Computers and the Humanities **35:** 193–214, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Computer-Based Authorship Attribution Without Lexical Measures

E. STAMATATOS, N. FAKOTAKIS and G. KOKKINAKIS
Dept. of Electrical and Computer Engineering, University of Patras, 265 00 – Patras, Greece
(E-mail: stamatatos@wcl.ee.upatras.gr)

Abstract. The most important approaches to computer-assisted authorship attribution are exclusively based on lexical measures that either represent the vocabulary richness of the author or simply comprise frequencies of occurrence of common words. In this paper we present a fully-automated approach to the identification of the authorship of unrestricted text that excludes any lexical measure. Instead we adapt a set of style markers to the analysis of the text performed by an already existing natural language processing tool using three stylometric levels, i.e., token-level, phrase-level, and analysis-level measures. The latter represent the way in which the text has been analyzed. The presented experiments on a Modern Greek newspaper corpus show that the proposed set of style markers is able to distinguish reliably the authors of a randomly-chosen group and performs better than a lexically-based approach. However, the combination of these two approaches provides the most accurate solution (i.e., 87% accuracy). Moreover, we describe experiments on various sizes of the training data as well as tests dealing with the significance of the proposed set of style markers.

1. Introduction

The vast majority of the attempts to attribute authorship deal with the establishment of the authorship of anonymous or doubtful literary texts. A typical paradigm is the case of the *Federalist Papers*, twelve of which are claimed by both Alexander Hamilton and James Madison (Mosteller and Wallace, 1984; Holmes and Forsyth, 1995). However, the use of such cases as testing-ground may cause some problems, namely:

- The number of candidate authors is usually limited (i.e., two or three). The tested technique, therefore, is likely to be less accurate in cases with more candidates (e.g., more than five).
- The literary texts are usually long (i.e., several thousands of words). Thus, a method requiring a quite high text-length in order to provide accurate results cannot be applied to relatively short texts.
- The literary texts often are not homogenous since they may comprise dialogues, narrative parts, etc. An integrated approach, therefore, would require the development of text sampling tools for selecting the parts of the text that best illustrate an author's style.

The lack of a formal definition of an author’s idiosyncratic style leads to its representation in terms of a set of measurable patterns (i.e., style markers). The most important approaches to authorship attribution are exclusively based on lexical measures that either represent the vocabulary richness of the author or simply comprise frequencies of occurrence of function (or context-free) words (Holmes, 1994). Tallentire (1973) claims that:

> “No potential parameter of style below or above that of the word is equally effective in establishing objective comparison between authors and their common linguistic heritage.”

However, the use of measures related to syntactic annotation has been proved to perform at least as well as the lexical ones. Baayen et al. (1996) used frequencies of use of rewrite rules as they appear in a syntactically annotated corpus. The comparison of their method with the lexically-based approaches for the *Federalist Papers* case shows that the frequencies with which syntactic rewrite rules are put to use perform better than word usage. On the other hand, they note:

> “We are not very optimistic about the use of fully automatic parsers, but follow-up research should not disregard this possibility.”

A typical approach to authorship attribution initially defines a set of style markers and then either counts manually these markers in the text under study or tries to find computational tools that can provide these counts reliably. The latter approach often requires manual confirmation of the automatically-acquired measures. In general, real natural language processing (NLP) (i.e., computational syntactic, semantic, or pragmatic analysis of text) is avoided since current NLP tools do not manage to provide very high accuracy dealing with unrestricted text. The use of computers regarding the extraction of stylometrics has been limited to auxiliary tools (e.g., simple programs for counting word frequencies fast and reliably). Hence, authorship attribution studies so far may be considered as *computer-assisted* rather than *computer-based*.

An alternative method aiming at the automatic selection of style markers has been proposed by Forsyth and Holmes (1996). In particular, they performed text categorization experiments (including authorship determination) letting the computer to find the strings that best distinguish the categories of a given text corpus by using the Monte-Carlo feature finding procedure. The reported results show that the frequencies of the automatically extracted strings are more effective than letter or word frequencies. This method requires minimal computational processing since it deals with low-level information. Although it is claimed that this information can be combined with syntactic and/or semantic markers, it is not clear how existing NLP tools could be employed towards this direction.

In this paper we present a fully-automated approach to the identification of authorship of unrestricted text. Instead of predefining a set of style markers and then trying to measure them as reliably as possible, we consider the analysis of the text by an already existing NLP tool and attempt to extract as many style markers as possible. In other words, the set of the style markers is adapted to the automatic analysis of the text.

Our method excludes any distributional lexical measure. Instead it is based on both low-level measures (e.g., sentence length, punctuation mark count, etc.) and syntax-based ones (e.g., noun phrase count, verb phrase count etc.). Additionally, we propose a set of style markers related to the particular method used for analyzing the text (analysis-level measures), i.e., an alternative way of capturing the stylistic information. The presented experiments are based on texts taken from a Modern Greek weekly newspaper. We show that the proposed set of style markers is able to distinguish reliably the authors of a randomly-chosen group and performs better than the lexically-based approaches.

This paper is organized as follows: the next Section contains a brief review of lexically-based authorship attribution studies. Section 3 describes our approach concerning both the extraction of style markers and the disambiguation method. Analytical experimental results are included in Section 4 while the conclusions drawn by this study are discussed in Section 5.

2. Lexically-Based Methods

The first pioneering works in authorship attribution had been based exclusively on low-level measures such as word-length (Brinegar, 1963), syllables per word (Fucks, 1952), and sentence-length (Morton, 1965). It is not possible for such measures to lead to reliable results. Therefore, they can only be used as complement to other, more complicated features. Currently, authorship attribution studies are dominated by the use of lexical measures. In a review paper Holmes (1994) asserts:

> "...yet, to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items."

There are two main trends in lexically-based approaches: (i) those that represent the vocabulary richness of the author and (ii) those that are based on frequencies of occurrence of individual words.

In order to capture the diversity of an author's vocabulary various measures have been proposed. The most typical one is the type-token ratio V/N where V is the size of the vocabulary of the sample text, and N is the number of tokens which form the sample text. Another way of measuring the diversity of the vocabulary is to count how many words occur once (i.e., *hapax legomena*), how many words occur twice (i.e., *dislegomena*) etc. These measures are strongly dependent on text-length. For example, Sichel (1986) shows that the proportion of the dislegomena is unstable for $N < 1{,}000$. In order to avoid this dependency many researchers have proposed functions that are claimed to be constant with respect to text-length. Typical paradigms are the *K* proposed by Yule (1944) and the *R* proposed by Honore (1979):

$$K = \frac{10^4(\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2}$$

$$R = \frac{(100 log N)}{(1 - (\frac{V_1}{V}))}$$

where V_i is the number of words used exactly i times in the text. In addition, there are approaches based on multivariate techniques, i.e., using more than one vocabulary richness function for achieving more accurate results (Holmes, 1992). However, recent studies have shown that the majority of these functions are not really text-length independent (Tweedie and Baayen, 1998). Moreover, the vocabulary richness functions are highly unstable for text-length smaller than 1,000 words.

Instead of counting how many words are used a certain number of times an alternative approach could examine how many times individual words are used in the text under study. The selection of context-free or function words that best distinguish a given group of authors requires a lot of manual effort (Mosteller and Wallace, 1984). Moreover, the function word set that manages to distinguish a given group of authors cannot be applied to a different group of authors with the same success (Oakman, 1980). Burrows (1987, 1992) used the frequencies of occurrence of sets (typically 30 or 50) of the most frequent words making no distinction between function-words and content-words. This seems to be the most promising method since it requires minimal computational cost and achieves remarkable results for a wide variety of authors. The separation of common homographic forms (e.g., the word “to” has a prepositional and an infinitive form) improves the accuracy. However, regarding a fully-automated system this separation demands the development of a reliable NLP tool able to recognize the appropriate word forms. Additionally, in case where the proper names have to be excluded from the high frequency set, an automatic name finder has also to be incorporated.

3. Our Approach

As mentioned above the set of style markers used in this study does not employ any distributional lexical measure. Instead it takes full advantage of the analysis of the text by a natural language processing tool. An overview of our approach is shown in Figure 1. In this section we first describe in brief the properties of this tool and then the set of style markers is analytically presented. Finally, we describe the classification method used in the experiments of the next section.

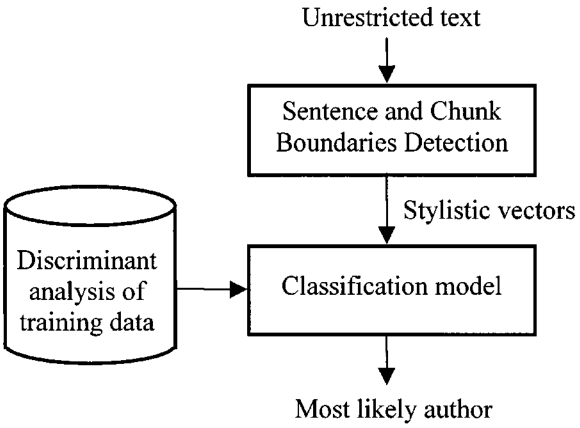
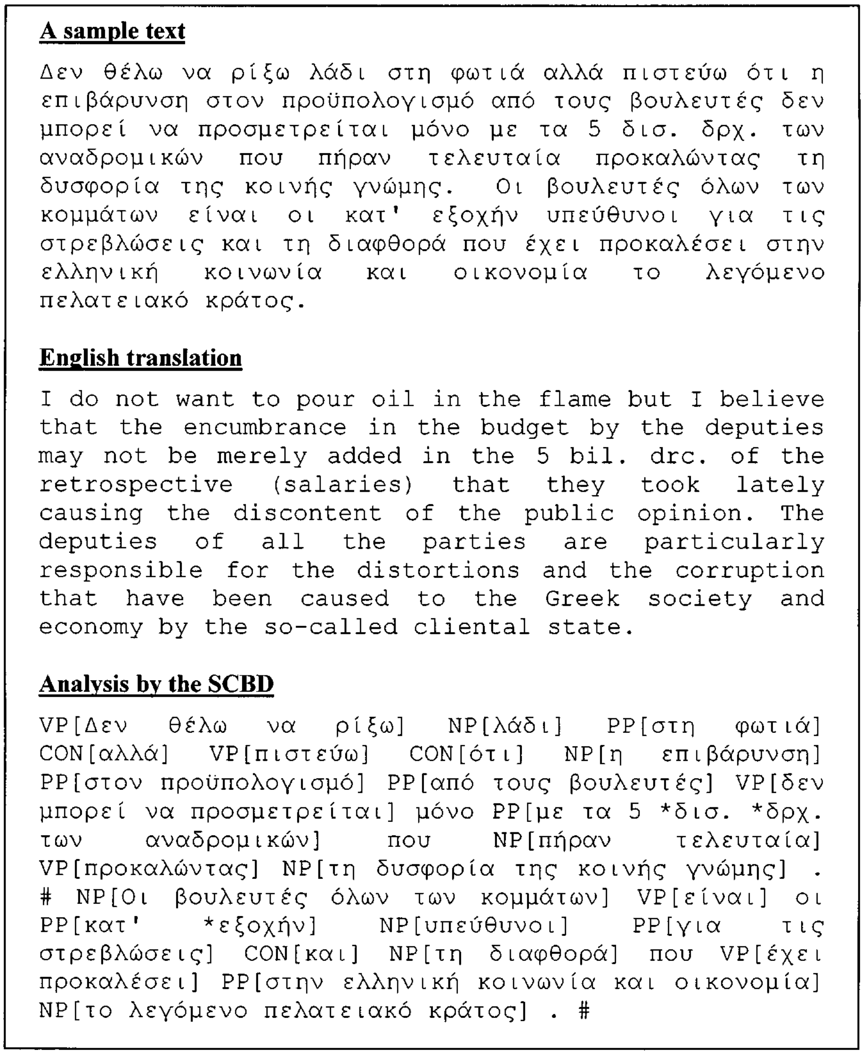

Unrestricted text
Sentence and Chunk Boundaries Detection
Stylistic vectors
Discriminant analysis of training data
Classification model
Most likely author

Figure 1. Overview of our approach.

3.1. TEXT ANALYSIS

The already existing NLP tool we used is a Sentence and Chunk Boundaries Detector (SCBD) able to analyze unrestricted Modern Greek text (Stamatatos et al., 2000). In more detail, this tool performs the following tasks:

- It detects the sentence boundaries in unrestricted text based on a set of automatically extracted disambiguation rules (Stamatatos et al., 1999b). The punctuation marks considered as potential sentence boundaries are: period, exclamation point, question mark, and ellipsis.
- It detects the chunk boundaries (i.e., non-overlapping intrasentencial phrases) within a sentence based on a set of keywords (i.e., closed-class words such as articles, prepositions, etc.) and common word suffixes taking advantage of the linguistic properties of Modern Greek (e.g., quasi-free word order, highly inflectional). Initially, a set of morphological descriptions is assigned to each word of the sentence not included in the keyword lexicon according to its suffix. If a word suffix does not match any of the stored suffixes then no morphological description is assigned. Such non-matching words are marked as special ones but they are not ignored in subsequent analysis. Then, multiple-pass parsing is performed (i.e., five passes). Each parsing pass analyzes a part of the sentence, based on the results of the previous passes, and the remaining part is kept for the subsequent passes. In general, the first passes try to detect simple cases that are easily recognizable, while the last passes deal with more complicated ones. Cases that are not covered by the disambiguation rules remain unanalyzed. The detected chunks may be noun phrases (NPs), prepositional phrases (PPs), verb phrases (VPs), and adverbial phrases (ADVPs). In addition, two chunks are usually connected by a sequence of conjunctions (CONs).



A sample text

Δεν θέλω να ρίξω λάδι στη φωτιά αλλά πιστεύω ότι η επιβάρυνση στον προϋπολογισμό από τους βουλευτές δεν μπορεί να προσμετρείται μόνο με τα 5 δισ. δρχ. των αναδρομικών που πήραν τελευταία προκαλώντας τη δυσφορία της κοινής γνώμης. Οι βουλευτές όλων των κομμάτων είναι οι κατ' εξοχήν υπεύθυνοι για τις στρεβλώσεις και τη διαφθορά που έχει προκαλέσει στην ελληνική κοινωνία και οικονομία το λεγόμενο πελατειακό κράτος.

English translation

I do not want to pour oil in the flame but I believe that the encumbrance in the budget by the deputies may not be merely added in the 5 bil. drc. of the retrospective (salaries) that they took lately causing the discontent of the public opinion. The deputies of all the parties are particularly responsible for the distortions and the corruption that have been caused to the Greek society and economy by the so-called cliental state.

Analysis by the SCBD

VP[Δεν θέλω να ρίξω] NP[λάδι] PP[στη φωτιά] CON[αλλά] VP[πιστεύω] CON[ότι] NP[η επιβάρυνση] PP[στον προϋπολογισμό] PP[από τους βουλευτές] VP[δεν μπορεί να προσμετρείται] μόνο PP[με τα 5 *δισ. *δρχ. των αναδρομικών] που NP[πήραν τελευταία] VP[προκαλώντας] NP[τη δυσφορία της κοινής γνώμης] . # NP[Οι βουλευτές όλων των κομμάτων] VP[είναι] οι PP[κατ' *εξοχήν] NP[υπεύθυνοι] PP[για τις στρεβλώσεις] CON[και] NP[τη διαφθορά] που VP[έχει προκαλέσει] PP[στην ελληνική κοινωνία και οικονομία] NP[το λεγόμενο πελατειακό κράτος] . #

Figure 2. Analysis of a sample text by the SCBD tool.

SCBD can cope rapidly with any piece of text, even ill-formed, and has been tested on an approximately 200,000 word corpus composed of journalistic text achieving 99.4% accuracy for sentence boundary detection as well as roughly 90% and 95% *recall* and *precision* results respectively for chunk boundary detection. An analysis example of a sample text is shown in Figure 2 (notice that non-matching words are marked with an asterisk and sentence boundaries are marked with a #). In order to allow the reader to understand the syntactic complexities a rough English translation is also provided.

3.2. STYLOMETRIC LEVELS

The style markers presented in this section try to exploit the output of SCBD and capture the useful stylistic information in any possible way. Towards this end we defined three stylometric levels. The first two levels dealing with the output produced by the SCBD, are:

- **Token-level:** The input text is considered as a sequence of tokens grouped in sentences. This level is based on the output of the sentence boundary detector. There are three such style markers:

Code	*Description*
M01	*detected sentences/words*
M02	*punctuation marks/words*
M03	*detected sentences/ potential sentence boundaries*

Detected sentences are the sentence boundaries found by SCBD while *words* is the number of word-tokens that compose the text. Sentence-length is a traditional and well-studied measure in authorship attribution studies and the use of punctuation is a very important characteristic of the personal style of an author. Moreover, regarding M03, any period, exclamation mark, question mark, and ellipsis is considered as potential sentence boundary. However, not all of them are actual sentence boundaries (e.g., a period may be included in a abbreviations). This marker is a strong stylistic indicator and is used here for first time.

- **Phrase-level:** The input text is considered as a sequence of phrases (i.e., chunks). Each phrase contains at least one word. This level is based on the output of the chunk boundary detector. There are ten such style markers:

Code	*Description*
M04	*detected NPs/total detected chunks*
M05	*detected VPs/total detected chunks*
M06	*detected ADVPs/ total detected chunks*
M07	*detected PPs/total detected chunks*
M08	*detected CONs/total detected chunks*
M09	*words included in NPs/detected NPs*
M10	*words included in VPs/detected VPs*
M11	*words included in ADVPs/detected ADVPs*
M12	*words included in PPs/detected PPs*
M13	*words included in CONs/detected CONs*

M04 to M08 are merely calculated by measuring the number of detected chunks of each category (i.e., NPs, PPs, etc.) as well as the total number of detected chunks. Moreover, the calculation of M09 to M13 requires the additional simple measure of the number of word-tokens that are included in chunk brackets for each category. Phrase-level markers are indicators of various stylistic aspects (e.g., syntactic complexity, formality, etc.).

Since SCBD is an automated text-processing tool, the style markers of the above levels are measured approximately. Depending on the complexity of the text in question the provided measures may vary from the real values which can only be measured manually. In order to face this problem we defined a third level of style markers:

- **Analysis-level:** It comprises style markers that represent the way in which the input text has been analyzed by SCBD. These markers are an alternative way of capturing the stylistic information that cannot be represented reliably by the two previous levels. There are 9 such style markers:

  Code *Description*

  M14 *detected keywords/words*. The number of the word-tokens found in the text that match an entry of the keyword lexicon is divided by the total word-tokens that compose the text.

  M15 *non-matching words/words*. The number of the word-tokens that do not match any entry of either the keyword or the suffix lexicon is divided by the total word-tokens that compose the text.

  M16 *words' morphological descriptions/words*. This marker requires the calculation of the number of the total morphological descriptions assigned to each word-token either by the keyword or the suffix lexicon.

  M17 *chunks' morphological descriptions/total detected chunks*. During the construction of a chunk, the morphological descriptions of the word-tokens that compose it are matched in order to form the morphological descriptions of the chunk. This marker requires the calculation of the total morphological descriptions of all the detected chunks.

  M18 *words remaining unanalyzed after pass 1/words*. The number of the word-tokens not included in any chunk brackets after the application of the first parsing pass is divided by the total number of the word-tokens that compose the text.

  M19 *words remaining unanalyzed after pass 2/words*. Same as above for the second parsing pass.

  M20 *words remaining unanalyzed after pass 3/words* Same as above for the third parsing pass.

  M21 *words remaining unanalyzed after pass 4/words*. Same as above for the fourth parsing pass.

  M22 *words remaining unanalyzed after pass 5/words*. Same as above for the fifth parsing pass.

M14 is an alternative measure of the percentage of common words (i.e., keywords) while M15 indicates the percentage of rare or foreign words in the input text. M16 is useful for representing the morphological ambiguity of the words and M17 indicates the degree in which this ambiguity has been resolved. Finally markers M18 to M22 indicate the syntactic complexity of the text. Since the first parsing passes analyze the most common cases, it is easy to understand

Table I. Values of the style markers for the sample text.

Code	Value	Code	Value	Code	Value	Code	Value
M01	0.03 (2/66)	M07	0.29 (7/24)	M13	1.00 (3/3)	M19	0.20 (13/66)
M02	0.08 (5/66)	M08	0.12 (3/24)	M14	0.54 (36/66)	M20	0.20 (13/66)
M03	0.50 (2/4)	M09	2.75 (22/8)	M15	0.05 (3/66)	M21	0.05 (3/66)
M04	0.33 (8/24)	M10	2.17 (13/6)	M16	1.62 (107/66)	M22	0.05 (3/66)
M05	0.25 (6/24)	M11	0.00	M17	1.83 (44/24)		
M06	0.00 (0/24)	M12	3.43 (24/7)	M18	0.29 (19/66)		

that a great part of a syntactically complicated text would not be analyzed by them (e.g., great values of M18, M19, and M20 in conjunction with low values of M21 and M22).

As can been seen each style marker is a ratio of two relevant measures. This approach was followed in order to achieve as text-length independent style markers as possible. Moreover, no distributional lexical measures are used. Rather, in the proposed style markers the word-token is merely used as counting unit. In order to illustrate the calculation of the proposed measures, we give the values of the complete set of style markers for the sample text of the Figure 2 in Table I.

The above analysis-level style markers can be calculated only when this particular computational tool (i.e., SCBD) is utilized. However, SCBD is a general-purpose tool and was not designed for providing stylistic information exclusively. Thus, any natural language processing tool (e.g., part-of-speech taggers, parsers, etc.) can provide similar measures. The appropriate analysis-level style markers have to be defined according to the methodology used by the tool in order to analyze the text. For example, some similar measures have been used in stylistic experiments in information retrieval on the basis of a robust parser built for information retrieval purposes (Strzalkowski, 1994). This parser produces trees in order to represent the structure of the sentences that compose the text. However, it is set to surrender attempts to parse clauses after reaching a timeout threshold. When the parser skips, it notes that in the parse tree. The measures proposed by Karlgren as indicators of clausal complexity are the average parse tree depth and the number of parser skips per sentence (Karlgren, 1999), which are analysis-level style markers.

It is worth noting that we do not claim that the proposed set of style markers is the optimal one. It could be possible, for example, to split M02 into separate measures such as periods per words, commas per words, colons per words, etc. In this paper our goal is to show how existing NLP tools can be used in authorship attribution studies and, moreover, to prove that an appropriately defined set of such style markers performs better than the traditional lexically-based measures.

3.3. CLASSIFICATION

The classification of the style marker vectors into the most likely author is performed using *discriminant analysis*. This methodology of multivariate statistics takes some training data, in other words a set of cases (i.e., style marker vectors) precategorized into naturally occurring groups (i.e., authors) and extracts a set of *discriminant functions* that distinguish the groups. The mathematical objective of discriminant analysis is to weight and linearly combine the discriminating variables (i.e., style markers) in some way so that the groups are forced to be as statistically distinct as possible (Eisenbeis and Avery, 1972). The optimal discriminant function, therefore, is assumed to be a linear function of the variables, and is determined by maximizing the between group variance while minimizing the within group variance using the training sample.

Then, discriminant analysis can be used for predicting the group membership of previously unseen cases (i.e., test data). There are multiple methods of actually classifying cases in discriminant analysis. The simplest method is based on the *classification functions*. There are as many classification functions as there are groups and each function allows us to compute classification scores for each case by applying the formula:

$$S_i = c_i + w_{i1}X_1 + w_{i2}X_2 + \ldots + w_{in}X_n$$

where $x_1, x_2, \ldots$, and x_n are the observed values of the independent variables (i.e., the style markers values) while $w_{i1}, w_{i2}, \ldots$, and w_{in} are the corresponding weights of those variables and c_i is a constant for the i-th group. S_i is the resultant classification score. Given the measures of the variables of a case, the classification scores are computed and the group with the highest score is selected.

However, in the experiments described in the next section we used a slightly more complicated classification method that is based on *Mahalonobis* distance (i.e., a measure of distance between two points in the space defined by multiple correlated variables). Firstly, for each group the location of the *centroids*, i.e., the points that represent the means for all variables in the multivariate space defined by the independent variables, is determined. Then, for each case the Mahalanobis distances from each of the group centroids are computed and the case is classified into the group with the closest one. Using this classification method we can also derive the probability that a case belongs to a particular group (i.e., *posterior probabilities*), which is roughly proportional to the Mahalanobis distance from that group centroid.

4. Experiments

4.1. CORPUS

The corpus used in this study comprises texts downloaded from the website[1] of the Modern Greek weekly newspaper entitled *TO BHMA* (the tribune). We selected

Table II. The structure of the Modern Greek weekly newspaper *TO BHMA*.

Section Code	**Title (translation)**	**Description**
A	ΤΟ ΒΗΜΑ (the tribune)	Editorials, diaries, reportage, politics, international affairs, sport reviews
B	ΝΕΕΣ ΕΠΟΧΕΣ (new ages)	Cultural supplement
C	ΤΟ ΑΛΛΟ ΒΗΜΑ (the other tribune)	Review magazine
D	ΑΝΑΠΤΥΞΗ (development)	Business, finance
E	Η ΔΡΑΧΜΗ ΣΑΣ (your money)	Personal finance
I	ΕΙΔΙΚΗ ΕΚΔΟΣΗ (special issue)	Issue of the week
S	ΒΙΒΛΙΑ (books)	Book review supplement
Z	ΤΕΧΝΕΣ ΚΑΙ ΚΑΛΛΙΤΕΧΝΕΣ (arts and artists)	Art review supplement
T	ΤΑΞΙΔΙΑ (travels)	Travels supplement

this particular newspaper since its website contains a wide variety of full-length articles and it is divided in specialized supplements. In more detail, this newspaper is composed of nine parts as it is shown in Table II. We chose to collect texts from the supplement B which includes essays on science, culture, history, etc. for three reasons:

- In such writings the idiosyncratic style of the author is not likely to be overshadowed by the characteristics of the corresponding text-genre.
- In general, the texts of the supplement B are written by scholars, writers, etc., rather than journalists.
- Finally, there is a closed set of authors that regularly contribute to this supplement. The collection of a considerable amount of texts by each author was, therefore, possible.

We selected 10 authors from the above set without taking any special criteria into account. Then, 30 texts of each author were downloaded from the website of the newspaper as shown in Table III. No manual text preprocessing nor text sampling was performed aside from removing unnecessary headings irrelevant to the text itself. All the downloaded texts were taken from issues published from 1997 till early 1999 in order to minimize the potential change of the personal style of an author over time. The last column of this table refers to the thematic area of the majority of the writings of each author. Notice that this information was not taken into account during the construction of the corpus. A subset of this corpus was used in the experiments of (Stamatatos et al., 1999a). Particularly, the presented corpus contains ten additional texts for each author.

Table III. The corpus consisting of texts taken from the weekly newspaper *TO BHMA*.

Code	**Author name**	**Texts**	**Total words**	**Average text-length (in words)**	**Thematic area**
A01	S. Alachiotis	30	30,137	1,005	Biology
A02	G. Babiniotis	30	34,747	1,158	Linguistics
A03	G. Dertilis	30	26,823	894	History, society
A04	C. Kiosse	30	50,670	1,689	Archeology
A05	A. Liakos	30	37,692	1,256	History, society
A06	D. Maronitis	30	17,166	572	Culture, society
A07	M. Ploritis	30	34,980	1,166	Culture, history
A08	T. Tasios	30	30,587	1,020	Technology, society
A09	K. Tsoukalas	30	41,389	1,380	International affairs
A10	G. Vokos	30	29,553	985	Philosophy
	TOTAL	300	333,744	1,112	

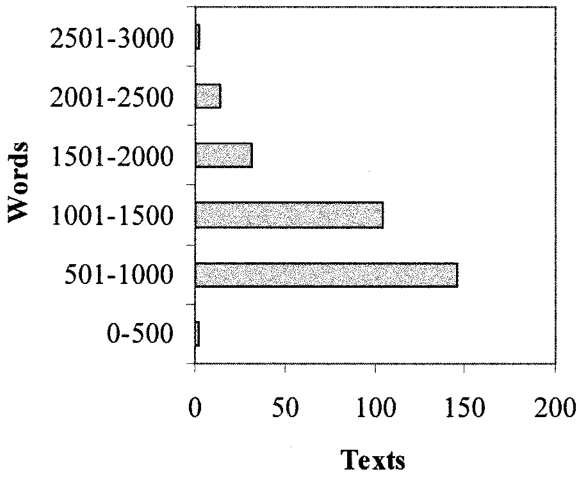


Figure 3. Text-length distribution in the corpus used in this study.

As can be seen, the text-length varies according to the author. There are three authors with average text-length shorter than 1,000 words (i.e., A03, A06, A10). The longest average text-length (i.e., of A04) is three times bigger than the shortest one (i.e., A06). Figure 3 presents the distribution of the corpus according to the text-length. Approximatelly 50% of the texts (i.e., 146 of 300) have a text-length shorter than 1,000 words.

Table IV. The fifty most frequent words of the training corpus in alphabetical order.

ακόμη	έχει	μια	που	τη
αλλά	η	μόνο	πρέπει	την
αν	ή	μπορεί	σε	της
από	ήταν	να	στα	τις
αυτή	θα	ο	στη	το
αυτό	και	οι	στην	τον
για	κατά	όμως	στις	του
δεν	κι	οποία	στο	τους
είναι	μας	όπως	στον	των
ένα	με	ότι	τα	ως

This corpus was divided into a training and a test corpus consisting of 20 and 10 texts respectively. The test corpus is the same one used in (Stamatatos et al., 1999a).

4.2. BASELINE

In order to set a baseline for the evaluation of the proposed method we decided to implement also a lexically-based approach. As aforementioned the two state-of-the-art methodologies in authorship attribution are the multivariate vocabulary richness analysis and the frequency of occurrence of the most frequent words.

The former approach is based on functions such as the Yule's *K*, the Honore's *R*, etc. in order to represent the diversity of the vocabulary used by the author. Several functions have been proved to be quite stable over text-length. However, the majority of them are quite unstable for text-length smaller than 1,000 words. Therefore, a method based on multivariate vocabulary richness analysis cannot be applied to our corpus since approximately 50% of the texts have a text-length smaller than 1,000 words (see Figure 3).

The latter approach has been applied to a wide variety of authors achieving remarkable results. It is based on frequencies of occurrence of the most frequent function words (typically sets of thirty or fifty most frequent words).

Initially, the fifty most frequent words in the training corpus were extracted. These words are presented in Table IV. No proper names are included in this list. We, then, performed discriminant analysis on the frequencies of occurrence of these words normalized by the text-length in the training corpus. The acquired classification models were, then, cross-validated on the test corpus. The confusion matrix of this experiment is shown in Table V.

Table V. The confusion matrix of the lexically-based approach (i.e., 50 style markers).

Actual	Guess										Error
	A01	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	**A10**	
A01	**5**	1	1	1	0	0	0	0	0	2	0.5
A02	1	**8**	0	1	0	0	0	0	0	0	0.2
A03	0	1	**3**	2	0	1	0	0	2	1	0.7
A04	0	0	0	**10**	0	0	0	0	0	0	0.0
A05	0	0	0	0	**9**	0	0	0	0	1	0.1
A06	2	0	1	1	0	**5**	1	0	0	0	0.5
A07	0	0	0	0	0	0	**10**	0	0	0	0.0
A08	0	1	0	0	0	0	0	**9**	0	0	0.1
A09	1	0	0	2	0	0	0	0	**7**	0	0.3
A10	0	0	0	0	0	0	1	0	1	**8**	0.2
										Average	0.26

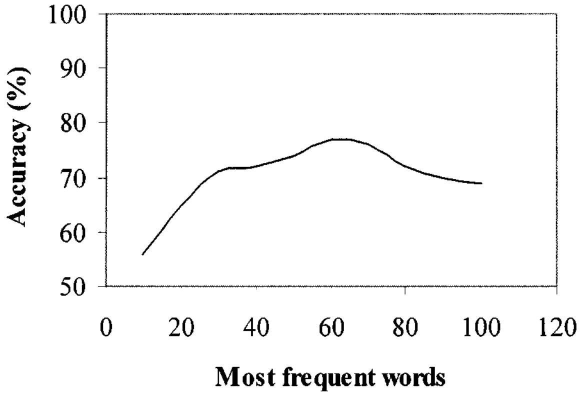


Figure 4. Classification accuracy for different sets of the most frequent words.

Each row contains the classification of the ten test texts of the corresponding author. The diagonal contains the correct classification. The lexically-based approach achieved 74% average accuracy. Approximately 65% of the average *identification error* (i.e., erroneously classified texts/total texts) corresponds to authors A01, A03, and A06 which have very short average text-length (see Table III).

Notice that the fifty most frequent words make up about 40% of all the tokens in the training corpus while one hundred most frequent words make up about 45%. In order to examine the degree to which the accuracy depends on the length of the set of the most frequent words, we performed the same experiment for different sets ranging from 10 to 100 most frequent words. The results are given in Figure 4. The best accuracy (77%) was achieved by using the sixty most frequent

Table VI. The confusion matrix of our approach (i.e., 22 style markers).

Actual	**Guess**										
	01	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	**A10**	**Error**
A01	**6**	0	0	3	0	0	0	0	0	1	0.4
A02	1	**9**	0	0	0	0	0	0	0	0	0.1
A03	2	0	**4**	0	1	0	0	2	1	0	0.6
A04	0	0	0	**10**	0	0	0	0	0	0	0.0
A05	0	0	0	0	**10**	0	0	0	0	0	0.0
A06	1	0	0	0	1	**7**	0	0	0	1	0.3
A07	0	0	0	0	0	0	**10**	0	0	0	0.0
A08	0	0	0	0	0	0	0	**10**	0	0	0.0
A09	0	0	2	0	0	0	0	0	**8**	0	0.2
A10	1	0	0	0	0	0	0	0	2	**7**	0.3
										Average	0.19

words. In general, the performance is not improved linearly by taking into account more words. According to our opinion, this is due to the training data overfitting of the classification model. Therefore, the more most frequent words taken into account (beyond a certain threshold), the less likely the achievement of reliable classification results in unseen cases.

4.3. PERFORMANCE

SCBD was used in order to analyze automatically both the training and test corpus and provide the vector of the 22 style markers for each text. In order to extract the classification models we performed discriminant analysis on the training corpus. The acquired models were, then tested on the test corpus. The results of that cross-validation procedure (i.e., the application of the classification procedure to unseen cases) are presented in the confusion matrix of Table VI. An average accuracy of 81% was achieved, which is 7% higher than that of the lexically-based approach. As in the case of this approach, the authors A01, A03, and A06 are responsible for approximately 65% of the average identification error.

We also performed a similar experiment combining our approach and the lexically-based one by using 72 style markers (i.e., the 50 most frequent word frequencies of occurrence plus our set of 22 style markers). Discriminant analysis was applied to the training corpus. The classification of the test corpus based on the models acquired by that training procedure is shown in Table VII. As can been seen this approach performs even better, i.e., it achieves an average accuracy of

Table VII. The confusion matrix of the combined approach (i.e., 72 style markers).

Actual	**Guess**										
	A01	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	**A10**	**Error**
A01	**6**	0	1	3	0	0	0	0	0	0	0.4
A02	0	**10**	0	0	0	0	0	0	0	0	0.0
A03	0	1	**6**	1	1	1	0	0	0	0	0.4
A04	0	0	0	**10**	0	0	0	0	0	0	0.0
A05	0	0	0	0	**10**	0	0	0	0	0	0.0
A06	0	0	0	1	0	**7**	0	0	2	0	0.3
A07	0	0	0	0	0	0	**10**	0	0	0	0.0
A08	0	0	0	0	0	0	1	**9**	0	0	0.1
A09	0	0	0	0	0	0	0	0	**10**	0	0.0
A10	0	0	1	0	0	0	0	0	0	**9**	0.1
										Average	0.13

87%, while the authors A01, A03, and A06 are responsible for approximately 85% of the average identification error.

These results show a strong dependency of the classification accuracy on the text-length. It seems that a text-length shorter than 1,000 words is not adequate for representing sufficiently the characteristics of the idiosyncratic style of an author by using either lexical measures, the presented set of style markers, or a combination of them.

4.4. TRAINING DATA SIZE

We conducted experiments with different sizes of the training data. In more detail, we trained our system using as training data subsets of the initial training corpus (i.e., 10 to 20 texts per author). Similar experiments were performed for both the lexically-based approach and the combination of the two approaches. The classification accuracy as a function of the training data size is presented in Figure 5.

The same training texts were used in all the three cases. Moreover, the test corpus was always the one used in the previously presented experiments (i.e., ten texts per author). In general, the accuracy was improved by increasing the training data. However, this improvement is not linear. Our approach presents the most stable performance since there are no significant differences between adjacent text measures. On the other hand, the lexically-based approach is quite unstable. For instance, using 15 texts per author the accuracy is practically the same as by using 10 texts per author. In general, our approach is more accurate than the lexical one

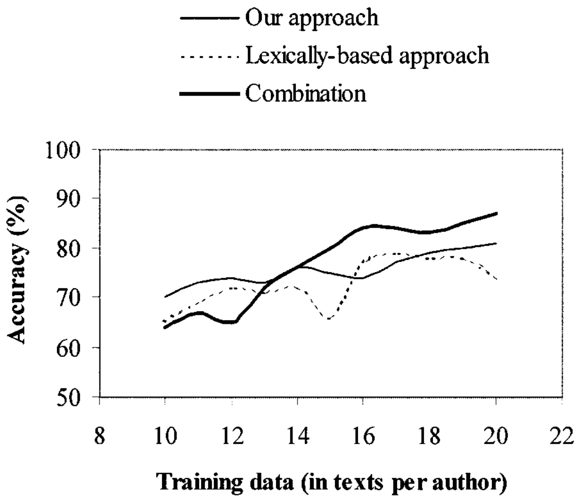


Figure 5. Classification accuracy for different sizes of training data.

(aside from two cases, i.e., 16 and 17 texts per author). The combined methodology is less accurate than the other two for training data smaller than 14 text per author. However, the results of the latter approach are quite satisfying when using more than 14 training texts per author.

Notice that Biber (1990, 1993) has shown that ten texts are adequate for representing the core linguistic features of a stylistic category. It has also to be underlined that in many cases there is only a limited number of texts available for training. As can been seen in Figure 5, our approach performs better than the other two using 10 texts per author as training corpus (i.e., 70% classification accuracy).

4.5. SIGNIFICANCE TEST

As aforementioned the proposed set of style markers is composed of three levels (i.e., token-level, phrase-level, and analysis-level). In order to illustrate the significance of each one of the proposed stylometric levels, the following experiment was conducted. We applied discriminant analysis to the entire training corpus (i.e., 20 texts per author) based on only one level per time. The obtained models were, then, used for classifying the test corpus. The results are shown in Figure 6. The classification accuracy achieved by the previous models (i.e., three-level approach, lexically-based approach, and combination of them) are also shown in that figure.

The most important stylometric level is the token-level since it managed to correctly classify 61 texts based on only 3 style markers. On the other hand, the phrase-level style markers managed to correctly classify 50 texts while the analysis-level ones identified correctly the authorship of 55 texts. It seems, therefore, that the analysis-level measures, which provide an alternative way of

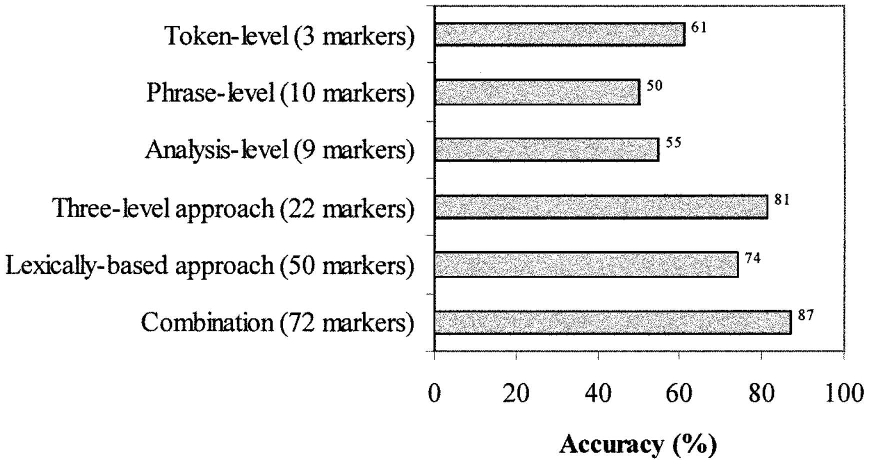


Figure 6. Classification accuracy of the tested models.

capturing the stylistic information, are more reliable than the measures related to the actual output of the SCBD (i.e., phrase-level markers).

In order to illustrate the disciminatory potential of any particular style marker, we performed analysis of variance (aka ANOVA). Specifically, ANOVA tests whether there are statistically significant differences among the authors with respect to the measured values of a particular marker. The results of the ANOVA tests are given in Table VIII. The F and r^2 values are indicators of importance. The greater the F value the more important the style marker. Moreover, r^2 measures the percentage of the variance among style marker values that can be predicted by knowing the author of the text.

As can been seen, the style markers M02, M03, M04, M07, M14, M17, M19, and M20 are the most significant as well as the best predictors of differences among the specific authors, since they have r^2 values greater than 50%. On the other hand, M08, M11, M12, M13, M21, and M22 are the less significant style markers, with r^2 values smaller than 20%. By excluding the latter style markers from the classification model (i.e., taking into account only the rest 16) an accuracy of 80% is achieved, i.e., slightly lower than taking all the proposed style markers into account. Hoewever, it has to be underlined that the presented ANOVA tests are valid only for that particular group of authors. Thus, a style marker that has been proved to be insignificant as regards a certain group of authors may be highly important considering a different group of authors.

Finally, the calculation of the average r^2 values for each stylometric level verifies the results of the Figure 6. Indeed, the average r^2 values of the token-level, phrase-level, and analysis-level style markers are 59.1%, 27.1%, and 41.7 respectively.

Table VIII. ANOVA tests for each style marker ($p < 0,0001$).

Style marker	**F**	**r^2(%)**
M01	26.5	45.2
M02	89.8	73.6
M03	45.2	58.4
M04	48.5	60.0
M05	14.4	30.8
M06	18.6	36.5
M07	35.9	52.7
M08	7.2	18.3
M09	9.5	22.3
M10	12.6	28.2
M11	2.3	6.8
M12	4.3	11.7
M13	3.3	9.3
M14	47.2	59.5
M15	25.6	44.3
M16	16.3	33.6
M17	34.5	51.7
M18	30.5	48.6
M19	33.9	51.3
M20	40.0	55.4
M21	5.9	15.5
M22	6.1	15.6

5. Discussion

We presented an approach to authorship attribution dealing with unrestricted Modern Greek texts. In contrast to other authorship attribution studies, we excluded any distributional lexical measure. Instead, a set of style markers was adapted to the automatic analysis of text by the SCBD tool. Any measure relevant to this analysis that could capture stylistic information was taken into account.

So far, the recent advances in NLP did not influence the authorship attribution studies since computers are used only for providing simple counts very fast. Real NLP is avoided despite the fact that various tools providing quite accurate results are nowadays available, at least at the syntactic level, covering a wide variety of natural languages. Just to name a few of them, Dermatas and Kokkinakis (1995) describe several accurate stochastic part-of-speech taggers for seven European languages. A language-independent trainable part-of-speech tagger proposed by Brill (1995) has been incorporated into many applications. Moreover, the systems

SATZ (Palmer and Hearst, 1997) and SuperTagger (Srinivas and Joshi, 1999) offer reliable solutions for detecting sentence boundaries and performing partial parsing, respectively. In this paper our goal was to show how existing NLP tools could be used for providing stylistic information. Notice that SCBD was not designed specifically to be used for attributing authorship. Towards this end, we introduced the notion of analysis-level measures, i.e., measures relevant to the particular method used by the NLP tool in order to analyze the text. The more carefully selected analysis-level measures are defined, the more useful stylistic information is extracted.

Among the three proposed stylometric levels, the token-level measures have been proved to be the most reliable discriminating factor. The calculation of these measures using SCBD is more accurate than the corresponding calculation of the phrase-level measures. Moreover, the analysis-level measures are more reliable than the phrase-level ones and play an important role in capturing the stylistic characteristics of the author.

Our methodology is fully-automated requiring no manual text pre-processing. However, we believe that the development of automatic text sampling tools which are able to detect the most representative parts of the text (i.e., the parts where the stylistic properties of the author is more likely to distinguish) can considerably enhance the performance. The text-length is a very crucial factor. Particularly, it seems that texts with less than 1,000 words are less likely to be correctly classified. On the other hand, such a lower bound cannot be applied in many cases. For example, half of the texts that compose the corpus used in this study do not fulfill this restriction.

All the presented experiments were based on unrestricted text downloaded from the Internet and a randomly-chosen group of authors. The proposed approach achieved higher accuracy than the lexically-based methodology introduced by Burrows (1987, 1992) that is based on the frequencies of occurrence of the fifty most frequent words. Moreover, our technique seems to be more robust for limited size of training data. However, the combination of these two approaches is the most accurate solution and can be used for reliable text categorization in terms of authorship. The presented methodology can also be used in *author verification* tasks, i.e., the verification of the hypothesis whether or not a given person is the author of the text under study (Stamatatos et al., 1999a).

The statistical technique of discriminant analysis was used as disambiguation procedure. The classification is very fast since it is based on the calculation of simple linear functions. Moreover, the training procedure does not require excessive computational and time cost and can be easily incorporated into a real-time application. However, we believe that a more complicated discrimination-classification technique (e.g., neural networks) could be applied to this problem with remarkable results.

Much else remains to be done as regards the explanation of the differences and the similarities between the authors. The presented methodology lacks any underlying linguistic theory since it is based on statistical measures. Thus, the interpretation of the statistical data (e.g., loadings of discriminant functions) would inevitably require subjective assumptions. Moreover, in case of texts written by more than one author, techniques that explore style variation within a single text have to be developed. We believe that the proposed approach can be used towards this end.

Note

[1] http://tovima.dolnet.gr

References

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

Biber, D. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variations." *Literary and Linguistic Computing*, 5 (1990), 257–269.

Biber, D. "Representativeness in Corpus Design." *Literary and Linguistic Computing*, 8 (1993), 1–15.

Brill E. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics*, 21(4) (1995), 543–565.

Brinegar, C. "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship." *Journal of the American Statistical Association*, 58 (1963), 85–96.

Burrows, J. "Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing*, 2(2) (1987), 61–70.

Burrows, J. "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information." *Literary and Linguistic Computing*, 7(2) (1992), 91–109.

Dermatas E. and G. Kokkinakis "Automatic Stochastic Tagging of Natural Language Texts." *Computational Linguistics*, 21(2) (1995), 137–164.

Eisenbeis, R. and R. Avery. *Discriminant Analysis and Classification Procedures: Theory and Applications.* Lexington, Mass.: D.C. Health and Co. 1972.

Forsyth, R. and D. Holmes. "Feature-Finding for Text Classification." *Literary and Linguistic Computing*, 11(4) (1996),163–174.

Fucks W. "On the Mathematical Analysis of Style." *Biometrica*, 39 (1952), 122–129.

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

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

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

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

Karlgren, J. "Stylistic Experiments in Information Retrieval." In *Natural Language Information Retrieval*. Ed. T. Strzalkowski, Kluwer Academic Publishers, 1999, pp. 147–166.

Morton A. "The Authorship of Greek Prose." *Journal of the Royal Statistical Society Series A*, 128 (1965), 169–233.

Mosteller, F. and D. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. MA: Addison-Wesley, Reading, 1984.

Oakman, R. *Computer Methods for Literary Research*. Columbia: University of South Carolina Press, 1980.

Palmer, D. and M. Hearst. “Adaptive Multilingual Sentence Boundary Disambiguation.” *Computational Linguistics*, 23(2) (1997), 241–267.

Sichel, H. “Word Frequency Distributions and Type-Token Characteristics.” *Mathematical Scientist*, 11 (1986), 45–72.

Srinivas, B and A. Joshi. “Supertagging: An Approach to Almost Parsing.” *Computational Linguistics*, 25(2) (1999), 237–265.

Stamatatos, E., N. Fakotakis and G. Kokkinakis. “Automatic Authorship Attribution.” In *Proc. of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL’99)*, 1999a, pp. 158–164.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. “Automatic Extraction of Rules for Sentence Boundary Disambiguation.” In *Proc. of the Workshop on Machine Learning in Human Language Technology, ECCAI Advanced Course on Artificial Intelligence (ACAI-99)*, 1999b, pp. 88–82.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. “A Practical Chunker for Unrestricted Text.” In *Proc. of the Second Int. Conf. on Natural Language Processing*, 2000.

Strzalkowski, T. “Robust Text Processing in Automated Information Retrieval.” In *Proc. of the 4^{th} Conf. On Applied Natural Language Processing*, 1994, pp. 168–173.

Tallentire D. “Towards an Archive of Lexical Norms: A Proposal.” In *The Computer and Literary Studies*. Eds. A. Aitken, R. Bailey, and N Hamilton-Smith, 1973, Edinburgh University Press.

Tweedie, F. and Baayen, R. “How Variable may a Constant be? Measures of Lexical Richness in Perspective.” *Computers and the Humanities*, 32(5) (1998), 323–352.

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

Computers and the Humanities **37:** 151–178, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Another Perspective on Vocabulary Richness

DAVID L. HOOVER
New York University, 19 University Place, New York, NY 10003, USA
E-mail: david.hoover@nyu.edu

Abstract. This article examines the usefulness of vocabulary richness for authorship attribution and tests the assumption that appropriate measures of vocabulary richness can capture an author's distinctive style or identity. After briefly discussing perceived and actual vocabulary richness, I show that doubling and combining texts affects some measures in computationally predictable but conceptually surprising ways. I discuss some theoretical and empirical problems with some measures and develop simple methods to test how well vocabulary richness distinguishes texts by different authors. These methods show that vocabulary richness is ineffective for large groups of texts because of the extreme variability within and among them. I conclude that vocabulary richness is of marginal value in stylistic and authorship studies because the basic assumption that it constitutes a wordprint for authors is false.

Key words: authorship attribution, lexical statistics, stylistics, vocabulary richness

1. Introduction

There has been considerable interest in recent years in the application of statistical techniques to literary texts, particularly in the area of authorship attribution. Although my own interest is not primarily in authorship, but rather in stylistic analysis, authorship attribution and stylistics share an interest in the size, coherence, and distribution of the vocabularies of texts and authors. Here I will focus primarily on measures of vocabulary richness and their potential usefulness in both spheres. If measures of vocabulary richness can reliably attribute texts to their authors, they may be of use in characterizing the styles of those authors; conversely, if they cannot do so, they are unlikely to be of any significant value in studies of style.[1]

Authors clearly differ in the sizes and structures of their vocabularies – some have large vocabularies and use many relatively infrequent words and others have smaller vocabularies and use many more frequent words. This has led to the reasonable assumption, often unstated, that vocabulary richness or concentration provides a kind of authorial wordprint that can distinguish authors from each other, an assumption made more reasonable by the unlikelihood that authors regularly control the richness of their vocabularies in a deliberate or conscious way. Word use that is not consciously controlled is likely to be automatic, habitual, and consistent.

The most obvious and basic measures of vocabulary richness are the number of different word types that a text contains and the closely related type/token ratio. Unfortunately, these measures depend, to a great extent, on the length of the text. Other simple measures of vocabulary richness are the number of *hapax legomena* (words occurring exactly once) and the number of *dis legomena* (words occurring exactly twice). Various mathematical transformations of the vocabulary size or type/token ratio, some of which will be discussed further below, have also been proposed. Still other measures of vocabulary richness reflect the randomness of a text; for example, by considering the probability of randomly drawing two identical tokens from it.[2] Whatever the methods of calculation, however, all of the proposed measures share the basic assumption that authors differ systematically in the richness of their vocabularies, and that the appropriate measure can capture something distinctive about the style of an author.

I test this basic assumption below. First I examine the relationship between perceived and actual vocabulary richness. Then, by examining the effects of doubling and combining texts, I demonstrate that some measures of vocabulary richness react in ways that are computationally predictable but seem peculiar and surprising from a common sense view of vocabulary richness and authorial style. After discussing some theoretical and empirical problems with some measures, I develop simple methods of testing their effectiveness in distinguishing texts by different authors and clustering texts by the same author. These simple methods allow for a broader examination of some relatively large groups of texts that shows that measures of vocabulary richness are very ineffective for such groups of texts because of the extreme variability in vocabulary richness both within and among texts. I conclude by arguing that the basic assumption that vocabulary richness constitutes a wordprint that can distinguish authors from each other is false, that measures of vocabulary richness are much less reliable and much less useful in distinguishing authors from each other than has been thought, and that they can be of only marginal value in stylistic and authorship studies.

2. Perceptions of Vocabulary Richness

Although a single measure of vocabulary richness that can characterize an author or text is an attractive idea, readers' perceptions about vocabulary richness are not necessarily accurate. For example, consider the following twelve texts: Faulkner, *Light in August*; James, *The Ambassadors*; Wilde, *The Picture of Dorian Gray*; Doyle, *The Return of Sherlock Holmes*; Stoker, *Dracula*; Woolf, *To the Lighthouse*; Chopin, *The Awakening*; Cather, *My Antonia*; Wells, *The War of the Worlds*; Kipling, *Kim*; London, *The Seawolf*; Lewis, *Main Street*. Readers will have different perceptions of the vocabulary richness of these texts, but very few will realize that they are listed in order of *increasing* vocabulary – here represented by the number of different types in the first 50,000 words, as shown in Figure 1.

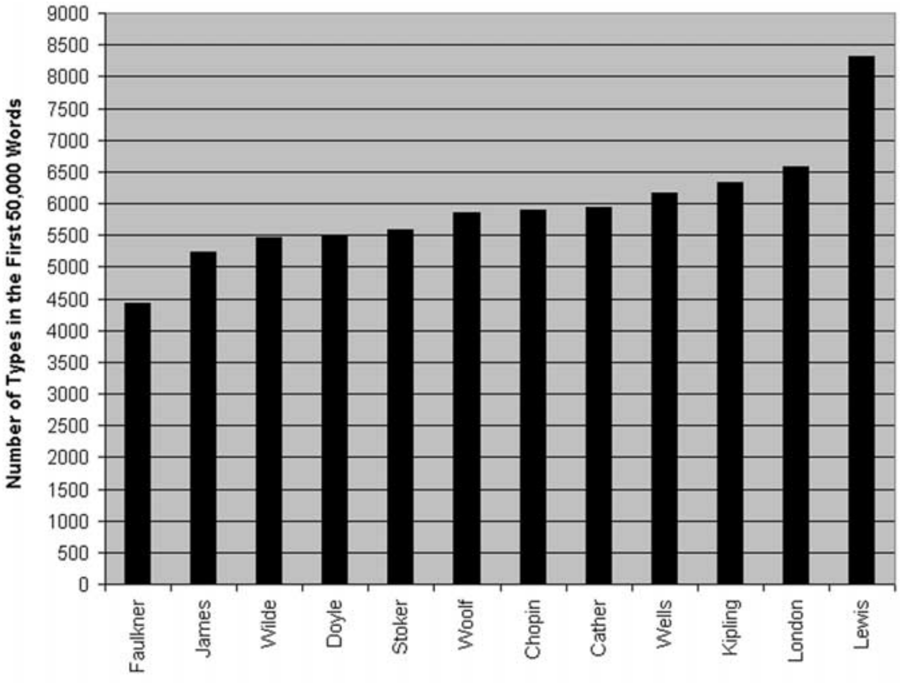


Figure 1. Vocabulary richness in twelve authors

In spite of a common perception that Faulkner and James have large vocabularies (perhaps because they seem "difficult"), the number of types in the first 50,000-word section is less than 4,500 for *Light in August,* less than 5,500 for *The Ambassadors*, but more than 8,300 for *Main Street*.[3] However, the mere failure of texts to have the sizes of vocabularies that readers might predict does not prove that vocabulary size does not reasonably characterize those texts.

3. Text Doubling and Combining and Vocabulary Richness

More problematic than the inaccuracy of readers' judgments about vocabulary richness are some peculiar effects that arise from the methods of calculation of some of the measures. Thoiron (1986) examines Simpson's Diversity, the probability of drawing two identical words from a text, and entropy, a measure of the disorder or randomness of a text.[4] Both measures, as Thoiron points out, are transparently related to intuitive concepts of vocabulary richness: the less probable it is that a pair of identical words will be drawn from a text, the richer the vocabulary; the more random or disordered a text, the richer the vocabulary (pp. 198–199). In spite of their clear conceptual interpretation, however, both measures, Thoiron argues, are flawed. He first shows that, if the total vocabulary of a text is kept constant, adding additional tokens of words that are already frequent in it causes Diversity to increase (increases the probability that any two words selected at random will be the same), marking the text as less rich, as one might expect. (This measure seems oddly named, since the higher the Diversity, the *less* diverse the text.) However, adding more tokens of infrequent words (turning *hapax legomena* into *dis legomena*, for example) makes the text richer. And this second effect, Thoiron quite reasonably argues, is counterintuitive, because it means that "a text T′, which is made up of a text T to which have been added some of its own constitutive elements, is richer than T" (p. 199).

Thoiron then shows that entropy also fails to react to textual modification as one might expect. When he successively adds short sections of a text to itself, contrary to expectation, the increasingly repetitive text does not show a gradual decrease in entropy, but rather "a more-or-less sinusoidal movement" (p. 200). Indeed, rather surprisingly, adding a text to itself (any number of times) has no effect at all on entropy, which is supposed to reflect vocabulary richness by measuring the disorder or randomness of the text. Thoiron rightly finds this troubling, asking, "Can one not consider as lexically poorer a text T′ which is merely made up of the repetition (twice or more) of every single item occurring in T?" (p. 200).

One might argue that a measure that attempts to capture an authorial wordprint should not react to text doubling. The added text is manifestly in the author's style, after all, and one would hope that adding more text that is *statistically* identical would not affect the vocabulary richness of a text. From a practical, common sense point of view, however, the presence of repeated passages that are identical in *content* in a literary text would surely be unusual, and would surely be seen as affecting its style. Furthermore, note that doubling the first half of a novel produces a text that is radically different from the original whole novel in vocabulary richness: the doubled text displays a much smaller vocabulary than the original novel because the second half of the novel adds a large number of new types.

Thoiron's experiments show that neither diversity nor entropy responds as one might intuitively expect to textual modifications that make the texts more repetitive, but a word of caution seems in order. His first experiment alters the ratio of *hapax legomena* to *dis legomena*, a measure that has itself been proposed as a stylistic marker. His second experiment produces a text with two identical halves. This eliminates all *hapax legomena*, thus producing a text so statistically bizarre as to be unprecedented: normally, roughly half of the types in a novel are *hapax legomena*. This possibility is so remote, in fact, that TACT (Version 2.1, Centre for Computing in the Humanities, University of Toronto) not unreasonably gives false statistics for such texts, reporting figures for *hapax legomena* and *dis legomena* that are actually figures for *dis legomena* and words occurring four times.

Taking Thoiron's experiment a step farther – by comparing the results of doubling a single text with the results of combining two texts – is instructive. For this experiment, I have selected 50,000-word sections of Woolf's *To the Lighthouse*, Lawrence's *Sons and Lovers*, James's *The Ambassadors* and *The Europeans*, and Lewis's *Main Street*, analyzing and combining them as displayed in Table I.

Table I. The effects of text-doubling and text-combining on measures of vocabulary richness

Text	Types	Tokens	Hapax Legom.	Dis Legom.	Herdan's V_m	Yule's K	Repeat	Skewness	Kurtosis	Word Length
Amb2	4687	50000	2314	767	0.0899	82.7345	26.151	16.2207	326.4262	4.2388
Amb2 doubled	4687	100000	0	2314	0.0899	82.8345	26.151	16.2207	326.4262	4.2388
Son1	5859	50000	3017	955	0.0943	90.3703	17.940	25.4412	883.8854	4.1749
Son1 doubled	5859	100000	0	3017	0.0943	90.4703	17.940	25.4412	883.8854	4.1749
Lih1	5851	50000	3044	1009	0.0954	92.5514	20.678	21.6911	609.3101	4.2897
Lih1 doubled	5851	100000	0	3044	0.0954	92.6514	20.678	21.6911	609.3101	4.2897
Son1+Lih1	9174	100000	4522	1490	0.0939	89.2401	19.212	29.5434	1168.6170	4.2323
Main1+Main2	12377	100000	6511	1993	0.0912	83.9095	19.135	38.3014	1891.0450	4.4411
Amb3	4663	50000	2315	747	0.0924	87.2501	24.178	16.7205	351.5496	4.1960
Eur1	4942	50000	2494	791	0.0905	83.6893	25.840	16.9638	358.2319	4.3279
Amb2+Amb3	6660	100000	3042	1104	0.0914	84.8767	25.126	19.6220	481.7780	4.2174
Amb2+Eur1	7322	100000	3485	1175	0.0893	80.9250	25.994	20.5814	528.1602	4.2834

Thoiron's point about the identity of entropy for texts and their doubles is actually more general than he indicates. As the first six rows of Table I show, TACT's figures for Herdan's V_m, word length, skewness, Kurtosis, and the repeat rate of the most frequent word (almost invariably *the*), all of which have been used as markers of style, are identical for texts and their doubles, and those for Yule's *K* are nearly identical.[5] It is surprising that measures intended to capture aspects of authorial style are completely insensitive to a transformation that intuitively seems to alter the style of the text. Other measures, such as Zipf's *Z*, the Carroll *TTR*, and Sichel's *S*, are altered, sometimes radically, by these transformations.[6]

Row seven of Table I shows that combining two texts by different authors but with very similar numbers of types, *hapax legomena*, and *dis legomena* produces very different results: for obvious reasons, the number of types in the combined text is much greater than for either of the doubled texts. The number of *hapax legomena* in the combined text is also much greater than the number of *dis legomena* in the doubled texts (the *hapax legomena* of the original texts become *dis legomena* in the doubled texts). Although the figures for a combination of *Sons and Lovers* and *To the Lighthouse* are much higher than for either novel doubled, the figures for the first two sections of *Main Street* show that a single novel with an exceptionally large vocabulary can produce even higher figures. This suggests a negative answer to a question that Holmes and Forsyth consider in their discussion of the Federalist Papers: "whether collaborative texts are always richer in vocabulary than texts from separate contributors" (1995, p. 117). If a text in an authorship attribution study has a substantially richer vocabulary than is found in the texts written by any of the claimants, joint authorship is clearly a strong possibility, but it assumes that authors are consistent in vocabulary richness. As we will see, however, this assumption cannot safely be made.

The last two rows of Table I show the results of combining sections of novels by one author that are similar in vocabulary richness. The vocabulary of a text formed by combining two sections is always smaller than the sum of the vocabularies of the sections because many words occur in both sections. The vocabulary of combined sections of *The Ambassadors*, for example, is only about 71% of the summed vocabularies of the sections, and the vocabulary of combined sections of *The Ambassadors* and *The Europeans* is about 76% of the summed vocabularies. The figure for combined sections of *Main Street* is about 75% and for combined sections of *Sons and Lovers* and *To the Lighthouse* is about 78%. A quick check of eighteen examples of combined sections of a single novel range from about 71% to 76% of the summed vocabularies, and ten examples of combined sections of novels by different authors range from about 76% to 81%. It is hardly surprising that combined sections of novels by different authors retain more of the summed vocabulary than do combined sections of the same novel. What seems more surprising is that it makes so little difference whether or not the two sections that are combined are parts of the same larger text, and whether or not they were written by the same author. One could choose novels in such a way as to maximize or minimize the combined vocabulary, of course, but the vocabularies of two texts (or two sections of the same text) by a single author are clearly very different.[7] This result may seem rather counterintuitive, but it is actually predictable from the large proportion of *hapax legomena* in texts. That is, since about half the types in any of these sections appear only once, combining any two sections will greatly increase the total vocabulary, regardless of the source of the sections.

4. Theoretical and Empirical Problems with Some Measures of Vocabulary Richness

In "How Variable May a Constant be? Measures of Lexical Richness in Perspective," Fiona J. Tweedie and R. Harald Baayen (1998) examine proposed measures of vocabulary richness. Since Yule's ground-breaking study in 1944, many constants have been proposed in an attempt to find one that is not affected by text length.[8] As Tweedie and Baayen note, it is easy to see that, the longer the text, the more slowly the total vocabulary grows, and hence the less rich the vocabulary becomes. The logical limit is reached when the author has used every word in his or her vocabulary. The notion of the "total vocabulary" of an author is more problematic than might appear, however, for authors normally learn new words during the writing of a novel, as Holmes (1994) notes, citing Brainerd (1988). Authors also forget words, or stop using them. In any case, it is clear that the rate of vocabulary growth normally slows as a text's length increases. Tweedie and Baayen present a thorough examination of the theoretical and empirical constancy of the various "constants," showing that some are not even theoretically constant, and that others are not constant when tested empirically (pp. 323–334).

Tweedie and Baayen also point out that the discourse-structure of texts violates the randomness assumption of the "urn" model underlying many discussions of vocabulary richness (pp. 333–334). Baayen (1993) states this problem succinctly and clearly: "Word types are re-used with more than chance frequency in texts. Once a particular topic is broached, the vocabulary items related to that topic have a substantially raised probability of being re-used" (pp. 360–361). Elsewhere, he has shown that the main source of divergence between the predicted and actual vocabulary size of a text is the use of words more frequently within some sections of a text, making those sections internally cohesive and also cohesive with each other (1996, pp. 458–460).

To examine the effects of discourse structure on vocabulary richness, Tweedie and Baayen perform sophisticated randomization experiments that uncover the behavior of the constants throughout texts. Their techniques allow them to plot trajectories for the constants (pp. 334–340), in a way that is reminiscent of Baayen's demonstration of different developmental profiles for the divergence from estimates of vocabulary size (1996, pp. 465–466). Tweedie and Baayen then use a partial randomization technique that allows the discourse-structure of the text to be reflected in a developmental profile of each constant throughout sixteen texts by eight authors, a technique that leads to "clearer differences in the vocabulary structure of texts" (p. 344). Their conclusion is that, although selected vocabulary richness constants capture some "aspects of authorial structure" and allow many of their sixteen texts to be grouped properly, they do not correctly group all of the texts by each author nor correctly separate all texts by different authors (pp. 345–348). Finally, they show that principal component analysis of the 100 most frequent function words does a better job of grouping and separating texts. Their concluding discussion is especially valuable in emphasizing the fact that two basic kinds of statistics provide a substantial amount of information about authorial style: measures such as K, D, and V_m reflect the rate at which words are repeated and constitute inverse measures of vocabulary richness; measures such as Z, b, and c are based on "probabilistic models for word frequency distributions" (p. 350) and measure vocabulary richness more directly.[9]

They argue that the use of many constants is not necessary, and that just two measures, Yule's K and Zipf's Z, capture a surprising amount of authorial style and "are two useful indicators of style," although they "should be used with care (given their within-text variability)" (p. 350). As we will see, however, vocabulary richness is a much less useful and a much more dangerous indicator of authorship and marker of style and than they suggest.

5. Simpler Techniques for Examining Vocabulary Richness

The statistical methods Tweedie and Baayen bring to their task are impressive. However, because the number of useful aspects of vocabulary structure is limited, and because the trajectories of Z and K alone are more accurate than the trajectories of all seventeen constants, duplicating their results with simpler and more accessible techniques should be possible. These simpler techniques facilitate the examination of larger sets of texts and a closer consideration of intratextual and intertextual variation.

The texts analyzed by Tweedie and Baayen are the following:

Baum, L. F.	*The Wonderful Wizard of Oz; The Marvelous Wizard of Oz*[10]
Brontë, E.	*Wuthering Heights*
Carroll, L.	*Alice's Adventures in Wonderland; Through the Looking-Glass and What Alice Found There*
Doyle, A. C.	*The Sign of Four; The Hound of the Baskervilles; The Valley of Fear*
James, H.	*Confidence; The Europeans*
St. Luke	*The Gospel According to St. Luke (KJV); The Acts of the Apostles (KJV)*
London, J.	*The Sea Wolf; The Call of the Wild*
Wells, H. G.	*The War of the Worlds; The Invisible Man*

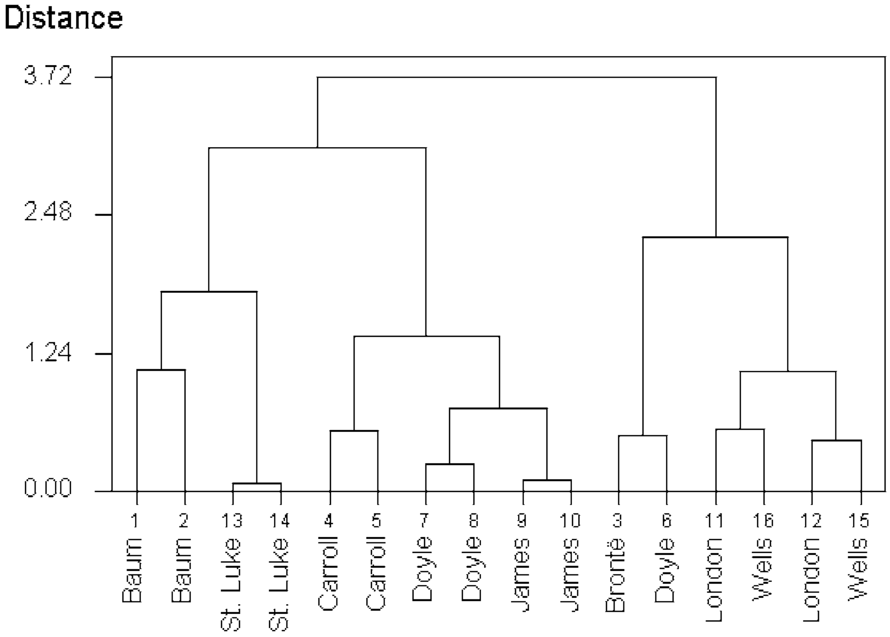


Figure 2. Cluster analysis of sixteen texts, based on the values of *Z(N)* and *K(N)* for the complete texts.

I downloaded the same texts that Tweedie and Baayen use (from different sources) and analyzed them with TACT, which produces statistics for types, *hapax legomena, dis legomena*, Yule's *K*, Herdan's V_m (a revision of *K*), the frequency of the most frequent word, and the repeat rate of the most frequent word (Lancashire, 1996, pp. 108–109). (Yule's *K* and Herdan's V_m should not be confused with Herdan's *C* or Rubet's *k*, which have very different derivations.[11]) First, consider Figure 2, which attempts to duplicate their results.[12] These results are quite similar to those of Tweedie and Baayen (348: Figure 16), and even slightly more accurate than their result for final values of *Z* and *K*, perhaps because I have standardized the variables to minimize the effect of the difference in size between *Z* and *K*. (The same analysis performed without standardized variables groups the same texts as their analysis.)

My attempt to duplicate this result with simpler techniques begins with an analysis of the first 24,000 words of each of their texts, roughly the length of the shortest text. Trimming the texts to equal size allows the number of types to be used as a direct measure of vocabulary richness and lays the groundwork for an examination of intratextual variability. Figure 3 presents a cluster analysis of the first 24,000-word section of each of the sixteen texts that tests the separation of texts and authors.

Although many of the texts are much longer than the excerpts examined in Figure 3, the number of types and the frequency of the most frequent word in initial

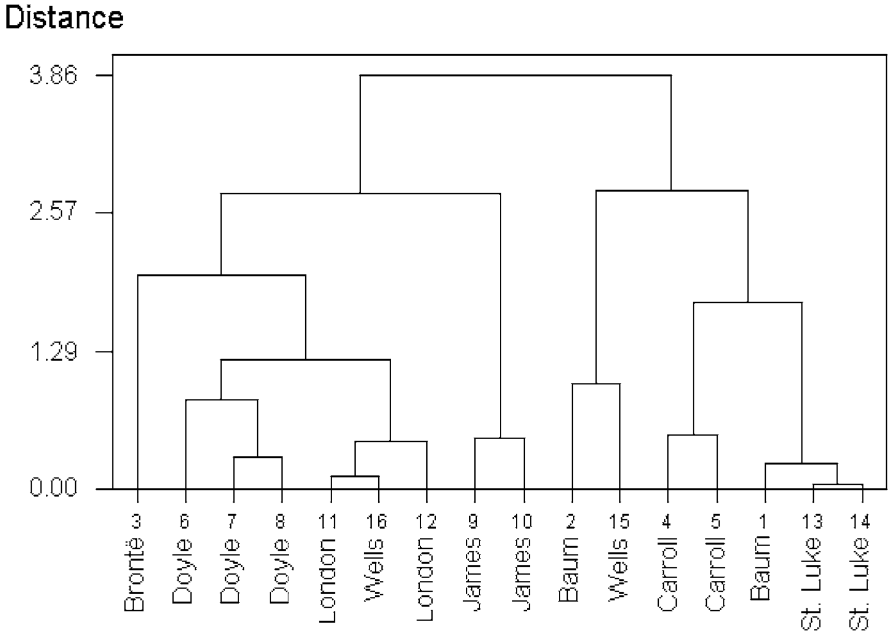


Figure 3. Cluster analysis of the first 24,000 words of sixteen texts, based on word types and the frequency of the most frequent word.

sections correctly cluster all sections of texts by Brontë, Carroll, Doyle, James, and St. Luke, and the sections by London are very close neighbors. This is a better result than Tweedie and Baayen achieve using the final values for all seventeen constants for each whole text, the trajectories of all seventeen consonants, or the final values of *Z* and *K*, and is about as good as their best results based on vocabulary richness, which use full trajectories of *Z* and *K* (348: Figure 16).[13]

Other vocabulary richness measures produced by TACT can be combined in various ways to test their effectiveness. Herdan's V_m and the frequency of the most frequent word, for example, produce groupings that are about as accurate as the one in Figure 3. The results for Yule's *K* and the frequency of the most frequent word are not as good, but adding the number of *dis legomena* to Herdan's V_m and the frequency of the most frequent word produces very good results, shown in Figure 4, in which the texts by Brontë, Carroll, Doyle, James, London, and St. Luke all cluster correctly, a result as accurate as any that Tweedie and Baayen produce, including that based on principal components analysis (p. 347).

Using equal-sized texts allows for the duplication of the results that Tweedie and Baayen achieve without requiring the calculation of seventeen constants. We can now examine what happens when the eight texts are cut into as many 24,000-word sections as possible and all of the sections are compared. If vocabulary richness measures truly capture authorial style or identity, the sections of single texts should cluster with each other and separate clearly from other texts even more strongly

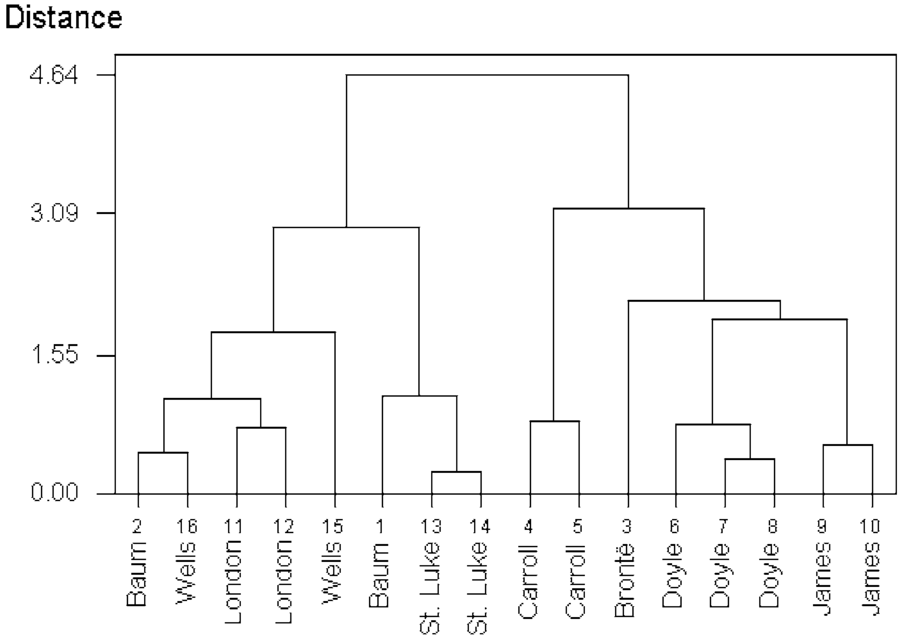


Figure 4. Cluster analysis of the first 24,000 words of sixteen texts, based on the frequency of the most frequent word, Herdan's V_m, and the number of *Dis Legomena*.

than do different texts by the same author. Figure 5 shows that the frequency of the most frequent word, Herdan's V_m, and the number of *dis legomena* correctly cluster all sections by Brontë, Doyle, Carroll, James, and the sections by St. Luke are nearest neighbors, both sections of Wells's *The War of the Worlds* cluster together, and Baum's *The Wonderful Wizard of Oz* forms its own cluster.[14] These results provide what initially seems to be rather striking support for the notion that vocabulary richness may be a marker of authorial style.[15]

If this kind of analysis proved to work as well on whole texts using the statistics produced by TACT as it does on equal-sized sections, it would be much simpler and more accessible than analyses that require tracing the trajectories of constants throughout texts. To test this possibility, I have analyzed all sixteen of the complete texts in TACT and performed the cluster analysis shown in Figure 6.

Note that the texts by Baum, Brontë, James, and St. Luke cluster correctly, and the texts by Doyle are close neighbors. (Yule's characteristic and the repeat rate of the most frequent word are very similar and only slightly less accurate.) These results are as accurate as those reported by Tweedie and Baayen for final values or full trajectories of all seventeen constants, and for final values of Z and K, and are only a little less accurate than the results using full trajectories for Z and K (p. 348).

These results suggest that vocabulary richness might be of significant use in studies of style and authorship attribution. One benefit (and temptation) of statistical programs, however, is that the discriminative power of any of the variables

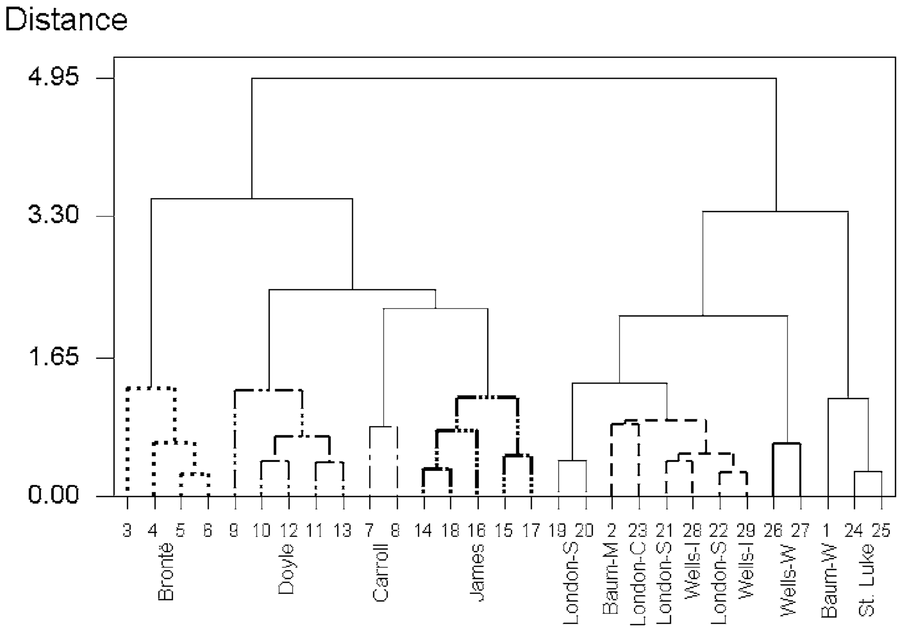


Figure 5. Cluster analysis of the twenty-nine 24,000-word sections of sixteen texts, based on the frequency of the most frequent word, Herdan's V_m, and the number of *Dis Legomena*.

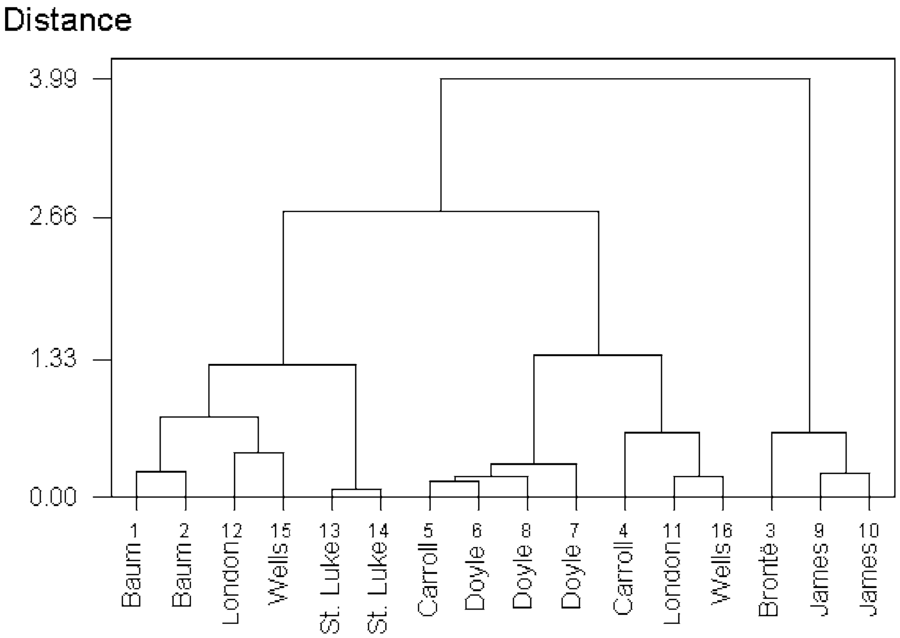


Figure 6. Cluster analysis of sixteen complete texts, based on Herdan's V_m and the repeat rate of the most frequent word.

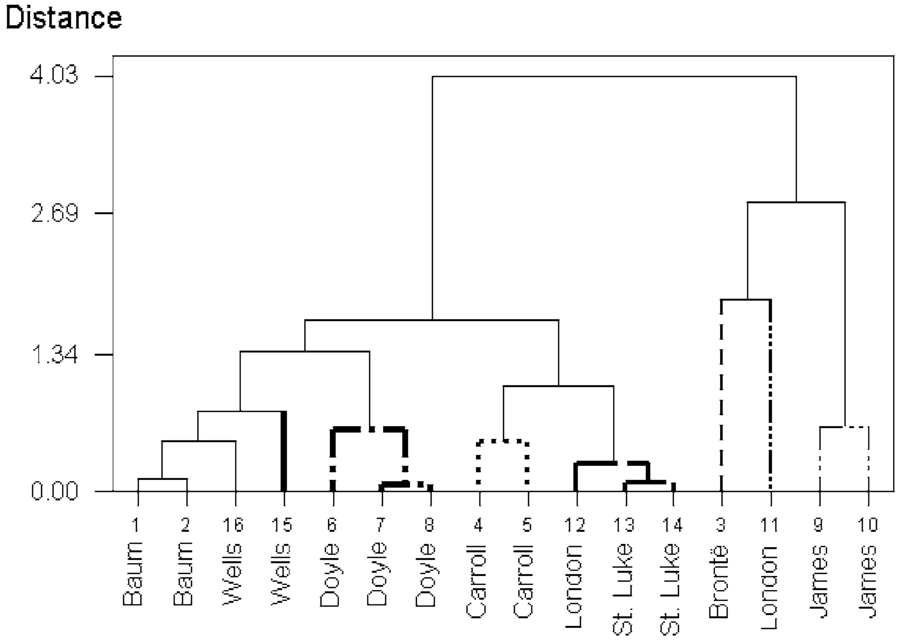


Figure 7. Cluster analysis of sixteen complete texts, based on word tokens and the repeat rate of the most frequent word.

present in the analysis can be tested. With the same texts clustering under so many different circumstances, it seems prudent to test the discriminative power of some less compelling variables. An analysis based on Herdan's V_m for initial *letters* and the repeat rate of the most frequent initial *letter* correctly clusters the texts by Brontë, Doyle, James, and St. Luke. Such variables may sometimes have legitimate discriminative value, but another possibility is that the texts being analyzed are so different that almost any characteristic will differentiate them from the other texts in the analysis.

Figure 7 shows another cluster analysis, in which the lengths of the texts (word tokens) and the repeat rate of the most frequent word are very effective in clustering the texts, failing only for London if interpreted as favorably as possible. This analysis allows the mere lengths of the texts to act as one of the variables, even though correcting for text length was the main reason for the creation of vocabulary richness constants in the first place. Yet it clusters the texts more accurately than Tweedie and Baayen's best results involving vocabulary richness. In fact, the results shown in Figure 7 are both very similar to and about as good as the results they achieve using principal component analysis of the 100 most frequent words of the texts (p. 347). Several other collections of measures produce similar results, and most of them include the repeat rate of the most frequent word. This suggests that the repeat rate is the most effective single measure among those tested here, but it is important to note that all of the most effective groups of measures include ones that, like the number of tokens, vary with text length. Indeed, the lengths of these fourteen books alone are fairly distinctive: the easiest way to tell a text by Brontë from one by Baum is by the thickness of the book. Even though it is true that some authors tend to write longer books than others, however, text length cannot be taken seriously as a general indication of authorship: even among the texts analyzed here, London's *The Sea Wolf* is more than three times as long as *The Call of the Wild.*

6. Vocabulary Richness Measures and Larger groups of Texts

Before drawing any rash conclusions from these strange results, it seems wiser to add some additional texts to the mix. I have added the thirty novels of my Novel Corpus because it is easily available and its texts have been extensively checked and analyzed (Hoover, 1999, pp. x–xii). Furthermore, this corpus is less diverse than the texts chosen by Tweedie and Baayen, and contains additional texts by James and Doyle, allowing further tests of correct clustering. The texts are as follows:

American texts

Winesburg, Ohio (Anderson, 1996) [1919]
My Antonia (Cather, 1996) [1918]
The House Behind the Cedars (Chestnutt, 1996) [1900]
The Awakening (Chopin, 1996) [1899]
The Red Badge of Courage (Crane, 1996b) [1895]
Sister Carrie (Dreiser, 1996) [1900]
Light in August (Faulkner, 1994) [1932]
The Damnation of Theron Ware (Frederic, 1996) [1896]
The Ambassadors (James, 1996) [1909]
Main Street (Lewis, 1996) [1920]
The Sea Wolf (London, 1996) [1904]
McTeague (Norris, 1996) [1899]
The Jungle (Sinclair, 1996) [1906]
The Tragedy of Pudd'nhead Wilson (Twain, 1996) [1894]
The Age of Innocence (Wharton, 1996) [1920]

British texts

Lord Jim (Conrad, 1996) [1900]
The Return of Sherlock Holmes (Doyle, 1996) [1901]
The Good Soldier (Ford, 1996) [1915]
Howards End (Forster, 1996) [1910]
Jude the Obscure (Hardy, 1996) [1896]
A Portrait of the Artist as a Young Man (Joyce, 1996) [1916]
Kim (Kipling, 1996) [1901]

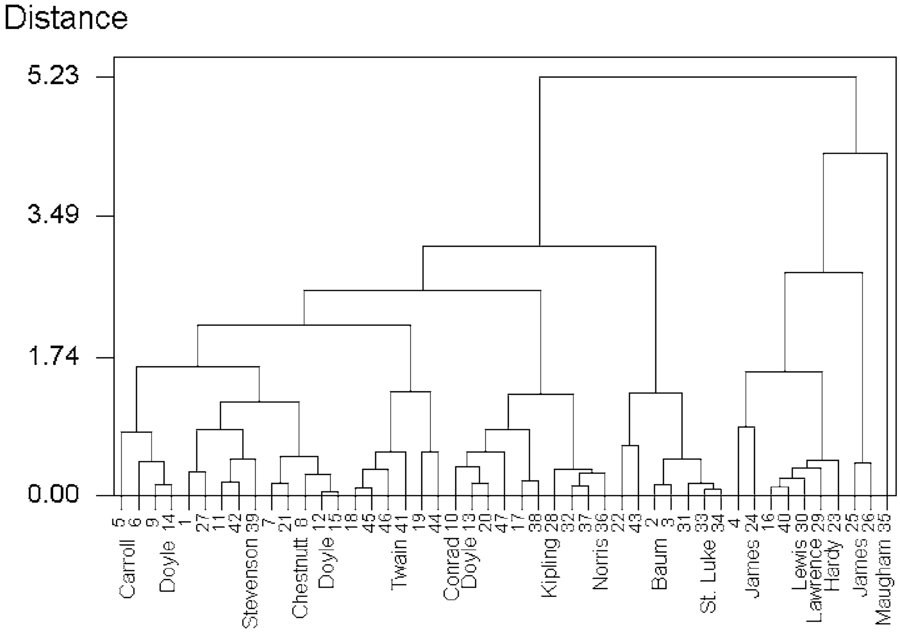
Sons and Lovers (Lawrence, 1996) [1913]
Of Human Bondage (Maugham, 1996) [1915]
Nineteen Eighty-Four (Orwell, 1994) [1949]
Treasure Island (Stevenson, 1996) [1883]
Dracula (Stoker, 1996) [1897]
The War of the Worlds (Wells, 1996) [1898]
The Picture of Dorian Gray (Wilde, 1996) [1891]
To the Lighthouse (Woolf, 1996) [1927]

To increase the number of authors represented by multiple texts, I have also added William Golding's *Freefall* and *The Inheritors* and Woolf's *The Voyage Out*, so that nine of thirty-five authors are represented by two or more texts in the resulting group of forty-seven texts (the sixteen used by Tweedie and Baayen plus the thirty-three mentioned above, less two texts that appear in both groups). Figure 8 shows the best results I have been able to achieve, correctly clustering all texts by only thirteen of the thirty-five authors (including cases in which a single text by an author forms its own cluster): Baum, Carroll, Luke, Chestnutt, Conrad, Hardy, Kipling, Lawrence, Lewis, Maugham, Norris, Stevenson, and Twain. Two of the four texts by Doyle (*Hound* and *Valley*), and two of the three texts by James (*Confidence* and *Europeans*) also cluster correctly.

The results using Herdan's V_m and the repeat rate of the most frequent word yield slightly poorer results. Other clusters of statistics are even less effective, and including additional variables generally causes fewer texts to cluster correctly, much as Tweedie and Baayen found that including all seventeen constants produced a less accurate result than did Z and K alone (p. 348).

The fact that the same texts by the same authors tend to cluster correctly in the various analyses above may suggest that these authors' styles are quite consistent. Given the ineffectiveness of the clustering overall, however, another possibility is that the texts by Baum, Carroll, and St. Luke are simply very different from the other texts in the study without being especially similar to each other. Let us extend the analysis to include an even larger group of texts: the forty-seven examined in Figure 8 plus Charlotte Brontë's *Jane Eyre* and *Shirley* and additional novels by Cather, Conrad, Forster, Hardy, Kipling, and Lewis. The addition of these novels creates a group of fifty-five complete texts by thirty-six authors, sixteen of whom are represented by more than one text. For this larger group of texts, a cluster analysis based on tokens and the repeat rate of the most frequent word is not very effective, clustering correctly all of the texts of only three authors (Baum, Carroll, and St. Luke), grouping two texts each by James and Doyle, and placing the single texts of eight authors in their own clusters. The best clustering, shown in Figure 9, is produced by seven measures that are somewhat questionable because they are not completely independent.

Figure 9 shows that all texts by Carroll, Cather, Forster, Hardy, Lewis, and St. Luke cluster accurately, as well as three texts by Doyle and two texts by James;

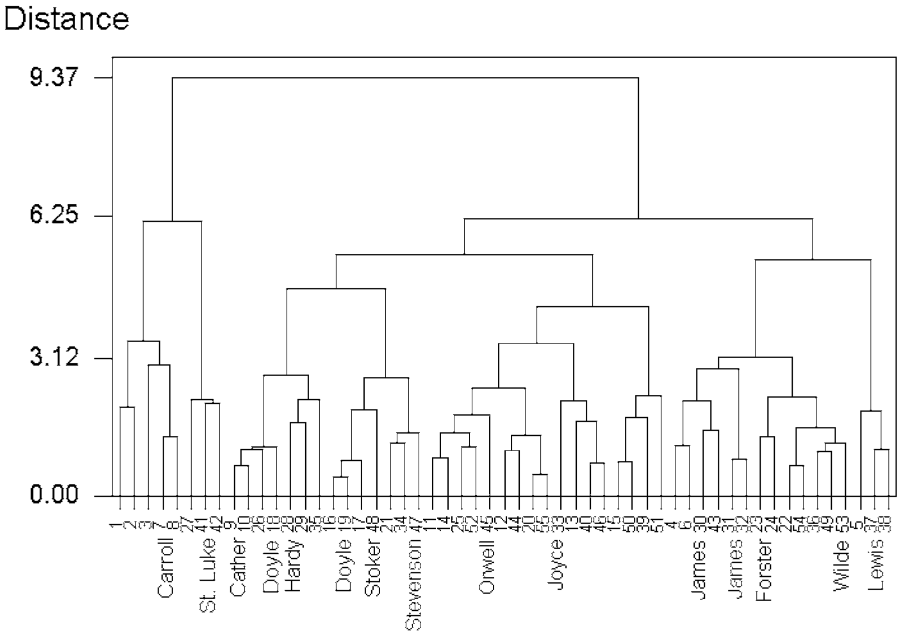


Text Key: Anderson, 1; Baum, 2–3; Brontë, 4; Carroll, 5–6; Cather, 7; Chestnutt, 8; Chopin, 9; Conrad, 10; Crane, 11; Conan Doyle, 12–15; Dreiser, 16; Faulkner, 17; Ford, 18; Forster, 19; Frederic, 20; Golding, 21–22; Hardy, 23; James, 24–26; Joyce, 27; Kipling, 28; Lawrence, 29; Lewis, 30; London, 31–32; St. Luke, 33–34; Maugham, 35; Norris, 36; Orwell, 37; Sinclair, 38; Stevenson, 39; Stoker, 40; Twain, 41; Wells, 42–43; Wharton, 44; Wilde, 45; Woolf, 46–47

Figure 8. Cluster analysis of forty-seven complete texts, based on word tokens and the repeat rate of the most frequent word.

the single texts by Joyce, Orwell, Stevenson, Stoker, and Wilde form their own clusters. Unfortunately, this means that the texts of only eleven of the thirty-six authors cluster correctly – not very encouraging results.

One final expansion of the number texts under analysis will point toward an explanation. Remember that the twenty-nine 24,000-word sections of the original sixteen texts clustered quite well (see Figure 5, above). It is unreasonable to expect measures of vocabulary richness to correctly cluster all 188 of the 24,000-word sections of the 55 texts that are analyzed in Figure 9, but an analysis using the same seven variables fails in a spectacular way.[16] In fact, Lewis Carroll is the only author represented by more than one text for whom all the sections of his (two) texts cluster correctly. For two authors represented by a single text, both sections of that text correctly constitute complete clusters: Sherwood Anderson's *Winesburg Ohio* and Stephen Crane's *The Red Badge of Courage*. In three more cases, both sections of a text constitute a complete cluster: William Golding's *The Inheritors*, H. G. Wells's *The War of the Worlds*, and Kipling's *The Jungle Book.* Two single-section



Text Key: Anderson, 1; Baum, 2–3; C. Brontë, 4–5; E. Brontë, 6; Carroll, 7–8; Cather, 9–10; Chestnutt, 11; Chopin, 12; Conrad, 13–14; Crane, 15; Doyle, 16–19; Dreiser, 20; Faulkner, 21; Ford, 22; Forster, 23–24; Frederic, 25; Golding, 26–27; Hardy, 28–29; James, 30–32; Joyce, 33; Kipling, 34–35; Lawrence, 36; Lewis, 37–38; London, 39–40; St. Luke, 41–42; Maugham, 43; Norris, 44; Orwell, 45; Sinclair, 46; Stevenson, 47; Stoker, 48; Twain, 49; Wells, 50–51; Wharton, 52; Wilde, 53; Woolf, 54–55

Figure 9. Cluster analysis of fifty-five complete texts, based on the ratio of *Hapax Legomena* to *Dis Legomena*, *Hapax Legomena* cubed times types squared, Herdan's V_m, Yule's *K*, Carroll *TTR* (types/square root of twice the tokens), word length, and the repeat rate of the most frequent word.

texts constitute whole clusters: Baum's *The Marvelous Wizard of Oz* and St. Luke's *Acts.* Finally, in several other cases, two or more sections of text(s) by the same author cluster together without constituting all the texts by that author or any one complete text. Clearly these measures of vocabulary richness (and word-length) capture some aspects of authorial style, but just as clearly, they fail to separate large numbers of texts by different authors or to cluster all sections of single texts together.

The number of possible combinations of variables that can be used for cluster analysis is so great that it is impractical to test them all. Furthermore, the effectiveness of different groups of variables is different for different groups of texts. After dozens of attempts, however, the best result I have been able to produce for the 188 sections of the fifty-five texts uses *W, H, K,* Skewness, word length, *hapax legomena*, and the frequency of the most frequent word, and is the same as the grouping just described, except that both of the texts by St. Luke cluster together, and Baum's *The Wonderful Wizard of Oz,* and Woolf's *To the Lighthouse* also constitute single clusters.[17]

7. Intra- and Inter-textual Variability in Vocabulary Richness

Examining measures of vocabulary richness over a moderately large number of texts makes their frequent failure to distinguish texts and authors seem less surprising, and even inevitable because the variation shown by a single text or a single author is often very great. For example, of the 188 sections of the fifty-five novels discussed above, sections of London's *The Seawolf* rank as low as 78th and as high as 144th in vocabulary, sections of Lawrence's *Sons and Lovers* rank as low as 28th and as high as 97th, sections of Virginia Woolf's *The Voyage Out* rank as low as 61st and as high as 135th, sections of Hardy's *Jude the Obscure* rank as low as 65th and as high as 159th, and sections of Joyce's *A Portrait of the Artist* rank as low as 49th and as high as 173rd. Different novels by the same author also vary greatly: Doyle's *The Hound of the Baskervilles* ranks 25th and 73rd, while *The Sign of Four* ranks 127th, Golding's *The Inheritors* ranks 4th and 7th while *Freefall* ranks 95th, 107th, and 110th, Kipling's *The Jungle Book* ranks 12th and 29th, while *Kim* ranks 102nd, 131st, 145th, and 161st. More concretely, the range in vocabulary for sections of Golding's two novels is from 2462 to 3949 words, and for sections of Kipling's two novels from 2935 to 4450 words, while eleven texts by eleven authors occupy ranks 99–109, with vocabularies ranging only from 3876 to 3945. And these are not tiny sections that might be expected to vary significantly – 24,000 words is about half the size of a short novel. If the vocabularies of sections of different texts by a single author can vary by more than 1500 words while the vocabularies of sections of texts by eleven different authors can vary by fewer than 70 words, there seems little hope that vocabulary richness alone can be safely used to determine authorship, or to illuminate an author's style.

Other measures of vocabulary richness are more complexly derived, but they display the same problem. For example, values of Z are much larger than the simple numbers of types, and have a much greater range, from about 9,800 to 113,000 for the 188 sections. Nevertheless, fifteen texts by fifteen authors occupy the fifteen ranks from 95 to 109, with a range of Z only from 34,831 to 38,805, while the ranks of Kipling's texts range from 20 to 171, with a range of Z from 18,524 to 62,596. The ranks of sections of Joyce's *A Portrait of the Artist* range from 72 to 178, with a range of Z from 29,230 to 71,119.

It seems clear that, as more and more texts are added to the comparison, the point is necessarily reached when no further distinctive values for the vocabulary richness measures are possible.[18] On the practical level, texts like those analyzed here show that it would be unwise for anyone doing authorship studies to place much confidence in the presence of a set of texts for a single claimant that display consistent figures for vocabulary richness: a disputed text displaying very different vocabulary richness cannot be reliably assumed to belong to a different author.

Various measures of vocabulary richness produce further interesting differences in how they rank texts on the basis of vocabulary richness – differences that reflect their radically different bases and methods of calculation. Tweedie and Baayen mention the fact that the seventeen measures they examine fall into four groups on the basis of how they rank their sixteen texts, and that only two of the groups are very effective at separating texts by different authors (p. 336). My own analysis of fifty-five complete texts confirms their groupings for the nine of their measures I have calculated. It is instructive, however, to examine the differences in ranking among the relatively effective measures a bit more closely. First, consider the rankings of the first 24,000-word section of each of the fifty-five texts examined above. Figure 10 shows that, for these texts of identical length, the rankings produced by *W, R, k,* and *C* are almost identical and match the ranking for types, Carroll's *TTR*, and the relatively unreliable *LN*.[19] The legend for the chart reads across and then down, with texts ordered according to how they rank in numbers of types. That is, LewisM has the largest number of types among the fifty-five sections, ranking second among all 188 sections; BaumW ranks 188th.[20] The measures Z and *H*, which belong to the same group, produce somewhat different rankings – in the case of *H*, quite different. This result is consistent with the fact that *H* is least like the other variables in its group (Tweedie and Baayen, 1998, p. 338). Yule's *K*, which belongs to the other fairly effective group of measures, predictably produces even more disparate rankings.

So far, these variations in richness order merely emphasize the fact that different measures of vocabulary richness measure different aspects of vocabulary structure. When we examine the rankings for the complete texts, however, the effect of the artificially identical lengths of the texts disappears, and Figure 11 shows a rather different pattern. For the whole texts, *R, k, Z,* and Carroll's *TTR* are quite consistent, while *W, C, H, K,* and *LN* produce wildly different rankings (the legend reads as for Figure 10, except that here the texts are ordered by their rank for R). This is further evidence for the failure of most of these measures to achieve independence from text length. The great variety in ranking also emphasizes the artificial nature of the measures and shows why using more of the measures does not produce more accurate groupings of texts. Finally, the wide disparity in rankings emphasizes how crucial the selection of constants can be in determining the outcome of an analysis.

8. Conclusion

What have we learned? Readers' perceptions about which texts or authors have large vocabularies are not necessarily accurate. Some measures fail to register even some extreme kinds of textual alteration that intuitively seem important to the overall style of a text. Because so many of the types in a text are *hapax legomena*, different texts and even different sections of a single text by one author are almost as different in vocabulary content as are texts by different authors, to say nothing of being different in vocabulary richness. As Tweedie and Baayen have shown, many

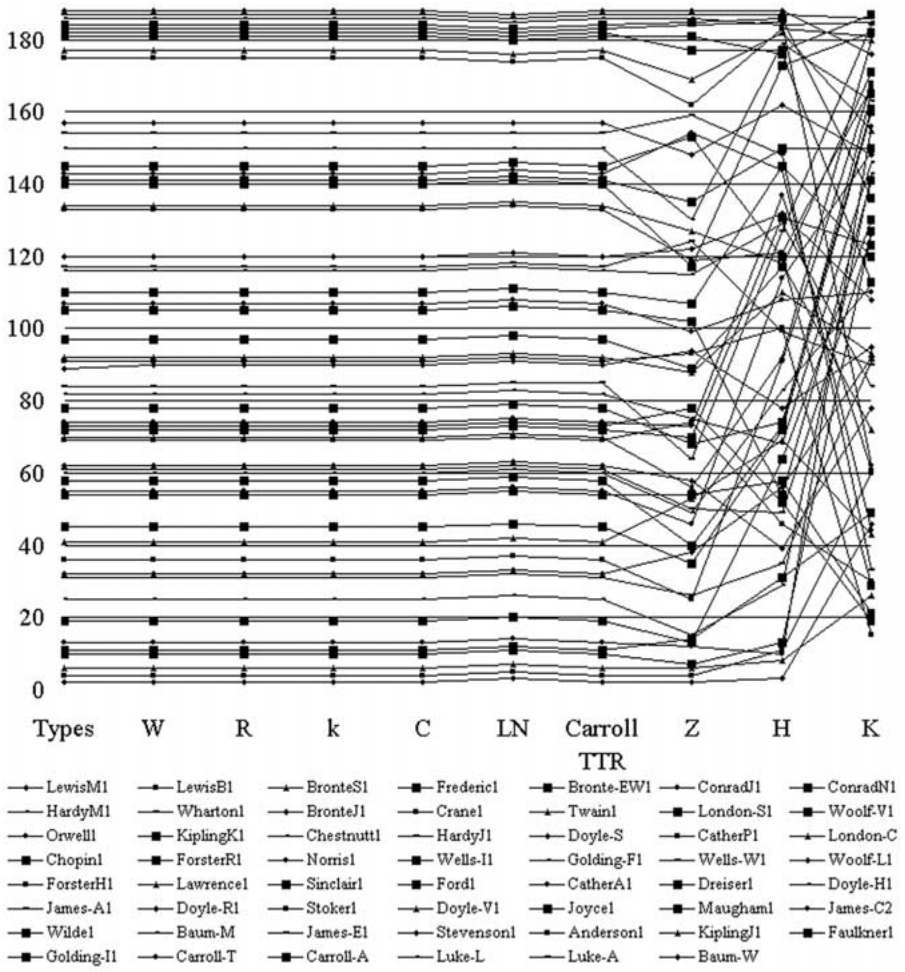


Figure 10. Lexical richness rankings for the first 24,000-word sections of fifty-five texts.

so-called constants either fail to be theoretically constant, or fail to be constant in practice; some do a poor job of clustering or differentiating texts, and using larger numbers of measures does not improve the effectiveness or accuracy of an analysis. We have also learned that some authors are relatively consistent in vocabulary richness across some texts and sections of texts, while other texts or sections by the same authors show differences that are quite extreme: an author's consistency across one group of texts is no guarantee that the next text by that author will be consistent with the others. Finally, we have learned that adding more texts to an analysis based on vocabulary richness reduces its accuracy, and that a fairly accurate and reliable analysis is possible only with a small and extremely various group of texts – texts for which such an analysis is least likely to be necessary or useful.

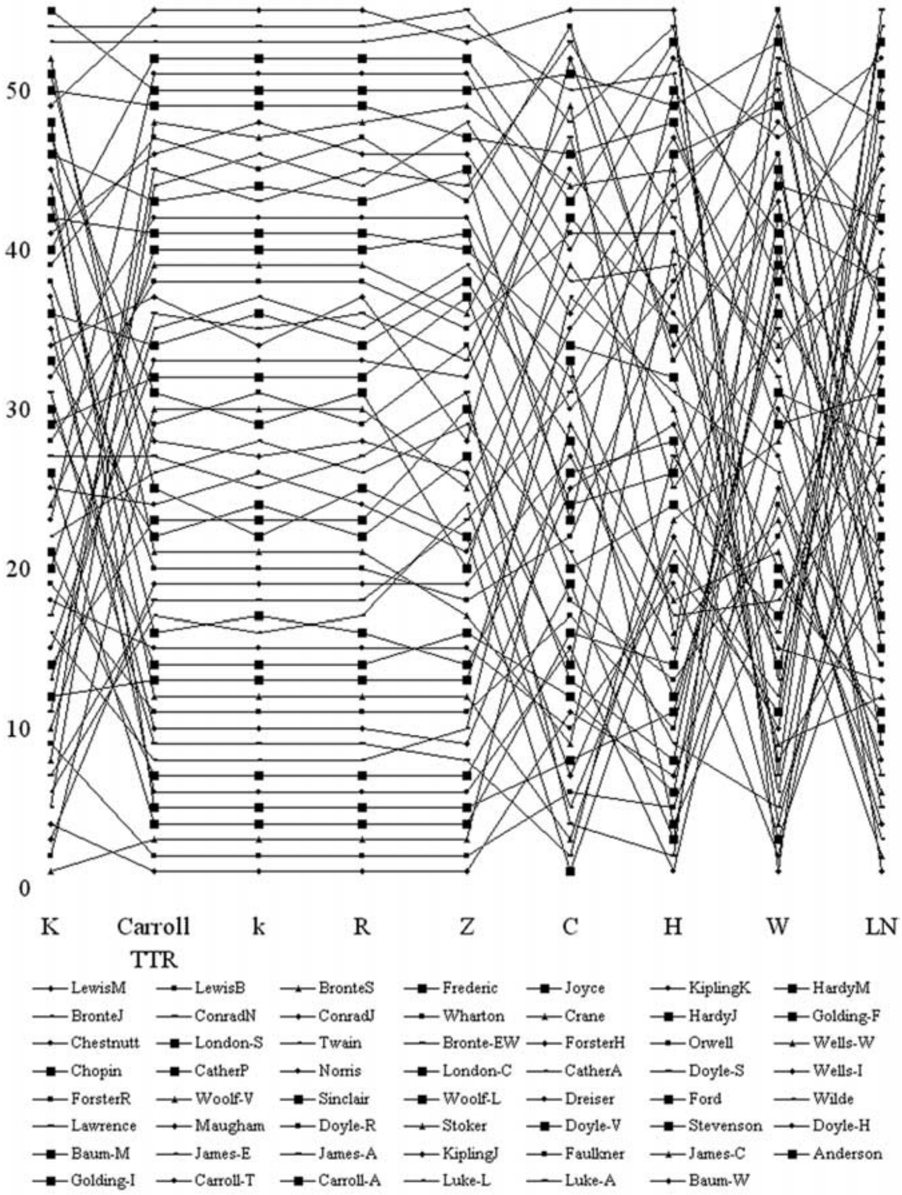


Figure 11. Lexical richness rankings for fifty-five whole texts.

Two final cluster analyses, both based on Herdan's V_m and the repeat rate of the most frequent word, dramatically illustrate the dangers of using vocabulary richness measures to group and distinguish texts: Figure 12 shows a group of fourteen texts by seven authors that cluster perfectly, and Figure 13 shows a group of sixteen texts by eight authors with no correct clusters at all. The chief determinant of the accuracy of clustering in an analysis based on vocabulary richness is simply

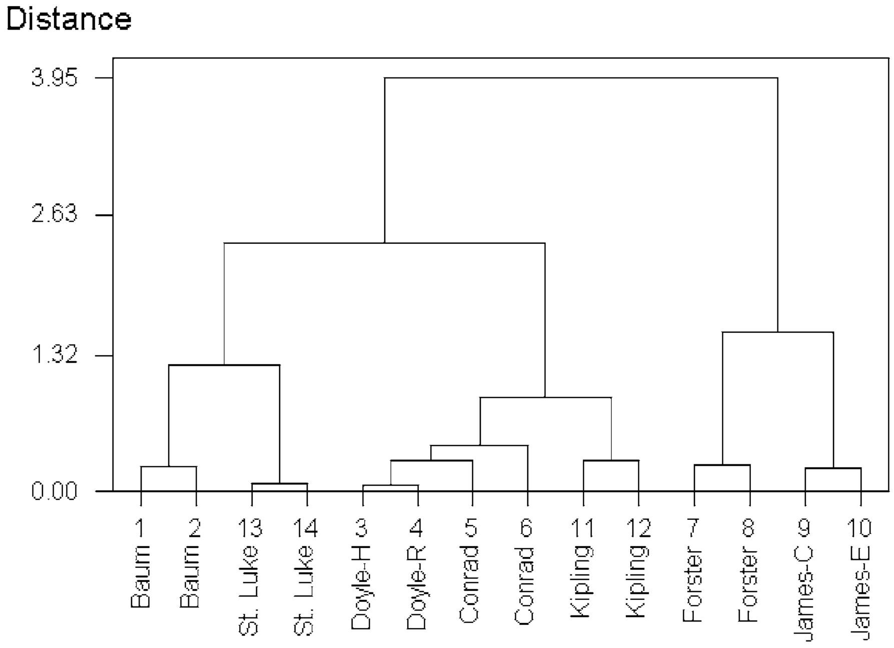


Figure 12. Cluster analysis of fourteen texts by seven authors based on Herdan's V_m and the repeat rate of the most frequent word: Best case scenario.

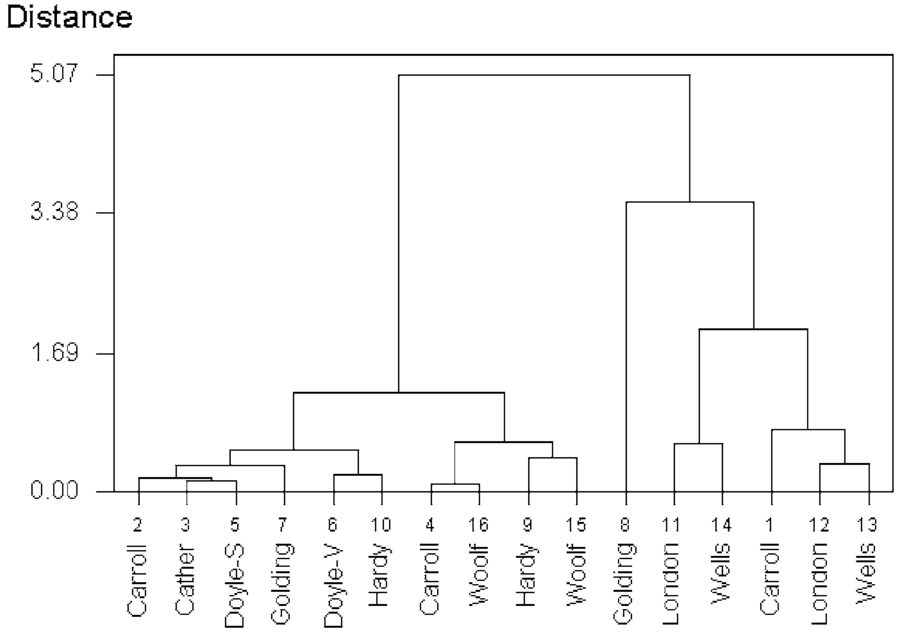


Figure 13. Cluster analysis of sixteen texts by eight authors based on Herdan's V_m and the repeat rate of the most frequent word: Worst case scenario.

the choice of texts to be analyzed. Had Tweedie and Baayen picked the texts in Figure 13 to analyze, their conclusions would have been radically different. In retrospect, this is hardly surprising. The tremendous variety of texts within their group of sixteen – from Early Modern English religious texts to children's literature to detective fiction to science fiction – is so great that a perceptive reader of the texts should be able to identify the author of nearly any 50-word passage from any of the texts.

Despite the attractiveness of measures of vocabulary richness, and despite the fact that they are sometimes effective in clustering texts by a single author and discriminating those texts from other texts by other authors, such measures cannot provide a consistent, reliable, or satisfactory means of identifying an author or describing a style. There is so much intratextual and intertextual variation among texts and authors that measures of vocabulary richness should be used with great caution, if at all, and should be treated only as preliminary indications of authorship, as rough suggestions about the style of a text or author, as characterizations of texts at the extremes of the range from richness to concentration. Perhaps their only significant usefulness is as an indicator of what texts or sections of texts may repay further analysis by more robust methods (see Hoover, 1999, pp. 79–113). Unfortunately, the long-cherished goal of a measure of vocabulary richness that characterizes authors and their styles appears to be unattainable. The basic assumption that underlies it is false.

Notes

[1] In "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" Craig (1999) provides a helpful and illuminating discussion of the linkage between authorial attribution and statistical stylistics (he addresses multivariate analysis of frequent words rather than vocabulary richness), a linkage that is also signally present in the work of John F. Burrows (1987, 1992; Burrows and Craig, 1994).

[2] Yule (1944) seems to have begun the search for a single constant that measures vocabulary richness independently of text length. His characteristic K achieves independence of text-length, and its calculation takes into account the frequencies of all of the words in a text. Tweedie and Baayen provide a useful overview of the origins of and formulas for the most important measures of vocabulary richness (1998, pp. 325–331).

[3] When I have presented similar lists of authors to English graduate students at New York University over the past fifteen years, Faulkner and James have invariably been among the authors predicted to have the largest vocabularies.

[4] Thoiron defines the two measures as follows (198, 200):
Diversity: $\sum i(i-1)V_i/N(N-1)$
Entropy: $-\sum V_i \cdot p_i \log p_i \quad \textit{where } p_i = i/N$ or
$\log N - ((1/N)(\sum i \cdot V_i \cdot \log i))$
Tweedie and Baayen define them slightly differently (1998, pp. 329–330):
Diversity: $\sum_{i=1}^{V(N)} V(i,N)\frac{i}{N}\frac{i-1}{N-1}$
Entropy: $\sum_{k=1}^{V(N)} -\log(p_k)p_k$

[5] Herdan's V_m is defined as follows: $\sqrt{\sum_{i=1}^{V(N)} V(i,N)(i/N)^2 - \frac{1}{V(N)}}$ (Tweedie and Baayen, 1998, p. 330). The repeat rate of the most frequent word is simply the number of tokens divided by

the frequency of the most frequent word. Kurtosis, a measure of the pitch of the word frequency distribution curve, and skewness, the peaking of the distribution at a value higher or lower than the mean, will not be discussed further. Yule's *K* is defined as follows: $10^4[-\frac{1}{N} + \sum_i V(i, N)(\frac{i}{N})^2]$ (Tweedie and Baayen, 1998, p. 330). The figures for *K* in Table 1 are independently calculated, rather than taken from TACT. I am grateful to *CHUM's* reviewers for pointing out that the figures for *K* that TACT produces (which are not affected by the doubling of a text) are erroneous.

[6] Zipf's *Z* is a free parameter of which the vocabulary of the text, *V(N)*, is a function: $V(N) = \frac{Z}{\log(p{*}Z)} \frac{N}{N-Z} \log N/Z$, where $p*$ "is the maximum sample relative frequency – the frequency of the most common word divided by the text length" (Tweedie and Baayen, 1998, p. 331). Carroll *TTR* is the number of types divided by the square root of twice the number of tokens, and Sichel's S is the ratio of *dis legomena* to total vocabulary size (p. 329).

[7] I should emphasize that I am making no claims about the statistical significance of any of these differences. Any statistical tests for significance would be better carried out on larger samples, and seem unnecessary for the rather general point I am making here about the differences and similarities among doubled and combined texts.

[8] Yule himself considered vocabulary concentration (a small, focused vocabulary) rather than vocabulary richness (a large, varied vocabulary) a mark of high quality (1944, pp. 122, 131); a high *K* value implies a small vocabulary. For fiction, however, a richer vocabulary is likely to be more highly valued.

[9] The measures *K, D,* V_m, and *Z* have been defined above. The measures *b* and *c*, which come from Sichel, are two free parameters related to vocabulary size as follows (Tweedie and Baayen, 1998, p. 331): $V(N) = \frac{2}{bc}[1 - e^{b(1-\sqrt{1+Nc})}]$.

[10] Tweedie and Baayen list this text as *Tip Manufactures a Pumpkinhead*, which seems to be a subtitle.

[11] The other constants are defined above; Herdan's *C* and Rubet's *k* as follows (Tweedie and Baayen, 1998, p. 327): $C = \frac{\log V(N)}{\log N} \quad k = \frac{\log V(N)}{\log(\log N)}$.

[12] Unless otherwise indicated, all cluster analyses were performed in Minitab using standardized variables (to reduce the effect of differences in variable size), complete linkage, and Euclidean distance.

[13] The repeat rate of the most frequent word is independent of the length of the text, so that it would seem a more appropriate measure to use than the frequency of the most frequent word. Nevertheless, when the texts being compared are of equal size, I have sometimes used the frequency of the most frequent word because it results in more accurate clustering.

[14] Minitab actually clusters the texts by St. Luke and Baum's *Wizard* together, although the separation between the two authors is fairly clear. Since more accurate clustering works against my argument, however, I have interpreted this and some other dendograms liberally. Tweedie and Baayen do not indicate the precise cluster membership in their dendograms, so that it is not possible to make fully accurate comparisons.

[15] To be attractive as indicators of authorial style, analyses involving vocabulary richness should ideally be at least 95% accurate, corresponding to $p < 0.05$. None of the analyses presented here achieve that level. As we will see, however, the problems with vocabulary richness are so severe that the issue of precise accuracy is not terribly important.

[16] Crane's *The Red Badge of Courage* has only 46138 words; I have added the beginning of "The Bride Comes to Yellow-Sky" (Crane, 1996a) to make forty-eight thousand words.

[17] *W* and *H* are defined as follows (Tweedie and Baayen, 1998, pp. 328–329):

$$W = N^{V(N)^{-a}}$$

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

[18] Principal Components analysis of the most frequent words of texts has a far greater potential for separating large numbers of texts because of the large number of variables involved. As Tweedie and

Baayen note, however, even a cluster analysis based on principal components fails to group all of the texts correctly (pp. 346–347). They do not give sufficient details about this part of their analysis to allow any firm conclusions; further work will be required to determine whether this local failure of principal components analysis is actually more general.

[19] *LN* is defined as follows (Tweedie and Baayen, 1998, p. 328): $LN = \frac{1-V(N)^2}{V(N)^2 \log N}$.

[20] The abbreviations for the texts should be transparent. Note that for James's *Confidence*, I have used the second section rather than the first. As occasionally happens, TACT's count for types is slightly different from my own (here 24,001 rather than 24,000), and even this small difference alters some of the measures of vocabulary richness. I have used the rankings from 1 to 188, rather than 1 to 55 so that the minor differences among the texts can be seen.

References

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

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

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

Brainerd B. (1988) Two Models for the Type-Token Relation with Time Dependant Vocabulary Reservoir. In Thoiron P., Serant D., Labbe D. (eds.), *Vocabulary Structure and Lexical Richness*, Champion-Slatkine, Paris.

Burrows J.F. (1987) *Computation into Criticism*. Clarendon Press, Oxford.

Burrows J.F. (1992) *Computers and the Study of Literature*. In Butler, pp. 167–204.

Burrows J.F., Craig D.H. (1994) Lyrical Drama and the 'Turbid Mountebanks': Styles of Dialogue in Romantic and Renaissance Tragedy. *Computers and the Humanities*, 28, pp. 63–86.

Craig H. (1999) Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them? *Literary and Linguistic Computing*, 14(1), pp. 103–113.

Craig H. (1999) Contrast and Change in the Idiolects of Ben Jonson Characters. *Computers and the Humanities*, 33, pp. 221–240.

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

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

Hoover D.L. (1999) *Language and Style in The Inheritors*. University Press of America, Lanham, MD.

Lancashire I., Bradley J., McCarty W., Stairs M., Wooldridge R.R. (1996) *Using TACT with Electronic Texts*. MLA, New York.

Minitab Release 12.2, Minitab, Inc., State College, Pennsylvania.

TACT Version 2.1, Centre for Computing in the Humanities, University of Toronto.

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

Tweedie F.J., Baayen R.H. (1998) How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32, pp. 323–352.

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

Tweedie F.J., Singh S., Holmes D.I. (1996) Neural Network Applications in Stylometry: *The Federalist Papers*. *Computers and the Humanities*, 30, pp. 1–10.

Yule G.U. (1994) *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.

Literary Texts Analyzed

The Acts of the Apostles (KJV). (2000) Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/] Downloaded January 12, 2000.

Anderson, Sherwood. (1996) *Winesburg, Ohio*. 1919. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded March 1, 1996.

Baum, L. Frank. (2000) *The Marvelous Land of Oz*. 1904. Online. Project Gutenberg. [ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg/etext93/] Downloaded January 12, 2000.

Baum, L. Frank. (2000) *The Wonderful Wizard of Oz*. 1900. Online. Robert Stockwell, Carnegie Mellon University. [http://www.cs.cmu.edu/People/rgs/] Downloaded January 12, 2000.

Brontë, Charlotte. (1998) *Jane Eyre*. 1846. Online. The English Server, Carnegie Mellon University. [http://english-www.hss.cmu.edu/fiction/] Downloaded December 18, 1998.

Brontë, Charlotte. (2000) *Shirley*. 1849. Online. The Brontë Sisters Web. [http://www. lang.nagoya-u.ac.jp/∼matsuoka/Bronte.html] Downloaded January 21, 2000.

Brontë, Emily. (2000) *Wuthering Heights*. 1847. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded January 14, 2000.

Carroll, Lewis. (2000) *Alice's Adventures in Wonderland*. 1866. Online. Project Gutenberg. Available: ftp://sunsite.unc.edu/pub/docs/books/gutenberg/etext97/. Downloaded January 12, 2000.

Carroll, Lewis. (2000) *Through the Looking-Glass and What Alice Found There*. 1862–63. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html/] Downloaded January 12, 2000.

Cather, Willa. (1996) *My Antonia*. 1918. Online. The English Server, Carnegie Mellon University. [http://english-server.hss.cmu.edu/fiction/] Downloaded May 16, 1996.

Cather, Willa. (2000) *The Professor's House*. 1925. Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/] Downloaded January 18, 2000.

Chestnutt, Charles W. (1996) *The House Behind the Cedars*. 1900. Athens: University of Georgia Press. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 23, 1996.

Chopin, Kate. (1996) *The Awakening and Selected Short Stories*. 1899. [New York: Bantam Books, 1988.] Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 16, 1996.

Conrad, Joseph. (1996) *Lord Jim*. 1900. N.p. [1961 reprint of the first edition.] Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 29, 1996.

Conrad, Joseph. (2000) *The Nigger of the Narcissus*. 1897. New York: Penguin Classics, 1987. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded January 21, 2000.

Crane, Stephen. (1996a) "The Bride Comes to Yellow-Sky". *McClure's Magazine* X, February 1898, 377–384. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 28, 1996.

Crane, Stephen. (1996b) *The Red Badge of Courage*. 1895. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 31, 1996.

Doyle, Sir Arthur Conan. (2000) *The Hound of the Baskervilles. The Strand Magazine*, August 1901–April 1902. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded January 12, 2000.

Doyle, Sir Arthur Conan. (1996) *The Return of Sherlock Holmes. The Strand Magazine*, October 1903. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 29, 1996.

Doyle, Sir Arthur Conan. (2000) *The Sign of Four*. 1890. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded January 12, 2000.

Doyle, Sir Arthur Conan. (2000) *The Valley of Fear. The Strand Magazine*, September 1914–May 1915. Online. University of Virginia Library. http://etext.lib.virginia.edu/modeng/modeng0.browse.html. Downloaded January 12, 2000.

Dreiser, Theodore. (1996) *Sister Carrie*. 1900. Online. Virginia Tech. University. [gopher://gopher.vt.edu:10010/10/33] Downloaded May 16, 1996.

Faulkner, William. (1994) *Light in August*. 1932. *Novels, 1930–1935*. New York: Library of America, 1985. Online. Oxford Text Archive. No longer available. Downloaded June 9, 1994.

Ford, Ford Maddox. (1996) *The Good Soldier*. 1915. Reprint, New York: Vintage, 1989. Scanned and corrected May 21, 1996.

Forster, E.M. (1996) *Howards End*. 1910. *Great Novels of E.M. Forster: Where Angels Fear to Tread, The Longest Journey, A Room with a View, Howards End*. New York: Caroll & Graff Publishers, Inc., 1992. Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/english/pd-modeng/bibl.html] Downloaded May 16, 1996.

Forster, E.M. (2000) *A Room with a View*. 1908. *Great Novels of E. M. Forster: Where Angels Fear to Tread, The Longest Journey, A Room with a View, Howards End*. New York: Caroll & Graff Publishers, Inc., 1992. Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/english/pd-modeng/bibl.html] Downloaded May 16, 1996.

Frederic, Harold. (1996) *The Damnation of Theron Ware*. 1896. Online. Project Gutenberg. http://www.promo.net/pg/list.html. Downloaded May 23, 1996.

Golding, William. (1960) *Free Fall*. 1959. New York: Harcourt, Brace, & World. Scanned and corrected January 6, 1997.

Golding, William. (1955b) *The Inheritors*. New York: Harcourt, Brace, & World. Created 1985.

Hardy, Thomas. (1996) *Jude the Obscure*. 1896. Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/english/pd-modeng/bibl.html] Downloaded May 31, 1996.

Hardy, Thomas. (2000) *The Mayor of Casterbridge*. Online. Wiretap. [gopher://wiretap.area.com/00/Library/Classic/] Downloaded January 17, 2000.

James, Henry. (1996) *The Ambassadors*. 1909. *The Novels and Tales of Henry James*. New York: Charles Scribner's Sons, 1907–17. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 16, 1996.

James, Henry. (2000) *Confidence*. 1879. Literary Classics of the United States, New York: Viking Press, 1983. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded January 12, 2000.

James, Henry. (2000) *The Europeans*. Boston: Houghton, Osgood and Company, 1878. Online. The Henry James scholar's Guide to Web Sites. [http://www.newpaltz.edu/~hathaway/] Downloaded January 12, 2000.

Joyce, James. (1996) *A Portrait of the Artist as a Young Man*. 1916. Online. Bibliomania. [http://www.bibliomania.com/Fiction/] Downloaded May 31, 1996.

Kipling, Rudyard. (2000) *The Jungle Book*. 1893. Online. Robert Stockwell, Carnegie Mellon University. [http://www.cs.cmu.edu/People/rgs/] Downloaded January 19, 2000.

Kipling, Rudyard. (1996) *Kim*. 1901. Online. Virginia Tech University. [gopher://gopher.vt.edu:10010/10/33] Downloaded May 16, 1996.

Lawrence, D.H. (1996) *Sons and Lovers*. New York: Viking Press, 1913. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 16, 1996.

Lewis, Sinclair. (2000) *Babbitt*. 1922. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded January 21, 2000.

Lewis, Sinclair. (1996) *Main Street*. 1920. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 16, 1996.

London, Jack. (2000) *The Call of the Wild*. 1903. Online. Wiretap. [gopher://wiretap.area.com/00/Library/Classic/] Downloaded January 8, 2000.

London, Jack. (1996) *The Sea Wolf*. 1904. Reprint, New York: Library of America, 1982. Online. Oxford Text Archive. [ftp://ftp.hti.umich.edu/pub/ota/public/] Downloaded May 31, 1996.

Luke (KJV). (2000) Online. Humanities Text Initiative, University of Michigan. [http://www.hti.umich.edu/] Downloaded January 12, 2000.

Maugham, W. Somerset. (1996) *Of Human Bondage*. Garden City, New York: Doubleday, Doran & Company, Inc., 1915. Online. Wiretap. [gopher://wiretap.spies.com:70/11/Books] Downloaded May 19, 1996.

Norris, Frank. (1996) *McTeague*. 1899. Reprint, New York: Rinehart & Co., 1958. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 31, 1996.

Orwell, George. (1994) *Nineteen Eighty-Four*. 1949. Reprint, New York: New American Library, 1961. Online. Oxford Text Archive. [http://ota.ahds.ac.uk/] Downloaded June 9, 1994.

Sinclair, Upton. (1996) *The Jungle*. 1906. [Reprint, New York: Signet, 1960?] Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 16, 1996.

Stevenson, Robert Louis. (1996) *Treasure Island*. 1883. Reprint, New York: Signet, 1981. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 29, 1996.

Stoker, Bram. (1996) *Dracula*. 1897. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 22, 1996.

Twain, Mark. (1996) *The Tragedy of Pudd'nhead Wilson*. 1894. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 31, 1996.

Wells, H.G. (2000) *The Invisible Man*. 1897. Best Science Fiction Stories of H. G. Wells, New York: Dover, 1966. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html. Downloaded January 12, 2000.

Wells, H.G. (1996) *The War of the Worlds*. 1898. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 31, 1996.

Wharton, Edith. (1996) *The Age of Innocence*. 1920. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 16, 1996.

Wilde, Oscar. (1996) *The Picture of Dorian Gray*. 1891. Reprint, New York: New American Library, 1962. Online. University of Virginia Library. [http://etext.lib.virginia.edu/modeng/modeng0.browse.html] Downloaded May 31, 1996.

Woolf, Virginia. (1996) *To the Lighthouse*. 1927. Reprint, New York: Harcourt, Brace, & World, 1955. Online. Oxford Text Archive. [http://ota.ahds.ac.uk/] Downloaded December 13, 1996.

Woolf, Virginia. (1996) *The Voyage Out*. 1915. Online. Project Gutenberg. [http://www.promo.net/pg/list.html] Downloaded May 11, 1996.