Authors’ version. Published in Computational Linguistics, 26(4), December 2000, pp. 471-495

Automatic Text Categorization in Terms of Genre and Author

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis

University of Patras

Dept. of Electrical & Computer Engineering

26500 Patras, Greece

Email: {stamatatos, fakotaki, gkokkin}@wcl.ee.upatras.gr

Abstract: The two main factors that characterize a text are its content and its style. Both of them can be used as categorization means. In this paper we present an approach to text categorization in terms of genre and author for Modern Greek. In contrast to hitherto stylometric approaches, we attempt to take full advantage of existing natural language processing (NLP) tools. To this end, we propose a set of style markers including analysis-level measures that represent the way in which the input text has been analyzed and capture useful stylistic information without additional cost. We present a set of small-scale but reasonable experiments in text genre detection, author identification as well as author verification tasks and show that the performance of the proposed method is better in comparison with the most popular distributional lexical measures, i.e., functions of vocabulary richness and frequencies of occurrence of the most frequent words. All the presented experiments are based on unrestricted text downloaded from the World Wide Web (WWW) without any manual text preprocessing or text sampling. Various performance issues regarding the training set size and the significance of the proposed

style markers are discussed. Our system can be used in any application that requires fast and easily adaptable text categorization in terms of stylistically homogeneous categories. Moreover, the procedure of defining analysis-level markers can be followed in order to employ already existing text processing tools for the extraction of useful stylistic information.

1. Introduction

The rapid expansion of the WWW in recent years has resulted in the creation of large volumes of text in electronic form. NLP applications such as information retrieval and information extraction have been developed to treat this information automatically. Since the Internet is a very heterogeneous domain, these applications usually involve text categorization tasks with the following desiderata:

- minimal computational cost,
- ability to handle real-world (or unrestricted) text, and
- either ease of adaptation to a certain domain/application or generality in order to cover a wide area of domains/applications.

The two main factors that characterize a text are its content and its style. Both of them can be used as categorization means. Nevertheless, the literature on computational stylistics is very limited in comparison to the work dealing with the propositional content of the text. This is due to the lack of a formal definition of style as well as to the inability of current NLP systems to incorporate stylistic theories that require complicated information. In contract to traditional stylistics based on formal linguistic theories, the use of statistical methods in style processing has proved to be a reliable approach (Biber

1995). According to stylostatisticians, a given style is defined by a set of measurable patterns, called style markers. In this study we adopt this definition.

Typical classificatory tasks in computational stylistics are the following:

- *Text genre detection:* It concerns the identification of the kind (or functional style) of the text (Karlgren and Cutting 1994; Michos et al. 1996; Kessler et al. 1997), and
- *Authorship attribution:* It concerns the identification of the author of the text (Holmes and Forsyth 1995; Baayen et al. 1996; Tweedie et al. 1996).

These tasks have so far been considered completely separate problems. A typical text categorization system utilizing stylistic analysis (i.e., either text genre or authorship identification) is usually based on the following modules:

a. **Extraction of style markers:** A set of quantifiable measures are defined and a text-processing tool is usually developed, to automatically count the markers.

b. **Classification procedure:** A disambiguation method (e.g., statistical, connectionist, etc.) is applied to classify the text in question into a predefined category (i.e., a text genre or an author).

The most important computational approaches to text genre detection focused on the use of simple measures that can be easily detected and reliably counted by a computational tool (Kessler et al. 1997). Towards this end, various sets of style markers have been proposed (Karlgren and Cutting 1994), which in essence are subsets of the set used by Biber (1995) who ranks registers along seven dimensions, by applying factor analysis to a set of lexical and syntactic style markers that had been manually counted. In general, the current text genre detection approaches try to avoid the employment of existing text processing tools rather than taking advantage of it.

Authorship attribution studies have focused on the establishment of the authorship of anonymous or doubtful literary texts, such as the *Federalist Papers*, 12 of which are of disputed authorship (Mosteller and Wallace, 1984; Holmes and Forsyth 1995). Typical methodologies deal with a limited number of candidate authors using long text samples of several thousands words. Almost all the approaches to this task are based mainly on distributional lexical style markers. In a review paper of authorship attribution studies Holmes (1994, page 87) claims:

> *"…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."*

To the best of our knowledge, there is still no computational system that can distinguish the texts of a randomly-chosen group of authors without having to be assisted by a human as concerns the selection of both the most appropriate set of style markers and the most accurate disambiguation procedure.

In this paper we describe an approach to text categorization in terms of genre and author based on a new stylometric method which utilizes already existing NLP tools. In addition to the style markers relevant to the actual output of the NLP tool (i.e., the analyzed text) we introduce the analysis-level style markers which represent the way in which the text has been analyzed by that tool. Such measures contain useful stylistic information and are easily available without additional computational cost.

The proposed technique is illustrated by applying it to text categorization tasks for Modern Greek corpora using SCBD, an already existing Sentence and Chunk Boundaries Detector in unrestricted Modern Greek text (Stamatatos et al. 2000). We present a set of small-scale but reasonable experiments in text genre detection, author identification, and author verification tasks and show that the proposed method is better in comparison with the most popular distributional lexical measures, i.e., functions of vocabulary richness and frequencies of occurrence of the most frequent words. Our approach is trainable and can be easily adapted to any set of stylistically homogeneous categories.

We begin by discussing work relevant to text genre detection and authorship attribution, focusing on the various types of style markers employed (Section 2). Next we describe the proposed solution for extracting style markers using existing NLP tools (Section 3) and apply our method to Modern Greek (Section 4), briefly describing the SCBD and proposing our set of style markers. The techniques for automatic categorization of the stylistic vectors are discussed in Section 5. Section 6 deals with the application of our approach to text genre detection, and Section 7, with authorship attribution, for both author identification and author verification. In Sections 8 and 9, we discuss important performance issues of the proposed methodology and the conclusions that can be drawn from this study.

2. Current Trends in Stylometry

The main feature that characterizes both text genre detection and authorship attribution studies is the selection of the most appropriate measures that reflect the style of the writing. Various sets have been proposed in the relevant literature. In this section, we try to classify the most popular of the hitherto proposed style markers taking into account the required information for their calculation rather than the task they have been applied to.

2.1 Token-level Measures

The simplest approach considers the sample text as a set of tokens grouped in sentences. Typical measures of this category are: word count, sentence count, character per word count, punctuation marks count, etc. Such features have been widely used in both text genre detection and authorship attribution research since they can be easily detected and computed. It is worth noting that the first pioneering works in authorship attribution, when no powerful computational systems were available, had been exclusively based on these measures. For example, Morton (1965) used sentence length for testing the authorship of Greek prose, Brinegar (1963) adopted word length measures, and Brainerd's (1974) approach was based on distribution of syllables per word. Although such measures seemed to work in specific cases, they became subject to heavy criticism for their lack of generality (Smith 1983, 1985).

2.2 Syntactic Annotation

The use of measures related to syntactic annotation of the text is very common in text genre detection. It has been proved that such measures provide very useful information for the exploration of the characteristics of style (Biber 1995). Typical paradigms are: passive count, nominalization count, and counts of the frequency of various syntactic categories (e.g., part-of-speech tags). Recently, syntactic information has also been applied to authorship attribution. Specifically, Baayen used frequencies of occurrence of rewrite rules, as they appear in a syntactically annotated corpus, and proved that they perform better than word frequencies (Baayen et al. 1996). However, their calculation requires tagged or parsed text. Current NLP tools are not able to provide accurate calculation results for many of the hitherto proposed style markers. In the study of register variation conducted by Biber (1995) a subset of the measures (i.e., the simplest ones) was calculated by computational tools and the remaining were counted manually. Additionally, the automatically acquired measures were counterchecked manually. Many researchers, therefore, try to avoid the use of features related to syntactic annotation in order to avoid such problems (Kessler et al. 1997). As a result, the recent advances in computational linguistics have not notably affected the research in computational stylistics.

2.3 Vocabulary Richness

Various measures have been proposed for capturing the richness or the diversity of the vocabulary of a text and they have been applied mainly to authorship attribution studies. The most typical measure of this category 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 of the sample text. Similar features are the *hapax legomena* (i.e., once-occurring words in the sample text) and the *dislegomena* (i.e., twice-occurring words in the sample text). Since text length dramatically affects these features, many researchers have proposed functions of these features that they claim are text length independent (Honore 1979; Yule 1944; Sichel 1975). Additionally, instead of using a single measure, some researchers were based on a set of such vocabulary richness functions in combination with multivariate statistical techniques in order to achieve better results in authorship attribution (Holmes 1992). In general, these measures do not require excessive computational cost for their calculation. However, according to results of recent studies, the majority of the vocabulary richness functions are highly text-length dependent and quite unstable for texts shorter than 1,000 words (Twedie and Baayen 1998).

2.4 Common Word Frequencies

Instead of using vocabulary distribution measures, some researchers counted the frequency of occurrence of individual words in the sample text. Such counts are a reliable discriminating factor (Karlgren and Cutting 1994; Kessler et al. 1997) and have been applied to many works in text genre detection. Their calculation is simple but non-trivial effort is required for the selection of the most appropriate words for a given problem. Morever, the words that best distinguish a given group of authors cannot be applied to a different group of authors with the same success (Holmes and Forsyth 1995). Oakman (1980) notes:

> *"The lesson seems clear not only for function words but for authorship word studies in general: particular words may work for specific cases such as 'The Federalist Papers' but cannot be counted on for other analyses."*

Furthermore, the results of such studies are highly language-dependent. Michos et al. (1996) introduces the idea of grouping certain words in categories such as idiomatic expressions, scientific terminology, formal words, etc. Although this solution is language independent, it requires the construction of a complicated computational mechanism for the automated detection of the categories in the sample text.

Alternatively, the use of sets (typically 30 or 50) of common high-frequency words has been applied mainly to authorship attribution studies (Burrows 1987). The application of a principal components analysis on the frequencies of occurrence of the most frequent words achieved remarkable results in plotting the texts in the space of the first two principal components, for a wide variety of authors (Burrows 1992). This approach is language independent and requires minimal computational cost. Various additional restrictions to this basic method have been proposed, e.g., separation of common homographic forms, removal of proper names from the most frequent word list, etc., aiming at the improvement of its performance. Regarding a fully-automated system, such restrictions require the employment of robust and accurate NLP tools.

3. The Proposed Method

Our method attempts to exploit already existing NLP tools for the extraction of the stylistic information. To this end, we use two types of measures as can be seen in figure 1:

- those that are relevant to the actual output of the NLP tool (i.e., usually tagged or parsed text), and
- those that are relevant to the particular methodology by which the NLP tool analyzes the text (analysis-level measures).

Thus, the set of style markers is adapted to a specific already-existing NLP tool taking into account its particular properties. Analysis-level measures capture useful stylistic information without additional cost. The NLP tool is not considered as a black box. Therefore, full access to its source code is required in order for one to define and measure analysis-level style markers. Moreover, tool-specific knowledge, rather than language-specific knowledge, is required for the definition of such measures. In other words, one can use this approach and define analysis-level measures based on his/her deep understanding of a particular NLP tool even he/she is not familiar with the natural language to which the methodology is to be applied.

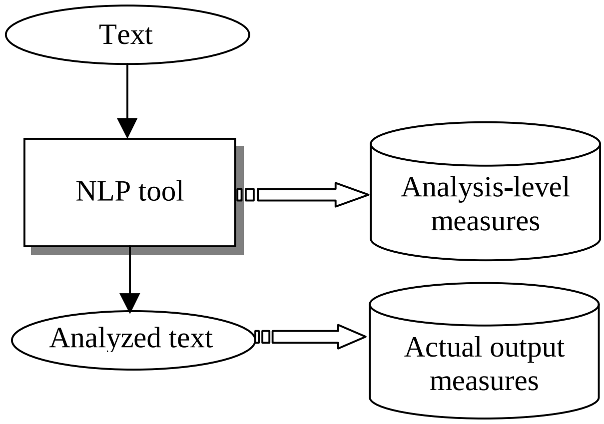

Fig. 1. The proposed method.

In order to illustrate the proposed method, we apply it to Modern Greek using SCBD, an already existing NLP tool able to detect sentence and chunk boundaries in unrestricted text, as described in the next section. In addition to a set of easily computable features (i.e., token-level and syntax-level measures) provided by the actual output of SCBD, we use a set of analysis-level features, i.e., measures that represent the way in which the input text has been analyzed by the SCBD.

The particular analysis-level style markers can be calculated only when this specific computational tool is utilized. However, SCBD is a general-purpose tool and was not designed for providing stylistic information exclusively. Thus, any NLP 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 (1999) as indicators of clausal complexity are the average parse tree depth and the number of parser skips per sentence which in essence are analysis-level style markers.

4. Style Markers for Modern Greek

As mentioned above, the subset of style markers used for Modern Greek depends on the text analysis by the specific NLP tool, the SCBD. Thus, before describing the set of style markers we used, we briefly present the main features of the SCBD in this section.

4.1 Description of SCBD

SCBD is a text-processing tool able to deal with unrestricted Modern Greek text. No manual preprocessing is required. It performs the following procedures:

- **Sentence boundary detection:** The following punctuation marks are considered as potential sentence boundaries: period, exclamation point, question mark, ellipsis. A set of automatically acquired disambiguation rules (Stamatatos et al. 1999) is, then, applied to every potential sentence boundary in order to locate the actual sentence boundaries. These rules utilize neither lexicons with specialized information nor abbreviation lists.
- **Chunk boundary detection Chunk boundary detection:** Intrasentential phrase detection is achieved through multiple-pass parsing making use of an approximately 450-keyword lexicon (i.e., closed-class words such as articles and propositions) and a 300-suffix lexicon containing the most common suffixes of Modern Greek words. Initially, using the suffix lexicon a set of morphological descriptions is assigned to any word of the sentence not included in the keyword lexicon. If the suffix of a word

does not match any of the entries of the suffix lexicon then no morphological description is assigned to this word. It is marked as a special word and is not ignored in subsequent analysis. Then, each parsing pass (five passes are performed) 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 are noun phrases (NPs), prepositional phrases (PPs), verb phrases (VPs), and adverbial phrases (APs). In addition, two chunks are usually connected by a sequence of conjunctions (CONs).

SCBD is able to cope rapidly with any piece of text, even ill formed, and its performance is comparable to more sophisticated systems that require more complicate resources. Figure 2 gives an overview of SCBD. An example of its output for a sample text together with a rough English translation (included in parantheses) is given below (note that special words, those that do not match with any of the stored suffixes, are marked with an asterisk):

VP[*Δεν θέλω να ρίξω* (I don’t want to pour)] NP[*λάδι* (oil)] PP[*στη φωτιά* (in the fire)] CON[*αλλά* (but)] VP[*πιστεύω* (I believe)] CON[*ότι* (that)] NP[*η επιβάρυνση* (the encumbrance)] PP[*στον προϋπολογισμό* (of the budget)] PP[*από τους βουλευτές* (by the deputies)] VP[*δεν μπορεί να προσμετρείται* (can not be measured)] *μόνο* (merely) PP[*με τα 5 *δισ. *δρχ. των αναδρομικών* (with the 5 bil. Dr. of the retroactive salaries)] *που* (that) NP[*πήραν τελευταία* (they took lately)] VP[*προκαλώντας* (causing)] NP[*τη δυσφορία της κοινής γνώμης* (the discontent of the public oppinion)] .

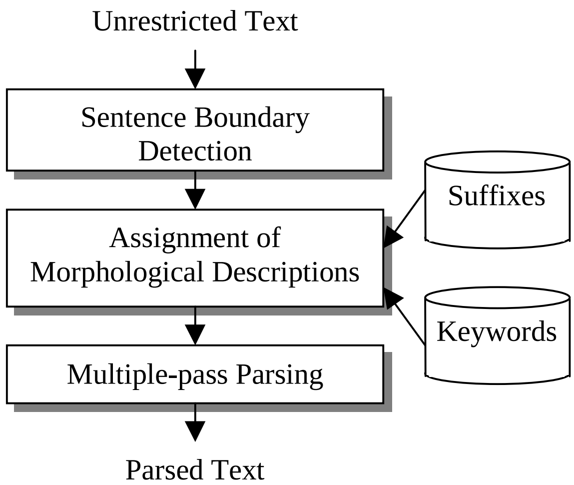

Fig. 2. The SCBD structure.

Regarding the calculation of the style markers described in the next section, it is worth noting that no modification has been performed to the structure of SCBD aside from adding simple functions for their measurement.

4.2 Stylometric-levels

Our aim during the definition of the set of style markers was to take full advantage of the analysis of the text by the SCBD. To this end, we included measures relevant to the actual output of this tool as well as measures relevant to the methodology used by SCBD in order to analyze the text. Specifically, the proposed set of style markers comprises three levels:

- **Token-level:** The sample text is considered as a set of tokens grouped in sentences. This level is based on the output of the sentence boundary detector:

Code	*Description*
M01	detected sentences / words[1]
M02	punctuation marks / words
M03	detected sentences / potential sentence boundaries

- **Phrase-level:** The sample text is considered as a set of phrases (i.e., chunks). This level is based on the output of the chunk boundary detector:

Code	*Description*
M04	detected NPs / total detected chunks
M05	detected VPs / total detected chunks
M06	detected APs / 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 APs / detected APs
M12	words included in PPs / detected PPs
M13	words included in CONs / detected CONs

- **Analysis-level:** It includes measures that represent the way in which the sample text has been analyzed by the particular methodology of the SCBD:

Code	*Description*
M14	detected keywords / words

[1] We consider words as word-tokens

M15 special words / words

M16 assigned morphological descriptions / words

M17 chunks' morphological descriptions / total detected chunks

M18 words remaining unanalyzed after pass 1 / words

M19 words remaining unanalyzed after pass 2 / words

M20 words remaining unanalyzed after pass 3 / words

M21 words remaining unanalyzed after pass 4 / words

M22 words remaining unanalyzed after pass 5 / words

It is clear that the analysis level contains extremely useful stylistic information. For example, M14 and M5 are valuable markers, which indicate the percentage of high-frequency words and the percentage of unusual words included in the sample text, respectively. M16 is a useful indicator of the morphological ambiguity of the words and M17 indicates the degree in which this ambiguity has been resolved. Moreover, 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 that a great part of a syntactically complicated text would not be analyzed by them (e.g., high values of M18, M19, and M20 in conjunction with low values of M21 and M22). Similarly, a syntactically simple text would be characterized by low values of M18, M19, and M20.

Note that all the proposed style markers are produced as ratios of two relative measures in order for them to be stable over the text-length. However, they are not standardized.

5. Text Categorization

The methodology described in the previous section provides a vector of 22 variables for each text. For classifying automatically this vector into one group (either genre or author) various techniques are nowadays available, stemmed from multivariate statistics (e.g., discriminant analysis), neural networks, and machine learning (e.g., decision trees). Recently, Yang (1999) studied the performance of several classifiers according to their application to text categorization tasks and concluded that all the tested methods perform comparably when the training set comprises over 300 instances per category. On the other hand, when the number of positive training instances per category is small (less than ten) a regression-like method called linear least-squares fit and k-nearest neighbors outperform neural networks and naive Bayes classifiers (Yang and Liu 1999).

In the present paper we use two well-known techniques of multivariate statistics: *multiple regression* and *discriminant analysis*. The response of these techniques is very fast since they are based on the calculation of simple linear functions. Moreover, their training procedures do not require excessive time or computational cost.

5.1 Multiple Regression

Multiple regression predicts values for a group of *response* (dependent) variables from a collection of *predictor* (independent) variable values (Edwards 1979). The response is expressed as a linear combination of the predictor variables, namely:

$$y_i=b_0 + z_1b_{1i} + z_2b_{2i} +\ldots+ z_rb_{ri} + e_i$$

where y_i is the response for the *i*-th category (i.e., text genre), z_1, z_2,..and z_r are the predictor variables (i.e., in our case r=22), b_0, b_{1i}, b_{2i},…, and b_{ri}, are the unknown coefficients calculated during the training procedure, and e_i is the random error. An indication of the goodness of fit of the model is provided by the *coefficient of determination* R^2 defined as follows:

$$R^2 = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})}{\sum_{j=1}^{n} (y_j - \bar{y})}$$

where n is the total amount of the training data (texts), $\bar{y}$ is the mean response, and finally, $\hat{y}_j$ and y_j are the estimated response and the training response value respectively. R^2 equals 1 if the fitted equation passes through all the data points, and, at the other extreme, equals 0.

Moreover, multiple regression can also be used for the estimation of the significance of the independent variables. In particular, the amount by which R^2 is reduced if a certain independent variable is deleted from the regression equation (in other words, the contribution of the independent variable to R^2) is represented by the squared semipartial correlation sr_i^2 (Tabachnick and Fidell 1996):

$$sr_i^2 = \frac{t_i^2}{df_{res}}(1 - R^2)$$

where t_i is the value of the t-statistic for the i-th variable and df_{res} are the residual degrees of freedom. Thus, the contribution of an independent variable to R^2 can be expressed as a function of the absolute value of t. The absolute t value of the j-th estimated regression coefficient b_j is calculated as follows:

$$t_{b_j} = \frac{b_j}{S_{b_j}}$$

where S_{b_j} is the standard error. The greater the t value, the more important the contribution of the independent variable (i.e., style marker) to the response value.

5.2 Discriminant Analysis

The mathematical objective of discriminant analysis is to weight and linearly combine the discriminating variables 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 a training sample.

Then, discriminant analysis can be used for predicting the group membership of previously unseen cases (i.e., test data) based on *Mahalonobis distance* (i.e., a measure of distance between two points in the space defined by multiple correlated variables). Initially, 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 closest group. The Mahalanobis distance d of a vector $\boldsymbol{x}$ from a mean vector $\boldsymbol{m}_x$ is given by the formula:

$$d^2 = (\boldsymbol{x} - \boldsymbol{m}_x)' C_x^{-1} (\boldsymbol{x} - \boldsymbol{m}_x)$$

where C_x is the covariance matrix of $\boldsymbol{x}$. 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. Discriminant analysis has been employed by researchers in automatic text genre detection (Biber 1993b; Karlgren and Cutting 1994) since it offers a simple and robust solution despite the fact that it presupposes normal distributions of the discriminating variables.

6. Text Genre Detection

6.1 Genre-based Corpus

Since no Modern Greek corpus covering a wide range of text genres was available we decided to compose one from scratch. The corpus used in experiments in (Michos et al. 1996) includes a limited number of carefully-selected and manually-edited texts divided in generic categories (e.g., journalistic, scientific, etc.). In general, the use of already-existing corpora not built for text genre detection (e.g., the *Brown* corpus) raise several problems since such categories may not be stylistically homogeneous (Kessler et al. 1997). The corpus used in our study contains texts that meet the following criteria:

- *Real-world text:* The texts have to be already in electronic form and due to this fact may be ill-formed.
- *Raw text:* Neither manually inserted tags nor other manual text-preprocessing restrictions are set.
- *Whole text:* Neither text-length limitations nor other manual text-sampling restrictions are set. In other words, a text has to be available as it appears in its source.

Code	Text genre	Texts	Words (average)	Source
G01	Press editorial	25	729	Newspaper *TO BHMA*
G02	Press reportage	25	902	Newspaper *TO BHMA*
G03	Academic prose	25	2,120	Journal of *ARCHIVES OF HELLENIC PATHOLOGY*
G04	Official documents	25	1,059	High Court decisions, Ministerial decisions
G05	Literature	25	1,508	Various pages
G06	Recipes	25	109	Magazine *NETLIFE*
G07	Curricula vitae	25	333	Various pages
G08	Interviews	25	2,625	Newspaper *TO BHMA*
G09	Planned speeches	25	2,569	Ministry of defense
G10	Broadcast news, scripted	25	137	Radio station *FLASH 9.61*

Table 1. The genre-based corpus.

Thus, we constructed a corpus by downloading texts from various WWW sites edited in Modern Greek trying to cover as many genres as possible. This corpus is shown in table 1. Notice that the complete set of text genres may differ significantly among two languages (Biber 1995). Nevertheless, they usually overlap to a great extent, especially for Indo-European languages. The set we propose, therefore, can be compared to the ones used in similar studies of English (Karlgren and Cutting 1994; Biber 1995). Additionally, no manual preprocessing was performed aside from removing unnecessary headings irrelevant to the text itself.

It has also to be underlined that the last three text genres (i.e., G08, G09, and G10) refer to spoken language that has been transcribed either before (i.e., planned speeches, broadcast news) or after (i.e., interviews) it has been uttered. On the other hand, G01 to G07 refer to written language.

The genre-based corpus was divided into a training and a test part of equal size. Ten texts per genre were included in the training corpus and ten texts per genre were included in the test corpus. The remaining five texts per genre were used only in the experiments of the Section 7.

6.2 Setting the Baseline

In order to evaluate the contribution of the proposed approach we decided to apply two previous stylometric approaches that are based on distributional lexical measures to the same testing ground: (i) a multivariate model of functions of vocabulary richness (Holmes 1992) and (ii) the frequencies of occurrence of the most frequent words (Burrows 1992). These two methods were selected since they are language-independent and require minimal computational cost.

For measuring the richness of the vocabulary we used a set of five functions, namely: *K* proposed by Yule (1944), *R* proposed by Honore (1979), *W* proposed by Brunet (1978), *S* proposed by Sichel (1975), and *D* proposed by Simpson (1949) which are defined as follows:

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

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

$$W = N^{V^{-\alpha}}$$

$$S = \frac{V_2}{V}$$

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

where V_i is the number of words used exactly i times (see section 2.3 for the definition of V and N) and α is a parameter usually fixed at 0.17. The same set of functions has been used by Baayen and his colleagues for similar purposes (Baayen et al. 1996). For every text these functions are calculated and a vector of five parameters is produced. These vectors can then be classified to the most likely genre by applying one of the classification techniques discussed in the previous section. Hereafter, this approach will be called VR (standing for vocabulary richness).

The second lexically-based method uses as style markers the frequencies of occurrence of the most frequent words of the training corpus. Typically, sets of 30 or 50 most frequent words are used (Baayen et al. 1996; Holmes and Forsyth 1995). For comparison purposes we employed two sets of common words based on 30 and 50 most frequent words of the training corpus, respectively. Thus, for each text a vector of 30 (or 50) parameters indicating the frequencies of the most frequent words of the training corpus (normalized by the text-length) are calculated. As above, these vectors can then be classified to the most likely genre. Hereafter, these two approaches will be called CWF-30 and CWF-50 (standing for common word frequencies and the number of the high frequency words), respectively.

6.3 Results

The entire corpus described in the previous section was analyzed by SCBD, which provided a vector of 22 parameters for each text automatically. The vectors of the training corpus were used in order to extract the classification model using both multiple regression and discriminant analysis. These classification models were then applied to the vectors of the test corpus for cross-validating their performance on unseen cases. The same training and test procedure was performed for the VR approach as well as for the CWF-30 and CWF-50 methods.

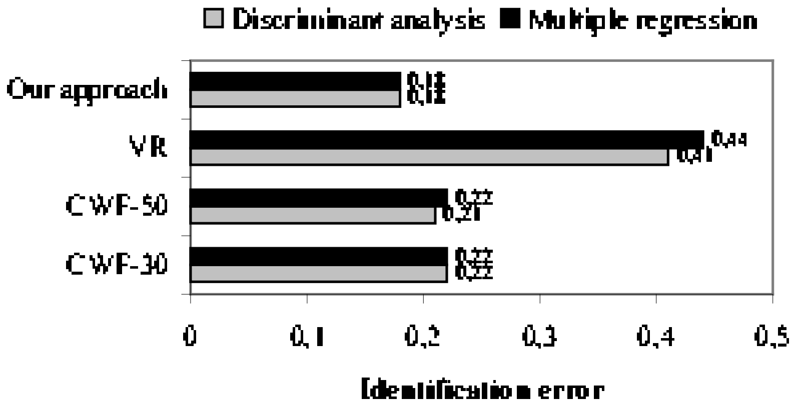

Fig. 3. Comparative results for text genre detection

Code	Identification error	
	Multiple regression	**Discriminant analysis**
G01	0.7	0.4
G02	0.2	0.1
G03	0.0	0.0
G04	0.1	0.2
G05	0.1	0.4
G06	0.0	0.0
G07	0.4	0.4
G08	0.1	0.0
G09	0.2	0.2
G10	0.0	0.1
Average	0.18	0.18

Table 2. The text genre detection results.

Comparative results in terms of identification error (i.e., erroneously classified texts / total texts) are given in figure 3. In general, discriminant analysis seems to be better able to distinguish the texts of the test corpus. The performance of the VR approach is quite poor. This is due to the limited text-length of the majority of the texts of the genre-corpus

(Tweedie and Baayen 1998). Moreover, our approach is more accurate than the CWF-30 and the CWF-50.

The identification error of our approach using both multiple regression and discriminant analysis is given in table 2. Although the average error is equal for the two methodologies there are significant differences in the disambiguation accuracy of certain text genres (see G01 and G05). In general, the error is more normally distributed using discriminant analysis. Moreover, approximately 60% of the identification error using multiple regression was caused by G01 and G07 while 65% of the identification error using discriminant analysis was caused by G01, G05, and G07. On the other hand, G04, G06, G08, and G10 are stylistically homogeneous to a great extend in both cases.

The complete identification results of our method using discriminant analysis are presented in the confusion matrix of the table 3. Each row represents a text genre under test and the columns represent the classification results of the test texts of that particular genre. The main misclassifications are as follows:

- *press editorial* → *press reportage*. Notice that the texts were taken from the same newspaper that is published on a weekly basis. In many cases therefore the reportage documents are characterized by reviewing a whole week and presenting some comments by the author.
- *curricula vitae* → *official documents*. Both are usually characterized by an abstract style.
- *literature* → *interviews* and *planned speeches*, in other words, two text genres of the spoken language that usually involve narration.

Actual	Classification										Total texts
	G01	**G02**	**G03**	**G04**	**G05**	**G06**	**G07**	**G08**	**G09**	**G10**	
G01	**6**	3	0	0	0	0	0	0	1	0	10
G02	0	**9**	0	0	0	0	0	0	0	1	10
G03	0	0	**10**	0	0	0	0	0	0	0	10
G04	0	1	0	**8**	0	0	0	0	0	1	10
G05	0	0	0	0	**6**	0	0	2	2	0	10
G06	0	0	0	0	0	**10**	0	0	0	0	10
G07	0	0	0	3	0	0	**6**	0	0	1	10
G08	0	0	0	0	0	0	0	**10**	0	0	10
G09	1	0	0	0	0	0	0	1	**8**	0	10
G10	0	0	0	1	0	0	0	0	0	**9**	10

Table 3. Confusion matrix for text genre detection using discriminant analysis.

Note that spoken language text genres (i.e., G08-G10) have a lower identification error rate, on average (0.10), than written language text genres (0.21) as calculated by either multiple regression or discriminant analysis.

The classification accuracy of our method related to the text-length for the text genre experiment using multiple regression is presented in figure 4. Due to the stylistic homogeneity of *recipes* and *broadcast news* the accuracy of texts shorter than 500 words (see table 1) is relatively high. In addition, texts over 1,500 words seem to be classified more reliably. Note that according to Biber (1990; 1993a) a text-length of 1,000 words is adequate for representing the distributions of many core linguistic features of a stylistic category.

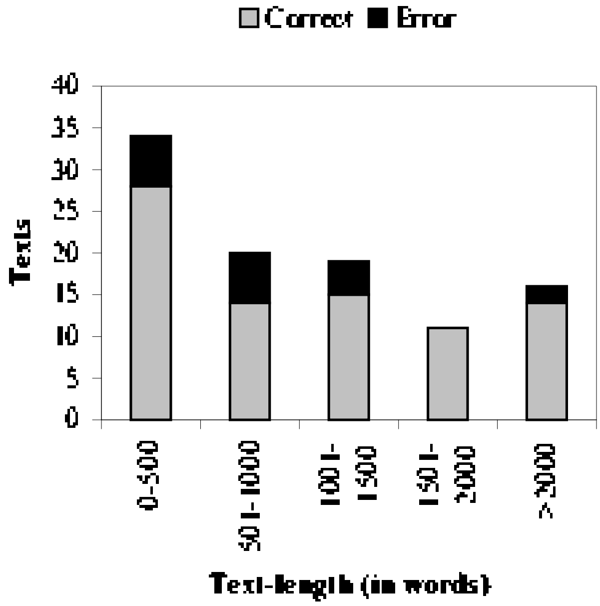

Fig. 4. Text-length related to accuracy for the text genre detection experiment.

7. Authorship Attribution

7.1 Author-based Corpus

In authorship attribution experiments we chose to deal with texts taken from newspapers since a wide variety of authors publish frequently their writings in the press. The collection of a considerable amount of texts of several authors was, therefore, easier. In particular, the corpus used for this study comprises texts downloaded from the WWW site of the Modern Greek weekly newspaper *TO BHMA*, (the tribune[2]). The structure of this newspaper is shown in table 4. We performed experiments based on two groups of authors, namely:

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

1. **Group A:** It consists of ten randomly selected authors whose writings are frequently found in section A. This section comprises texts written mainly by journalists on a variety of current affairs. Moreover, a certain author may sign texts from different text genres (e.g., editorial, reportage, etc.). Note that in many cases such writings are highly edited in order to conform to a predefined style, thus washing out specific characteristics of the authors which complicate the task of attributing authorship.
2. **Group B:** It consists of ten randomly selected authors whose writings are frequently found in section B. This supplement comprises essays on science, culture, history, etc., in other words, writings in which the idiosyncratic style of the author is not overshadowed by functional objectives. In general, the texts included in the B section are written by scholars, rather than by journalists.

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

Table 4. The structure of the Modern Greek weekly newspaper *ΤΟ ΒΗΜΑ*.

Analytical information on the author-based corpus can be found in table 5. All the downloaded texts were taken from issues published within 1998 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. The author-based corpus was divided into a training and a test part of equal size (i.e., ten texts per author for training and ten texts per author for test).

Group	Code	Author name	Texts	Words (average)	Thematic area
A	A01	N. Nikolaou	20	797	Economy
	A02	N. Marakis	20	871	International affairs
	A03	D. Psychogios	20	535	Politics
	A04	G. Bitros	20	689	Politics, society
	A05	D. Nikolakopoulos	20	1,162	Politics, society
	A06	T. Lianos	20	696	Society
	A07	K. Chalbatzakis	20	1,061	Technology
	A08	G. Lakopoulos	20	1,248	Politics
	A09	R. Someritis	20	721	Politics, society
	A10	D. Mitropoulos	20	888	International affairs
B	B01	D. Maronitis	20	589	Culture, society
	B02	M. Ploritis	20	1,147	Culture, history
	B03	K. Tsoukalas	20	1,516	International affairs
	B04	C. Kiosse	20	1,741	Archaeology
	B05	S. Alachiotis	20	958	Biology
	B06	G. Babiniotis	20	1,273	Linguistics
	B07	T. Tasios	20	1,049	Technology, society
	B08	G. Dertilis	20	916	History, society
	B09	A. Liakos	20	1,291	History, society
	B10	G. Vokos	20	1,002	Philosophy

Table 5. The author-based corpus.

7.2 Author Identification

Similarly to the text genre detection experiment, the entire corpus was analyzed automatically by the SCBD and, afterwards, we used the stylistic vectors of the training corpus in order to train the classification model for each group separately based on multiple regression as well as discriminant analysis. The acquired models were then cross-validated by applying them to the test corpus of the corresponding group. The same procedure was followed based on the VR, CWF-30, and CWF-50 approaches. Comparative results in terms of the identification error for the groups A and B are given in figures 5 and 6, respectively. As in the case of text-genre detection, the VR method achieved far lower accuracy results than the others. The performance of the CWF-30 and CWF-50 is significantly better in group B in comparison with that of group A. In both groups our approach achieved the best performance.

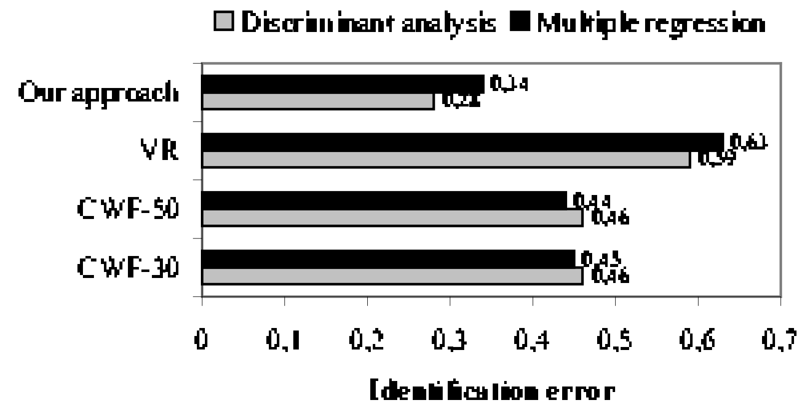

Fig. 5. Comparative results for authorship identification in g*roup A.*

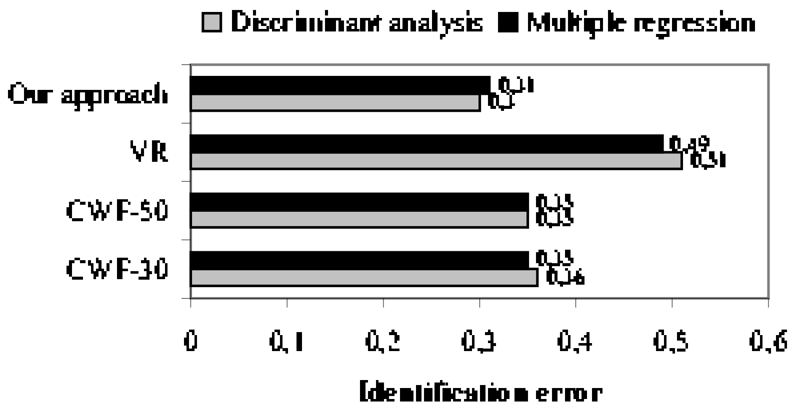

Fig. 6. Comparative results for authorship identification in g*roup B.*

The identification error rates of our approach using both multiple regression and discriminant analysis are presented in Table 6. For group A, there are significant differences in the accuracy of the two techniques. However, three authors (A01, A03, and

A06) are responsible for approximately 50% of the average error rate, probably because the average text length of these authors is relatively short, i.e., shorter than 800 words (see table 5).

Code	Identification error		Code	Identification error	
	Multiple regression	**Discriminant analysis**		**Multiple regression**	**Discriminant analysis**
A01	0.5	0.4	B01	0.7	0.6
A02	0.3	0.2	B02	0.0	0.0
A03	0.6	0.5	B03	0.2	0.4
A04	0.2	0.1	B04	0.1	0.1
A05	0.3	0.3	B05	0.7	0.4
A06	0.7	0.5	B06	0.3	0.3
A07	0.3	0.3	B07	0.0	0.1
A08	0.1	0.1	B08	0.6	0.6
A09	0.2	0.3	B09	0.1	0.1
A10	0.2	0.1	B10	0.4	0.4
Average	0.34	0.28	Average	0.31	0.30

Table 6. The author identification results for both group A and group B.

On the other hand, both techniques give similar disambiguation results as regards group B. A considerable percentage of the average error is caused by the authors B01, B05, and B08 (i.e., 65% for multiple regression, 55% for discriminant analysis). As above, these authors have a relatively short text-length in average, i.e., shorter than 1,000 words.

It seems, therefore, that text length is a crucial factor in identifying the stylistic features that characterize a certain author. Classification accuracy for both groups using multiple regression related to text length is presented in more detail in Figure 7. Approximately 80% (i.e., 53 out of 65) of the total erroneously classified texts are shorter than 1,000 words.

Moreover, the accuracy results for the two groups are comparable. In fact, the best results have been achieved under discriminant analysis for group A. This fact verifies that the proposed set of style markers is capable of capturing the underlying stylistic features that characterize the author of a text even in the case of dealing with texts taken from various text genres. Note that CWF-30 and CWF-50 failed to achieve comparable performance for groups A and B.

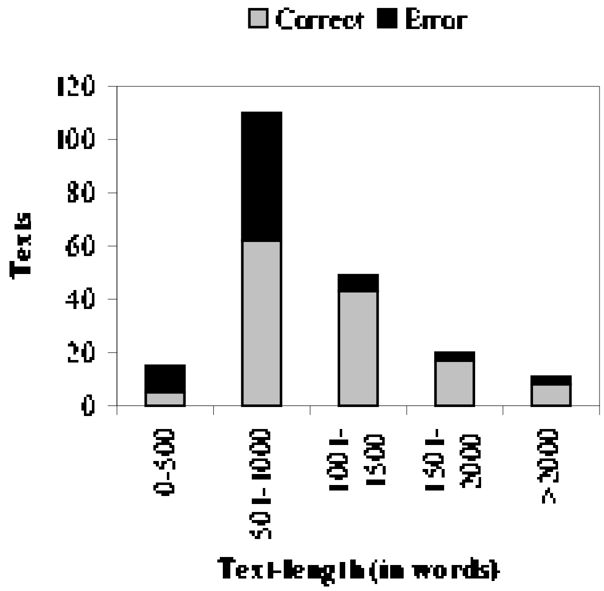

Fig. 7. Text-length related to accuracy for the author identification experiments.

7.3 Author Verification

Instead of trying to identify the most likely author of a given text taking into account a given group of authors (i.e., the *author identification* problem), a lot of applications require the verification of the hypothesis whether or not a given person is the author of the text (i.e., the *author verification* problem). In such cases, the classification procedure is less complicated since the possible answers are only two: *yes*, i.e., the author in question is indeed the person who wrote the text, or *no*, i.e., the text was not written by this person.

The implementation of an automatic author verification system has two main tasks. In particular, it requires:

- The development of a *response function* for a given author. As regards a certain text, this function must provide a response value based on the vector of the style markers of the text.
- The definition of a *threshold value*. Any text whose response value is greater than the one of the threshold is accepted as written by the author in question. Otherwise it is rejected.

Additionally, for measuring the accuracy of an author verification method as regards a certain author, *False Rejection* (FR) and *False Acceptance* (FA) can be used. These measures are widely utilized in the area of speaker verification in speech processing (Fakotakis et al. 1993) and they are defined as follows:

FR = rejected texts of the author / total texts of the author

FA = accepted texts of other authors / total texts of other authors

In our study, we used the response functions taken from the application of multiple regression to the group A and group B, as described in the previous section. On the other hand, the selection of a threshold value depends highly on the application. Some applications require either minimal FR or minimal FA while others require minimal *mean error*, i.e., (FR+FA)/2.

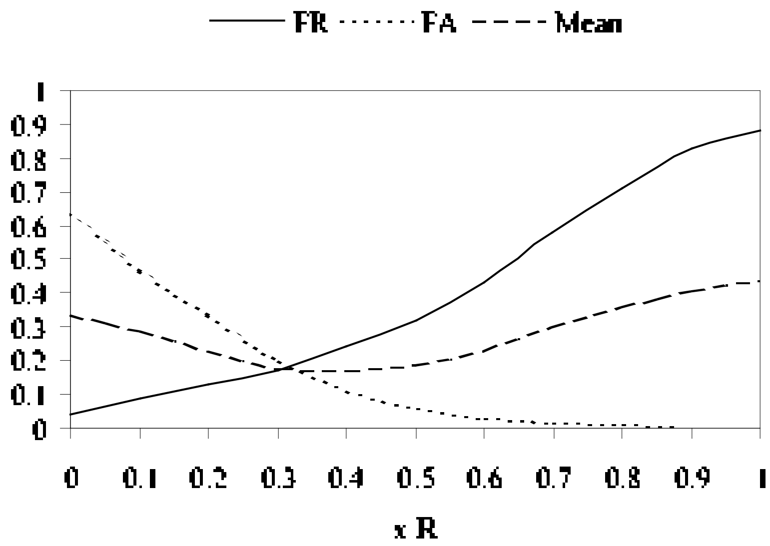

Fig. 8. FR, FA, and mean error for *group A* related to threshold values expressed as subdivisions of *R*.

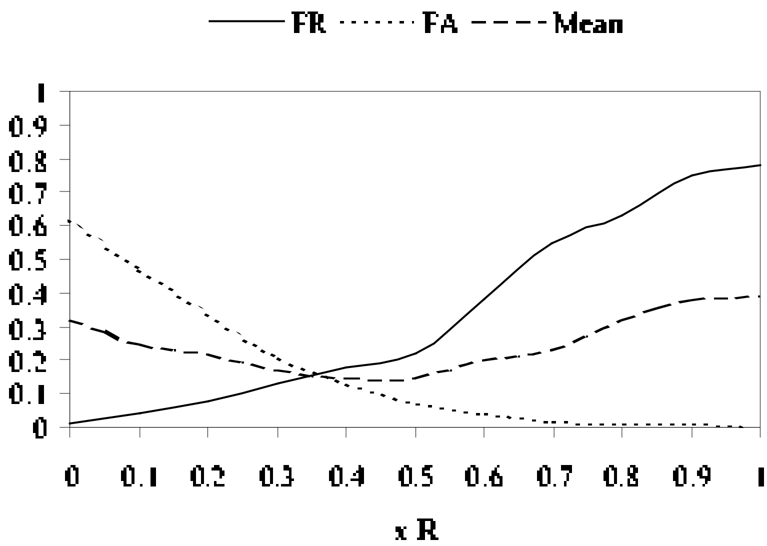

Fig. 9. FR, FA, and mean error for *group B* related to threshold values expressed as subdivisions of *R*.

We chose to express the threshold value as a function of the *multiple correlation coefficient* $R = +\sqrt{R^2}$ of the regression functions (see Section 5.1) since it measures the degree in which the regression function fits the training data. It equals 1 if the fitted equation passes through all the data points and at the other extreme, equals 0, as already mentioned for R^2. Figures 8 and 9 depict the variation of the average FR, FA, and the mean error values for the test corpus of group A and group B, respectively, using various subdivisions of R as threshold. Notice that the presented evaluation used texts within the same group of authors for testing (i.e., closed-set evaluation). Low values of threshold correspond to minimal FR while high values of threshold correspond to minimal FA. As can be seen, the minimal mean error corresponds to threshold values between 0.4R and 0.5R for both groups.

Code	***R*/2**	**FR**	**FA**	**Code**	***R*/2**	**FR**	**FA**
A01	0.33	0.5	0.033	B01	0.32	0.3	0.022
A02	0.33	0.3	0.011	B02	0.42	0.0	0.044
A03	0.36	0.6	0.044	B03	0.33	0.0	0.155
A04	0.36	0.2	0.111	B04	0.33	0.1	0.089
A05	0.35	0.3	0.067	B05	0.28	0.6	0.144
A06	0.35	0.7	0.044	B06	0.36	0.2	0.011
A07	0.34	0.2	0.044	B07	0.38	0.0	0.022
A08	0.31	0.1	0.111	B08	0.30	0.6	0.100
A09	0.35	0.2	0.055	B09	0.36	0.0	0.055
A10	0.35	0.1	0.089	B10	0.40	0.4	0.033
Average	0.35	0.32	0.061	Average	0.35	0.22	0.068

Table 7. The author verification results for both groups (threshold=$R/2$).

The FR and FA values for group A and group B using 0.5R as threshold are given in table 7. The greatest part of the total FR in both groups accounts for the authors characterized by short text-length (i.e., group A: A01, A03, and A06, group B: B01, B05, and B08) as in the case of author identification. On the other hand, FA seems to be highly relevant to the threshold value. The smaller the threshold value, the greater the false acceptance.

8. Performance Issues

8.1 Training Set Size

Our study makes use of ten training texts from each category (i.e., either a text genre or an author) in order to extract the appropriate coefficients. This assessment meets the criteria of a system that requires easy adaptation of the text-categorization methodology to a certain domain. Note that Biber (1990; 1993a) claims that it is possible to represent the distributions of many core linguistic features of a stylistic category based on relatively few texts from each category, i.e., as few as ten texts. However, it is interesting to explore the way in which the identification error is affected by increasing the training data.

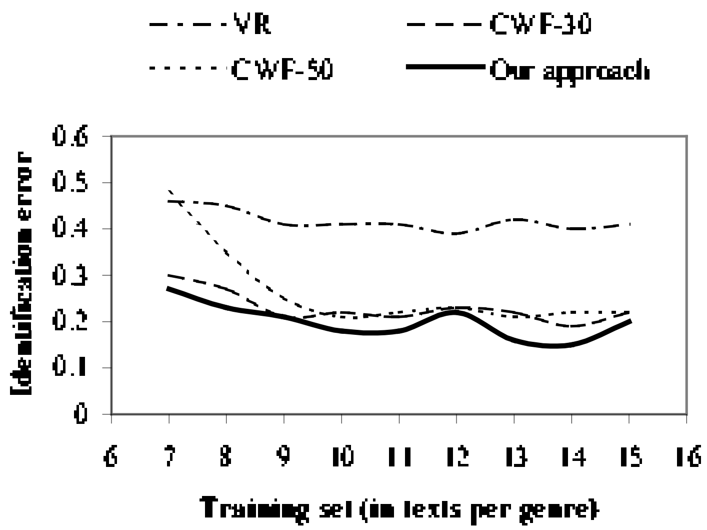

Fig. 10. The identification error of the text genre detection experiment related to the training set size.

To this end, we performed experiments on text genre detection using multiple regression based on variable training data. Specifically, we varied the training corpus, including 7 to 15 texts for each genre, but used the same test corpus of 10 texts for all of the experiments. The same procedure was followed for the lexically-based approaches VR, CWF-30, and CWF-50. Comparative results of the average identification error related to the training set size are shown in figure 10.

The performance of the VR is not affected significantly by increasing the training set size. On the other hand, the identification error of CWF-30, CWF-50, and that of our approach is generally reduced by increasing the number of texts used for training. The performance of CWF-30 is more stable in comparison to CWF-50 but is lower than the one achieved by our set of style markers.

The best results are achieved by our approach using 14 training texts per genre (i.e., only 15 out of 100 texts misclassified). However, the identification error is not continuously decreasing from 11 to 15 training texts. Thus, the identification error using 12 as well as 15 training texts for each category is greater than the one attained by using 10 texts. It is clear that a quite satisfactory accuracy can be achieved with only ten training texts.

8.2 Significance of Style Markers

The proposed set of style markers is divided into three levels, namely: token-level, syntax-level, and analysis-level. Thus, it would be useful to calculate the contribution of each marker, and consequently of each level, to the classification procedure. To this end, we used the absolute *t* values of the linear regression functions that indicate the contribution of each independent variable to the response value (see Section 5.1).

The average absolute t values of the 22 style markers, taking into account the regression functions for both text genre and author identification experiments, are presented in Table 8. In both cases, the most important stylometric level is the token-level, while the syntax-level contribute the least to the final response. On the other hand, M02, M12, and M15 are the most important style markers for text genre detection (i.e., average t > 1.50) while the token-level measures M01, M02, and M03 are the most valuable measures for authorship attribution (for the specific groups of authors).

Stylometric level	**Style marker**	**Absolute *t* values (average)**	
		Text genre detection	**Authorship attribution**
Token-level	M01	1.06	1.80
	M02	2.52	1.85
	M03	1.43	1.98
	Level average	1.67	1.88
Syntax-level	M04	0.57	0.76
	M05	0.58	0.77
	M06	0.56	0.77
	M07	0.57	0.75
	M08	0.57	0.76
	M09	0.77	0.98
	M10	0.93	0.85
	M11	0.59	0.90
	M12	1.72	1.07
	M13	0.67	0.97
	Level average	0.75	0.86
Analysis-level	M14	1.03	1.30
	M15	2.11	1.05
	M16	1.45	0.79
	M17	1.08	1.42
	M18	0.72	1.06
	M19	1.14	0.84
	M20	1.00	0.86
	M21	0.81	0.90
	M22	0.65	0.84
	Level average	1.11	1.01

Table 8. Absolute *t* values (in average) for the regression functions of both text genre detection and authorship attribution.

8.3 Defective Computational Analysis

The set of style markers is provided by the SCBD, an existing computational tool. To explore the degree to which the accuracy results are dependent on the accuracy of the

SCBD, we created an artificial defect in the output of the SCBD by corrupting the sentence and chunk boundary detection procedures:

- regarding the sentence boundary detection procedure, only periods were considered as denoting a potential sentence boundary, and
- the fifth parsing pass was excluded from the chunk boundary detection procedure.

These changes decreased significantly the accuracy of the output of the SCBD. We performed again the text genre experiment using multiple regression based on the defective data. The average identification error was increased approximately 25% (i.e., new identification error=0.23). As it was expected, the accuracy of the text-categorization methodology strongly depends on the accuracy of the SCBD.

Stylometric level	**Absolute *t* (average)**	
	Regular analysis	**Defective analysis**
Token-level	1.67	1.55
Syntax-level	0.75	0.97
Analysis-level	1.11	1.29

Table 9. Absolute values of *t* (in average) of the stylometric levels for both regular and defective analysis.

However, it has to be underlined that the contribution of the stylometric-levels to the final response has also been changed. Table 9 shows average absolute *t* values for both the regular and the defective computational analysis. Although the token-level measures are still the most important ones as regards their contribution to the response, the disproportion between them and both the analysis-level and the syntax-level measures has considerably decreased.

9. Conclusions

In this paper we presented an approach to text categorization in terms of stylistically homogeneous categories, either text genres or authors. The results of applying this methodology to text genre detection and author identification/verification experiments are strongly encouraging and outperform existing lexically-based methods. Since the stylistic differences are clearer among text genres, the results achieved in text genre detection are considerably better than the ones of authorship attribution tasks. However, in both cases a limited number of text genres or authors are responsible for the greatest part of the identification error.

As it can be seen in figures 4 and 7 text-length plays an important role especially in the case of author identification. A lower boundary of 1,000 words for each text seems reasonable for assuring improved performance. Nevertheless, when dealing with real-world text, this lower bound is not always possible to reach. The corpora used in all the presented experiments consist of real-world texts downloaded from the Internet without any manual text preprocessing or text sampling limitations. The majority of these texts have an average text-length shorter than 1,000 words.

As shown in the presented experiments, our method can be applied to a randomly-selected group of stylistically homogeneous categories without any manual adaptation restriction. A training corpus consisting of ten texts per category is adequate for achieving relatively high classification accuracy. We attempted to take advantage of existing NLP tools by using analysis-level style markers that provide useful stylistic information without any additional cost. In essence such measures represent the way in which the text has been analyzed by the computational tool. We proved that these measures are more important, as regards their contribution to the final response, than the ones related to the actual output of the tool on the syntactic level (see table 8).

Much work remains to be done as regards the stylistic interpretation of the acquired results and the automatic extraction of stylistic conclusions related to the text itself as well as its author. Such stylistic conclusions could explain the differences and the similarities among various genres or authors on a formal basis. Moreover, the definition of a basic text-length unit would open the way to the exploration of the variation of style within a single text. This procedure could assist to the detection of certain sections of the input text where the useful information is more likely to be included. We believe that such tasks can be performed by using a set of style markers similar to the proposed one. Finally, the combination of our approach with lexically-based methods, such as CWF-30, can result to a very reliable text categorization system in terms of stylistically homogeneous categories.

Acknowledgement

We would like to thank the anonymous CL reviewers for their valuable and insightful comments. Their suggestions have greatly improved an earlier draft of this paper.

References

Baayen, Harald; Van Halteren, Hans; and Tweedie Fiona. (1996). Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3): 121-131.

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

Biber, Douglas. (1993a). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8: 1-15.

Biber, Douglas. (1993b). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, 19(2): 219-242.

Biber, Douglas. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge University Press.

Brainerd, Barron. (1974). *Weighting Evidence in Language and Literature: A Statistical Approach.* University of Toronto Press.

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

Brunet, Ettienne. (1978). *Vocabulaire de Jean Giraudoux: Structure et Evolution.* Slatkine.

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

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

Edwards, Allen F. (1979). *Multiple Regression and the Analysis of Variance and Covariance.* San Francisco, CA: W. H. Freeman.

Eisenbeis, Robert A.; and Avery, Robert B. (1972). *Discriminant Analysis and Classification Procedures: Theory and Applications.* Lexington, Mass.: D.C. Health and Co.

Fakotakis, Nikos; Tsopanoglou, Anastasios; and Kokkinakis, George (1993). A Text-independent Speaker Recognition System Based on Vowel Spotting. *Speech Communication*, 12: 57-68.

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

Holmes, David I. (1994). Authorship Attribution. *Computers and the Humanities*, 28: 87-106.

Holmes, David I.; and Forsyth, Richard S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2): 111-127.

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

Karlgren, Jussi; and Cutting, Douglass. (1994). Recognizing text Genres with Simple Metrics Using Discriminant Analysis. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, 1071-1075.

Karlgren, Jussi. (1999). Stylistic Experiments in Information Retrieval. In *Natural Language Information Retrieval*, edited by T. Strzalkowski, Kluwer Academic Publishers, 147-166.

Kessler, Brett; Nunberg, Geoffrey; and Schutze, Hinrich. (1997). Automatic Detection of Text Genre. In *Proc. of 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97)*, 32-38.

Michos, Stefanos; Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George. (1996). An Empirical Text Categorizing Computational Model Based on Stylistic Aspects. In *Proc. of the 8th Conference on Tools with Artificial Intelligence (ICTAI'96)*, 71-77.

Morton, Andrew Q. (1965). The Authorship of Greek Prose. *Journal of the Royal Statistical Society, Series A*, 128: 169-233.

Mosteller, Fredrick; and Wallace, David. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Addison-Wesley, Reading, MA.

Oakman, Robert L. (1980) *Computer Methods for Literary Research.* Columbia, University of South Carolina Press.

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

Simpson, Edward H. (1949). Measurement of Diversity. *Nature*, 163: 688.

Smith, M.W.A. (1983). Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Literary and Linguistic Computing Bulletin*, 11: 73-82.

Smith, M.W.A. (1985). An Investigation of Morton's Method to Distinguish Elizabethan Playwrights. *Computers and the Humanities*, 19(1): 3-21.

Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George. (1999). Automatic Extraction of Rules for Sentence Boundary Disambiguation. In *Proc. of the Workshop in Machine Learning in Human Language Technology*, *Advance Course on Artificial Intelligence (ACAI'99),* 88-92.

Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George. (2000). A Practical Chunker for Unrestricted Text. In *Proc. of the 2nd Int. Conf. on Natural Language Processing*, 139-150.

Strzalkowski, Tomek. (1994). Robust Text Processing in Automated Information Retrieval. In *Proc. of the 4th Conf. On Applied Natural Language Processing (ANLP'94),* 168-173.

Tabachnick, Barbara G.; and Fidell, Linda S. (1996). *Using Multivariate Statistics* (third edition). HarperCollins College Publishers.

Tweedie, Fiona; Singh, Sameer; and Holmes, David I. (1996) Neural Network Applications in Stylometry: The Federalist Papers. *Computers and the Humanities,* 30(1): 1-10.

Tweedie, Fiona; and Baayen, Harald. (1998). How Variable may a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5): 323-352.

Yang, Yiming. (1999). An Evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, 1(1): 69-90.

Yang, Yiming; and Liu, Xin (1999). A re-examination of text categorization methods. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 42-49.

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