The Effects of Lexical Specialization on the Growth Curve of the Vocabulary

R. Harald Baayen*
Max Planck Institute for
Psycholinguistics

The number of different words expected on the basis of the urn model to appear in, for example, the first half of a text, is known to overestimate the observed number of different words. This paper examines the source of this overestimation bias. It is shown that this bias does not arise due to sentence-bound syntactic constraints, but that it is a direct consequence of topic cohesion in discourse. The nonrandom, clustered appearance of lexically specialized words, often the key words of the text, explains the main trends in the overestimation bias both quantitatively and qualitatively. The effects of nonrandomness are so strong that they introduce an overestimation bias in distributions of units derived from words, such as syllables and digrams. Nonrandom word usage also affects the accuracy of the Good-Turing frequency estimates which, for the lowest frequencies, reveal a strong underestimation bias. A heuristic adjusted frequency estimate is proposed that, at least for novel-sized texts, is considerably more accurate.

1. Introduction

When reading through a text, word token by word token, the number of different word types encountered increases, quickly at first, and ever more slowly as one progresses through the text. The number of different word types encountered after reading N tokens, the vocabulary size $V(N)$, is a function of N. Analytical expressions for $V(N)$ based on the urn model are available. A classic problem in word frequency studies is, however, that these analytical expressions tend to overestimate the observed vocabulary size, irrespective of whether these expressions are nonparametric (Good 1953; Good and Toulmin 1956; Muller 1979; Brunet 1978) or parametric (Sichel 1986; Khmaladze and Chitashvili 1989; Chitashvili and Baayen 1993) in nature.

Although the theoretical or expected vocabulary size $\mathrm{E}[V(N)]$ generally is of the same order of magnitude as the observed vocabulary size, the lack of precision one observes time and again casts serious doubt on the reliability of a number of measures in word frequency statistics. For instance, Baayen (1989, 1992) and Baayen and Renouf (1996) exploit the Good-Turing estimate for the probability of sampling unseen types (Good 1953) to develop measures for the degree of productivity of affixes, Baayen and Sproat (to appear) apply this Good-Turing estimate to obtain enhanced estimates of lexical priors for unseen words, and the Good-Turing estimates also play an important role for estimating population probabilities (Church and Gale 1991). If a simple random variable such as the vocabulary size reveals consistent and significant deviation from its expectation, the accuracy of the Good-Turing estimates is also called into question. The aim of this paper is to understand why this deviation between the-

* Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. E-mail: baayen@mpi.nl

ory and observation arises in word frequency distributions, and in this light evaluate applications of the Good-Turing results.

The remainder of this paper is structured as follows. In Section 2, I introduce some basic notation and the expressions for the growth curve of the vocabulary with which we will be concerned throughout, including a model proposed by Hubert and Labbe (1988), which, by introducing a smoothing parameter, leads to much-improved fits. Unfortunately, this model is based on a series of unrealistic simplifications, and cannot serve as an explanation for the divergence between the observed and expected vocabulary size. In Section 3, therefore, I consider a number of possible sources for the misfit in greater detail: nonrandomness at the sentence level due to syntactic structure, nonrandomness due to the discourse structure of the text as a whole, and nonrandomness due to thematic cohesion in restricted sequences of sentences (paragraphs). Section 4 traces the implications of the results obtained for distributions of units derived from words, such as syllables and digrams, and examines the accuracy of the Good-Turing frequency estimates. A list of symbols is provided at the end of the paper.

2. The Growth Curve of the Vocabulary

Let N be the size of a text in word tokens, and let V denote the total number of different word types observed among the N word tokens. Roughly half of the word types occur only once, the so-called **hapax legomena**, others occur with higher frequencies.[1] Let $V(N,1)$ denote the number of once-occurring types among N tokens, and, similarly, let $V(N,f)$ denote the number of types occurring f times after sampling N tokens. The expected number of different types $\mathrm{E}[V(M)]$ for $M < N$ conditional on the **frequency spectrum** $\{V(N,f)\}, f = 1, 2, 3, \ldots$ can be estimated by

$$\mathrm{E}[V(M)] = V - \sum_{f} V(N,f)\left(1 - \frac{M}{N}\right)^{f}. \tag{1}$$

A proof for (1) is presented in the appendix.

Figure 1 illustrates the problems that arise when (1) is applied to three texts, *Alice in Wonderland*, by Lewis Carroll (upper panels), *Moby Dick* by Herman Melville (middle panels), and *Max Havelaar* by Multatuli (the pseudonym of Eduard Douwes Dekker, bottom panels).[2] All panels show the sample size N on the horizontal axis. Thus the horizontal axis can be viewed as displaying the "text time" measured in word tokens. The vertical axis of the left-hand panels shows the number of observed word types (dotted line) and the number of types predicted by the model (solid line) obtained using (1). These panels reveal that the expected vocabulary size overestimates the observed vocabulary size for almost all of the 40 equidistant measurement points. To the eye, the overestimation seems fairly small. Nevertheless, in absolute terms the expectation may be several hundreds of types too high, and may run up to 5% of the total vocabulary size.

[1] The type definition I have used throughout is based on the orthographic word form: *house* and *houses* are counted as two different types, *houses* and *houses* as two tokens of the same type. No lemmatization has been attempted, first, because the probabilistic aspects of the problem considered here are not affected by whether or not lemmatization is carried out, and second, because it is of interest to ascertain how much information can be extracted from texts with minimal preprocessing.

[2] These texts were obtained by anonymous ftp from the Project Gutenberg at obi.std.com. The header of the electronic version of *Moby Dick* requires mention of E.F. Tray at the University of Colorado, Boulder, who prepared the text on the basis of the Hendricks House Edition.

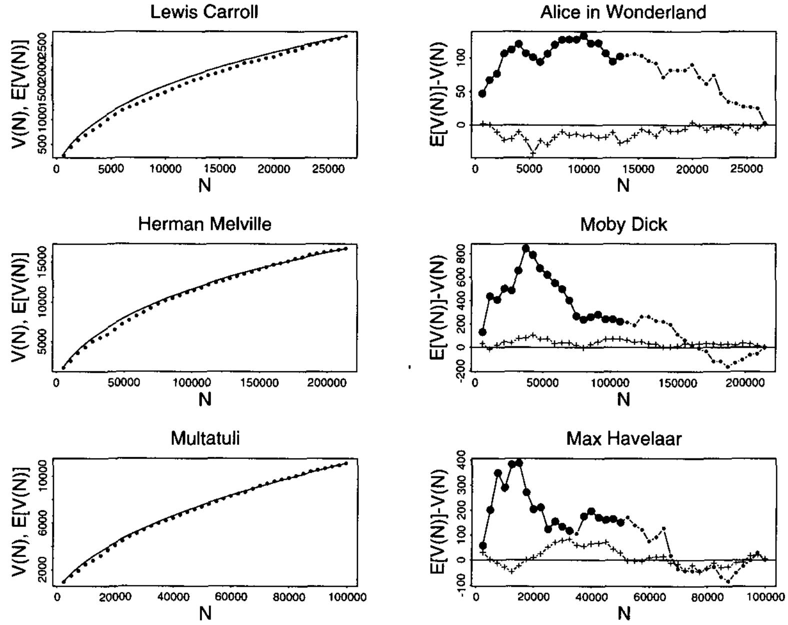

Figure 1
The growth curve of the vocabulary. Observed vocabulary size $V(N)$ (dotted lines) and expected vocabulary size $E[V(N)]$ (solid lines) for three novels (left-hand panels) and the corresponding overestimation errors $E[V(N)] - V(N)$ (dotted lines) and their sentence-randomized versions ("+"-lines, see Section 3.1) (right-hand panels).

The right-hand panels of Figure 1 show the overestimation error functions $E[V(N)] - V(N)$ corresponding to the left-hand panels using dotted lines. For the first 20 measurement points, the instances for which $E[V(N)]$ diverges significantly from $V(N)$ are shown in bold.[3] Clearly, the divergence is significant for almost all of the first 20 measurement points. This suggests informally that the discrepancy between $E[V(N)]$ and $V(N)$ is significant over a wide range of sample sizes.

2.1 The Model Proposed by Hubert and Labbe

The problem of the systematic estimation error of $E[V(N)]$ has been pointed out by Muller (1979) and Brunet (1978), who hypothesize that lexical specialization is at issue. In any text, there are words the use of which is mainly or even exclusively restricted to a given subsection of that text. Such locally concentrated clusters of words are at odds with the randomness assumption underlying the derivation of (1), and may be the cause of the divergence illustrated in Figure 1. Following this line of reasoning, Hubert and Labbe (1988) propose a model according to which (1) should be modified

[3] Since the expression for an estimate of the variance of $V(N)$ figuring in the Z-scores used here requires knowledge of $E[V(2N)]$, the significance of the divergence for the second 20 measurement points is not available. For technical details, see Chitashvili and Baayen (1993).

as follows (see the appendix for further details):

$$\mathrm{E}_{HL}[V(M)] = p\frac{M}{N}V + (1-p)V - (1-p)\sum_f V(N,f)\left(1-\frac{M}{N}\right)^f . \tag{2}$$

Hubert and Labbe's model contains one free parameter, the **coefficient of vocabulary partition** *p*, an estimate of the proportion of specialized words in the vocabulary. Given *K* different text sizes for which the observed and expected vocabulary sizes are known, *p* can be estimated by minimizing the mean squared error (MSE)

$$\frac{\sum_{k=1}^{K}(V(M_k) - \mathrm{E}[V(M_k)])^2}{K} \tag{3}$$

or the chi-square statistic

$$\sum_{k=1}^{K}\frac{(V(M_k) - \mathrm{E}[V(M_k)])^2}{\mathrm{E}[V(M_k)]} \tag{4}$$

(conveniently ignoring that the variance of $V(M)$ increases with M; see Chitashvili and Baayen [1993]). For *Alice in Wonderland,* minimalization of (4) for $K = 40$ leads to $p = 0.16$, and according to this rough estimate of goodness-of-fit the revised model fits the data very well indeed ($X^2_{(39)} = 3.58, p > 0.5$). For *Moby Dick,* however, the chi-squared statistic suggests a significant difference between the observed and expected vocabulary sizes ($X^2_{(39)} = 172.93, p < 0.001$), even though the value of the p parameter (0.12) leads to a fit that is much improved with respect to the unadjusted growth curve ($X^2_{(39)} = 730.47$). Closer inspection of the error pattern of the adjusted estimate reveals the source of the misfit: for the first 12 measurement points, the observed vocabulary size is consistently overestimated. From the 14th observation onwards, the Hubert-Labbe model consistently underestimates the real vocabulary size. Apparently, the development of the vocabulary in *Moby Dick* can be modeled globally, but local fluctuations introducing additional points of inflection into the growth curve are outside its scope—a more detailed study of the development of lexical specialization in the narrative is required if the appearance of these points of inflection are to be understood.

In spite of this deficiency, the Hubert-Labbe curve appears to be an optimal smoother, and this suggests that the value obtained for the coefficient of vocabulary partition *p* is a fairly reliable estimate of the extent to which a text is characterized by lexical specialization. In this light, the evaluation by Holmes (1994), who suggests that *p* might be a useful discriminant for authorship attribution studies, is understandable. Unfortunately, the assumptions underlying (2) are overly simplistic, and seriously call into question the reliability of *p* as a measure of lexical specialization, and the same holds for the explanatory value of this model for the inaccuracy of $\mathrm{E}[V(N)]$.

2.2 Problems with the Hubert and Labbe Model

One highly questionable simplification underlying the derivation of (2) spelled out in the appendix is that specialized words are assumed to occur in a single text slice only. Consider Figure 2, which plots the number of times *Ahab* appears in 40 successive, equally sized text slices that jointly constitute the full text of *Moby Dick.* The dotted line reveals the main developmental pattern (time-series smoothing using running medians). Even though Ahab is one of the main characters in *Moby Dick,* and even though his name certainly belongs to the specialized vocabulary of the novel, Ahab is

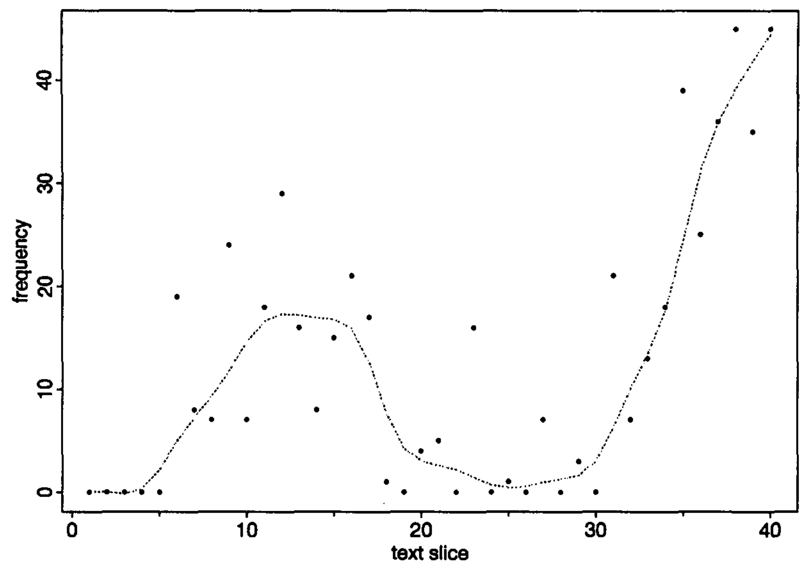

Figure 2
Nonrandom word usage illustrated for *Ahab* in *Moby Dick*. The horizontal axis plots the 40 equally sized text slices, the vertical axis the frequency of *Ahab* in these text slices. The dotted line represents a time-series smoother using running medians (Tukey 1977).

not mentioned by name in one text slice only, as the Hubert-Labbe model would have. What we find is that he is not mentioned at all in the first five text slices. Following this we observe a series of text slices in which he appears frequently. These are in turn succeeded by slices in which Ahab is hardly mentioned, but he reappears in the last part of the book, and as the book draws to its dramatic close, the frequency of *Ahab* increases to its maximum. This is an illustration of what Indefrey and Baayen (1994) refer to as **inter-textual cohesion**: the word *Ahab* enjoys specialized use, but it occurs in a series of subtexts within the novel as a whole, contributing to its overall cohesion. Within text slices where *Ahab* is frequently mentioned, the **intra-textual cohesion** may similarly be strengthened. For instance, *Ahab* appears to be a specialized word in text slice 23, but he is mentioned only in passing in text slice 25. His appearance in the two text slices strengthens the intertextual cohesion of the whole novel, but it is only the intra-textual cohesion of slice 23 that is raised. The presence of inter-textual cohesion in addition to intra-textual cohesion and the concomitant phenomenon of global lexical specialization suggest that in order to understand the discrepancy between $V(N)$ and its expectation, a more fine-grained approach is required.

A second question concerns how lexical specialization affects the empirical growth curve of the vocabulary. Inspection of plots such as those presented in Figure 1 for *Alice in Wonderland* suggests that the effects of lexical specialization appear in the central sections of the text, as it is there that the largest differences between the expected and the observed vocabulary are to be observed—differences that are highly penalized by the MSE and chi-squared techniques used to estimate the proportion of specialized words in the vocabulary. Unfortunately, the central sections are not necessarily the ones characterized by the highest degree of lexical specialization. To see this, consider Figure 3, which plots the difference between the expected number of new types using

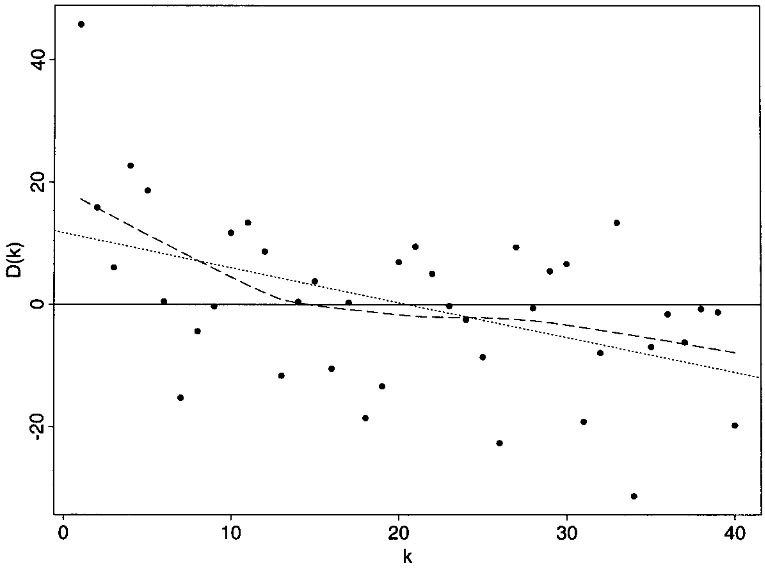

Figure 3
Error scores for the influx of new types in *Alice in Wonderland*. The $k = 1, 2, \ldots, 40$ text slices are displayed on the horizontal axis, the progressive difference scores $D(k)$ are shown on the vertical axis. The dashed line represents a nonparametric scatterplot smoother (Cleveland 1979), the dotted line a least squares regression line (the negative slope is significant, $F(1,38) = 11.07, p < .002$).

(1) and the observed number of new types for the successive text slices of *Alice in Wonderland*. More precisely, for each text slice k, $k = 1, \ldots, 40$, we calculate the **progressive difference error scores** $D(k)$, $k = 1 \ldots 40$:

$$D(k) = \{\mathrm{E}[V(M_k)] - \mathrm{E}[V(M_{k-1})]\} - \{V(M_k) - V(M_{k-1})\}. \tag{5}$$

Note that in addition to positive difference scores, which should be present given that $\mathrm{E}[V(M_k)] > V(M_k)$ for most, or, as in *Alice in Wonderland*, for all values of k, we also have negative difference scores. Text slices containing more types than expected under chance conditions are necessarily present given the existence of text slices k for which $\mathrm{E}[V(M_k)] - V(M_k) > 0$: the total number of types accumulated over the 40 text slices has to sum up to $V(N)$. Figure 3 shows that the expected numbers of new word types are overestimated for the initial part of the novel, that the theoretical estimates are fairly reliable for the middle section of the novel, while the final chapters show a slightly greater increase in the number of new types than expected under chance conditions. If lexical specialization affects the influx of new types, its effects appear not in the central sections of the novel as suggested by Figure 1, but rather in the beginning and perhaps at the end. This finding seriously questions the appropriateness of using the growth curve of the vocabulary for deriving a measure of lexical specialization.

A third question arises with respect to how one's measure of lexical concentration is affected by the number of text slices K. In Hubert and Labbe's model, the optimal value of the p parameter is independent of the number of text slices K for not-too-small K ($K > 10$). Since the expected growth curve and the observed growth curve are completely fixed and independent of K—the former is fully determined by the frequency spectrum of the complete text, the latter is determined by the text itself—the choice of K influences only the number of points at which the divergence between the two curves is measured. Increasing the number of measurement points increases the degrees of freedom along with the deviance, and the optimal value of the p parameter remains virtually unchanged. But is this a desirable property for a measure of lexical specialization? Even without taking the effects of inter-textual cohesion into account, and concentrating solely on local specialization and intra-textual cohesion, formulating lexical specialization in terms of concentration at a particular point in the text is unrealistic: it is absurd to assume that all tokens of a specialized word appear in one chunk without any other intervening words. A more realistic definition of (local) lexical specialization is the concentration of the tokens of a given word within a particular text slice. In such an approach, however, the size of the text slice is of crucial importance. A word appearing only in the first half of a book enjoys some specialized use, but to a far lesser extent than a word with the same frequency that occurs in the first half of the first chapter only. In other words, an approach to lexical specialization in terms of concentration of use is incomplete without a specification of the unit of concentration itself.

3. Sources of Nonrandomness

To avoid these problems, I will now sketch a somewhat more fine-grained approach to understanding why $V(N)$ and its expectation diverge, adopting Hubert and Labbe's central insight that lexical specialization can be modeled in terms of local concentration. Consider again the potential sources for violation of the randomness assumption underlying the derivation of $\mathrm{E}[V(N)]$. At least three possibilities suggest themselves: syntactic constraints on word usage within sentences, global discourse organization, and local repetition. I will consider these possibilities in turn.

3.1 Syntactic Constraints

Syntactic constraints at the level of the sentence introduce many restrictions on the occurrence of words. For instance, in normal written English, following the determiner *the* the appearance of a second instance of the same determiner (as in this sentence), is extremely unlikely. According to the urn model, however, such a sequence is likely to occur once every 278 words (the relative frequency of *the* in English is approximately 0.06), say once every two pages. This is not what we normally find. Clearly, syntax imposes severe constraints on the occurrence of words. Does this imply that the urn model is wrong? For individual sentences, the answer is undoubtedly yes. But for more global textual properties such as vocabulary size, a motivated answer is less easy to give. According to Herdan (1960, 40), reacting to Halle's criticism of the urn model as a realistic model for language, there is no problem, since statistics is concerned with form, not content.[4] Whatever the force of this argument may be, Figure 1 demonstrates clearly that the urn model lacks precision for our data.

In order to ascertain the potential relevance of syntactic constraints referred to by Halle, we may proceed as follows: If sentence-level syntax underlies the misfit between the observed and the expected vocabulary size, then this misfit should remain visible for randomized versions of the text in which the sentences have been left unchanged, but in which the order of the sentences has been permuted. If the misfit disappears,

[4] M. Halle, "In defence of the number two," in *Studies Presented to J. Whatmough*, The Hague, 1957, quoted in Herdan, 1960, page 40.

we know that constraints the domain of which are restricted to the sentence can be ruled out.

The results of this randomization test applied to *Alice in Wonderland, Moby Dick,* and *Max Havelaar* are shown in the right-hand panels of Figure 1 by means of "+" symbols. What we find is that following sentence randomization, all traces of a significant divergence between the observed and expected vocabulary size disappear. The differences between $\mathrm{E}[V(N)]$ and $V(N)$ are substantially reduced and may remain slightly negative, as in *Alice in Wonderland,* or slightly positive, as for *Moby Dick,* or they may fluctuate around zero in an unpredictable way, as in *Max Havelaar*. Since we are left with variation that is probably to be attributed to the particularities of the individual randomization orders, we may conclude that at the global level of the text as an (unordered) aggregate of sentences, the randomness assumption remains reasonable. The nonrandomness at the level of sentence structure does not influence the expected vocabulary size. As a global text characteristic, it is probably insensitive to the strictly local constraints imposed by syntax. Apparently, it is the sequential order in which sentences actually appear that crucially determines the bias of our theoretical estimates. There are at least two domains where this sequential order might be relevant: the global domain of the discourse structure of the text as a whole, and the more local domain of relatively small sequences of sentences sharing a particular topic.

To explore these two potential explanatory domains in detail, we need a method for linking topical discourse structure and local topic continuity with word usage. Lexical specialization, informally defined as topic-linked concentrated word usage, and formalized in terms of underdispersion, provides us with the required tool.

3.2 Lexical Specialization

Recall that the word *Ahab* is unevenly distributed in *Moby Dick.* Given its high frequency (510), one would expect it to occur in all 40 text slices, but it does not. In fact, there are 11 text slices where Ahab is not mentioned at all. Technically speaking, *Ahab* is **underdispersed**. If there are many such words, and if these underdispersed words cluster together, the resulting deviations from randomness may be substantial enough to become visible as a divergence between the observed and theoretical growth curves of the vocabulary.

In order to explore this intuition, we need a reliable way to ascertain whether a word is underdispersed. Let the dispersion d_i of a word ω_i be the number of different text slices in which ω_i appears. Analytical expressions for $\mathrm{E}[d_i]$ and $\mathrm{VAR}[d_i]$ are available (Johnson and Kotz 1977, 113–114), so that in principle Z-scores can be calculated. These Z-scores can then be used to ascertain which words are significantly underdispersed in that they occur in significantly too few text slices given the urn model (cf. Baayen, 1996). Unfortunately, dispersions deviate substantially from normality, so that Z-scores remain somewhat impressionistic. I have therefore used a randomization test to ascertain which words are significantly underdispersed.

The randomization test proceeded as follows: The sequence of words of a text was randomized 1,000 times. For each permutation, the dispersion of each word type in that particular permutation was obtained. For each word, we calculated the proportion of permutations for which the dispersion was lower than or equal to the empirical dispersion. For *Ahab,* all 1,000 permutations revealed full dispersion ($d = 40$), which suggests that the probability that the low empirical dispersion of *Ahab* ($d = 28$) is due to chance is (much) less than .001.[5] The content words singled out as being signifi-

[5] I am indebted to an anonymous referee for pointing out to me that Z-scores are imprecise. I am

cantly underdispersed at the 1% level (the significance level I will use throughout this study for determining underdispersion) reveal a strong tendency to be **key words**. For instance, for *Moby Dick*, the ten most frequent underdispersed content words are *Ahab, boat, captain, said, white, Stubb, whales, men, sperm*, and *Queequeg*. The five most frequent underdispersed function words are *you, ye, such, her*, and *any*.[6]

The number of chunks in which an underdispersed word appears, and the frequencies with which such a word appears in the various chunks, cannot be predicted on the basis of the urn model. (Instead of the binomial or Poisson models, the negative binomial has been found to be a good model for such words, see, e.g., Church and Gale [1995]). Before studying how these words appear in texts and how they affect the growth curve of the vocabulary, it is useful to further refine our definition of underdispersion.

Consider again the distribution of the word *Ahab* in Figure 2. In text slice 25, *Ahab* occurs only once. Although this single occurrence contributes to the inter-textual cohesion of the novel as a whole, it can hardly be said to be a key word within text slice 25. In order to eliminate such spurious instances of key words, it is useful to set a frequency threshold. The threshold used here is that the frequency of the word in a given text slice should be at least equal to the mean frequency of the word calculated for the text slices in which the word appears. More formally, let $f_{i,k}$ be the frequency of the i-th word type in the k-th text slice, and define the indicator variable $d_{i,k}$ as follows:

$$d_{i,k} = \begin{cases} 1 & \text{iff } \frac{f_i}{d_i} \geq f_{i,k} \text{ and } \omega_i \text{ underdispersed} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The number of underdispersed types in text slice k, $VU(k)$, and the corresponding number of underdispersed tokens, $NU(k)$, can now be defined as

$$VU(k) = \sum_i d_{i,k} \tag{7}$$

$$NU(k) = \sum_i d_{i,k} \cdot f_{i,k}. \tag{8}$$

3.3 Lexical Specialization and Discourse Structure

We are now in a position to investigate where underdispersed words appear and how they influence the observed growth curve of the vocabulary. First consider Figure 4, which summarizes a number of diagnostic functions for *Alice in Wonderland*. The upper panels plot $VU(k)$ (left) and $NU(k)$ (right), the numbers of underdispersed types and tokens appearing in the successive text chunks. Over sampling time, we observe a slight increase in both the numbers of tokens and the numbers of types. Both trends are significant according to least squares regressions, represented by dotted lines ($F(1,38) = 6.591, p < .02$ for $VU(k)$; $F(1,38) = 16.58, p < .001$ for $NU(k)$). A time-series smoother using running medians (Tukey 1977), represented by solid lines,

similarly indebted to Fiona Tweedie, who suggested the use of the randomization test. Comparison of the results based on Z-scores (see Baayen, to appear) and the results based on the randomization test, however, reveal only minor differences that leave the main patterns in the data unaffected.

[6] The present method of finding underdispersed words appears to be fairly robust with respect to the number of text slices K. For different numbers of text chunks, virtually the same high-frequency words appear to be underdispersed. The number of text chunks exploited in this paper, 40, has been chosen to allow patterns in "sampling time" to become visible without leading to overly small text slices for the smaller texts.

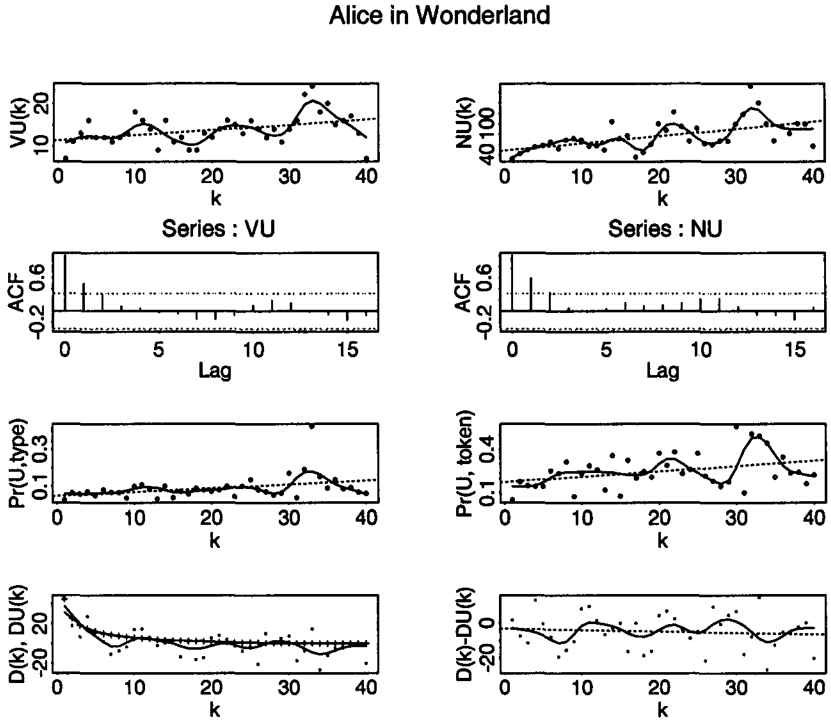

Figure 4
Diagnostic functions for *Alice in Wonderland*. *VU*(*k*) and *NU*(*k*): numbers of underdispersed types and tokens in text slice *k*; ACF: auto-correlation function; Pr(U, type) and Pr(U, token): proportions of underdispersed types and tokens; *D*(*k*) and *DU*(*k*): progressive difference scores for the overall vocabulary and the underdispersed words.

suggests a slightly oscillating pattern. At least for a time lag of 1, this finds some support in the autocorrelation functions, shown in the second line of panels of Figure 4. Clearly, key words are not uniformly distributed in *Alice in Wonderland*. Not only does the use of key words in one text slice appear to influence the intensity with which key words are used in the immediately neighboring text slices, but as the novel proceeds key words appear with increasing frequency.

How does this nonrandom organization of key words in the discourse as a whole influence $V(N)$? To answer this question, it is convenient to investigate the nature of the new types that arrive with the successive text slices. Let

$$\Delta V(M_k) = V(M_k) - V(M_{k-1}) \qquad (9)$$

denote the number of new types observed in text slice k, and let

$$\Delta VU(M_k) = VU(M_k) - VU(M_{k-1}) \qquad (10)$$

denote the number of new underdispersed types for text slice k. The proportion of new underdispersed types in text slice k on the total number of new types, $\mathrm{Pr}(U,\ \mathrm{type}, k)$ is given by

$$\mathrm{Pr}(U,\ \mathrm{type}, k) = \frac{\Delta VU(k)}{\Delta V(k)}. \qquad (11)$$

The plot of $\mathrm{Pr}(U,\ \mathrm{types}, k)$ is shown on the third row of Figure 4 (left-hand panel). According to a least squares regression (dotted line), there is a significant increase in the proportion of underdispersed new types as k increases ($F(1,38) = 5.804, p < .05$). The right-hand side counterpart shows a similar trend for the word tokens that is also supported by a least squares regression ($F(1,38) = 5.681, p < .05$). Here, the proportion of new underdispersed tokens on the total number of new tokens is defined as

$$\Pr(U,\ \text{token}, k) = \frac{\sum_i nU_{i,k}}{\sum_i n_{i,k}}, \tag{12}$$

with

$$n_{i,k} = \begin{cases} f_{i,k} & \text{iff } \sum_{m=1}^{k-1} f_{i,m} = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

and

$$nU_{i,k} = \begin{cases} f_{i,k} & \text{iff } \sum_{m=1}^{k-1} f_{i,m} = 0 \text{ and } d_{i,k} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The increase in the proportions of new underdispersed types and tokens shows that the pattern observed for the absolute numbers of types and tokens observed in the top panels of Figure 4 persists with respect to the new types and tokens.

We can now test to what extent the underdispersed types are responsible for the divergence of $\mathrm{E}[V(N)]$ and its expectation by comparing the progressive difference scores $D(k)$ defined in (5) with the progressive difference scores for the subset of the underdispersed words $DU(k)$, defined as

$$DU(k) = \mathrm{E}[VU(k)] - \mathrm{E}[VU(k-1)] - \Delta VU(k). \tag{15}$$

The two progressive difference score functions are shown in the bottom left panel of Figure 4, and the **residuals** $D(k) - DU(k)$ are plotted in the bottom right-hand panel. The residuals do not reveal any significant trend ($F(1,38) < 1$), which suggests that the underdispersed vocabulary is indeed responsible for the main trend in the progressive difference scores $D(k)$ of the vocabulary and hence for the divergence between $\mathrm{E}[V(N)]$ and $V(N)$. In the next section, I will argue that intra-textual cohesion is in large part responsible for the general downward curvature of $DU(k)$. In what follows, I will first present an attempt to understand the differences in the error scores $\mathrm{E}[V(N)] - V(N)$ shown in Figure 1 as a function of differences in the use of key words at the discourse level.

In *Alice in Wonderland,* key words are relatively rare in the initial text slices. As a result, these text slices reveal fewer types than expected under chance conditions. Consequently, $V(N)$ is smaller than $\mathrm{E}[V(N)]$. For increasing k, as shown in the upper right panel of Figure 1, the divergence between $V(N)$ and its expectation first increases—the initial text slices contain the lowest numbers of underdispersed types and tokens—and then decreases as more and more underdispersed words appear. Thus the semi-circular shape of the error scores $\mathrm{E}[V(N)] - V(N)$ shown in Figure 1 is a direct consequence of the topical structure at discourse level of *Alice in Wonderland.*

The error scores $\mathrm{E}[V(N)] - V(N)$ for *Moby Dick* and *Max Havelaar* shown in Figure 1 reveal a different developmental profile. In these novels, the maximal divergence appears early on in the text, after which the divergence decreases until, just before the end, $V(N)$ becomes even slightly larger than its expectation. Is it possible to understand this qualitatively different pattern in terms of the discourse structure of these novels? First, consider *Moby Dick.* A series of diagnostic plots is shown in Figure 5. The numbers of underdispersed types and tokens $VU(k)$ and $NU(k)$ reveal some variation, but unlike in *Alice in Wonderland,* there is only a nonsignificant trend

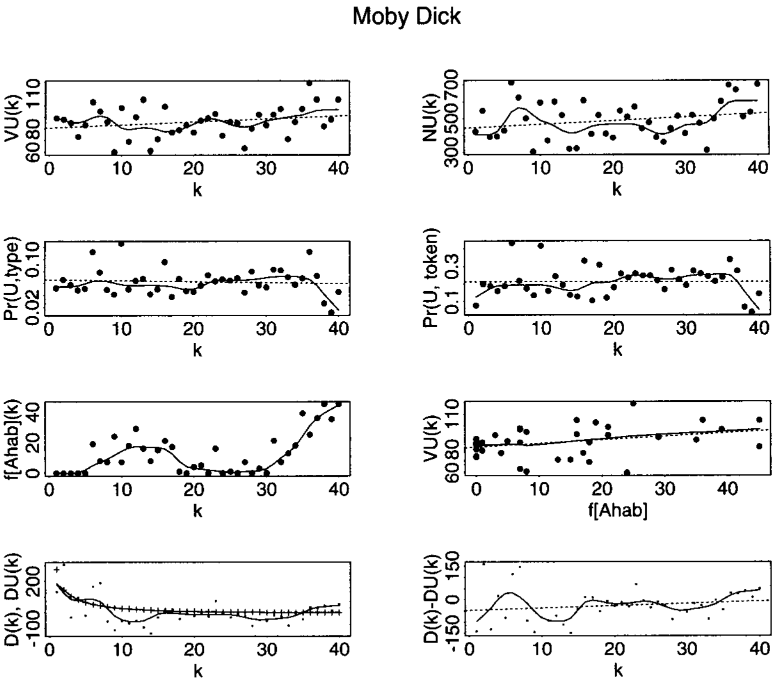

Figure 5
Diagnostic functions for *Moby Dick*. $VU(k)$ and $NU(k)$: numbers of underdispersed types and tokens in text slice k; Pr(U, type) and Pr(U, token): proportions of underdispersed types and tokens; $D(k)$ and $DU(k)$: progressive difference scores for the overall vocabulary and the underdispersed words; f[Ahab](k): frequency of *Ahab* in text slice k.

($F(1,38) = 2.11, p > .15$ for $VU(k)$, $F(1,38) = 1.98, p > .15$ for $NU(k)$) for underdispersion to occur more often as the novel progresses. The absence of a trend is supported by the proportions of underdispersed types and tokens, shown in the second row of panels ($F < 1$ for both types and tokens). In the last text slices, underdispersed words are even underrepresented. The bottom panels show that the progressive difference scores $DU(k)$ for the underdispersed words capture the main trend in the progressive difference scores of the total vocabulary $D(k)$ quite well: The residuals $D(k) - DU(k)$ do not reveal a significant trend ($F(1,38) = 1.08$, $p > .3$).

Interestingly, the use of underdispersed words in *Moby Dick* is to some extent correlated with the frequency of the word *Ahab*, with respect to both types and tokens ($F(1,38) = 4.61, p < .04, r^2 = .11$ for $VU(k)$; $F(1,38) = 10.77, p < .003, r^2 = .22$ for $NU(k)$. The panels on the third row of Figure 5 show the frequencies of *Ahab* (left) and $VU(k)$ as a function of the frequency of *Ahab* (right). A nonparametric time series smoother (solid line) supports the least squares regression line (dotted line). In other words, the key figure of *Moby Dick* induces a somewhat more intensive use of the key words of the novel.

The nonuniform distribution of *Ahab* sheds some light on the details of the shape of the difference function $E[V(N)] - V(N)$ shown in Figure 1. The initial sections do not mention Ahab, it is here that $D(k)$ reveals its highest values, and here too we find the largest discrepancies between $E[V(N)]$ and $V(N)$. By text slice 20, Ahab has been firmly established as a principal character in the novel, and the main key words have

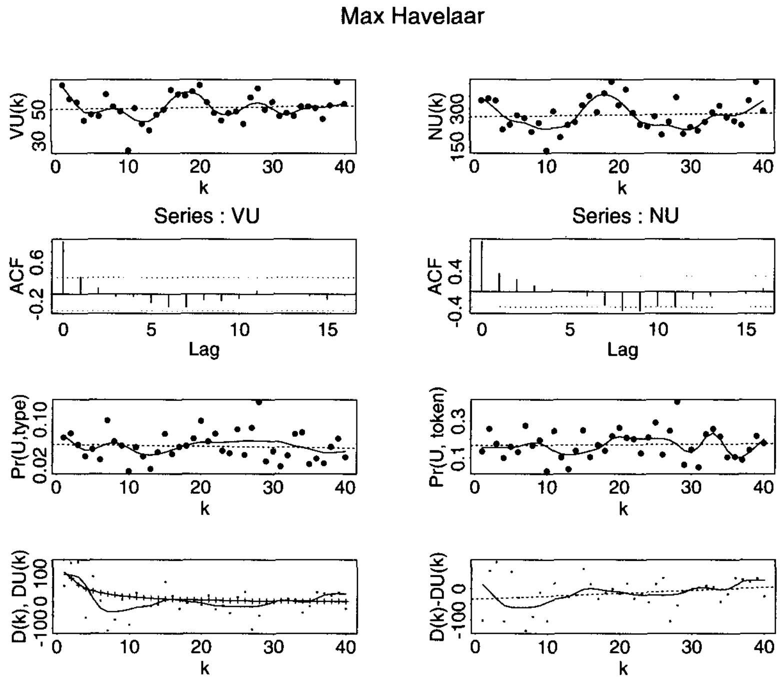

Figure 6
Diagnostic functions for *Max Havelaar*. *VU*(*k*) and *NU*(*k*): numbers of underdispersed types and tokens in text slice *k*; ACF: auto-correlation function; Pr(U, type) and Pr(U, token): proportions of underdispersed types and tokens; *D*(*k*) and *DU*(*k*): progressive difference scores for the overall vocabulary and the underdispersed words.

appeared. The overestimation of the vocabulary is substantially reduced. As the novel draws to its dramatic end, the frequency of *Ahab* increases to its maximum. The plots on the first row of Figure 5 suggest that underdispersed types and tokens are also used more intensively in the last text slices. However, the proportions plots on the second row show a final dip, suggesting that at the very end of the novel, a more than average number of normally dispersed new types appears. Considered together, this may explain why at the very end of the novel the expected vocabulary slightly underestimates the observed vocabulary size, as shown in Figure 1.

Finally, consider the diagnostic plots for *Max Havelaar*, shown in Figure 6. The time series smoother (solid line) for the absolute numbers of underdispersed types ($VU(k)$) and tokens ($NU(k)$) suggests an oscillating use of key words without any increase in the use of key words over time (the dotted lines represent the least squares regression lines, neither of which are significant: $F < 1$ in both cases). This oscillatory structure receives some support from the autocorrelation functions shown in the second row of panels. Especially in the token analysis, there is some evidence for positive autocorrelation at lag 1, and for a negative polarity at time lags 8 and 9. No trend emerges from the proportions of new underdispersed types and tokens (third row, $F < 1$ in both analyses). A comparison of the progressive difference scores $D(k)$ and $DU(k)$ (bottom row) shows that the underdispersed words are again largely responsible for the large values of $D(k)$ for small k. No significant trend remains in the residuals $D(k) - DU(k)$ ($F(1,38) = 1.848, p > .15$).

Figure 1 revealed that $\mathrm{E}[V(N)] - V(N)$ is largest around text slices 3 to 7, but becomes negative for roughly the last third of the novel. This pattern may be due to the oscillating use of key words in *Max Havelaar*. Although there is a fair number of key words in the first few text chunks, the intensity of key words drops quickly, only to rise again around chunk 20. Thus, key words are slightly underrepresented in the first part of the novel, allowing the largest divergence between the expected and observed vocabulary size to emerge there.

3.4 The Paragraph as the Domain of Topic Continuity

The preceding analyses all revealed violations of the randomness assumption underlying the urn model that originate in the topical structure of the narrative as a whole. I have argued that a detailed analysis of the distribution of key word tokens and types may shed some light on why the theoretical vocabulary size sometimes overestimates and sometimes underestimates the observed vocabulary size. We are left with the question of to what extent repeated use of words within relatively short sequences of sentences, henceforth for ease of reference **paragraphs**, affects the accuracy of $\mathrm{E}[V(N)]$. I therefore carried out two additional analyses, one using five issues of the Dutch newspaper Trouw, and one using the random samples of the Dutch newspaper De Telegraaf available in the Uit den Boogaart (1975) corpus. For both texts, no overall topical discourse structure is at issue, so that we can obtain a better view of the effects of intra-textual cohesion by itself.

For each newspaper, the available texts were brought together in one large corpus, preserving chronological order. Each corpus was divided into 40 equally large text slices. The upper left panel of Figure 7 shows that in the consecutive issues of Trouw (March 1994) the expected vocabulary size differs significantly from the observed vocabulary size for all of the first 20 measurement points, the domain for which significance can be ascertained (see footnote 3). The upper right panel reveals that for the chronologically ordered series of samples from De Telegraaf in the Uit den Boogaart corpus (268 randomly sampled text fragments with on average 75 word tokens) only 3 text chunks reveal a significant difference between $\mathrm{E}[V(N)]$ and $V(N)$. The bottom panels of Figure 7 show the corresponding plots of the progressive difference scores for the complete vocabulary ($D(k)$, ".") and underdispersed words ($DU(k)$, "+"). The least squares regression lines (dotted) for $D(k)$, supported by nonparametric scatterplot smoothers (solid lines), reveal a significant negative slope ($F(1,38) = 6.89, p < .02$ for Trouw, $F(1,38) = 10.99, p < .001$ for De Telegraaf). The residuals $D(k) - DU(k)$ do not reveal any significant trends ($F < 1$ for both newspapers). Note that for De Telegraaf $DU(k)$ does not capture the downward curvature of $D(k)$ as well as it should for large k. This may be due to the relatively small number of words that emerge as significantly underdispersed for this corpus.

Figure 7 shows that intra-textual cohesion within paragraphs is sufficient to give rise to substantial deviation between $\mathrm{E}[V(N)]$ and $V(N)$ in texts with no overall discourse organization. Within successive issues of a newspaper, in which a given topic is often discussed on several pages within the same newspaper, and in which a topic may reappear in subsequent issues, strands of inter-textual cohesion may still contribute significantly to the large divergence between the observed and expected vocabulary size. It is only by randomly sampling short text fragments, as for the data from the Uit den Boogaart corpus, which contains samples evenly spread out over a period of one year, that a substantial reduction in overestimation is obtained. Note, however, that even for the corpus data we again find that the expectation of $V(N)$ is consistently too high. Within paragraphs, words tend to be reused more often than expected under change conditions. This reuse pre-empts the use of other word tokens, among which

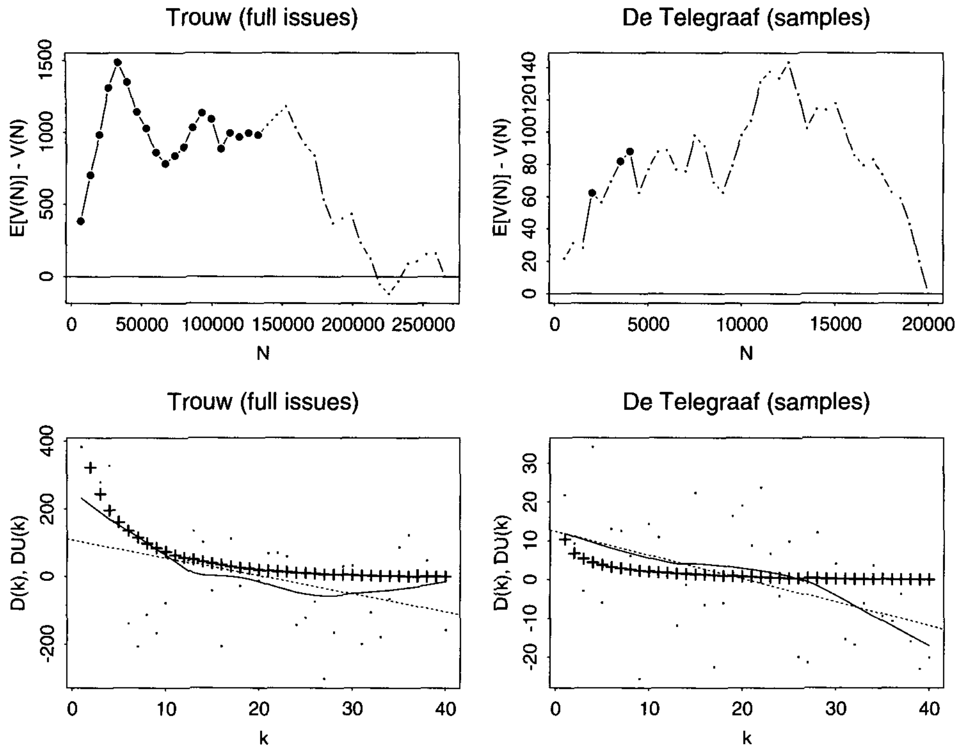

Figure 7
Diagnostic plots for two Dutch newspapers. The difference between the expected and observed vocabulary size for the Trouw data (five issues from March 1994) and the random samples of De Telegraaf in the Uit den Boogaart coipus (upper panels; significant differences are highlighted for the first 20 measurement points). The bottom panels show the progressive difference error scores for the total vocabulary ($D(k)$) and for the subset of underdispersed words ($DU(k)$). The dotted line is a least squares regression, the solid line a nonparametric scatterplot smoother.

tokens of types that have not been observed among the preceding tokens, and leads to a decrease in type richness. Since intra-textual cohesion is also present in the texts of novels, we may conclude that the overestimation bias in novels is determined by a combination of intra-textual and inter-textual cohesion.

4. Implications

We have seen that intra-textual and inter-textual cohesion lead to a significant difference between the expected and observed vocabulary size for a wide range of sample sizes. This section addresses two additional questions. First, to what extent does the nonrandomness of word occurrences affect distributions of units selected or derived from words? Second, how does cohesive word usage affect the Good-Turing frequency estimates?

4.1 Word-derived Units

First consider the effect of nonrandomness on the frequency distributions of morphological categories. The upper panels of Figure 8 plot the difference between the expected and observed vocabulary size for the morphological category of words with

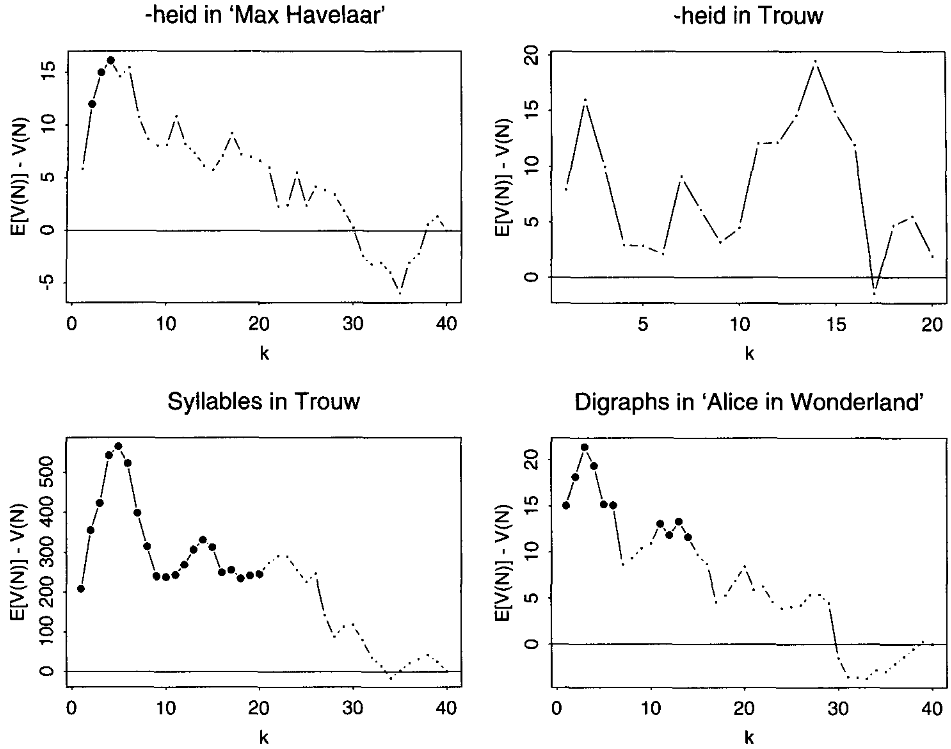

Figure 8
Diagnostic plots for affixes, syllables, and digraphs. The difference between the expected and observed vocabulary size for the morphological category of words with the Dutch suffix *-heid* '-ness' in *Max Havelaar* (upper left) and in Trouw (upper right), for syllables in Trouw (lower left), and for digraphs in *Alice in Wonderland*. Significant differences are shown in bold for the first half of the tokens.

the Dutch suffix *-heid*, which, like *-ness* in English, is used to coin abstract nouns from adjectives (e.g., *snelheid*, 'speed', from *snel*, 'quick'). The plots are based on samples consisting of all and only those words occurring in *Max Havelaar* (upper left) and Trouw (upper right) that belong to the morphological category of *-heid*, ignoring all other words, and preserving their order of appearance in the original texts. The sample of *-heid* words in *Max Havelaar* consisted of 640 tokens representing 260 types, of which 146 hapax legomena. From Trouw, 1145 tokens representing 394 types were extracted, among which 246 hapax legomena.

In *Max Havelaar*, a number of words in *-heid*, such as *waarheid* 'truth' and *vrijheid* 'freedom', are underdispersed key words. Not surprisingly, this affects the growth curve of *-heid* itself. For small values of k, we observe a significant divergence between $E[V(N)]$ and $V(N)$. In the newspaper Trouw, where *-heid* words do not play a central role in an overall discourse, no significant divergence emerges. Nevertheless, we again observe a consistent trend for the expected vocabulary size to overestimate the actual vocabulary size.

Figure 8 also plots the development of the vocabulary of syllables in Trouw (bottom left), and the development of the vocabulary of digraphs in *Alice in Wonderland* (bottom right). The "texts" of syllables and digraphs preserve the linear order of the texts from which they were derived. For both digraphs (80,870 tokens representing 398 types, of which 30 hapax legomena) and syllables (470,520 tokens, 6,748 types, and 1,909 hapax legomena), Figure 8 reveals significant deviation in the first half of both texts. This suggests that the nonrandomness observed for words carries over to word-based units such as digraphs and syllables.

4.2 Accuracy of Good-Turing Estimates

Samples of words generally contain—often small—subsets of all the different types available in the population. The probability mass of the unseen types is generally large enough to significantly bias population probabilities estimated from sample relative frequencies. Good (1953) introduced an adjusted frequency estimate (which he credits to Turing) to correct this bias. Instead of estimating the probability of a word with frequency f by its sample relative frequency

$$p_f = \frac{f}{N}, \tag{16}$$

Good suggests the use of the adjusted estimate

$$p_f^*(N) = \frac{1}{N} \cdot \frac{(f+1)\mathrm{E}[V(N,f+1)]}{\mathrm{E}[V(N,f)]}. \tag{17}$$

A closely related statistic is the probability $\mathcal{P}(N)$ of sampling a new, unseen type after N word tokens have been sampled:

$$\mathcal{P}(N) = \frac{\mathrm{E}[V(N,1)]}{N}. \tag{18}$$

These estimates are in wide use (see, e.g., Church and Gale [1991] for application to bigrams, Bod [1995] for application to syntax, and Baayen [1992] and Baayen and Sproat [1996] for application to morphology). Hence, it is useful to consider in some detail how their accuracy is affected by inter-textual and intra-textual cohesion. To this end, I carried out a short series of experiments of the following kind.

Assume that the Trouw data used in the previous section constitute a population of $N = 265{,}360$ word tokens from which we sample the first $N/2 = 132{,}680$ words. For the Trouw data, this is a matter of stipulation, but for texts such as *Moby Dick* or *Alice in Wonderland,* an argument can be made that the novel is the true population rather than a sample from a population. For the present purposes, the crucial point is that we now have defined a population for which we know exactly what the population probabilities—the relative frequencies in the complete texts—are.

First consider how accurately we can estimate the vocabulary size of the population from the sample. The expression for $\mathrm{E}[V(N)]$ given in (1) that we have used thus far does not allow us to extrapolate to larger sample sizes. However, analytical expressions that allow both interpolation (in the sense of estimating $V(N)$ on the basis of the frequency spectrum for sample sizes $M < N$) and extrapolation (in the sense of estimating $V(M)$ for $M > N$) are available (for a review, see Chitashvili and Baayen [1993]). Here, I will make use of a smoother developed by Sichel (1986). The three parameters of this smoother are estimated by requiring that $\mathrm{E}[V(N)] = V(N)$, that $\mathrm{E}[V(N,1)] = V(N,1)$, and by minimizing the chi-square statistic for a given span of frequency ranks.

The upper left panel of Figure 9 shows that it was possible to select the parameters of Sichel's model such that the observed frequencies of the first 20 frequency ranks $(V(N,f), f = 1,\ldots,20)$ do not differ significantly from their model-dependent

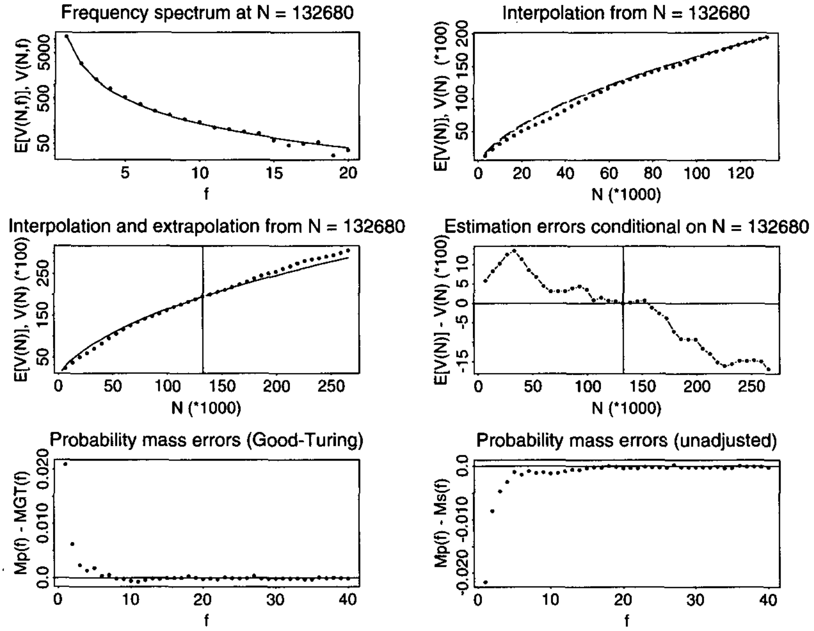

Figure 9
Interpolation and extrapolation from sample (the first half of the Trouw data) to population (the complete Trouw data). E[$V(N,f)$] and $V(N,f)$: expected and observed frequency spectrum; E[$V(N)$] and $V(N)$: expected and observed numbers of types; $Mp(f)$: population probability mass of the types with frequency f in the sample; $MGT(f)$: Good-Turing estimate of $Mp(f)$; $Ms(f)$: unadjusted sample estimate of $Mp(f)$.

expectations $E_S[V(N,f)]$.[7] The upper right panel shows that interpolation on the basis of Sichel's model (dashed line) is virtually indistinguishable from interpolation using (1) (dotted line). The observed vocabulary sizes are represented by large dots. As expected, both (1) and the parametric smoother reveal the characteristic overestimation pattern.

The center panels of Figure 9 show that the overestimation characteristic for interpolation is reversed when extrapolating to larger samples. For extrapolation, underestimation is typical. The dotted line in the left-hand panel represents the observed vocabulary size of the complete Trouw text, the solid line shows the result from interpolation and extrapolation from $N = 132{,}680$. The right-hand panel highlights the corresponding difference scores. For $N = 265{,}360$, the error is large: 5.5% of the actual vocabulary size.

Having established that E[$V(N)$] underestimates $V(N)$ when extrapolating, the question is how well the Good-Turing estimates perform. To determine this, I will consider the probability mass of the frequency classes $V(M,f)$ for $f = 1 \ldots 40$. Let

$$M_{GT}(f,M) = V(M,f) \cdot p_f^*(M) \tag{19}$$

[7] The fit ($X^2(18) = 9.93, p > .9$) was obtained for the parameter values $\alpha = 0.291, \gamma = -0.7$, and $b = 0.011$.

be the joint Good-Turing probability mass of all types with frequency f in the sample of $M = 132{,}680$ tokens, and let $M_p(f)$ be the joint probability mass of exactly the same word types, but now in the population ($N = 265{,}360$ tokens):

$$M_p(f) = \frac{\sum_i \mathrm{I}_{[f(i,M)=f]} \cdot f(i,N)}{N}, \qquad (20)$$

with $f(i,X)$ the frequency of the i-th type in a sample of X tokens. The bottom left panel of Figure 9 shows that for the first frequency ranks f, the Good-Turing estimate $M_{GT}(f,M)$ underestimates the probability mass of the frequency class in the population. For the higher-frequency ranks, the estimates are fairly reliable. The bottom right panel of Figure 9 plots the corresponding errors for the unadjusted sample probability estimate

$$M_s(f,M) = \frac{f}{M} V(M,f), \qquad (21)$$

which overestimates the population values. Surprisingly, the unadjusted estimates overestimate the population values to roughly the same extent that the adjusted estimates lead to underestimation. A heuristic estimate,

$$M_h(f,M) = \frac{V(M,f)}{\mathrm{E}_S[V(M,f)]} \cdot \frac{(f+1)\mathrm{E}_S[V(M,f+1)] + f\mathrm{E}_S[V(M,f)]}{2M} \qquad (22)$$

the mean of $M_s(f,M)$ and $M_{GT}(f,M)$, appears to approximate the population relative class frequencies $M_p(f)$ reasonably well, as shown in Table 1 for the Trouw data as well as for *Alice in Wonderland*, *Moby Dick*, and *Max Havelaar*. For $f > 5$, as shown in Figure 10, the heuristic estimate remains a reasonable compromise.

We have seen that both $\mathrm{E}[V(N)]$ and the Good-Turing estimates $M_{GT}(f,M)$ (especially for $f \leq 5$) lead to underestimation of population values. Interestingly, $\mathcal{P}(M)$ overestimates the probability mass of unseen types. For the Trouw data, at $M = 132{,}680$ we count 11,363 hapax legomena, hence $\mathcal{P}(M) = 0.0856$. However, the probability mass of the types that do not appear among the first 132,680 tokens, $M(0)$, is much smaller: 0.0609. Table 1 shows that $\mathcal{P}(M)$ similarly leads to overestimation for *Alice in Wonderland, Moby Dick*, and *Max Havelaar*. To judge from Table 1, the Good-Turing estimate $M_{GT}(1,M)$ is an approximate lower bound and the unadjusted estimate $M_s(1,M)$ a strict upper bound for $M_p(0)$.

It is easy to see why $\mathcal{P}(N)$ is an upper bound for coherent text by focusing on its interpretation. Given the urn model, the probability that the first token sampled represents a type that will not be represented by any other token equals $V(N,1)/N$. By symmetry, this probability is identical to the probability that the very last token sampled will represent an unseen type. This probability approximates the probability that, after N tokens have been sampled, the next token sampled will be a new type. However, this interpretation hinges on the random selection of word tokens, and this paper presents ample evidence that once a word has been used it is much more likely to be used again than the urn model predicts. Hence, the probability that after sampling N tokens the next token represents an unseen type is less than $V(N,1)/N$. Due to intra-textual and inter-textual cohesion, the $V(N) - V(N,1)$ types that have already been observed have a slightly higher probability of appearing than expected under chance conditions, and consequently the unseen types have a lower probability.

Summing up, the Good-Turing frequency estimates are severely effected by the cohesive use of words in normal text. In the absence of probabilistic models that take cohesive word usage into account, estimates of (relative) frequencies remain heuristic

Table 1

Comparison of probability mass estimates for frequencies $f = 1, \ldots, 5$ using the smoother $\mathrm{E_S}[V(N,f)]$ of Sichel (1986). The probability mass of unseen types, $M_p(0)$, is also tabulated. Notation: $M_{GT}(f,M)$: Good-Turing estimate; $M_s(f,M)$: sample estimate; $M_h(f,M)$: heuristic estimate; $M_p(f)$: population mass. For *Max Havelaar*, a sample comprising the first third of the novel was used, for the other texts, a sample consisting of the first half of the tokens was selected.

f	$V(N,f)$	$\mathrm{E_S}[V(N,f)]$	$M_{GT}(f,M)$	$M_s(f,M)$	$M_h(f,M)$	$M_p(f)$
Alice in Wonderland						
0						0.0630
1	885	885.00	0.0411	0.0668	0.0540	0.0560
2	287	272.27	0.0328	0.0434	0.0381	0.0372
3	147	137.27	0.0277	0.0333	0.0305	0.0293
4	97	85.52	0.0255	0.0293	0.0274	0.0289
5	68	59.55	0.0230	0.0257	0.0243	0.0228
Moby Dick						
0						0.0350
1	5,914	5,914.04	0.0366	0.0553	0.0460	0.0472
2	2,035	1,958.22	0.0272	0.0381	0.0326	0.0331
3	990	932.22	0.0218	0.0278	0.0248	0.0251
4	601	549.44	0.0187	0.0225	0.0206	0.0210
5	416	366.22	0.0168	0.0195	0.0181	0.0179
Max Havelaar						
0						0.0921
1	3,513	3,513.01	0.0494	0.1058	0.0776	0.0692
2	908	821.04	0.0362	0.0547	0.0455	0.0411
3	346	362.65	0.0240	0.0313	0.0276	0.0246
4	214	208.95	0.0213	0.0258	0.0235	0.0216
5	157	137.84	0.0203	0.0236	0.0220	0.0189
Trouw						
0						0.0609
1	11,363	11,363.05	0.0431	0.0856	0.0644	0.0639
2	2,941	2,856.65	0.0297	0.0443	0.0370	0.0359
3	1,338	1,276.07	0.0233	0.0303	0.0268	0.0256
4	826	737.69	0.0206	0.0249	0.0227	0.0219
5	532	487.51	0.0172	0.0200	0.0186	0.0190

in nature. For the frequencies of types occurring at least once in the sample, the average of the sample and Good-Turing adjusted frequencies is a useful heuristic. For estimates of the probability of unseen types, the sample and Good-Turing estimates provide approximate upper and lower bounds.

5. Discussion

Words do not occur randomly in texts. This simple fact is difficult to take into account in statistical models of word frequency distributions. Hence, it is often ignored, in the hope that violations of the randomness assumption will not seriously affect the accuracy of quantitative measures and estimates.

The goal of this paper has been to explore in detail the consequences of intra-

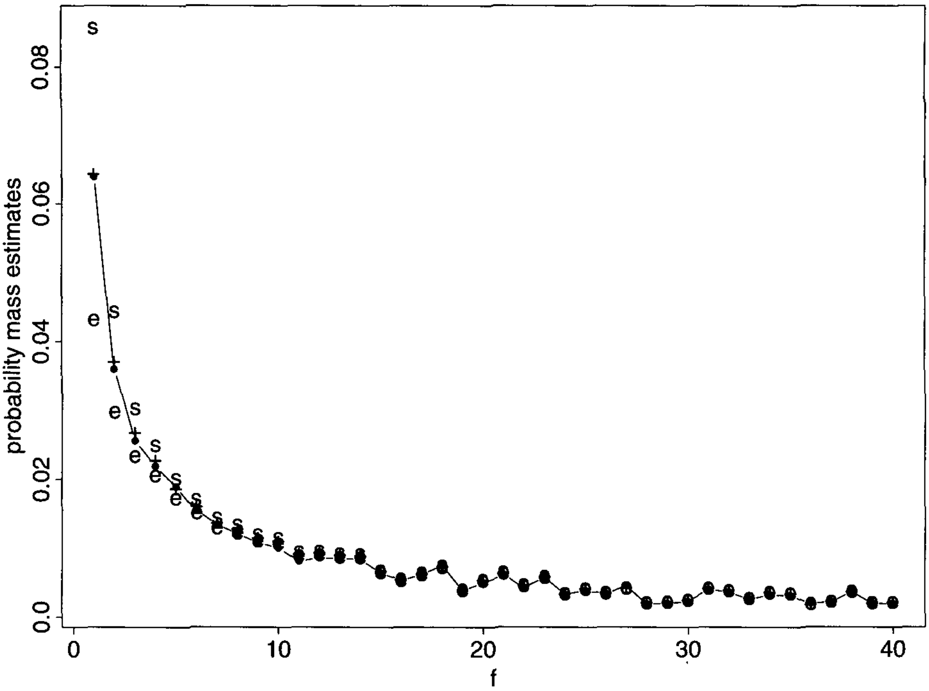

Figure 10
Frequency class probability mass estimates for the first 40 frequency ranks in a sample of $M = 132{,}680$ of the Trouw data. The dots denote the probability mass $M_p(f)$ in the full text ($N = 265{,}360$) of the words with frequency f in the sample. The Good-Turing estimates $M_{GT}(f, M)$ are represented by "e," the sample estimates $M_s(f, M)$ by "s," and the heuristic estimate $M_h(f, M)$ by "+".

textual and inter-textual cohesion on the accuracy of theoretical estimates of vocabulary size, the growth rate of the vocabulary, and Good-Turing adjusted frequency estimates, in the belief that knowledge of how nonrandomness might affect these measures ultimately leads to a better understanding of the conditions under which these measures may, or may not, be reliable.

Analyses of three novels, five consecutive issues of the Dutch newspaper Trouw, and the chronologically ordered samples of the Dutch newspaper De Telegraaf in the Uit den Boogaart corpus, all revealed systematic overestimation for the expected vocabulary size. Further analyses of subsets of derived words, syllables, and digrams showed that the overestimation bias reappears in units derived from words when these words occur in normal, cohesive text.

The overestimation bias disappears when the order of the sentences is randomized. This indicates that the bias should not be attributed to syntactic and semantic constraints on word usage operating within the sentence. Instead, the bias arises due to intra-textual and inter-textual cohesion. In sequences of sentences, words are more likely to be reused than expected under chance conditions. Coherent discourse requires local topic continuity. This intra-textual cohesion gives rise to a substantial part of the overestimation bias, a bias that leads to significant deviations even when small text fragments of some 75 words are selected randomly from a newspaper.

In addition to intra-textual cohesion, there are words that contribute to the cohesion of the discourse as a whole. Detailed analyses of how these key words appear over sampling time in the novels reveal marked differences in their distributions. These differences in turn shed light on the details of the differences in the patterns of estimation errors $\mathrm{E}[V(N)] - V(N)$ that characterize the texts. The progressive difference scores of the key words, the deviation scores for the expected and observed numbers of new types appearing in the successive text slices, reveal a pattern that is highly similar to the same scores for the vocabulary as a whole, both qualitatively and quantitatively. This supports the hypothesis that the key words are primarily responsible for the deviation of the expected vocabulary size from its expectation.

Nonrandomness in word usage not only introduces a bias with respect to the expected vocabulary size—overestimation when interpolating and underestimation when extrapolating, it also affects the accuracy of the Good-Turing estimates. To correct for an overestimation bias, Good (1953) introduced adjusted estimates, building on the assumption that word usage is to all practical purposes random. These adjusted estimates, however, appear to overshoot their mark for continuous text in that they underestimate the population relative frequencies to roughly the same extent that the unadjusted probabilities lead to overestimation, especially for the lowest frequencies. Again, the effect of inter-textual and intra-textual cohesion manifests itself. Once used, words tend to be used again, and this leads to a somewhat higher relative population frequency than expected. The other side of the same coin is that Good's estimate for the probability mass of unseen types, $\mathcal{P}(N)$, is an upper bound. The words that have already been used have a raised probability of being used again. Hence, the probability for unseen types to appear is lowered.

There are two major ways to deal with the effects of nonrandomness in word usage on the accuracy of statistical estimates. First, by randomly sampling individual sentences instead of sequences of sentences, the effects of intra-textual and inter-textual cohesion will be largely eliminated. With the increasingly large corpora that are becoming available at present, enhanced sampling methods should pose no serious problem. For literary studies, however, the discourse structure of a text is part and parcel of the object of study itself. Here, the use of the heuristically adjusted estimates proposed in Section 4.2 may prove to be useful.

Finally, the investigation of the distribution of key words may turn out to be a useful tool for investigating the structure of literary texts, a tool that may lead to an improved understanding of the role of lexical specialization in shaping the quantitative developmental structure of the vocabulary.

Appendix

Equation (1) can be derived as follows; see Good 1953; Good and Toulmin 1956; Kalinin 1965: Let $f(i, M)$ denote the frequency of ω_i in a sample of M tokens ($M < N$), and define

$$X_i = \begin{cases} 1 & \text{if } f(i,M) = m \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Denoting the probability of ω_i by p_i, the expected total number of word types with frequency m in a sample of M tokens, $\mathrm{E}[V(M,m)]$, is given by

$$\begin{aligned} \mathrm{E}[V(M,m)] &= \mathrm{E}[\sum_i \mathrm{X}_i] \\ &= \sum_i \mathrm{E}[\mathrm{X}_i] \end{aligned}$$

$$
\begin{aligned}
&= \sum_i 1 \cdot \Pr(X_i = 1) + 0 \cdot \Pr(X_i = 0) \\
&= \sum_i \binom{M}{m} p_i^m (1 - p_i)^{M-m}, \qquad (24)
\end{aligned}
$$

where we assume that the frequencies $f(i, M)$ are independently and identically binomially (M, p_i) distributed. The expected overall number of different types in the sample, irrespective of their frequency, follows immediately:

$$
\begin{aligned}
\mathrm{E}[V(M)] &= \mathrm{E}[\sum_{m \geq 1} V(M, m)] \\
&= \sum_{m \geq 1} \sum_i \binom{M}{m} p_i^m (1 - p_i)^{M-m} \\
&= \sum_i (1 - (1 - p_i)^M). \qquad (25)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathrm{E}[V(M, m)] &= \sum_i \frac{(\lambda_i M)^m}{m!} e^{-\lambda_i M} \\
\mathrm{E}[V(M)] &= \sum_i (1 - e^{-\lambda_i M}). \qquad (26)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathrm{E}[V(M)] &= \sum_{i=1}^{V(N)} (1 - e^{-\lambda_i M}) \\
&= \sum_{i=1}^{V(N)} (1 - e^{-\frac{f(i,N)}{N} M}) \\
&= V(N) - \sum_{f=1} V(N,f) e^{-\frac{f}{N} M}. \qquad (27)
\end{aligned}
$$

In the last step, all $V(N,f)$ types sharing the same frequency f have been grouped together. Note that when the N tokens themselves constitute a sample from a larger population, $\mathrm{E}[V(M)]$ is in fact an estimate.

The derivation of (27) uses an urn model in which words are sampled with replacement. A model in which words are sampled without replacement is more precise. For instance, for a randomly reordered text, the likelihood that a hapax-legomenon in the full text that appears in the first M tokens will reappear among the remaining $N - M$ tokens is greater than zero in a model that assumes constant probabilities, contrary to fact. For large N and M, however, the binomial probabilities (sampling with replacement) are a good approximation of the hypergeometric probabilities (sampling without replacement).

Finally note that (27) suggests that, under randomness, and conditional on the words appearing in the sample of N tokens, $f(i,M)$ can alternatively be viewed as a binomially distributed random variable with parameters M/N and $f(i,N)$ for $f(i,N) \ll M,N$ (Muller 1977):

$$\begin{aligned}\mathrm{E}[V(M)] &= V - \sum_f V(N,f)e^{-\frac{M}{N}f} \\ &\approx V - \sum_f V(N,f)\left(1-\frac{M}{N}\right)^f . \end{aligned} \tag{28}$$

The modification of (28) proposed by Hubert and Labbe (1988) requires the assumption that all the tokens of a word type with specialized use occur in a single text slice. Let the total number of words in the set $\mathcal{S}$ of types with specialized use be pV, and also assume that the text slices in which these specialized words appear are randomly distributed over the text. Let

$$X_i = \begin{cases} 1 & \text{if } \omega_i \in \mathcal{S} \text{ and } \omega_i \text{ occurs in } P_1 \\ 0 & \text{otherwise,} \end{cases} \tag{29}$$

and let

$$Y_i = \begin{cases} 1 & \text{if } \omega_i \notin \mathcal{S} \text{ and } \omega_i \text{ occurs in } P_1 \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

The overall number of types in P_1 is $\sum_i X_i + \sum_j Y_j$. If $\omega_i \in \mathcal{S}$, its f_i tokens ($f_i \ll M$) will all appear in the same part of the text. The probability that they will appear in P_1 is $\frac{M}{N}$. Hence

$$\begin{aligned}\mathrm{E}_{HL}[V(M)] &= \mathrm{E}\left[\sum_i X_i + \sum_j Y_j\right] \\ &= \sum_{\omega_i \in \mathcal{S}} \Pr(X_i = 1) + \sum_{\omega_j \notin \mathcal{S}} \Pr(Y_j = 1) \\ &= \sum_{\omega_i \in \mathcal{S}} \frac{M}{N} + \sum_{\omega_j \notin \mathcal{S}} \left(1 - \left(1-\frac{M}{N}\right)^{f_j}\right) \\ &= pV\frac{M}{N} + (1-p)V - \sum_{\omega_j \notin \mathcal{S}} \left(1-\frac{M}{N}\right)^{f_j} \\ &= pV\frac{M}{N} + (1-p)V - \sum_f (1-p)V(N,f)\left(1-\frac{M}{N}\right)^f \\ &= pV\frac{M}{N} + (1-p)V - \sum_f (1-p)V(N,f)e^{-\frac{M}{N}f}. \end{aligned} \tag{31}$$

Note the implicit assumption that the same proportion of the $V(N,f)$ word types with frequency f is specialized, irrespective of the value of f.

List of Symbols

d_i	dispersion of word type i
$d_{i,k}$	indicator variable for underdispersion of type i in chunk k
$D(k)$	progressive difference score for text slice k
$DU(k)$	progressive difference score of underdispersed words at chunk k
$\Delta V(k)$	number of new types at k
$\Delta VU(k)$	number of new underdispersed types at k
$\mathrm{E}[X]$	expectation of X
$\mathrm{E}_{HL}[V(M)]$	expectation of $V(M)$ in the Hubert-Labbe model
$\mathrm{E}_{S}[V(N,f)]$	expectation of $V(N,f)$ given Sichel's (1986) model
f	token frequency of a word
$f(i,M)$	token frequency of i-th word type in sample of size M
$f_{i,k}$	token frequency of i-th word in the k-th text slice
i	index for word types $1,\ldots,V$
k	index for text slices $1,\ldots,K$
K	number of text slices
m	token frequency of a word in sample of size M
M	sample size in tokens when contrasting two sample sizes ($M < N$)
$M_p(f)$	population probability mass of frequency class f
$M_{GT}(f,M)$	Good-Turing sample estimate of $M_p(f)$ in sample of size M
$M_s(f,M)$	sample estimate of $M_p(f)$
$M_h(f,M)$	heuristic estimate of $M_p(f)$ (mean of $M_s(M,f)$ and $M_{GT}(M,f)$)
$n_{i,k}$	indicator variable for type i appearing first in chunk k
$nU_{i,k}$	indicator variable for type i appearing first in chunk k and i being underdispersed in k
N	number of word tokens in the sample
$NU(k)$	number of underdispersed tokens in chunk k
p	Hubert-Labbe coefficient of vocabulary partition
p_f	sample probability (f/N)
p_i	probability of ω_i
$p_f^*(M)$	Good-Turing adjusted probability for sample of size M
$\Pr(U,\ \text{token},\ k)$	proportion of new underdispersed tokens at k
$\Pr(U,\ \text{type},\ k)$	proportion of new underdispersed types at k
$\mathcal{P}(N)$	growth rate of the vocabulary ($\mathrm{E}[V(N,1)]/N$)
$V(N)$	number of different word types among N tokens
$V(N,f)$	number of types with frequency f in a sample of N tokens
$V(M_k)$	number of types in the first $\frac{kN}{K}$ tokens
$VU(k)$	number of underdispersed types in chunk k

Acknowledgments

I am indebted to Ken Church, Anne Cutler, James McQueen, Rob Schreuder, Richard Sproat, Fiona Tweedie, and five anonymous reviewers for valuable comments and discussion.

References

Baayen, R. Harald. 1989. *A Corpus-based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Ph.D. thesis, Free University, Amsterdam.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In G. E. Booij and J. van Marle, editors, *Yearbook of Morphology 1991.* Kluwer Academic Publishers, Dordrecht, pages 109–149.

Baayen, R. Harald. 1996. The randomness assumption in word frequency statistics.

In G. Perissinotto, editor, *Research in Humanities Computing* 5. Oxford University Press, Oxford.

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

Baayen, R. Harald and Richard Sproat. 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. *Computational Linguistics*, 22(2):155–166.

Bod, Rens. 1995. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. University of Amsterdam: Institute for logic, language and computation, Amsterdam.

Brunet, Etienne. 1978. *Le vocabulaire de Jean Giraudoux*, volume 1 of *TLQ*. Slatkine, Genève.

Chitashvili, Revas J and R. Harald Baayen. 1993. Word frequency distributions. In G. Altman and L. Hřebíček, editors, *Quantitative Text Analysis*. Wissenschaftlicher Verlag Trier, Trier, pages 54–135.

Church, Kenneth and William Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

Church, Kenneth and William Gale. 1995. Poisson mixtures. *Journal of Natural Language Engineering*, 1(2):163–190.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

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

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

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

Holmes, David. I. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Hubert, Pierre and Dominique Labbe. 1988. A model of vocabulary partition. *Literary and Linguistic Computing*, 3:223–225.

Indefrey, Peter and R. Harald Baayen. 1994. Estimating word frequencies from dispersion data. *Statistica Neerlandica*, 48:259–270.

Johnson, Norman L. and Samuel Kotz. 1977. *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, New York.

Kalinin, V. M. 1965. 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. Steklov Institute of Mathematics 79, American Mathematical Society.

Khmaladze, Estate V and Revas J. Chitashvili. 1989. Statistical analysis of large number of rare events and related problems. *Transactions of the Tbilisi Mathematical Institute*, 91:196–245.

Muller, Charles. 1977. *Principes et méthodes de statistique lexicale*. Hachette, Paris.

Muller, Charles. 1979. *Langue française et linguistique quantitative*. Slatkine, Genève.

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

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.

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