orlov's original article was [Orl82] but we have not managed to collect it and we learned partially about its contents from [Sam88]. Here, some possible reconstruction of the Orlov's reasoning will be presented.

Using (33), one can compute number of tokens N in the text as:

$$N = \int_0^A f(r)dr \approx \int_0^A \left[\frac{A+\rho}{r+\rho} \right]^{1+\epsilon} dr = \frac{\rho}{\epsilon} \left[\frac{A+\rho}{\rho} \right]^{1+\epsilon} \left[1 - \left[\frac{\rho}{A+\rho} \right]^{\epsilon} \right]. \tag{34}$$

If one considers an ensemble of texts of variable size N written in the same language, it is reasonable to assume that the same grammar is obeyed in the whole ensemble. The conservation of grammar implies the stability of probability estimators for the functional words which occupy constantly the same lowest ranks. Thus:

$$\frac{f(r)}{N} \approx \text{const} \quad \text{for} \quad r = 0, 1, 2, ..., K, \tag{35}$$

where K is some small natural number and N is assumed to be *any* length of text in the ensemble. If $N \to \infty$, however, $f(r)/N \to \text{const}$ for all r. Thus:

$$\epsilon \to \text{const} \quad \text{and} \quad \rho \to \text{const} \quad \text{for} \quad N \to \infty. \tag{36}$$

Postulates (35), (36) when applied to (33), (34) are equivalent to the following three postulates:

1. There is such $N = N_0$ that $\epsilon = 0$.
2. For all N, it is $f(0)/N = \text{const}$.
3. For all N, it is $f(1)/N = \text{const}$.

For $N = N_0$, let write down $A = A_0$, $\rho = \rho_0$. Combining postulates 2 and 3, one obtains $f(0)/f(1) = \text{const}$. Inserting (33) for any N and for $N = N_0$, and preserving terms linear in $1/\rho$ yields:

$$\rho = (1+\epsilon)\rho_0. \tag{37}$$

Parameter N_0 can be rewritten by means of A_0 and ρ_0 as:

$$N_0 \approx \int_0^{A_0} \left[\frac{A_0+\rho_0}{r+\rho_0} \right] dr = \rho_0 \left[\frac{A+\rho_0}{\rho_0} \right] \ln \left[\frac{A_0+\rho_0}{\rho_0} \right]. \tag{38}$$

Combining directly postulate 2 with (33) for any N and for $N = N_0$ gives:

$$\rho_0 \ln\left[\frac{\rho_0}{A_0+\rho_0}\right] = \frac{\rho}{\epsilon}\left[\left[\frac{\rho}{A+\rho}\right]^{\epsilon} - 1\right]. \tag{39}$$

It is convenient to define:

$$\lambda = 1 - \frac{\epsilon x}{1+\epsilon}, \tag{40}$$

$$x = \ln\left[\frac{A_0+\rho_0}{\rho_0}\right]. \tag{41}$$

Then:

$$1+\epsilon = \frac{x}{x-1+\lambda}. \tag{42}$$

Equations (39) and (37) yield:

$$\left[\frac{A+\rho}{\rho}\right]^{1+\epsilon} = \left[1 - \left[\frac{\epsilon}{1+\epsilon}\right]\ln\left[\frac{A_0+\rho_0}{\rho_0}\right]\right]^{-\frac{1+\epsilon}{\epsilon}} = [e(\lambda)]^x, \tag{43}$$

where function $e(\lambda)$ is defined as:

$$e(\lambda) = \lambda^{1/(\lambda-1)}. \tag{44}$$

Resuming, one obtains:

$$f(r) \approx [e(\lambda)]^x \left[\frac{\rho_0\left[\frac{x}{x-1+\lambda}\right]}{r+\rho_0\left[\frac{x}{x-1+\lambda}\right]}\right]^{\left[\frac{x}{x-1+\lambda}\right]}, \tag{45}$$

$$N \approx [e(\lambda)]^x \rho_0 x, \tag{46}$$

$$A \approx \rho_0\left[\frac{x}{x-1+\lambda}\right]\left[[e(\lambda)]^{(x-1+\lambda)} - 1\right], \tag{47}$$

where A was computed from property $f(A) = 1$.

In equations (45)-(47) three parameters appear: ρ_0, x and λ. The status of ρ_0 and x is different from λ. Parameters ρ_0 and x are the properties of a given language and they are constant in the whole ensemble of texts in that language. Parameter λ is a function of the size of text N and the two other parameters. Since $e(\lambda) > 1$ then our model can be only applied if $N > \rho_0 x$. For $N \to \rho_0 x$, it is $\lambda \to \infty$. For $N \to \infty$, it is $\lambda \to 0$. In order to compute λ as the function of N, ρ_0, and x, it is necessary to find the inverse of $e(\lambda) = \lambda^{1/(\lambda-1)}$. The inverse of $e(\lambda)$ is not a simply computable function of its argument but there is some good elementary approximation, which is presented in appendix A.

6 Summary

Some basic quantitative description of language texts is given by Zipf's law (1). The law initially observed in linguistics is a general law of statistics with lots of appearances in various domains as given on-line at [Li]. Zipf's distribution (1), which was generalized by Mandelbrot as power law (2), is the default distribution for ranking in zeroth order Markov models (e.g. in the texts of random letters and spaces as shown in [Li92]).

In spite of carrying substantial grammatical and semantical structure, the natural and artificial languages seem to exhibit still a lot of chaos typical for simple, low-order Markov models. Approximating the linguistic reality by these models and using various techniques of mathematical analysis, one can find some sensible predictions for the parameter dependence.

All this information is derived assuming extremely simple language models and it cannot be perceived as terribly informative for many problems in NLP. The low order Markovian properties of texts may be worth attention since they found the basic level of statistical properties of language performance and they cause some interesting limitations for the feasibility of endeavors in computational linguistics. On the other hand, some of the limitations may be not so severe as the learning algorithms will use better and better means for overcoming the sparseness of data.

A Approximating the inverse of $e(\lambda) = \lambda^{1/(\lambda-1)}$

The function defined by (44) is very closely related to the definition of base e of natural logarithm. Actually:

$$e(1) = e. \tag{48}$$

Function $e(\lambda)$ is an easily computable function of λ. Unfortunately the inverse is not true. Quantity λ is not a simply computable function of $e(\lambda)$. There is, however, easily invertible and good approximation $\bar{e}(\lambda)$:

$$\bar{e}(\lambda) = 1 + \frac{e-2}{\sqrt{\lambda}} + \frac{1}{\lambda}. \tag{49}$$

One can define the relative error of $\bar{e}(\lambda)$ as:

$$b(\lambda) = \frac{\bar{e}(\lambda) - e(\lambda)}{\bar{e}(\lambda)}. \tag{50}$$

Function $e(\lambda)$ has domain $\lambda \in \{0, \infty\}$. In this domain, the following substitution is convenient:

$$\lambda = \frac{1-u}{1+u}, \tag{51}$$

where $u \in \{-1, 1\}$. Let $\bar{b}(u) = b(\lambda)$. Then $\bar{b}(u) = 0$ for $u = -1, 0, 1$. Explicitly:

$$\bar{b}(u) = 1 - \frac{\left[\frac{1-u}{1+u}\right]^{-1/2u}}{\sqrt{\frac{1-u}{1+u}} + (e-2) + \sqrt{\frac{1+u}{1-u}}}, \tag{52}$$

so $\bar{b}(u) = \bar{b}(-u)$. Furthermore, looking at the plot of $\bar{b}(u)$ one sees that $0 \leq \bar{b}(u) < 0.04$. Resuming, $\bar{e}(\lambda)$ is a very good approximation of $\lambda^{1/(\lambda-1)}$.

Approximation $\bar{e}(\lambda)$ is a very useful one, too. In order to find λ for given $\bar{e}(\lambda)$, one just should realize that definition (49) is a quadratic equation for $1/\sqrt{\lambda}$ and it can be solved:

$$\lambda = \frac{4}{\left[2 - e + \sqrt{e^2 - 4e + 4\bar{e}(\lambda)}\right]^2} \tag{53}$$

In formula (53), only this one of two possible solutions was chosen which reproduces $\lambda = 1$ for $\bar{e}(\lambda) = e$.

References

[GA82] H. Guiter and M. V. Arapov, editors. *Studies on Zipf's Law.* Wissenschaftlicher Verlag Trier, 1982.

[GLSW96] R. Günther, L. Levitin, B. Schapiro, and P. Wagner. Zipf's law and the effect of ranking on probability distributions. *International Journal of Theoretical Physics*, 15:395, 1996.

[Li] Wentian Li. References on Zipf's law. URL: http://linkage.rockefeller.edu/wli/zipf/.

[Li92] Wentian Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38:1842, 1992.

[Man54] Benoit Mandelbrot. Structure formelle des textes et communication. *Word*, 10:1, 1954.

[Man83] Benoit B. Mandelbrot. *The Fractal Geometry of Nature.* W. H. Freeman, 1983.

[Mil65] George Miller. Introduction. In George Kingsley Zipf, editor, *Human Behavior and the Principle of Least Effort.* The MIT Press, 1965.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, 1999.

[Orl82] Ju. K. Orlov. Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? In Ju. K. Orlov, M. G. Boroda, and I. S. Nadarejšvili, editors, *Sprache, Text, Kunst.* Wissenschaftlicher Verlag Trier, 1982.

[Sam72] Jadwiga Sambor. *Słowa i liczby. Zagadnienia językoznawstwa statystycznego.* Ossolineum, 1972.

[Sam88] Jadwiga Sambor. Lingwistyka kwantytatywna — stan badań i perspektywy rozwoju. *Biuletyn Polskiego Towarzystwa Językoznawczego*, XLI:47, 1988.

[Zip35] George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Houghton Mifflin, 1935.

[Zip49] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, 1949.