# Statistical Models for Word Frequency Distributions: A Linguistic Evaluation

Harald Baayen

*Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands*
*e-mail: baayen@mpi.nl*

**Abstract:** Three models for word frequency distributions, the lognormal law, the generalized inverse Gauss-Poisson law and the extended generalized Zipf's law are compared and evaluated with respect to goodness of fit and rationale. Application of these models to frequency distributions of a text, a corpus and morphological data reveals that no model can lay claim to exclusive validity, while inspection of the extrapolated theoretical vocabulary sizes raises doubts as to whether the urn scheme with independent trials is the correct underlying model for word frequency data. The role of morphology in shaping word frequency distributions is discussed, as well as parallelisms between vocabulary richness in literary studies and morphological productivity in linguistics.

**Key Words:** word frequency distribution, lognormal, generalized inverse Gauss-Poisson, extended generalized Zipf's law, vocabulary richness, morphological productivity, goodness of fit

## 1. Introduction

Word frequency distributions have been studied intensively from both literary and linguistic perspectives. In literary studies, word frequency distributions have been used to obtain estimates of an author's vocabulary (e.g. Menard, 1983; Efron and Thisted, 1976; Muller, 1979) or to obtain some more or less invariant characteristic measure of the distribution (e.g. Yule, 1944; Guiraud, 1954; Brunet, 1978). In linguistic studies, word frequency distributions have been studied for corpora (Carroll, 1967) as well as for subsets of words selected according to some linguistic criterion (e.g. nouns [Yule, 1944], abstract nouns in *-ness* and *-ity* [Harwood and Wright, 1956] or 'coverbs' [Roy, 1976]). Baayen (1989, 1991b) and Baayen and Lieber (1991) studied the word frequency distributions of morphological categories with respect to their productivity. Interpreting the notion of productivity as the statistical readiness (Bolinger, 1948) with which new words are formed spontaneously and unintentionally (Schultink, 1961; Baayen and Lieber, 1991), they found that the growth rate of the vocabulary is a useful quantitative measure for the degree of productivity of a word formation rule. Another way in which the productivity of a word formation rule can be evaluated is to consider the number of potential words the rule might give rise to. This is the way in which the question of how to estimate the theoretical vocabulary size re-appears in linguistics.

Since the reliability of estimates of the theoretical vocabulary size depend on the assumptions one is prepared to make concerning the distribution 'law' underlying the frequency data, it is important to subject statistical models that allow the theoretical vocabulary size to be estimated[1] to a detailed analysis of their rationale, goodness-of-fit and predictive adequacy. This is the main aim of the present paper. A second aim is to point out some similarities between the frequency distributions of well-written literary texts and productive word formation processes on the one hand, and

---

*R. Harald Baayen received his PhD at the Free University, Amsterdam, where he was involved in research on morphological productivity. He is now at the Max-Planck Institute for Psycholinguistics, Nijmegen, participating in a project on computational modelling of lexical representation and process.*

*Computers and the Humanities* **26**: 347—363, 1993.

between those of large corpora and unproductive word formation processes on the other.

The paper is structured as follows. In section 2 some necessary objects and notations are introduced. Section 3 discusses Carroll's (1967) lognormal law, Sichel's (1975, 1986) generalized inverse Gauss-Poisson law and Orlov and Chitashvili's (1983b) extended generalized Zipf's law. The role of morphology and semantics in shaping word frequency distributions is sketched in section 4, followed by a discussion of the results obtained with respect to the theoretical vocabulary size in section 5.

## 2. Word Frequency Distributions

Once the criteria for distinguishing between word types — in the present study, dictionary entries or lemmas — have been established, one can count the number of occurrences or tokens for each type in a text. Two ways of summarizing word frequency counts are relevant here. A rank-frequency distribution is obtained when the frequency $f_i$ of the $i^{th}$ type is viewed as a function of its rank $i$, the types being ranked such that $f_i \geqslant f_{i+1}$ for all $i$. A grouped frequency distribution is obtained when the number of types $n_r$ for which $f_i = r$ are grouped together in frequency class $r$. Expressions for the rank-frequency distribution can be transformed into expressions for the grouped frequency distribution. For instance, the Zipf-Mandelbrot law (Mandelbrot, 1962)

$$f_i = \frac{K}{(i+B)^\gamma}, \tag{1}$$

with $\gamma$ a parameter of type richness, $B$ a parameter introduced to account for systematic departure from Zipf's law $f_i = K/i^\gamma$ at the head of the distribution, and $K$ a normalizing constant, is stated in terms of the rank-frequency distribution. It is reformulated in terms of the grouped frequency distribution as follows:

$$\begin{aligned} \mathrm{E}[n_r] &= \sum_i \mathrm{I}[f_i \geqslant r] - \sum_i \mathrm{I}[f_i \geqslant r+1] \\ &= \left(\frac{r}{K}\right)^\gamma - B - \left(\frac{r+1}{K}\right)^\gamma + B \\ &= K^{-\gamma}[r^\gamma - (r+1)^\gamma]. \end{aligned} \tag{2}$$

Note that the parameter $B$ disappears in the expression for $\mathrm{E}[n_r]$. This illustrates a general property of models phrased in terms of the grouped frequency distribution, namely that they are useful for the study of the lower frequency types only.

The parametric models to be discussed in this paper will be evaluated on the basis of their rationales on the one hand, and on the basis of the goodness-of-fit on the other. Denoting the observed vocabulary size at sample size $N$ by $V(N)$ and writing $n_r(N)$ for the number of types with frequency $r$ in a sample of $N$ tokens, we evaluate the goodness-of-fit by means of the test statistic

$$X^2_{N,k} \equiv (\bar{x} - \bar{\mu})^T (\sigma_{ij})^{-1} (\bar{x} - \bar{\mu}), \tag{3}$$

with $\bar{x}$ and $\bar{\mu}$ the vectors

$$\begin{aligned} &(V(N), n_1(N), n_2(N), \ldots, n_k(N)) \\ &(\mathrm{E}[V(N)], \mathrm{E}[n_1(N)], \mathrm{E}[n_2(N)], \ldots, \mathrm{E}[n_k(N)]) \end{aligned} \tag{4}$$

respectively, and $(\sigma_{ij})$ the corresponding covariance matrix (Morrison, 1976). If the model has $a$ parameters, $X^2_{N,k}$ is $\chi^2_{k+1-a}$ distributed. Expressions for the covariances $\sigma_{ij}$ can be found in Good and Toulmin (1956) and in in 't Veld (1984). Note that the test statistic

$$Q_k = \sum_{r=1}^{k} \frac{(n_r(N) - \mathrm{E}[n_r(N)])^2}{\mathrm{E}[n_r]} + \\ + \frac{(n_r(N)^+ - \mathrm{E}[n_r(N)]^+)^2}{\mathrm{E}[n_r(N)^+]} \tag{5}$$

cannot be used. Contrary to what is often assumed in the literature (see e.g. Sichel, 1975, 1986; Muller, 1979), $Q_k$ is not $\chi^2_k$ distributed: $(n_1, n_2, \ldots, n_k, n_k^+)$ should not be confused with $(X_1, \ldots, X_k, X_k^+)$, where $X_1, \ldots, X_k, X_k^+$ are multinomially distributed with parameters $N, \pi_1, \pi_2, \ldots, \pi_k, 1 - \Sigma_{i=1}^k \pi_i$. But while $\Sigma_{i=1}^k X_i + X_k^+ = N$, we have that $\Sigma\, n_r(N) = V(N)$, itself a random variable depending on $N$. In addition, the fact that each $n_r(N)$ has its own variance should be taken into account.

Parameter estimation will be carried out by requiring that $\mathrm{E}[V(N)] = V(N)$ and that $\mathrm{E}[n_1(N)] = n_1$, and by minimization of $X^2_{N,k}$ in case there are more than two parameters. This procedure

ensures that gross departures of the vocabulary size and the vocabulary growth rate are avoided.

## 3. Statistical Models for Word Frequency Distributions

The parametric models to be discussed in this section are the lognormal model (Herdan, 1960; Carroll, 1967), Sichel's (1975, 1986) generalized inverse Gauss-Poisson law and Orlov and Chitashvili's (1982ab, 1983ab) extended generalized Zipf's law. This section presents brief summaries of these models and their rationales, together with an evaluation in terms of the goodness of fit obtained for various word frequency distributions. Section 4 evaluates the rationales from a linguistic perspective, and the predictive accuracy of these models with respect to the theoretical vocabulary size is studied in section 5.

### 3.1. *The lognormal law*

Herdan (1960, 42—58) and Carroll (1967) have argued that word frequency distributions are governed by the lognormal law. Consider the structural token distribution

$$\Psi(\pi) = \sum_i \pi_i \mathrm{I}[\pi_i \leqslant \pi], \tag{6}$$

a distribution characterized by the property

$$\frac{\Psi(\pi_j) - \Psi(\pi_{j-1})}{\pi_j} = n_{\pi_j}, \tag{7}$$

where $\pi_j$ is the first probability greater than $\pi_{j-1}$ and $n_{\pi_j}$ the number of types with probability $\pi_j$. In the case of the lognormal model, the structural token distribution is approximated by the continuous expression

$$\Psi(\pi) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\pi} \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2} dx. \tag{8}$$

We can now write the expressions of the compound Poisson law (Yule, 1944)

$$\mathrm{E}[n_r(N)] = \sum_i \frac{(\pi_i N)^r}{r!} e^{-\pi_i N} \tag{9}$$

$$\mathrm{E}[V(N)] = \sum_i (1 - e^{-\pi_i N}) \tag{10}$$

in the sense of Stieltjes integral as

$$\mathrm{E}[n_r(N)] = \int_0^{\infty} \frac{(\pi N)^r}{r!} e^{-\pi N} \frac{1}{\pi} d\Psi(\pi) \tag{11}$$

$$\simeq \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\infty} \frac{(xN)^r}{x^2 r!} e^{-xN - \frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2} dx$$

$$\mathrm{E}[V(N)] = \int_0^{\infty} (1 - e^{-\pi N}) \frac{1}{\pi} d\Psi(\pi) \tag{12}$$

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

The parameters $\mu$ and $\sigma$ are estimated by solving

$$\mathrm{E}[n_1(N)] = n_1(N)$$
$$\mathrm{E}[V(N)] = V(N). \tag{13}$$

Carroll (1967) develops an algorithm for obtaining estimates of the population mean and variance that does not make use of (11) and (12). Using property (7) of the structural token distribution (6), he considers the distribution of the logarithmic transform $y = \log(\pi)$, obtaining estimates of the number of types $n_\pi$ in the interval $(\log(\pi) - \varepsilon, \log(\pi) + \varepsilon)$ by dividing $\Pr(\log(\pi) - \varepsilon \leqslant \log(\pi) \leqslant \log(\pi) + \varepsilon)$ by $\pi$. By partitioning the area under the normal curve of $\log(\pi)$ corresponding to the interval $(-\infty, 0)$ into a large number of areas $A_i$, followed by summation of the fractions $A_i/\pi_i$, the theoretical vocabulary size $S$ is calculated. Once the areas $A_i$ and the corresponding probabilities $\pi_i$ are fixed, $\mathrm{E}[n_r(N)]$ can be obtained using the compound Poisson law (9). In order to allow comparison with Carroll's (1967) data we have used this algorithm for the analyses reported below. Consequently means and variances in the following discussion should be understood as having been calculated for the lognormal transform $y = \log(\pi)$.

Carroll (1967) is, to our knowledge, the first to have observed that for word frequency distributions sample relative frequencies are biased estimates of population probabilities.

This is clear from the fact that the minimum value of a word probability computed from a sample is $1/N$, where

> $N$ is the size of the sample. There will be a large number of word types in the population that will not appear even once in the sample. The probability that a word type of a given probability will appear once or more in the sample is a function of that probability; only the more frequent words will have very high probabilities of appearing at least once in a sample of moderate size. This fact is the explanation for the tendency of the lognormal plots of moderately sized samples to bend downwards at their lower end . . . . (1967, p. 408)

Khmaladze and Chitashvili (1989) show that this bias is due to the large number of very low frequency types characteristic for lexical frequency distributions, and work out its statistical consequences.[2] The extent of the bias caused by the large number of rare words in most word frequency distributions can be illustrated by comparing the estimates $\hat{\mu}$ and $\hat{\sigma}$ obtained by (13) with the estimates $m$ and $s$ based on the sample relative frequencies. As shown in Table 1, the two kinds of estimates diverge considerably, illustrating the necessity of the estimation procedure developed by Carroll.

TABLE 1
Correct and biased estimates of the parameters of the lognormal law for the Cobuild corpus, Pushkin's 'The captain's Daughter,' and the Dutch derivational suffixes *-je* (*huisje*, 'small house'), *-ing* (*generering*, 'generation'), *-er* (*loper*, 'walker') and *-heid* (*goedheid*, 'goodness').

| | $\hat{\mu}$ | $\hat{\sigma}$ | $m$ | $s$ |
|---|---|---|---|---|
| Cobuild | −3.3220 | 1.0062 | −6.9533 | 7.1189 |
| Pushkin | −3.0290 | 1.0970 | −6.7842 | 4.4401 |
| *-je* | −2.9324 | 0.9382 | −6.2268 | 2.0899 |
| *-ing* | −2.4780 | 0.8055 | −5.9132 | 1.6608 |
| *-er* | −2.1900 | 0.9500 | −5.2636 | 1.6936 |
| *-heid* | −2.0800 | 1.1450 | −4.9102 | 2.7167 |

The highest frequency types tend to appear with somewhat higher frequencies than one would expect on the basis of the lognormal hypothesis. Herdan (1960) seeks to explain this fact by calling attention to the exceptional frequential properties of function words, typically the highest frequency words in the distribution. Removal of the function words from the distribution, he argues, will bring the resulting distribution of content words in line with the lognormal curve. Unfortunately, this solution is somewhat unsatisfactory since it is often only the last two or three highest frequency types that are exceptional in my data. The problem is not related to function words as such — many function words are not exceptional at all. Moreover, a similar upward curvature can be observed for the distributions of morphological categories, distributions in which no function words are involved. The problem is a problem of discretization: modelling a discrete random variable by a continuous one leads to a smooth line where in the discrete case one finds abrupt jumps at the right hand side of the graph. In fact, the lognormal model does not rule out the possibility of a word type having a frequency exceeding the sample size. This illustrates a general property of the models discussed here, namely that they are inaccurate for the study of the highest frequency types. However, since the model may give a fairly accurate characterization of the left hand side of the distribution, and may thus be a useful tool for estimating $S$, it is worthwhile to consider the goodness of fit in some more detail.

In order to assess the goodness of fit of the lognormal model to the Pushkin data, we compared the observed vocabulary $V(N)$ and the numbers $n_r(N)$ of types occurring $r$ times for $r = 1, 2, \ldots, 15$ with the corresponding expected values using (3). The results are somewhat disappointing: $\chi^2 = 38.99$, $q = 0.000366$. For the distribution of written language in the Cobuild corpus (Sinclair, 1987) the fit is even worse: $\chi^2 = 5195.30$, $q = 0.000000$. Although the extremely high $\chi^2$ value may in this case be due to the circumstance that in general it is extremely difficult to obtain acceptable fits for very large samples, we shall see that a reasonable fit can be obtained with the extended generalized Zipf's law. The high $\chi^2$ value obtained for the Cobuild corpus data forces us to conclude that the lognormal model is not the correct distribution here.

Surprisingly, a very good fit is obtained for the Dutch suffix *-heid*, used to coin abstract nouns from adjectives, such as *snelheid*, 'speed,' from *snel*, 'quick.' Here $\chi^2 = 5.94$, $q = 0.97$. This extremely high value of $q$ cannot be attributed wholly to the small size of the distribution ($N =$

2251), since for the Dutch suffix *-er* (e.g. *schrijv-er*, 'writer'), which creates agent nouns from verbs, the $\chi^2$ value equals 37.13 ($q = 0.001$) for only slightly larger $N$ (2345), while for the diminutive suffix *-je* (e.g. *huis-je*, 'small house') we have that $q = 0.06$ for $N = 2580$.

Comparing the $q$ values obtained for the distributions listed in Table 2 with the corresponding vocabulary growth rates $n_1/N$ suggests that there is a positive correlation between goodness of fit and growth rate, such that samples with higher growth rates are more likely to be modelled by the lognormal law than samples with low growth rates. An observation in favor of this tentative correlation concerns the shape of the lognormal curve of the Dutch nominalizing suffix *-ing* ($N = 7881$) shown in Figure 1. Note that after $r = 20$ the token distribution shows a steady upward curvature that does not harmonize well with the lognormal hypothesis. Not surprisingly, the $\chi^2$ value obtained is high ($\chi^2 = 78.45$, $q = 0.000000$). These findings suggest that the lognormal model may be a reasonable model for perhaps literary texts (Pushkin) but certainly not for corpora (Cobuild), for affixes with a high degree of productivity (*-heid*) but not for affixes with a low vocabulary growth rate (*-ing*).

Finally, consider the question in what way the lognormal hypothesis might shed light on the factors shaping word frequency distributions. Carroll (1969), in answer to criticism of, for example, Mandelbrot (1962), that application of the lognormal model to word frequency distributions amounts to 'curve fitting' without any intrinsic motivation, develops the following rationale. Suppose that the choice for a particular vocabulary item $w$ is determined by a series of binary choices, and suppose that the 'choice probabilities' corresponding to each choice constitute a random variable $X$ that is symmetrically distributed around 0.5, each probability $p$ having a complementary probability $1 - p$. The probability of selecting $w$ is now given by

$$\Pr(w) = \prod_{j=1}^{m} X_j, \tag{14}$$

with $m$ the depth of the decision tree. Assuming that $\log(X)$ is $\mathcal{N}(\mu, \sigma^2)$ distributed, $\log(\prod_{j=1}^{m} X_j) = \sum_{j=1}^{m} \log(X_j)$ is lognormally distributed with parameters $m\mu$ and $m\sigma^2$. Carroll (1969) considers in detail possible densities for $X$ for fixed and variable decision path length $m$, obtaining results that suggest that some form of asymptotic lognormal generating function might well give rise to adequate fits to observed data.

This rationale has some intuitive appeal in the case of word frequency distributions obtained for word association experiments, and might be reasonable for continuous text, assuming that the conditions for selecting a particular word change as we move through the text, and including different semantic and grammatical features in the decision tree. Interestingly, this rationale may shed some light on why good fits are obtained for *-heid*, *-je* and perhaps *-er*, while the model appears to be less well suited for dealing with *-ing* or the distribution of monomorphemic nouns in Dutch. Since in the case of *-heid* and *-je* the semantics of the relevant morphological categories are highly

TABLE 2

Parameters, growth rate $n_1/N$, sample size $N$ and goodness of fit statistics for selected word frequency distributions: the lognormal model. Dutch N: monomorphemic nouns (Dutch) in the Eindhoven corpus.

| distribution | $N$ | $\hat{\mu}$ | $\hat{\sigma}$ | $n_1/N$ | $\chi^2$ | $q$ | df |
|---|---|---|---|---|---|---|---|
| *-heid* | 2251 | −2.0800 | 1.1450 | 0.114 | 5.94 | 0.967845 | 14 |
| *-je* | 2580 | −2.9324 | 0.9384 | 0.253 | 22.92 | 0.061521 | 14 |
| *-er* | 2345 | −2.1900 | 0.9500 | 0.093 | 37.14 | 0.000703 | 14 |
| Pushkin | 28471 | −3.0290 | 1.0970 | 0.084 | 38.99 | 0.000366 | 14 |
| Dutch N | 37836 | −2.4395 | 0.8691 | 0.008 | 49.00 | 0.000009 | 14 |
| *-ing* | 7881 | −2.4780 | 0.8055 | 0.038 | 78.45 | 0.000000 | 14 |
| Cobuild | 15713145 | −3.3220 | 1.0062 | 0.000 | 263.77 | 0.000000 | 18 |

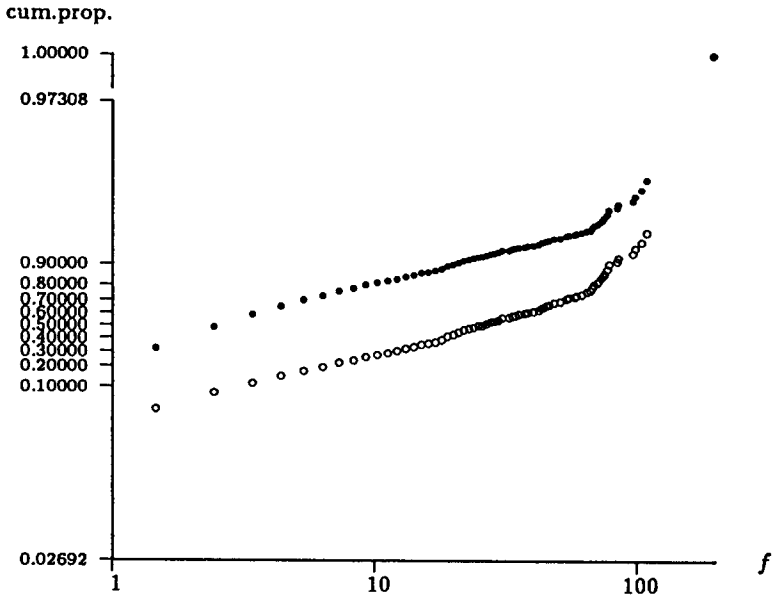


Figure 1. Lognormal plot for *-ing* nominalizations in the Eindhoven corpus. The lower curve represents the token distribution, the upper curve the type distribution. The horizontal axis is scaled logarithmically, the vertical axis is scaled proportional to the error function. Under the lognormal hypothesis, the two curves should show up as parallel straight lines.

transparent, the choice for a particular derived item can be understood as being conditioned by a particular node in the decision tree, in the sense that almost all abstract nouns or diminutives in the language belong to these morphological categories, which therefore can be viewed as constituting natural branches of the decision tree. Assuming that such natural branches are isomorphic with the tree itself, we have that these categories should again be lognormally distributed. Of course, many branches in the decision tree will be highly heterogeneous with respect to the morphological constituency of their elements. The low $q$ value obtained for *-ing* may well be correlated with the fact that this nominalizing suffix is substantially affected by type and token blocking (van Haeringen, 1971; Rainer, 1988) and to some extent by loss of semantic transparency, so that there is no branch in the decision tree with only or predominantly formations in *-ing*. In the light of Carroll's rationale, such distributions must be considered as artificial groupings of lexical items rather than as natural semantic sets in the decision tree. If this line of reasoning is correct, obtaining a good lognormal fit to a morphological frequency distribution would by itself be a litmus test for productivity.

3.2. *The generalized inverse Gauss-Poisson law*

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

$$G(\pi) = \sum_i \mathrm{I}[\pi_i \leqslant \pi]. \tag{15}$$

Given $G(\pi)$, the expressions (9) and (10) can be rewritten in the sense of Stieltjes integral as

$$\mathrm{E}[V(N)] = \int_0^\infty (1 - e^{-N\pi}) dG(\pi) \tag{16}$$

$$\mathrm{E}[n_r(N)] = \int_0^\infty \frac{(N\pi)^r}{r!} e^{-N\pi} dG(\pi). \tag{17}$$

Writing $\psi(\pi)$ for $G'(\pi)/S$, the (normalized) probability of selecting at random a type $i$ for which $\pi_i \leqslant \pi$, Sichel (1975, 1986), following up a suggestion by Good (1953), reports that excel-

lent results are obtained for

$$\psi(\pi) = \frac{(2/bc)^{\gamma}}{2K_{\gamma}(b)}\, \pi^{\gamma-1} e^{\left(-\frac{\pi}{c}-\frac{b^2 c}{4\pi}\right)}, \qquad (18)$$

where $K_{\gamma}(b)$ is the modified Bessel function of the second kind of order $\gamma$ and argument $b$. Hence we have

$$\mathrm{E}[n_r(N)] \simeq S \int_0^{\infty} \frac{(N\pi)^r}{r!}\, e^{-N\pi} \psi(\pi) d\pi \qquad (19)$$

$$\mathrm{E}[V(N)] \simeq S \int_0^{\infty} (1 - e^{-N\pi}) \psi(\pi) d\pi, \qquad (20)$$

Given the mean of the distribution $\psi(\pi)$,

$$\mathrm{E}[\pi] = \frac{bc}{2}\, \frac{K_{\gamma+1}(b)}{K_{\gamma}(b)}, \qquad (21)$$

and using

$$\mathrm{E}[\pi] = \frac{1}{S} \sum_{i=1}^{S} \pi_i = \frac{1}{S}, \qquad (22)$$

$S$ can be determined as the reciprocal of $\mathrm{E}[\pi]$. Solving (20) leads to

$$\mathrm{E}[V(N)] =$$

$$\frac{2}{bc}\, \frac{K_{\gamma}(b)}{K_{\gamma+1}(b)} \left[ 1 - \frac{K_{\gamma}(b\{1+cN\}^{1/2})}{(1+cN)^{\gamma/2} K_{\gamma}(b)} \right]. \qquad (23)$$

Let

$$\alpha(r, N) = \frac{\mathrm{E}[n_r(N)]}{\mathrm{E}[V(N)]} \qquad (24)$$

denote the ratio of the number of types with frequency $r$ in the sample to the number of different types in the sample. This ratio can be rewritten as

$$\alpha(r, N) = \frac{1}{(1-\theta_N)^{-\gamma/2} K_{\gamma}(\alpha_N (1-\theta_N)^{1/2}) - K_{\gamma}(\alpha_N)} \cdot \frac{(0.5\alpha_N \theta_N)^r}{r!}\, K_{\gamma+r}(\alpha_N), \qquad (25)$$

where $\alpha_N = b\sqrt{(1+cN)}$ and $\theta_N = cN/(1+cN)$. Note that the parameters $\alpha_N$ and $\theta_N$, introduced for simplification, are functions of the sample size $N$, while the parameters $b$, $c$ and $\gamma$ are population invariants. As before, parameters are estimated by requiring

$$\mathrm{E}[n_1(N)] = \alpha(1, N)\mathrm{E}[V(N)] = n_1(N)$$
$$\mathrm{E}[V(N)] = V(N). \qquad (26)$$

Simplified expressions can be obtained when $\gamma$ is fixed a priori at $-0.5$, in which case (26) completely determines all parameters. When $\gamma$ is free, it is chosen such that the $\chi^2$ value is minimized while satisfying (26). Note that although a different structural distribution is involved, Sichel's model avoids direct estimation of population probabilities on the basis of sample relative frequencies $f_i/N$ in the same way as Carroll's lognormal model. Of course, both models cannot avoid using sample data to estimate the structural distribution, and the precision with which the population parameters are estimated will depend on the extent to which sample parameters such as $n_1(N)$ and $V(N)$ deviate from the corresponding expectations.

Table 3 summarizes the results obtained with this model for a number of word frequency distributions. No accurate fit can be obtained for the written language of the Cobuild corpus. In fact, the best fit (in the sense of $\chi^2$) has a point of inflection at $r = 2$ that is absent in the data. Evidently, the model thinks that the rare types in the distribution should be nearly exhausted, contrary to fact. Interestingly, we have found that the low value of $q$ obtained for the Cobuild data is not due to the size of the sample. When smaller random samples taken by sampling without replacement are considered of 30000 or 1000000 tokens, the minimal $\chi^2$ values obtained remain unacceptably high. This suggests informally that either no satisfactory fit is obtained for whatever sample size, or that a reliable fit is obtained, in which case the parameters $\gamma$, $b$ and $c$ are to all practical purposes independent of the sample size $N$.

An important property of Sichel's model is that it allows for the possibility that the mode of the frequency spectrum is situated at some $r > 1$. Frequency distributions with this characteristic are typical of 'pathological language' (Mandelbrot, 1962) in the case of text counts, and of unproductive morphological categories and sets of simplex items as they occur in large corpora (Baayen,

TABLE 3
Parameters, sample size and goodness of fit statistics for selected word frequency distributions: the generalized inverse Gauss-Poisson distribution. Dutch N: monomorphemic nouns (Dutch) in the Eindhoven corpus.

| | $N$ | $\gamma$ | $b$ | $c$ | $\chi^2$ | $q$ | df |
|---|---|---|---|---|---|---|---|
| *-heid* | 2251 | −0.725 | 0.035341 | 0.084489 | 7.53 | 0.8729 | 13 |
| *-je* | 2580 | −0.50 | 2.859e-7 | 0.005644 | 19.95 | 0.0965 | 13 |
| *-er* | 2345 | −0.36 | 0.001963 | 0.016792 | 10.38 | 0.6628 | 13 |
| *-ing* | 7881 | −0.40 | 0.109813 | 0.009787 | 9.38 | 0.7436 | 13 |
| Dutch N | 37836 | −0.35 | 0.081843 | 0.007995 | 12.87 | 0.4577 | 13 |
| Pushkin | 28471 | −0.85 | 0.034795 | 0.022650 | 24.13 | 0.1409 | 18 |
| Cobuild | 15713145 | −0.1 | 0.030076 | 0.000353 | 920.38 | 0.0000 | 18 |

1989). Unfortunately, the grouped frequency distributions with shifted modes that have come to my attention are subject to so much noise that it is extremely difficult to trace whether a particular theoretical model is valid.

One serious drawback of Sichel's inverse Gauss-Poisson distribution is that it has no rationale. From a linguistic point of view, the absence of a rationale brings application of the model uncomfortably close to 'curve fitting,' however useful that may be when one is interested in estimating $S$.

### 3.3. *The extended generalized Zipf law*

Orlov and Chitashvili (1982ab, 1983ab) develop a model that is a generalization of Zipf's law. Recalling the notation $\alpha(r, N)$ for the ratio of the number of types occurring $r$ times in a sample of size $N$ to the total number of types occurring in that sample, the generalized Zipf's law states that for some sample size $Z$

$$\alpha(r, Z) = \frac{\int_0^\infty \frac{(\pi Z)^r}{r!} e^{-\pi Z} dG(\pi)}{\int_0^\infty (1 - e^{-\pi Z}) dG(\pi)} = \frac{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^\alpha}{(1+y)^{r+1}(1+y)^\beta} dy}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-1}}{(1+y)^{\beta+1}} dy}. \quad (27)$$

Note that $Z$ does not appear in the right hand side of (27). In fact, the sample size $Z$ is uniquely determined by the fact that (27) holds. Conversely, larger or smaller samples from the same population will not be adequately characterized by (27). By making use of the non-parametric extrapolation result (Good and Toulmin, 1956; Kalinin, 1965)

$$\mathrm{E}[n_r(N)] = \sum_{j \geq r} \mathrm{E}[n_j(Z)] \binom{j}{r} t^r (1-t)^{j-r}, \quad (28)$$

where $t = N/Z$, the following expressions for the expectations of $V(N)$ and $n_r(N)$ can be obtained for what we will refer to as the extended generalized Zipf's law:

$$\mathrm{E}[n_r(N)] = C(Z, \alpha, \beta, \gamma) t^r \int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^\alpha}{(t+y)^{r+1}(1+y)^{\beta+1}} dy \quad (29)$$

$$\mathrm{E}[V(N)] = C(Z, \alpha, \beta, \gamma) t \int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-1}}{(t+y)(1+y)^\beta} dy, \quad (30)$$

where

$$C(Z, \alpha, \beta, \gamma) = \frac{V^{(Z)}}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-1}}{(1+y)^{\beta+1}} dy}. \quad (31)$$

The expected number of types for the sample size $Z$, denoted by $V^{(Z)}$, is estimated by

$$V^{(Z)} = Z \frac{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-1}}{(1+y)^{\beta+1}} dy}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1} y^{\alpha-2}}{(1+y)^{\beta+Z\hat{p}^*}} \left[(1+y)^{Z\hat{p}^*} - 1 - \frac{Z\hat{p}^* y}{1+y}\right] dy}, \quad (32)$$

with $\hat{p}^*$ denoting the maximal sample relative frequency, a population constant for not too small $N$. This completes the formal description of this model.

The way in which the extended generalized Zipf's law is obtained can be justified by considering the so-called triangle scheme (or scheme of series) experiment model. For example. the Poisson distribution $\Pi(\lambda)$ is a good approximation to the binomial distribution when $N \to \infty$ and $\pi \to 0$. For fixed $\pi_k$, a particular Poisson approximation $\Pi_k(\lambda_k)$ is appropriate only for some single value of $N$, since in general $\lambda = N\pi$. Suppose $\Pi_k(\lambda_k)$ gives a good fit for $N = Z$, then for $N \neq Z$ we have that $\Pi_k(t\lambda_k)$ is valid for sample size $N$ when $t = N/Z$. This is the way in which the parameter $t$ should be understood in the case of the generalized Zipf's law, which should not be interpreted as some limiting distribution for $N \to \infty$ but as a 'limiting' distribution for $N \to Z$.

We consider the goodness-of-fit for the model with the parameter $\gamma$ fixed at unity. Note that $\gamma$ completely disappears from (27), so that we are in fact dealing with a three parameter model, the extended counterpart of the Waring-Herdan-Muller distribution (Muller, 1979). In this case $S$ is given by

$$S = \frac{V^{(Z)}\beta}{\beta - \alpha}. \quad (33)$$

Table 4 summarizes the results obtained. No satisfying fits ensued for the suffixes *-je* and *-er*. In the case of *-er*, it appears that the extended Waring-Herdan-Muller model is inadequate. Possibly, allowing the parameter $\gamma$ to assume values other than unity may lead to better results. In the case of *-je*, however, the failure to obtain a good fit can be traced to the expression for $V^{(Z)}$ (32), which is computationally unsatisfactory for small $\alpha$ and $\beta$. In fact, machine precision errors give rise to theoretically impossible negative values for $V^{(Z)}$ precisely in the area of parameter space where a good fit for *-je* is most likely to be found. For the other morphological distributions good fits are obtained. Note that a satisfactory fit was obtained for the Pushkin data with $\alpha = 1$, in which case the model simplifies to the extended version of the Yule-Simon model, as we shall see below. Finally note that the fit obtained for the Cobuild data ($q = 0.0016$ for $r = 1$—$40$) is not unreasonable for a 15.7 million corpus.

We now turn to the rationale of the generalized Zipf's law, a model subsuming a range of word frequency laws that appear as limiting distributions of linguistically motivated stochastic processes. In its simplest form, with $\alpha = \beta = \gamma$ fixed at unity, $\alpha(r)$

TABLE 4

Sample size, parameters and goodness-of-fit statistics for selected word frequency distributions: the extended generalized Zipf's law with $\gamma = 1$. Dutch N: monomorphemic nouns (Dutch) in the Eindhoven corpus.

| | $N$ | $\alpha$ | $\beta$ | $t$ | $\chi^2$ | $q$ | df |
|---|---|---|---|---|---|---|---|
| *-je* | 2580 | 0.8675 | 0.7280 | 1.050 | 154.36 | 0.0000 | 13 |
| *-er* | 2345 | 0.5700 | 3.3170 | 0.006 | 66.67 | 0.0000 | 13 |
| *-ing* | 7881 | 0.8500 | 2.4126 | 0.500 | 4.84 | 0.9786 | 13 |
| *-heid* | 2251 | 0.8000 | 8.1121 | 0.010 | 8.93 | 0.7779 | 13 |
| Dutch N | 37836 | 0.8500 | 1.8052 | 3.000 | 15.71 | 0.2651 | 13 |
| Pushkin | 28411 | 1.0000 | 5.5420 | 0.057 | 24.18 | 0.1491 | 18 |
| Cobuild | 15713145 | 0.9100 | 2.9520 | 12.000 | 68.75 | 0.0016 | 38 |

reduces to Zipf's law (Zipf, 1935):

$$\alpha(r) = \frac{1}{r(r+1)}. \tag{34}$$

Particular choices for $\alpha$, $\beta$ and $\gamma$ lead to the following generalizations:

1. Yule-Simon (Simon, 1955) ($\alpha = \gamma = 1$)

$$\alpha(r) = \frac{\beta}{(r+\beta-1)(r+\beta)}, \tag{35}$$

2. Waring-Herdan-Muller (Herdan, 1960, 1964; Muller, 1979) ($\gamma = 1$)

$$\alpha(r) = \frac{\Gamma(\beta+1)\alpha}{\Gamma(\beta+1-\alpha)} \cdot \frac{\Gamma(r+\beta-\alpha)}{\Gamma(r+\beta+1)}, \tag{36}$$

3. Karlin-Rouault (Rouault, 1978) ($\beta = 0$, $\gamma = 1$)

$$\alpha(r) = \frac{\alpha\Gamma(r-\alpha)}{\Gamma(1-\alpha)\Gamma(r+1)}, \tag{37}$$

4. Zipf-Mandelbrot (Mandelbrot, 1962) ($\alpha = \beta = 1$)

$$\alpha(r) = \frac{1}{r^{\gamma}} - \frac{1}{(r+1)^{\gamma}}. \tag{38}$$

Let us briefly review the rationales for these models.

The Yule-Simon model appears as the limiting form (under the condition of equilibrium) of a stochastic process that is constructed to reflect the way in which an author writes a text. It explores the consequences of assuming (i) that there is a constant probability $\alpha$ of using a new type in the text, and (ii) that the probability of re-using any of the types that already occurred $r$ times in the text is proportional to $rn_{r,N}$. This is equivalent to fixing the probability of any particular type $i$ for which $f_{i,N} = r$ proportional to the frequency $f_{i,N}$. Thus we have that the probability of selecting type $i$ at sampling stage $N$ is given by

$$p_{i,N} = \mathrm{I}[f_{i,N} > 0](1-\alpha)\frac{f_{i,N}}{N} + \mathrm{I}[f_{i,N} = 0]\alpha. \tag{39}$$

Lánský and Radil-Weiss (1980) discuss a generalization of Simon's original scheme by allowing the re-use of any type that has already appeared $r$ times to be some function $\phi$ of $n_r$. Rewriting $\phi$ in terms of the probability of selecting a particular item $i$ for which $f_{i,N} = r$ we obtain

$$p_{i,N} = \mathrm{I}[f_{i,N} > 0](1-\alpha)\phi_{i,N}(f_{i,N}) + \\ + \mathrm{I}[f_{i,N} = 0]\alpha. \tag{40}$$

We may construct $\phi$ as a linear function of $f_{i,N}$:

$$\phi_{i,N}(f_{i,N}) = C_N^{-1}\left(a_i + b_i \frac{f_{i,N}}{N}\right), \tag{41}$$

with $C_N$ the appropriate normalizing factor and $a_i$ and $b_i$ varying for each type $i$. In its simplest form, $a_i = a_j$, $b_i = b_j$ for all $i, j$, it can be shown that $\alpha(r)$ can be expressed as (36) (Khmaladze and Chitashvili, 1989). Thus the Waring-Herdan-Muller law appears as a generalization of Simon's model.

The Karlin-Rouault distribution appears as the limiting form in the Markov scheme for generating words as strings of letters. Note that the Karlin-Rouault distribution is a special case of the Waring-Herdan-Muller model ($\alpha$ is free, $\beta$ is fixed at 0). Interestingly, the Karlin-Rouault law defines the prototypical LNRE distribution, in that there is a formal proof that the law of large numbers is not valid for distributions governed by (37) (Khmaladze and Chitashvili, 1989).

The Zipf-Mandelbrot law is obtained when assumptions concerning optimization of cost of coding and information transmission are added to the Markovian word formation scheme.

While the generalized Zipf's law itself is supported by a series of well-motivated, although undoubtedly highly simplified, rationales, we are still left with the question of how to interpret and motivate the parameter $t$ of the extended generalized Zipf's law. Orlov (1983a, b) suggests that the sample size $Z$ defines an optimal frequential balance for literary texts. For instance, in the case of Pushkin's *The Captain's Daughter* ($t = 0.057$, $\alpha = 1$, $\beta = 5.542$, $q = 0.1491$), he would argue that the Yule-Simon model describes the virtual size of the text, a text size not reached by far in this relatively short novel, but nevertheless a sample size that the author would have considered as ideal for a larger novel on the same subject. More generally, Orlov claims that rich texts are characterized by $t <= 1$, and that poor or repetitive texts have $t \gg 1$. He predicts that short stories

will show up with rather small values of $t$, while well-written voluminous novels will reach completion at approximately the characteristic sample size $Z$. Conversely, long winded novels, as well as large corpora, are predicted to show up with $t$ values substantially larger than unity. These predictions are born out for our data. For instance, the Cobuild distribution requires $t = 12$ where Pushkin's novel has $t = 0.057$. A similar inversion with respect to the value of $t$ can be observed for productive versus unproductive morphological categories: for productive *-heid* $t$ is small (0.010), for 'unproductive' monomorphematic nouns $t = 3.0$. This suggests that $t$ appears as a parameter of lexical richness c.q. productivity.

Having obtained an interpretation for $t$, we may proceed to inquire what factors necessitate its introduction. Since $t$ specifies the distance a particular distribution is removed form the sample size at which the generalized Zipf's law is valid, it can be viewed as a measure of the extent to which the rationale of the model is a realistic one. Perhaps the most important property of these rationales is that they are valid for limiting distributions for $N \rightarrow \infty$, often under conditions of equilibrium. Since these conditions are not met by empirical distributions, the introduction of $t$ serves to allow 'ideal' theoretical limiting distributions obtained under simplified assumptions to describe frequency distributions at particular stages of their development through (sampling) time.

## 4. Morphology and Semantics

Although the rationales discussed above give some indication of the kind of factors that shape the grouped frequency distribution, it is fruitful to return to the rank-frequency distribution to consider in some more detail how semantic and morphological factors codetermine the 'morphology' of the rank-frequency distribution. This will serve as a point of departure for evaluating the rationales discussed in section 3.

The problems at hand are best introduced with reference to Figure 2. The left hand plot shows the rank-frequency distribution of monomorphematic content words in a 1,000,000 sample of Dutch. The right hand graph summarizes the distribution of all types in this sample, including function words and morphologically complex formations. The question with which we will be concerned is how to account for the differences between the two curves. None of the rationales for word frequency distributions discussed is of any help. Simon's stochastic process is indifferent to the properties of its items, and Mandelbrot's Markovian source for words as strings of phonemes does not take morphological structure into account. It is also unclear in what way Carroll's rationale for the lognormal model might be of relevance here.

Taking up the issue of morphological structure first, recall that we have considered two kinds of word frequency distributions, distributions of running text and distributions of morphological

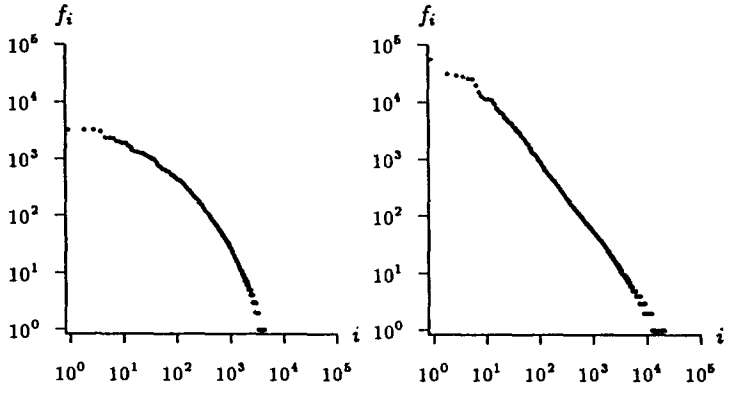


Figure 2. Rank-frequency curves for Dutch stems. The left hand graph presents the distribution of monomorphemic content words, the right hand graph the complete distribution, including function words and morphologically complex formations.

categories. The analysis of the frequential characteristics of morphological categories reveals that each category has its own (conditional) growth rate and theoretical vocabulary size, depending on the productivity and extent of use of the category. Within texts as wholes morphological categories again appear at different rates. From the textual point of view, the growth rate $\mathscr{P}_V(N)$ of the vocabulary as a whole,

$$\mathscr{P}_V(N) \equiv \frac{\mathrm{E}[n_1(N)]}{N} \tag{42}$$

can be analyzed as the sum of the (non-conditional) growth rates

$$\mathscr{P}_{c_i}(N) \equiv \frac{\mathrm{E}[n_{c_{i,1}}(N)]}{N}$$

of the individual morphological categories $c_i$ in the language:

$$\mathscr{P}_V(N) = \sum_i \frac{\mathrm{E}[n_{c_{i,1}}(N)]}{N}. \tag{43}$$

The contribution of morphology to $\mathscr{P}_V(N)$ is substantial: for the Cobuild data on written English the morphologically complex types occurring once represent 64.39% of all types occurring once only, with the contribution of once occurring compounds being seriously underestimated due to compounds with intervening space characters in the orthography not having been recognized as such in the CELEX analysis. The difference in the tails of the distributions of Figure 2 can therefore be traced to the substantial influx of morphologically complex words. Evidently, for a rationale for word frequency distributions to be acceptable from a linguistic point of view, the role of morphology should be taken into account.

We now turn to the divergence between the two curves of Figure 2 for the highest ranks *i*. Recall that none of the models discussed above has anything to say about the frequential behavior of these types. Nevertheless, this behavior remains of interest, the more so since Mandelbrot's law (1) explicitly deals with the systematic departure of the highest frequency types from Zipf's law by means of the parameter *B*. Mandelbrot derived (1) invoking external principles such as 'optimal coding' and 'maximalization of information transmission.' Miller (1957) criticized these external principles as 'straining one's credulity,' showing that (1) appears under the assumption of random spacing for the case in which all letters are equiprobable. Rouault (1978), however, has shown that the limiting form of any Markovian source for word frequency distributions is given by (37) and not by (38) or (34). From this point of view, it is profitable to consider whether Mandelbrot's external principles of minimization of cost of coding and maximalization of information transmission might not be supported by language-internal evidence. An informal suggestion in this direction is developed in Baayen (1991a), where the density[3] structure of the lexicon is used as a criterion for evaluating the explanatory value of models for word frequency distributions. Interestingly, a Markovian source for words as strings of phonemes or letters gives rise to word distributions with density effects (Nusbaum, 1985). Unfortunately, the frequency-density correlation is not modelled correctly, the density effects emerging in distorted form. This can be traced to the unnatural surplus of hapax legomena appearing in this word formation scheme. Hence some way of enforcing a more intensive use of the word types that have already appeared in the generation process is required. Since the Karlin-Rouault law and the Yule-Simon law both appear as special limiting forms of the Waring-Herdan-Muller law, one possibility that suggests itself is to combine a Markovian word generator with a stochastic process of the kind suggested by Simon. The Markovian word generator can be thought of as defining a probability distribution that reflects the relative ease with which (monomorphemic) words can be pronounced by the human vocal tract, while the Simonian stochastic process can be interpreted as simulating factors pertaining to language use, relatively independently of the pronounceability of these words. Baayen (1991a) reports a computer simulation in which an initial frequency distribution ($f_i$) was obtained by means of a Markov process generating a large subset of phonotactically legal (possible) Dutch words. This

initial distribution ($f_i$) served as the starting point for a stochastic process defined by

$$p_{i,N} = \mathrm{I}[f_{i,N} > 0] \frac{(\alpha - 1)}{C_N} \frac{rn_{r,N}}{N} \log\left(\frac{rn_{r,N}}{N}\right) + \\ + \mathrm{I}[f_{i,N} = 0]\alpha \frac{q_i}{\Sigma_j q_j \mathrm{I}[f_{j,N} = 0]} \quad (44)$$

where $C_N$ is the normalizing constant

$$C_N = -\sum_r \frac{rn_{r,N}}{N} \log\left(\frac{rn_{r,N}}{N}\right)$$

and ($q_i$) the initial (Markovian) probability distribution of types. Qualitatively satisfying results were obtained for the distribution of monomorphemic content words of Dutch summarized in Figure 2, both with respect to the overall shape of the rank-frequency curve as with respect to the frequency-density correlation.

The motivation for choosing the entropy function

$$H_{r,N} = -\frac{rn_{r,N}}{N} \log\left(\frac{rn_{r,N}}{N}\right) \quad (45)$$

for Lánský and Radil-Weiss's (1980) $\phi$ function is of main interest here. It is a semantically motivated means to obtain a better trade-off in the distribution between maximalization of information transmission and optimization of the cost of coding this information. In order to minimize the cost of coding, formalizing the cost of coding for word $y$ as $C(y) = -\log(\Pr(y))$, the highest frequency words should be re-used. In order to maximize information transmission, on the other hand, the lowest frequency types should be re-used ($H_{r,N}$ is maximal for uniformly distributed $rn_{r,N}/N$). Thus we have two conflicting requirements, which balance out in favor of a more intensive use of the lower and intermediate frequency ranges given $H_{r,N}$. Interestingly, $H_{r,N}$ is motivated on language-internal grounds. The use of $H_{r,N}$ implies that higher frequency words contribute less to the average amount of information than might be expected on the basis of their relative frequencies. This harmonizes well with the greater number of (shades of) meaning that higher frequency words are known to have (see e.g. Reder, Anderson and Bjork, 1974; Paivio, Yuille and Madigan, 1968). Since a greater number of meanings implies an increased contextual dependency for interpretation, the amount of information contributed by such types out of context (under conditions of statistical independence) is less than what would be predicted on the basis of their relative frequencies. The results obtained suggest informally that the semantics of the higher frequency words codetermine the shape of the head of the rank-frequency distribution of (monomorphemic) content words in Figure 2. For formal modelling of this semantic effect the limiting properties of (44) should be studied, or preferably, in order to avoid the unnatural constant vocabulary growth rate $\alpha$ given with (44), the stochastic process defined by

$$p_{i,N} = \frac{1}{C'_N} \left\{ \mathrm{I}[f_{i,N} = 0] q_i - \right. \\ \left. - \mathrm{I}[f_{i,N} > 0] \frac{rn_{r,N}}{N} \log\left(\frac{rn_{r,N}}{N}\right) \right\}, \quad (46)$$

with $C'_N$ the appropriate normalizing factor, as suggested by Khmaladze and Chitashvili (1989) in general for dynamic models of this kind. Note that the parameter $\alpha$ has been eliminated, and that the probability of using new words decreases with increasing $N$, as required.

Finally, note that the introduction of function words into the distribution greatly reduces the downward curvature at the head of the rank-frequency distribution, as can be seen when the two graphs of Figure 2 are compared. Interestingly, function words are generally semantically well-defined, implying that they should not be governed by (45). In turn, this leads to the prediction that they should appear with higher frequencies than content words, as is indeed the case.

In sum, we have argued that the existing rationales for word frequency distributions are too simplistic from a linguistic point of view in that they neglect the semantic and morphological factors which codetermine the shape of word frequency distributions.

## 5. Estimating the Theoretical Vocabulary Size

The three parametric models discussed in the present paper all allow the theoretical vocabulary size to be estimated. Since there are instances where each model is found to give a reasonable fit, we select the model for which the $q$ value is maximal for the estimation of $S$, this being the model which has the maximum likelihood of being correct. Selection according to the criterion of maximum $q$ shows (see Table 5) that the lognormal model has the weakest coverage, the other two models being roughly equivalent as to their range of applications.

Although the fits obtained are quite good, it is of interest to ascertain whether the predictions about $S$ are reliable. First consider the morphological categories listed in the first half of Table 5, for which $\hat{S}$ is calculated on the basis of the Dutch Eindhoven corpus (600,000 tokens, Uit den Boogaart, 1975). When we compare $\hat{S}$ with the number of types $V_l$ listed in the CELEX database (which combines counts for a 42 million corpus with information taken from the van Dale dictionary (van Sterkenburg and Pijnenburg, 1984), we observe substantial differences. In the case of *-heid, -je* and *-er* we seem to be dealing with overestimation. In the light of their high degree of productivity, however, it may well be that the dictionary-based estimates are too low — it is not sensible nor feasible for a dictionary to list all possible (and mostly completely predictable) formations with these suffixes. Note that the diminutive suffix *-je*, which is extremely productive in Dutch, appears with a value for $\hat{S}$ that approximates 'infinity,' the number of possible types predicted on the basis of recursion by the calculus of morphology for productive affixation in general.

Unfortunately, the number of types $S$ is seriously underestimated in the case of *-ing*[4] and monomorphemic nouns. This discrepancy can be traced to three factors. First, due to its smallish size, the Eindhoven corpus covers only a small range of the topics that are discussed in the language at large. Hence the estimates of $S$ may be accurate only for the kind of language used to discuss the relatively limited range of topics that appear in the Eindhoven corpus. Second, the dictionary count overestimates the number of types available to individual speakers. Generally, speakers are versed in only a limited number of fields of expertise. Their vocabularies will only contain those types that pertain to the fields they have mastered. When the dictionary count is used to estimate $S$, it is tacitly assumed that the 'ideal' speaker is knowledgeable in all these technical areas, contrary to fact. Hence it may be unrealistic to compare estimates based on the Eindhoven corpus with the dictionary counts, especially so in the case of monomorphemic nouns. Third, the possibility that the fundamental but irrealistic assumption underlying all of the models discussed in the present paper, namely that words occur independently in texts, introduces a bias. Word types are re-used with more than chance frequency in texts. Once a particular topic is

TABLE 5

Goodness of fit $q$, sample vocabulary size $V$ and estimates of the theoretical vocabulary size $S$ for the lognormal law (L), the generalized inverse Gauss-Poisson law (GP) and the extended generalized Zipf's law (Z). The last column lists external estimates $V_l$ of the theoretical vocabulary size.

| distribution | model | $q$ | df | $V$ | $\hat{S}$ | $V_l$ |
|---|---|---|---|---|---|---|
| *-heid* | L | 0.97 | 14 | 466 | 3888 | 2399 |
| *-je* | GP | 0.10 | 13 | 1031 | 1239156496 | — |
| *-er* | GP | 0.66 | 13 | 460 | 1620 | 1342 |
| *-ing* | Z | 0.98 | 13 | 942 | 1772 | 2897 |
| Dutch N | GP | 0.46 | 13 | 1495 | 1876 | 4008 |
| Pushkin | Z | 0.15 | 18 | 4783 | 14590 | 21197 |
| Cobuild | Z | 0.05 | 18 | 29086 | 30920 | 31101 |

broached, the vocabulary items related to that topic have a substantially raised probability of being re-used. This has the effect of lowering the estimated growth rate of the vocabulary and introducing a bias in the estimation of $S$. Hence $S$ as estimated by the models studied here should be interpreted as a lower bound for the theoretical vocabulary size.[5]

Finally, consider the Pushkin and Cobuild data in the second half of Table 5. For Pushkin's novel, the fact that $\hat{S} \ll V_l$, where $V_l$ is based on a count of types in Pushkin's complete works (Orlov, 1983b, p. 204), should probably be traced to the difficulty of generalizing to an author's vocabulary on the basis of a single text belonging to one particular literary genre only. As to the Cobuild data, it is interesting to observe that a 15.7 million word count allows a reasonable prediction of the number of lemmas available in the CELEX database.[6]

The results obtained illustrate a simple methodological point, namely that the assumptions underlying a statistical model should really be satisfied if it is to be a reliable tool. In the present case, the mathematically convenient but linguistically unrealistic assumption of statistical independence gives rise to the paradoxical situation that, even though excellent fits are obtained, the theoretical vocabulary size need not be estimated accurately. Although a lot of progress has been made in the area of word frequency distributions since Zipf's early studies, the main challenge for future research in this area is to construct linguistically less naive models that do not build on the unrealistic assumption that in language words appear at random.

## Acknowledgements

The author is indebted to Rezo Chitashvili and Bert Hoeks for many stimulating discussions on the topics of this paper.

## Notes

[1] Non-parametric methods for obtaining estimates of the theoretical vocabulary size $S$ on the basis of the grouped frequency distribution are developed in Good and Toulmin (1956), Efron and Thisted (1976), Kalinin (1965) and in 't Veld (1984). Unfortunately, the expressions obtained for $S$ do not lend themselves to empirical calculation, which is the reason that this paper focusses on parametric models.

[2] Khmaladze and Chitashvili (1989) present a detailed analysis of distributions with Large Numbers of Rare Events. They show that theoretical LNRE distributions can be defined for which the law of large numbers is not valid, in that sample relative frequencies cannot be used to estimate population probabilities. To all practical purposes, the same holds for many empirical word frequency distributions, even though the mathematical conditions defining the LNRE property are not rigorously met.

[3] Defining a *neighbor* of a target word $i$ as word that differs in exactly one phoneme (or letter) from $i$, it can be observed (Landauer and Streeter, 1973) that higher frequency words have more neighbors than lower frequency words, and that higher frequency words have higher frequency neighbors than lower frequency words. These density effects are weak but significantly present.

[4] Interestingly, *-ing* has been listed more exhaustively than *-er*. A count of types in the 42 million INL corpus available under CELEX reveals 842 types in *-er* and 2036 in *-ing*. Comparing this with the 1342 and 2897 types found in the dictionary, it appears that the types in *-er* in the corpus represent 62.7% of the types in the dictionary. For *-ing* the corresponding percentage is 70.3%. The difference in coverage is significant ($Z = 4.85$).

[5] Conversely, the interpolated values of $V$ for $N' < N$ tend to be too large. The same problem has been observed for Muller's (1977) non-parametric reduction method, which is based on the binomial probability distribution (see e.g. Brunet, 1978; Ratkowsky, 1988; Martin, 1970). Interestingly, the parametric models discussed in the present paper give rise to interpolation curves that are virtually indistinguishable from those obtained on the basis of Muller's technique, provided that the fit to the grouped frequency distribution is sufficiently accurate.

[6] The CELEX database contains all lemmas found in the *Longman Dictionary of Contemporary English*, London: Longman, 1978, and in the *Oxford Advanced Learner's Dictionary of Current English*, Oxford, OUP, 1974.

## References

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

Baayen, R.H., and Lieber, R. "Productivity and English Derivation: A Corpus Based Study." *Linguistics*, 29 (1991), 801—43.

Baayen, R.H. "A Stochastic Process for Word Frequency Distributions." In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.* Ed. D.E. Appelt. 1991 (a), pp. 271—78.

Baayen, R.H. "A Quantitative Approach to Morphological Productivity." In *Yearbook of Morphology 1991.* Eds. G.E. Booij and J. van Marle. Dordrecht: Kluwer, 1991 (b), 109—49.

Bolinger, D.L. "On Defining the Morpheme." In *Forms of English. Accent, Morpheme, Order.* Ed. D.L. Bolinger. Cambridge, MA: Harvard University Press, 1948, pp. 183—89.

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

Carroll, J.B. "On Sampling from a Lognormal Model of Word Frequency Distribution." In *Computational Analysis of Present-Day American English.* Eds. H. Kučera and W.N. Francis. Providence: Brown University Press, 1967, pp. 406—24.

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

Efron, B., and Thisted, R. "Estimating the Number of Unseen Species: How many Words did Shakespeare Know?" *Biometrika*, 63 (1976), 435—47.

Good, I.J. "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika*, 40 (1953) 237—64.

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

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

Haeringen, C. B. van "Het Achtervoegsel -ing: Mogelijkheden en Beperkingen." *De Nieuwe Taalgids*, 64 (1971), 449—68.

Harwood, F.W., and Wright, A.M. "Statistical Study of English Word Formation." *Language*, 32 (1956), 260—73.

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

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

Hill, B. M. "A Theoretical Derevation of the Zipf (Pareto) Law." In *Studies on Zipf's Law.* Eds. H. Guiter and M.V. Arapov. Bochum: Brockmeyer, 1983, pp. 53—64.

Kalinin, V.M. "Functionals Related to the Poisson Distribution, and Statistical Structure of a Text." In *Articles on Mathematical Statistics and the Theory of Probability.* Ed. J.V. Finnik. Providence, RI: American Mathematical Society, 1965, pp. 202—20.

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

Landauer, T.K., and Streeter, L.A. "Structural Differences Between Common and Rare Words: Failure of Equivalence Assumptions for Theories of Word Recognition." *Journal of Verbal Learning and Verbal Behavior*, 12 (1973), 119—31.

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

Mandelbrot, B. "On the Theory of Word Frequencies and on Related Markovian Models of Discourse." In *Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics.* Vol. XII. Ed. R. Jakobson. Providence, RI: American Mathematical Society, 1962, pp. 190—219.

Martin, W. *Analyse van een Vocabularium met behulp van een computer.* Brussels: AIMAV, 1970.

Menard, N. *Mesure de la Richesse Lexicale. Théorie et Vérifications Expérimentales. Etudes Stylométriques et Sociolinguistiques.* Genève: Slatkine-Champion, 1983.

Miller, G.A. "Some Effects of Intermittent Silence." *The American Journal of Psychology*, 52 (1957), 311—14.

Miller, G.A., Newman, E.B., and Friedman, E.A. "Length-Frequency Statistics for Written English." *Information and Control*, 1 (1958), 370—89.

Morrison, D.F. *Multivariate Statistical Methods.* Tokyo: McGraw-Hill Kogakusha, 1976.

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

Muller, C. "Du Nouveau sur les Distributions Lexicales: La Formule de Waring-Herdan." In *Langue Française et Linguistique Quantitative.* Ed. C. Muller. Genève: Slatkine, 1979, pp. 177—95.

Nushbaum, H.C. "A Stochastic Account of the Relationship between Lexical Density and Word Frequency." *Research on Speech Perception, Progress Report # 11.* 1985, Indiana University.

Orlov, J.K. "Dynamik der Häufigkeitsstrukturen." In *Studies on Zipf's Law.* Eds. H. Guiter and M.V. Arapov. Bochum: Brockmeyer, 1983, pp. 116—53.

Orlov, J.K. "Ein Model der Häufigkeitsstruktur des Vokabulars." In *Studies of Zipf's Law.* Eds. H. Guiter and M.V. Arapov. Bochum: Brockmeyer, 1983, pp. 154—233.

Orlov, J.K., and Chitashvili, R.Y. "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 (1982a), 297—300.

Orlov, J.K., and Chitashvili, R.Y. "On Some Problems of Statistical Estimation in Relatively Small Samples." *Bulletin of the Academy of Sciences, Georgia*, 108.3 (1982b), 513—16.

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

Orlov, J.K., and Chitashvili, R.Y. "Generalized Z-Distribution Generating the Well-Known 'Rank-Distributions'." *Bulletin of the Academy of Sciences, Georgia*, 110.2 (1983b), 268—72.

Paivio, A., Yuille, J.C., and Madigan, S. "Concreteness, Imagery and Meaningful Values for 925 Nouns." *Journal of Experimental Psychology Monograph 76.* I, Pt. 2. 1968.

Rainer, F. "Towards a Theory of Blocking: The Case of Italian and German Quality Nouns." *Yearbook of Morphology*, 1 (1988), 155—85.

Ratkowsky, D. "The Travaux de Linguistique Quantitative." (Book Review.) *Computers and the Humanities*, 22 (1988), 77—85.

Reder, L.M., Anderson, J.R., and Bjork, R.A. "A Semantic Interpretation of Encoding Specificity." *Journal of Experimental Psychology*, 102 (1974), 648—56.

Rouault, A. "Loi de Zipf et Sources Markoviennes." *Ann. Inst. H. Poincaré*, 14 (1978), 169—88.

Roy, G-R. *Contribution à l'Analyse de Syntagme Verbal.*

*Étude Morphosyntaxique et Statistique des Coverbes.* Paris: Klincksieck, 1976.

Schultink, H. "Produktiviteit als Morfologisch Fenomeen." *Forum der Letteren,* 2 (1961), 110—25.

Sichel, H.A. "On a Distribution Law for Word Frequencies." *Journal of the American Statistical Association,* 70 (1975), 542—47.

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

Simon, H.A. "On a Class of Skew Distribution Functions." *Biometrika,* 42 (1955), 435—40.

Sinclair, J.M., ed. *Looking Up: An Account of the Cobuild Project in Lexical Computing.* London: Collins, 1987.

Sterkenburg, P.G.J., and Pijnenburg, W.J.J. *van Dale Groot woordenboek van hedendaags Nederlands.* Utrecht: Van Dale Lexicografie, 1984.

Uit den Boogaart, P.C. *Woordfrequenties in Gesproken en Geschreven Nederlands.* Utrecht: Oosthoek, Scheltema and Holkema, 1975.

Veld, R in 't. *Hoe willekeurig kiest een schrijver zijn woorden? Een urn model voor onderzoek naar de frequenties van woorden, munten, achternamen en vissen.* Doctoral dissertation. University of Amsterdam, 1984.

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

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