The randomness assumption in word frequency statistics

R. Harald Baayen

Max Planck Institute for Psycholinguistics, Nijmegen

1 Introduction

The mathematical and computational tools available for the study of word frequency distributions have become increasingly powerful since Zipf published his seminal studies some 60 years ago (Zipf 1935, 1949). The first frequency counts were obtained manually, either by going through a text and filing new words and updating the frequencies of words already encountered on slips of paper, or by going through (manually compiled) concordances. The first statistician to study word frequency distributions, G. U. Yule, obtained the data for his book on "The statistical study of literary vocabulary" (Yule, 1944) in this way. The first frequency dictionary of Dutch, "De meest voorkomende woorden en woordcombinaties in het Nederlandsch", was similarly compiled manually by De la Court in 1937.

The first frequency list of Dutch obtained by means of a computer was compiled at the Mathematical Centre in 1965 by van Berckel, Brandt Corstius, Mokken, and van Wijngaarden. By 1967, Kučera and Francis had compiled a corpus of one million wordforms for English, and had published frequency counts and analyses in their famous "Computational Analysis of present-day American English" (Kučera and Francis, 1967). This prompted the construction of a slightly smaller corpus (727000 wordforms) of similar design for Dutch by the 'Werkgroep Frequentie-Onderzoek Nederlands', leading to the publication of "Woordfrequenties in geschreven en gesproken Nederlands" (Uit den Boogaart, 1975) and "Spreektaal. Woordfrequenties in Gesproken Nederlands" (de Jong, 1979). The most recent frequency information for Dutch is available in the CELEX lexical database (Burnage 1990), which can be queried on-line in the Netherlands, and of which a version on CD-ROM is also available (Baayen, Piepenbrock and van Rijn, 1993). The frequency counts in the CELEX database, which also contains information on spelling, phonology, morphological structure and syntactic features, are based on a corpus of 42 million wordforms compiled by the Institute for Dutch Lexicology in Leiden.

The transition from printed frequency lists based on relatively small corpora to on-line lexical databases based on corpora of tens of millions of words is accompanied by an ever increasing body of texts available in electronic form. Some collections of texts are made accessible via sophisticated software that enables users to search for words or word collocations. Typically, the matches found are presented with some preceding and following context.

For Dutch, the Institute for Dutch Lexicology (INL) has recently made a corpus of 5 million wordforms available for such on-line queries. Similarly, a Dutch newspaper, 'de Volkskrant', is now available on CD-ROM. The software facilitating access to 'De Volkskrant' and to the INL on-line corpus has as a serious drawback that the user is denied access to the texts themselves. Access to the full text, however, is especially critical for the question addressed in this study, namely the randomness assumption underlying all presently available statistical models for word frequency distributions.

Word frequency models build on the fundamental assumption that word tokens occur randomly in texts. It is clear that for natural language this assumption is too strong. The syntax of natural languages imposes severe constraints on where words can occur. For instance, following the Dutch determiner *de*, adjectives and nouns, but not verbs, are allowed (*de lamp, de felle lamp, *de schijnt*). Similarly, semantic constraints and principles of discourse organization may severely limit the way in which words occur in texts. This raises the question to what extent the predictions of theoretical models can be relied on, especially since it is known that the interpolated vocabulary size tends to seriously overestimate the observed vocabulary size (Brunet 1978, Hubert and Labbe 1988, Labbe and Hubert 1993). The aim of this paper is to trace the source of this overestimation, and to evaluate its consequences for the application of word frequency models in lexical statistics.

To do so, we need access to complete texts in electronic form. Fortunately, collections of raw electronic texts without limiting software-guided access are available by anonymous ftp. The Oxford Text Archive, at black.ox.ac.uk, the Gutenberg Project at mrcnext.cso.uiuc.edu, and the Online Book Initiative at obi.std.com have brought together large numbers of electronic texts, most of which are in English, ranging from election speeches by Clinton to electronic Startrek novels, and from Milton's 'Paradise Lost' to the Book of Mormon. From the Project Gutenberg, I obtained an electronic copy of *Alice in Wonderland*, by Lewis Carroll, and a copy of *Moby Dick*, by Herman Melville.[1] The Online Book Initiative has recently made available the first complete text of a Dutch novel to come to my attention, *Max Havelaar* by Multatuli, which I have also included in my analyses.

My discussion is structured as follows. In section 2, I introduce some basic expressions for the expectation and variance of the vocabulary size V_N as a function of the number of word occurrences N in the sample, and of the frequency spectrum, the number of different word types $V_N(m)$ with frequency m, again as a function of N. In section 3, the randomness assumption is tested by studying the development of the vocabulary in the three texts mentioned above. For each of these texts, it is shown that the observed and expected values diverge significantly for a large range of values of N. The goal of section 4 is to trace the source of this misfit, which may arise due to syntactic and

[1]The header of the electronic version of Melville's *Moby Dick* requires that I mention that this version was prepared by E. F. Tray at the University of Colorado, Boulder, on the basis of the Hendricks House Edition.

semantic constraints operating at the sentence level, to lexical specialization (Brunet 1978, Hubert and Labbe 1988), or to the discourse organization of the text. I will show that it is the way in which discourse is developed over time that gives rise to the misfit. The consequences of these findings are discussed in section 5.

2 Word Frequency Models

A text can be viewed as an ordered sequence of occurrences (or tokens) of words

$$(w_1, w_2, w_3, \ldots, w_N).$$

Usually, the number V of distinct words, the so-called word types, in the observed vocabulary

$$(A_1, A_2, A_3, \ldots, A_V)$$

is much smaller than the sample size N, due to the repeated occurrence of many word types. Let $f_N(A_i)$ denote the frequency with which word type A_i occurs in a sample of size N. Expressions for the numbers of different word types occurring for arbitrary sample sizes, as well as expressions for the numbers of different word types occurring with some specified frequency at a given sample size have been available since Good (1953), Kalinin (1965), and Good and Toulmin (1976) (see Chitashvili and Baayen (1993) for a review of word frequency models). In this section I introduce the expressions required for studying in what way the randomness assumption is violated in written texts.

Let $f_N(A_i) = m$ denote the event that word type A_i occurs with frequency m in a sample of N tokens. The expected total number of such word types, $\mathrm{E}[V_N(m)]$, is given by

$$\begin{aligned} \mathrm{E}[V_N(m)] &= \mathrm{E}[\sum_i \mathrm{I}_{[f_N(A_i)=m]}] \\ &= \sum_i \binom{N}{m} p(A_i)^m (1-p(A_i))^{N-m}. \end{aligned} \tag{1}$$

Note that the assumption that $f_N(A_i)$ is $\mathrm{bin}(N, p(A_i))$ distributed implies that the tokens of A_i occur randomly in the text. The expected overall number of different types in the sample, irrespective of their frequency, follows immediately:

$$\begin{aligned} \mathrm{E}[V_N] &= \mathrm{E}[\sum_{m\geq 1} V_N(m)] \\ &= \sum_{m\geq 1}\sum_i \binom{N}{m} p(A_i)^m (1-p(A_i))^{N-m} \\ &= \sum_i (1-(1-p(A_i))^N). \end{aligned} \tag{2}$$

For large N and small p, binomial probabilities can be approximated by Poisson probabilities, leading to the simplified expressions

$$\begin{aligned} \mathrm{E}[V_N(m)] &= \sum_i \frac{(\lambda(A_i)N)^m}{m!} e^{-\lambda(A_i)N} \\ \mathrm{E}[V_N] &= \sum_i (1 - e^{-\lambda(A_i)N}). \end{aligned} \tag{3}$$

Conditional on a given frequency spectrum $(V_N(m), m = 1, 2, \ldots)$, the vocabulary size $\mathrm{E}[V_M]$ for sample size $M < N$ equals

$$\begin{aligned} \mathrm{E}[V_M] &= \sum_{i=1}^{V_N} (1 - e^{-\lambda(A_i)M}) \\ &= \sum_{i=1}^{V_N} (1 - e^{-\frac{f_N(A_i)}{N}M}) \\ &= V_N - \sum_{m=1} V_N(m) e^{-\frac{M}{N}m}. \end{aligned} \tag{4}$$

Note that (4) suggests that, under randomness, and conditional on the words appearing in the first N tokens, $f_M(A_i)$ can alternatively be viewed as a binomially distributed random variable with parameters M/N and $f_N(A_i)$.

The Poisson approximation is especially useful for obtaining expressions for covariances:

$$\begin{aligned} &\mathrm{COV}(V_N(m), V_N(k)) = \\ &= \mathrm{COV}(\sum_{i=1}^{S} \mathrm{I}_{[f_N(A_i)=m]}, \sum_{j=1}^{S} \mathrm{I}_{[f_N(A_j)=k]}) \\ &= \sum_i \sum_j \mathrm{E}[\{\mathrm{I}_{[f_N(A_i)=m]}\} \cap \{\mathrm{I}_{[f_N(A_j)=k]}\}] \\ &\quad - \sum_i \sum_j \mathrm{E}[\mathrm{I}_{[f_N(A_i)=m]}]\mathrm{E}[\mathrm{I}_{[f_N(A_j)=k]}] \\ &= \sum_{i \neq j} \sum \frac{(\lambda(A_i)N)^m}{m!} \frac{(\lambda(A_j)N)^k}{k!} e^{-\lambda(A_i)N - \lambda(A_j)N} + \\ &\quad \delta_{mk}\mathrm{E}[V_N(m)] - \mathrm{E}[V_N(m)]\mathrm{E}[V_N(k)] \\ &= \sum_i \sum_j \frac{(\lambda(A_i)N)^m}{m!} \frac{(\lambda(A_j)N)^k}{k!} e^{-\lambda(A_i)N - \lambda(A_j)N} - \\ &\quad \sum_i \frac{(\lambda(A_i)2N)^{m+k}}{(m+k)!} \binom{m+k}{m} \frac{1}{2^{m+k}} e^{-2\lambda(A_i)N} + \\ &\quad \delta_{mk}\mathrm{E}[V_N(m)] - \mathrm{E}[V_N(m)]\mathrm{E}[V_N(k)] \end{aligned}$$

$$= \delta_{mk}\mathrm{E}[V_N(m)] - \binom{m+k}{m}\frac{1}{2^{m+k}}\mathrm{E}[V_{2N}(m+k)]. \quad (5)$$

Let S denote the number of different word types in the population from which a given text is sampled. Since $\mathrm{E}[V_N] = S - \mathrm{E}[V_N(0)]$,

$$\mathrm{VAR}(V_N) = \mathrm{VAR}(V_N(0)) = \mathrm{E}[V_{2N}] - \mathrm{E}[V_N]. \quad (6)$$

3 Testing the randomness assumption

The left hand panels of Figure 1 show the characteristic divergence between the observed and expected vocabulary size measured at 40 equally spaced intervals for Lewis Carroll's *Alice in Wonderland* (top), Herman Melville's *Moby Dick* (middle), and Multatuli's *Max Havelaar* (bottom). The type definition used here is a very simple one in which distinct strings represent different types. No morphological preprocessing has been applied. Hence *house* and *houses* are counted as two different types. The expected vocabulary size $\mathrm{E}[V_M]$ was obtained using (4), for each novel conditioning on the frequency spectrum of the complete text.

Note that for all three novels the difference between the expected and observed vocabulary size tends to be substantial for a large range of values of M ($M < N$). In the case of *Alice in Wonderland*, the expected vocabulary size exceeds the observed vocabulary size for the full range of values of M. For *Moby Dick* and *Max Havelaar*, this divergence is reversed for large M, where the expected vocabulary size is smaller than the observed vocabulary size. For the first 20 measurement points, (6) can be used to estimate the variance of V_N, so that standardized scores $Z = (V_N - \mathrm{E}[V_N])/\sqrt{\mathrm{VAR}[V_N]}$ can be obtained. Measurement points for which $|Z| > 1.96$ have been highlighted. For the three novels studied here, all Z-scores obtained, except for one text size in Multatuli's *Max Havelaar*, are smaller than -1.96, suggesting informally that the divergence between the observed and expected growth curves is significant for at least the first half of the text.

4 Tracing the source of the misfit

We have seen that the predictions derived from the basic model for word frequency distributions, essentially a simple urn model (without replacement), diverge substantially from the empirical intermediate vocabulary sizes. Instead of rejecting the model as unfit for the study of actual language data, it is useful to study the source of the misfit in some more detail, as this may shed some light on the conditions under which the model might remain valid.

There are three possible sources for the divergence between the empirical and expected vocabulary growth curves. Syntactic and semantic constraints at the level of the sentence are in conflict with the randomness assumption. These constraints might give rise to the observed misfit. Alternatively, it has been claimed that lexical specialization is at issue here. If the use of specialized words is restricted to particular text fragments, as it often appears to be, the

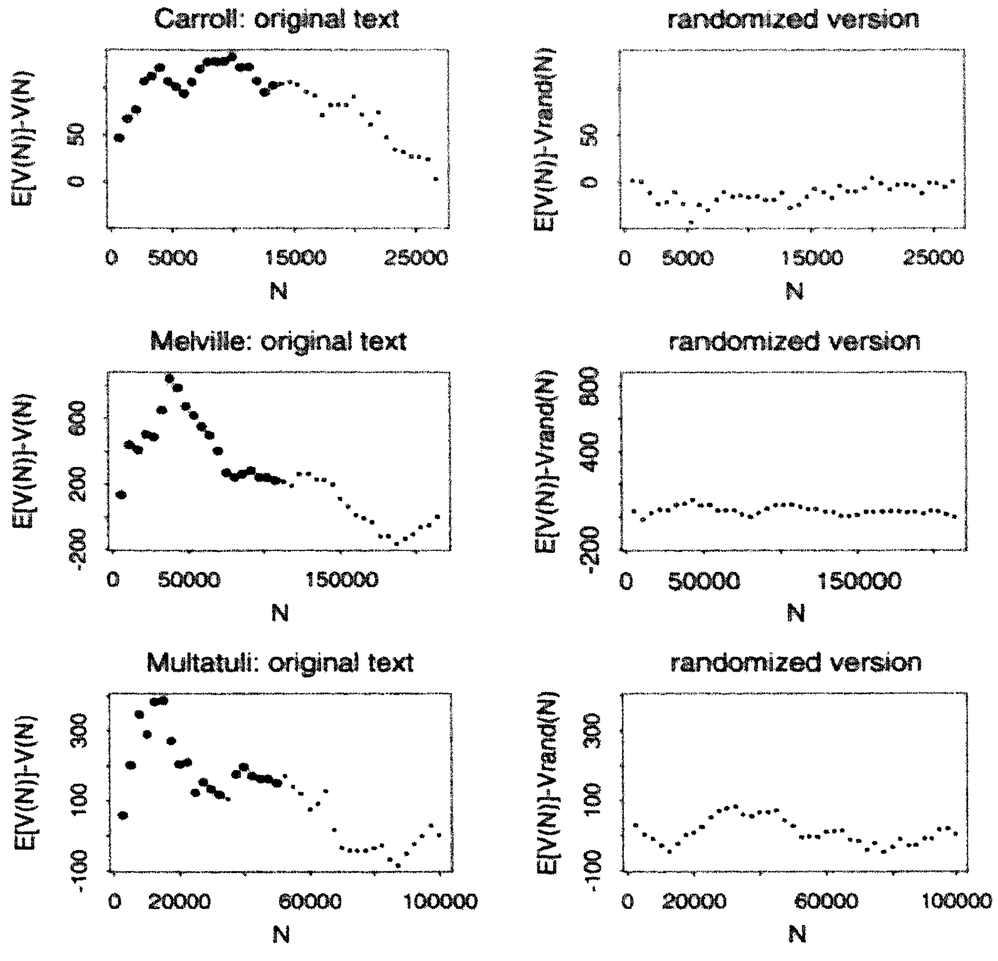

FIGURE 1. The size of the divergence between the empirical and expected vocabulary size $\mathrm{E}[V_M] - V_M$ for 40 equally spaced measurement points for L. Carroll's *Alice in Wonderland*, H. Melville's *Moby Dick*, and Multatuli's *Max Havelaar* (left column), and the size of this difference for a version of the novel in which the order of the sentences but not the order of the words in the sentences was randomized (right column). Significant differences have been highlighted.

uneven, clustered occurrence of the tokens of these types may underlie the misfit. Finally, it might be the case that the discourse organization of the text induces a non-random development of the vocabulary. I will explore these possibilities in turn.

4.1 Syntactic and semantic constraints

In order to trace the possible role of syntactic and semantic constraints, I made artificial versions of the three novels in which the order of the sentences was randomized, while keeping the order of the words in the sentences unchanged. Table 1 summarizes for each text the number of tokens N, the number of types V, the number of sentences s and the mean sentence length msl. The mean

novel	N	V	s	msl
Carroll	26611	2695	2323	11.45
Melville	213756	16741	16307	13.11
Multatuli	99819	11126	6791	14.69

TABLE 1. Number of tokens N, number of types V, number of sentences s and mean sentence length msl for Lewis Carroll's *Alice in Wonderland*, Herman Melville's *Moby Dick*, and Multatuli's *Max Havelaar*.

sentence length ranges between 11 and 15 words per sentence. Given these far from trivially small mean sentence lengths, syntactic and semantic constraints at the sentence level cannot but be operative. If their presence induces the misfit between the observed and expected vocabulary size, the randomized versions of the novels should show a similar pattern as found in the left hand panels of Figure 1.

The right hand panels of Figure 1 plot the results obtained. For all novels, the divergence between the observed and expected vocabulary sizes is substantially reduced. For all measurement points, the Z-score did not reach significance ($|Z| < 1.96$). Moreover, the direction of the difference appears to vary randomly, yielding largely negative scores for *Alice in Wonderland*, generally positive scores for *Moby Dick*, and both negative and positive scores for *Max Havelaar*. These results show that syntactic and semantic constraints at the sentence level can be ruled out as factors responsible for the lack of goodness-of-fit.

4.2 Lexical specialization

It has been argued that lexical specialization is to be held responsible for this lack of goodness-of-fit (Brunet 1978, Labbe and Hubert, 1993). The argument is based on the observation that the curve of V_N often reflects differences between texts when texts of different authors, or even different texts of the same author are studied jointly. To illustrate this simple observation, I concatenated Carroll's *Alice in Wonderland*, Baum's *The Wizard of Oz*, a collection of election speeches by Clinton, and Barrie's *Peter Pan*.[2] The observed and predicted vocabulary growth curves are shown in Figure 2. A marked discontinuity in the growth curve can be observed at the second vertical line, where the officialese of Clinton's election speeches succeeds *Alice in Wonderland* and *The Wonderful Wizard of Oz*. The specialized, concentrated use of officialese in the third partition of this artificial text gives rise to both substantial quantitative as well as qualitative differences between the observed and expected growth curves.

In this example, it is evident that the texts have not been sampled from the same population. Different authors will generally tend to use different sets of words. In addition, present-day officialese can hardly be compared with books

[2]The last three texts were obtained from the Project Gutenberg, the Online Book Initiative, and the Oxford Text Archive, respectively.

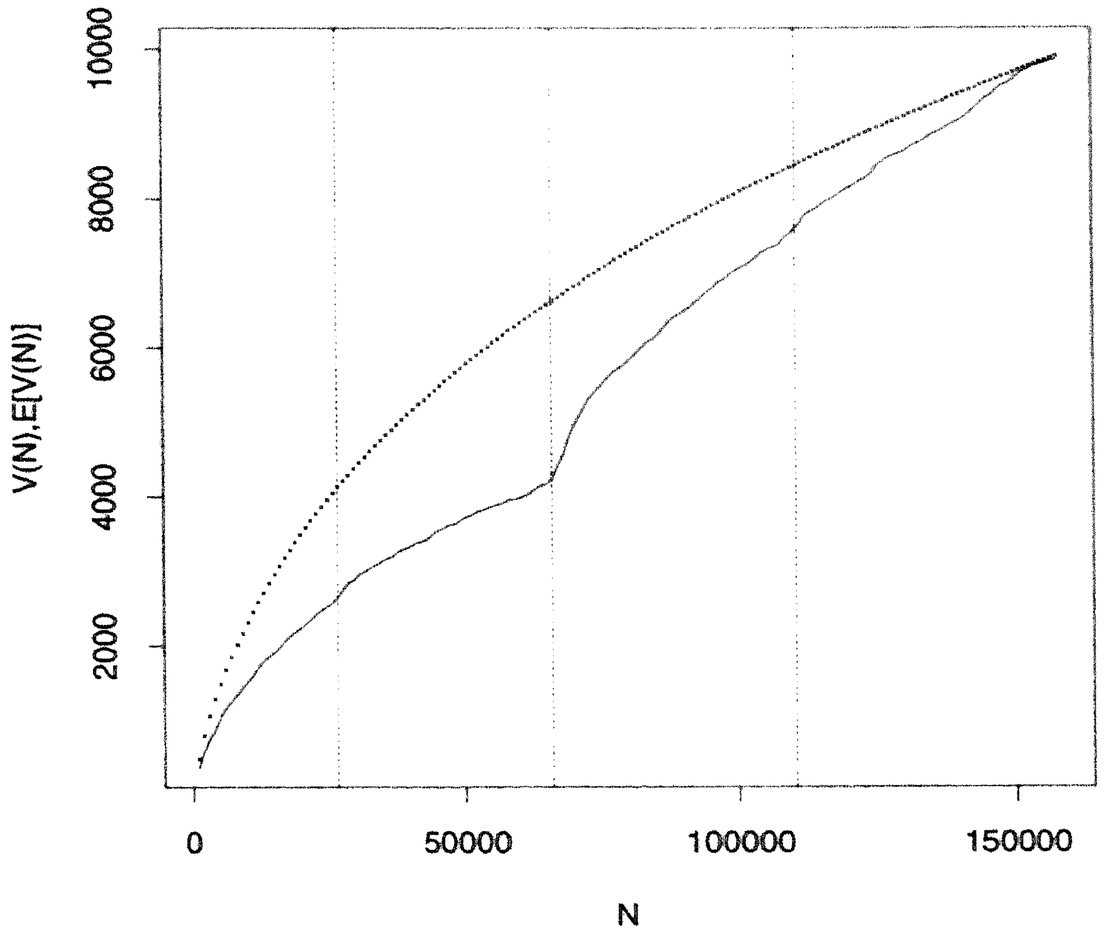

FIGURE 2. Empirical (solid line) and expected (dotted line) growth curves for the concatenated texts of L. Carroll's *Alice in Wonderland*, L. F. Baum's *The Wonderful Wizard of Oz*, election speeches by B. Clinton, and J. M. Barrie's *Peter Pan*, for 160 measurement points. Dotted vertical lines indicate the transition points between texts.

written for children more than 70 years ago. The substantial misfit comes as no surprise. Within a single novel, the effects of lexical specialization will not be as extreme. At first sight, there are two ways in which lexical specialization might violate the randomness assumption. It might be that lexical specialization is characteristic of certain parts of the text, but not for others, as in the above artificial example. Alternatively, lexical specialization, although uniformly distributed in the text, might as such give rise to the misfit between the observed and expected vocabulary size. If lexical specialization leads to local concentration of the tokens of a specialized type, this local concentration might imply that within the relevant text slice tokens that would otherwise have been free to represent additional non-specialized types are now allocated to one specialized type. For text slices with specialized words, this would result in a lower expected value for the vocabulary size.

This idea has been formalized by Hubert and Labbe (1988) and Labbe and Hubert (1993), who present the following modification of (4):

$$\mathrm{E}[V_M] = p\frac{M}{N}V_N + (1-p)\{V_N - \sum_m (V_N(m)e^{-\frac{M}{N}m})\}. \tag{7}$$

To obtain (7), assume that all tokens of a specialized word occur jointly in a particular fragment of the text. Also assume that a proportion p of all V_N types in the text enjoy specialized use, and that this specialization affects the same proportion p of the $V_N(m)$ types for all m. Finally, assume that the chunks of tokens of the specialized word types $(S_i, i = 1, 2, \ldots, pV_N)$ appear randomly distributed over the text. If so,

$$\begin{aligned}
\mathrm{E}[V_M] &= \mathrm{E}[\sum_{i=1}^{pV_N} \mathrm{I}_{[f_M(S_i)>0]} + \sum_{i=1}^{(1-p)V_N} \mathrm{I}_{[f_M(A_i)>0]} \\
&= \sum_{i=1}^{pV_N} \frac{M}{N} + \sum_{i=1} (1-p)V_N(1 - e^{-\frac{M}{N}f_N(A_i)}) \\
&= \frac{M}{N}pV_N + (1-p)V_N - \sum_m (1-p)V_N(m)e^{-\frac{M}{N}m}.
\end{aligned} \tag{8}$$

For K measurement points $(M_k, k = 1, 2, \ldots, K, M_k < N)$, Labbe and Hubert (1993) determine p by minimizing the chi-squared statistic

$$\sum_{k=1}^{K} \frac{(V_{M,k} - \mathrm{E}[V_{M,k}])^2}{\mathrm{E}[V_{M,k}]}, \tag{9}$$

conveniently ignoring that the variance of $\mathrm{E}[V_{M,k}]$ increases with M. In this way, much improved and often excellent fits can be obtained. For instance, for *Alice in Wonderland*, the optimal value of p for $K = 40$ equals 0.16, and the fit obtained is a perfect smoothed curve through the observed values of $V_{M,k}$ $(\chi^2_{(39)} = 3.58, p > 0.05)$. These results would suggest that lexical specialization as such violates the randomness assumption and gives rise to the discrepancy between the observed and expected vocabulary growth curves. Unfortunately, some of the assumptions underlying (7) are questionable.

First, for the majority of texts, the number of so-called hapax legomena, $V_N(1)$, accounts for roughly half the number of types V_N. Hapaxes, by virtue of occurring once only, cannot enjoy specialized use, if the operationalization of lexical specialization in terms of the bundled occurrence of all the tokens of a given type in a particular segment of the text is not to be trivialized. Second, if text slices in which specialized words occur are characterized by a deficit in the number of types, there should also be text slices with a surplus of types — the successive increments in the vocabulary size sum up to V_N for both the expected and the observed counts. If the text slices with a surplus of types also occur randomly in the text, it may well be that the effects of

lexical specialization are counterbalanced by effects of lexical richness. If so, no discrepancy between theory and observation should arise. Third, observe that in Figure 1 $\mathrm{E}[V_M] - V_M$ tends to be negative for large M in the novels by Melville and Multatuli. Application of (7) and (9) shows that for *Moby Dick* the optimal choice for Labbe and Hubert's parameter p, 0.12, does not yield an acceptable fit ($\chi^2_{(39)} = 162.79, p < 0.001$), and the same holds for *Max Havelaar*, $\chi^2_{(39)} = 92.58, p < 0.001$ for the Labbe and Hubert parameter $p = 0.10$. If the modification of (4) proposed by Labbe and Hubert (1993) is has any validity at all, this validity is restricted to texts with the developmental profile of *Alice in Wonderland* only. Texts with skewed profiles such as observed for *Moby Dick* and Max Havelaar cannot be analyzed in this way. We may conclude that if lexical specialization is to lead to violation of the randomness assumption, specialized types should not be randomly distributed in the text.

4.3 Discourse Structure

To test for possible effects of lexical specialization as a function of the discourse structure of the text, we need a formal definition of lexical specialization. Given the intuitive idea that lexical specialization implies a significant concentration in the occurrences of a word, we can define lexical specialization in terms of underdispersion. If a text is divided into K text slices, the dispersion d_i of word A_i is defined as the number of text slices in which this word occurs. If a word's dispersion is smaller than expected under chance conditions, it is underdispersed. To test whether A_i is significantly underdispersed, the test statistic

$$Z_i = \frac{d_i - \mathrm{E}[d_i]}{\sqrt{\mathrm{VAR}[d_i]}} \tag{10}$$

can be used. Since we have no reason to suppose that overdispersion occurs, we may assume that A_i is significantly underdispersed at the 5% level when $Z_i < -1.645$.

Expressions for $\mathrm{E}[d_i]$ and $\mathrm{VAR}[d_i]$ can be obtained using occupancy theory (Johnson and Kotz 1977: 113-114). Let X denote the number of text slices unoccupied by a token of word A_i with frequency $f_N(A_i)$. On partitioning a text into K slices, we can express X as the sum of the individual unoccupied slices:

$$X = \sum_{k=1}^{K} X_k, \tag{11}$$

with

$$X_k = \begin{cases} 0 & \text{if } A_i \text{ appears in the } k^{th} \text{ text slice,} \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

The number of text slices occupied by at least one token of A_i equals $d_i = K - X$. Since $\Pr(X_k = 1) = (1 - p_k)^{f_N(A_i)}$, with p_k the probability of assigning

a word token to the k^{th} text slice, we find that

$$\begin{aligned} \mathrm{E}[d_i] &= \mathrm{E}[K - \sum_{k=1}^{K} X_k] \\ &= K - \sum_{k=1}^{K} \mathrm{E}[X_k] \\ &= K - \sum_{k=1}^{K} (1 - p_k)^{f_N(A_i)}. \end{aligned} \tag{13}$$

When a given word token is equally likely to be assigned to any of the text slices, (13) reduces to

$$\mathrm{E}[d_i] = K(1 - (1 - \frac{1}{K})^{f_N(A_i)}). \tag{14}$$

After allotting N word tokens to K equiprobable text slices, each text slice will contain on average N/K word tokens. This allows us to use (14) to estimate the expected dispersion of all types A_i for the 40 equally large text slices of *Alice in Wonderland*, *Moby Dick*, and *Max Havelaar*.

The variance of d_i is obtained as follows.

$$\begin{aligned} \mathrm{VAR}[X] &= \mathrm{VAR}[\sum_{k=1}^{K} X_k] \\ &= \sum_k \mathrm{VAR}[X_k] + 2 \sum_{n<m} \sum \mathrm{COV}(X_n X_m) \\ &= \sum_k (\mathrm{E}[X_k^2] - (\mathrm{E}[X_k])^2) \\ &\quad +2 \sum_{n<m} \sum (\mathrm{E}[X_n X_m] - \mathrm{E}[X_n]\mathrm{E}[X_m]). \end{aligned} \tag{15}$$

As X_k^2 is nonzero only when $X_k = 1$, $\mathrm{E}[X_k^2] = \mathrm{E}[X_k]$. Similarly, we have that $\mathrm{E}[X_n X_m] = 1$ iff $X_n = X_m = 1$, and hence

$$\begin{aligned} \mathrm{E}[X_n X_m] &= \Pr(X_n X_m = 1) \\ &= \Pr(\{X_n = 1\} \cap \{X_m = 1\}) \\ &= (1 - p_n - p_m)^{f_N(A_i)}. \end{aligned} \tag{16}$$

This leads directly to

$$\begin{aligned} \mathrm{VAR}[X] &= \sum_{k=1}^{K} (1 - p_k)^{f_N(A_i)}(1 - (1 - p_k)^{f_N(A_i)}) \\ &+2 \sum_{n<m} \sum (1 - p_n - p_m)^{f_N(A_i)} - (1 - p_n)^{f_i}(1 - p_m)^{f_N(A_i)}. \end{aligned} \tag{17}$$

For equally sized text slices, $\mathrm{VAR}[d_i] = \mathrm{VAR}[K - X] = \mathrm{VAR}[X]$ is simplified to

$$\begin{aligned}\mathrm{VAR}[d_i] &= K(1-\frac{1}{K})^{f_N(A_i)} + K(K-1)(1-\frac{2}{K})^{f_N(A_i)} \\ &\quad -K^2(1-\frac{1}{K})^{2f_N(A_i)}. \end{aligned} \tag{18}$$

For each of the 40 text slices of *Alice in Wonderland*, *Moby Dick*, and *Max Havelaar*, I calculated the number of significantly underdispersed words. To study the relation between the growth of the vocabulary and the amount of underdispersion, it is useful to compare, for each successive text slice, the influx of new types with the influx of new underdispersed types. In order to compare observed with empirical values, it is convenient to introduce two difference functions. Let $D_V(k)$ denote the difference between the expected and observed number of new types in text slice k,

$$D_V(k) = (\mathrm{E}[V_{M,k}] - \mathrm{E}[V_{M,k-1}]) - (V_{M,k} - V_{M,k-1}), \tag{19}$$

and let

$$D_U(k) = (U_{M,k} - U_{M,k-1}) - (\mathrm{E}[U_{M,k}] - \mathrm{E}[U_{M,k-1}]), \tag{20}$$

with $U_{M,k}$ the number of underdispersed types in the k^{th} text slice, denote the difference between the observed and expected numbers of new underdispersed types in text slice k. Figure 3 plots $D_V(k)$ (small dots) and $D_U(k)$ (large dots) and the corresponding smoothed curves using running medians (Tukey, 1977) for our three texts. In each case, we find that the two curves tend to be each other's mirror images. Especially for the first 7 measurement points, $D_V(k)$ tends to be large and $D_U(k)$ small. In other words, in the initial parts of these novels, both new types and significantly underdispersed types are scarce. In later parts of the novels, there is a tendency for the expected increase in vocabulary to slightly underestimate the empirical increase, and it is here that the empirical numbers of underdispersed words are slightly higher than expected.

This pattern of results suggests that lexical specialization, defined in terms of significant underdispersion, is not randomly distributed in the text, and that it is the scarcity of significant underdispersion in the initial segments of the text, combined with a deficit in type richness, that gives rise to the divergence between the observed and expected vocabulary growth curves. In hindsight, it is obvious that lexical specialization and vocabulary richness go hand in hand. When a particular topic is discussed in detail, key words for that topic will be used intensively. These key words are the significantly underdispersed words of this study (see Baayen, 1994, for detailed discussion). At the same time, additional vocabulary is called upon, without which the many facets of the topic that make it worth mentioning could not be discussed.

What we find, then, is that the organization of texts at the discourse level is at issue. In the initial sections of the text, the reader is introduced gently to the fictive world of the novel. Here, large numbers of specialized words, both

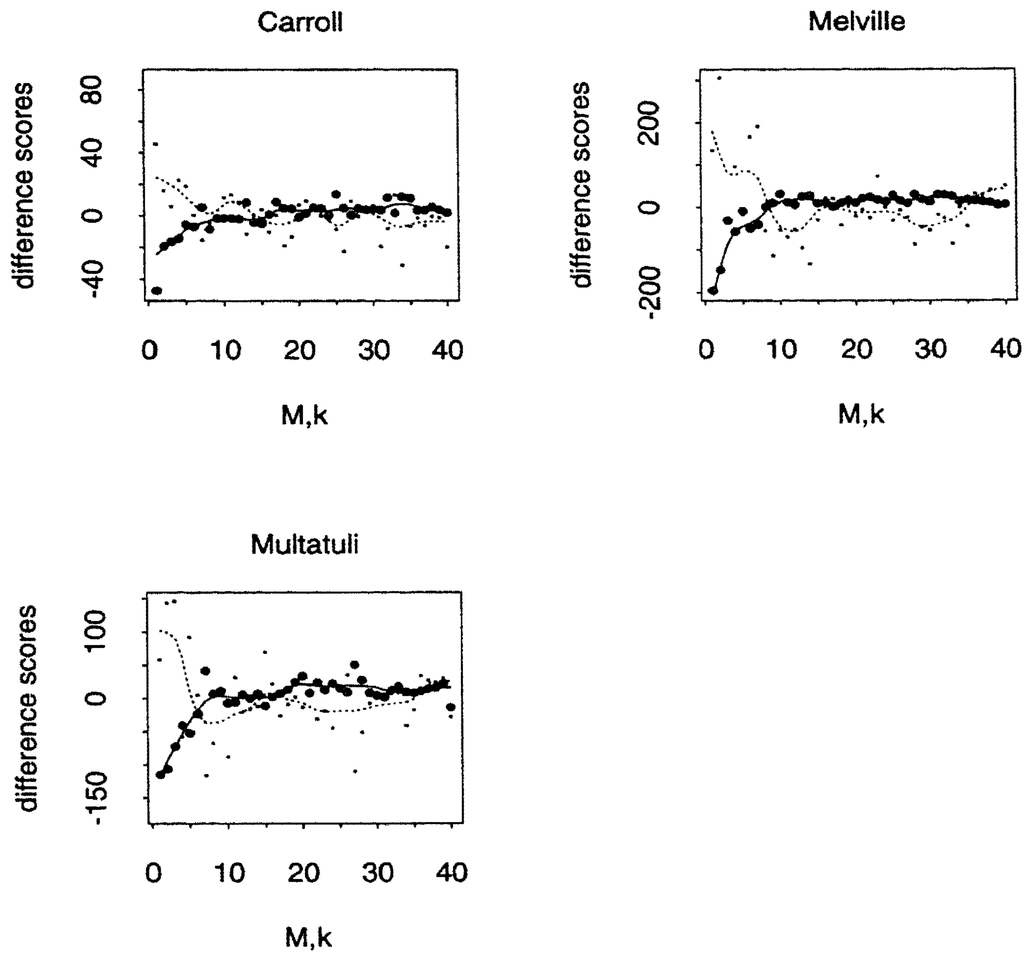

FIGURE 3. Difference scores $D_V(k)$ (dotted line, small dots) and $D_U(k)$ (solid line, large dots) for L. Carroll's *Alice in Wonderland*, H. Melville's *Moby Dick*, and Multatuli's *Max Havelaar*.

the hard-worked underdispersed words, as well as the specialized low-frequency words their use brings along, are avoided. Once the general topic domain has been established, specialized vocabulary is put to use to elaborate more specific topics in full.

5 DISCUSSION

I have shown that the lack of goodness-of-fit of any probabilistic model for word frequency distributions of texts that assumes that words occur randomly in texts is due to the way in which texts are structured on the discourse level. Syntactic and semantic constraints operating on the sentence level, as well as lexical specialization by itself, do not play a significant role.

This finding has important consequences for the statistical analysis of word frequency distributions, as it shows that the theoretical predictions of the urn model will be accurate either if the textual materials studied do not have the

discourse structure observed for the novels studied here, or if this discourse structure is irrelevant to the question at hand. The first possibility arises in studies where corpora are investigated. Corpora such as Uit den Boogaart (1975) and Kučera and Francis (1965) are collections of randomly sampled short text fragments of approximately the same length. No discourse organization will be present in the sequence of such fragments. Hence the model will accurately predict the observed vocabulary size for $M < N$.

The second possibility arises when the model is used to obtain estimates of population parameters that are relatively independent of discourse organization. For instance, in studies of vocabulary richness (Good and Toulmin 1976, Efron and Thisted 1976, Sichel 1986), the number of different word types in an author's vocabulary is estimated on the basis of one or more of his texts. If the number of different word types an author chooses to use to discuss a particular topic domain is independent of the way in which he structures the text to facilitate comprehension for the reader, then the rhetorical structure of the text becomes irrelevant when one's aim is to estimate the size of the vocabulary the author had at his disposal for discussing this topic domain, including the words he knew but did not use.

Summing up, the finding that the randomness assumption is violated at the level of discourse structure implies that word frequency models for which this assumption is crucial can nevertheless be reliably applied in corpus-based studies and in studies of lexical richness.

6 Epilogue

Having come to the end of my discussion of the randomness assumption in word frequency statistics, I would like to add a few words on the occasion of Cor Baayen's retirement as scientific director of the Centre for Mathematics and Computer Science.

As mentioned in the introduction, the first computerized frequency list of Dutch was compiled in 1965 at the Mathematical Centre, the name of the Centre for Mathematics and Computer Science at that time. The director of the institute, Aad van Wijngaarden, one of the pioneers of computer science in the Netherlands, had a keen interest in language in general, and in lexicology and etymology in particular. Not surprisingly, the first study of the Dutch language in which the computer was used as a tool for obtaining word frequency counts and for carrying out morphological analyses to appear in print was a Mathematical Centre Tract (van Berckel et al., 1965).

Van Wijngaarden's successor as director of the Mathematical Centre was Cor Baayen. While sharing the same interest in historical linguistics and etymology, Cor was well aware of the importance of methods of formal logic as tools for the analysis of problems of ambiguity and scope that arise at the level of the syntax and semantics of natural language, and he has stimulated research in the interdisciplinary domain of language, logic and computer science throughout his directorship.

Looking back, it is clear that the study of natural language in the Netherlands has profited from the erudition and breadth of vision of the Mathematical Centre's last scientific directors. It is to be hoped that the future CWI will be able to demonstrate a similar breadth of vision, stimulating the use of new mathematical techniques not only in the sciences and in engineering, but also in the humanities.

References

[Baaed] R. H. Baayen. The effects of lexical specialization on the growth curve of the vocabulary. (submitted), September 1994 submitted.

[BPvR93] R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX lexical database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.

[Bru78] E. Brunet. *Le Vocabulaire de Jean Giraudoux*, volume 1 of *TLQ*. Slatkine, Genève, 1978.

[Bur88] Burnage. *CELEX; A guide for users.* Centre for Lexical Information, Nijmegen, 1988.

[CB93] R. J. Chitashvili and R. H. Baayen. Word frequency distributions. In G. Altmann and L. Hřebíček, editors, *Quantitative Text Analysis*, pages 54–135. Wissenschaftlicher Verlag Trier, Trier, 1993.

[DlC37] J. F. H. A. De la Court. *De meest voorkomende woorden en woordcombinaties in het Nederlandsch.* Volkslectuur, Batavia, 1937.

[ET76] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63:435–447, 1976.

[Goo53] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

[GT56] I. J. Good and G. H. Toulmin. The number of new species and the increase in population coverage, when a sample is increased. *Biometrika*, 43:45–63, 1956.

[HL88] P. Hubert and D. Labbe. Un modèle de partition du vocabulaire. In D. Labbe, P. Thoiron, and D. Serant, editors, *Etudes sur la richesse et les structures lexicales*, pages 93–114. Slatkine-Champion, Paris, 1988.

[JK77] N. L. Johnson and S. Kotz. *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory.* John Wiley & Sons, New York, 1977.

[Jon79] E. D. de Jong. *Spreektaal. Woordfrequenties in Gesproken Nederlands.* Oosthoek, Scheltema en Holkema, Utrecht, 1979.

[Kal65] V. M. Kalinin. Functionals related to the poisson distribution, and statistical structure of a text. In J. V. Finnik, editor, *Articles on Mathematical Statistics and the Theory of Probability*, pages 202–220, Providence, Rhode Island, 1965. Steklov Institute of Mathematics 79, American Mathematical Society.

[KF67] H. Kučera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, 1967.

[LH93] D. Labbe and P. Hubert. La richesse du vocabulaire (vocabulary richness). Centre de Recherche sur le Politique, l'Administration et le Territoire, October 1993.

[Sic86] H. S. Sichel. Word frequency distibutions and type-token characteristics. *Mathematical Scientist*, 11:45–72, 1986.

[Tuk77] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass., 1977.

[UdB75] P. C. Uit den Boogaart, editor. *Woordfrequenties in Gesproken en Geschreven Nederlands*. Oosthoek, Scheltema & Holkema, Utrecht, 1975.

[vBBCMvW65] J. A. Th. M. van Berckel, H. Brandt Corstius, R. J. Mokken, and A. van Wijngaarden. *Formal Properties of Newspaper Dutch*. Mathematisch Centrum, Amsterdam, 1965.

[Yul44] G. U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.

[Zip35] G. K. Zipf. *The Psycho-Biology of Language*. Houghton Mifflin, Boston, 1935.

[Zip49] G. K. Zipf. *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*. Hafner, New York, 1949.