A Simple LNRE Model for Random Character Sequences

Stefan Evert
evert@ims.uni-stuttgart.de

IMS - University of Stuttgart - Azenbergstr. 12 - D-70174 Stuttgart - Germany

Abstract

This paper describes a population model for word frequency distributions based on the Zipf-Mandelbrot law, corresponding to the word frequency distribution induced by a random character sequence. The model, which has convenient analytical and numerical properties, is shown to be adequate for the description of language data extracted by automatic means from large text corpora. It can thus be used to study the problems faced by the statistical analysis of such data in the field of natural-language processing.

Keywords: lexical statistics, LNRE models, Zipf-Mandelbrot law, random text, cooccurrence statistics

1 Introduction to lexical statistics and LNRE models

Most work in the area of lexical statistics is based on random sampling with replacement.[1] This model assumes a population of types $w_1, \cdots, w_S$ with occurrence probabilities $\pi_1, \cdots, \pi_S$. S is called the population size and may be infinite ($S = \infty$) in the case of a countably infinite population. The probabilities π_i are the parameters of this model and must satisfiy

$$\pi_1 + \cdots + \pi_S = 1. \tag{1}$$

It is convenient to assume that they are arranged in descending order, i.e. $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_S$. The random selection of a token from this population is described by a random variable $X : \Omega \to \{1, \ldots, S\}$.[2] A value of $X = k$ implies that the selected token is of type w_k, and the distribution of X is given by $P\big(X = k\big) = \pi_k$ for $k \in \{1, \ldots, S\}$. A random sample of size N corresponds to the N-fold independent repetition of this experiment, i.e. to independent random variables $\mathbf{X} = (X_1, \ldots, X_N)$ with distributions identical to that of X.

In lexical statistics, a text sample of N tokens (which may be anything ranging from orthographic words, over words belonging to a specific morphological category, to word pairs representing cooccurrences) is interpreted as such a random sample $\mathbf{X}$. In this view, two major goals of the statistical analysis are: (i) Draw inferences about the population parameters from the observed data, which are then interpreted in light of the research question. An example is the estimation of the population size S, which may correspond to the number of different word types that a particular word-formation process can generate (e.g. Baayen, 2001, Sec. 6.2) or to the size of an author's vocabulary (e.g. McNeil, 1973). (ii) Given the estimated population parameters (or, more generally, assumptions about these parameters), predict the behaviour of various observable quantities. Such quantities correspond to random variables in the random

[1] For all the concepts and results introduced in this section, see (Baayen, 2001). The notation has been adopted from the same source with minor changes.

[2] When $S = \infty$, $\{1, \ldots, S\}$ stands for the set $\mathbb{N}$ of all natural numbers.

sample model, for which expectations and variances can be computed. A typical example is the prediction of vocabulary growth curves, which measure the increase in the number of observed types when the sample size N is increased (see Baayen, 2001).

Since the random variables $(X_1, \ldots, X_N)$ are jointly independent, the sequential ordering of the tokens in the sample $\mathbf{X}$ provides no information about the population parameters.[3] It is therefore sufficient to consider the type frequencies f_i for $i \in \{1, \ldots, S\}$ (in the mathematical terminology, $\mathbf{f} = (f_1, \ldots, f_S)$ is a sufficient statistic for the random sample $\mathbf{X}$). The frequency f_i is the number of tokens in the sample $\mathbf{X}$ belonging to type w_i, or formally

$$f_i := \sum_{k=1}^{N} I_{[X_k=i]}, \tag{2}$$

where

$$I_{[X_k=i]} := \begin{cases} 1 & X_k = i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

is the general notation for indicator variables (= random variables with range $\{0, 1\}$). Each f_i is a binomially distributed random variable with success probability π_i, i.e.

$$P\big(f_i = k\big) = \binom{N}{k} (\pi_i)^k (1 - \pi_i)^{N-k} \tag{4}$$

for $k \in \{1, \ldots, N\}$. It has to be kept in mind, though, that the f_i are not mutually independent. The mathematical analysis is considerably simplified when conditioning on a fixed sample size N is avoided, i.e. one assumes that the sample size is itself a Poisson-distributed random variable with mean N. The type frequencies then become *independent* Poisson-distributed random variables with

$$P\big(f_i = k\big) = e^{-N\pi_i} \frac{(N\pi_i)^k}{k!}. \tag{5}$$

This approach (henceforth called independent Poisson sampling) is quite natural when studying the number of different species in biological samples, where the total number of specimens in the sample is obviously subject to random variation and cannot be fixed in advance.[4] Independent Poisson sampling can also be applied in lexical statistics, especially for large N. For instance, the unconstrained sample size with mean $N = 1\,000\,000$ has a standard deviation of $\sigma = \sqrt{N} = 1\,000$; therefore, the observed sample size will almost certainly deviate from N by less than 1%.[5] In the following, I will always assume independent Poisson sampling, as does Baayen (2001). The expectation of f_i is then $E\big[f_i\big] = N\pi_i$, and its variance is $VAR\big[f_i\big] = N\pi_i$.

Since it is usually not known which one of the observed types is the i-th population type w_i, the set of observed frequencies cannot be matched directly against the random variables f_i. It is common practice to arrange the observed frequency values in descending order $f_1^* \geq f_2^* \geq \ldots$, which is called a Zipf ranking. Although π_1 is the highest type probability, and $E\big[f_1\big]$ the

[3] This ordering can be used to test the adequacy of the random sample model, though, e.g. with the dispersion test described in (Baayen, 2001, Sec. 5.1).

[4] A good deal of the work on word frequency distributions originates in this area, e.g. Good (1953), Holgate (1969), Engen (1974).

[5] The same argument shows that great care has to be taken when Equation (5) is applied to small samples. For $N = 1\,000$, the standard deviation is $\sigma = \sqrt{1\,000} \approx 31.62$ and deviations as far as 10% from the mean N have to be expected.

highest expectation, w_1 need not be the most frequent observed type corresponding to f_1^*, and the same holds for all f_i^*. A better approach is to look at other summary statistics that can be directly observed without an exact knowledge of the population types.

In order to do so, collect all types w_i with the same frequency $f_i = m$ into the frequency class m. The class size V_m, i.e. the number of different types in the frequency class m, can be easily determined from the observed sample. In the random sample model, it is given by the random variable

$$V_m := \sum_{i=1}^{S} I_{[f_i=m]}. \tag{6}$$

The sequence of all class sizes $(V_1, V_2, \dots)$ is called the frequency spectrum. Note that all but finitely many of the V_m equal zero (in particular, the largest non-empty frequency class is $V_{f_1^*}$). Using the same definition, V_0 is the number of unobserved types, which cannot be determined from the sample. The vocabulary size V is the total number of types observed in the sample:

$$V := \sum_{i=1}^{S} I_{[f_i>0]}. \tag{7}$$

The frequency spectrum is related to V and N through the identities $V = \sum_{m=1}^{\infty} V_m$ and $N = \sum_{m=1}^{\infty} m V_m$. The expectations of V and V_m can easily be computed from (5):

$$E\big[V_m\big] = \sum_{i=1}^{S} e^{-N\pi_i} \frac{(N\pi_i)^m}{m!} \quad \text{and} \quad E\big[V\big] = \sum_{i=1}^{S} \left(1 - e^{-N\pi_i}\right), \tag{8}$$

but it is more difficult to obtain variances and the exact distributions (see Baayen, 2001).

As noted before, it is impossible to estimate the large number of probability parameters directly from a sample. It is therefore necessary to formulate a population model with a small number of parameters: once these have been estimated, the hypothesised distribution of the probability parameters π_i can be computed. Following Baayen (2001), I use the term LNRE model for such a population model.[6] While it is in principle possible to formulate an LNRE model directly for the type probability parameters (e.g. Holgate, 1969), it is usually more convenient to use the structural type distribution, which is a step function given by

$$G(\rho) := \big|\{i \in \{1, \dots, S\} \mid \pi_i \geq \rho\}\big|. \tag{9}$$

$E\big[V_m\big]$ and $E\big[V\big]$ can then be expressed in terms of Stieltjes integrals

$$E\big[V_m\big] = \int_0^{\infty} \frac{(N\pi)^m}{m!} e^{-N\pi}\, dG(\pi), \quad E\big[V\big] = \int_0^{\infty} (1 - e^{-N\pi})\, dG(\pi) \tag{10}$$

(Baayen, 2001, 47f). Most LNRE models approximate $G(\rho)$ by a continuous function with type density function $g(\pi)$, i.e.

$$G(\rho) = \int_{\rho}^{\infty} g(\pi)\, d\pi. \tag{11}$$

[6] LNRE stands for Large Number of Rare Events, a term introduced by Khmaladze (1987). It refers to the very large number of types with low occurrence probabilities that are characteristic of word frequency distributions and the associated population models.

Note the use of $+\infty$ as an upper integration limit although all type probabilities must fall into the range $0 \leq \pi \leq 1$. This device allows for more elegant mathematical formulations, but care has to be taken that $G(1) \ll 1$ (otherwise the LNRE model would predict the existence of types with $\pi > 1$). For an LNRE model based on a type density function $g(\pi)$, the expectations of V_m and V become

$$E\big[V_m\big] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi, \quad E\big[V\big] = \int_0^\infty (1 - e^{-N\pi}) g(\pi)\, d\pi. \tag{12}$$

Equation (1) leads to the normalisation condition

$$\int_0^\infty \pi \cdot g(\pi)\, d\pi = 1, \tag{13}$$

and the population size is given by $S = \int_0^\infty g(\pi)\, d\pi$.

2 Random character sequences and the Zipf-Mandelbrot law

Zipf's law (Zipf, 1949), which states that the frequency of the r-th most frequent type is proportional to $1/r$, was originally formulated for the Zipf ranking of observed frequencies ($f_r^* \approx Cr^{-1}$) and (more or less equivalently) for the observed frequency spectrum ($V_m \approx C/m(m+1)$). In its first form, Zipf's law describes a fascinating property of the higher-frequency words in a language, for which explanations related to Zipf's principle of least effort have been put forward (e.g. Mandelbrot, 1962; Powers, 1998). In its second form, it is a statement about the enormous abundance of lowest-frequency types, which has many consequences for the statistical analysis and for applications in natural-language processing.

It has long been known that the word frequency distributions obtained from random text are strikingly similar to Zipf's law (Li, 1992; Miller, 1957). Formally, random text is understood as a character sequence generated by a Markov process, with word boundaries indicated by a special "space" character. Rouault (1978) shows that, under very general conditions, this segmented character sequence is equivalent to a random sample of words (with replacement, corresponding to the model introduced in Section 1) and that the population probabilities of low-frequency types asymptotically satisfy the Zipf-Mandelbrot law

$$\pi_i = \frac{C}{(i+b)^a} \tag{14}$$

with parameters $a > 1$ and $b > 0$ (Baayen, 2001, 101ff). In Sections 3 and 4, I will formulate LNRE models for random character sequences based on the Zipf-Mandelbrot law. Although Baayen remarks that "for Zipf's harmonic spectrum law and related models, no complete expression for the structural type distribution is available" (Baayen, 2001, 94), this need not discourage us: (14) refers to the population parameters rather than to the observed Zipf ranking. The Zipf-Mandelbrot law for random text is a population model, while the original formulation of Zipf's law and its variants (Baayen, 2001, 94f) have a purely descriptive nature.

These considerations open up an entirely new perspective on Zipf's law: If an LNRE model based on (14) can be shown to agree with the observed data, we must conclude that – as far as

statistical analysis is concerned – such language data is not substantially different from random text. As a consequence, the statistical analysis faces all the problems of making sense from random noise, and these problems can be predicted with the LNRE models of Sections 3 and 4.

One of the characteristics of random text is an infinite population size, since there can be words of arbitrary length, leading to an extremely skewed LNRE distribution. It has often been noted that this does not accord well with real-world data, especially when there are narrow restrictions and the data have been cleaned up manually. Examples are studies of (morphological) productivity (e.g. Baayen and Renouf, 1996) or the word frequency distributions of small literary texts (see Baayen, 2001). However, the situation is different when one considers "raw" data obtained from a large corpus of hundreds of millions of words, which is the input that statistical methods in natural-language processing typically have to deal with. The similarity to random text becomes even more striking for combinations of two or more words (cf. Baayen, 2001, 221). Most techniques for the extraction of collocations from text corpora apply statistical independence tests to such base material (e.g. Evert and Krenn, 2001), and are thus also affected by the consequences of the Zipf-Mandelbrot law. Ha *et al.* (2002) demonstrate such an effect for Mandarin Chinese ideographs: while the number of different graphs is comparatively small and does not exhibit an LNRE distribution, the situation changes when sequences of two or more such graphs are examined. The longer the sequences, the more closely their frequency distribution agrees with the Zipf-Mandelbrot law.

3 The Zipf-Mandelbrot (ZM) LNRE model

In order to derive a useful LNRE model from the Zipf-Mandelbrot law, it is necessary to reformulate (14) in terms of a type density function $g(\pi)$. The structural type distribution corresponding to the Zipf-Mandelbrot law is a step function with $G(\pi_i) = i$ (since there are exactly i types with $\pi \geq \pi_i$, namely $w_1, \ldots, w_i$). Solving (14) for i, we obtain

$$G(\pi) = \frac{C^{1/a}}{\pi^{1/a}} - b \tag{15}$$

for $\pi = \pi_i$, and $G(\pi)$ is constant between these steps. Differentiation of (15) suggests a type density of the form

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & 0 \leq \pi \leq B \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

with two free parameters $0 < \alpha < 1$ and $B > 0$.[7] The normalising constant C can be determined from (13):

$$1 = \int_0^B \pi g(\pi)\, d\pi = \int_0^B C\pi^{-\alpha}\, d\pi = C \cdot \left[\frac{\pi^{1-\alpha}}{1-\alpha}\right]_0^B = C \cdot \frac{B^{1-\alpha}}{1-\alpha} \tag{17}$$

which evaluates to

$$C = \frac{1-\alpha}{B^{1-\alpha}}. \tag{18}$$

[7] The constraints on the parameter α follow from $0 < 1/a < 1$. C is a normalising constant and will be determined from (13). The upper cutoff point B is necessary since the model would predict types with probability $\pi > 1$ otherwise. B should roughly correspond to the probability π_1 of the most frequent type.

The ZM model describes an infinite population, since $S = \int_0^B g(\pi)\,d\pi = \infty$, and its structural type distribution

$$\begin{aligned} G(\rho) &= \int_\rho^B g(\pi)\,d\pi = C \cdot \int_\rho^B \pi^{-\alpha-1}\,d\pi = C \cdot \left[\frac{\pi^{-\alpha}}{-\alpha}\right]_\rho^B \\ &= \frac{C \cdot \rho^{-\alpha}}{\alpha} - \frac{C \cdot B^{-\alpha}}{\alpha} = \frac{C/\alpha}{\rho^\alpha} - \frac{1-\alpha}{B \cdot \alpha} \end{aligned}$$

is identical to (15) with $a = \alpha^{-1}$ and $b = (1-\alpha)B^{-1}\alpha^{-1}$ for any values of ρ where $G(\rho) \in \mathbb{N}$. Thus, (16) can indeed be understood as a continuous extension of the Zipf-Mandelbrot law.

$$\begin{aligned} E\big[V_m\big] &= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\,d\pi = \frac{C}{m!}\int_0^B (N\pi)^m e^{-N\pi}\pi^{-\alpha-1}\,d\pi \\ &= \frac{C}{m!}\int_0^{NB} t^m e^{-t}\left(\frac{t}{N}\right)^{-\alpha-1}\frac{1}{N}\,dt = \frac{C}{m!}N^\alpha\int_0^{NB} t^{m-\alpha-1}e^{-t}\,dt \\ &\approx \frac{C}{m!}N^\alpha\int_0^\infty t^{m-\alpha-1}e^{-t}\,dt \end{aligned}$$

In the second line, the substitution $t := N\pi$ has been made. The approximation in the last line is justified for $NB \gg m$ (which should always be the case for the large samples that are of interest here) where the integral $\int_{NB}^\infty t^{m-\alpha-1}e^{-t}\,dt$ is vanishingly small. Thus, $E\big[V_m\big]$ is reduced to the gamma integral $\int_0^\infty t^{m-\alpha-1}e^{-t}\,dt = \Gamma(m-\alpha)$ (Weisstein, 1999, *s.v. Gamma Function*) and we obtain the concise expression

$$E\big[V_m\big] = \frac{C}{m!} \cdot N^\alpha \cdot \Gamma(m-\alpha). \tag{19}$$

The computation of $E\big[V\big]$ involves an improper integral solved by partial integration:

$$\begin{aligned} E\big[V\big] &= \int_0^\infty (1-e^{-N\pi})g(\pi)\,d\pi \approx CN^\alpha\int_0^\infty (1-e^{-t})t^{-\alpha-1}\,dt \\ &= CN^\alpha \cdot \lim_{A\downarrow 0}\left(\int_A^\infty t^{-\alpha-1}\,dt - \int_A^\infty e^{-t}t^{-\alpha-1}\,dt\right) \\ &= CN^\alpha \cdot \lim_{A\downarrow 0}\left(\left[\frac{t^{-\alpha}}{-\alpha}\right]_A^\infty - \left[e^{-t}\frac{t^{-\alpha}}{-\alpha}\right]_A^\infty - \int_A^\infty e^{-t}\frac{t^{-\alpha}}{-\alpha}\,dt\right) \\ &= CN^\alpha \cdot \lim_{A\downarrow 0}\Bigg(\underbrace{\big(1-e^{-A}\big)\cdot\frac{A^{-\alpha}}{\alpha}}_{=O(A^{1-\alpha})\to 0} + \underbrace{\frac{\Gamma(1-\alpha,A)}{\alpha}}_{\to\Gamma(1-\alpha)/\alpha}\Bigg) \end{aligned}$$

where $\int_A^\infty e^{-t}t^{-\alpha}\,dt = \Gamma(1-\alpha,A)$ is the upper incomplete gamma function (Weisstein, 1999, *s.v. Incomplete Gamma Function*). This leads to

$$E\big[V\big] = C \cdot N^\alpha \cdot \frac{\Gamma(1-\alpha)}{\alpha}. \tag{20}$$

Consequences of (19) and (20) are the recurrence relation

$$\frac{E\big[V_{m+1}\big]}{E\big[V_m\big]} = \frac{\Gamma(m+1-\alpha)}{(m+1)!} \cdot \frac{m!}{\Gamma(m-\alpha)} = \frac{m-\alpha}{m+1}, \tag{21}$$

a relative frequency spectrum

$$\frac{E\big[V_m\big]}{E\big[V\big]} = \frac{\alpha \cdot \Gamma(m-\alpha)}{\Gamma(m+1) \cdot \Gamma(1-\alpha)} \tag{22}$$

which is independent of the sample size N (cf. Baayen, 2001, 118), and a power law

$$E\big[V(N)\big] = C' \cdot N^{\alpha} \quad \text{with} \quad 0 < \alpha < 1 \tag{23}$$

for the vocabulary growth curve. Equation (23) is known as Herdan's law (Herdan, 1964) in quantitative linguistics and as Heaps' law (Heaps, 1978) in information retrieval.

The appeal of the ZM model lies in its mathematical elegance and numerical efficiency. Computation of the expected frequency spectrum and similar statistics is fast and accurate, using implementations of the complete and incomplete gamma function that are provided by many scientific libraries. Moreover, due to the simple form of $g(\pi)$ many other important integrals such as

$$E\big[V_{m,\rho}\big] = \int_0^{\rho} \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi \tag{24}$$

for $0 < \rho < B$ have closed-form solutions and can be studied analytically.

4 The finite Zipf-Mandelbrot (fZM) LNRE model

Although the ZM model is theoretically well-founded as a model for random character sequences, its assumption of an infinite vocabulary is unrealistic for natural-language data. In order to achieve a better approximation of such frequency distributions, the finite ZM model introduces an additional lower cutoff point $A > 0$ for the type density:

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & A \le \pi \le B \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

which implies that there are no types with probability $\pi < A$ in the population. The normalising constant C is determined from (13) as

$$C = \frac{1-\alpha}{B^{1-\alpha} - A^{1-\alpha}}, \tag{26}$$

and the population size is

$$S = \frac{C}{\alpha} \cdot (A^{-\alpha} - B^{-\alpha}) = \frac{1-\alpha}{\alpha} \cdot \frac{A^{-\alpha} - B^{-\alpha}}{A^{1-\alpha} - B^{1-\alpha}}. \tag{27}$$

Again, the structural type density $G(\rho)$ is identical to (15), with $G(\rho) = S$ for $\rho \leq A$. The expectations of V_m and V are calculated to be

$$E\big[V_m\big] = \frac{C}{m!} \cdot N^\alpha \cdot \Gamma(m - \alpha, NA), \tag{28}$$

$$E\big[V\big] = C \cdot N^\alpha \cdot \frac{\Gamma(1-\alpha, NA)}{\alpha} + \frac{C}{\alpha \cdot A^\alpha} \left(1 - e^{-NA}\right). \tag{29}$$

There are no simple expressions for the recurrence relation (21) and the relative frequency spectrum (22). Although much of the mathematical elegance of the ZM model has been lost, the fZM model is still numerically efficient, and many integrals like (24) have closed-form expressions involving incomplete gamma functions.

5 Some other (related) LNRE models

Rouault (1978) has studied the properties of random character processes and shown that their observed relative frequency spectrum V_m/V converges to the expression (22) predicted by the ZM model for $N \to \infty$. This result provides theoretical support for its use as a model of large samples of random (or nearly random) text.

In the literature on lexical statistics, models for word frequency distributions are often based on a power law similar to (16). Sometimes, a decay factor $e^{-\lambda\pi}$ is used instead of the arbitrary cutoff point B. Such a model is introduced by Good (1953, 248) as a "Pearson Type III" distribution for $\alpha < 0$, and generalised to the range $0 < \alpha < 1$ by Engen (1974). Multiplication with a second decay factor $e^{-\mu/\pi}$ instead of the lower cutoff point A of the fZM model leads to Equation (50) of Good (1953, 249). Good refers to it as a mixture "between Pearson's Types III and V" and remarks that it is "analytically unwieldy". Sichel (1971, 1975) works out this model under the name Generalized Inverse Gauß-Poisson (GIGP), using the type density

$$g(\pi) = \frac{(2/bc)^{\gamma+1}}{2K_{\gamma+1}(b)} \pi^{\gamma-1} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}} \tag{30}$$

with parameters b, c, and γ (Baayen, 2001, Sec. 3.2.2). Carroll (1967) and Holgate (1969) assume a log-normal distribution for the population frequencies of words or species in a biological sample, citing Preston (1948) for a theoretical motivation. The resulting type density is

$$g(\pi) = \frac{1}{\sigma\sqrt{2\boldsymbol{\pi}}} \frac{1}{\pi^2} e^{-\frac{1}{2\sigma^2}(\log\pi - \mu)^2} \tag{31}$$

with parameters μ and σ (Baayen, 2001, Sec. 3.2.1).[8] Baayen (2001, Sec. 3.2.3) presents several other models based on variants of Zipf's law for the expected frequency spectrum at a certain Zipf sample size Z (see also Good, 1953, 249). Although the expectations of V_m and V can be computed for arbitrary sample sizes N using extrapolation techniques, none of these models can be reformulated as a population model (Baayen, 2001, 94). Implementations of the GIGP, log-normal, and several of the Zipf models are available in the `lexstats` package distributed with (Baayen, 2001). Of the various Zipf models, the Yule-Simon model (Simon, 1960) is found to be useful and numerically manageable.

[8] The constant $\boldsymbol{\pi}$ is printed in bold font to distinguish it from the type probability π.

6 Empirical data

In order to see how well the ZM and fZM models describe real-world data, they have been applied to nouns and adjective-noun cooccurrences extracted from the 100-million word British National Corpus (BNC) and a 225-million word corpus of German newspaper text from the 1990's (HGC). The following four data sets were used. **BNC-N**: 19 million instances of nouns extracted from the BNC corpus, and filtered with regular expressions to weed out non-words ($N = 19 \times 10^6, V = 217\,527$). **HGC-N**: 48 million instances of nouns extracted from the HGC corpus, and checked with a morphological analyser (Lezius *et al.*, 2000) ($N = 48 \times 10^6, V = 1\,556\,203$). **BNC-AN**: 4 million instances of adjacent adjective-noun pairs from the BNC corpus. Both the adjective and the noun were checked with regular expressions ($N = 4 \times 10^6, V = 1\,391\,498$). **HGC-AN**: 12 million instances of adjectives modifying nouns within a noun phrase, extracted using part-of-speech patterns. This simple extraction method has been found to reach excellent precision (Evert and Kermes, 2003). Both the adjective and the noun were validated with a morphological analyser ($N = 12 \times 10^6, V = 3\,621\,708$).

The Herdan law and the size-invariant relative frequency spectrum, which are characteristic properties of the ZM model, have repeatedly been critcised as unrealistic (e.g. Baayen, 2001, 118). Figure 1 shows the development of the relative frequency spectrum up to $m = 5$ for the HGC-AN data set (left panel). After approximately 2 million tokens, the relative spectrum has converged and is nearly constant afterwards. Likewise, the relative error of the Herdan law $E\big[V(N)\big] = C \cdot N^\alpha$ with $\alpha = 0.87$ (determined by linear regression) remains below 1% after the first 4 million tokens (right panel). Together with similar results for the other three data sets, this is a strong indication that the ZM and fZM models may indeed be well suited for the type of frequency data represented by these data sets.

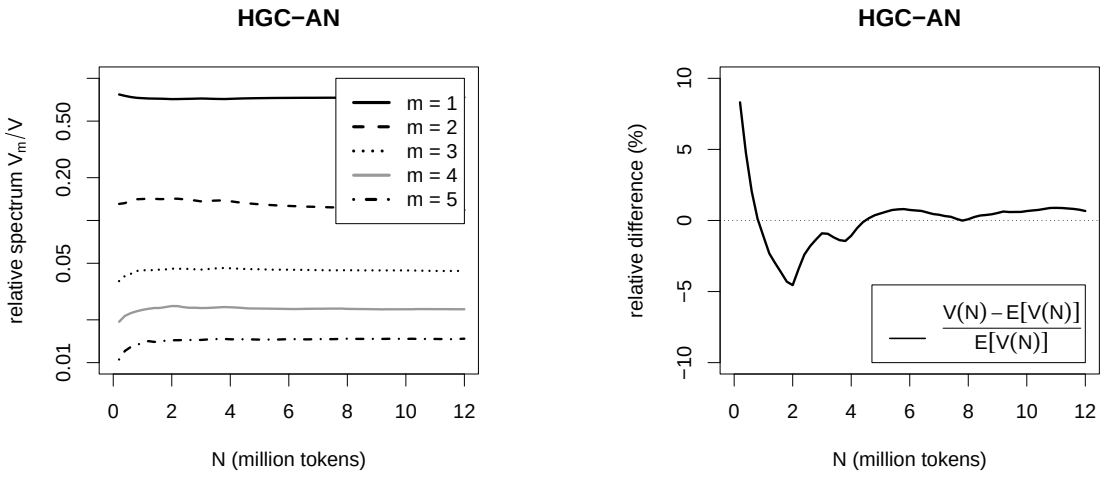

Figure 1: *Development of relative frequency spectrum and relative error of Herdan law (Heaps' law) with* $\alpha = 0.87$ *for the* HGC-AN *data set.*

7 Evaluation of the Zipf-Mandelbrot models

Both the ZM and the fZM model were implemented using the freely available statistical computation software R,[9] and fitted to the four data sets described in Section 6. For the infinite ZM model, the parameter α can be estimated directly from (22) for $m = 1$:

$$\alpha = \frac{E\big[V_1\big]}{E\big[V\big]} \approx \frac{V_1}{V} \tag{32}$$

(see also Rouault, 1978, 172). However, Equation (32) turned out to give unsatisfactory results, so the parameters for both models were estimated through non-linear minimisation of a goodness-of-fit chi-squared statistic for the first 15 spectrum elements, with the additional constraint $E\big[V\big] = V$. Goodness-of-fit was measured with a multi-variate chi-squared test, following Baayen (2001, Sec. 3.3) and using the `lexstats` implementation. The results are shown in Table 1.[10]

	ZM model		fZM model		
data set	α	χ^2_{14}	α	S	χ^2_{13}
BNC-N	0.4686416	80.75	0.4728356	4 021 728	22.20
HGC-N	0.6181580	27015.67	0.6663519	16 325 666	591.72
BNC-AN	0.7145849	313472.66	0.9168508	9 048 002	9364.46
HGC-AN	0.7441247	441448.77	0.9134667	37 983 975	1855.59

Table 1: *Estimated Zipf parameter α, population size S, and goodness-of-fit statistic χ^2 for the ZM and fZM models applied to the four data sets of Section 6.*

The fZM model gives considerably better approximations of the observed frequency spectrum than the ZM model, especially for the adjective-noun data sets where the distribution of population probabilities is much more skewed (indicated by a larger value of α). It is worth noting that the fZM model is entirely consistent with the BNC-N data set: $\chi^2_{13} = 22.20$ corresponds to a p-value of $p \approx 0.0524$ and the model is thus accepted at the 5% significance level.

A graphic representation of the accordance between the expected and observed frequency spectrum for the HGC-AN data set is shown in Figure 2. Surprisingly, the estimated lower cutoff points ($A = 9.267 \times 10^{-9}$ for BNC-AN and $A = 1.576 \times 10^{-9}$ for HGC-AN) are already quite close to the observed relative frequency of the hapax legomena ($p = 1/N$). According to the predicitions of the fZM model, increasing the sample 100-fold ($N \approx 10^9$) would already leave the LNRE zone, with all expected frequencies greater than 1 (cf. Baayen, 2001, Sec. 2.4).

A possible explanation for this counter-intuitive result is provided by term clustering effects, which violate the randomness assumption and cause the number V_1 of hapax legomena to be less than predicted by a random sample model. Such clustering effects can be detected with a dispersion test as described by (Baayen, 2001, Sec. 5.1). For the HGC-AN data set, a highly

[9] `http://www.r-project.org/`

[10] Note that the χ^2 statistic for the ZM model has $df = 14$ because 2 parameters were estimated from the observed spectrum. Likewise, the statistic for the fZM model with 3 estimated parameters has $df = 13$.

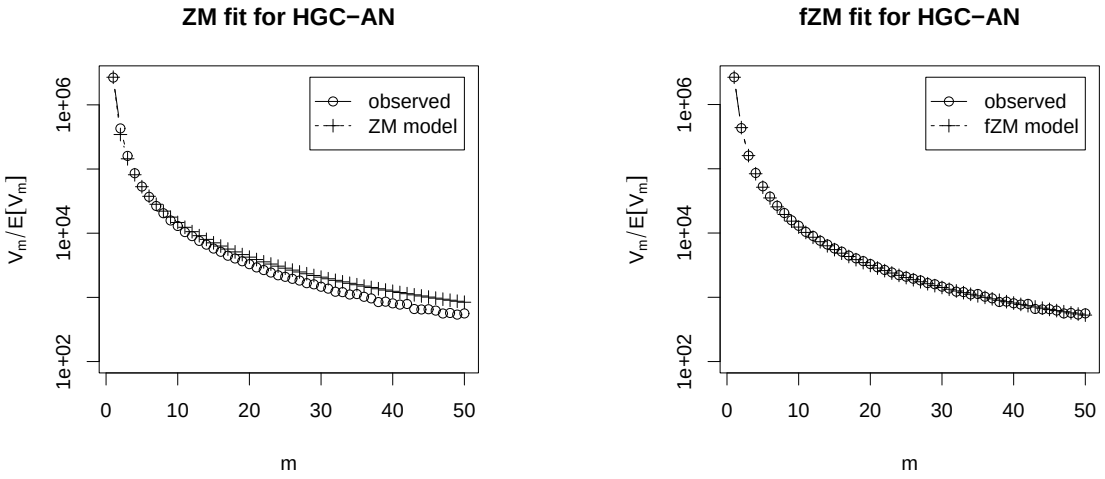

Figure 2: *Expected frequency spectrum of ZM (left panel) and fZM (right panel) models compared to observed spectrum for the* HGC-AN *data set (logarithmic scale, up to* $m = 50$*).*

significant deviation from the randomness assumption ($p \approx 0$) was found with this test.[11]

For comparison, I also applied the GIGP, log-normal, and Yule-Simon models (see Section 5) to the four data sets, using the implementations included in the `lexstats` package with automatic parameter optimisation. Goodness-of-fit results for the HGC-AN data set range from $\chi^2_{14} = 259800.05$ (log-normal) to $\chi^2_{13} = 63531.61$ (Yule-Simon); for the numerically simpler GIGP model with $\gamma = -0.5$, no reasonable fit could be achieved. Results for the BNC-N data set range from $\chi^2_{13} = 36562.75$ (log-normal) to $\chi^2_{13} = 1267.81$ (GIGP).

To conclude, the ZM model with its elegant formulation and convenient analytical properties achieves a goodness-of-fit comparable to that of the other LNRE models. The less elegant fZM model consistently outperforms its competitors and has the additional benefit of a fast and robust numerical implementation.

References

Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Baayen, R. Harald and Renouf, Antoinette (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, **72**(1), 69–96.

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

Engen, Steinar (1974). On species frequency models. *Biometrika*, **61**(2), 263–270.

[11] For the application of the dispersion test, the $N = 12 \times 10^6$ tokens were divided into 12 000 equally-sized chunks of 1 000 tokens each. The probability that a dis legomenon (a type with $f = 2$) is underdispersed, i.e. both its occurrences are in the same chunk, is found to be $p = 8.325 \times 10^{-5}$. Therefore, about 36 of the $V_2 = 430\,277$ dis legomena are expected to be underdispersed if the randomness assumption is correct. However, in the HGC-AN data set, there were 35 677 underdispersed dis legomena, which is a highly signifi cant deviation from the expected value (according to a binomial test).

Evert, Stefan and Kermes, Hannah (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, pages 83–86.

Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**(3/4), 237–264.

Ha, Le Quan, Sicilia-Garcia, E. I., Ming, Ji, and Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of COLING 2002*, Taipei, Taiwan.

Heaps, H. S. (1978). *Information Retrieval – Computational and Theoretical Aspects*. Academic Press.

Herdan, G. (1964). *Quantitative Linguistics*. Buttersworths, London.

Holgate, P. (1969). Species frequency distributions. *Biometrika*, **56**(3), 651–660.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Lezius, Wolfgang, Dipper, Stefanie, and Fitschen, Arne (2000). IMSLex – representing morphological and syntactical information in a relational database. In U. Heid, S. Evert, E. Lehmann, and C. Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 133–139, Stuttgart, Germany.

Li, Wentian (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, **38**(6), 1842–1845.

Mandelbrot, Benoit (1962). On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI.

McNeil, Donald R. (1973). Estimating an author's vocabulary. *Journal of the American Statistical Association*, **68**, 92–96.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, **52**, 311–314.

Powers, David M. W. (1998). Applications and explanations of Zipf's law. In D. M. W. Powers, editor, *Proceedings of New Methods in Language Processing and Computational Natural Language Learning*, pages 151–160. ACL.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Rouault, Alain (1978). Lois de Zipf et sources markoviennes. *Annales de l'Institut H. Poincaré (B)*, **14**, 169–188.

Sichel, H. S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In N. F. Laubscher, editor, *Proceedings of the Third Symposium on Mathematical Statistics*, pages 51–97, Pretoria, South Africa. C.S.I.R.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.

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

Weisstein, Eric W. (1999). *Eric Weisstein's World of Mathematics*. Wolfram Inc. On-line resource at `http://mathworld.wolfram.com/`.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.