Word Frequency Distributions

Revas J. Chitashvili, Tbilisi
R. Harald Baayen, Nijmegen

1. Introduction

Word frequency distributions have been studied from a variety of perspectives. In literary studies, word frequency distributions have attracted the attention of scholars interested in authorship attribution and vocabulary richness (Orlov 1983b, Muller 1977, 1979, Menard 1983, Thisted and Efron 1987, Herdan 1960, 1964). Psychologists have long been interested in word frequencies since word frequency is one of the most robust and important predictors of response time in a variety of experimental tasks addressing on-line word production and word recognition (Carroll 1969, 1970, Scarborough et al. 1977, Whaley 1978). Recently, word frequency distributions have also been exploited for the study of morphological productivity, the extent to which various word formation processes are alive in the language and may be expected to give rise to new (morphologically complex) formations (Baayen 1992, 1993a). This paper focusses on the probabilistic properties of word frequency distributions and on the statistical techniques developed for their analysis. Some attempt will be made, however, to understand the typical statistical properties of word frequency distributions of running texts in terms of the morphological structure of the constituent words and the productivity of the underlying word formation processes.

Our discussion is structured as follows. Section 2 introduces various ways of describing empirical word frequency distributions as well as a number of 'laws' that have been advanced in the literature as governing these distributions. Section 3 develops a stochastic approach to word frequency distributions. The multinomial and Poisson models are introduced as means for obtaining theoretical expressions for the expected vocabulary and the frequency spectrum as functions of the sample (text) size. The important concept of the Large Number of Rare Events Zone (LNRE ZONE) is introduced. It is shown that many empirical samples are located in this zone where relative sample frequencies are biased estimates of population probabilities. The consequences for the construction of theoretical models for word frequency distributions are considered in detail. Three such models for LNRE distributions are discussed, Carroll's (1967, 1969) lognormal 'law', Sichel's (1975, 1986) generalized inverse Gauss-Poisson 'law', and Orlov and Chitashvili's (1982a, 1982b, 1983a, 1983b) extended generalized Zipf's 'law'. In section 4 rationales for the lognormal 'law' and various extensions

of Zipf's 'law' are discussed. Section 5 outlines how statistical analyses with LNRE models may be carried out. Finally, section 6 discusses the relation between morphological productivity and the LNRE property of running texts.

Since our aim is to give a bird's eye view of the main results obtained in the study of word frequency distributions, mathematical proofs have been ommitted, many of the results reviewed here being common knowledge ever since Yule's (1944) seminal study and the important papers by Good (1953), Good and Toulmin (1956) and Kalinin (1965). For an in-depth mathematical discussion of the to our mind central notion of LNRE distributions the reader is referred to Khmaladze and Chitashvili (1989), part of which has appeared in English as Khmaladze (1987).

2. LNRE Features of Word Frequency Distributions

In this section we introduce some general properties of word frequency distributions. We first present some techniques for describing the frequency spectrum, and then turn to review some of the 'laws' supposedly governing word frequency distributions suggested in the literature.

2.1. The Frequency Spectrum

We can view a running text as an ordered sequence of word tokens

$$(w_1, w_2, \ldots, w_N).$$

Usually the observed (or empirical) vocabulary $\underline{\hat{V}}$,

$$\underline{\hat{V}} = (A_1, A_2, \ldots, A_{\hat{V}}), \tag{1}$$

the (arbitrarily ordered) set of different words (or word types) used in the text, or, alternatively,

$$\underline{\hat{V}}_o = (A_{(1)}, A_{(2)}, \ldots, A_{(\hat{V})}), \tag{2}$$

the set of word types ordered according to their (token) frequencies,

$$f_N(A_{(1)}) \geq f_N(A_{(2)}) \geq \ldots \geq f_N(A_{(\hat{V})}), \tag{3}$$

contains a much smaller number $\hat{V}$ of elements then the sample size (or text size) N. This makes it convenient to present a text in the form of an array $\underline{A}$

$$\underline{A} = \left(\begin{array}{lllllll} A_{(1)}: & \tau_1(A_{(1)}) & \tau_2(A_{(1)}) & \cdots & \cdots & & \tau_{f(A_{(1)})}(A_{(1)}) \\ A_{(2)}: & \tau_1(A_{(2)}) & \tau_2(A_{(2)}) & \cdots & \tau_{f(A_{(2)})}(A_{(2)}) & & \\ & \cdots & \cdots & \cdots & & & \\ A_{(i)}: & \tau_1(A_{(i)}) & \tau_2(A_{(i)}) & & & & \\ A_{(i+1)}: & \tau_1(A_{(i+1)}) & \tau_2(A_{(i+1)}) & & & & \\ & \cdots & & & & & \\ A_{(j)}: & \tau_1(A_{(j)}) & & & & & \\ A_{(j+1)}: & \tau_1(A_{(j+1)}) & & & & & \\ & \cdots & & & & & \\ A_{(\hat{V}-1)}: & \tau_1(A_{(\hat{V}-1)}) & & & & & \\ A_{(\hat{V})}: & \tau_1(A_{(\hat{V})}) & & & & & \end{array} \right) \tag{4}$$

in which $\tau_1(A_{(i)})$, $\tau_2(A_{(i)})$,... indicate the positions of word $A_{(i)}$ in the text. For instance, $\tau_7(A_{(i)}) = 137$ denotes that word $A_{(i)}$ occurred for the seventh time on the 137th stage (trial)). Note that in (4) the highest frequency type is on the first line and that the so-called hapax legomena, the types occuring once only, occupy lines j down to $\hat{V}$.

Corresponding to the sample frequencies we have the sample relative frequencies $\hat{p}(A_i)$ and $\hat{p}(A_{(i)})$ for the unordered and frequentially ordered vocabulary items respectively:

$$\hat{p}(A_i) = \frac{f_N(A_i)}{N} \quad (cf.\ (1)) \tag{5}$$

$$\hat{p}(A_{(i)}) = \frac{f_N(A_{(i)})}{N} \quad (cf.\ (2)) = \hat{p}_N\{i\}\ . \tag{6}$$

The information contained in $\underline{A}$ can be used for various purposes. For instance, the transition probabilities

$$\hat{p}_N(A_i|A_j) = \frac{1}{f_N(A_j)} \sum_{n=1}^{f_N(A_j)} \sum_{k=1}^{f_N(A_i)} \mathbb{I}_{[\tau_k(A_i) = \tau_n(A_j)+1]} \tag{7}$$

can be used to study dependencies between words as they occur in some text. In this paper we focus on the analysis of the frequency distribution, i.e. the set

$$(f_N(A_1), f_N(A_2), \ldots, f_N(A_{\hat{V}}))$$

of lengths of rows in array $\underline{A}$. To restrict attention to the frequency distribution is to use the information which is invariant with respect to permutations of elements both in the sample and in the vocabulary.

The characteristic feature of the samples (texts, morphological categories) we are to investigate in this paper is that besides the elements with high (token) frequencies (e.g. $\hat{p}_N(A_{(1)}) \approx 0.05$), in the above array the upper rows of substantial length, we observe many elements that occur only once, twice, etc. Crucially, these events constitute a significant part of the vocabulary. Often the number of elements occurring only once approximates half the observed vocabulary size. We will refer to distributions with this characteristic as Large Number of Rare Events (LNRE) distributions.

The frequency distribution can be presented in at least four equivalent forms:

1. The frequency spectrum:

Let $\hat{V}_N(m)$ denote the number of elements of the vocabulary which occurred m times in a sample of size N:

$$\hat{V}_N(m) = \sum_{i\geq1} \mathbb{I}_{[f(A_i) = m]}\ , \quad m = 1, 2, \ldots, \tag{8}$$

where $\mathbb{I}_{[f(A_i) = m]}$ is the indicator of the event $[f(A_i) = m]$, i.e.

$$\mathbb{I}_{[f(A_i) = m]} = \begin{cases} 1 & \textit{if } f(A_i) = m \\ 0 & \textit{otherwise.} \end{cases}$$

It is easy to observe that the (empirical) vocabulary size for sample size N is given by

$$\hat{V}_N = \sum_{m\geq1} \hat{V}_N(m). \tag{9}$$

We shall often make use of the relative frequency spectrum

$$\hat{\alpha}_N(m) = \frac{\hat{V}_N(m)}{\hat{V}_N}, \quad m = 1,2,\dots \; . \tag{10}$$

Figure 1 illustrates these functions for the English suffix *-ness* as it appears in the Cobuild corpus (Sinclair 1987). (Here, and in all examples to follow, the frequency count is lemma-based, inflectional variants of a stem being counted as tokens of one and the same lemma type.) Note that the number of hapaxes $\hat{V}_N(m)$ is approximately $\hat{V}_N/2$.

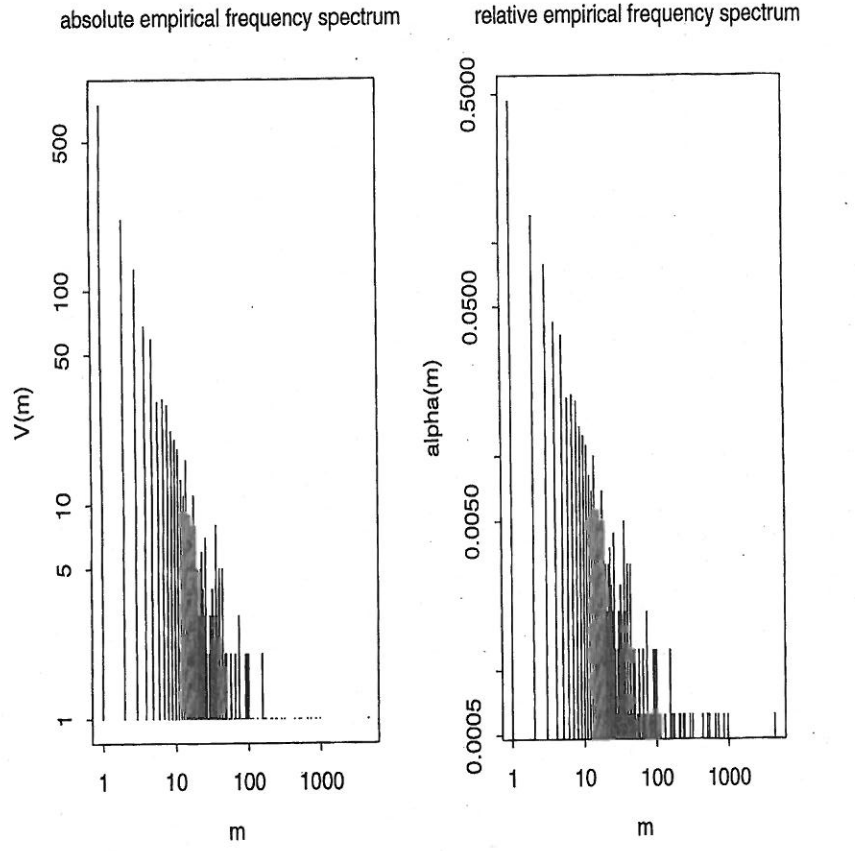


Figure 1. Absolute and relative frequency spectrum for the English suffix *-ness* as it appears in the Cobuild corpus, plotted on a double logarithmic scale.

2. The rank frequency distribution:

Given that the elements of the vocabulary

$$\hat{V} = (A_1, A_2, \dots , A_{\hat{V}})$$

are ordered according to decreasing frequency as specified in (3), we can denote any word by its rank r (its row number in $\underline{A}$) in the resulting list:

$$f_N\{r\} = f_N(A_{(r)}), \quad r = 1,2,\dots \; . \tag{11}$$

This way of representing word frequency distributions is well known from the early studies by Zipf (1935) onwards.

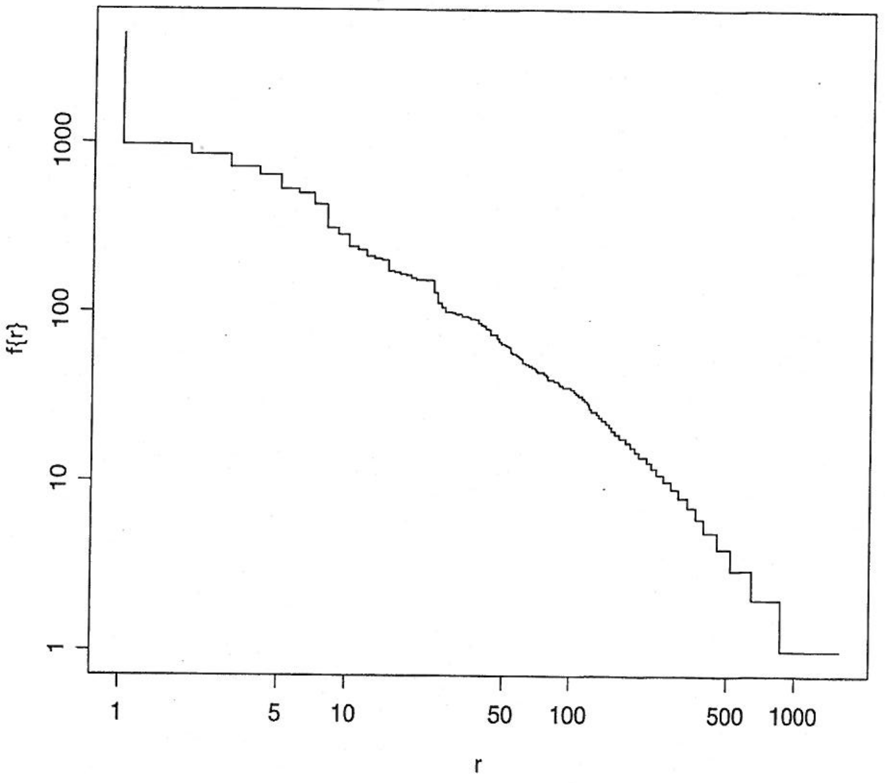


Figure 2. Rank-frequency plot for the English suffix *-ness* as it appears in the Cobuild corpus. The X-axis and the Y-axis are scaled logarithmically.

Sometimes it is more natural to consider the relative rank-frequency distribution

$$\hat{p}_N\{r\} = \frac{f(A_{(r)})}{N} .$$

Thus

$$(\hat{p}_N\{r\},\ 1 \le r \le \hat{V}_N)$$

is the ordered set of relative frequencies

$$(\hat{p}_N(A_i),\ \ 1 \le i \le \hat{V}_N).$$

Note that, as shown in figure 2, it is often convenient (and more demonstrative) to present graphs of the rank frequency distribution (or of the structural distributions to be discussed below) in a double logarithmic scale, that is, to consider the transformed step function

$$log_a\ \hat{p}_N\{[a^x]\},\ \ x \ge 0$$

of a variable $x = log\ r$, where we use the notation $[a^x]$ to denote the integer part of a^x. Usually, e or 10 are chosen for the logarithmic base a.

3. The empirical structural type distribution

The cumulative type frequency or empirical structural type distribution is defined in terms of the type probability p in the sample:

$$\hat{G}_N(p) = \sum_{i\ge1} \mathbb{I}_{[f_N(A_i) \ge Np]} = \sum_{i\ge1} \mathbb{I}_{[\hat{p}_N(A_i) \ge p]}. \qquad (12)$$

In (12), $\hat{G}_N(p)$ denotes the number of elements of the vocabulary which occurred at least Np times in the sample (text).

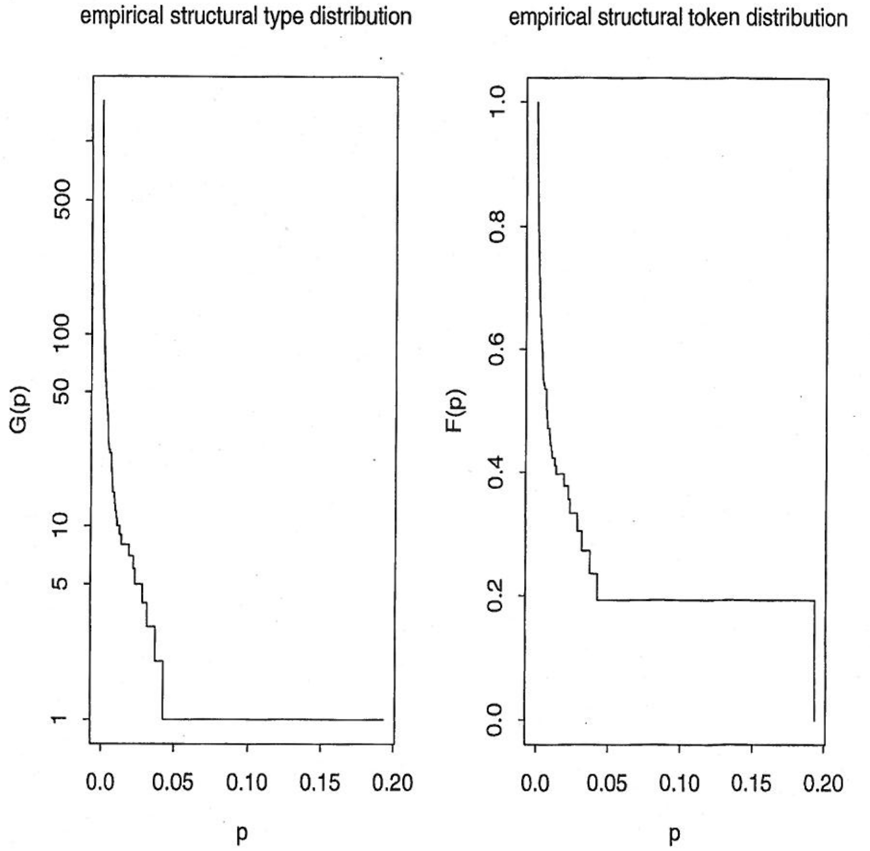


Figure 3. The empirical structural type and token distributions for the English suffix *-ness*.

4. The empirical structural token distribution

The cumulative token frequency or empirical structural token distribution is defined as

$$\hat{F}_N(p) = \sum_{i\ge1} \hat{p}_N(A_i)\mathbb{I}_{[\hat{p}_N(A_i) \ge p]}. \qquad (13)$$

So $\hat{F}_N(p)$ is the relative frequency of those tokens in the sample which are instances of types with a relative frequency not less then p. Sometimes we will refer to both $\hat{G}$ and $\hat{F}$ as empirical structural distributions. Figure 3 plots these functions for the suffix *-ness*. Note how the presence of a single very high frequency type effects a sizeable difference in the shape of the two graphs.

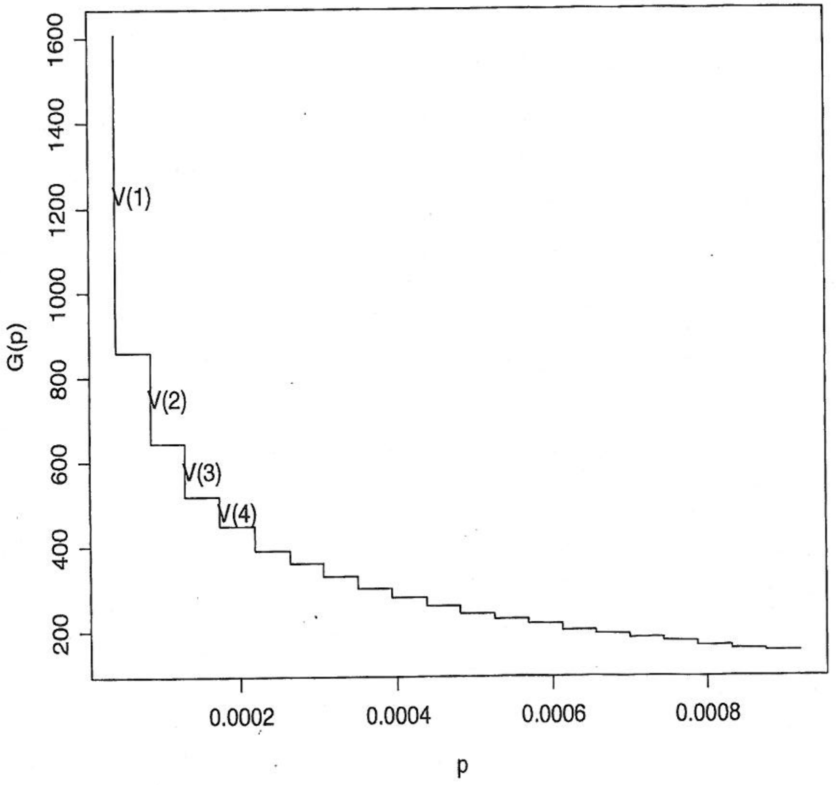


Figure 4. The relation between the frequency spectrum and the empirical structural token distribution. $\hat{G}_N(p)$ is shown for the first 23 distinct values of p for the suffix *-ness*. The corresponding spectrum elements have been added for $m = 1, ..., 4$.

These four ways of representing the frequency distribution are fully equivalent. This becomes apparent when we make explicit the relations that hold between them:

(a) The terms of the frequency spectrum $(\hat{V}_N(m))_{m\ =\ 1,2,...}$ can be expressed in terms of the empirical structural type distribution,

$$\hat{V}_N(m) = \hat{G}_N(\frac{m}{N}) - \hat{G}_N(\frac{m+1}{N}), \quad m = 1,2,...$$

as shown in figure 4. Equivalently, we have

$$\hat{G}_N(\frac{m}{N}) = \sum_{k \geq m} \hat{V}_N(k)$$

(b) The empirical structural type and token distributions are related by the equality

$$\hat{G}_N(p) = \sum_{q \geq p} \frac{1}{q} \Delta \hat{F}_N(q),$$

where $\Delta \hat{F}_N(q)$ is a finite difference (i.e. the jump value) of the step function $\hat{F}_N(p)$ at a point q. Or, equivalently,

$$\hat{F}_N(p) = \frac{1}{N} \sum_{m \geq Np} m \hat{V}_N(m).$$

(c) The structural distribution $\hat{G}_N(p)$ is an inverse function to the rank-frequency distribution $f_N\{r\}$, i.e.

$$\hat{G}_N\left(\frac{f_N\{r\}}{N}\right) = \hat{G}_N(\hat{p}_n\{r\}) = r, \quad r = 1,2,... . \tag{14}$$

Some probabilistic meaning can be given to the relative frequency spectrum $\hat{\alpha}_N(m)$. If the empirical vocabulary of distinct word types is conceived of as constituting the experimental population from which we are sampling a type at random, then is the probability that some word type having token frequency m will be chosen, or, equivalently, $\hat{G}_N(p)/\hat{V}_N$ is the probability that some word type having a relative frequency of at least p will be chosen. A stochastic interpretation of the token probability distribution $\hat{F}_N(p)$ is given in section 3.1.3.

2.2. Laws Proposed for Frequency Spectra

A number of simple analytical expressions have been suggested in the literature for 'theoretical laws', either for the relative frequency spectrum or for rank-frequency distributions. In terms of the relative spectrum these 'laws' can be presented as follows:

1. Zipf (Zipf 1935)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{1}{m(m+1)}, \tag{15}$$

2. Yule (Yule 1924; Simon 1955)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\Gamma(\beta+1)\Gamma(m)\beta}{\Gamma(m+\beta+1)}, \quad (\beta > 0), \tag{16}$$

3. Yule-Simon (Simon 1956, 1960)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\beta}{(m+\beta-1)(m+\beta)}, \quad (\beta > 0)\ , \tag{17}$$

4. Waring-Herdan-Muller (Herdan 1960, 1964; Muller 1979)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\Gamma(\beta+1)\alpha}{\Gamma(\beta+1-\alpha)} \cdot \frac{\Gamma(m+\beta-\alpha)}{\Gamma(m+\beta+1)}, \quad (0<\alpha<1,\ \beta>\alpha), \tag{18}$$

5. Karlin-Rouault (Rouault 1978)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\alpha\Gamma(m-\alpha)}{\Gamma(1-\alpha)\Gamma(m+1)}, \quad (0 < \alpha < 1), \tag{19}$$

6. Zipf-Mandelbrot (Mandelbrot 1962)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{1}{m^{\gamma}} - \frac{1}{(m+1)^{\gamma}}, \quad (\gamma > 0). \tag{20}$$

We will refer to these 'laws' as the Zipfian family of models.

Graphs of these 'laws' for varying parameter values are shown in figure 5 and figure 6. Note that in the case of the Waring-Herdan 'law' increasing the value of α leads to an increase in type richness, as evidenced by the values of $\alpha_N(1)$ and the ratio $r_V = \beta/(\beta - \alpha)$ by which $\hat{V}_N$ has to be multiplied to obtain the theoretical vocabulary V. Decreasing β similarly leads to higher values of V. Also observe that especially high values of V are obtained when the difference between β and α is small. Finally, note that for the Karlin-Rouault 'law', the parameter α equals $\alpha_N(1)$.

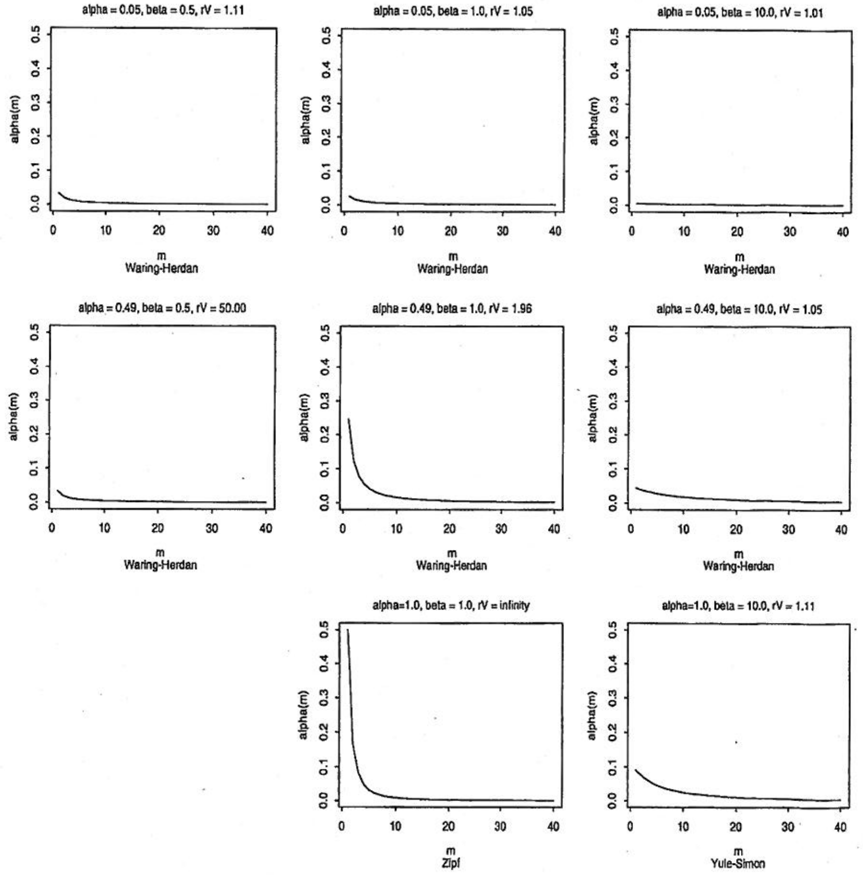


Figure 5. The Zipfian family of 'laws' for selected parameter values: Waring-Herdan, Yule-Simon and Zipf.

The corresponding 'laws' for the rank-frequency distribution may be obtained using the relation (14) between the cumulative structural distribution G and the rank-frequency distribution $p\{r\}$. In fact, in the case of Zipf 's 'law', for instance, the corresponding model for the structural distribution $\hat{G}(p)$ should be any function $\hat{G}(r)$ with the property

$$\frac{\hat{G}(\frac{m}{N}) - \hat{G}(\frac{m+1}{N})}{\hat{G}(\frac{1}{N})} = \frac{1}{m(m+1)} .$$

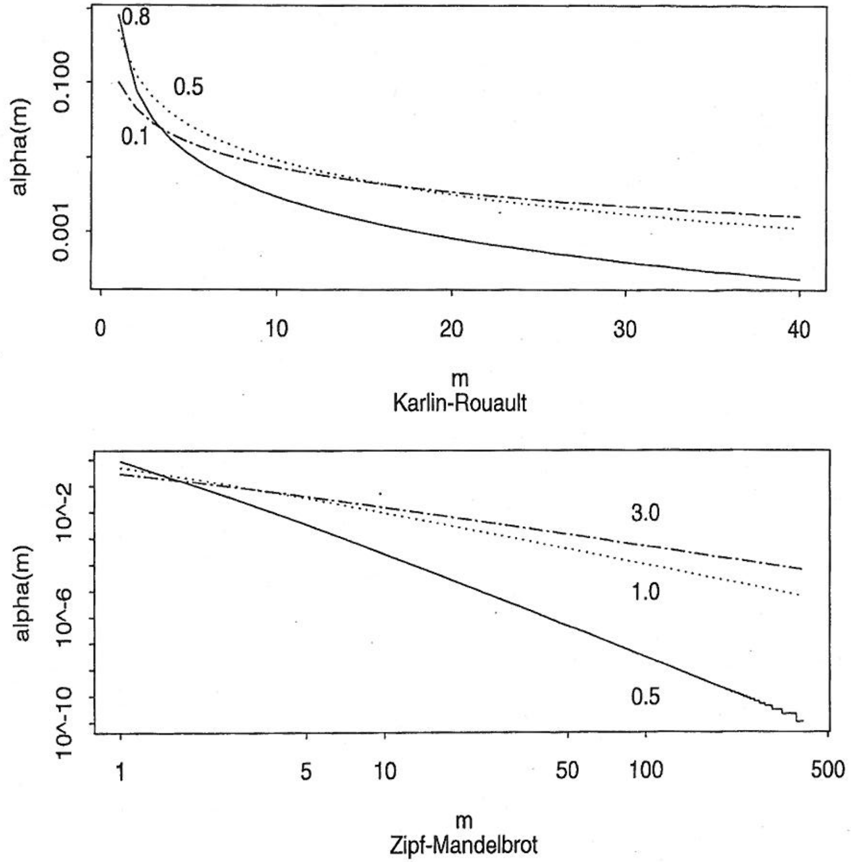


Figure 6. The Zipfian family of 'laws' for selected parameter values: Karlin-Rouault and Zipf-Mandelbrot.

The solution is simply

$$\hat{G}(\frac{m}{N}) = \frac{NC}{m} + B, \tag{21}$$

with some parameters (C,B). Again using relation (14) we find that the corresponding rank-frequency distribution should have the form

$$\hat{p}_N\{r\} \approx p\{r\} = \frac{C}{r - B}$$

as a solution for the equation

$$\frac{C}{\hat{p}_N\{r\}} + B = r.$$

In the case of the Zipf-Mandelbrot 'law' we similarly have

$$\overline{G}\left(\frac{m}{N}\right) = \frac{C}{\left(\frac{m}{N}\right)^{\gamma}} + B \tag{22}$$

and

$$\hat{p}_N\{r\} \approx p\{r\} = \left(\frac{C}{r - B}\right)^{1/\gamma} \tag{23}$$

for some parameters C and B. Graphs of these distributions for varying choices of the parameters are shown in figure 7. Note that small values of γ effect a downward curvature for the lower ranks r without influencing the shape of the curve for the higher ranks. We will return to the independence of the head and tail of the frequency distribution in section 3.3.3.

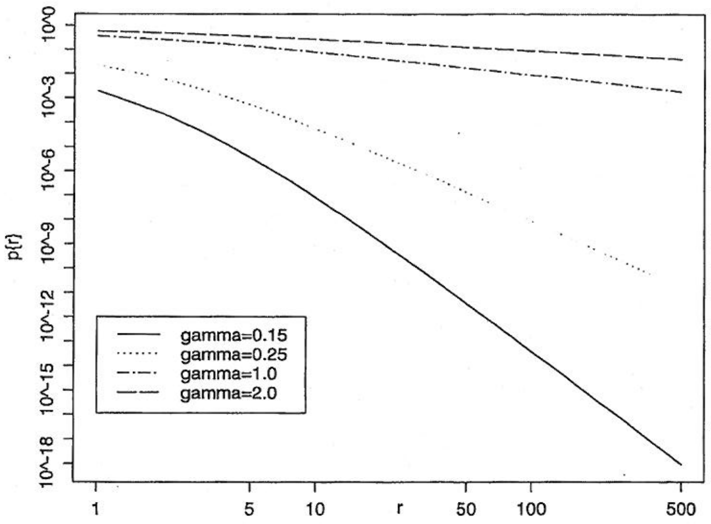


Figure 7. Rank-frequency distributions: Zipf and Zipf-Mandelbrot distributions for $B = -1.5$, $C^{1/\gamma} = 0.9$ and varying γ.

3. Stochastic Modelling of LNRE

In this section we first introduce expressions for the structural distributions using the multinomial and Poisson models. We then define the concepts of the LNRE ZONE and the generalized structural distribution. Finally, we consider the rationale for generalized structural distributions using an asymptotic approach.

3.1. The Structural Distribution in the Classical Scheme

3.1.1. The Multinomial Model

Even though the four forms in which we may represent the frequency distribution are fully equivalent in that they summarize exactly the same information, we will focus on the structural distributions since it is the structural distributions which contain explicit expressions for the relevant probabilistic characteristics.

Assuming the classical scheme of independent identically distributed trials, let

$$(\mathrm{P}(A_i),\ 1 \le i \le V)$$

be the probability distribution over the set

$$\underline{V} = (A_1, A_2, \ldots, A_V)$$

of elements of the theoretical vocabulary.

As direct analogues for the empirical structural distributions we consider the following expressions:

$$G(p) = \sum_{i=1}^{V} \mathbb{I}_{[p(A_i) \ge p]}, \quad p \ge 0, \tag{24}$$

$$F(p) = \sum_{i=1}^{V} p(A_i)\mathbb{I}_{[p(A_i) \ge p]}, \quad p \ge 0\ . \tag{25}$$

The functions $G(p)$ and $F(p)$ can be interpreted in the same way as their empirical analogues $\hat{G}_N(p)$ and $\hat{F}_N(p)$, be it that the general population of words is considered instead of the experimental sample population. We shall refer to these functions as the theoretical structural type and token distributions, or, alternatively, as the theoretical cumulative type and token probability distributions.

Let's now consider how the theoretical and empirical distributions are related. It is a well known fact that the vector of frequencies

$$(f_N(\mathrm{A}_1),\ \mathrm{f}_N(\mathrm{A}_2),\ \ldots,\ \mathrm{f}_N(\mathrm{A}_V)) \tag{26}$$

is multinomially distributed:

$$\boldsymbol{Pr}(f_N(A_i) = n_i\ ,\ \ 1 \le i \le V) = \binom{N}{n_1,\ n_2,\ \ldots,\ n_V} \prod_{i=1}^{V} p(A_i). \tag{27}$$

For the important special case of binomial probabilities and the corresponding upper-cumulative probabilities we will use the notations

$$B(N,\ m,\ p) = \binom{N}{m} p^m (1 - p)^{N-m} \tag{28}$$

$$B^+(N,\ m,\ p) = \sum_{k \ge m} \binom{N}{k} p^k (1 - p)^{N-k}\ . \tag{29}$$

Similarly, trinomial probabilities will be referred to as

$$T(N,\ m,\ l,\ p,\ q) = \binom{N}{m,\ k} p^m q^k (1 - p - q)^{N-m-k}\ .$$

For the expected values and covariances of the indicators

$$\mathbb{I}_{[f_N(A_i) = m]},\ m \ge 1$$

we have

$$\boldsymbol{E}\, \mathbb{I}_{[f_N(A_i) = m]} = \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} = B(N,\ m,\ p(A_i)) \tag{30}$$

and

$$\begin{aligned} \boldsymbol{COV}(\,\mathbb{I}_{[f_N(A_i) = m]},\ \mathbb{I}_{[f_N(A_j) = k]}) &= \delta_{ij}\delta_{mk} B(n,\ m,\ p(A_j)) \\ &+ (1 - \delta_{ij}) T(N,\ m,\ k,\ p(A_i),\ p(A_j)) \\ &- B(N,\ m,\ p(A_i)) B(N,\ k,\ p(A_j)), \end{aligned} \tag{31}$$

where

$$\delta_{ij} = \begin{cases} 1 & \textbf{\textit{if } } i = j \\ 0 & \textbf{\textit{otherwise.}} \end{cases}$$

Similarly,

$$\boldsymbol{E}\,\mathbb{I}_{[f_N(A_i) \geq m]} = B^+(N,\ m,\ p(A_i))$$

and

$$\begin{aligned} \boldsymbol{COV}(\mathbb{I}_{[f_N(A_i) \geq m]}\ ,\ \mathbb{I}_{[f_N(A_j) \geq k]}) \quad &= \delta_{ij} B^+(N,\ \max(m,\ k),\ p(A_i)) \\ &+ (1 - \delta_{ij}) \sum_{l \geq m, r \geq k} T(N,\ l,\ r,\ p(A_i),\ p(A_j)) \\ &- B^+(N,\ m,\ p(A_i))B^+(N,\ k,\ p(A_j)). \qquad (32) \end{aligned}$$

Now the expected values for the frequency spectrum and the empirical distributions $\hat{G}_N(p)$ and $\hat{F}_N(p)$ can be expressed as

$$\begin{aligned} V_N(m) = \boldsymbol{E}\hat{V}_N(m) \quad &= \boldsymbol{E} \sum_i \mathbb{I}_{[f_N(A_i) = m]} \\ &= \sum_i \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \qquad (33) \end{aligned}$$

$$\begin{aligned} V_N = \boldsymbol{E}\hat{V}_N \quad &= \boldsymbol{E} \sum_{m \geq 1} \hat{V}_N(m) \\ &= \boldsymbol{E} \sum_m \sum_i \mathbb{I}_{[f_N(A_i) = m]} \\ &= \sum_{m \geq 1} \sum_i \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i (1 - (1 - p(A_i))^N) \qquad (34) \end{aligned}$$

$$\begin{aligned} G_N(p) \ &= \boldsymbol{E}\hat{G}_N(p) = \sum_i \sum_{m \geq Np} \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i B^+(N,\ Np,\ p(A_i)) \qquad (35) \end{aligned}$$

$$\begin{aligned} F_N(p) = \boldsymbol{E}\hat{F}_N(p) \quad &= E\frac{1}{N} \sum_{m \geq Np} m\hat{V}_N(m) \\ &= \sum_i \sum_{m \geq Np} \frac{m}{N} \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i p(A_i) B^+(N - 1,\ Np - 1,\ p(A_i)) \qquad (36) \end{aligned}$$

and

$$\begin{aligned} &\boldsymbol{COV}(\hat{V}_N(m),\ \hat{V}_N(k)) = \\ &= \sum_i \binom{N}{m} (p(A_i))^m (1 - p(A_i))^{N - m} \\ &+ \sum_{ij} \binom{N}{m,\ k} (p(A_i))^m (p(A_j))^k (1 - p(A_i) - p(A_j))^{N-m-k} \\ &- \sum_i \binom{N}{m,\ k} (p(A_i))^{m + k} (1 - 2p(A_i))^{N - m - k} \\ &- \sum_i \binom{N}{m} (p(A_i))^m (1 - p(A_i))^{N-m} \sum_i \binom{N}{k} (p(A_i))^k (1 - p(A_i))^{N-k}. \quad (37) \end{aligned}$$

Similar expressions can be obtained for the covariances of other characteristics of the frequency distribution $(\boldsymbol{COV}(\hat{F}_N(p),\ \hat{F}_N(p')))$ for instance) as linear combinations of the spectrum, but we omit them as we will be using simpler versions based on the Poisson model.

We shall now make a rather formal step to rewrite these expressions in integral form to show that (any) probabilistic characteristics of frequency distributions can be expressed in terms of the corresponding theoretical structural distributions. This will also allow us to express further generalizations in a natural way.

Since the theoretical structural type distribution *G(p)* is a (nonincreasing) step function defined on the interval [0,1] with jumps at the points $(p_1, p_2, \ldots, p_V)$,

$$\Delta G(p_i) = G(p_i) - G(p_i+), \qquad (38)$$

where $p_i+ = \lim_{p \downarrow p_i} G(p)$, and similarly for the structural token distribution,

$$\Delta F(p_i) = p_i(G(p_i) - G(p_i+)), \qquad (39)$$

sums of the form

$$S = \sum_{i=1}^{V} h(p_i),$$

with h some function of p, can be written as Stieltjes integrals:

$$\begin{aligned} S &= \sum_{p} h(p)\Delta G(p) \\ &= \int_0^1 h(p)dG(p); \\ S &= \sum_{p} h(p)\frac{\Delta F(p)}{p} \\ &= \int_0^1 h(p)\frac{dF(p)}{p}. \end{aligned}$$

We can now rewrite the expected frequency distribution as

$$\begin{aligned} V_N(m) &= E\hat{V}_N(m) = \int_0^1 \binom{N}{m} p^m(1 - p)^{N-m} dG(p) \\ &= \int_0^1 \binom{N}{m} p^m(1 - p)^{N-m} \frac{dF(p)}{p} \end{aligned} \qquad (40)$$

$$V_N = E\hat{V}_N = \int_0^1 (1 - (1 - p)^N)dG(p) \qquad (41)$$

$$G_N(p) = E\hat{G}_N = \int_0^1 B^+(N, Np, q)dG(q) \qquad (42)$$

$$F_N(p) = E\hat{F}_N(p) = \int_0^1 qB^+(N - 1, Np - 1, q)dG(q). \qquad (43)$$

In the same way the covariances can be presented as

$$\begin{aligned} COV(\hat{V}_N(m), \hat{V}_N(k)) &= \delta_{ij} E\hat{V}_N(m) \\ &+ \int_0^1\int_0^1 T(N, m, k, p, q)dG(p)dG(q) \\ &- \int_0^1 T(N, m, k, p, p)dG(p) \\ &- E\hat{V}_N(m)E\hat{V}_N(k). \end{aligned} \qquad (44)$$

Note that we do not exclude the possibility that the theoretical vocabulary may be infinite, i.e. $V = \infty$. This is the reason that we consider the upper cumulative distributions G and F, using for brevity the notation dG, dF instead of ($-dG$), ($-dF$).

3.1.2. The Poisson Model

Generally, we may consider the multinomial model within the framework of the Poisson model of the (sampling) experiment. In fact, if we assume that the frequencies of the vocabulary elements

$$(f_t(A_1), f_t(A_2), ..., f_t(A_V)),$$

are independent Poisson processes in continuous time $t \geq 0$ with parameters (intensities)

$$(\lambda(A_1), \lambda(A_2), ..., \lambda(A_V)),$$

then the vector of frequencies

$$(f_{T_N}(A_1), f_{T_N}(A_2), ..., f_{T_N}(A_V))$$

observed at moments in time when the number of observed tokens is increasing,

$$T_N = \min\{t: \sum_{i=1}^{V} f_t(A_i) = N\},$$

is multinomially distributed according to the probability distribution

$$(p(A_i) = \frac{\lambda(A_i)}{\sum_{j=1}^{V}\lambda(A_j)}, \quad 1 \le i \le V).$$

Interestingly, for LNRE samples it may be assumed that for the terms of the frequency spectrum the multinomial and the Poisson schemes are (asymptotically) equivalent. In the Poisson scheme the expressions for the expected values become simpler. In particular, we now have

$$\begin{aligned} E\hat{V}_t(m) &= \int_0^\infty \frac{(\lambda t)^m}{m!} e^{-\lambda t} dG(\lambda) \\ &= \int_0^\infty \Pi(t, m, \lambda) dG(\lambda) \end{aligned} \tag{45}$$

$$E\hat{V}_t = \int_0^\infty (1 - e^{-\lambda t}) dG(\lambda) \tag{46}$$

$$E\hat{G}_t(\lambda) = \int_0^\infty \Pi^+(t, t\lambda, x) dG(x) \tag{47}$$

$$E\hat{F}_t(\lambda) = \int_0^\infty x\Pi^+(t, t\lambda - 1, x) dG(x), \tag{48}$$

where $(\hat{G}_T(\lambda), \hat{F}_t(\lambda))$ and $(G(\lambda), F(\lambda))$ play the role of empirical and theoretical distributions for type and token intensities in the general population. We use the notations

$$\Pi(t, m, \lambda) = \frac{(\lambda t)^m}{m!} e^{-\lambda t} \tag{49}$$

$$\Pi^+(t, m, \lambda) = \sum_{k \ge m} \Pi(t, k, \lambda) \tag{50}$$

for the Poisson probabilities and the corresponding upper sums.

The expressions for covariances are simplified significantly since the trinomial distribution is substituted formally as follows:

$$\binom{N}{m,\ k}(p(A_i)^m(p(A_j))^k(1-p(A_i)-p(A_j))^{N-m-k} \approx \frac{(\lambda(A_i)t)^m}{m!}\frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_i)t - \lambda(A_j)t}.$$

Hence, for instance,

$$\begin{aligned} COV(\hat{V}_t(m), \hat{V}_t(k)) &= \sum_i \frac{(\lambda(A_i)t)^m}{m!} e^{-\lambda(A_i)t} \\ &+ \sum_{ij} \frac{(\lambda(A_i)t)^m}{m!}\frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_i)t-\lambda(A_j)t} \\ &- \sum_i \frac{(\lambda(A_i)t)^{m+k}}{(m+k)!}\binom{m+k}{m}\frac{1}{2^{m+k}} e^{-2\lambda(A_i)t} \\ &- \sum_i \frac{(\lambda(A_i)t)^m}{m!} e^{-\lambda(A_i)t} \sum_j \frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_j)t} \\ &= E\hat{V}_t(m) - \binom{m+k}{m}\frac{1}{2^{m+k}} E\hat{V}_{2t}(m+k). \end{aligned} \tag{51}$$

For reasons of expositional clarity we will henceforth use the more traditional notation p (probability) instead of λ (intensity) in expressions making use of the Poisson model, even though p may now range over the whole interval $[0,\infty)$ rather than $[0,1]$. Similarly pN will replace λt.

3.1.3. Stochastic Interpretation of the Token Probability Distribution

Further insight into the structural token distribution can be gained by investigating how the text may be generated stochastically. Consider associating with any word token w_n, $1 \le n \le N$ in the running text both its relative frequency and its (theoretical) probability, such that the text is viewed as a series of triplets (word token, relative frequency, probability):

$$\begin{matrix} w_1, & w_2, & \dots, & w_N \\ \hat{p}_N(w_1), & \hat{p}_N(w_2), & \dots, & \hat{p}_N(w_N) \\ p(w_1), & p(w_2), & \dots, & p(w_N). \end{matrix}$$

The probabilities on the second row are sampled from a population with distribution $\hat{F}$(without replacement). The probabilities on the third row are sampled from a general population with distribution F (with replacement).

Now the following scheme for the stochastic generation of texts can be suggested. At each n-th stage of the experiment, first generate the (random) probability p_n according to the distribution $F(p)$, then choose some appropriate word type A_i from all words in the vocabulary for which $p(A_i) = p(w_n) = p_n$. To define the last step of this experimental scheme somewhat more precisely, let

$$V(p,\ q) = (A_i\!: p \leq p(A_i) \leq q)$$

be the part of the vocabulary $\underline{V}$ consisting of words with probability falling in the interval $[p,\ q]$. Obviously, the number of elements $V(p,\ q)$ in this 'subvocabulary' is just

$$\begin{aligned} V(p,\ q) &= \sum_i \mathbb{I}_{[p < p(A_i) \leq q]} \\ &= \int_p^q dG(x) \\ &= \int_p^q \frac{1}{x} dF(x). \end{aligned} \tag{52}$$

Now it is easy to see that the initial multinomial scheme of experiment is equivalent to that described above if only word w_n is supposed to be chosen by chance (i.e., according to the uniform distribution) from the subvocabulary $V(p_n, q_n)$. In that case the variables $(w_1, w_2, \ldots)$ are independently and identically distributed. In addition,

$$\begin{aligned} \boldsymbol{Pr}(w_n = A_i) &= \frac{1}{V(p(A_i),\ p(A_i))} \int_{p(A_i)}^{p(A_i)} dF(p) \\ &= \frac{\Delta F(p(A_i))}{\Delta G(p(A_i))} = p(A_i). \end{aligned}$$

Thus the running text can be viewed as the realization of an experiment governed by two stochastic mechanisms:

1. the token probability distribution $F(p)$ to generate the probabilities p_n at each stage, and
2. the (conditional) distribution $W(A \mid p)$ to generate words w_n from subvocabularies corresponding to the probability p_n occurring at this stage.

Moreover, in as far as we are restricting ourselves to the analysis of frequency distributions, and since the particular character of the second mechanism (notably the assumption of independence) does not affect the conclusions made on the basis of the frequency distribution data, we can accept far more general hypotheses concerning the nature of the word distribution scheme, the only requirement being that for any interval $[p,\ q]$ the elements of the subvocabulary $V(p,\ q)$ should be uniformly distributed over the set

$$\tau_1(p,\ q),\ \tau_2(p,\ q),\ \tau_3(p,\ q),\ \ldots$$

of positions through the running text at which the occurring probabilities p_n fell in the interval $[p,\ q]$.

3.1.4. Interpolation

We sometimes need expressions for the vocabulary or the frequency spectrum for sample sizes smaller than N. More precisely, if $(\hat{V}_n(m),\ m = 1,2,\ldots)$ is a frequency spectrum observed on a sample of size N, the question is how to estimate the frequency spectrum $(\hat{V}_n(m),\ m = 1,2,\ldots)$ for a subsample of the size n. The formula

$$\hat{V}_{N,n}(m) = \sum_{j \geq m} \hat{V}_N(j) \binom{j}{m} \left(\frac{n}{N}\right)^m \left(1 - \frac{n}{N}\right)^{j - m} \tag{53}$$

gives the best solution to this problem: $\hat{V}_{N,n}(m)$is a conditional expectation of the spectrum $\hat{V}_n(m)$ given the observed spectrum $(\hat{V}_N(k),\ k \geq 1)$:

$$\hat{V}_{N,n}(m) = \boldsymbol{E}(\hat{V}_n(m) | \hat{V}_N(k),\ k \geq 1),$$

that is optimal in the mean squares sense.

To see this, consider a finite population of size N consisting of $\hat{V}_N$ types of elements

$$(A_1,\ A_2,\ \ldots,\ A_{\hat{V}_N})$$

with corresponding frequencies

$$(f_N(A_1),\ \ldots,\ f_N(A_{\hat{V}_N})).$$

Let some sample of size n be taken from this population without replacement.

Denote by $\hat{V}_{N,n}(k, l)$ the number of elements with a frequency k in the population which occur l times in the sample, i.e.

$$\hat{V}_{N,n}(k, l) = \sum_{i} \mathbb{I}_{[f_N(A_i) = k,\, f_n(A_i) = l]}. \tag{54}$$

Evidently the spectrum terms in the sample can be presented as sums

$$\hat{V}_n(l) = \sum_{i} \mathbb{I}_{[f_n(A_i) = l]} \tag{55}$$

$$= \sum_{k \geq l} \hat{V}_{N,n}(k, l). \tag{56}$$

It can be shown that the (matrix) statistic $\hat{V}_{N,n}(k, l)$, $l \leq k$, $1 \leq k$, is distributed by the compound hypergeometric law, i.e.

$$\boldsymbol{Pr}(\hat{V}_{N,n}(k, l) = m_{k,l},\ l \leq k;\ 1 \leq k \leq N) = \binom{N}{n}^{-1} \prod_{k=1}^{N} \frac{\hat{V}_N(k)!}{m_{k,1}! \,\ldots\, m_{k,k}!} \prod_{l=1}^{k} \binom{k}{l}^{m_{k,l}} \tag{57}$$

on the domain

$$(m_{k,l}\colon \sum_{l=1}^{k} m_{k,l} = \hat{V}_N(k),\ \sum_{k} \sum_{l} m_{k,l} = n).$$

For sufficiently large sample sizes (n, N) the vectors

$$(\hat{V}_{N,n}(k, l),\ l \leq k),\ k = 1, 2, \ldots$$

are independent in k and multinomially distributed. As a result formula (53) can be derived as well as an expression for the expected vocabulary,

$$\hat{V}_{N,n} = \sum_{j \geq 1} \hat{V}_N(j) \left(1 - \left(1 - \frac{n}{N}\right)^{j}\right) \tag{58}$$

3.2. The LNRE ZONE

According to the law of large numbers, we have that for any probability distribution

$$(p(A_i),\ 1 \leq i \leq V)$$

with a finite vocabulary V the relative sample frequencies will converge to the population probabilities for ever increasing sample size N:

$$\hat{P}_N(A_i) \rightarrow Pr(A_i)$$

in probability as $N \rightarrow \infty$. As a simple consequence,

$$\hat{V}_N(m) \rightarrow 0$$

for all m. If so, the relative expected spectrum

$$\alpha_N(m) = \frac{E\hat{V}_N(m)}{E\hat{V}_N} = \frac{V_N(m)}{V_N}$$

may coinside with most of the 'laws' (15-20) only for finite samples ($N < \infty$). If one of these 'laws' appears for $N = \infty$, then the general population must be necessarily infinite too. In qualitative terms, a sample in the LNRE ZONE can be defined as a sample for which (a) the sample size is large enough to allow the estimation of the first terms of the probability rank distribution (the big probabilities), but where (b) the first terms of the relative frequency spectrum take on significant values. The questions we shall try to give an answer to by applying the analytical expressions for the expected spectrum can be formulated as follows:

1. What is the empirical criterion for the LNRE ZONE? In other words, how can we ascertian whether a sample is located in the LNRE ZONE?
2. What is the theoretical definition for the LNRE ZONE? In particular, does a theoretical distribution exist which realizes some given 'law' either on finite or on infinite samples, such that the coincidence

   $$\alpha_N(m) = \alpha(m),\ m = 1, 2, \ldots$$

   takes place for 'laws' $\alpha(m)$ such as (15-20)?

3.2.1. Locating Samples with Respect to the LNRE ZONE

In this section we address the first question, proposing two methods for ascertaining whether a sample is located in the LNRE ZONE. The first method makes use of the way the frequency spectrum develops through sampling time, the second method gauges the extent to which the sample relative frequencies are biased estimates of the population probabilities.

The following reasoning corresponds to the intuitive understanding that the LNRE ZONE must be located in the neighborhood of sample sizes where the (relative) expected spectrum terms achieve their peaks. Using the Poisson model for the terms of the expected spectrum and differentiating in the variable N we find that

$$\frac{d}{dN}V_N = \frac{1}{N}V_N(1) \tag{59}$$

$$\frac{d}{dN}V_N(m) = \frac{1}{N}[mV_N(m) - (m+1)V_N(m+1)],\ m = 1,2,\dots \tag{60}$$

$$\frac{d}{dN}\alpha_N(m) = \alpha_N(m)[m - \alpha_N(1)] - (m+1)\alpha_N(m+1),\ m = 1,2,\dots \tag{61}$$

Denote by N^*_m, $m = 1, 2, \dots$ and $\overline{N}_m$, $m = 1, 2, \dots$ the values of those sample sizes where the terms of the absolute ($V_N(m)$) or relative ($\alpha_N(m)$) expected spectrum are achieving their maximums respectively. Interestingly, the dynamic behavior of these functions is characterized by the following property. At the time moment $N = N^*_1$ at which the expected number of hapax legomena (the words occuring only once) $V_N(1)$ reaches its maximum the number of hapaxes is exactly twice that of the dislegomena $V_N(2)$: from

$$\frac{d}{dN}V_N(1) = 0$$

we have by (60) that

$$\frac{V_N(1)}{N} - \frac{2V_N(2)}{N} = 0,$$

hence $V_N(1) = 2V_N(2)$ at N^*_1. Similarly, the number of dislegomena increases until at time moment $N = N^*_2$ it becomes 3/2 of the expected number of words occurring three times, $V_N(3)$, and so on. The moments of peaks of the relative expected spectrum $\alpha_N(m)$ are arranged in the same way but with some (often substantial) anticipation,

$$\overline{N}_1 \le N^*_1,\ \overline{N}_2 \le N^*_2, \dots$$

Now suppose that Zipf 's law is realized for some sample size Z. When we substitute the expression 1/[m(m+1)] in the right hand side of (61) and take

$$\frac{d}{dN}\alpha_N(m) = 0,$$

then it is easy to see that

$$\frac{d}{dN}\alpha_N(1)\Big|_{N=Z} < 0$$

and that

$$\frac{d}{dN}\alpha(2)\Big|_{N=Z} = 0.$$

Hence we have that

$$\overline{N}_1 \le Z = \overline{N}_2 \le N^*_1.$$

Thus Z appears as the sampling time at which the relative number of expected dislegomena E $\alpha_N(2)$ achieves its maximum.

Given some observed frequency spectrum ($\hat{V}_N(1)$, $\hat{V}_N(2)$,...),we may test whether a sample is located in the LNRE zone at sample size N by inquiring whether the number of hapaxes and dislegomena are still increasing. If $N \le N^*_1$, that is,

$$\frac{1}{N}(\hat{V}_N(1) - 2\hat{V}_N(2)) > 0, \tag{62}$$

we know that the number of hapaxes is still increasing; if $N^*_1 \le N \le N^*_2$, that is,

$$\begin{cases} \frac{1}{N}(\hat{V}_N(1) - 2\hat{V}_N(2)) < 0 \\ \frac{1}{N}(2\hat{V}_N(2) - 3\hat{V}_N(3)) > 0 \end{cases} \tag{63}$$

we know that the number of hapaxes has passed its maximum while the number of dislegomena is still increasing. We will refer to the 'time' interval $(0, N_1^*]$ as the central LNRE ZONE and to the interval $(N_1^*, N_k^*]$ for small k as the late LNRE ZONE.

To see how this test can be applied to actual data, consider figures 8 and 9.

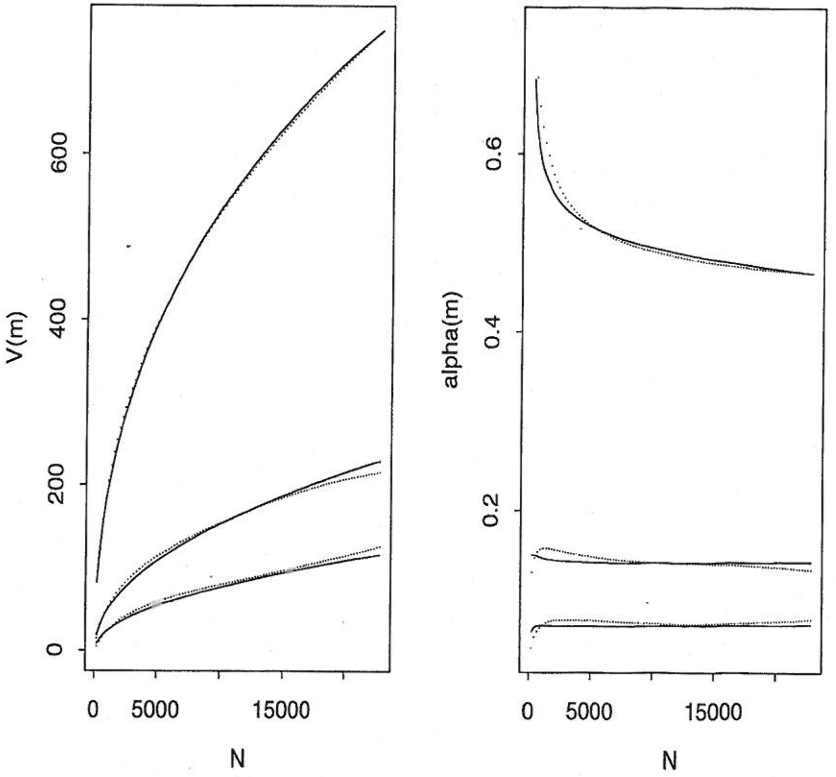


Figure 8. The development of the absolute and relative frequency spectra (m = 1, 2, 3) through sampling time: the productive English suffix *-ness*. The continuous lines are calculated on the basis of the inverse Gauss-Poisson 'law', the dotted lines are obtained by hypergeometric interpolation.

Figure 8 shows how the first three spectrum elements develop through sampling time for the productive English suffix *-ness*. Since $V_N(1)$ and $V_N(2)$ are still increasing at the observed sample size, we may conclude that this sample is located in the central LNRE ZONE. Next consider the corresponding graphs for

the unproductive English prefix *en-* (figure 9). According to the Gauss-Poisson model (see sections 3.2.2 and 5.1), the sample is at a position far beyond the late LNRE ZONE. The hypergeometric interpolation curves appear to be less useful here, due to the presence of extra maxima which are brought about by the combined presence of a substantial number of very high frequency words and a smallish number of low frequency words. In this case, the early maxima observed for small N are indicative of the sample's location outside the late LNRE ZONE. Note that the relative spectrum elements reach their maxima far earlier than the absolute spectrum elements, which is the reason why the test is formulated in terms of $V_N(m)$ rather than in terms of $\alpha_N(m)$.

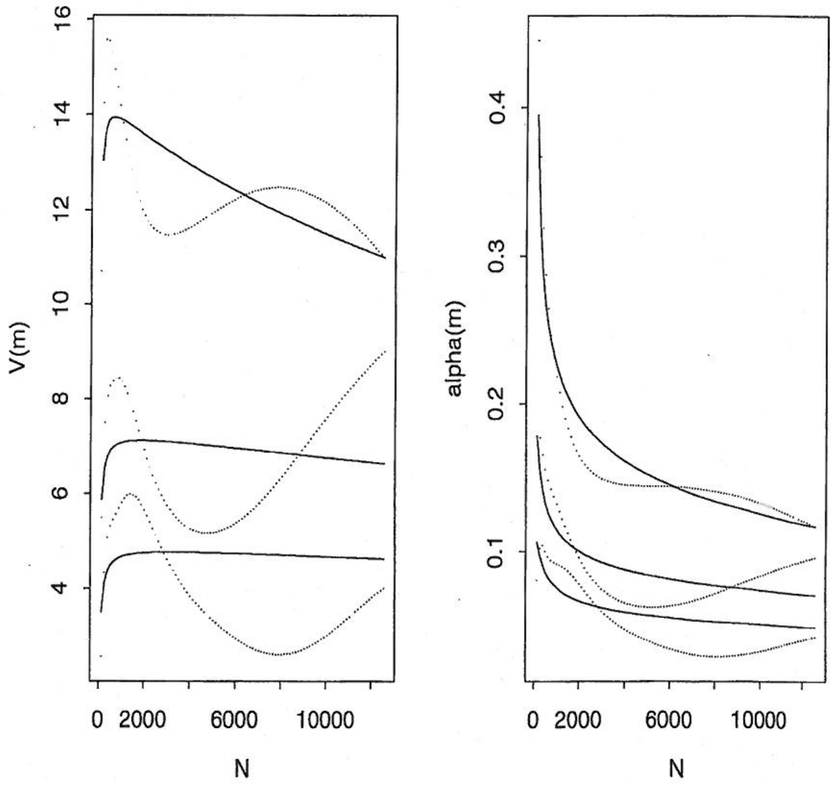


Figure 9. The development of the absolute and relative frequency spectra (m = 1, 2, 3) through sampling time: the unproductive English prefix *en-*. The continuous lines are obtained using the inverse Gauss-Poisson 'law'. The dotted lines are obtained by hypergeometric interpolation.

For not too small samples located beyond the LNRE ZONE the absolute and relative spectrum elements are (globally) decreasing functions of N. Note that the extreme case when

$$\overline{N}_1 = \overline{N}_2 = \ldots$$

can take place only if

$$\overline{N}_1 = \overline{N}_2 = \ldots = \infty,$$

in which case the relative expected spectrum terms are increasingly converging as the sample size $N \to \infty$. If some limiting 'law' $\alpha(m)$ is to hold for $N \to \infty$,

$$\alpha(m) = \lim_{N\to\infty} \alpha_N(m)$$

then the (stationarity) condition

$$\lim_{N\to\infty} \frac{d}{dN}\alpha_N(m) = \alpha(m)(m - \alpha(1)) - (m + 1)\alpha(m + 1), \tag{64}$$

should necessarily be satisfied. In other words, the growth rates of the spectrum elements should no longer vary with N. The unique solution to (64) is

$$\alpha(m) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}.$$

Thus Karlin-Rouault's 'law' appears as the only parametric family of limiting 'laws'. In section 4.2 we shall give a description of the theoretical structural distributions which can realize this law. First, however, we consider an alternative method for establishing whether a sample is located in the LNRE ZONE.

From a standard asymptotic point of view we may consider ourselves as situated outside the LNRE ZONE when, roughly speaking, we can convince ourselves that the empirical distribution $(\hat{p}_N(A_i))$ is so close to the theoretical distribution that we can allow ourselves to replace theoretical expectations by empirical ones. If the sample is located outside of the LNRE ZONE, the expected spectrum elements can be approximated by the expressions

$$\hat{E}\hat{V}_N(m) = \sum_i \frac{(\hat{p}_N(A_i)N)^m}{m!} e^{-\hat{p}_N(A_i)N}$$

$$= \int_0^\infty \frac{(\hat{p}_N N)^m}{m!} e^{-\hat{p}_N N} dG_N(\hat{p}_N)$$

$$\hat{E}\hat{V}_N = \sum (1 - e^{-\hat{p}_N(A_i)N})$$

$$= \int_0^\infty (1 - e^{-\hat{p}_N N}) dG_N(\hat{p}).$$

We can use the differences between the expected values

$$\begin{aligned} &\boldsymbol{E}\hat{V}_N(m) - \boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_M(m) \\ &\boldsymbol{E}\hat{V}_N - \boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N \end{aligned}$$

to evaluate the accuracy of the approximation and the extent of the bias introduced by estimating population probabilities by sample relative frequencies. Focussing on $\boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N$, the expected vocabulary at the sample size N if instead of the theoretical probabilities the empirical distribution $(\hat{p}_N(A_i), 1 \le i \le \hat{V})$is used to simulate the experiment on the same sample size, we find that

$$\begin{aligned} \boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N &= \boldsymbol{E}\sum_{m\ge1} (1 - e^{-m}) \hat{V}_N(m) \\ &= \sum_{m\ge1} (1 - e^{-m}) \boldsymbol{E}\hat{V}_N(m) \\ &= \boldsymbol{E}\hat{V}_{(1-e^{-1})N} . \end{aligned}$$

In other words, if the expected vocabulary at the smaller sample size $0.63N$ is approximately the same as for the sample size N, the sample is not located in the LNRE ZONE. This state of affairs obtains only when $\frac{d}{dM} V_{M|M=0.63\,N} \approx 0$.

A computationally convenient test is to consider the ratio

$$C_L = \frac{\hat{V}_N - \sum_m (1 - e^{-m})\hat{V}_N(m)}{\hat{V}_N} = \sum_m \hat{\alpha}_N(m) e^{-m}, \tag{65}$$

large values of which can be used to identify the LNRE zone. By way of example, suppose Zipf 's law is valid in the zone we are situated in. We then obtain

$$
\begin{aligned}
E\hat{E}V(N) &\approx \sum_{m\geq 1}(1 - e^{-m})E\left[V(N)\frac{1}{m(m + 1)}\right]\\
&= E[V(N)(e - 1)(1 - \ln(e - 1))]\\
&\sim 0.5EV(N),
\end{aligned}
$$

so that in this case we are loosing about half of the vocabulary on the assumption that we are not positioned in the LNRE ZONE. Some typical empirical examples are presented in table 1. The sample of words prefixed with *en-* appears with the lowest score for C_L. This accords well with our previous findings concerning the very early stage at which $V_N(1)$ achieves its maximum for *en-* (see figure 9). Although low values of C_L are typical for unproductive affixes, they are rarely observed for texts. However, we would not be surprised to find that the very large corpora that are at present being compiled ($N \gg 300{,}000{,}000$) will be located outside the central LNRE ZONE and probably outside the late LNRE ZONE as well.

Table 1. Some typical C_L values for various kinds of word frequency distributions

sample type	C_L	V	n_1
English *-ness (Cobuild)*	0.195	1607	749
English *en-*	0.059	94	11
Dutch *-heid*	0.228	466	256
Durch *-ing*	0.147	942	302
Carroll's *Alice in Wonderland*	0.163	1930	721
Bronte's *Wuthering Heights*	0.165	6420	2427
Pushkin's *Captain's Daughter*	0.213	4783	2384

3.2.2. Generalized Structural Distributions

We have already discussed the fact that Rouault's 'law' appears as the only limiting 'law' for $N \to \infty$. We now turn to consider the question whether theoretical structural distributions can be found that realize some 'law' for a finite, specific sample size Z. The unique solution for G as the (unknown) structural distribution appearing in the formula for the relative expected spectrum using the Poisson model,

$$
\alpha(m) = \frac{\displaystyle\int_0^\infty \frac{(pZ)^m}{m!}e^{-pZ}dG(p)}{\displaystyle\int_0^\infty (1 - e^{-pZ})dG(p)}, \qquad (66)
$$

given that one of the Zipfian laws (15-20) is substituted for $\alpha(m)$ in (66), takes the parameterized form

$$
G(p) = C\int_p^\infty e^{-Zpx}\,\frac{(\log(1 + x))^{\gamma - 1}x^{\alpha - 1}}{(1 + x)^{\beta + 1}}dx, \qquad (67)
$$

with some constant C (cf. Orlov and Chitashvili 1983b). In fact, if we substitute $G(p)$ in (67) into (66), the relative expected spectrum $\alpha(m)$ can be expressed as

$$
\alpha(m, \alpha, \beta, \gamma) = \frac{\displaystyle\int_0^\infty \frac{(\log(1 + x))^{\gamma - 1}x^{\alpha}}{(1 + x)^{m + \beta + 1}}dx}{\displaystyle\int_0^\infty \frac{(\log(1 + x)^{\gamma - 1}x^{\alpha - 1}}{(1 + x)^{\beta + 1}}dx}. \qquad (68)
$$

All laws (15-20) appear as special submodels for particular choices of the parameters α, β and γ (Zipf: $\alpha = \beta = \gamma = 1$, Yule: $\alpha = \beta$, $\gamma = 1$, Yule-Simon: $\alpha = 1$, $\gamma = 1$, Waring-Herdan: $\gamma = 1$, Karlin-Rouault: $\beta = 0$, $\gamma = 1$, Zipf-Mandelbrot: $\alpha = \beta = 1$). Unfortunately, expression (67) does not represent any real structural distribution because

1. $G(p)$ is not a step (or step-wise constant) function,
2. the distribution

$$
F(p) = \int_p^\infty x\,dG(x)
$$

may not be a normalized distribution, and

3. the theoretical vocabulary $V = G(0)$ may be infinite.

Nevertheless, the reasoning presented above at least makes it natural to admit generalized forms for structural distributions so long as they allow us to formulate expressions for the expected spectrum at prescribed sample sizes.

In addition to the Zipfian family (15-20) defined by (67), to which we shall refer as the generalized Zipf 's structural distribution, two other structural distributions, i.e. decreasing functions $G(p)$ of a general nature, should be mentioned. These distributions, the lognormal distribution (Herdan 1960, Carroll 1967, 1969)

$$G(p) = \frac{1}{\sigma\sqrt{2\pi}} \int_p^\infty \frac{1}{x^2} e^{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx \tag{69}$$

and the generalized inverse Gauss-Poisson distribution (Sichel 1976,1986)

$$G(p) = \frac{2^\gamma}{(bc)^{\gamma+1} K_{\gamma+1}(b)} \int_p^\infty x^{\gamma-1} e^{-\frac{x}{c}-\frac{b^2c}{4x}} dx, \tag{70}$$

where $K_\gamma(b)$ is the modified Bessel function of the second kind of order γ and argument b, allow the expected spectrum to be defined as

$$V_N(m) = \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} dG(p). \tag{71}$$

In both cases the structural distributions may be presented in the form

$$G(p) = \frac{Z}{\sigma\sqrt{2\pi}} \int_{pZ}^\infty \frac{1}{y^2} e^{-\frac{(\log y)^2}{2\sigma^2}} dy = ZG^*(pZ) \tag{72}$$

and

$$G(p) = \frac{2^\gamma}{cb^{\gamma+1} K_{\gamma+1}(b)} \int_{pZ}^\infty y^{\gamma-1} e^{-y-\frac{b^2}{4y}} dy = ZG^o(pZ) \tag{73}$$

with the parameters $Z = e^{-\mu}$ and $Z = 1/c$ playing the role of the sample-locator defining the sample's position with respect to the LNRE ZONE.

3.2.3. Simulating Generalized Laws

We now present an algorithm by which an experiment (in the framework of the multinomial model) could be simulated (approximately) corresponding to some generalized structural distribution. In other words, given the generalized probability type distribution $G^o(p)$ defined by the relation

$$\alpha(m) = \int_0^\infty \frac{p^m}{m!} e^{-p} dG^o(p), \tag{74}$$

for some relative spectrum 'law' $\alpha(m)$, we want to construct (for a sufficiently large N) the set of probabilities

$$(p_{i,N},\ 1 \le i \le V) \tag{75}$$

such that the corresponding structural distribution

$$G(p) = \sum_{i=1} \mathbb{I}_{[p_{i,N} > p]} \tag{76}$$

realizes on a sample of size N the relative expected spectrum

$$\alpha_N(m) = \frac{V_N(m)}{V_N} = \frac{\sum_{i=1}^{V} \frac{(p_{i,N}N)^m}{m!} e^{-p_{i,N}N}}{\sum_{i=1}^{V} (1 - e^{-p_{i,N}N})}$$

$$= \frac{\int_0^\infty \frac{p^m}{m!} e^{-p} dG(\frac{p}{N})}{\int_0^\infty (1 - e^{-p}) dG(\frac{p}{N})} \approx \int_0^\infty \frac{p^m}{m!} e^{-p} dG^o(p) = \alpha(m), \tag{77}$$

with G^0 the standardized correlate of G. To do this, construct for $\varepsilon > 0$ the sequence

$$(\lambda_i(\varepsilon),\ 1 \le i \le V)$$

from the relations

$$\begin{aligned} G^0(\lambda_1(\varepsilon)) &= \varepsilon \\ G^0(\lambda_i(\varepsilon)) &= G^0(\lambda_{i-1}(\varepsilon)) + \varepsilon,\ i \ge 2, \end{aligned} \tag{78}$$

where

$$V = V_\varepsilon = \min(k: \sum_{i=1}^{k}(1 - e^{-\lambda_i(\varepsilon)}) \geq \frac{1}{\varepsilon}). \tag{79}$$

Now define

$$n_\varepsilon = [\sum_{i=1}^{V}\lambda_i(\varepsilon)]$$

and construct the probabilities by the formula

$$p_{i,N} = \frac{\lambda_i(\varepsilon_N)}{n_{\varepsilon_N}}, \quad 1 \leq i \leq n_{\varepsilon_N} \tag{80}$$

with ε_N chosen so to satisfy

$$\varepsilon_N = \max(\varepsilon: n_\varepsilon \geq N). \tag{81}$$

3.3. Asymptotic Approach

Under what conditions can the use of generalized structural distributions be justified? This question is discussed in section 3.3.1. Section 3.3.2 considers the accuracy of the theoretical models, and section 3.3.3 calls attention to the independence of the high and low frequency 'tails' of LNRE distributions.

3.3.1. The Triangle Scheme of Experiment

We may justify generalized structural distributions (or generalized population probability distributions) by using the asymptotic approach argument. Although the LNRE ZONE is usually located at rather early stages of (imaginable) experiments, the samples in which the characteristic features of LNRE distributions are present are often large enough to apply the asymptotic analysis.

Within the framework of the classical scheme of experiment, the only way to justify generalized distributions is to admit the so-called triangle scheme of experiment, i.e. to consider the asymptotic scheme when (i) the normalized theoretical structural distribution

$$\frac{G^Z\left(\frac{p}{Z}\right)}{\int_0^\infty (1 - e^{-x})dG^Z\left(\frac{x}{Z}\right)};$$

$p = \lambda = 1/\pi$

indexed by some parameter Z, approaches some generalized structural distribution G^0, and when (ii) the sample size N is taken in the neighbourhood of Z (considered as the center of the LNRE ZONE), i.e. $N \approx Z$.

We assume that the token probabilities p_n are independent and identically distributed. Informally speaking, we find that at first sight there appears to be no distinction between the following suggestions:

A. The author selects some structural distribution $F(p)$ and then generates (creates) a text of some sufficiently large size N according to this distribution;

B. The author determines some sample size Z (the desired horizon) and chooses the structural distribution intending to get some (desired) frequency distribution 'law' on a sample of size Z and then generates a text of a size $N \asymp Z$.

But the distinction becomes obvious when we set the problem in asymptotic scheme. In fact, let some 'law' $\alpha(m)$, $m \geq 1$ be fixed. Now let the problem of the existence of a structural distribution realizing this law be stated in an asymptotic form, i.e., does a sequence $G^Z(p)$, $Z \gg 1$ of structural distributions exist such that the relative expected spectrum

$$\alpha_Z(m) \approx \alpha(m), \quad m \geq 1. \tag{82}$$

But this approximation takes place if and only if the normalized structural distribution is approximated (for $Z \gg 1$) by some (generalized and normalized) distribution $G^0(p)$,

$$G^0(p) \approx \frac{G^Z(\frac{p}{Z})}{\int_0^\infty (1 - e^{-x})dG^Z(\frac{x}{Z})}, \tag{83}$$

where $G^0(p)$ is uniquely determined from the equation

$$\alpha(m) = \int_0^\infty \frac{p^m}{m!}e^{-p}dG^0(p), \quad m = 1, 2, \ldots \tag{84}$$

In other words, $G^Z(p)$ with the property (82) for sufficiently large Z can be represented as

$$G^Z(p) \approx V_Z G^0(pZ), \tag{85}$$

where

$$V_Z = \int_0^\infty (1 - e^{-x}) dG^Z\left(\frac{x}{Z}\right)$$

is the expected vocabulary on the sample size Z and with the generalized structural distribution G^0 defined by (83). Thus to use the generalized structural distribution is equivalent to accept the hypothesis:

$$G(p) = G^Z(p)$$

with $G^Z(p)$ satisfying (82) for some 'law' $\alpha(m)$. Note that for the transition from a discrete step function for the structural distribution to a continuous function G^0 to be justified in the triangle scheme, the parameter Z, where $Z = e^{-\mu}$ for the lognormal model and $Z = 1/c$ for the generalized inverse Gauss-Poisson model, should assume a value not too different from N - the ratio $t = N/Z$ should not be too small or too large.

Thus for samples of size $N \gg 1$ we have two possibilities for the asymptotics of the relative expected spectrum. If G is fixed, that is, if we drop the index Z from G^Z, we again have Rouault's 'law'

$$\alpha(m) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}$$

as the only limiting distribution. If we allow G to be parameterized for Z such that (83) is satisfied, the triangle scheme leads to the following expression for $\alpha_N(m)$:

$$\alpha_N(m) \approx \frac{\int_0^\infty \frac{(pt)^m}{m!} e^{-pt} dG^0(p)}{\int_0^\infty (1 - e^{-pt}) dG^0(p)}, \tag{86}$$

a parametric family of 'laws' extended in sampling time and parametrized by Z, the 'Zipf' size, or equivalently by the parameter $t = N/Z$. For the generalized Zipf 's 'law' (68) the extended version takes the form

$$\alpha_N(m, \alpha, \beta, \gamma, t) = t^{m-1} \frac{\int_0^\infty \frac{[\ln(1 + y)]^{\gamma - 1} y^{\alpha}}{(t + y)^{m+1}(1 + y)^{\beta+1}} dy}{\int_0^\infty \frac{[\ln(1 + y)]^{\gamma - 1} y^{\alpha - 1}}{(t + y)(1 + y)^{\beta}} dy}. \tag{87}$$

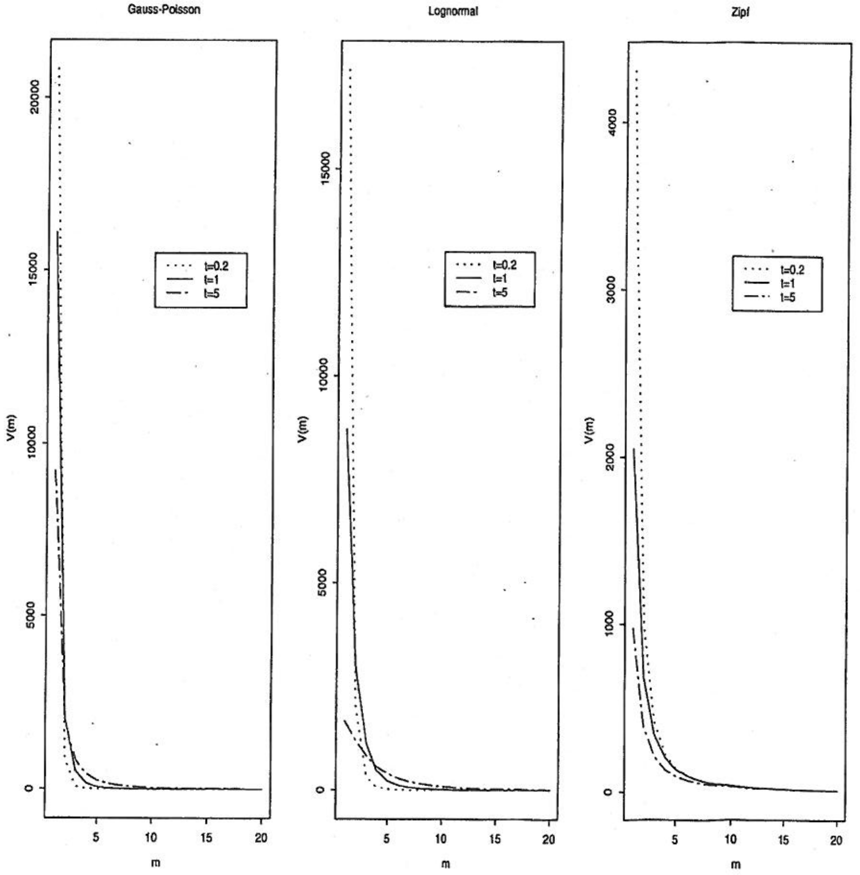


Figure 10. The role of the parameter t in the extended generalized Zipf 's 'law', the lognormal 'law' and the generalized inverse Gauss-Poisson 'law'.

For Sichel's Gauss-Poisson 'law', $\alpha_N(m)$ can be expressed as

$$\alpha_N(m,\gamma,b,t) = \frac{1}{(1-\frac{t}{t+1})^{-\gamma/2}K_\gamma\left(b\sqrt{(1+t)(1-\frac{t}{1+t})}\right) - K_\gamma(b\sqrt{1+t})}.$$

$$\frac{(0.5b\sqrt{1+t}\frac{t}{1+t})^m}{m!}K_{\gamma+m}(b\sqrt{1+t}).$$

For Carroll's lognormal model we finally have that

$$\alpha_N(m,\sigma,t) = \frac{\int_0^\infty \frac{1}{p^2}\frac{(pN)^m}{m!}e^{-pN-\frac{1}{2}\left(\frac{\log(pN/t)}{\sigma}\right)^2}dp}{\int_0^\infty \frac{1}{p^2}(1 - e^{-pN})e^{-\frac{1}{2}\left(\frac{\log(pN/t)}{\sigma}\right)^2}dp}. \tag{88}$$

Figure (10) illustrates the role of the parameter t for these three 'laws'. The Gauss-Poisson 'law' is shown for $\gamma = -0.5$ and $b = 0.01$, the lognormal 'law' for $\sigma = 1$ and the Waring-Herdan 'law' for $\alpha = \beta = 1$ (Zipf). For all models, increasing t leads to theoretical distributions in which the lowest frequency types play less prominent roles, as expected for samples moving away from the LNRE ZONE.

3.3.2. The Accuracy of Theoretical Models

In the LNRE ZONE the accuracy of theoretical models for frequency distributions can be treated in the gaussian framework. Formally, if

$$N \gg 1,\ V_N \gg 1,\ \frac{E\hat{V}_N(m)}{V_N} \asymp \alpha(m) > 0,$$

then

$$\left(\frac{\hat{V}_N(m) - E\hat{V}_N(m)}{\sqrt{V_N}},\ m \geq 1\right) \overset{D}{\sim} N(0,R), \tag{89}$$

by which we mean that the normalized spectrum is approximated by the gaussian vector with 0 mean and covariance matrix

$$R_{m,k} = \delta_{m,k}\alpha_N(m) - \binom{m+k}{k}\frac{1}{2^{m+k}}\alpha_{2N}(m+k). \tag{90}$$

With respect to the sequence ($\hat{V}_N$, $1 \leq n \leq N$)of observed values of the empirical vocabulary volumes through sampling time we have

$$\left(\frac{\hat{V}_n - V_n}{\sqrt{V_n}},\ 1 \leq n \leq N\right) \overset{D}{\sim} N(0,\overline{R}), \tag{91}$$

where n denotes the current sample size and where the covariance matrix

$$\overline{R}_{n,k} = \frac{COV(\hat{V}_n - V_n, \hat{V}_k - V_k)}{V_N},\quad 1 \leq n,\ k \leq N$$

can be given in the form

$$\overline{R}_{n,k} = V_{n+k} - V_{\max(n,k)},\quad 1 \leq n,\ k \leq N. \tag{92}$$

If we use the interpolation formula

$$\hat{V}_{N,n} = \sum_{j\geq 1}\hat{V}_N(j)\left(1 - (1 - \frac{n}{N})^j\right)$$

for the vocabulary growth curve to estimate the accuracy of the model, we may use the approximation

$$\left(\frac{\hat{V}_{n,N} - V_n}{\sqrt{V_N}},\ 1 \leq n \leq N\right) \overset{D}{\sim} N(0,\tilde{R}) \tag{93}$$

with covariance matrix

$$\tilde{R}_{n,k} = V_{n+k} - V_{n+k-\frac{nk}{N}},\quad 1 \leq n,\ k \leq N. \tag{94}$$

3.3.3. The Distribution of the Frequency Spectrum

Even though the lower elements of the frequency spectrum are the most important for LNRE samples, more global analyses of frequency distributions including the highest frequency terms are by no means devoid of interest. Speaking in terms of the structural type frequency distribution $\hat{G}_N(p)$, the theoretical models considered above were intended to give satisfactory approximations for the left hand tails, i.e.

$$G(\frac{p}{N}) \asymp E\hat{G}_N(\frac{p}{N}) .$$

To test some theoretical model for the structural distribution *G(p)* on the whole range of values $0 \leq p \leq 1$ it is useful to know that the differences

$$\Delta_N^+(p) = (\hat{G}_N(p) - E\hat{G}_N(p)) \tag{95}$$

$$\Delta_N^-(p) = (\hat{G}_N(\frac{p}{N}) - E\hat{G}_N(\frac{p}{N})) \tag{96}$$

are asymptotically gaussian with variances

$$\sigma^2\Delta_N^+(p) \sim \frac{1}{N} \tag{97}$$

$$\sigma^2\Delta_N^-(p) \sim EV_N , \tag{98}$$

and that, significantly, these differences are not correlated so that

$$COV(\sqrt{N}\Delta_N^+(p), \frac{1}{\sqrt{EV_N}}\Delta_N^-(p)) \approx \sqrt{\frac{EV_N}{N}} \int_0^\infty x(1 - \Pi^+(p,x))dG^0(x). \tag{99}$$

The important conclusion from this fact, which might be expected intuitively, is that mathematical models for tail and high frequency zones can be suggested independently. In particular, if some analytical expression ($\alpha(m)$, $m = 1, 2, ...$) is suggested for the relative expected spectrum, then we may write

$$E\hat{G}_N(p) = E\hat{V}_N(\sum_{m \geq pN} \alpha_N(m)) + \Delta_N(p), \tag{100}$$

where $\Delta_N(p)$ is intended to improve the fit for not small values of p, with the only property that

$$\frac{1}{E\hat{V}_N}\Delta_N(\frac{p}{N}) \rightarrow 0, N \rightarrow \infty. \tag{101}$$

Note that $\Delta_N(p)$ is exactly the parameter B in the models (21) and (22) discussed in section 2.2. This extra parameter can be used, in particular, to improve the fit with the theoretical rank probability distribution at the left hand tail (i.e. the high probability region) without affecting the low frequency zone.

4. LNRE Models and their Rationales

In this section we shall present and try to systematize different mathematical models intended as analytical tools for LNRE samples. Since the empirical frequency distribution is the main object for mathematical modelling, the practical output of any such mathematical model is to suggest some analytical expression such as Zipf 's law for the frequency distribution as described by the rank frequency distribution, the frequency spectrum, or the cumulative type or token frequency distributions. By the interpretation of the corresponding analytical expressions these models can be divided into three essentially different classes:

1. Models which consider the analytical expressions used to approximate the rank-frequency distribution as structural probability distributions of a general population. Typically, such models focus on developing stochastic schemes generating such populations.

2. Models which consider the analytical expressions used to approximate the frequency spectrum as limiting distributions that characterize the equilibrium state. Typically, these models focus on stochastic schemes leading to the desired steady (equilibrium) state.

3. Models which consider the analytical expressions used to approximate the frequency spectrum as expected values for finite samples. These models focus on general population models realizing these laws on finite samples.

4.1. Mandelbrot and Miller

Mandelbrot's rank-probability distribution

$$p\{r\} = \frac{C^{1/\gamma}}{(r + B)^{1/\gamma}} \tag{102}$$

or the corresponding structural type distribution

$$G(p) = \frac{C}{p^\gamma} + B$$

has proved to be a good enough approximation for a number of observed distributions $\hat{p}_N\{r\}$ or $\hat{G}_N(p)$. Hence, from the point of view of traditional probabilistic modelling, it seems natural to be interested in general populations with a rank-probability distribution of this form. Two approaches in this direction should be mentioned. Mandelbrot (1953,1962) has shown that the significance of the distribution *p{r}* can be explained by its optimality property of maximizing the information contained in a message constructed of words as sequences of letters indexed by different costs. Miller (1957) presented a pure probabilistic model where *p{r}* appears as a rank probability distribution of words viewed as sequences of letters chosen by chance at each stage of an experiment (as if a monkey were typing text), and where the parameters (*B, C, γ*) depend on the probability of a blank space and the number of letters.

But the significant bias between theoretical and empirical distributions symptomatic for LNRE renders such interpretations unconvincing, at least in the framework of the classical scheme of experiment. In fact, if the Zipf-Mandelbrot 'law' is taken as the theoretical probability distribution, then the relative frequency spectrum terms are converging to the expressions

$$\frac{\hat{V}_N(m)}{\hat{V}_N} = \hat{\alpha}_N(m) \rightarrow \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)} , \tag{103}$$

i.e. to the Karlin-Rouault 'law', instead of to the expression

$$\alpha(m) = \frac{1}{m^\gamma} - \frac{1}{(m + 1)^\gamma} ,$$

the Zipf-Mandelbrot 'law' in terms of the spectrum which might be expected.

4.2. Rouault

As mentioned above, the law (103) is the only limiting expression for the relative expected spectrum when sampling from a general population with a fixed structural probability distribution. It can be shown (Rouault 1978; Khmaladze and Chitashvili 1989) that the necessary and sufficient condition on the structural distribution *G(p)* when the limit (103) exists is the property of tails

$$G(p) = p^{-\alpha}\mathcal{L}(p) \tag{104}$$

with some $0 < \alpha < 1$, and some at $p = 0$ slowly increasing function $\mathcal{L}$, i.e.

$$\frac{\mathcal{L}(cp)}{\mathcal{L}(p)} \rightarrow 1, \; p \rightarrow 0$$

for each $c > 0$, as e.g. in the case that (102) takes place. In Khmaladze and Chitashvili (1989) it is shown that condition (104) is even necessary to have positive limits

$$\lim_{N\rightarrow\infty} \alpha_N(m) > 0, \quad r = 1,2,\ldots .$$

The aim of the mathematical models to be considered here is to suggest some natural stochastic scheme which provides the general population with property (104). The most complete is the markovian model of word generation considered by Rouault (1978), who generalized Miller's stochastic scheme. Let

$$\mathcal{L} = \{L_0, L_1, L_2, \ldots\}$$

be the set of elements (letters) including the blank space L_0 which occur according to some transition probability $p_{i,j}$. A particular word *A* can be viewed as a (finite) sequence of letters limited by two blanks:

$$A = [L_0 L_{i_1} L_{i_2}, \ldots, L_{i_k} L_0] .$$

Such a word has probability

$$p(A) = p_{0,i_1} p_{i_1,i_2} p_{i_2,i_3} \cdots p_{i_k,0} .$$

To form an idea of the structure of token probabilities p_n, $1 \leq n \leq N$ over a sample of size *N*, let x_t be the Markov chain realization of the procedure generating the running text as a sequence of letters. Let

$$\tau_1, \tau_2, \ldots, \tau_N$$

be the successive moments when blanks occur in this sample. We can present the sample of token words and the corresponding token probabilities as

$$\begin{aligned} w_1 &= (x_1, x_2, \ldots, x_{\tau_1}) \\ p(w_1) &= \prod_{t=1}^{\tau_1 - 1} p(x_t, x_{t+1}) \\ w_2 &= (x_{\tau_1}, x_{\tau_1+1}, \ldots, x_{\tau_2}) \\ p(w_2) &= \prod_{t=\tau_1}^{\tau_2 - 1} p(x_t, x_{t+1}). \end{aligned}$$

Rouault (1978) shows that the collection of such probabilities over all words (of any length) possesses (through its structural distribution) the property (104). Note that Miller's (1957) model is a special case of Rouault's scheme with p_{0i} and p_{ij} not depending on i, j.

Thus, if the dynamics of a population are governed by the simplest multiplication rule, according to which a particular element enters the population or is re-used with a constant probability not depending on the current state of the system, then (103) expresses the only possible equilibrium state distribution. However, the class of equilibrium distributions can be essentially enlarged if more general birth and death stochastic schemes of population dynamics are considered.

4.3. Dynamic Models: Simon, Waring-Herdan

We have seen that within the classical scheme of independent and identically distributed observations the majority of Zipfian frequency distribution laws, with as the only exception Rouault's 'law', can be considered as analytical expressions for the relative expected spectrum on finite sample sizes in the asymptotic setting of the triangle scheme. In other words, we are dealing with 'laws' that do not have the property of stability with respect to changing sample sizes. In order to justify these 'laws' as laws expressing the equilibrium (or steady state) property of well organized systems, we therefore need more general stochastic processes as models for the formation of populations with large numbers of different elements. In this section we consider a number of such processes which can be viewed as providing rationales for a number of 'laws' of the Zipfian family.

The dynamic modelling idea becomes very natural if we look on the frequency

dynamics in the classical scheme. In fact, the frequency of some word A_i can be calculated recursively

$$f_n(A_i) = f_{n-1}(A_i) + \varepsilon^i_N, \quad n = 1,2,\ldots\ , \tag{105}$$

with the increments ε^i_n taking only two values (0, 1) independently of the frequency structure

$$\mathscr{F}_n = (f_n(A_i),\ 1 \le i)$$

of the sample at the current stage n. In other words, the conditional probability coincides with the unconditional one, and is constant:

$$\boldsymbol{Pr}(\varepsilon^i_n = 1 | \mathscr{F}_n) = \boldsymbol{Pr}(\varepsilon^i_n = 1) = \boldsymbol{Pr}(A_i). \tag{106}$$

The specifics of the kind of dynamic modelling considered here lies in the assumption that this probability may depend on the current state of the 'system', and in particular on the frequency of the element A_i:

$$\boldsymbol{Pr}(\varepsilon^i_n = 1 | \mathscr{F}_n) = \boldsymbol{Pr}(A_i | \mathscr{F}_n). \tag{107}$$

One of the versions of Simon's (1955, 1960) models can be presented as the simplest (but nevertheless very natural) example. The conditional probability of the 'birth' of the element A_i is defined as

$$\boldsymbol{Pr}(A_i | \mathscr{F}_n) = q \mathbb{I}_{[f_n(A_i) = 0]} \frac{p(A_i)}{\sum_j \mathbb{I}_{[f_n(A_j) = 0]} p(A_j)} + (1 - q) \mathbb{I}_{[f_n(A_i) > 0]} \frac{f_n(A_i)}{n}\ , \tag{108}$$

where $0 \le q \le 1$ stands for the probability that some new element from the vocabulary with the frequency $f_n(A_i) = 0$ can be included into population, and where with probability $1 - q$ some already present element can be re-used proportional to its frequency in the sample. From the simple relation

$$\mathbb{I}_{[f_{n+1}(A_i) = m]} = \mathbb{I}_{[f_n(A_i) = m]}(1 - \mathbb{I}_{[\varepsilon^i_i = 1]}) + \mathbb{I}_{[f_n(A_i) = m-1]} \mathbb{I}_{[\varepsilon^i_n = 1]} \tag{109}$$

the recursive relations for the spectrum terms can be derived,

$$\hat{V}_{n+1}(m) = \hat{V}_n(m) + \hat{V}_n(m-1)\frac{(1-q)(m-1)}{n} - \hat{V}_n(m)\frac{(1-q)m}{n} + \varepsilon_n(m), \quad (110)$$

where the additive term $\varepsilon_n(m)$ is conditionally centered

$$E(\varepsilon_n(m)|\mathscr{F}_n) = 0. \quad (111)$$

For the expected spectrum we have the recursive equations

$$V_{n+1}(m) = V_n(m) + V_n(m-1)\frac{(1-q)(m-1)}{n} - V_n(m)\frac{(1-q)m}{n} \quad (112)$$

$$V_{n+1}(1) = V_n(1) + 1 - V_n(1)\frac{1-q}{n} \quad (113)$$

$$V_{n+1} = V_n + q. \quad (114)$$

Now recursive equations for the relative expected spectrum can be obtained

$$\alpha_{n+1}(m) = \alpha_n(m) + \frac{1}{n}[(1-q)(m-1)\alpha_n(m-1) - ((1-q)m + 1)\alpha_n(m)] \quad (115)$$

which tells us that the limiting law

$$\alpha_n(m) \to \alpha(m),\ n \to \infty \quad (116)$$

- in fact it can be shown that this convergence and moreover the law of large numbers

$$\frac{\hat{V}_n(m)}{\hat{V}_n} \to \alpha(m),\ n \to \infty \quad (117)$$

takes place here - should be the solution of the equilibrium or steady state equation

$$(1 - q)(m - 1)\alpha(m - 1) = ((1 - q)m + 1)\alpha(m). \quad (118)$$

The solution to (118),

$$\alpha(m) = \frac{\Gamma(1 + \frac{1}{1-q})}{1 - q} \cdot \frac{\Gamma(m)}{\Gamma(m + \frac{1}{1-q} + 1)}, \quad (119)$$

is a particular instantiation of Simon's (1960) model.

For a slightly more general model where an element with frequency m, $m > 0$ can be re-used proportionally to its relative frequency but in which the probability that a new word occurs on the n-th stage is proportional to the total probability of unused words, we have

$$Pr(A_i|\mathscr{F}_n) = \frac{rp(A_i)\mathbb{I}_{[f_n(A_i) = 0]} + \frac{f_n(A_i)}{n}\mathbb{I}_{[f_n(A_i) > 0]}}{r\sum_j p(A_j)\mathbb{I}_{[f_n(A_j) = 0]} + 1} \quad (120)$$

Note that the probability of generating new words will now eventually decrease to 0. If the structural distribution

$$G(p) = \sum_{i\geq 1} \mathbb{I}_{[p(A_i) \geq p]}$$

of the general population satisfies the condition (104), then the steady state law is

$$\alpha(m) = \frac{\Gamma(m)\Gamma(1 + \alpha r)}{\Gamma(m + 1 + \alpha r)}, \quad (121)$$

the beta function of Yule (1924).

Both the Yule distribution (119) and what we have called the Yule-Simon 'law'

$$\alpha(m) = \frac{\beta}{(m + \beta - 1)(m + \beta)},$$

which we have found to be the more useful expression for the analysis of texts, are special cases of the 'law' advanced by Simon (1960:69) on the basis of a birth and death process model for the population dynamics,

$$\alpha(m) = A\lambda^m B(m + c, d - c + 1),$$

with $B(.,.)$ the Beta-function and with parameters c, d, λ defining the birth and death probabilities and with normalizing constant A. For specially chosen parameters and λ fixed at unity both models can be derived. Interestingly, the Waring-Herdan-Muller model, which includes the two above versions of Simon's model, can be obtained along similar lines when the probability of re-using some word is a linear function of the frequency of that word (see Khmaladze and Chitashvili, 1989).

These dynamic (birth and death) equilibrium models are satisfactory for the LNRE analysis of biological (distribution of species), social (distribution of incomes), psychological (distribution of responses to some stimulus), and a number of other living and technical systems. Unfortunately, lexical samples generally do not show the tendency to equilibrium state, and though the LNRE features may be displaid clearly, the frequency distributions change considerably with changing sample size.[1] This makes it natural to view such samples as being located in the LNRE zone and to apply models for general populations realizing appropriate frequency distributions on finite samples. In the following sections, we discuss three such models.

4.4. The Lognormal Model (Herdan, Carroll)

In this model a word token is chosen on each stage of the experiment as a result of some recursive procedure. Unlike the Markovian scheme underlying Rouault's 'law', this procedure is interpreted as a choice between words rather than as a word generation procedure. The token probabilities are now expressed as products of transition probabilities which are themselves random.

In the lognormal model as considered by Carroll (1967, 1969) the structure of the vocabulary is described in terms of a decision tree. (More general schemes can be considered than that proposed by Carroll, here we will limit ourselves to Carroll's approach.) Assume that we have a binary decision tree where the paths leading to the leaves of the tree, the elements of the vocabulary, may have different lengths. Let $\vec{y} = y_s, (s = 0, 1, 2, ...)$ denote some path from the root of the tree to some leaf, i.e. the sequence of decisions made at the different levels of the tree, with $y_s \; \varepsilon \; \{0, 1\}$ indicating the possible decisions on each stage. For each path $\vec{y}$ we define some stopping moment $\tau(\vec{y})$ indicating the length of the path. Each path uniquely determines some word $A(y)$ as a result of the decision procedure.

To define probabilities of words let $(\vec{\pi} = \pi_s, s = 0,1,2,...)$ be decision probabilities corresponding to the stages $s = 0, 1, ...$. Also assume that these probabilities are randomly distributed according to some distribution function $\phi(\pi)$, $(0 \leq \pi \leq 1)$ on the interval [0, 1]. Suppose finally that y_s as well as π_s are independent. Then the probability of word $w = w(\vec{y})$, given the probabilities $\vec{\pi}$ equals

[1] For a dynamic model which combines the Mandelbrot/Miller/Rouault approach with that of Simon without imposing the equilibrium constraint, the reader is referred to Baayen (1991), a simulation study that focusses on the similarity relations between words in the lexicon.

$$p(w) = \prod_{s=0}^{\tau_{\vec{y}}} (\pi_s)^{y_s} (1 - \pi_s)^{1 - y_s}. \tag{122}$$

The token probability distribution can now be expressed as

$$F(p) = \boldsymbol{Pr}(p(w) \geq p) = \boldsymbol{Pr}\left(\sum_{s=0}^{\tau_{\vec{y}}} [y_s \log(\pi_s) + (1 - y_s)\log(1 - \pi_s)] \geq \log p\right)$$

$$= \boldsymbol{Pr}\left(\sum_{s=0}^{\tau_{\vec{y}}} [\hat{\theta}(s) \geq \log p\right). \tag{123}$$

The mean and the variance of the random variable $\hat{\theta}$ are easily calculated:

$$\theta = \boldsymbol{E}\hat{\theta} = E[\pi \log \pi + (1 - \pi)\log(1 - \pi)]$$

$$= \int_0^1 [x \log x + (1 - x)\log(1 - x)] d\Phi(x) \tag{124}$$

$$\sigma_\theta^2 = \boldsymbol{VAR}\,\hat{\theta} = \int_0^1 [x(\log x)^2 + (1 - x)(\log(1 - x))^2] d\Phi(x) - \theta^2 \tag{125}$$

We need some conditions on the stopping moment τ and on the variance of the distribution Φ. Suppose that (i) τ is a Markov moment. This is a conventional assumption in the theory of stochastic processes when sums of random numbers of random variables are investigated. In the present case, the assumption that τ is a Markov moment implies that for any path $\vec{y}$ the event $[\tau = k]$, given the path $(y_s, 0 \leq s \leq k)$, does not depend on the future values $(y_s, k + 1 \leq s \leq N)$. This condition is met when, for instance, τ does not depend on the path $\vec{y}$ at all, as is assumed in Carroll (1969). Also suppose that (ii) the expected path-length, the time needed to come to one of the leaves of the decision tree, is sufficiently large ($E\tau \gg 1$) and that the variance is relatively small, as in the case of τ having a Poisson distribution:

$$\frac{\boldsymbol{VAR}(\tau)}{E\tau^2} \ll 1.$$

Finally, suppose that

$$\boldsymbol{VAR}\,\hat{\theta} \rightarrow \frac{\boldsymbol{VAR}(\tau)}{\boldsymbol{E}\tau}.$$

Under these assumptions the token probability distribution indexed by the parameter Z can asymptotically be expressed as

$$F(p) = F^Z(p) \rightarrow \boldsymbol{Pr}(\sigma N^Z - \log(Z)) \geq \log(p)$$

where we introduce the notation $Z = e^{-\theta E\tau}$, and where N^Z is an asymptotically standard gaussian variable. Equivalently,

$$\begin{aligned} F(p) &= F^Z(p) \\ &= \frac{1}{\sqrt{2\pi}\,\sigma}\int_p^\infty \frac{1}{x} e^{-\frac{(\log x - \log Z)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\frac{p}{Z}}^\infty \frac{1}{x} e^{-\frac{(\log x)^2}{2\sigma^2}} dx. \end{aligned} \tag{126}$$

Thus the lognormal model is justified asymptotically in the framework of the triangle scheme discussed above.

4.5. The Generalized Inverse Gauss-Poisson 'Law'

The motivation for the generalized inverse Gaussian-Poisson distribution (Sichel, 1986), which is presented by the structural distribution

$$G(p) = \frac{(2/bc)^\gamma}{2K_\gamma(b)}\int_\lambda^\infty x^{\gamma-1}\exp\left(-\frac{x}{c} - \frac{\sigma^2 c}{4x}\right)dx, \tag{127}$$

seems formal, though as special cases it includes e.g. the Γ-distribution ($b = 0$) and the distribution of an inverse of a Gaussian random variable ($c \rightarrow \infty$, $\sigma^2 \rightarrow 0$, $\sigma^2 c$ = const).

4.6. The Generalized Zipf's 'Law'

The basis for the rationale of the generalized Zipf 's model, presented by the structural distribution

$$G(p) = C\int_0^\infty e^{-Zpx}\frac{(\log(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx,$$

is clear enough: this is the unique parametric family of structural distributions which can realize on a finite sample of a particular size Z (Z being one of its parameters) a desired representative of the Zipfian family of 'laws' in terms of the relative expected spectrum.

5. Statistical Analysis with LNRE Models

In this section the information needed for the application of parametric LNRE models to statistical data analysis is presented. In section 5.1 we discuss the expressions for the various theoretical characteristics in which we are interested. We present some expressions for covariances in section 5.2. Section 5.3 outlines a number of ways in which the parameters of theoretical models may be estimated. Section 5.4 briefly discusses how to estimate confidence regions for estimated parameters. Goodness-of-fit tests for theoretical models are given in section 5.5. Section 5.6 contains some suggestions how to compare LNRE samples. Finally, the software known to us for the modelling of LNRE distributions is discussed in section 5.7 and applied to a number of empirical distributions.

To make these sections as independent of the other parts as possible and thus more convenient for application, we give some expressions in detail even though they can be found in previous sections. For ease of presentation, we will phrase the discussion in terms of a general three-parameter model with the structural distributions

$$\begin{aligned} G(p) &= G(p;\ \alpha,\ \beta,\ \gamma) \\ F(p) &= F(p;\ \alpha,\ \beta,\ \gamma) \\ \phi(p) &= \phi(p;\ \alpha,\ \beta,\ \gamma), \end{aligned} \tag{128}$$

where $\phi(p)$ is the density function of the token probability distribution $F(p)$,

$$\phi(p) = \frac{d}{dp}F(p).$$

5.1. Expressions for the Theoretical Spectrum

5.1.1. Nonparametric Expressions

General expressions for the expected frequency spectrum and the expected empirical vocabulary for three parameter models can be presented in integral form:

$$\begin{aligned} V_N(m) &= V_N(m;\ \alpha,\ \beta,\ \gamma) = E\hat{V}_N(m) \\ &= \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} dG(p) \\ &= \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} \frac{1}{p} dF(p) \\ &= = \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} \frac{1}{p} \phi(p) dp. \end{aligned} \tag{129}$$

for the expected frequency spectrum, and

$$\begin{aligned} V_N(0) &= V_N(0;\ \alpha,\ \beta,\ \gamma) = E\hat{V}_N \\ &= \int_0^\infty (1 - e^{-pN}) dG(p) \\ &= \int_0^\infty (1 - e^{-pN}) \frac{1}{p} dF(p) \\ &= \int_0^\infty (1 - e^{-pN}) \frac{1}{p} \phi(p) dp. \end{aligned} \tag{130}$$

for the expected empirical vocabulary. Note that for notational convenience the expected empirical vocabulary is denoted by $V_N(0)$. Thus the vector $(V_N^{(m)}$, $m = 0, 1, 2, \ldots, M)$ denotes the first M elements of the theoretical frequency spectrum and the expected empirical vocabulary jointly. The same holds for its empirical analogue, $(\hat{V}_N^{(m)}$, $m = 0, 1, 2, \ldots, M)$.

5.1.2. Parametric Expressions

We now present explicit expressions for the three parametric families of structural distribution models, the lognormal model, the inverse generalized Gauss-Poisson model, and the generalized Zipf model.

The lognormal model. Carroll's (1967, 1969) lognormal model is defined by the structural token probability distribution

$$F(p) = \frac{1}{\sigma\sqrt{2\pi}} \int_p^\infty \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx. \tag{131}$$

The expected spectrum and vocabulary for the sample size N can be expressed in integral form:

$$\begin{aligned} V_N(m) &= E\hat{V}_N(m) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{(xN)^m}{x^2 m!} e^{-xN - \frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx \end{aligned} \tag{132}$$

$$V_N = E\hat{V}_N \tag{133}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty (1 - e^{-xN}) \frac{1}{x^2} e^{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx. \tag{134}$$

The theoretical vocabulary (number of types) is obtained by considering V_N in the limit for $N \to \infty$:

$$V = \lim_{N \to \infty} V_N = e^{\frac{\sigma^2}{2} - \mu}.$$

Note that the parameter μ, the mean value of $log\ p$ in the general population, typically is a negative number < -1.

The generalized inverse Gauss-Poisson model. Sichel's (1975, 1986) generalized inverse Gauss-Poisson 'law' is based on the structural type distribution

$$G(p) = \cancel{V} \frac{(2/bc)^\gamma}{2K_\gamma(b)} \int_p^\infty x^{\gamma - 1} e^{\left(-\frac{x}{c} - \frac{b^2 c}{4x}\right)} dx\ , \tag{135}$$

where $K_\gamma(b)$ is the modified Bessel function of the second kind of order γ and argument b. The theoretical vocabulary V, the number of types in the population, can be determined on the basis of the normalizing argument. In fact, since

$$\int_0^\infty dF(p) = \int_0^\infty p\,dG(p) = 1,$$

we can easily find the expression for V:

$$V = \frac{2}{bc}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}. \tag{136}$$

For the expected vocabulary and the (relative) expected spectrum for arbitrary sample size N explicit formulas in terms of the Bessel function can be found:

$$\begin{aligned} V_N(0) &= \frac{2}{bc}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}\left[1 - \frac{K_\gamma(b\sqrt{1+cN})}{(1+cN)^{\gamma/2}K_\gamma(b)}\right]. \\ V_N(m) &= \frac{V_N(0)}{(1-\theta_N)^{-\gamma/2}K_\gamma(\alpha_N(1-\theta_N)^{1/2}) - K_\gamma(\alpha_N)}\frac{(0.5\alpha_N\theta_N)^m}{m!}K_{\gamma+m}(\alpha_N), \quad (137) \\ \alpha_N(m) &= \frac{1}{(1-\theta_N)^{-\gamma/2}K_\gamma(\alpha_N(1-\theta_N)^{1/2}) - K_\gamma(\alpha_N)}\frac{(0.5\alpha_N\theta_N)^m}{m!}K_{\gamma+m}(\alpha_N), \end{aligned}$$

where the parameters $\alpha_N = b(1 + cN)^{1/2}$ and $\theta_N = cN/(1 + cN)$ are introduced for notational simplicity. Note that the parameters α_N and θ_N are functions of the sample size N, while the parameters b, c and γ are population invariants.

The extended generalized Zipf's 'law'. Orlov and Chitashvili (1982a,b, 1983 a,b) develop a model that is a generalization of Zipf's law. For this model the structural probability type distribution

$$G(p) = C\int_p^\infty e^{-Zpx}\frac{(\ln(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx,$$

[illegible] p.224: (19)

where C is a normalizing coefficient (defined below), is characterized by the property that it realizes on the sample size Z the relative expected spectrum

$$\alpha_Z(m) = \frac{\displaystyle\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha}}{(1+y)^{m+1}(1+y)^{\beta}}dy}{\displaystyle\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(1+y)^{\beta+1}}dy}.$$

A number of known 'laws' are included as special cases. The following expressions for the expected vocabulary V_N and frequency spectrum terms $V_N(m)$ can be obtained:

$$\begin{aligned} V_N(m) &= E\hat{V}_N(m) & (138) \\ &= C(Z, \alpha, \beta, \gamma)t^m\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha}}{(t+y)^{m+1}(1+y)^{\beta+1}}dy & (139) \\ V_N &= E\hat{V}_N & (140) \\ &= C(Z, \alpha, \beta, \gamma)t\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(t+y)(1+y)^{\beta}}dy & (141) \end{aligned}$$

where $t = N/Z$ and where the coefficient C is defined by

$$C(Z, \alpha, \beta, \gamma) = \frac{V_Z}{\displaystyle\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(1+y)^{\beta+1}}dy}. \tag{142}$$

The expected number of types V_Z for the sample size Z can be determined by the normalizing argument, namely from the relation

$$N = \sum_{m=1}^{N\hat{p}_N\{1\}} m\hat{V}_N(m),$$

where $\hat{p}_N\{1\}$ is the maximal observed relative frequency. Application of this relation to the expected frequency spectrum for $Z = N$ leads to the following expression

$$V_Z = Z \frac{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-1}}{(1+y)^{\beta+1}} dy}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-2}}{(1+y)^{\beta+Z\hat{p}_N\{1\}}} [(1+y)^{Z\hat{p}_N\{1\}} - 1 - \frac{Z\hat{p}_N\{1\} y}{1+y}] dy} . \quad (143)$$

For the important case of the extended Waring-Herdan-Muller law ($\gamma = 1$), all formulas are significantly simplified:

$$V_N(m) = C(Z, \alpha, \beta) t^m \int_0^\infty \frac{y^\alpha}{(t+y)^{m+1}(1+y)^{\beta+1}} dy \quad (144)$$

$$V_N = C(Z, \alpha, \beta) t \int_0^\infty \frac{y^{\alpha-1}}{(t+y)(1+y)^\beta} dy \quad (145)$$

with

$$C(Z, \alpha, \beta) = \frac{V_Z}{\int_0^\infty \frac{y^{\alpha-1}}{(1+y)^{\beta+1}} dy}.$$

The theoretical vocabulary V is finite if $\beta > \alpha$ and can be expressed as

$$V = \frac{V_Z \beta}{\beta - \alpha}.$$

If, furthermore, $\alpha = 1$ (the extended Yule-Simon law), then the expression for V_Z can be approximated by

$$V_Z \approx \frac{Z}{\beta \ln(Z\hat{p}_N\{1\})} . \quad (146)$$

5.2. Expressions for Covariances

The covariances between the terms $(\hat{V}_N, \hat{V}_N(m), m = 1, 2, ...)$, i.e. the autocovariances and crosscovariances for varying sample sizes, can be presented in terms of the expected values of the spectrum terms, as shown in sections 3.1 and 3.3.2. When convenient, they can be stated in integral form, applying the parametric and non-parametric representations discussed above.

Given the vector statistic $(\hat{V}_N^{(m)} = 0, 1, 2, ..., M)$ and the expressions for $V_N(m)$, $m = 0, 1, 2, ...$, the corresponding covariance matrix $R_{m,k}(N, M)$ is easily calculated:

$$R_{m,k}(N, M) = (\mathbf{COV}(\hat{V}_N(m), \hat{V}_N(k)))_{k,m = 0,1,\ldots,M} =$$

$$= \begin{cases} \delta_{m,k} V_N(m) - \binom{m+k}{m} \frac{1}{2^{m+k}} V_{2N}(m+k) & \textit{for} \quad m, k = 1,2,...,M. \\ -\frac{1}{2^m} V_{2N}(m) & \textit{for } m = 0, k = 1,2,...,M. \\ V_N - V_{2N} & \textit{for } m = 0, k = 0. \end{cases} \quad (147)$$

For two different samples with size N and n, $N \leq n$ we have

$$\mathbf{COV}(\hat{V}_n(m), \hat{V}_N(k)) = V_n(m) \binom{m}{k} \left(\frac{N}{n}\right)^k \left(1 - \frac{N}{n}\right)^{m-k} -$$
$$V_{N+n}(m+k) \binom{m+k}{m} \left(\frac{N}{N+n}\right)^k \left(1 - \frac{N}{N+n}\right)^m$$

$$\mathbf{COV}(\hat{V}_n, \hat{V}_N(k)) = \left(\frac{N}{n+N}\right)^k V_{n+N}(k)$$

$$\mathbf{COV}(\hat{V}_n(m), \hat{V}_N) = \left(\frac{n}{n+N}\right)^m V_{n+N}(m) - \left(1 - \frac{N}{n}\right)^m V_n(m)$$

$$\mathbf{COV}(\hat{V}_n, \hat{V}_N) = V_{n+N} - V_{\min(N,n)}.$$

In section 3.1.4 we considered the interpolation problem. The results obtained there can be generalized, so that for arbitrary n, N the recursive relations

$$V_n(m) = \sum_{j \geq m} V_N(j) \binom{j}{m} \left(\frac{n}{N}\right)^m \left(1 - \frac{n}{N}\right)^{j-m} \quad (148)$$

$$V_n = \sum_{j \geq 1} V_N(j) \left(1 - \left(1 - \frac{n}{N}\right)^j\right) \quad (149)$$

between the expected spectrum terms can be defined (Good and Toulmin 1956; Kalinin 1965). The autocovariance of $\hat{V}_{N,n}$ equals

$$\boldsymbol{COV}(\hat{V}_{N,n}, \hat{V}_{N,k}) = V_{n+k} - V_{n+k-\frac{nk}{N}},\ 1 \leq n,\ k \leq N. \tag{150}$$

The mean square deviation of $\hat{V}_{N,n}$ from the "true" value of the vocabulary on the sample of size n (the interpolation accuracy) can be presented as

$$\boldsymbol{E}(\hat{V}_{N,n} - \hat{V}_{n})^2 = V_{2n-\frac{n^2}{N}} - V_{N}\ . \tag{151}$$

Expected spectrum elements for sample sizes $N' > N$ are often required in the formulas for variances and covariances. Unfortunately, the nonparametric expressions (148) and (149) become unstable for $n > 2N$ (see e.g. Good and Toulmin 1956), even though for instance (149) still possesses some optimality property: it gives the best linear extrapolation whereas the optimal extrapolation formula

$$\hat{V}_{N,n} = \boldsymbol{E}(\hat{V}_n | \hat{V}_N(k),\ k \geq 1)$$

is strictly nonlinear for $n \geq N$ and rather complicated for an exact calculation. Perhaps the best way to proceed is to use the simple extrapolation formulas based on some parametric model and to substitute the estimated parameters $(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)$ for their theoretical counterparts (α, β, γ). In the case of extrapolated vocabulary sizes,

$$\hat{V}_{n,N} = V_n(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)\ ,$$

the accuracy of the predicted values can be gauged by considering

$$D_{n,N} = \boldsymbol{E}(\hat{V}_n - \hat{V}_{n,N})^2\ ,$$

where $\hat{V}_{n,N}$ can be approximated for sufficiently large N by

$$\hat{V}_{n,N} \approx V_n + (\hat{\alpha}_N - \alpha)\dot{V}_n^1 + (\hat{\beta}_N - \beta)\dot{V}_n^2 + (\hat{\gamma}_N - \gamma)\dot{V}_{N}^3, \tag{152}$$

where

$$V_n = V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^1 = \frac{\partial}{\partial\alpha}V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^2 = \frac{\partial}{\partial\beta}V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^3 = \frac{\partial}{\partial\gamma}V_n(\alpha, \beta, \gamma).$$

Note that to use this accuracy expression, we must again replace the parameters (α, β, γ) in the right hand side of $D_{n,N}$ by their estimators $(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)$.

5.3. Parameter Estimation

Several procedures can be suggested for estimating the parameters of a word frequency 'law'.

5.3.1. Method 1

The simplest way is to require that the first (three) 'most remarkable' terms of the frequency spectrum, that is, the vector $(\hat{V}_N(m),\ m = 0, 1, 2)$, should coincide with their expected values:

$$\left\{\begin{aligned} \hat{V}_N(0) &= V_N(0;\ \alpha, \beta, \gamma) \\ \hat{V}_N(1) &= V_N(1;\ \alpha, \beta, \gamma) \\ \hat{V}_N(2) &= V_N(2;\ \alpha, \beta, \gamma) \end{aligned}\right\} \Rightarrow (\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N), \tag{153}$$

where we denoted the resulting parameter estimators by

$$(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N).$$

Note that the number of equations equals the number of parameters.

5.3.2. Method 2

A more global, though rather complicated algorithm can be used which takes more terms of the spectrum into consideration. We fix some number $M \geq 3$ of terms of the vector

$$(\hat{V}_N^{(m)},\ m = 0,1,2,\dots,M)$$

and construct the chi-square statistic

$$\chi^2_{(M-3)} = \sum_{0\le m,k\le M} (\hat{V}_N(m) - V_N(m))R^{-1}_{m,k}(N,\ M)(\hat{V}_N(k) - V_N(k)) \qquad (154)$$

where $R^{-1}_{m,k}(N,M)$, $0 \le m,\ k \le M$ is the inverse of $R_{m,k}(N,M)$. We then search for the estimators

$$(\alpha^*_N,\ \beta^*_N,\ \gamma^*_N)$$

for which $\chi^2_{(M-3)}$ is minimal.

5.3.3. Method 3

A method that we have found to be especially useful is to fix one parameter, say γ, and to choose the other two parameters such that

$$\begin{Bmatrix} \hat{V}_N(0) = V_N(0;\ \alpha,\ \beta,\ \gamma) \\ \hat{V}_N(1) = V_N(1;\ \alpha,\ \beta,\ \gamma) \end{Bmatrix} \qquad (155)$$

is satisfied. Following this, γ is varied (and α and β adjusted to satisfy (155) such that the value of $\chi^2_{(M-3)}$ is minimal.

5.3.4. Method 4

As a modification of method 2, the estimator

$$(\alpha^{**}_N,\ \beta^{**}_N,\ \gamma^{**}_N)$$

can be constructed so as to minimize the chi-square statistic for the differences between the (nonparametric) interpolated vocabulary growth curve

$$\hat{V}_{N,n} = \sum_{m\ge1} \hat{V}_N(m)\left(1 - (1 - \frac{n}{N})^m\right)$$

and its expectation with respect to the parametric model. Thus the estimators

$$(\alpha^{**}_N,\ \beta^{**}_N,\ \gamma^{**}_N)$$

are chosen such that

$$\tilde{\chi}^2_{(M-3)} = \sum_{1\le i,j\le M} (\hat{V}_{n_i,N} - V_{n_i})\tilde{R}^{-1}_{n_i,n_j}(N,M)(\hat{V}_{n_j,N} - V_{n_j}) \qquad (156)$$

is minimal, where $\tilde{R}^{-1}_{n_i n_j}(N,\ M)$, $1 \le n_i,\ n_j \le N$ is the inverse of the covariance matrix

$$\tilde{R}_{n_i n_j}(N,M) = \boldsymbol{COV}(\hat{V}_{n_i,N},\ \hat{V}_{n_j,N}) = V_{n_i+n_j} - V_{n_i+n_j-\frac{n_i n_j}{N}},\ 1 \le n_i,\ n_j \le N.$$

5.4. Confidence Intervals

For completeness, we briefly discuss how confidence regions for the estimators can be constructed. To do so, we need the matrix $\dot{V}(M,3) = \dot{V}_m(M,3)_{m=0,1,2,\dots,M}$ of partial derivatives of the expected spectrum with respect to the parameters:

$$\dot{V}(M,\ 3) = \begin{Bmatrix} \frac{\partial}{\partial\alpha}V_N(0;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(0;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(0;\ \alpha,\ \beta,\ \gamma) \\ \frac{\partial}{\partial\alpha}V_N(1;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(1;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(1;\ \alpha,\ \beta,\ \gamma) \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial\alpha}V_N(M;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(M;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(M;\ \alpha,\ \beta,\ \gamma) \end{Bmatrix}. \qquad (157)$$

Then for the parameter estimators

$$(\hat{\alpha}_N,\ \hat{\beta}_N,\ \hat{\gamma}_N)$$

the normal distribution can be assumed

$$\begin{pmatrix} \hat{\alpha}_N - \alpha \\ \hat{\beta}_N - \beta \\ \hat{\gamma}_N - \gamma \end{pmatrix} \overset{D}{\rightarrow} N(0, \hat{C}) , \tag{158}$$

with the covariance matrix

$$\hat{C} = (\hat{C}_{ij})_{1 \le i,j \le 3} = (\dot{V}(3,3))^{-1} R(N,3)(\dot{V}(3,3))^{-1}. \tag{159}$$

For the estimators

$$(\alpha^*_N, \beta^*_N, \gamma^*_N)$$

the normal distribution

$$\begin{pmatrix} \hat{\alpha}^*_N - \alpha \\ \hat{\beta}^*_N - \beta \\ \hat{\gamma}^*_N - \gamma \end{pmatrix} \overset{D}{\rightarrow} N(0, \hat{C}^*) , \tag{160}$$

can be used with the covariance matrix

$$\hat{C}^* = (\hat{C}^*_{ij})_{1 \le i,j \le 3} = [\dot{V}(M,3) R^{-1}(N,M) \dot{V}(M,3)]^{-1}. \tag{161}$$

With $M = 3$ these covariances obviously coincide, but if $M > 3$ then the estimators

$$(\alpha^*_N, \beta^*_N, \gamma^*_N)$$

are characterized by the narrower confidence region.

5.5. Goodness-of-fit Test for Models

The minimal values of the χ^2 statistics can be used to test whether the chosen parametric model fits the data. For instance, if estimation method 2 is used, then the minimal value of $\chi^2_{(M-3)}$ obtained when the parameters (α, β, γ) are substituted by their estimators $(\hat{\alpha}^*_N, \hat{\beta}^*_N, \hat{\gamma}^*_N)$ should be less then the desired significance level of the χ^2 distribution with *M-3* degrees of freedom.

Note also that some particular parametric model, satisfactory for the first M terms of the spectrum statistics

$$(\hat{V}_N, \hat{V}_N(m), 1 \le m \le M - 1)$$

may not be acceptable in a global sense for the whole vector

$$(\hat{V}_N(m), 1 \le m).$$

5.6. Comparing Samples

Two samples can be compared to establish the identity of the (theoretical) probability distributions of the corresponding general populations, for instance for the purpose of authorship determination.

Let two samples of sizes N^1 and N^2 be given with, generally speaking, different vocabularies, as in the case that texts written in different languages are compared:

$$\underline{V}^i = (A^i_1, A^i_2, \ldots, A^i_V), \; i = 1,2.$$

The corresponding frequencies, rank frequency distributions and frequency spectra are, for $i = 1,2$:

$$f^i_{N^i}(A^i_1), f^i_{N^i}(A^i_2), f^i_{N^i}(A_{\hat{V}^i})$$

$$f^i_{N^i}(A^i_1) \ge f^i_{N^i}(A^i_2) \ge \ldots \ge f^i_{N^i}(A_{\hat{V}^i}),$$

$$\hat{V}^i_{N^i}(m) = \sum_{j \ge 1} \mathbb{I}_{[f(A^i_j) = m]}, \; m = 1, 2, \ldots$$

$$\hat{V}^i_{N^i} = \sum_{m \ge 1} \hat{V}^i_{N^i}(m).$$

We must distinguish several ways in which the comparison problem can be stated in terms of the corresponding theoretical models expressed in the form of probability distributions

$$(P^i(A^i_j), 1 \le j \le V), \; i = 1,2$$

or structural probability distributions

$$G^i(p) = \sum_{j=1}^{V^i} \mathbb{I}_{[p^i(A_j^i) \geq p]}, \; p \geq 0, \; i = 1,2.$$

First consider the case for which the vocabularies are identical,

$$A_j^1 = A_j^2 = A_j, \; j = 1,2,\ldots ,$$

the two texts being written in one and the same language. It is natural to construct this comparison problem in terms of the hypothesis that the (individual) probabilities coincide:

$$P^1(A_j) = P^2(A_j), \; 1 \leq j \leq V.$$

Since high and low frequencies can be considered as independent for LNRE samples (see section 3.3.3), we can focus on the left hand (high probabilities) or on the right hand (low probabilities) tails of a rank probability distribution. We can apply e.g. the standard χ^2 test to check the coincidence of high probabilities. The testing of the right hand tails is quite nontrivial, however.

To do this, consider the united sample of size $N = N^1 + N^2$ with frequencies

$$f_N(A_1) = f^1_{N^1}(A_1) + f^2_{N^2}(A_2),$$

$$f_N(A_2) = f^1_{N^1}(A_2) + f^2_{N^2}(A_2), \; \ldots ,$$

$$f_N(A_{\hat{V}}) = f^1_{N^1}(A_{\hat{V}}) + f^2_{N^2}(A_{\hat{V}}).$$

Introduce the joint frequency spectrum

$$\hat{V}_N(m, k, l) = \sum_{j \geq 1} \mathbb{I}_{[f_N(A_j) = m, \, f^1_{N^1}(A_j) = k, \, f^2_{N^2}(A_j) = l]}, \; m = 1, 2, \ldots, k+l = m,$$

the number of elements which appear k times in the first and l times in the second sample, and let $\hat{V}_N(m)$ be a frequency spectrum on the united sample:

$$\hat{V}_N(m) = \sum_{k+l=m} \hat{V}_N(m, k, l).$$

Applying the scheme of sampling without replacement presented in section 3.1.4, according to which, for large enough N, the vector

$$\hat{V}_N(m, k, l), \; 0 \leq k \leq m, \; k + l = m$$

is multinomially distributed given $\hat{V}_N(m)$, the following series of χ^2 statistics can be suggested for the test of comparison:

$$\chi^2(m) = \sum_{k=0}^{m} \frac{(\hat{V}_N(m, k, l) - B(m, k, \frac{N^1}{N})\hat{V}_N(m))^2}{B(m, k, \frac{N^1}{N})\hat{V}_N(m)}$$

in particular,

$$\chi^2(1) = \frac{1}{N^1 N^2} \frac{(N^1 \hat{V}_N(1, 0, 1) - N^2 \hat{V}_N(1, 1, 0))^2}{\hat{V}_N(1, 0, 1) + \hat{V}_N(1, 0, 1)}.$$

For equal sample sizes ($N^1 = N^2$), the distance between two samples measured in terms of the hapaxes, the number of elements that appeared in exactly one of the samples only, is expressed by the ratio

$$\chi^2(1) = \frac{(\hat{V}_N(1, 0, 1) - \hat{V}_N(1, 1, 0))^2}{\hat{V}_N(1, 0, 1) + \hat{V}_N(1, 1, 0)}.$$

By successively checking the admissibility of the values

$$\chi^2(1), \; \chi^2(1) + \chi^2(2), \; \chi^2(1) + \chi^2(2) + \chi^2(3), \ldots$$

with respect to the critical levels of the χ^2-distribution with 1, 3, 6, ... degrees of freedom respectively, we are able to accept with increasing accuracy the hypothesis of coincidence on the tails of the probability distributions studied.

Next consider the case that the samples have been obtained from general populations with different vocabularies (texts written in different languages). It is reasonable to analyse this problem in terms of structural probability distributions. Of course, we can do this even when the vocabularies are the same, in which case we accept the identity of the theoretical models and state that although the individual probabilities may be different, the rank probability distributions coincide.

To check the coincidence of the right hand tails of rank probability distributions we must compare the components of the frequency spectra $\hat{V}^1_{N^1}(m)$ and $\hat{V}^2_{N^2}(m)$, $m \geq 1$. To construct the χ^2 statistic for the difference of the spectrum terms (even in the case that the sample sizes are the same), we need the covariance matrix $\boldsymbol{COV}(\hat{V}^i_{N^i}(m),\ \hat{V}^i_{N^i}(k))$, which is itself unknown. We can apply formula (147), which represents this matrix in terms of the expected spectrum components corresponding not only to the sample sizes N^1 and N^2, but to $2N^1$ and $2N^2$ as well. The natural way to proceed is to use (149) to obtain $V_{2N^i}(m)$, $m \geq 1$, substituting the observed values for the expected ones. Unfortunately, we are again confronted with the problem of the instability of (149) for $n \geq 2N$, apart from the necessity of taking into account the differences between the sample sizes.

To avoid these difficulties, the following construction scheme of a χ^2-distance between the samples can be suggested. First construct the interpolated vocabulary growth curves for both samples $i = 1, 2$:

$$\hat{V}^i_{N^i,n} = \sum_{j\geq 1} \hat{V}^i_{N^i}(j)\left(1 - (1 - \frac{n}{N})^j\right),\ 1 \leq n \leq N^i. \tag{162}$$

Next construct the estimations $\hat{Q}^i_{N^i}(n, k)$, $1 \leq n,\ k \leq N^i$, $i = 1,2$ for the covariances

$$\boldsymbol{COV}(\hat{V}^i_{N^i,n},\ \hat{V}^i_{N^i,k}),\ 1 \leq n,\ k \leq N^i$$

using

$$\hat{Q}^i_{N^i}(n,\ k) = \hat{V}^i_{N^i,n+k} - \hat{V}^i_{N^i,n+k-\frac{nk}{N^i}}$$

(cf. sections 3.3.2 and 5.2). Finally, fix some sample size values on which the interpolated vocabulary growth curves for two samples should be compared,

$$1 \leq n_1 \leq n_2 \leq n_3 \leq \ldots \leq \min(N^1,\ N^2)\ ,$$

and construct the χ^2-distance

$$D^{1,2}(N^1,\ N^2) = \sum_{j,k} (\hat{V}^1_{N^1,n_j} - \hat{V}^2_{N^2,n_j}) Q^{(-1)}(n_j, n_k)(\hat{V}^1_{N^1,n_k} - \hat{V}^2_{N^2,n_k})$$

where $Q^{(-1)}$ is the inverse matrix of the sum of the matrices

$$\hat{Q}^1_{N^1}(n,\ k) + \hat{Q}^2_{N^2}(n,\ k).$$

In other words, the identity of the theoretical structural distributions is checked by the χ^2-distance between the interpolated vocabulary growth, and the abovementioned difficulty is avoided because the maximal index $n_j + n_k$ is selected to be less than $\min(N^1,\ N^2)$ in the expression of $\hat{Q}^i_{N^i}(n_j,\ n_k)$.

Up till now we have focussed on non-parametric tests for the identity of two samples. If some parametric family of (structural) probability distribution is accepted as satisfactory for both samples, then the comparison tests can be improved (become more powerfull) if they are based on the comparison of the estimated parameters. Let

$$(\hat{\alpha}^i_{N^i},\ \hat{\beta}^i_{N^i},\ \hat{\gamma}^i_{N^i},\ i = 1,2)$$

be the estimated parameters constructed by one of the estimation schemes discussed above. As the χ^2-distance between the samples we can consider the quadratic form (a χ^2-statistic with three degrees of freedom)

$$(X,\ AX) = \sum_{i,j=1}^{3} X_i A^{(-1)}_{i,j} X_j$$

where

$$X_1 = \hat{\alpha}^1_{N^1} - \hat{\alpha}^2_{N^2}$$

$$X_2 = \hat{\beta}^1_{N^1} - \hat{\beta}^2_{N^2}$$

$$X_3 = \hat{\gamma}^1_{N^1} - \hat{\gamma}^2_{N^2}$$

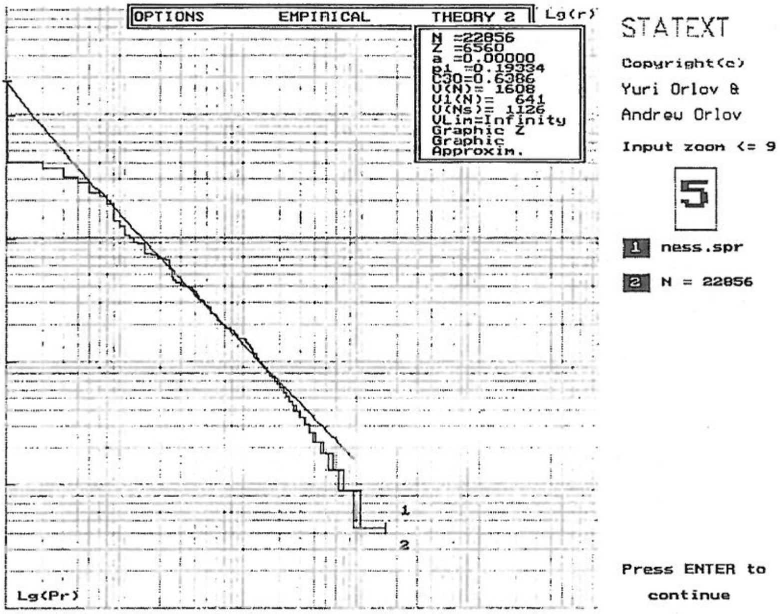
and where $A^{(-1)}$ is the inverse matrix of the sum of covariance matricies of the estimators.

5.7. Software

Software for carrying out LNRE analyses is currently being developed by various researchers. J.K.Orlov and A.J.Orlov have completed a program (STATEXT) for IBM-compatible PC that estimates the parameters for the extended Zipf and Yule-Simon 'laws'. The output of the program is a plot of *log(r)* versus *log(p*

tribution. The plot shows that there is considerable divergence for the lowest probabilities. A similar plot using the Yule-Simon 'law' is, at least to visual inspection, quite satisfactory.

At the Institute of Mathematics of the Georgian Academy of Sciences (Tbilisi), the first author has initiated a more general project for the investigation of various aspects of LNRE distributions such as modelling of LNRE samples, comparison of samples, distribution and interaction between words, and the analysis of word frequency distributions using the generalized Zipf's 'law'. With respect to the analysis of the frequency spectrum, a program has been developed for PC that estimates parameters for the Zipf and Yule-Simon 'laws'. Like STATEXT, it plots the empirical and expected rank frequency curves, but in addition it calculates confidence intervals for the frequency spectrum as well as the goodness-of-fit in terms of the test statistic $\chi^2_{(M-h)}$, with h the number of parameters. For the data on *-ness* the fit obtained for the Yule-Simon model is shown in table 2. The parameters of the Yule-Simon ($\hat{\beta}$ = 1.009, $\hat{\imath}$ = 3.342) model were estimated using the Tbilisi program (χ^2 = 65.66, q « 0.001).



```
N =22856
Z =6560
a =0.00000
p1 =0.19334
C30=0.6386
V(N)= 1608
V1(N)=  641
V(Ns)= 1126
VLim=Infinity
Graphic Z
Graphic
Approxim.
```

Figure 11. Rank-probability plot of the extended Zipf 's 'law': STATEXT applied to the English suffix *-ness.*

A semi-automatic program for estimating the parameters of the extended Waring-Herdan 'law' has been developed at the Max-Planck Institute for Psycholinguistics by the present authors. This program, which runs under UNIX, allows one to interactively search through the parameter space for parameter values satisfying $V_N(0) = \hat{V}_N(0)$ and $V_N(1) = \hat{V}_N(1)$ and minimalizing $\chi^2_{(M-h)}$. Thusfar, attempts to develop a fully automatic estimation procedure have failed, due to the for numerical calculation infelicitous expression for V^Z (143) and the bounded parameter ranges. The results obtained for English *-ness* are summarized in table 2. The fit, with the estimated parameters ($\hat{\alpha}$ = 0.712, $\hat{\beta}$ = 1.075, $\hat{\imath}$ = 0.1) is optimal in the chi-square sense, other choices of the parameters leading to higher values of the χ^2 statistic. Since $\chi^2_{(4)}$ = 8.21, q = 0.084, we may be confident that a reasonable fit has been obtained.

Table 2. Observed and estimated frequency spectrum: *-ness*

m	$\hat{V}_N(m)$	$V_N(m)$			
		Yule-Simon	Waring-Herdan	Lognormal	Gauss-Poisson
1	749	646	748	523	749
2	215	257	228	226	229
3	126	144	110	130	116
4	68	94	65	86	73
5	59	66	44	62	51
6	30	50	32	47	38
7	31	39	24	37	30
8	29	31	19	30	24
9	22	25	15	25	20
10	20	21	13	21	17

The authors have completed a fully automatic estimation and evaluation programs for the lognormal 'law' and the Gauss-Poisson 'law'. The results obtained for *-ness* can be found in table 2. For the lognormal 'law', the parameter values $\hat{\mu}$ = -5.0, $\hat{\sigma}$ = 2.570 lead to a minimal chi-square value ($\chi^2_{(4)}$ = 206.73) that, unfortunately, fails to meet any standards of acceptability (q = 0.000). The lowest chi-square value for the Gauss-Poisson 'law', $\chi^2_{(4)}$ = 6.292, q = 0.178, was obtained for the parameters $\hat{\gamma}$ = 0.5, $\hat{b}$ = 0.0092, $\hat{c}$ = 0.0264. Evidently, the Gauss-Poisson 'law' provides the best fit. Perhaps not surprisingly, the 'law' with the smallest number of parameters, the lognormal 'law', fails to meet the simultaneous requirements $\hat{V}_N = V_N$ and $\hat{V}_N(1) = V_N(1)$ [1].

[1] In this paper, the lognormal 'law' is fitted to the data using the expressions (132), the integrals being evaluated numerically by means of Romberg integration (see Press et al. 1988). The results obtained contrast with those reported in Baayen (1993b). Using the approximation method suggested by Carroll (1967), he obtained reasonable fits for the more

Table 3. Observed and estimated frequency spectrum: *en-*.

m	$\hat{V}_N(m)$	$V_N(m)$		
		Waring-Herdan	Lognormal	Gauss-Poisson
1	11	6	11	11
2	9	4	8	7
3	4	3	6	5
4	2	2	4	4
5	1	2	4	3

By way of comparison, consider table 3, which lists the results obtained for the unproductive prefix *en-*. The lognormal 'law' ($\hat{\mu}$ = -2.0, $\hat{\sigma}$ = 2.335, $\chi^2_{(3)}$ = 4.27, q = 0.234) does much better than the extended Waring-Herdan 'law' ($\hat{\alpha}$ = 0.3, $\hat{\beta}$ = 1.0027, $\hat{\imath}$ = 10, $\chi^2_{(3)}$ = 34.47, q = 0.000), which fails to provide a parameter set that simultaneously satisfies the equations $\hat{V}_N = V_N$ and $\hat{V}_N(1) = V_N(1)$. Given that *en-* is located outside the (late) LNRE ZONE (see table 1 and section 3.2.1), and given that the extended Waring-Herdan model is tightly linked with the LNRE ZONE, the lack of accuracy - note the large value of $\hat{\imath}$ - is to be expected. The most accurate fit is again provided by the Gauss--Poisson 'law' ($\hat{\gamma}$ = -0.0005, $\hat{b}$ = 0.02289, $\hat{c}$ = 0.0683, $\chi^2_{(3)}$ = 3.31, q = 0.346), but even here the extremely low value of $\hat{\gamma}$ and the slightly too high values for m = 3, 4, 5 suggest that this 'law' is stretched to, or perhaps beyond its limits in its attempt to model the frequency spectrum of this unproductive prefix. (For a more detailed comparison of these models with respect to goodness-of-fit for a variety of samples the reader is refered to Baayen 1993b).

6. Morphology and the LNRE ZONE

In the previous sections the frequency spectra of the English affixes *-ness* and *en-* have been analyzed in some detail. The suffix *-ness*, a typical example of a productive affix, is characterized by a frequency distribution that is dominated by low-frequency types. Not surprisingly, the theoretical vocabulary as estimated by the Gauss-Poisson 'law' exceeds the observed vocabulary by a factor 5

productive affixes. The more rigidly defined methods used in the present paper, however, suggest that for the more productive affixes the lognormal 'law' fails to reach the same level of accuracy as the Waring-Herdan and Gauss-Poisson 'laws'.

(V = 8261, $\hat{V}_N$ = 1607). In contrast, the frequency distribution of the unproductive prefix *en-* is dominated by the higher-frequency types and the theoretical vocabulary V = 114, again calculated using the Gauss-Poisson 'law', is only slightly larger than the observed vocabulary $\hat{V}_N$ = 94.

Interestingly, frequency spectra of full texts resemble the spectrum of the morphological category of nouns in *-ness*, the spectrum of the prefix *en-* being quite atypical. Since the large numbers of rare types appearing in the frequency spectra of the productive morphological categories as realized in some text necessarily appear as the 'rare events' of the frequency spectrum of the text as a whole, it seems natural to explore the hypothesis that productive word formation processes anchor texts in the LNRE ZONE. We will investigate this possibility by analysing the morphological constituency of the words appearing in two 'texts', E.Bronte's 'Wuthering Heights' ($N \approx$ 120,000), the full text of which was obtained by anonymous ftp from the Online Book Initiative at obi.std.com, and the Dutch INL corpus ($N \approx$ 40,000,000), using the word frequencies as given in the CELEX lexical database (Burnage, 1990).

6.1. The Development of Morphology in Bronte's 'Wuthering Heights'

First consider E.Bronte's novel. According to the tests developed in section 3.2.1, we are dealing with a text that is located in the central LNRE ZONE: C_L = 0.165, the number of hapaxes is increasing ($\frac{d}{dN} V_N(1) = \frac{1}{N}(V_N(1) - 2V_N(2))$= (1/119321)*(2427 - 2*973) > 0), and in addition, the theoretical vocabulary V as calculated using the Gauss-Poisson 'law' encompasses some 12,150 word types, a number exceeding by roughly a factor 2 the observed vocabulary (6420).

One way of investigating the extent to which productive morphological rules may be held responsible for this novel's location in the central LNRE ZONE is to focus on how morphology contributes to the growth rate $\frac{d}{dN} V_N$of the vocabulary. Using (59), we may estimate the growth rate by $\frac{\hat{V}_{119321}(1)}{119321}$ = 0.02. Note that the growth rate, when viewed as a function of the sample size, is completely determined by the number of hapax legomena. Thus it seems natural to approach the question of whether morphology effects a text's location in the LNRE ZONE by investigating what proportion of the hapax legomena are morphologically complex, since this will allow us to gauge the extent to which word formation gives rise to the substantial growth rate of Bronte's vocabulary as it unfolds through 'text time'. The left hand graph of figure 12 plots the fraction

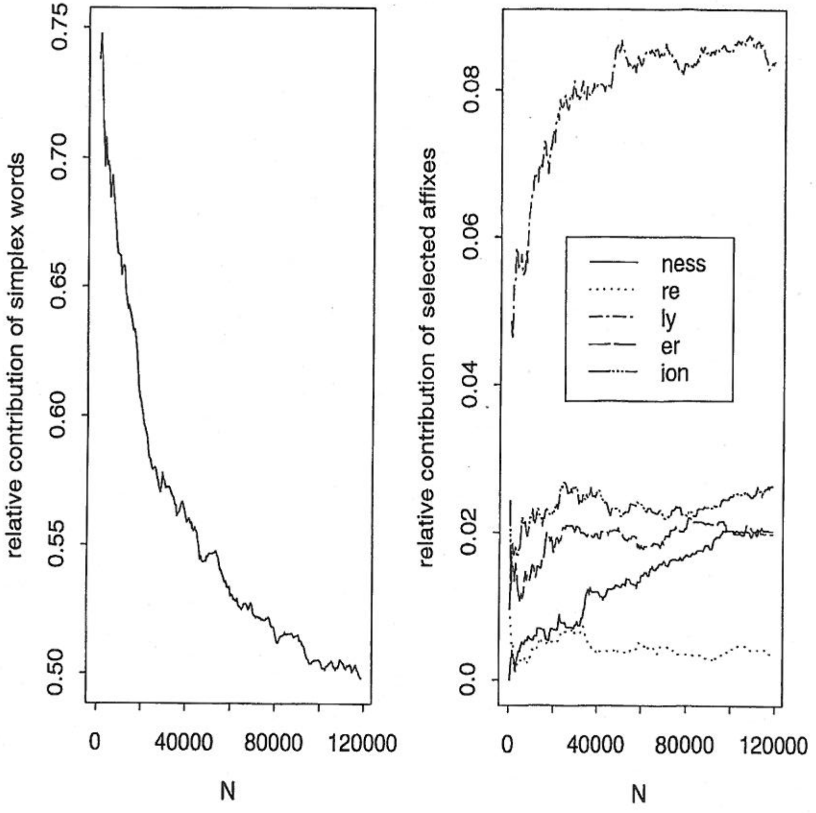


Figure 12. The relative contribution of simplex words and a selected set of affixes to the growth rate of the vocabulary in E.Bronte's 'Wuthering Heights'. The measurement interval for N equals 500 word tokens.

$$\hat{H}_N^{(s)} = \frac{\hat{V}_N^{(s)}(1)}{\hat{V}_N(1)}$$

of hapax legomena that are monomorphemic or simplex *(s)* as a function of the text size N. The right hand graph plots the fraction

$$\hat{H}_N^{(e)} = \frac{\hat{V}_N^{(e)}(1)}{\hat{V}_N(1)}$$

of polymorphemic hapaxes for selected affixes e. What we find is that $\hat{H}_N^{(s)}$ is a decreasing function of N, whereas $\hat{H}_N^{(e)}$ increases with N, notably so for highly productive suffixes like *-ness* and especially *-ly*. Note that for the full novel, the relative contribution of morphology is substantial:

$$\hat{H}_N^{(\bar{s})} = \sum_e \hat{H}_N^{(e)} = 0.502.$$

Bronte's novel is too small to allow us to investigate how the relative contribution of morphology to the growth rate will develop for larger samples. The left hand graph of figure 12, however, suggests that a further increase in the relative contribution of morphology may be expected for larger texts. To gain some insight into the 'limiting' properties of $\hat{H}_N^{(s)}$ and $\hat{H}_N^{(\bar{s})}$ we therefore analyze the frequency distribution of a much larger sample, the INL corpus of written Dutch.

6.2. Morphology in the INL Corpus

The INL corpus, compiled by the Dutch Institute for Lexicography, contains roughly 40,000,000 wordforms. With the exception of the hapaxes, the frequencies of the words occurring in this corpus as well as detailed information on the orthographical, morphological and phonological properties of these words are available in the CELEX lexical database. The first spectrum elements of the frequency distribution of the INL corpus are presented in table 4. Even though the hapax legomena are not registered in the CELEX lexical database, the available spectrum elements allow us to ascertain that even this moderately large text corpus is located in the LNRE ZONE. For instance, using (63) we find that $\frac{d}{dN} V_N(2) > 0$. In addition, $C_L = 0.022$, which is quite high given that some 65,000 types have been registered. Inspection of the morphological constituency of the dislegomena reveals that $\hat{H}_N^{(s)} = 291/7264 = 0.04$, a substantially lower value than the corresponding value for Bronte's 'Wuthering Heights', $540/973 = 0.555$ ($p < 0.001$). This suggests informally that the asymptotic value of $\hat{H}_N^{(s)}$ for $N \rightarrow \infty$ will tend to zero.

Table 4. The tail of the frequency distribution of the lemmas registered in the CELEX lexical database.

m	1	2	3	4	5	6	7	8	9	10
$\hat{V}_N(m)$	-	7269	4355	3433	2569	2296	1834	1646	1391	1313

The crucial relation between morphology and the LNRE property of texts can also be approached from a slightly different angle. Figure 13 plots the degree of morphological complexity, measured in terms of the number of morphological constituents of a word, as a function of the frequency of that word, using a non-parametric regression technique (see Haerdle 1991). Clearly, morphological complexity is a decreasing function of word frequency, an illustration on the morphological level of Zipf's 'law of abbreviation' (Zipf 1935). Table 5 illustrates the same point in a slightly different way: a relatively small set of monomorphemic words accounts for the bulk of all tokens, morphological complexity being relatively scarce token-wise but, paradoxically, frequentially dominant type-wise.

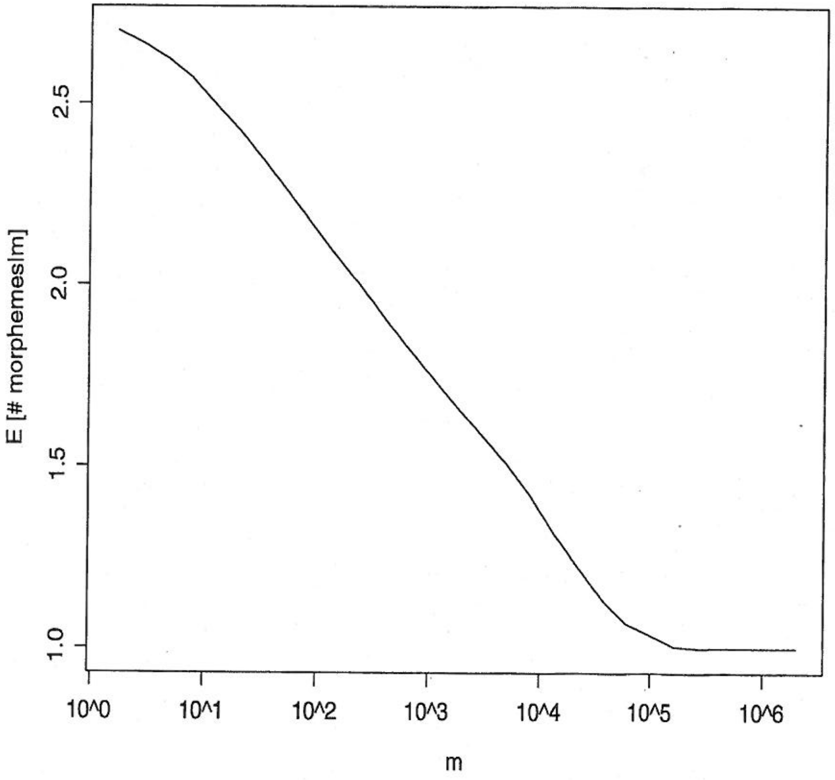


Figure 13. The number of constituent morphemes as a function of (log) word frequency (ln m) in the INL corpus. (WARPing approximation of the Nadaraya-Watson estimator using an Epanechnikov kernel, a bin width 0.5 and a window width 2.0).

Table 5. Statistics for the morphological constituency of the lemmas in the CELEX lexical database.

Morphology	N	$\hat{V}_N$	$\hat{V}_N(2)$
monomorphemic words	30197189	8083	293
compounding	1959754	32622	5324
derivation	2713506	13656	1144
synthetic compounding	85025	1206	159
undefined	3464960	9795	679
Total	38420434	65362	7599

6.3. Productive Rules as LNRE Generators

We have seen that word formation rules play a crucial role in anchoring texts in the LNRE ZONE. This result seriously questions the validity of the rationals for the Rouault, Mandelbrot and Waring-Herdan 'laws' discussed in section 4. The main problem with these rationals is that they fail to take into account what Martinet (1965) has called the 'double articulation' of language, the fact that language is structured on two relatively autonomous planes, the phonological plane and the morphology-syntax plane. Since Markovian models in which words appear as strings of letters focus exclusively on the phonological plane, they cannot and in simulations do not give rise to lexica with realistic frequency-length characteristics, nor can the similarity relations in the lexicon be modelled adequately (see Baayen 1991 for detailed discussion).

Interestingly, the defining characteristic of monomorphemic words is, from a quantitative point of view, that the associated theoretical vocabulary is strictly finite. In contrast, productive morphological categories can be argued to be, at least in theory, infinite. To see this, first consider the simplex words registered in the Ascot version of the Longman dictionary of Contemporary English and the Oxford Advanced Learner's Dictionary. Table 6 summarizes the first ten spectrum elements. The first value of m for which $\frac{d}{dN}V_N(m) > 0$ equals 6, indicating that this sample is located outside the late LNRE ZONE. This is confirmed by a comparison of the observed number of types $\hat{V}_N = 11869$ and the approximated theoretical vocabulary size $V \approx 12{,}000$, calculated on the basis of a rather bad Gauss-Poisson fit ($\chi^2_{(18)} = 836.87$, $\hat{\gamma} = -0.01$, $\hat{b} = 0.0229$, $\hat{c} = 0.00067$). The nearly identical sizes of the observed and theoretical vocabularies is reminiscent of what we have observed for the unproductive prefix en-. In both cases we are dealing with strictly finite populations.

Table 6. The tail of the frequency distribution of the monomorphemic lemmas in the Longman dictionary of Contemporary English and the Oxford Advanced Learner's Dictionary in the Cobuild corpus as registered in the CELEX lexical database.

m	1	2	3	4	5	6	7	8	9	10
$\hat{V}_N(m)$	357	336	259	264	261	225	182	193	163	169

Next consider the Dutch diminutive *-tje*, an extremely productive derivational suffix. Using the Uit den Boogaart (1977) corpus, the Gauss-Poisson 'law' predicts a theoretical vocabulary of 1,239,156,496 types ($\chi^2_{(13)} = 19.95$, $q = 0.0965$), a value large enough to substantiate the claim that unrestricted productivity gives rise to infinite populations. From this point of view, the following formal definition for LNRE distributions (Khmaladze & Chitashvili 1989), which can be realized only for infinite V (see section 3.2),

$$\lim_{N \to \infty} \alpha_N(1) > 0 \qquad (163)$$

appears to be useful as a formal definition of productivity as well. Of course, many productive categories do not meet this strict probabilistic definition. In the case of *-ness*, for instance, the theoretical vocabulary is approximately 8,000, exceeding the observed vocabulary by 'only' a factor 5. Even in the case of the highly productive English adverbial suffix *-ly* ($\hat{V}_N = 3914$) the estimated theoretical vocabulary equals a 'mere' 24,000 types (Gauss-Poisson estimation, $\chi^2_{(18)} = 166.09$). Observe, however, that *-ly* by itself potentially generates a morphological category with twice as many types as estimated for the monomorphemic English words discussed above (see table 6). This suggests that, even when (163) is not strictly met, the very large numbers of 'morphologically possible words' defined by all word formation rules of the language jointly, will anchor running text in the LNRE ZONE for substantial values of N. How large N should be for a text to move out of the late LNRE ZONE into the 'Law of Large Numbers ZONE' is at present unclear. Perhaps the huge corpora that are at present being compiled, such as the British National Corpus and the International Corpus of English, will shed more light on this issue, that, for as yet, has to be left unresolved.

References

Baayen, R.H. (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Diss. Free University, Amsterdam, 1989.

Baayen, R.H. (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics: 271-278.* Berkeley.

Baayen, R.H. (1992). A Quantitative Approach to Morphological Productivity. In: G.E.Booij & J.van Marle (eds.), *Yearbook of Morphology 1991: 109-149.* Dordrecht: Kluwer.

Baayen, R.H. (1993a). On frequency, transparency and productivity. In: G.E. Booij & J.van Marle (eds.), *Yearbook of Morphology 1992: 181-208.* Dordrecht: Kluwer.

Baayen, R.H. (1993b). Statistical models for word frequency distributions: a linguistic evaluation. *Computers and the Humanities 26, 331-347.*

Baayen, R.H. & Lieber, R. (1991). Productivity and English Derivation: A corpus bases study. *Linguistics, 29, 801-843.*

Brunet, E. (1978). *Le Vocabulaire de Jean Giraudoux. Structure et Évolution.* Genève: Slatkine.

Burnage, G. (1990). *CELEX; A guide for users.* Nijmegen: Centre for Lexical Information.

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

Carroll, J.B. (1969). A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions. *Research Bulletin - Educational Testing Service. Princeton, November 1969.*

Carroll, J.B. (1970). An alternative to Juilland's Usage Coefficient for Lexical Frequencies, and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior 3, 61-65.*

Efron, B. & Thisted, R. (1976). Estimating the Number of Unseen Species: How many Words did Shakespeare Know? *Biometrika 63, 435-447.*

Good, I.J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika 40, 237-264.*

Good, I.J. & Toulmin, G.H. (1956). The Number of New Species and the Increase in Population Coverage, when a Sample is Increased. *Biometrika 43, 45-63.*

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire.* Paris: Presses Universitaires de France.

Haerdle, W. (1991). *Smoothing Techniques With Implementation in S.* Berlin: Springer.

Herdan, G. (1960). *Type-Token Mathematics*. The Hague: Mouton.

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

Kalinin, V.M. (1965). Functionals Related to the Poisson Distribution, and Statistical Structure of a Text. In: J.V. Finnik (ed.), *Articles on Mathematical Statistics and the Theory of Probability: 202-220*. Providence, Rhode Island: American Mathematical Society.

Khmaladze, E.V. (1987). *The Statistical Analysis of Large Number of Rare Events*. Report MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

Khmaladze, E.V. & Chitashvili, R.J. (1989). Statistical Analysis of Large Number of Rare Events and Related Problems. *Transactions of the Tbilisi Mathematical Institute 91, 196-245.*

Lánský, P. & Radil-Weiss, T. (1980). A Generalization of the Yule-Simon Model, with Special Reference to Word Association Tests and Neural Cell Assembly Formation. *Journal of Mathematical Psychology, 21, 53-65.*

Mandelbrot, B. (1953). *An information theory of the statistical structure of language*. In: Jackson, W.E. (ed.), *Communication Theory: 503-512*. New York, Academic Press.

Mandelbrot, B. (1962). On the Theory of Word Frequencies and on Related Markovian Models of Discourse. In: R. Jakobson (ed.), *Structure of Language and its Mathematical Aspects: 190-219*. Proceedings of Symposia in Applied Mathematics, Vol. XII. Providence, Rhode Island: American Mathematical Society.

Martinet, A. (1965). *La linguistique synchronique: études et recherches*. Paris: Presses Universitaires de France.

Menard, N. (1983). *Mesure de la Richesse Lexicale. Théorie et vérifications expérimentales. Etudes stylométriques et sociolinguistiques*. Genève: Slatkine-Champion.

Miller, G.A. (1957). Some Effects of Intermittent Silence. *The American Journal of Psychology 52, 311-314.*

Muller, C. (1977). *Principes et Méthodes de Statistique Lexicale*. Paris: Hachette.

Muller, C. (1979). Du nouveau sur les distributions lexicales: la formule de Waring-Herdan. In: C. Muller. (ed.), *Langue Française et Linguistique Quantitative: 177-195*. Genève: Slatkine.

Orlov, J.K. (1983a). Dynamik der Häufigkeitsstrukturen. In: H. Guiter & M.V. Arapov (eds.), *Studies on Zipf's Law:* 116-153. Bochum: Brockmeyer.

Orlov, J.K. (1983b). Ein Model der Häufigkeitsstruktur des Vokabulars. In: H. Guiter & M.V. Arapov (eds.), *Studies on Zipf's Law: 154-233*. Bochum: Brockmeyer.

Orlov, J.K. & Chitashvili, R.Y. (1982a). On the Distribution of Frequency Spectrum in Small Samples from Populations with a Large Number of Events. *Bulletin of the Academy of Sciences, Georgia 108.2, 297-300.*

Orlov, J.K. & Chitashvili, R.Y. (1982b). On Some Problems of Statistical Estimation in Relatively Small Samples. *Bulletin of the Academy of Sciences, Georgia 108.3, 513--516.*

Orlov, J.K. and Chitashvili, R.Y (1983a). On the Statistical Interpretation of Zipf's Law. *Bulletin of the Academy of Sciences, Georgia 109.3, 505-508.*

Orlov, J.K. and Chitashvili, R.Y. (1983b). Generalized Z-Distribution Generating the Well-Known 'Rank-Distributions'. *Bulletin of the Academy of Sciences, Georgia 110.2, 269-272.*

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1988). *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Rouault, A. (1978). Loi de Zipf et sources markoviennes. *Annales de l'Institute H. Poincare 14, 169-188.*

Scarborough, D.L., Cortese, C. & Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance 3.1, 1-17.*

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

Sichel, H.A. (1986). Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist 11, 45-72.*

Simon, H.A. (1955). On a Class of Skew Distribution Functions. *Biometrika 42, 435-440.*

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

Sinclair, J.M. (ed.) (1985). *Looking Up: An Account of the Cobuild Project in Lexical Computing*. London: Collins.

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

Uit den Boogaart, P.C. (ed.) (1975). *Woordfrekwenties in gesproken en geschreven Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

Whaley, C.P. (1978). Word-Nonword Classification Time. *Journal of Verbal Learning and Verbal Behavior 17, 143-154.*

Yule, G.U. (1924). A mathematical theory of evolution, based on the conclusions of Dr. J.C.Willis, F.R.S. *Philosophical Transactions of the Royal Society of London Ser. B, 213, 21-87.*

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zipf, G.K. (1935). *The Psycho-Biology of Language*. Boston: Houghton Mifflin.