Computers and the Humanities **35:** 193–214, 2001.

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.

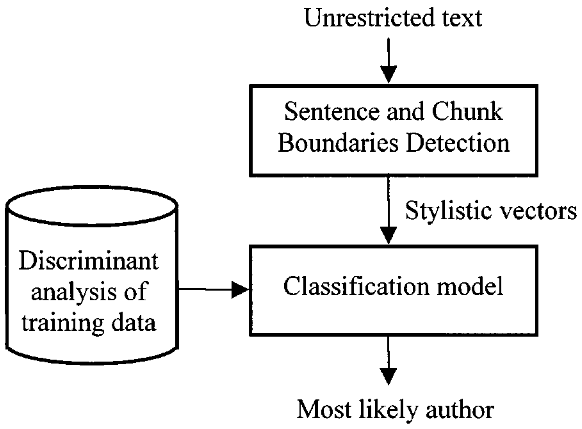
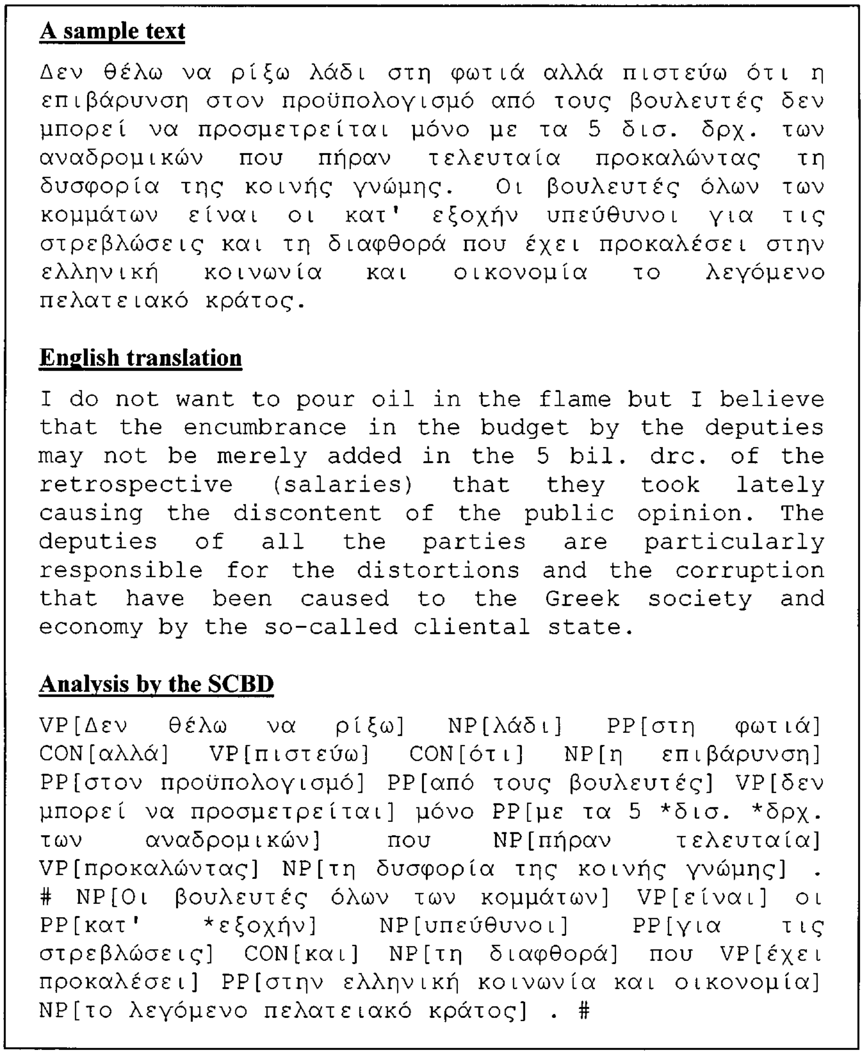

Unrestricted text
Sentence and Chunk Boundaries Detection
Stylistic vectors
Discriminant analysis of training data
Classification model
Most likely author

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	

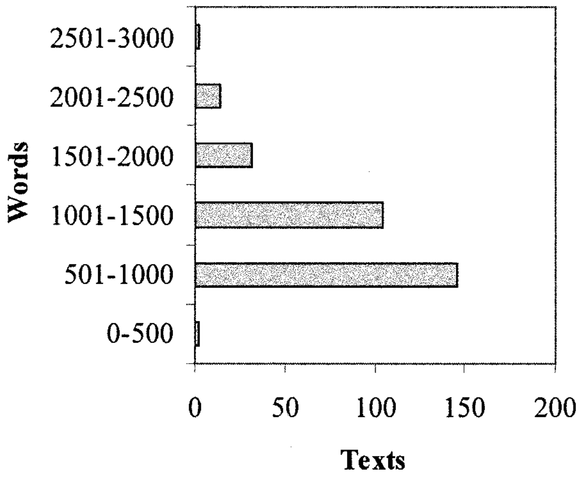


Figure 3. Text-length distribution in the corpus used in this study.

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**										
	A01	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	**A10**	**Error**
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

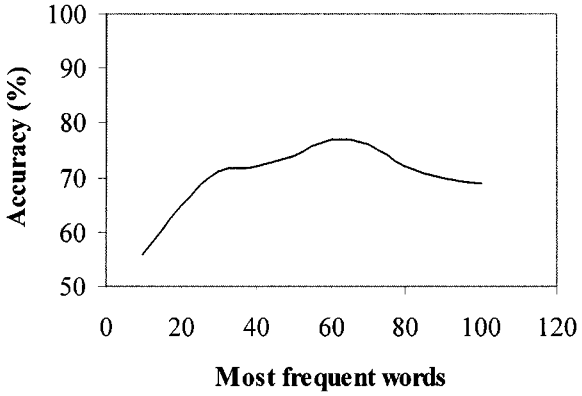


Figure 4. Classification accuracy for different sets of the most frequent words.

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

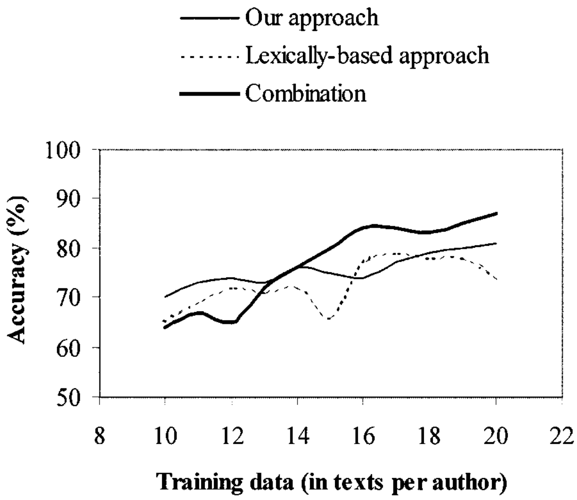


Figure 5. Classification accuracy for different sizes of training data.

(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

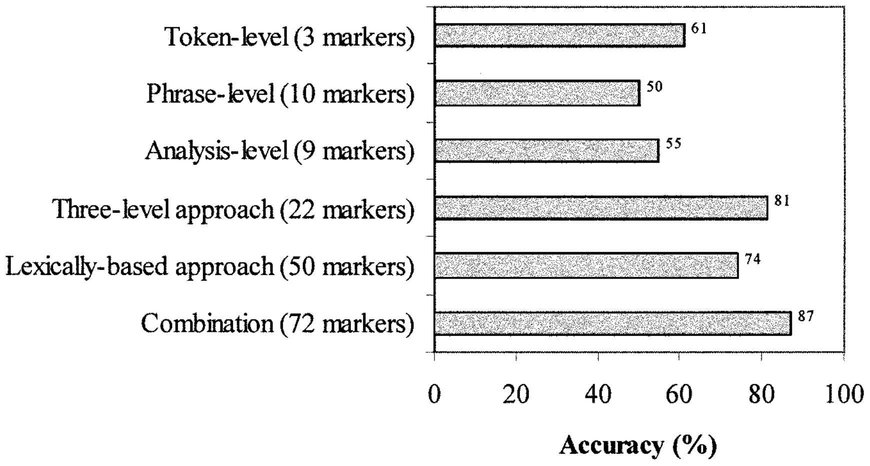


Figure 6. Classification accuracy of the tested models.

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 4th 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.