Computers and the Humanities **32:** 323–352, 1998.

How Variable May a Constant be? Measures of Lexical Richness in Perspective

FIONA J. TWEEDIE[1] and R. HARALD BAAYEN[2]
[1]*University of Glasgow, United Kingdom;* [2]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

Key words: lexical statistics, Monte Carlo methods, vocabulary richness

Abstract. A well-known problem in the domain of quantitative linguistics and stylistics concerns the evaluation of the lexical richness of texts. Since the most obvious measure of lexical richness, the vocabulary size (the number of different word types), depends heavily on the text length (measured in word tokens), a variety of alternative measures has been proposed which are claimed to be independent of the text length. This paper has a threefold aim. Firstly, we have investigated to what extent these alternative measures are truly textual constants. We have observed that in practice all measures vary substantially and systematically with the text length. We also show that in theory, only three of these measures are truly constant or nearly constant. Secondly, we have studied the extent to which these measures tap into different aspects of lexical structure. We have found that there are two main families of constants, one measuring lexical richness and one measuring lexical repetition. Thirdly, we have considered to what extent these measures can be used to investigate questions of textual similarity between and within authors. We propose to carry out such comparisons by means of the empirical trajectories of texts in the plane spanned by the dimensions of lexical richness and lexical repetition, and we provide a statistical technique for constructing confidence intervals around the empirical trajectories of texts. Our results suggest that the trajectories tap into a considerable amount of authorial structure without, however, guaranteeing that spatial separation implies a difference in authorship.

1. Introduction

A time-honoured problem in the domain of quantitative linguistics is the evaluation of the lexical richness of texts. An obvious measure of lexical richness is the number of different words that appear in a text. Unfortunately, a text's vocabulary size depends on its length. Ever since Yule (1944)'s seminal study, a central question has been how to measure lexical richness by means of a statistic that does not depend on text length.

A great many textual measures are now available. Although these measures have gained some acceptance as length-invariant statistics, a number of researchers (Weitzman, 1971; Ménard, 1983; Orlov, 1983; Thoiron, 1986; Baayen, 1989; Cossette, 1994) have expressed doubts about the length-invariance of at least some of them. In this paper we will show that nearly all available measures are highly

dependent on text length. More specifically, we argue that there are two issues that need to be taken into account when evaluating the reliability of a given measure. Firstly, is a given statistic mathematically constant, given the simplifying, but technically convenient assumption that words are used randomly and independently? We will show that most proposed constants are in theory not constant at all. Secondly, how is a constant affected by violations of the randomness assumption in actual texts? Even those few measures that are theoretically truly constant might still reveal significant dependence on text length when applied to real texts. We will show that this is indeed the case: all measures reviewed here are subject to the effects of non-random word use. Our conclusion will be, therefore, that it is extremely hazardous to use lexical 'constants' to compare texts of different length.

For some measures of lexical richness, such as, for instance, the type-token ratio, this dependence on text length is well-known (see, e.g., Holmes, 1994, pp. 95–97). Unfortunately, the type to token ratio is still in use as a traditional stylometric measure (as in, e.g., Whissell, 1996, p. 259), and the same holds for its inverse, the mean word frequency (as in, e.g., Martindale and McKenzie, 1995, p. 261), without explicit reference to the role of text length or any explicit discussion of normalization with respect to text length. The theoretical dependence on the text length of almost all other measures reviewed in this paper also questions the legitimacy of their use in authorship studies (as in, e.g., Holmes, 1992, and Holmes and Forsyth, 1995). The first, negative, goal of this paper, then, is to advise against the use of lexical constants without correcting for possible differences in text length.

A second, positive, goal of this paper is to investigate to what extent lexical constants might still be of use in lexicometric studies. A possibility that we explore in detail is to turn this dependence on text length to our advantage by considering how the values of constants develop through the text. A case study of a sample of texts reveals that constants in works by different authors tend to change in different ways. We shall say that they exhibit different developmental profiles, that is, the plot of the constant against the text length tends to have a different shape in works by different authors. Conversely, texts by the same author tend to have remarkably similar developmental profiles. This suggests that these developmental profiles can be used as textual characteristics, rather than individual values of the constants for the full texts. We will therefore present methods for obtaining confidence intervals for such developmental profiles. In addition, we shall introduce the idea of partial randomisations, where text is permuted in sections to allow for confidence intervals to be constructed around the empirical values of the measures. We will show how discourse structure (the non-random patterning of sentences in narrative texts) can be taken into varying degrees of account in the construction of the confidence intervals.

Our comparisons suggest a classification of constants into disjunct families capturing different aspects of lexical use. For each family we will identify the statistic that in our experience has that greatest discriminatory power. We will

present the information captured by these measures in the form of trajectory plots, which allow us to take the information from both families of measures into account simultaneously.

Finally, we will evaluate our attempt at enhancing the reliability of lexical constants by means of an authorship attribution study that compares the classificatory power of lexical constants with the classificatory power of the most frequent function words as suggested by Burrows (1989). Our data suggest that the use of two independent constants that each are truly constant at least in theory uncovers a reasonable amount of authorial structure, but that optimal precision is obtainable only by taking many more lexical variables (such as the relative frequencies of function words) into account.

2. Measures of Lexical Richness

We begin our overview of measures of lexical richness by considering the most fundamental measure of all, the vocabulary size itself. The vocabulary size depends on the text length, N. As we read through a text, N increases from 1 to the total number of word tokens in the text. A word token is an instance of a particular word type. For instance, the preceding sentence contains two tokens of the type *a*. As the text length increases, the number of different word types encountered also increases, quickly at first, then more slowly as additional text is read. The first panel of Figure 1 illustrates this functional dependence of the number of types on the number of tokens for Lewis Carroll's *Alice's Adventures in Wonderland.* The horizontal axis displays the text length in word tokens, the vertical axis shows the vocabulary size in word types. The second panel plots the growth rate of the vocabulary

$$P(N) = \frac{V(1, N)}{N} \tag{1}$$

as a function of N(Good, 1953; Chitashvili and Baayen, 1993), where $V(i, N)$ denotes the number of types occurring i times in the text at length N. The number of types occurring once, $V(1, N)$ is generally referred to as the number of *hapax legomena*. This plot highlights the diminishing rate at which the vocabulary increases through the text.

The dynamics of vocabulary development affect two other simple statistics; the mean word frequency,

$$MWF(N) = \frac{N}{V(N)}, \tag{2}$$

and its reciprocal, the type token ratio,

$$TTR(N) = \frac{V(N)}{N}. \tag{3}$$

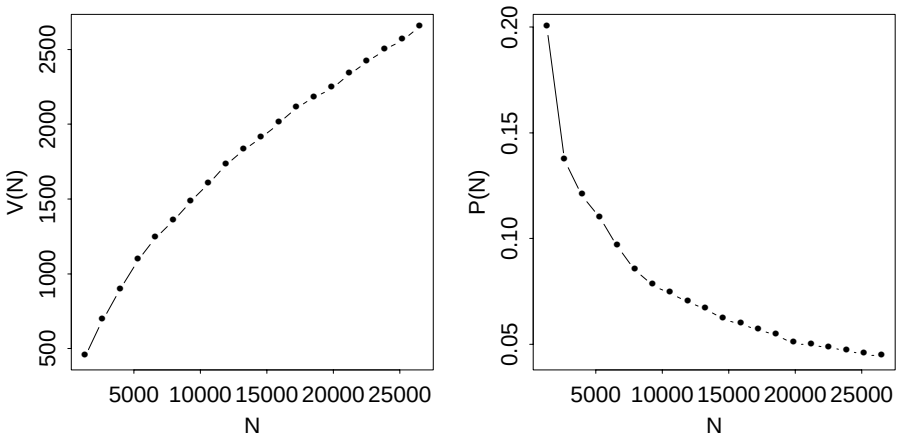

Figure 1. The vocabulary size $V(N)$, and its growth rate $P(N)$ as a function of text length N for *Alice's Adventures in Wonderland* at twenty equally-spaced measurement points.

We have made explicit in our notation that these two measures are functions of N, a property that they inherit from the vocabulary size $V(N)$. Baker (1988), rather confusingly, presented the mean word frequency as a measure of vocabulary richness; he calls it *Pace*. However, his calculations indicate that he is in fact using the type-token ratio.

The dependency of these two measures on N is illustrated in Figure 2. The inherent variability of the mean word frequency exemplifies the nature of the problem which has led to the development of a series of alternative lexical measures: the failure of the most obvious summary statistics for lexical richness to characterise a text irrespective of its length.

Three lines of approach have been taken to obtain measures that are independent of N. In the first instance, simple functions of N and $V(N)$, such as the square root and the logarithm, are used to eliminate the curvature of $V(N)$ illustrated in Figure 1. In the second approach the spectrum elements, $V(i, N)$, the numbers of types occurring i times in a sample of length N, are taken into account. Finally, the parameters of probabilistic models for lexical distributions can be considered. We will discuss each of these approaches in turn.

2.1. MEASURES BASED ON SIMPLE TRANSFORMATIONS

Seven measures are expressed in terms of simple transformations of $V(N)$ and N. All these measures can be described as arising from attempts to fit simple mathematical functions to the curve of the vocabulary size $V(N)$ as a function of N.

Guiraud (1954) proposed the following text characteristic:

$$R = \frac{V(N)}{\sqrt{N}}. \tag{4}$$

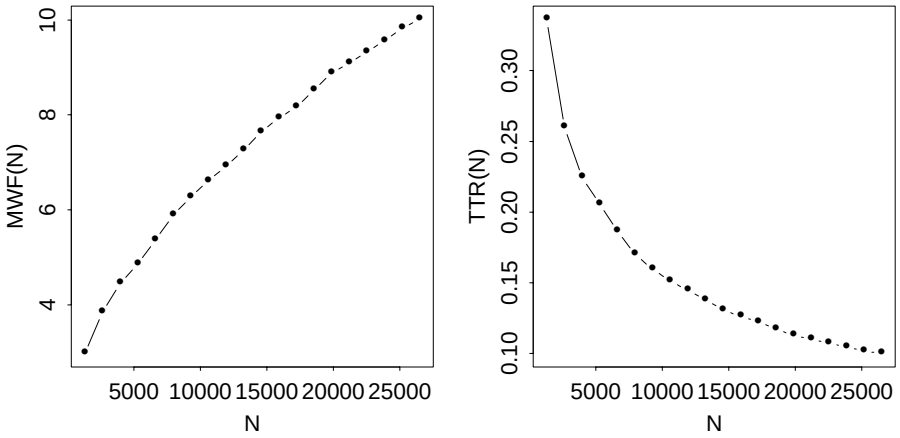

Figure 2. The mean word frequency $MWF(N)$, and its reciprocal type-token ratio $TTR(N)$ as functions of text length N for *Alice's Adventures in Wonderland* at twenty equally-spaced measurement points.

This constant implies that the vocabulary size is proportional to the square root of the text length:

$$V(N) = R\sqrt{N}.$$

A second measure was introduced by Herdan in 1960 and 1964 and is defined as:

$$C = \frac{\log V(N)}{\log N}. \tag{5}$$

Here, the vocabulary size is assumed to be a simple power function of N:

$$V(N) = N^C.$$

Dugast (1979, 23) cites Rubet's *A Dynamical Study of Word Distribution* as modifying equation (5) to produce:

$$k = \frac{\log V(N)}{\log(\log N)}, \tag{6}$$

where the vocabulary size is assumed to be a power function of $\log N$:

$$V(N) = \log^k N.$$

Maas (1972) proposed an associated relationship between V and N, where

$$a^2 = \frac{\log N - \log V(N)}{\log^2 N}. \tag{7}$$

This expression is a modification of Rubet's k. To see this, we rewrite (7) in the form

$$V(N) = N \log^{-a^2} N,$$

with $-a^2 = k$. A notational variant of Maas' constant was proposed by Dugast (1978, 1979):

$$U = \frac{\log^2 N}{\log N - \log V(N)}, \tag{8}$$

or, equivalently,

$$V(N) = N \log^{-1/U} N,$$

which implies that Maas' a^2 is the same as Dugast's $1/U$.

Tuldava (1977) cites work published by Luk"janenkov and Nesitoj in 1975 which proposes

$$LN = \frac{1 - V(N)^2}{V(N)^2 \log N} \tag{9}$$

where $V(N)$ is related to the square root of $\log N$:

$$V(N) = \frac{1}{\sqrt{1 + LN \log N}}$$

Finally, in 1978, Brunet introduced a parametric expression

$$W = N^{V(N)^{-a}}, \tag{10}$$

where a is usually set to -0.172, which amounts to the claim that a change in text length can be accounted for in terms of a change in the base of the logarithm:

$$V(N) = \left(\frac{\log W}{\log N}\right)^a = \log_N^a W.$$

2.2. MEASURES USING ELEMENTS OF THE FREQUENCY SPECTRUM

We now introduce measures that make use of elements of the frequency spectrum, $V(i, N)$. Honoré (1979) proposed a measure which assumes that the ratio of *hapax legomena*, $V(1, N)$, to the vocabulary size, i.e. the growth rate, is constant with respect to the logarithm of the text size:

$$\frac{V(1, N)}{V(N)} = a + b \log N.$$

For $a = 1$ and $b = 100/H$, we can reformulate this as

$$H = 100\frac{\log N}{1 - \frac{V(1,N)}{V(N)}}, \tag{11}$$

which is the form in which H was originally introduced. It follows, if H is truly constant, that

$$V(N) = \frac{V(1, N)}{1 - \log\left(N^{\frac{H}{100}}\right)}.$$

Sichel (1975) observed that the ratio of *dis legomena*, $V(2, N)$ to the vocabulary size is roughly constant across a wide range of sample sizes:

$$S = \frac{V(2, N)}{V(N)}, \tag{12}$$

or equivalently,

$$V(N) = \frac{V(2, N)}{S}.$$

He suggested that the constancy of this statistic at certain text sizes might be useful for comparing texts of different lengths. This observation had also been made by Michéa in 1969 and 1971, who proposed to use the reciprocal of S as a textual measure:

$$M = \frac{V(N)}{V(2, N)}. \tag{13}$$

In addition to measures that make use of specific spectrum elements in combination with N and $V(N)$, there is a family of measures that takes all spectrum elements into account. This family was introduced by Good (1953) and is defined as:

$$\begin{aligned} c_{s,t} &= \sum_{k=1}^{V(N)} (-\log p_k)^s p_k^t \\ &= \sum_{i=1}^{N} V(i, N)[-\log(i/N)]^s (i/N)^t. \end{aligned} \tag{14}$$

The second expression for $c_{s,t}$ is obtained by grouping the $V(i, N)$ types with frequency i and probability i/N. Perhaps the best known member of the $c_{s,t}$ family is the entropy, $c_{1,1}$,

$$\begin{aligned} E &= \sum_{k=1}^{V(N)} -\log(p_k) p_k \\ &= \sum_{i=1}^{N} V(i, N) \left(-\log \frac{i}{N}\right) \frac{i}{N}, \end{aligned} \tag{15}$$

a measure for the average amount of information, widely used in information theory. In lexical statistics, the first 'Characteristic Constant' proposed in the literature is a variant of $c_{0,2}$. In 1944, Yule argued that K,

$$\begin{aligned} K &= 10^4 \frac{[\sum_{i=1}^{N} V(i,N)(i/N)^2] - N}{N^2} \\ &= 10^4 \left[-\frac{1}{N} + \sum_i V(i,N) \left(\frac{i}{N}\right)^2 \right], \end{aligned} \tag{16}$$

is a text characteristic that is independent of text length, N. For $N \to \infty$, and disregarding the scaling factor 10^4, $K \to c_{0,2}$. A closely related measure is Simpson's D:

$$D = \sum_{i=1}^{V(N)} V(i,N) \frac{i}{N} \frac{i-1}{N-1}, \tag{17}$$

and, in an attempt to correct perceived flaws in the derivation of K, Herdan proposed the following modification of K in 1955:

$$V_m = \sqrt{\sum_{i=1}^{V(N)} V(i,N)(i/N)^2 - \frac{1}{V(N)}}. \tag{18}$$

Disregarding the 10^4 scaling factor, V_m is related to K as:

$$V_m^2 = K + \left(\frac{1}{N} - \frac{1}{V(N)} \right). \tag{19}$$

K, D and V_m are measures of the rate at which words are repeated, and can therefore be considered as inverse measures of lexical richness.

2.3. PARAMETERS OF PROBABILISTIC MODELS

All the measures considered thus far seek to characterise the properties of the frequency spectrum by means of simple summary statistics and by expressions ranging over all spectrum elements. Another line of approach is to make use of probabilistic models for word frequency distributions that provide explicit expressions for the vocabulary size and the spectrum elements by means of a limited number of formal parameters. For word frequency distributions, which fall into the class of Large Number of Rare Event (LNRE) distributions, three models are available (Baayen, 1993; Chitashvili and Baayen, 1993). In this paper we will consider two computationally tractable sub-models; Orlov's generalised Zipf model and Sichel's generalised inverse Gauss-Poisson model.

According to the generalised Zipf distribution (Orlov, 1983), V(N) is a function of one free parameter, Z:

$$V(N) = \frac{Z}{\log(p^*Z)} \frac{N}{N-Z} \log(N/Z). \tag{20}$$

This parameter specifies the text length at which Zipf's law in its simplest form,

$$V(i, N) \propto \frac{1}{i(i+1)},$$

holds. We can interpret Z as a measure of lexical richness: an increase in Z leads to an increase in $V(N)$. The second parameter in (20), p^*, is the maximum sample relative frequency – the frequency of the most common word divided by the text length. At least in theory, p^* is independent of the length of the text and can be regarded as a fixed parameter or text characteristic.

Turning to Sichel's generalised inverse Gauss-Poisson model, we can express the vocabulary size as a function of N with two free parameters, b and c (with the third parameter of the general model, γ, held at -0.5 for computational tractability):

$$V(N) = \frac{2}{bc}\left[1 - e^{b(1-\sqrt{1+Nc})}\right]. \tag{21}$$

The fraction $2/bc$ represents the number of different words in the population of the author's vocabulary, ν, from which the $V(N)$ words used in a given text are a sample. Clearly, the population vocabulary ν is itself a measure of lexical richness, as are the parameters b and c. As b and c become smaller, the population number of types increases, along with the number of words observed in the text, $V(N)$.

3. The Variability of Lexical Constants

Having completed our survey of proposed length-independent measures of lexical richness, we now consider to what extent these measures are truly independent of the sample size, N. We are not the first to cast doubt on the constancy of measures of lexical richness. Orlov (1983) shows that Guiraud's R is a convex function of N, and he points out that Herdan's C is likewise slightly convex. The constancy of C is also questioned by Weitzman (1971). Ménard (1983) also finds that R and C are variable, and he questions the whole rationale of Michéa's M. According to Thoiron (1986), the 'sensitivity' of D (and, by implication, that of K) to the text length 'cannot be totally disproved' (p. 198). Thoiron also points out that the entropy E changes along with N (see also Johnson, 1979). Brunet's W and Dugast's U come under the scrutiny of Cossette (1994), who finds them to vary with the text length. To our knowledge, the measures H, S, Z, and the parameters of the inverse Gauss-Poisson model have received general acceptance as length-invariant measures.

In what follows, we address the problem of length-invariance for all constants, using two complementary approaches. We will first study the behaviour of the constants from a mathematical point of view, using simple randomisation techniques. We will then proceed to show how the values of these constants are affected by the non-random way in which words are used in actual coherent prose.

3.1. THEORETICAL CONSTANCY

In order to evaluate the mathematical properties of the constants, we will follow Yule (1944) in making the simplifying assumptions that words are used randomly and independently in texts, assumptions which lead to the urn model (Johnson and Kotz, 1977). When we apply the urn model to lexical data, the use of a word can be modelled as the random selection of a marble from an urn. The urn typically contains a large number of marbles of various colours. Some colours appear on many marbles, others on just a few. The urn model lies at the basis of a great many analytical expressions for word frequency distributions (see, e.g., Good, 1953, Good and Toulmin, 1956, and Chitashvili and Baayen, 1993). In this study, we have opted to use randomisation techniques to investigate the behaviour of our constants across a wide range of text lengths.

The randomisation technique that we have used is a very simple one. The basic step is to randomly permute the order in which the words appear in a text. Following permutation, we calculate the values of a given constant for a pre-specified number (K_0) of text lengths, the points at which we measure the values of our textual statistics. This procedure is repeated many times, and leads to a distribution of that constant at each measurement point. From this distribution, we can obtain estimates of a constant's mean and a 95% confidence interval for each of these points. The confidence interval is constructed by ordering all the values obtained for the constant at that point, followed by the removal of the top and bottom 2.5%. Thus for 1000 permutations, the lower confidence limit is made up of the values of the 25th element at each measurement point, the upper confidence limit being the values of the 976th element. This is known as a Monte Carlo (MC) confidence interval.

Figure 3 illustrates the extent to which the constants vary with N for the text of *Alice's Adventures in Wonderland* (obtained from the Oxford Text Archive)[1] using twenty equally-spaced measurement points and 5000 randomisations. It is clear that W, E and LN are monotonically increasing with text length, N, while C, H and U are monotonically decreasing.[2] In addition, R and S rise to a maximum, then decrease with increasing N. Only K is constant across all text lengths, although Z appears constant from measurement point $k = 5$. The parameters of the inverse Gauss-Poisson model, rather than being constant, seem to vary considerably; b is a monotonically increasing function, while c decreases with N.[3] The source of this apparent theoretical non-constancy of LNRE parameters is considered in detail in Baayen and Tweedie (1998).

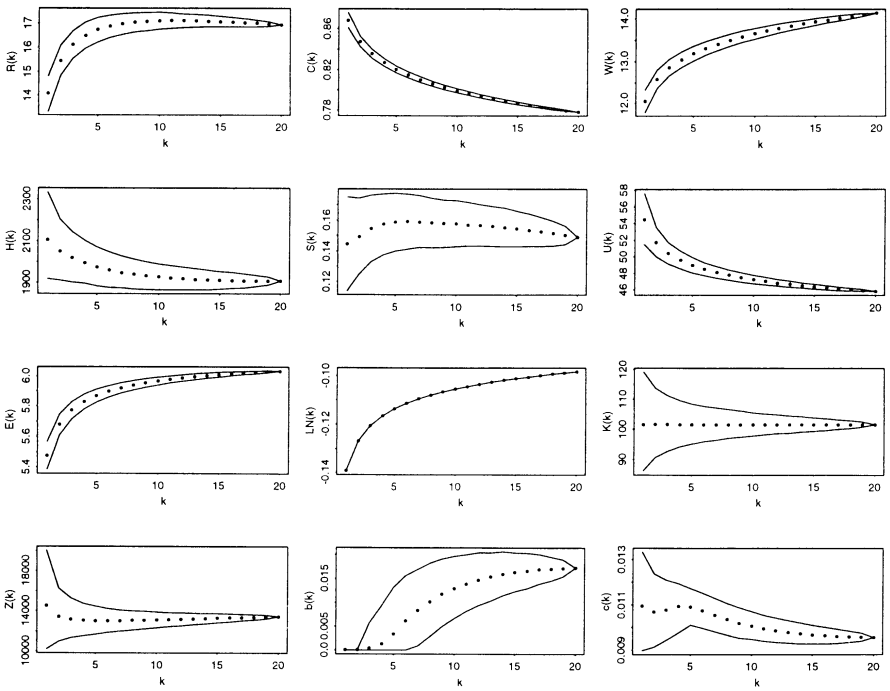

Figure 3. The dependence of selected constants on N in *Alice's Adventures in Wonderland* using Monte Carlo estimation. For $k = 1, 2, \ldots, 20$ measurement points, the dots represent the mean values for 5000 permutations, and the solid lines the upper and lower limits of the 95% Monte Carlo confidence interval. The x-axis is measured in 20 equally-sized chunks of text, thus k increases in steps of $N/20 = 1325$ words.

Summing up, in theory, with the exception of K and Z, all constants systematically change as a function of the text length N, as shown by our Monte Carlo means across twenty measurement points. In the remainder of this paper, we will make this dependence on text length explicit in our notation, writing $H(N)$ instead of H, and similarly for all other measures.

3.2. EMPIRICAL CONSTANCY

The next issue to be addressed is the potential effect on the constancy of our constants of non-random word use in coherent prose. Coherent prose does not consist of a string of randomly chosen words and sentences. There is structure to coherent prose, not only syntactic structure at the level of the sentence, but also structure at the level of textual organization. *Alice's Adventures in Wonderland*, for instance, is a narrative text organized in a series of chapters, with themes introduced in the opening chapter that are picked up again in the closing chapter. We will use the term 'discourse structure' to refer to this non-random textual organization of texts.

The discourse structure of texts is at odds with the randomness assumption that lies at the heart of the theoretical constancy of textual measures. We will examine the potential effect of violating the randomness assumption on the constancy of our measures by calculating the values of the constants for the actual text of *Alice's Adventures in Wonderland* and comparing them with the confidence intervals obtained above.

Figure 4 shows these empirical values along with the randomisation confidence intervals from Figure 3. The constants $R(N)$, $C(N)$, $W(N)$, $U(N)$, $K(N)$ and $Z(N)$ all exhibit significant divergence from their theoretical values. The values for the entropy, $E(N)$, track the lower confidence limit, while those for $c(N)$ are slightly higher than their upper confidence limit. Of all the constants examined here, only four; $H(N)$, $LN(N)$, $S(N)$ and $b(N)$, appear to behave in a similar way in running text as they do under the assumptions of the urn model for not too small N. These examples suggest that it should not be taken for granted that discourse structure leaves the constancy of lexical measures unaffected.

Summing up, what our data suggest is that some constants ($R(N)$, $C(N)$ and $E(N)$) vary with the text length N in theory and also depart from their expected values given the urn model in real text. The constant $LN(N)$ is very variable, yet so constrained in nature that the observed value must fall inside its confidence interval. Others ($K(N)$, $D(N)$ and $Z(N)$) are truly constant, or nearly constant in the case of $Z(N)$, in theory, but may reveal significant deviation from their expected values in actual text. Finally, the parameters of Sichel's model ($b(N)$ and $c(N)$), which in theory should be truly constant, also revealed systematic dependency on the text length for both the empirical data and the Monte Carlo simulations.

The main point of this section has been a negative one: almost all constants that have been proposed in the literature change systematically with the text length. The aim of the following sections is to ascertain to what extent constants can nevertheless be used in stylometric studies. Section 4 addresses the question of how the within-text variability of a given constant relates to its between-text variability. Section 5 introduces a method for testing whether texts differ significantly with respect to the empirical variability of a given constant. Section 6, finally, compares the efficacy of constants as a means for clustering texts by author with the efficacy of using the relative frequencies of the highest-frequency function words (Mosteller and Wallace, 1964; Burrows, 1989).

4. Developmental Profiles

Thus far we have considered the variability of constants in a single text, that of *Alice's Adventures in Wonderland*. We cannot know whether or not the variability demonstrated above, the within-text variability, severely affects the usefulness of these constants as text characteristics unless we compare the values obtained from this text with those from other texts, the between-text variability. If the within-text variability is small compared with the between-text variability, then the constant

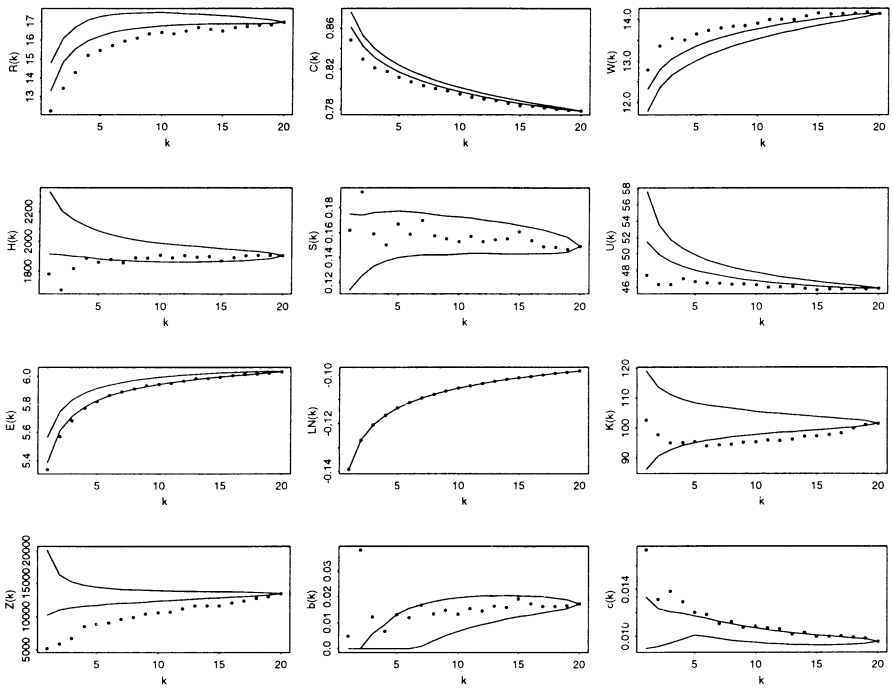

Figure 4. The dependence of lexical constants on N in *Alice's Adventures in Wonderland*. The points represent the values observed for the original text. The solid lines represent the upper and lower limits of the 95% Monte Carlo confidence interval previously shown in Figure 3.

may be of discriminatory use. On the other hand, if the within-text variability is large compared with the between-text variability, then, even when theoretically constant, the measure would be unsuitable for quantitative stylistic purposes.

Table I details the texts that we have used in this paper to investigate the issue of within and between-author variability. We have chosen eight authors and sixteen works, two works by each author except for a single work from Emily Brontë and three from Sir Arthur Conan Doyle. The texts were obtained from the Oxford Text Archive and vary in length from *The Acts of the Apostles* with 24246 words to the 116534 words that comprise *Wuthering Heights*. This data set allows us to examine the behaviour of the constants between and within a variety of texts and authors.

Figures 5 and 6 show the results of computing the values of selected constants for a text from each author. Only one text per author is plotted here for expositional clarity. It is clear that some measures assume consistently different values for different authors. In other cases the within-author variability may be large, but there is clear separation between developmental profiles from different authors. The exception to this is $S(N)$, where the within-author variability is as large as the between-author variability and no authorial structure can be seen in the graph. Hence we may conclude that $S(N)$ does not seem to be suitable for between-author discrimination.

Table I. The texts used in this study

Author	Title	*N*	Key
L. F. Baum	The Wonderful Wizard of Oz	39282	b1
	Tip Manufactures a Pumpkinhead	41571	b2
E. Brontë	Wuthering Heights	116534	B1
L. Carroll	Alice's Adventures in Wonderland	26505	a1
	Through the Looking-glass and What Alice found there	29053	a2
A. Conan Doyle	The Sign of Four	43125	c1
	The Hound of the Baskervilles	59233	c2
	The Valley of Fear	57746	c3
H. James	Confidence	76512	j1
	The Europeans	59800	j2
St Luke	Gospel according to St Luke (KJV)	25939	L1
	The Acts of the Apostles (KJV)	24246	L2
J. London	The Sea Wolf	105925	l1
	The Call of the Wild	31891	l2
H. G. Wells	The War of the Worlds	60187	w1
	The Invisible Man	48599	w2

In addition, examination of the developmental profiles shows that certain constants appear to measure the same facet of the vocabulary spectrum. For example, the orderings of the texts are very similar for constants $R(N)$, $C(N)$ and $W(N)$. $K(N)$ and $D(N)$ also have the same orderings, although they are different from that of the $R(N)$ group. In order to have a more objective classification of constants into groups, we carried out a cluster analysis.[4] The ordering of texts generated by each constant was examined and each text given a rank corresponding to its position. This produced seventeen ranks for each text which were then subjected to cluster analysis. The results are shown in Figure 7. It is clear that there are four main clusters; one containing $K(N)$, $D(N)$ and $V_m(N)$, the next with $c(N)$ and $LN(N)$ while the third cluster contains the remaining constants, with the exception of $b(N)$, $S(N)$ and $M(N)$ which fall into the final cluster. We noted above that $S(N)$ was not able to distinguish between authors, and inspection of the orderings for the other constants in the second and fourth clusters confirm that these measures are less good at separating authors. The rest of the constants fall into two groups representing the first and third clusters in Figure 7.

It thus appears that the constants are able to separate to some extent texts by different authors. However, it may be possible that texts by the same author are also teased apart by this method. To examine this, we shall introduce the rest of our sample texts. Due to the size of the graphs, we shall only plot the values for $V(N)$, $W(N)$, $K(N)$ and $Z(N)$. These are shown in Figures 8 and 9. Texts belonging to

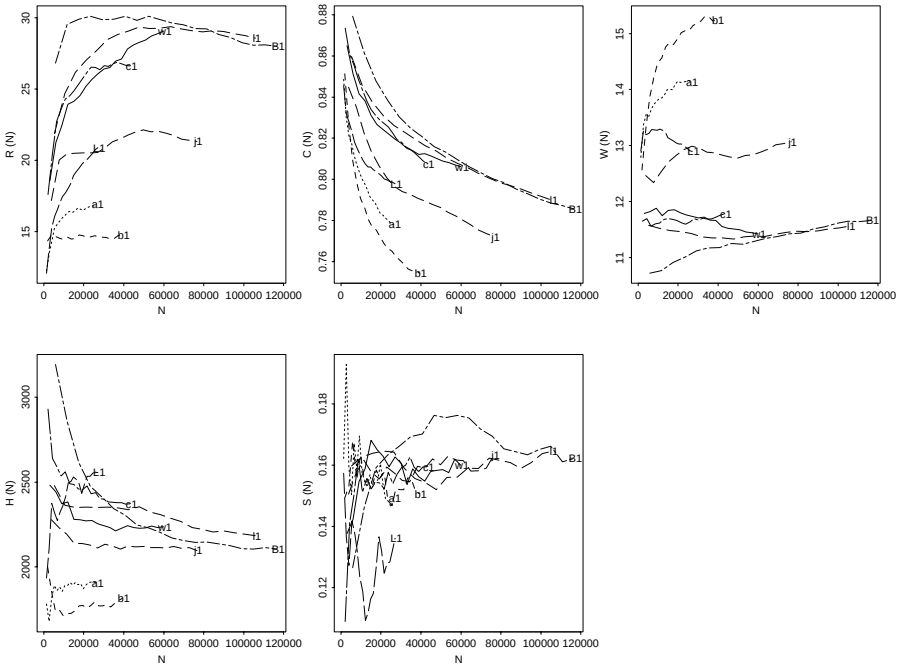

Figure 5. The behaviour of constants in works by different authors.

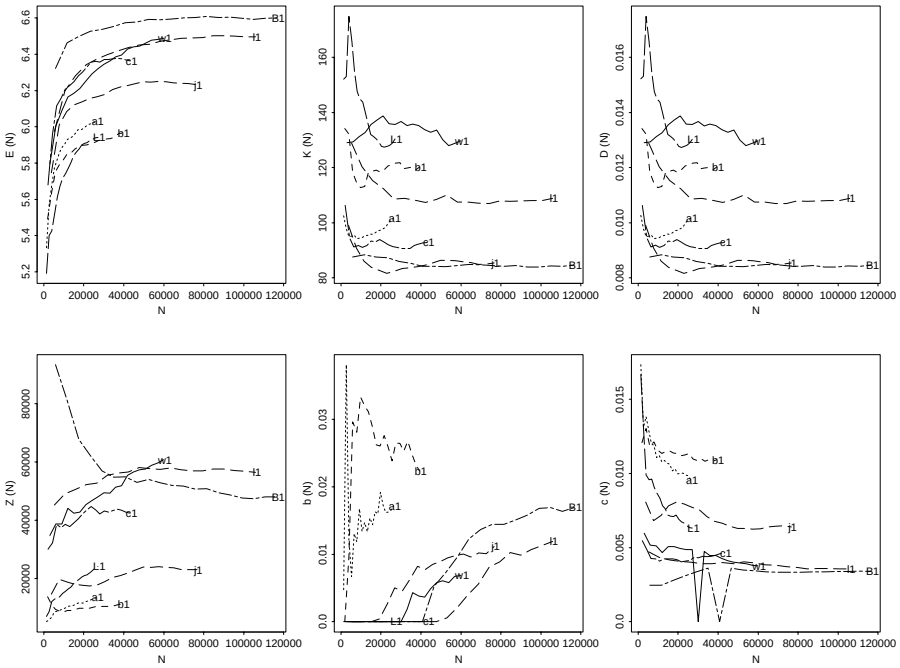

Figure 6. The behaviour of constants in works by different authors.

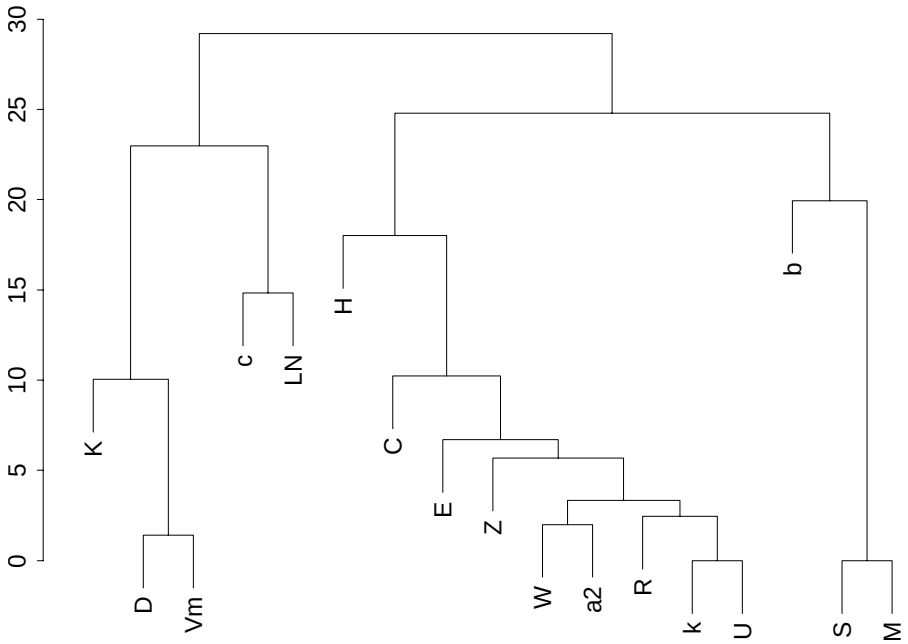

Figure 7. The classification of constants into families by their ordering of the texts listed in Table I. `Vm` represents Herdan's V_m and `a2` represents Maas' a^2.

the same author have the same type of line. It can be seen from the $V(N)$ graph that texts by the same author have similar vocabulary structures. The three constants plotted tease apart the vocabulary structure more clearly for visual inspection and allow us to examine it in detail. The plot of $W(N)$ shows that the texts by St Luke (`L1` and `L2`), Carroll (`a1` and `a2`), James (`j1` and `j2`), Wells (`w1` and `w2`) and London (`l1` and `l2`) have very similar developmental profiles. Two of the Conan Doyle texts (`c2` and `c3`) appear to be similar, yet a third appears much lower on the graph while the Baum texts (`b1` and `b2`) are quite disparate.

Turning to the graph of $K(N)$ shown in the first panel of Figure 9, the Baum texts have been placed close together and all three Conan Doyle texts have been united, while the London and especially the Wells texts have been pulled apart. We noted above that $K(N)$ belonged to a different family of constants and that it concentrates on the structure of the high-frequency words. It appears that London and Wells use high-frequency words in different ways in their two books under consideration here. The graph of $Z(N)$ has a similar ordering to that of $W(N)$, with some minor changes.

The information provided by the plots of the different measures can be summarised by considering a single representative function from both the major groups of measures. We chose $K(N)$ to represent its group; while $K(N)$, $V_m(N)$ and $D(N)$ give very similar orderings and are theoretically constant, $K(N)$ antedates $D(N)$ by five years and Herdan's $V_m(N)$ by eleven years. For the other group,

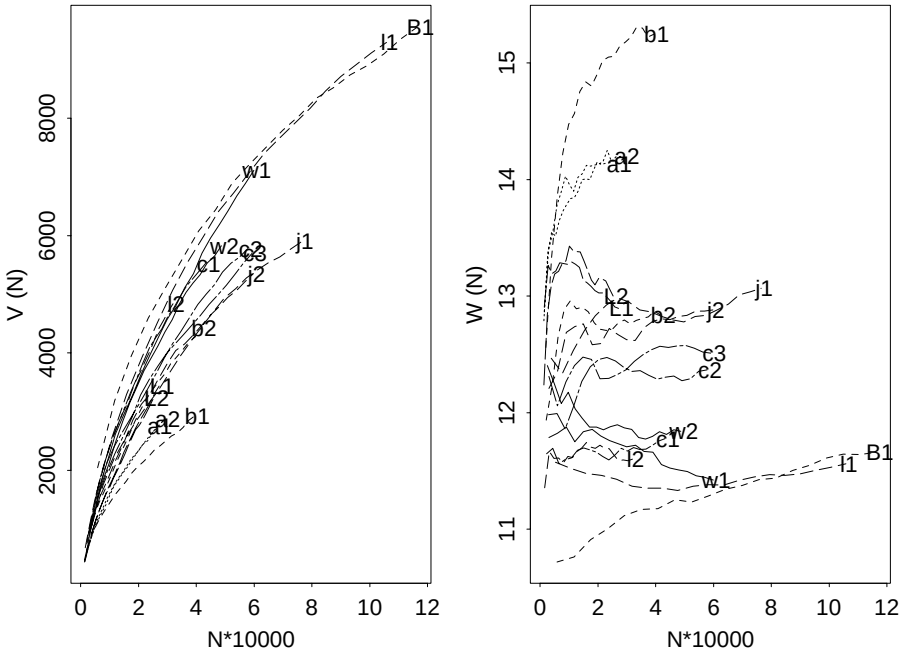

Figure 8. The behaviour of $V(N)$ and $W(N)$ in several works by different authors. Table I details the codes used for the texts.

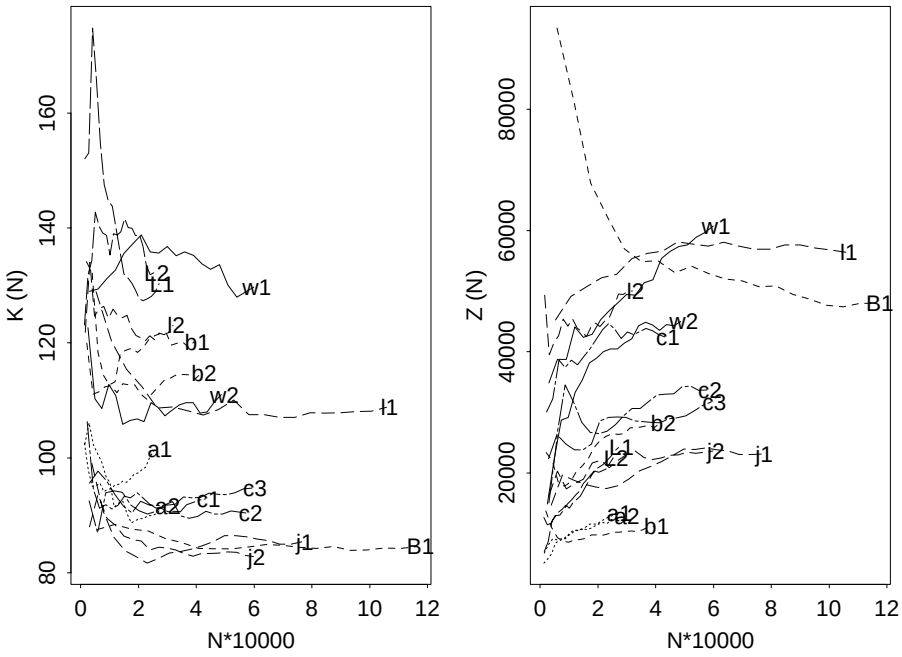

Figure 9. The behaviour of $K(N)$ and $Z(N)$ in several works by different authors. Table I details the codes used for the texts.

which contains most of the other functions, we chose $Z(N)$ due to its mathematical derivation and its theoretical constancy.

4.1. TRAJECTORIES

The figures exhibited in the previous section show, for a single constant, the variation found as one scans through a selection of texts. We found that there were two main families of constants which can be represented by the values of $Z(N)$ and $K(N)$. Rather than plotting these values in different graphs, we can combine them into a scatter-plot as shown in Figure 10. In this two-dimensional plane, texts are more clearly separated.

Here the whole trajectory of the text can be taken into account. The endpoint of the trajectory is marked by the text code, so that the direction of the development through the text can be traced. We can identify areas of the $Z - K$ space occupied by various authors.

It can be seen, for example, that the text of Brontë's *Wuthering Heights* (`B1`) has a stable value of K throughout the text, while the value for Z decreases through the text, as indicated by the movement from right to left in the lower right corner of Figure 10. Almost all the other texts move in the opposite direction on the Z axis, reflecting that their values of Z increase as the text length increases.

In addition, convex hulls drawn around each trajectory show that each author tends to occupy a unique space in the $Z - K$ plane, with the exception of some overlap between the initial values of *The Wizard of Oz* (`b1`) and the first of the Luke texts (`L1`). The convex hull of the second Wells text (`w2`) also almost completely encloses the second Baum text (`b2`), while *The Call of the Wild* (`l2`) is situated almost exactly between the Wells texts. The Wells pair are the most disparate of texts by the same author. Text `w1`, *The War of the Worlds*, has much higher values of $K(N)$ than text `w2`, *The Invisible Man*, throughout the text. While values of $Z(N)$ for later text in `w2` overlap early values in `w1`, for the most part `w1` has higher values of $Z(N)$. Thus, although written by the same author, *The War of the Worlds* has a much higher repeat rate (reflected in the higher values of $K(N)$) and a greater lexical richness as measured by $Z(N)$ than *The Invisible Man*.

5. The Comparison of Developmental Profiles

In the previous section we examined the behaviour of lexical constants in a selection of texts. We found that, in general, the developmental profiles of texts by different authors could be distinguished. Texts written by the same author were, for the most part, coincident, with some exceptions. Thus far, our impressions have been subjective ones; if we wish to compare texts objectively we must find a statistical method for their comparison. In this section we will present two such comparison techniques using text randomisations. The first considers a randomisation of the whole text, as was carried out for the text of *Alice's Adventures*

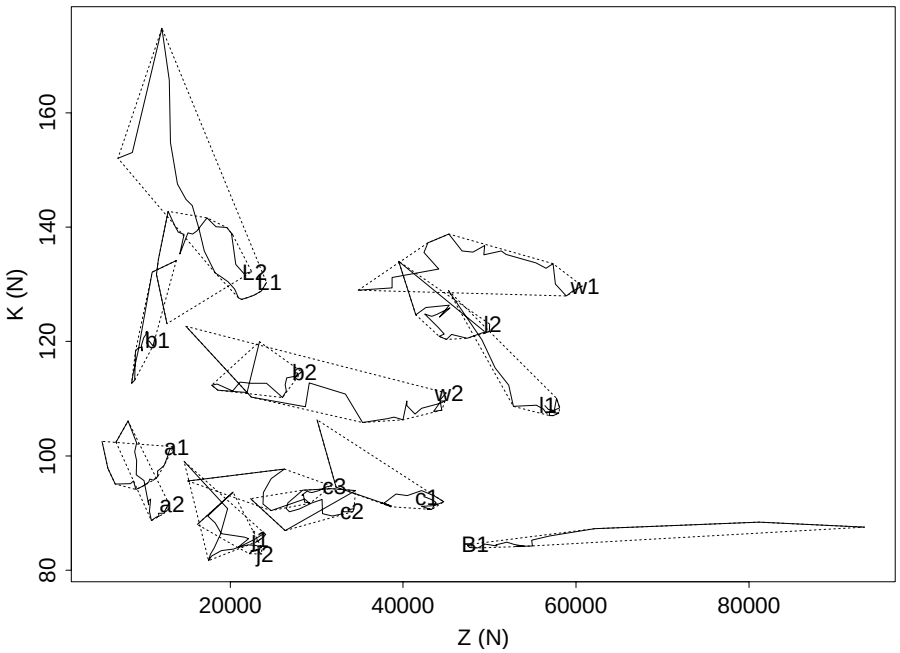

Figure 10. The behaviour of $Z(N)$ and $K(N)$ (solid lines) and their convex hulls (dotted lines) in texts by different authors.

in Wonderland in section 3.1. We will thus compare the expected values of the constants under the urn model. The second technique will make much more use of the empirical data in the construction of the text randomisations. We will consider comparisons between texts in our data set, for both between- and within-author cases.

5.1. FULL RANDOMISATION

Figure 3 showed the theoretical values of the constants for the text of *Alice's Adventures in Wonderland*, along with 95% MC confidence intervals. In order to compare two texts, we plot the results from both on the same graph, as shown in Figure 11.

It can be seen that, in many cases, the confidence intervals for the two texts do not separate at any point during the text. For others, the confidence intervals separate, for example, $K(N)$ splits at the seventh measurement point, at around 10,000 words of text. It is interesting that, although the values of the constants observed in the text are often outside their confidence intervals, their developmental profiles are similar.

In the above we have considered two texts by Lewis Carroll. Figure 12 shows the $Z - K$ trajectories of theoretical values and MC confidence intervals for all the texts in our data set. The theoretical invariability of $K(N)$ can be clearly seen, as

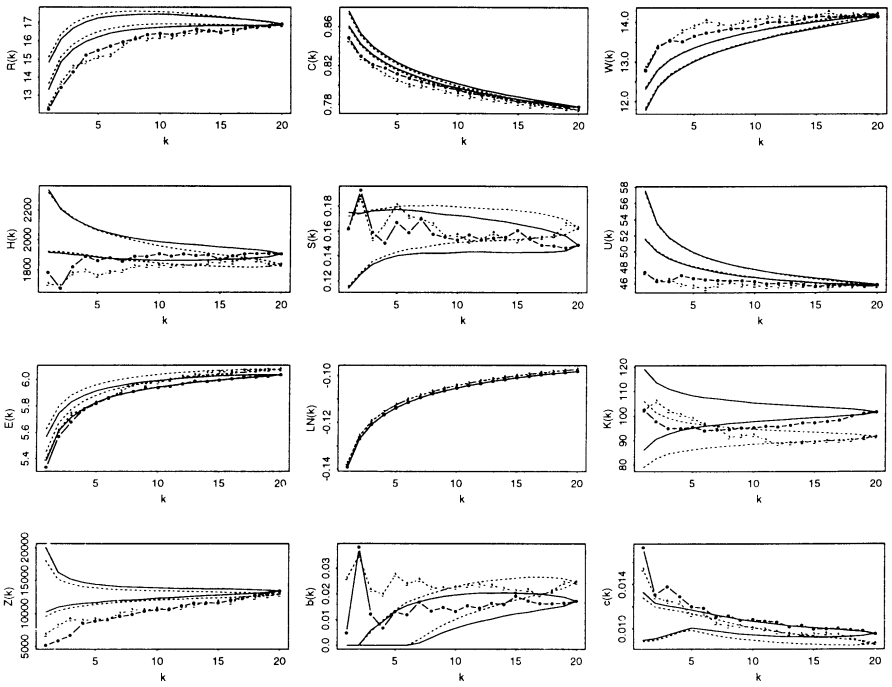

Figure 11. The dependence of selected constants on N in *Alice's Adventures in Wonderland* and *Through the Looking-glass.* The large and small points represent the values observed for the texts of *Alice's Adventures in Wonderland* and *Through the Looking-glass* respectively. The solid lines represent the upper and lower confidence limits for *Alice's Adventures in Wonderland* and the dotted lines the limits for *Through the Looking-glass*.

the mean values of $K(N)$ in the simulations are horizontal lines, representing no change in the $K(N)$ axis. It can be seen however, that $Z(N)$ does change through the text, increasing in most cases, with the exceptions of the texts by Baum (`b1` and `b2`), Carroll (`a1` and `a2`), and the first of the Wells texts (`w1`). The confidence intervals are often rather wide, and overlap in a fair number of cases, suggesting that the developmental profiles are much more similar to each other than suggested by the trajectories and their convex hulls themselves.

However, the plotted confidence intervals are generated from randomised texts; we saw in Figure 3 that the empirical values of the constants are often very different from their randomised values. The empirical profiles are determined by two factors: the vocabulary structure, in particular its richness and repetitiveness; and the discourse structure employed by the author. It is possible that the simplifying assumptions of the urn model which destroy the discourse structure are hampering our ability to distinguish between authors. We will therefore consider another technique which remains faithful to the empirical values while generating confidence intervals.

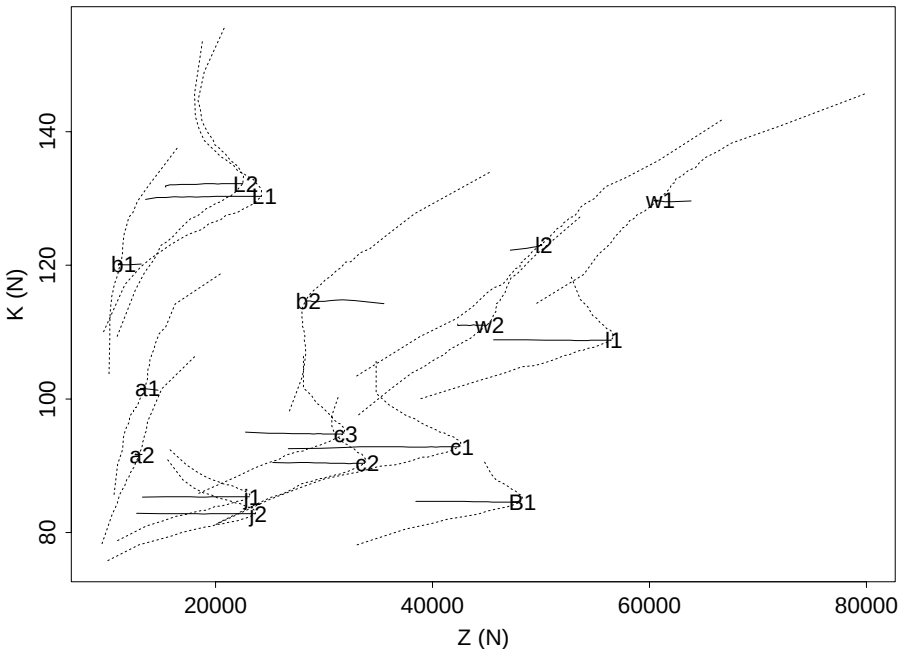

Figure 12. The mean behaviour of $Z(N)$ and $K(N)$ in various texts (solid lines) with MC confidence intervals (dotted lines).

5.2. PARTIAL RANDOMISATION

We saw in sections 3.1 and 3.2 above that the empirical values of the constants found in coherent text are often very different from the values found when the assumptions associated with the urn model are made. We would like to be able to obtain confidence intervals for the empirical values, thus employing the inherent discourse structure found in the text, a structure which is partialled out by full-text randomisation (Baayen, 1996). But, in order to construct empirical confidence intervals it is nonetheless necessary to perform some kind of randomisation and re-sampling of the text. We propose a method that lies between the empirical values and the full text randomisation.

The main idea behind our proposal is that of the *randomisation window*. Rather than randomising the full text, we will only permute sections of the text surrounding measurement points. The rest of the text remains unaltered, allowing discourse structure to be maintained. We will define the width of this window in terms of the measurement points in the texts, but this is not strictly necessary.

The general formula for the permuted region is

$$N_k \pm \frac{TN}{2K_0}, \tag{22}$$

where N_k is the kth measurement point out of K_0, in this paper $K_0 = 20$, and T is the size of the randomisation window.

For example, with a randomisation window of size 1, and the first measurement point at word 1325, as found in *Alice's Adventure's in Wonderland*, the text from word 662 (1325 – 1*26505/2*20) to word 1987 (1325 + 1*26505/2*20) would be permuted. Thus a word present in the first 1987 words has a chance of being counted at the first measurement point at word 1325. For the second measurement point, at word 2650, the text between words 1988 and 3312 would be permuted, and so on. Care must be taken when T is greater than 1 to ensure that the values calculated at subsequent measurement points are not compromised by permutations around the point of immediate interest.

A randomisation window of size 1 will allow a minimum level of randomisation to take place; words in the randomised text are constrained to remain very close to their textual positions. Increases in T will gradually release the constraints of discourse structure, allowing words to move more and more freely throughout the text. A randomisation window where $T = K_0$, the number of measurement points in the text, is equivalent to the full text randomisation described above.

It is not possible to take measurements at all K_0 points in the text. The final measurement point is at the end of the text and as there is no text after this point, no randomisations can be made and thus no confidence interval can be generated. As T increases, so the increasing width of the randomisation window invalidates measurements at the edges of the text where there is insufficient text before or after the measurement points for the randomisation to occur. We will, however, be able to construct MC confidence intervals around the central part of the text.

Figure 13 shows how changes in the window size, T, affect the MC confidence intervals for $K(N)$ in *Alice's Adventures in Wonderland* and *Through the Looking-glass and what Alice found there*. It can be seen that as T increases, the means of the randomised values become closer to the final (theoretically-constant) value of $K(N)$.

We can also plot the confidence intervals derived from the partial randomisation in the $Z-K$ plane. Figure 14 shows the texts plotted in this way for a randomisation window of size 5. The letters indicating the texts are plotted at the end point of the actual text, hence text `w1` can be seen to have a lot of movement still at the end of the text, while others, such as `c1` have fairly stable values of $Z(N)$ and $K(N)$ in the latter part of the text. The Wells texts remain split by the London texts, otherwise the confidence regions for each text are completely separate from those of texts by other authors.

The comparison of Figures 12 and 14 makes it clear that taking the discourse structure of the texts into account leads to improved confidence intervals. Figure 14 shows clearer differences in the vocabulary structure of the texts. While randomising the full text allows us to examine gross differences between texts and authors, for finer comparisons it is necessary to allow for the discourse structure that the author has imposed.

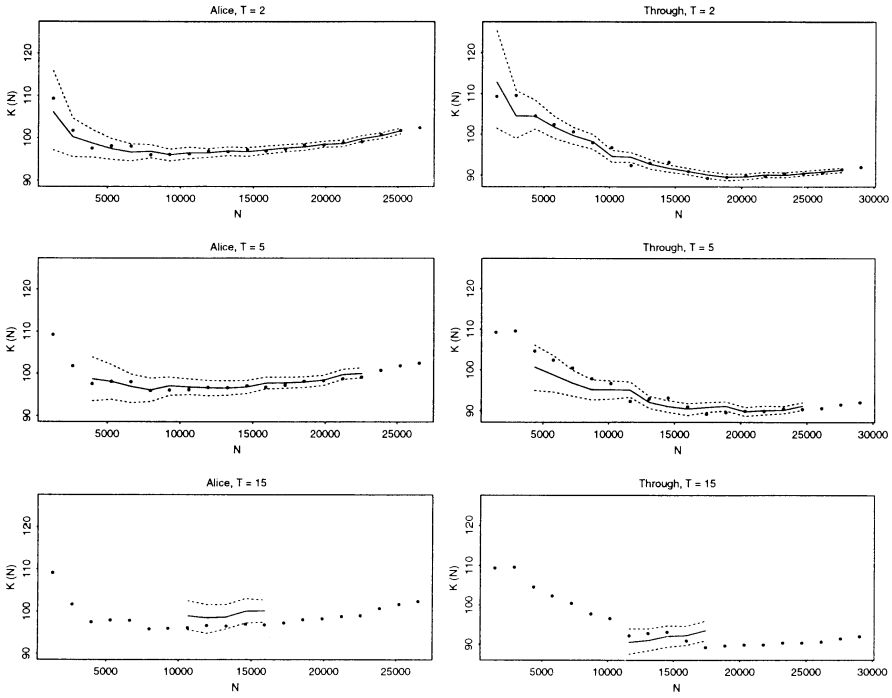

Figure 13. The behaviour of $K(N)$ in *Alice's Adventures in Wonderland* and *Through the Looking-glass*(dots) for $T = 2$, 5 and 15 with MC means and confidence intervals (solid and dotted lines).

6. Constants and Function Words

The preceding sections have illustrated that lexical constants, notably $Z(N)$ and $K(N)$, capture aspects of authorial structure. Complete authorial separation was not obtained, however: the texts by Wells and London, for instance, do not separate well in the plane spanned by $Z(N)$ and $K(N)$. The question that remains to be answered is whether this failure is due to a lack of discriminatory power on the part of the constants, or whether this lack of separation is in fact due to the actual stylometric similarity of the texts by Wells and London.

In order to answer this question, we compare the discriminatory power of lexical constants with the discriminatory power of the highest-frequency function words. Mosteller and Wallace (1964) were among the first to call attention to the discriminatory potential of function words, which tap into the (more or less) unconscious syntactic and stylistic habits of authors. Burrows (1989), Holmes and Forsyth (1995), Baayen et al. (1996), and Tweedie et al. (1998) use the relative frequencies of the 50 or 100 most-frequent function words in principal components analysis as authorial fingerprints. There is a growing consensus that this is a powerful methodology that captures many details of authorial structure. We have therefore selected

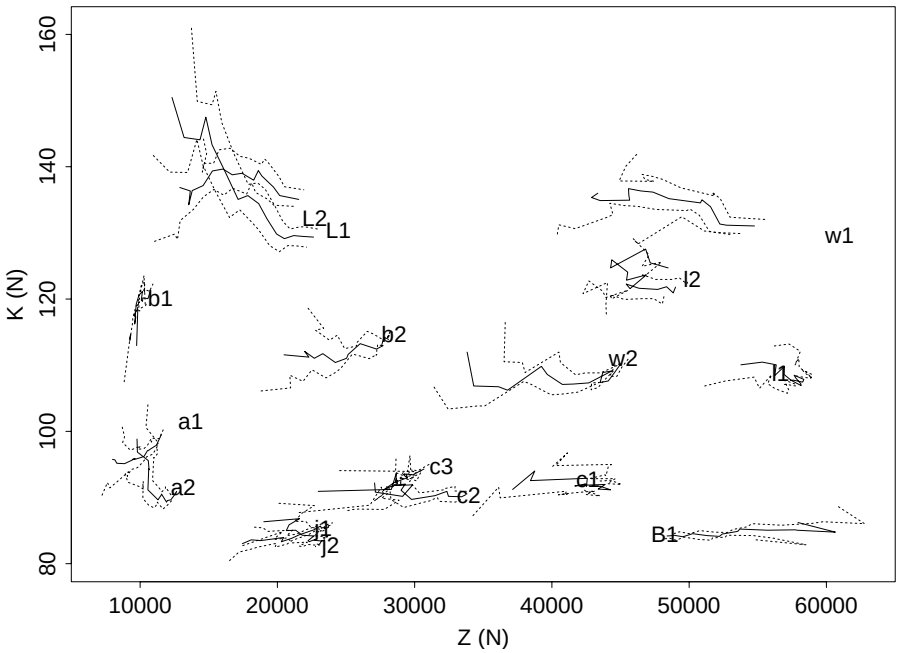

Figure 14. Mean values (solid lines), and upper and lower 95% confidence limits (dotted lines) of $Z(N)$ and $K(N)$ in texts by various authors with randomisation window size $T = 5$.

this methodology as a baseline for studying the usefulness of lexical constants as stylometric measures.

From the texts listed in Table I, we selected the 100 most-frequent function words common to all texts. For each function word and each text, we calculated the relative frequency of that function word in that text. In this way we obtained a matrix of 16 texts by 100 function words. This matrix was subjected to a principal components analysis, which resulted in eight significant principal components that described 74.07% of the original variation.[5] Principal components analysis is a dimension-reducing technique, thus instead of each text representing a point in a 100-dimensional space, each text is now a point in an 8-dimensional space. The coordinates of the texts in this 8-dimensional space were subjected to a cluster analysis, the results of which are shown in Figure 15.

The clustering obtained reflects the authorial provenance of our texts, with the exception of the texts by London (`l1` and `l2`) one of which clusters with our text by Brontë (`B1`), and one of which adjoins the cluster of texts by Wells (`w1` and `w2`). In order to compare this analysis with the previous analyses based on lexical constants, we carried out the cluster analyses summarized in Figure 16.

The top left panel in Figure 16 shows a cluster analysis of the five significant principal components of the final values of all of the constants detailed in this paper. The principal components described 91.48% of the variation within this data. It can be seen that the texts by Carroll (`a1` and `a2`) cluster together, as do the texts

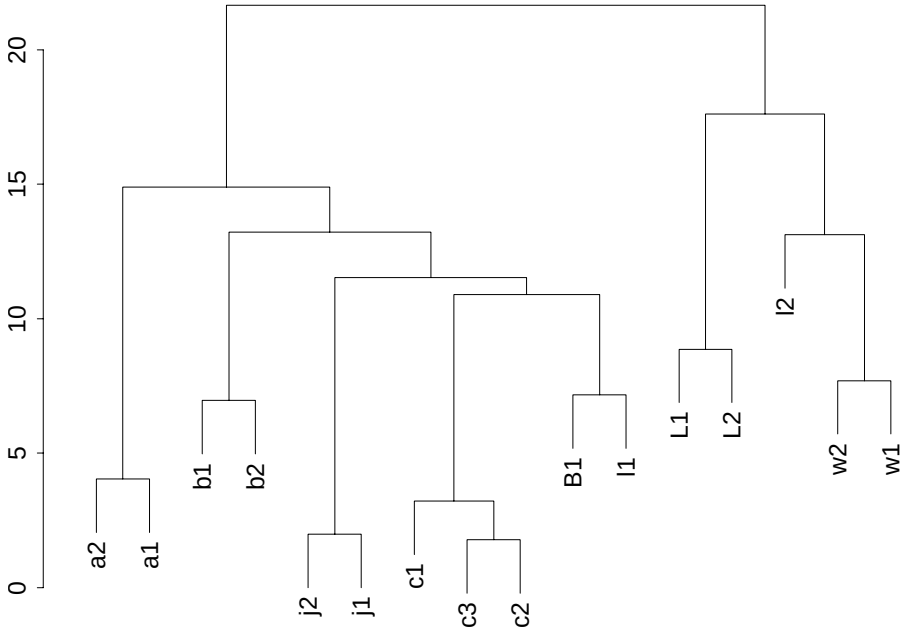

Figure 15. Authorial structure as revealed by a cluster analysis of the coordinates of texts in the space spanned by the 100 most-frequent function words after dimension reduction by means of principal components analysis.

by James (`j1` and `j2`), St Luke (`L1` and `L2`) and two of the Conan Doyle texts (`c2` and `c3`). However, the other texts cluster more closely with texts by other authors. A similar pattern is found in the top-right dendrogram where the five significant principal components of the full trajectories of all the constants are examined. These principal components describe 72.62% of the variation in the trajectories. This dendrogram may offer a slightly better categorisation as the texts by London (`l1` and `l2`) and those by Wells (`w1` and `w2`) are within the same larger cluster.

The lower panels of Figure 16 illustrate the clustering when only $Z(N)$ and $K(N)$ are used, rather than all of the seventeen constants that we have examined. The left panel shows the dendrogram resulting from cluster analysis of the final values of $Z(N)$ and $K(N)$. No principal components analysis is required here, as we have two values only from each text. It can be seen again that texts by Carroll, St Luke, James and two of the Conan Doyle texts are nearest-neighbours in this analysis. As in the panel above, other texts cluster with texts by other authors. The final panel in the lower-right is the result of a cluster analysis performed on the four significant principal components of the trajectories of $Z(N)$ and $K(N)$, describing 86.04% of the variation. This dendrogram gives us the best results of the four; the texts that have clustered in the previous panels do so, as do the texts by Baum (`b1` and `b2`).

Baayen and Tweedie (1998) use Linear Models and Repeated Measures techniques to analyse the $Z(N)$ values of a similar group of texts.[6] They find significant

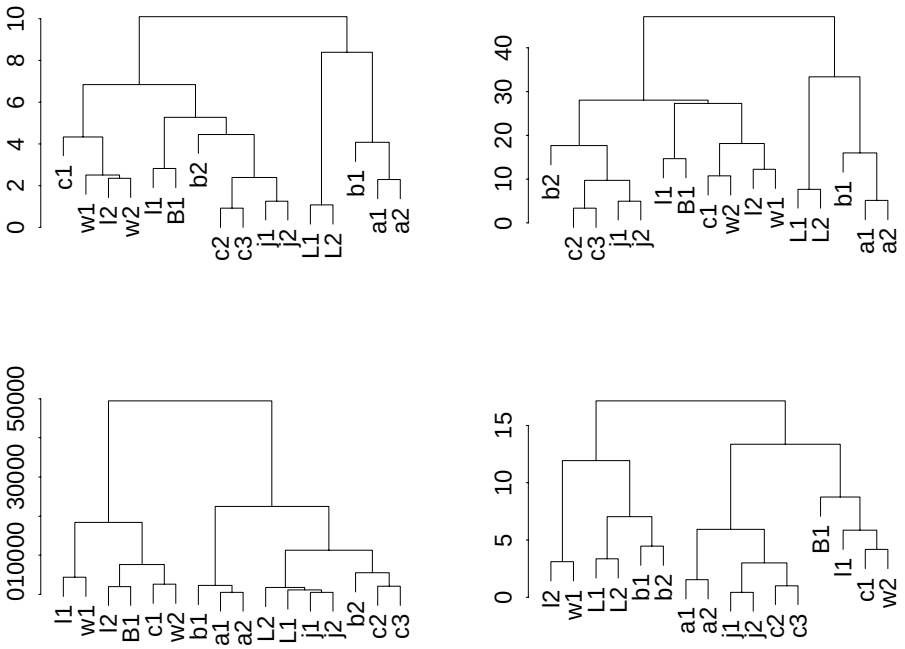

Figure 16. Authorial structure as revealed by a cluster analysis of the coordinates of texts in the space spanned by lexical constants. The top left panel shows the results obtained from the final values of all the constants examined, the top right panel has the results from using the full trajectories of all of the constants. The lower panels show the results obtained when the values of just $Z(N)$ and $K(N)$ are considered; the left panel uses only the final values, the right panel the full trajectories.

differences between the trajectories of the two texts by Wells, the texts by Baum, the texts by Conan Doyle (`c1` and `c3`), and the texts by London. This is reflected in the fact that texts by these authors are generally found in different clusters in the dendrograms in Figure 16. In addition, no significant differences were found between the texts by Carroll, the texts by St. Luke and the texts by James. Again, this is reflected in our dendrograms by the fact that these texts appear as nearest neighbours in each case.

The four dendrograms which make up Figure 16 allow us to make the following observations. Analysis of the trajectories produces results that are at least as good as analysis of the final values. Analysis of the constants $Z(N)$ and $K(N)$ is as effective as the analysis of all of the constants that we have considered. The best classification is obtained using the trajectories of $K(N)$ and $Z(N)$. However, none of these analyses produces results as good as those based on the function words as shown in Figure 15.

These observations draw us to the following conclusions. First, the trajectories of $Z(N)$ and $K(N)$ provide as good an indication of vocabulary richness as can be obtained from a larger number of constants. Second, analyses using function words can produce better results than analyses using measures of vocabulary

richess. Third, some authors, for example London and Wells, are stylometrically very similar, illustrating that authorial differences can be and often are visible quantitatively in word use, but that this is not always the case.

7. Discussion

We started this paper by describing a number of measures of lexical richness that have been proposed in the literature. Some of these were based on simple curve fitting of the number of types $V(N)$ as a function of the number of tokens N, others made use of elements of the frequency spectrum, while the final set were parameters of Large Number of Rare Event distributions. In general, these measures have been assumed to be constant with respect the text length, with only a little doubt being cast upon them.

Many of these measures are based on the urn model assumption, that is that words occur randomly in text. In order to examine the *theoretical* constancy of the measures, we used randomisation techniques to simulate the urn model. Almost all of the so-called constants varied as the text length increased. Turning to measures which are theoretically constant, $K(N)$, $D(N)$, $Z(N)$, $b(N)$ and $c(N)$, the first three are indeed constant in theory, while the parameters of Sichel's model were found to be heavily dependent on the text length.

While the urn model allows for simplicity in modelling, we have not taken into account the non-randomness of words in coherent prose. When the empirical values of the text constants are compared with the theoretical values, they frequently fall outside the 95% MC confidence limits established. Even measures which appeared to be theoretically constant exhibit dependency on the text length when empirical values are calculated. It is clear that discourse structure has a large effect on these measures. This aspect is discussed further in Baayen and Tweedie (1998).

We then considered the between- and within-author variation exhibited by the measures of lexical richness in texts. It became clear that various measures give rise to the same ordering of texts. The measures can be divided into two major groups; the first containing $K(N)$, $D(N)$ and $V_m(N)$; the second being made up of the other measures with the exception of $c(N)$, $LN(N)$, $b(N)$, $S(N)$ and $M(N)$. $LN(N)$, $S(N)$ and $M(N)$ turned out to be ineffective at discriminating between authors, while the orderings expressed by $b(N)$ and $c(N)$ are suspect due to the absence of fits for some of the texts, and the unclear interpretation of $b(N)$ and $c(N)$ themselves. The measures $K(N)$ and $Z(N)$ were chosen to represent the two main groups as both are theoretically constant, while the other members of the second group all displayed a systematic theoretical dependency on the text length. The groupings can be exploited by plotting the values of $Z(N)$ against those for $K(N)$. This leads to a plot where, with some exceptions, each authorial group occupies a separate space in the $Z - K$ plane.

In order to compare the developmental profiles we again used the Monte Carlo technique to produce confidence intervals around the theoretical values of con-

stants from texts by different authors. However, we had already established that the empirical values of the constants could diverge from their theoretical ones. To allow for confidence intervals around the empirical values of the constants we therefore introduced the idea of partial randomisations, where only a small section of the text is permuted. The influence of discourse structure can be changed by changing the size of the permuted region. These confidence intervals can also be plotted in the $Z - K$ plane, resulting in groups of texts by the same author occupying the same space. However, as the texts by Wells and London show, authors can still significantly change their style across works in their canon.

In order to gauge the discriminatory potential of lexical constants vis-á-vis other methods, we compared these results to those obtained by means of a principal components analysis of the relative frequencies of the 100 highest-frequency function words. For our data set, the function words provide a more precise authorial classification. At the same time, just the two measures $Z(N)$ and $K(N)$ already reveal some major patterns of authorial structure.

To conclude, our results question two aspects of the use of the so-called constants. Firstly, we have shown that the assumption that measures of lexical richness are independent, or roughly independent of text length is invalid. The values of almost all the proposed measures change substantially in systematic ways with text length. It is thus necessary to correct for text length, or to consider the developmental profiles or trajectories of the full text.

Secondly, our results question the usefulness of including many different 'constants' in authorship attribution studies (e.g. Holmes, 1992; Holmes and Forsyth, 1995) as we have shown that there are two useful families which measure the two facets of vocabulary structure: richness and repeat rate. With only two independent constants, the use of a great many different lexical constants in authorship attribution studies is unnecessary.

Finally, compared to an analysis of 100 function words, it is surprising how much authorial structure is already captured by just two measures, $Z(N)$ and $K(N)$. We conclude that $Z(N)$ and $K(N)$ are two useful indicators of style that should be used with care (given their within-text variability) and in conjunction with the many other indicators of style (such as the relative frequencies of function words) that are currently available for stylometric analyses.

Notes

[1] http://ota.ahds.ac.uk.

[2] The confidence interval surrounding $LN(k)$ is so narrow in relation to the variability found in values of the constant that in the figure, the interval appears to have no width. For example, the biggest difference between the upper and lower confidence intervals for $V(k)$ results in a change of $7 * 10^{-9}$ in the value of $LN(k)$. Changes in the vocabulary size between texts, and even between authors, will produce changes of this order of magnitude, which are close to being undetectable when the value of $LN(k)$ varies between -0.14 and -0.10 in the whole text of, in this case, *Alice's Adventures in Wonderland.*

[3] The parameters b and c of Sichel's model are estimated by iteration such that $E[V(N)] = V(N)$ and $E[V(1, N)] = V(1, N)$ (see Sichel, 1986, for details). For small values of k, and thus N, no solution is available for b and c that meets these requirements. The means and confidence intervals that we present in this study are conditional on the availability of a fit.

[4] The cluster analyses in this paper use complete linkage and the Euclidean distance metric.

[5] The principal components analyses in this paper are carried out on the correlation matrix rather than the covariance matrix of the variables, thus allowing for the different size of the variables.

[6] In order to balance their experimental design, Baayen and Tweedie (1998) did not analyse the text by Brontë (`B1`) nor the second of the Conan Doyle texts (`c2`).

References

Baayen, R. H. *A Corpus-based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. PhD thesis, Amsterdam: Free University, 1989.

Baayen, R. H. "Statistical Models for Word Frequency Distributions: A Linguistic Evaluation". *Computers and the Humanities* 26 (1993), 347–363.

Baayen, R. H. "The Effect of Lexical Specialisation on the Growth Curve of the Vocabulary". *Computational Linguistics* 22 (1996), 455–480.

Baayen, R. H. and F. J. Tweedie. "The Sample-size Invariance of LNRE Model Parameters: Problems and Opportunities". *Journal of Quantitative Linguistics* 5 (1998).

Baayen, R. H., H. van Halteren and F. J. Tweedie. "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution". *Literary and Linguistic Computing* 11(3) (1996), 121–131.

Baker, J. C. "Pace: A Test of Authorship Based on the Rate at Which New Words Enter the Author's Text". *Literary and Linguistic Computing* 3(1) (1988), 136–139.

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

Burrows, J. F. " 'An Ocean Where Each Kind . . . ': Statistical Analysis and Some Major Determinants of Literary Style". *Computers and the Humanities* 23(4–5) (1989), 309–321.

Chitashvili, R. J. and R. H. Baayen. "Word Frequency Distributions". In *Quantitative Text Analysis*. Eds. G. Altmann and L. Hrebícek, Trier: Wissenschaftlicher Verlag Trier, 1993.

Cossette, A. *La Richesse Lexicale et sa Mesure*. Number 53 in Travaux de Linguistique Quantitative. Paris: Slatkine-Champion, Geneva, 1994.

Dugast, D. "Sur quoi se fonde la notion d'étendue théoretique du vocabulaire?". *Le francais moderne* 46(1) (1978), 25–32.

Dugast, D. *Vocabulaire et Stylistique. I Théâtre et Dialogue*. Travaux de Linguistique Quantitative. Paris: Slatkine-Champion, Geneva, 1979.

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

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

Herdan, G. "A New Derivation and Interpretation of Yule's Characteristic *K*". *Zeitschrift für Angewandte Mathematik und Physik* (1955).

Herdan, G. *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. The Hague, The Netherlands: Mouton & Co., 1960.

Herdan, G. *Quantatative Linguistics*. London: Butterworth, 1964.

Holmes, D. I. "A Stylometric Analysis of Mormon Scripture and Related Texts". *Journal of the Royal Statistical Society Series A* 155(1) (1992), 91–120.

Holmes, D. I. "Authorship Attribution". *Computers and the Humanities* 28(2) (1994), 87–106.

Holmes, D. I. and R. S. Forsyth. "The Federalist Revisited: New Directions in Authorship Attribution". *Literary and Linguistic Computing* 10(2) (1995), 111–127.

Honoré, A. "Some Simple Measures of Richness of Vocabulary". *Association for Literary and Linguistic Computing Bulletin* 7(2) (1979), 172–177.

Johnson, N. L. and S. Kotz. *Urn Models and their Application. An Approach to Modern Discrete Probability Theory*. New York: John Wiley and Sons, 1977.

Johnson, R. "Measures of Vocabulary Diversity". In *Advances in Computer-aided Literary and Linguistic Research*. Eds. D. E. Ager, F. E. Knowles and M. W. A. Smith, AMLC, 1979.

Maas, H.-D. "Zusammenhang zwischen wortschatzumfang und länge eines textes". *Zeitschrift für Literaturwissenschaft und Linguistik* 8 (1972), 73–79.

Martindale, C. and D. McKenzie. "On the Utility of Content Analysis in Author Attribution: The Federalist". *Computers and the Humanities* 29 (1995), 259–270.

Ménard, N. *Mesure de la Richesse Lexicale. Théorie et vérifications expérimentales. Etudes stylométriques et sociolinguistiques*. Number 14 in Travaux de Linguistique Quantitative. Paris: Slatkine-Champion, Geneva, 1983.

Michéa, R. "Répétition et variété dans l'emploi des mots". *Bulletin de la société de linguistique de Paris* (1969).

Michéa, R. "De la relation entre le nombre des mots d'une fréquence déterminée et celui des mots différents employés dans le texte". *Cahiers de Lexicologie* (1971).

Mosteller, F. and D. L. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, 1964.

Orlov, Y. K. "Ein modell der häufigkeitsstruktur des vokabulars". In *Studies on Zipf's Law*. Bochum: Brockmeyer, 1983, pp. 154–233.

Sichel, H. S. "On a Distribution Law for Word Frequencies". *Journal of the American Statistical Association* 70 (1975), 542–547.

Sichel, H. S. "Word Frequency Distributions and Type-token Characteristics". *The Mathematical Scientist* 11 (1986), 45–72.

Simpson, E. H. "Measurement of Diversity". *Nature* 163 (1949), 168.

Thoiron, P. "Diversity Index and Entropy as Measures of Lexical Richness". *Computers and the Humanities* 20 (1986), 197–202.

Tuldava, J. "Quantitative Relations between the Size of the Text and the Size of Vocabulary". *SMIL Quarterly, Journal of Linguistic Calculus* 4 (1977).

Tweedie, F. J., D. I. Holmes and T. N. Corns. "The Provenance of *De Doctrina Christiana*, Attributed to John Milton: A Statistical Investigation". *Literary and Linguistic Computing* 13(2) (1998), 77–87.

Weitzman, M. "How Useful is the Logarithmic Type-token Ratio?". *Journal of Linguistics* 7 (1971), 237–243.

Whissell, C. "Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon". *Computers and the Humanities* 30(3) (1996), 257–265.

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