Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution

HARALD BAAYEN
Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands

HANS VAN HALTEREN
Catholic University of Nijmegen, Nijmegen, The Netherlands

FIONA TWEEDIE
The University of the West of England, Bristol, UK

Abstract

This paper reports an experiment in authorship attribution in which statistical measures and methods that have been widely applied to words and their frequencies of use are applied to rewrite rules as they appear in a syntactically annotated corpus. The outcome of this experiment suggests that the frequencies with which syntactic rewrite rules are put to use provide a better clue to authorship than word usage. Complementary methods focusing on the high-frequency head and the low-frequency tail of the distribution independently reveal a higher resolution than traditional word-based analyses, and promise enhanced accuracy for authorship attribution.

1. Introduction

A number of recent contributions to authorship attribution are based on words and their frequencies of occurrence (see, for example, Burrows (1992, 1993), Holmes and Forsyth (1995) and Holmes (1994) for a general review of methods for authorship attribution). This comes as no surprise, as the statistical analysis of word frequencies requires minimal textual preprocessing. Nevertheless, precisely those words which have proved to have a high discriminatory resolution in the seminal work by Burrows (1992, 1993), the so-called function words (*a*, *the*, *that*, *and*, *but*, . . ., etc.), appear to tap into the use of syntax. This suggests it might be profitable to study the use of syntax directly.

We designed a statistical experiment using syntactically annotated corpus material to investigate the discriminatory potential of syntactic rewrite rules for authorship attribution. We followed tradition, as exemplified in the study by Mosteller and Wallace (1964), in that we compared texts of unknown authorship with texts of which authorship is beyond doubt. In our experiment, however, the authorship of all texts was known, albeit initially only to the experiment leader, van Halteren, and not to Tweedie and Baayen, who carried out the analyses. This considerably simplifies the process of evaluating the accuracy of the methods we have used.

The texts, their syntactic annotation, and the details of the design of our statistical experiment, are introduced in Section 2. In Section 3, we set ourselves a baseline by studying the accuracy of word-based methods. In Section 4, we show that analyses that exploit the frequencies of syntactic rewrite rules instead of the frequencies of words have an increased discriminatory potential. Some evidence that the frequencies of rewrite rules are less subject to intratextual variation than the frequencies of words is discussed in Section 5. Our conclusions are presented in Section 6.

2. Experimental Design

Our investigation makes use of material from the Nijmegen corpus (Keulen, 1986). This corpus comprises a series of texts and text fragments ranging from texts on cell biology (scientific), psychology (popular scientific), and literary criticism to crime fiction, drama, and on-line tennis reports, all originating in the mid-sixties (see Table 1). Texts from this corpus have been syntactically annotated with two different analysis systems, all texts with the CCPP system (Keulen, 1986) and a substantial subset with the TOSCA system (Oostdijk, 1991).

For our main experiment, we restrict ourselves to two texts from the Nijmegen corpus, both of the same register, crime fiction. We make this choice because of the results of a pilot study, described in Section 2.1, which revealed that differences in register may override differences in authorship: texts in different registers or text types by one author may differ more than texts written by different authors in the same text type.

A positive side-effect of this restriction to the crime fiction texts is that we can use the TOSCA annotation, which is more detailed and which uses a more consistent descriptive model. Section 2.2 describes the TOSCA analyses and the form in which they were used in the experiment. The exact setup of the authorship attribution experiment, finally, is described in Section 2.3.

2.1 *Register and Authorship*

It is well-known that not only differences between authors, but also differences in register or text type are reflected in the relative frequencies of linguistic variables, many of which are syntactic in nature (Biber, 1995). Before considering questions of authorship, we therefore need to have some idea of the range of variation in the use of language for one author writing in different registers, and for different authors writing in

Correspondence: R. Harald Baayen, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD, Nijmegen, The Netherlands.

Table 1 Texts in the Nijmegen corpus used in the pilot study on register and author-specific variation

Register	Author	Size in words	Number of chunks	Code
Crime fiction	Allingham (1965)	20,244	10	1
Crime fiction	Innes (1966)	21,516	10	2
Literary criticism	Stewart (1963)	20,736	10	3
Popular scientific	Brown (1963)	20,035	10	4
Scientific	Paul (1965)	19,370	9	5
Drama	Livings (1962)	12,099	6	6
Drama	Livings (1963)	5,708	2	7
Tennis reports	Wimbledon Final (1968)	3,993	1	8
Tennis Reports	Wightman Cup (1968)	2,084	1	9

the same register. Hence, we ran a pilot study on the full range of texts available in the Nijmegen corpus, in which we examined the relative frequencies of the most frequent function words and their CCPP word category codes. These texts are listed in Table 1, together with their size in word tokens. Of special interest for our examination is that Innes and Stewart are one and the same person: for his fictional work, Stewart uses the pseudonym Innes.

We divided each text into chunks of exactly 2,000 words and discarded the remaining smaller final chunks. For each of the resulting fifty-nine chunks, we calculated the relative sample frequency of the fifty most frequent words in the pooled texts. Thus, each text fragment appeared as a single observation in a fifty-dimensional space. We used principal components analysis[1] to explore the most important dimensions of variation. Five components emerged from the analysis, the first explaining 33.5%, the second component 9.9%, and the third 8.6% of the variance, respectively. The fourth and fifth principal components jointly account for 8.4% of the variance. Figure 1 presents the scatterplots for the first three principal components.

First consider the panel for PCA 1 and PCA 2, the upper left-hand panel of Fig. 1. Clearly, chunks from the same text tend to cluster. For instance, the chunks of the text on cell biology (5, scientific) all occur at the centre right-hand edge. This suggests that the way in which an author exploits the most frequent function words and their associated syntactic constructions is reasonably consistent across a text. Interestingly, some gross differences in register are nicely reflected in the use of the highest-frequency words. For instance, the drama chunks (6, 7) cluster together in the upper left corner, the crime fiction appears more or less in the lower left quadrant of the plot (1, 2), and the scientific texts (3, 4, 5) appear at the right. Importantly, Fig. 1 clearly reveals that for one author differences in register can be much stronger than differences within a register between texts of different authors. The literary criticism of Stewart/Innes (3) patterns more closely with the chunks on psychology (4, popular scientific) and even cell biology (5, scientific) than with his own novel (1), which is much more similar to the other novel, written by Allingham (2).

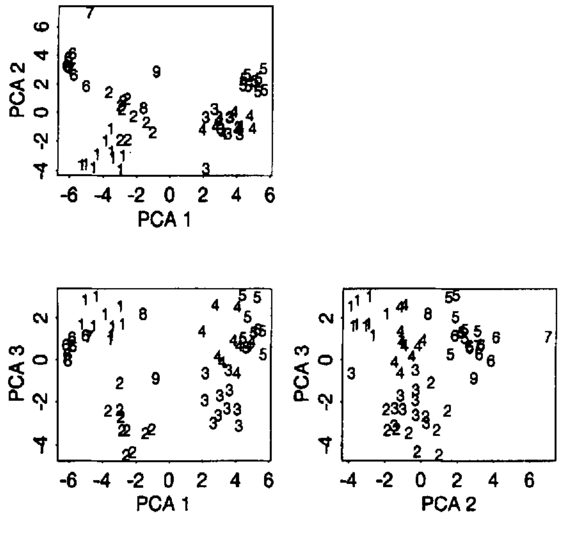

Fig. 1 Scatterplot matrix for the first three principal components based on the fifty most frequent words in the pooled vocabulary. (1,2: crime fiction; 3: literary criticism; 4: popular scientific; 5: scientific; 6,7: drama; 8,9: tennis reports.)

On the other hand, the panels on the bottom row of Fig. 1 show that when we take the third principal component into account, the texts by Stewart/Innes (2, 3) reveal negative scores where the other texts tend to show up with positive scores. In fact, on PCA 3, the chunks by Stewart/Innes and the texts by Allingham are quite well separated. Thus there are subspaces where the texts of one author may cluster, side by side with subspaces where they pattern quite differently. Figure 1 suggests that the first two dimensions, the dimensions that account for most of the variance, primarily capture variation in register. Scientific prose and literary criticism appear with positive scores; drama, tennis reports, and crime fiction, which contain substantial amounts of direct speech, appear with negative scores on PCA 1. The biology texts and the drama are separated from the psychology and crime fiction, respectively by PCA 2. But after removal of this register-bound variance, author-specific differences emerge on PCA 3 (see Binongo (1994) for similar conclusions).

Although the third principal component separates the two novels quite well, the question remains to what extent this success is co-determined by the properties of the texts from other registers that happen to be included in the analysis. Would similar results have been obtained if a random selection of texts from another random set of registers had been included? We have left this question to future research, and have opted for a controlled experiment on authorship attribution within a single register.[2] In what follows, we will therefore concentrate on the novels by Allingham and Innes, both crime fiction.

2.2 *Syntactic Annotation*

For our experiment we had available 20,000 words of running text each from M. Innes' *The Bloody Wood*

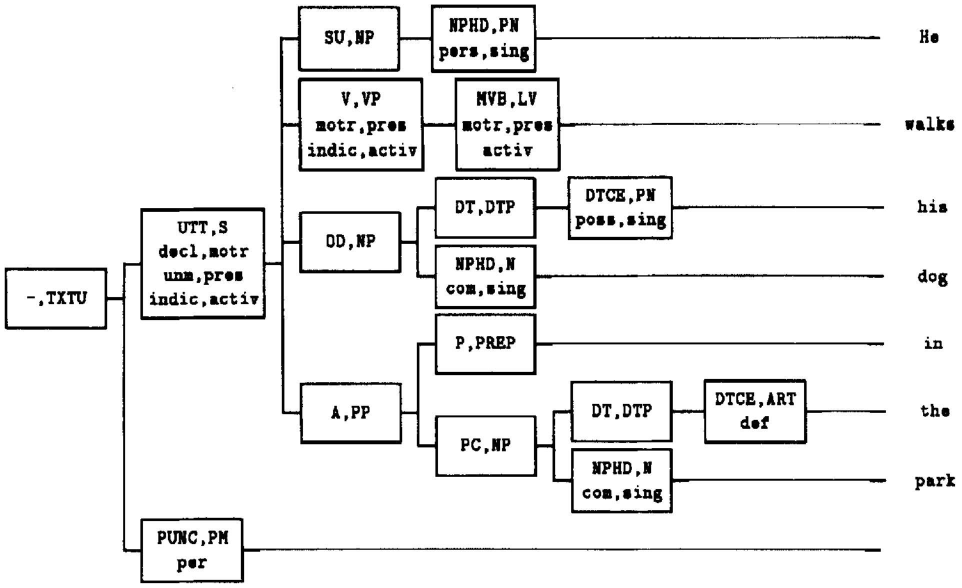

Fig. 2 A sample analysis tree for the sentence *He walks his dog in the park* using TOSCA annotation

(henceforth sample A) and from M. Allingham's *The Mind Readers* (henceforth sample B). Both samples have been syntactically annotated with the TOSCA annotation scheme. In this section, we outline this annotation scheme and the way in which we transformed the annotated texts into sequences of rewrite rules.

The syntactic annotation is exemplified in Fig. 2, which shows the analysis assigned to the sentence *He walks his dog in the park.* On each node we find labels describing the semantic and syntactic properties of the corresponding constituent. The label on the right in the first line in the node represents the syntactic category, i.e. the general nature of the constituent itself. The label on the left in the first line represents the syntactic function, i.e. the role the constituent plays in the immediately dominating larger constituent. The labels on second and further lines represent additional attributes deemed interesting in the descriptive model. Consider, for instance, the node immediately to the left of the word *park.* For this node, the syntactic category is 'Noun' (N), meaning that the constituent *park* itself is a noun. The syntactic function is 'Noun Phrase Head' (NPHD), meaning that the constituent *park* functions as the head of the constituent *the park.* The attributes for *park*, finally, are 'Common' (com) and 'Singular' (sing). A full list of all labels is given in Tables A1–A3 in the Appendix.

In order to enable the application of techniques for words and their frequencies of use to syntactic trees it was necessary to translate part of the information present in each analysis tree into a pseudo-word sequence. The question, then, is which part. We have used two criteria to decide which information to include. The first criterion requires selection of the most important information. The second criterion requires the resulting pseudo-words to be as similar to normal words as possible. The second criterion led us to exploit the individual rewrites (combinations of a node and its immediate constituents), since these are the building blocks of the tree in the same way as words are the building blocks of sentences. The first criterion led us to focus first on the category label (e.g. NP), then on the function label (e.g. SU) and only last on the attribute labels (e.g. sing) on the nodes.

For an exact choice of the information to use we counted the number of pseudo-word tokens and types. The total number of rewrite tokens in the two texts of our main experiment was 46,403.[3] Using only the category labels, e.g. (for a noun phrase like *the park*)

NP → DTP + N

led to 2318 types. Adding the function labels at the right hand side, e.g.

NP → DT:DTP + NPHD:N

increased this number to 2732. Addition of the function label on the left hand side as well,

PC:NP → DT:DTP + NPHD:N

raised the number to 4194. As the resulting type-token ratio was fairly close to that for the normal words of our samples, this is the labeling we decided to use. For the example sentence of Fig. 2 this leads to the following sequence of rewrites:

-:TXTU → UTT:S + PUNC:PM
UTT:S → SU:NP + V:VP + OD:NP + A:PP
SU:NP → NPHD:PN
V:VP → MVB:LV
OD:NP → DT:DTP + NPHD:N
DT:DTP → DTCE:PN
A:PP → P:PREP + PC:NP
PC:NP → DT:DTP + NPHD:N
DT:DTP → DTCE:ART

The most frequent rewrites are present in both samples. The first one missing in sample A is the 59th most frequent one,

UTT:COORD → CJ:S + COOR:CONJN + CJ:S

as in *Edward accused her and Sam backed his cousin up*, which occurs eighty-five times in sample B. The first one missing in sample B is the 231st most frequent one,

RPDU:CLOID → DIFU:REACT + PUNC:PM + DIFU:REACT

as in *'Yes, indeed,' Mrs Gillingham corroborated*, which occurs fourteen times in sample A. Even these simple numbers are already indicative of the pattern that will become apparent below: A is not as strongly focused on the highest-frequency rewrite rules as B, and instead shows a greater richness in the use of low-frequency rewrites.

We translated the syntactic rewrite information in the samples into pseudo-words. The main reason for this was that the existing software is likely to expect words rather than the complex (and long) expressions that make up rewrites. For the translation, we sorted the rewrites accordingly to their frequency (cumulative over both samples) and named them accordingly. Thus, the most frequent rewrite becomes W0001, the second most frequent one W0002, etc. as shown in Table 2. The translated rewrite rules were presented in the original order in which they appear in the samples. In addition, text unit separators (S) were inserted to indicate which pseudo-words together formed a pseudo-sentence (i.e. which rewrites jointly form an analysis tree). As a result, the experimenters received the following kind of data: S W0084 W3165 W0048 S W0021 W0061 W0002 W0001 W0031 W0019 S W0010 . . .

Table 2 The ten most frequent rewrite rules and their pseudo-word codes

Pseudo-word	Frequency	Rewrite rule
W0001	4670	V:VP → MVB:LV
W0002	3566	SU:NP → NPHD:PN
W0003	2674	DT:DTP → DTCE:ART
W0004	1948	A:AVP → AVHD:ADV
W0005	1729	A:PP → P:PREP + PC:NP
W0006	1435	V:VP → OP:AUX + MVB:LV
W0007	1395	NPPR:AJP → AJHD:ADJ
W0008	1172	DT:DTP → DTCE:PN
W0009	1017	PC:NP → DT:DTP + NPHD:N
W0010	1016	-:TXTU → UTT:S + PUNC:PM

2.3 *Design of the Main Experiment*

For our experiment, we had available the samples A and B in normal word form (with TOSCA wordclass tags) and in pseudo-word form as described in Section 2.2. For the evaluation of authorship attribution techniques, we split the two texts into fourteen labelled samples and six unlabelled test samples. The two pseudo-texts were both divided into ten parts, such that a new part was initiated at the first text unit separator after 2,500 pseudo-words (including separators). All parts were about the same size, except for the tenth part of sample B, which contained only 2,254 pseudo-words. The normal word versions were split in such a way that they represented the same stretch of text as the corresponding pseudo-word samples. The first seven parts of each pseudo-text were provided as labeled samples: A1–A7 and B1–B7. The remaining six parts were provided as test samples: Q1 (=A10), Q2 (=B10), Q3 (=B8), Q4 (=A8), Q5 (=B9) and Q6 (=A9). All correspondence information was withheld from the experimenters.

We now have six test samples for determining the discriminatory accuracy of authorship attribution techniques. Demanding correct attribution of, for instance, five out of six samples is not sufficient since the probability of getting at least five right by random assignment is 7/64 (0.109), a value which in our opinion is too high. In order for a technique to be found sufficiently accurate, therefore, it must provide the correct attribution for all six test samples. The probability of getting this result purely by chance is a mere 1/64 (0.016). Although this is already a rather rigorous criterion, we also wanted to ensure independence of accuracy of assignment and our particular choice of unknown text fragments. We therefore further required that a successful method should in fact group all twenty samples including the test samples into two clearly distinguishable clusters.

3. Setting the Baseline: Word-Based Methods

The intuition underlying our approach to authorship attribution by means of the frequencies of rewrite rules is that the rewrite rules are a more precise clue to authorship than the function words that have been exploited in the seminal studies by Burrows (1992, 1993). In order to evaluate potential gains in accuracy by changing from function words to rewrites, we need to know the success rate of word-based authorship attribution for the same task. We therefore carried out two analyses, one based on measures of vocabulary richness along the lines of Holmes and Forsyth (1995), and one based on the fifty most frequent words following the approach of Burrows (1992, 1993).

3.1 *Measures of Vocabulary Richness*

Various measures of vocabulary richness have recently been applied to questions of authorship attribution (see Holmes and Forsyth (1995) for application to the *Federalist Papers*, Holmes (1992) for Mormon scripture, and Holmes and Singh (1995) for aphasic speech patterns). These measures are of interest because, unlike measures such as the sample mean frequency, they are robust with respect to differences in text size. In this study, we consider five of these measures, the first of which was proposed by Yule (1994). It is defined as:

$$K = 10^4 \frac{\sum_{i=1}^{v} i^2 V(i, N) - N}{N^2}, \tag{1}$$

with N the number of tokens, $V(i, N)$ the number of types which occur i times in a sample of N tokens, and v the highest frequency of occurrence. Another measure was proposed by Simpson (1949), who focused on the probability that two words randomly selected from the text are the same. This measure is defined as

$$D = \sum_{i=1}^{v} V(i, N) \; \frac{i(i-1)}{N(N-1)}. \tag{2}$$

The values of both D and K are primarily determined by the high end of the frequency distribution structure. They quantify the repeat rate of the samples.

In order to consider the low frequency end of the distribution, we also include measures proposed by Honoré (1979) and Sichel (1975). Honoré's measure,

$$R = 100 \frac{\log N}{1 - \frac{V(1,N)}{V(N)}}, \tag{3}$$

where $V(N)$ denotes the number of different types, was used initially to examine the vocabulary of Latin judicial authors. R takes into account the probability that the author will re-use a given type in the text rather than choosing a new one. Its dependence on $V(1, N)$, the number of hapax legomena, may add useful information. Another measure that is sensitive to the low end of the frequency distribution was proposed by Sichel (1975):

$$S = V(2, N)/V(N). \tag{4}$$

By means of this measure we take the number of dis legomena, the words which appear twice in the text, into account.

Finally, we examined a variable which has measured vocabulary richness with success in various field. Proposed by Brunet (1978), it is defined as:

$$W = N^{V(N)^{-a}}, \tag{5}$$

where a is a parameter, usually fixed at 0.17, such that W is approximately constant and independent of N. Values for K, D, R, S, and W were calculated from the word frequency distributions of the twenty text samples in our experiment. In this way, we obtained twenty observations in a five-dimensional space. We used principal components analysis to select the most relevant dimensions. The analysis revealed three significant dimensions, which explain 48.5%, 41.5%, and 9.3% of the variance, respectively. The first two dimensions are shown in panel A of Fig. 3. The measures W (r = 0.956) and R (r = −0.915) are strongly correlated with the first dimension (PCA 1), K and D (r = −0.947 for both variables) are strongly correlated with the second dimension (PCA 2). This figure shows that there is a general separation of authors on the second dimension. Nevertheless, the test samples Q2 and Q5, both by B, cluster with the fragments of A, and sample A3 patterns with B on the second dimension instead of with A. Thus, this approach fails to meet our criteria for accurate classification in two ways. It fails to achieve a high accuracy for the test samples (a misclassification rate of 2/6), and it fails to separate all fragments adequately (a misclassification rate of 1/14). This result seriously questions the discriminatory potential of methods based on summary statistics of vocabulary richness for purposes of authorship attribution.

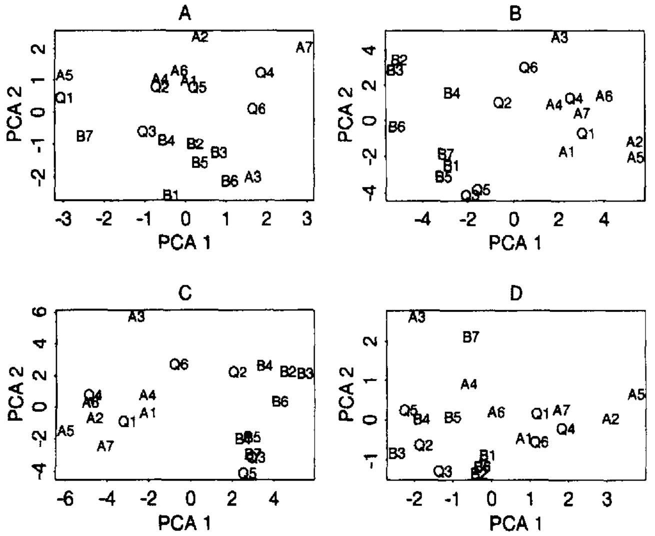

Fig. 3 Known and test samples in the plane spanned by the first two principal components. Panel A: word-based analysis of five text characteristics; panel B: analysis of the relative frequencies of the fifty most frequent function words, without distinguishing homographs with respect to category and attribute labels; panel C: analysis of the relative frequencies of the fifty most frequent function words, homographs with respect to category and attribute labels distinguished; panel D: rewrite-based analysis of five text characteristics.

3.2 *The Fifty Most Frequent Words*

We next carried out two principal components analyses on the individual sample relative frequencies of the fifty most frequent function words in the pooled samples A1–7 and B1–7. In the first analysis, we considered the frequencies of the fifty most frequent words (excluding proper names) without considering their word category and attribute labels. All twenty text samples were jointly considered in the PCA. Panel B of Fig. 3 plots all twenty samples in the plane spanned by the first two principal components, which explain 24.3% and 12.6% of the variance. Four other principal components emerged as explaining at least 5% of the variance. Jointly, they accounted for 28.2% of the variance. Only the first principal component separates the texts by A from the texts by B. This figure shows that this raw analysis does remarkably well. Except for the test sample Q2, which is by B, but which, according to a discriminant analysis with the labelled samples as training samples, is equally likely to be by A or B, both the known samples and the test samples are well separated in two distinct clusters.

In the second analysis, homographs receiving different word category and attribute labels in the TOSCA analysis were analysed as different words. For instance, *and* with the code CONJN ('conjunction'), and *and* with the code CON ('connective') were counted as two different types. This approach is more in line with the methodology of Burrows (1992, 1993), who, for example, carefully distinguishes between subordinating *that* and demonstrative *that*.

Panel C of Fig. 3 shows that all samples by A and B are now well separated in the plane spanned by the first two principal components, which account for 28.8% and 12.2% of the variance, respectively. It is the first component that crucially distinguishes between the two authors. The function words that are most highly correlated with this component are listed in Table 3. The texts by author B tend to make more use of *him* and of coordination with *and* and *but* (as in *the cat* AND *the mouse*, while author A favors *but* and *and* as connectives (as in BUT *he said that*. . .), as well as the auxiliary *would*. What we find, then, is that the CON and CONJN functions of *and* and *but* should be carefully distinguished if the required level of accuracy is to be achieved. Note, however, that the distinction between

Table 3 The function words revealing the highest correlations with the first principal component

Function word	Code	Correlation
but	CON	−0.86
would	AUX(indic, mod, past)	−0.85
and	CON	−0.79
that	CONJN(subord)	−0.78
but	CONJN(coord)	0.77
and	CONJN(coord)	0.82
him	PN(pers, sing)	0.83

and and *but* as conjunctions and connectives is a subtle one that does not appear in, for example, Burrows (1992). In fact, this distinction is entirely due to the descriptive model used in the TOSCA syntactic analysis. The crucial role of this distinction in our analysis underlines the importance of the syntactic environment in which a function word appears. Furthermore, if the accuracy of attribution increases by introducing more and more syntactic distinctions into the word based analysis, then this supports our hypothesis that it is useful to consider the frequencies of syntactic constructions directly rather than indirectly via the function words.

4. Syntax-Based Methods

To evaluate the potential of a syntax-based approach, we proceed as follows. In Section 4.1, we evaluate the use of measures of vocabulary richness, but now applied to the frequency distributions of syntactic rewrite rules. In Section 4.2, we consider the discriminatory potential of the highest-frequency rewrite rules. Finally, in Section 4.3 we investigate how the lowest-frequency rewrite rules can be exploited for authorship attribution.

4.1 *Measures of Vocabulary Richness for Rewrites*

As before, values for K, D, R, S, and W were calculated, this time on the basis of the syntactic rewrite frequency data from the twenty text samples. A principal components analysis revealed two important dimensions, of which the first explained 61.7% of the variance, and the second 21.7%. The resulting plot is shown as panel D in Fig. 3. It is clear that a fairly good inter-authorial separation is achieved by the two components. Except for B7, all samples by B cluster in the lower left corner. The samples by A appear with the higher values on both dimensions. Examination of the correlation structure indicates that R (r = 0.890) and K (r = − 0.828) are almost wholly responsible for the first principal component, while D (r = − 0.558) and S (r = 0.523) are responsible for variation in the second. The unknown samples are clearly within the A and B regions of the plot and their attributions are correct.

A and B differ most notably in their K and R values. Most texts by A have high values of R and low values of K, the reverse holds for the texts by B. In other words, the samples by A are characterized by high proportions of hapax legomena (R) and by a low repeat rate (K). For B, repetition and relatively little innovation are typical. The same kind of pattern appears in a weakened form on the second dimension, where positive scores indicate a high proportion of dislegomena [$S = V(2,N)/V(N)$] and a low repeat rate (D).

The principal components analysis is successful in correctly classifying all the unknown samples, and thereby satisfies our first criterion for accuracy. However, even though all but one of the twenty samples is correctly classified, a success rate of 95%, our second criterion is not satisfied. A different selection of unknown text fragments might have included the misclassified text B7, and hence would have led to a failure with respect to our first criterion. Nevertheless, it is clear that we have achieved an accuracy which is considerably improved compared with the corresponding word-based analysis discussed in Section 3.1.

4.2 *The Fifty Most Frequent Rewrite Rules*

Are the fifty most frequent rewrite rules in the pooled vocabulary of rewrite rules of authors A and B equally informative as the fifty most frequent words? The results of a principal components analysis of the relative frequencies of the fifty most frequent rewrites in our samples are summarized in panel A of Fig. 4. It can be seen that the samples from the second text (B1–B7, Q2, Q3, Q5) cluster in the upper left corner of the plot, while the fragments by A appear to the right. While the two components shown jointly explain 36% of the variance, it is clear from this figure that the first principal component is the main discriminator. It is highly correlated with the rewrite rules W0001 (r = − 0.832) and W0002 (r = − 0.817) listed in Table 1, as well as with W0015 (r = 0.826, PC:NP → NPHD:N) and W0013 (r = 0.758, SUB:SUBP → SBHD:CONJN). Thus, it would appear that B is using W0001 and W0002 often, while A is using W0013 and W0015 more frequently. The only fragment by A that strays in the direction of the B cluster is A3.

Panel B of Fig. 4 shows that, according to a discriminant analysis, A3 reliably sides with the samples by A.

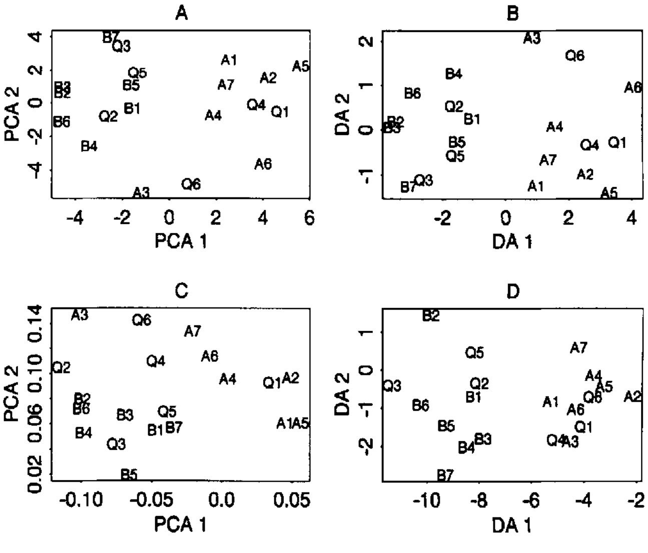

Fig. 4 Principal components analysis and discriminant scores for known and test samples. Panel A: Principal components analysis based on the sample relative frequencies of the fifty highest-frequency rewrite rules; panel B: discriminant analysis of the PCA scores of panel A; panel C; principal components analysis of the $\mathcal{H}_{i,j}$ scores; panel D: discriminant analysis of the $\mathcal{H}_{i,j}$ scores of UTT.COORD, CJ.S, and RPDU.S.

We may therefore conclude that an analysis of the highest-frequency rewrites satisfies both our criteria for accurate authorship attribution: correct assignment of all test samples, as well as a good separation of all known samples. Note that the syntax-based analysis does not point to the use of *and* and *but* as conjunctions and connectives as the most important clues to authorship, as emerged in the enhanced function-word analysis. Instead, the use of verb phrases with no auxiliary verbs and the use of subjects realized by pronouns emerge as primary discriminants. This suggests that the two approaches may be complementary.

4.3 *The Discriminatory Potential of the Lowest-Frequency Rewrite Rules*

The methods that we have used thus far exploit differences in the frequencies of the highest-frequency rewrite rules, either directly, as in the analysis based on the highest-frequency rewrites, or indirectly, via measures of the repeat rate such as D and K, which are also influenced mainly by the higher-frequency types. In this section, we pursue the hypothesis that robust clues to authorship identity should also emerge on the basis of the hapax legomena, the rewrite rules with the lowest possible frequency of use. This hypothesis is grounded in three considerations. First, if authors can be distinguished on the basis of the highest-frequency rewrites, they should also be distinguishable given the lowest-frequency rewrites. Second, words in the highest frequency ranges often have properties that are atypical for the population as a whole (see Baayen and Sproat, 1996). Hence, it is potentially rewarding to examine whether enhanced discriminatory power can be obtained by turning to the lowest-frequency types. Third, since the likelihood of storage in memory increases with frequency of use, and since awareness builds on memory, it is in the highest frequency ranges that conscious and deliberate wording and syntactic phrasing may be expected, leading to variation that is a function of, for example, narrative development rather than of an author's unconscious habitual use of language. Taken jointly, these considerations, which pertain primarily to word usage, but which may also carry over to the highest frequency rewrites, suggest that the lowest frequency ranges might provide a clue to authorship that is less contaminated by conscious rhetorical manipulation and thematic structuring that probably affect the higher-frequency units of analysis.

In Section 3.1, we have seen that global measures such as S and R are not sensitive enough for our purposes. These measures, which are functions of the numbers of dislegomena and hapax legomena respectively, have been developed as characteristic 'constants' that should reveal minimal variation as a function of the sample size N. Possibly, this property of constancy underlies their low discriminatory potential. Hence, it is useful to consider statistics for low-frequency units that are more sensitive to variations in lexical richness.

Among the low-frequency units, the hapax legomena, the units which occur once only, are of special interest. Good (1953) has shown that the likelihood of observing an unseen type is estimated by the ratio of hapax legomena to the total number of tokens: $V(1,N)/N$. In other words, $P(N) = V(1,N)/N$ estimates the rate at which new units appear, the rate at which the vocabulary of units increases. With respect to distributions of syntactic rewrite rules, this growth rate $P(N)$ estimates the probability that an author will produce a new rewrite rule that she/he has not yet used before. In other words, $P(N)$ taps into an author's syntactic creativity, and can be used to gauge how well an author has mastered the possibilities offered by the grammar.

Does $P(N)$ have a good discriminatory resolution for authorship attribution for our experiment? Text A appears to make a more productive use of syntax than text B, as both $V(N)$, the total number of different construction types, and $P(N)$ are significantly higher for A [$V(N) = 2114$, $P(\mathrm{N}) = 0.090$] than for B [$V(N) = 1883$, $P(N) = 0.074$].[4]

Not surprisingly, this difference in construction richness carries over to the seven known samples of A and B. After correcting for the differences in size of the twenty text samples, a classification tree analysis (Breiman *et al.*, 1984) on the basis of $P(N)$ correctly assigns all test samples. This positive result is counterbalanced by a rather imperfect classification of the known samples, for which the same classification tree reveals a misclassification rate of 2/14. Interestingly, using $V(N)$ instead of $P(N)$, again corrected for differences in sample size, a misclassification rate for the known samples of 1/14 is obtained. As before, all test samples are correctly assigned to their respective authors. Although $P(N)$ and $V(N)$ clearly capture important differences between our two authors, they are by themselves unable to satisfy the criteria we have set ourselves, namely, to obtain a classification with a misclassification rate of 0/20.

To increase our sensitivity to author-specific differences in the use of the lowest-frequency rewrite rules, a subclassification of the hapax legomena is required. To do so, we sorted all rewrite rules, irrespective of their frequency, according to their left-hand side, the information appearing to the left of the arrow in the rewrite rule. Some left-hand sides L appear in a great many different rewrite rules, others appear in just a few rules. We selected the left-hand sides with more than ten different right-hand sides for further analysis. There were forty-nine such left-hand sides in the pooled twenty text fragments. Let $L_i (i = 1, 2, \ldots, 49)$ denote the set of rewrite rules with the i-th left hand side, and let $h_{i,j} (j = 1, 2, \ldots, 20)$ denote the number of rewrite rules in text sample j belonging to set L_i that occur once only in sample j and that do not occur in any of the other text samples (a hapax legomenon occurring in sample j). Furthermore, let

$$\mathcal{H}_{i,j} = \frac{h_{i,j}}{\sum_{i=1}^{49} h_{i,j}} \qquad (6)$$

be the relative frequency of hapax legomena, in text j falling in category L_i with respect to the total number of hapax legomena in j summed over all forty-nine categories. The relative frequency $\mathcal{H}_{i,j}$ measures the extent to which the syntactic creativity unique to a particular author (or text sample) manifests itself in the i-th set of rewrite rules.

A principal components analysis on the (20,49) matrix of relative frequencies $\mathcal{H}_{i,j}$ revealed the pattern shown in panel C of Fig. 4. The first principal component is highly correlated with the left-hand side UTT:S ($r = 0.96$), the second principal component with the left hand sides CJ:CL ($r = -0.66$), RPDU:S ($r = 0.63$), RPGT:S ($r = -0.63$) and V:VP ($r = 0.63$). All test samples are correctly classified, and the samples by A and those by B appear well separated by the diagonal from upper left to lower right. This visual impression is strongly supported by discriminant analyses. Panel D of Fig. 4 shows the separation effected by applying discriminant analysis to the raw scores of three left-hand rewrite sides, UTT.COORD, CJ.S, and RPDU.S, which are already sufficient to distinguish between the samples by A and those by B. Many other (small) subsets of left-hand sides can be found that also separate the two authors quite well, including the rewrite sets singled out by the principal components analysis. Similar separation of the two authors can be achieved by applying discriminant analysis to the PCA scores. Interestingly, this is the only principal components analysis where we found that standardization (that is, using the correlation matrix instead of the covariance matrix) did not lead to improved results. This suggests that this technique, which satisfies our two criteria for accurate authorship attribution (correct assignment of test samples, and overall classification accuracy) is more robust than the techniques that exploit the highest-frequency rewrite rules.

5. Variability in Word Usage and the Use of Syntax

Our experiment suggests that syntactic annotation provides excellent clues for authorship attribution, and that measures focusing on syntactic creativity are especially promising. But why do analyses that are increasingly sensitive to syntactic differences lead to more accurate results? First, as we have hypothesized above, by focusing on syntax we are tapping into the more abstract, largely unconscious and hence most revealing habits of our authors. Second, the use of syntactic rules might be subject to intra-textual variation to a lesser extent than the use of words. This would tie in with our first explanation: as a textual property becomes more abstract and less directly manipulable, it is more likely to be uniformly (randomly) distributed. And if it is more uniformly distributed, it is a more reliable clue for authorship attribution.

The hypothesis of the more uniform use of syntax can be tested by making use of a technique developed in Baayen (1996). It is a well-known property of word frequency distributions that the expected vocabulary size calculated for the initial M tokens of a text may differ substantially from the observed vocabulary size $V(M)$. Baayen shows that narrative structure and thematic organization and the concomitant non-uniform use of key words in novels give rise to a development of the vocabulary size that may differ markedly from the development expected on the basis of the urn model for word frequency distributions.

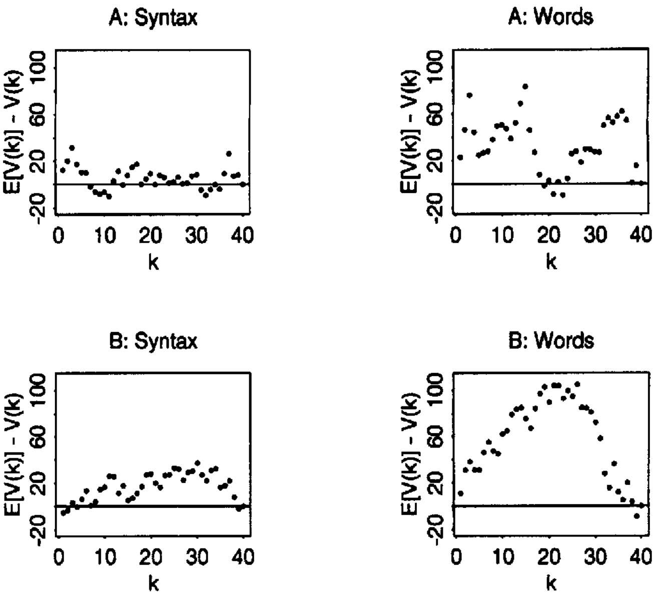

Fig. 5 Interpolation accuracy measured by $E[V(k)] - V(k)$ for the complete texts of A (top) and B (bottom) for words (right-hand panels) and rewrite rules (left-hand panels). Observations at $k = 1 \ldots 40$ equally spaced intervals in 'token time'.

The same technique can be applied to the vocabulary of rewrites. Figure 5 shows the difference between the expected and observed 'vocabulary size' ($E[V(k)] - V(k)$) for $k = 1 \ldots 40$ equidistant measurement points in 'token time' for author A (top) and author B (bottom) using words (right) and rewrite rules (left). Observe that the error scores are significantly larger for words than for rewrite rules. This observation holds not only for the absolute magnitudes. The relative size of the error scores compared with $V(k)$ is also significantly larger for words than for rewrites.[5] What these findings suggest is that the use of syntax is indeed more uniform than word usage. As we expected, narrative structure influences the non-random way in which we use words to a much greater extent than the way in which we use rewrite rules. This is in line with our intuition that as we move from the relatively concrete domain of words into the more abstract domain of syntax, the use of elementary units becomes less subject to conscious manipulation and thematic development.

Note, however, that this does not imply that the use of syntax in our texts is completely random. We still find that for smaller sample sizes the expected number of types is larger than the observed number of types. And more detailed investigations reveal that there are rewrite rules (fifteen in A, ten in B) of which it is clear that their use is significantly ($P < 0.01$) concentrated in a smaller number of fragments than expected under chance conditions.[6]

Interestingly, there are also differences between authors A and B in the developmental profile of the error scores displayed in Fig. 5. The bottom panels (B) reveal a semi-circular pattern both for words and for rewrite rules. The pattern for A is far less regular, notably for words, but also to some extent for rewrites. The pattern for B appears to be characteristic for fairly straightforward narrative development, while the pattern observed for A suggests a more complex narrative structure (see Baayen, 1996). Although the effects of narrative structure cannot be eliminated by turning from words to syntax, it nevertheless appears to be the case that syntax-based analyses are less likely to be foiled by idiosyncrasies of particular parts of the text than analyses based on word usage.

6. Discussion

Our investigation has revealed a consistent pattern in the differences between the authors A and B. A has a greater vocabulary size than B, both with respect to words and with respect to syntactic constructions. Similarly, A makes more use of morphologically complex words than B,[7] and finally, narrative development in A appears to be more complex than in B. Across the board, A reveals a more creative use of the possibilities of English. Since A (Stewart/Innes) is a literary critic as well as a writer of crime fiction, this difference comes as no surprise. It is this difference that has enabled us to tease apart text samples written by Innes from text samples written by B (Allingham).

However, in the current study, the discovery of such a consistent pattern of differences is of secondary importance. After all, this is basically a methodological study, and the focus has been on how the differences between authors are best observed. To this end, we have compared traditional word-based methods with syntax-based methods. An increase in classification accuracy could be observed for the more syntactically aware methods. We interpret this result as confirming our initial intuition that the use of function words for classification purposes is an economical way of tapping into the use of syntax, but that the direct examination of the frequencies of syntactic constructions leads to a higher discriminatory resolution. This hypothesis received explicit support from an additional analysis demonstrating that the use of rewrite rules is less variable within texts than the use of words. This indicates that more robust results may be expected with syntax-based methods than with word-based methods. In addition, both the high-frequency head of the rewrite frequency distribution as well as its low-frequency tail provide independent converging evidence for authorship, thus confirming the reliability of the syntactic approach.

Our analyses have furthermore revealed that methods based on the frequencies of large numbers of types, either high-frequency types or the lowest frequency types, are substantially more accurate than methods based on summary statistics of vocabulary richness. The latter methods pick up the main trends, but they fail in that they give rise to higher misclassification rates.

Since all this means that the proposed syntax-based methods need an advanced level of annotation, we are faced with the question how these methods can be used in actual authorship attribution practice. With the general lack of syntactically annotated text material, it is unlikely that the works in question are available in such an annotated form. An optimistic solution would be to mark exactly those properties which our experiment shows to have high discriminatory value, e.g. coordination at utterance level (UTT:COORD). It is very likely, however, that the most discriminatory properties will vary with the author pair (or group). Until more research has shown that there is indeed a specific set of highly discriminatory properties, it is safest to annotate the texts as extensively as possible[8] and let the statistical techniques select what is most interesting. Such annotation will generally entail substantial time investment,[9] but it does lead to an increase in classificatory accuracy.

Apart from the advantages of syntax-based methods our analyses have also shown the need for closer examination of the relative importance of register-specific and author-specific variation. Our exploratory pilot study revealed that register variation masks author-specific variation on the most important principal components. Nevertheless, after register variation has been partialed out, author-specific variation emerges with surprising clarity. Evidently, more systematic investigation of author-specific variation against the background of register variation is extremely promising for cross-register authorship attribution. In the light of the overall success of our experiment, syntactically annotated multi-register corpora such as those used by Biber (1995) in his studies of register variation are especially promising as frames of reference for questions of inter-register authorship attribution.

Notes

1. Unless stated otherwise, all principal components analyses made us of standardized scores (i.e. analysed the correlation structure instead of the covariance structure). In almost all our analyses, standardization led to greatly improved classifications.
2. The attribution task in this experiment is obviously much simpler than in the case that a large number of texts from many different authors written over a large number of years have to be compared. We cannot guarantee that techniques that yield accurate results in this experiment will also be useful in more complex cases. We may, however, expect that techniques that are inaccurate even for the present simple case will not be accurate either for more complex cases. Furthermore, it seems plausible that techniques that prove to be more accurate for simple cases might also prove to be more accurate for complex ones.
3. Note that the number of rewrites (46,403) is 4% smaller than the number of tokens (48,477), including punctuation), but 16% larger than the number of words (39,866, no punctuation). This means that the sample sizes in the experiment will vary with type of data used. The number of rewrites is equal to the number of non-leaf nodes in the analysis trees. In a descriptive model which uses consistent binary branching, the number of rewrites would be equal to the number of tokens minus the number of trees. However, in the TOSCA model, which allows multiple branching as well as single branching, the number of rewrites cannot be expressed as a simple function of the number of tokens. Multiple branching decreases the number of rewrites, whereas single branching increases the number of rewrites. These two effects appear to balance out for our texts, where 48,477 tokens in 3688 sentences lead to an expected number of 44,789 rewrites for binary branching. The actual number, 46,403, is only 4% higher.

[4] In both cases, $P < 0.001$, proportions test.

[5] Two-tailed paired t-tests on $|E[V(N)] - V(N)|$ for A and B revealed significant differences: $t_{(39)} = 7.98$, $P < 0.001$ for A and $t_{(39)} = 8.70$, $P < 0.001$ for B. For the relative error $|E[V(N)] - V(N)|/V(N)$, two-tailed paired t-tests similarly revealed significantly better accuracy for syntax-based estimates: $t_{(39)} = 6.18$, $P < 0.001$ for A, and $t_{(39)} = 6.34$, $P < 0.001$ for B.

[6] See Baayen (1996) for the permutation test for underdispersion used here.

[7] For instance, in the known samples, A uses significantly more derived words in *-ness*, *-less*, *un-*, *in-*, *-able*, *-ity*, *-(at)ion*, *-ize*, *-ian*, *-ment*, *-ly*, and *-er* than B ($P < 0.001$, two-sided proportions test with continuity correction). In line with results reported in Baayen (1994), especially the use of adverbial *-ly* appears to be a good discriminant, leading to an overall misclassification rate of only 1/20.

[8] A variable here, obviously, is the annotation scheme to be used for this. We have shown the TOSCA scheme to yield good results, even without the use of the attribute labels (any readers wishing to apply this scheme should contact Hans van Halteren). We expect other schemes (as long as they are fairly detailed) to do as well, although the differences in annotation may lead to shifts in the most discriminatory properties. It is not possible to compare text properties if the texts have not all been annotated in the same scheme.

[9] We are not very optimistic about the use of fully automatic parsers, but follow-up research should not disregard this possibility.

References

Allingham, M. (1965). *The Mind Readers.* Chatto & Windus, London.

Baayen, R. H. (1994). Derivational Productivity and Text Typology. *Journal of Quantitative Linguistics*, 1: 16–34.

——(1996). The Effects of Lexical Specialization on the Growth Curve of the Vocabulary. *Computational Linguistics* 22.4, in press.

——and Sproat, R. (1996). Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms. *Computational Linguistics* 22.1, in press.

Biber, D. (1995). *Dimensions of Register Variation.* Cambridge University Press, Cambridge.

Binongo, J. N. G. (1994). Joaquin's 'Joaquinesqerie', 'Joaquinesqerie's Joaquin: A Statistical Study of a Filipino Writer's Style. *Literary and Linguistic Computing*, 9: 267–279.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.

Brown, J. (1963). *Techniques of Persuasion.* Penguin Books, Harmondsworth.

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

Burrows, J. F. (1992). Computers and the Study of Literature. In C. S. Butler (ed.), *Computers and Written Texts*, Blackwell, Oxford, pp. 167–204.

——(1993). Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative. In S. Hockey and N. Ide (eds), *Research in Humanities Computing '92*, Oxford University Press, Oxford, in press.

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

Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society (A)*, 155: 91–120.

——(1994). Authorship Attribution. *Computers and the Humanities*, 28: 87–106.

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

——and Singh, S. (1995). A Stylometric Analysis of Conversational Speech of Aphasic Patients. *Technical Report No. 5, Faculty of Computer Studies and Mathematics*, University of the West of England, Bristol, December 1995.

Honoré, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7.2: 172–177.

Innes, M. (1966). *The Bloody Wood.* Victor Gollancz, London.

Keulen, F. (1986). The Dutch Computer Corpus Pilot Project. In J. Aarts and W. Meijs (eds) *Corpus Linguistics. II. New studies in the analysis and exploitation of computer corpora*, Rodopi, Amsterdam.

Livings, H. (1962). *Stop it, Whoever You Are.* Penguin Books, Harmondsworth.

——(1963). *Nil Carborundum.* Penguin Books, Harmondsworth.

Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference. The Case of the Federalist Papers.* Springer, New York.

Oostdijk, N. (1991). *Corpus Linguistics and the Automatic Analysis of English.* Rodopi, Amsterdam.

Paul, J. (1965) *Cell Biology.* Heineman Educational Books, London.

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

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

Stewart, J. I. M. (1963). *Eight Modern Writers.* Clarendon Press, Oxford.

Wimbledon Final (1968). BBC TV 1. A transcript of part of the commentary on the match between Laver and Roche on July 5, 1968. Commentators: D. Maskell, J. Kramer, and D. Coleman.

Wightman Cup. (1968). BBC TV 1. A transcript of part of the commentary on the singles matches of Christine Jones versus Nancy Richey and Virginia Wade versus Mary Anne Eisel; and on the doubles match of Winnie Shaw and Virginia Wade versus Nancy Richey and Mary Anne Eisel. Commentators: P. West and D. Maskell.

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

Appendix: The subset of Syntactic Annotation Labels (TOSCA) Used in this Paper

Table A1 Category labels

Abbreviation	Full name
ADJ	Adjective
ADV	Adverb
AJP	Adjective phrase
ART	Article
AUX	Auxiliary
AVP	Adverb phrase
CL	Clause
CLOID	Clausoid
CON	Connective
CONJN	Conjunction
COORD	Coordination
DET	Determiner
DTP	Determiner phrase
LV	Lexical verb
N	Noun
NP	Noun phrase
PM	Punctuation mark
PN	Pronoun
PP	Prepositional phrase
PREP	Preposition
REACT	Reaction signal
S	Sentence
SUBP	Subordinator phrase
TXTU	Textual unit
VP	Verb phrase

Table A2 Function labels

Abbreviation	Full name
A	Adverbial
AJHD	Adjective phrase head
AVHD	Adverb phrase head
CJ	Conjoin
COOR	Coordinator
DIFU	Discourse function
DT	Determiner
DTCE	Central determiner
MVB	Main verb
NPHD	Noun phrase head
OD	Direct object
P	Preposition
PC	Prepositional complement
PUNC	Punctuation
RPDU	Reported utterance
RPGT	Reporting tail
SBHD	Subordinator phrase head
SU	Subject
SUB	Subordinator
UTT	Utterance
V	Verb

Table A3 Attribute labels

Abbreviation	Full name
activ	active
com	common
coord	coordinating
decl	declarative
def	definite
indic	indicative
mod	modal
motr	monotransitive
past	past
per	period
pers	personal
poss	possessive
pres	present
sing	singular
subord	subordinating
unm	unmarked