Automatic Text Categorization in Terms of Genre and Author

Efstathios Stamatatos*
University of Patras

Nikos Fakotakis†
University of Patras

George Kokkinakis‡
University of Patras

The two main factors that characterize a text are its content and its style, and both can be used as a means of categorization. In this paper we present an approach to text categorization in terms of genre and author for Modern Greek. In contrast to previous 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, and author verification tasks and show that the proposed method performs better than 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 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 extract useful stylistic information using existing text processing tools.

1. Introduction

The rapid expansion of the World Wide Web (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 or application or generality in order to cover a wide range of domains or applications.

* University of Patras, Department of Electrical & Computer Engineering, 26500 Patras, Greece. E-mail: stamatatos@wcl.ee.upatras.gr.
† University of Patras, Department of Electrical & Computer Engineering, 26500 Patras, Greece. E-mail: fakotaki@wcl.ee.upatras.gr.
‡ University of Patras, Department of Electrical & Computer Engineering, 26500 Patras, Greece. E-mail: gkokkin@wcl.ee.upatras.gr.

The two main factors that characterize a text are its content and its style, both of which can be used for categorization purposes. 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 contrast 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 the stylostatisticians, a given style is defined as a set of measurable patterns, called **style markers**. We adopt this definition in this study.

Typical classificatory tasks in computational stylistics are the following:

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

1. **Extraction of style markers:** A set of quantifiable measures are defined and a text-processing tool is usually developed, to automatically count them.
2. **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 have focused on the use of simple measures that can be easily detected and reliably counted by a computational tool (Kessler, Nunberg, and Schütze 1997). To this end, various sets of style markers have been proposed (Karlgren and Cutting 1994), all of which are, in essence, subsets of the set used by Biber (1995), who ranked 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 using existing text processing tools rather than taking advantage of them.

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 thousand 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) 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" (p. 87).

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 requiring human assistance in 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 that 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 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.

To illustrate, we apply the proposed technique to text categorization tasks for Modern Greek corpora using an already existing sentence and chunk boundaries detector (SCBD) in unrestricted Modern Greek text (Stamatatos, Fakotakis, and Kokkinakis 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 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. 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 already 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 used 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, namely, those that reflect the style of the writing. Various sets have been proposed in the literature. In this section, we classify the most popular of the proposed style markers, taking into account the information required 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, and punctuation marks count. 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, were based exclusively 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. 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, Van Halteren, and Tweedie (1996) used frequencies of occurrence of rewrite rules as they appear in a syntactically annotated corpus and proved that they perform better than word frequencies. Their calculation requires tagged or parsed text, however. Current NLP tools are not able to provide accurate calculation results for many of the previously 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, Nunberg, and Schütze 1997). As a result, the recent advances in computational linguistics have not notably affected 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., words occurring once in the sample text) and the dislegomena (i.e., words occurring twice 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 (Honoré 1979; Yule 1944; Sichel 1975). Additionally, instead of using a single measure, some researchers have used a set of such vocabulary richness functions in combination with multivariate statistical techniques to achieve better results in authorship attribution (Holmes 1992). In general, these measures are not computationally expensive. 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 (Tweedie and Baayen 1998).

2.4 Common Word Frequencies

Instead of using vocabulary distribution measures, some researchers have counted the frequency of occurrence of individual words in the sample text. Such counts are a reliable discriminating factor (Karlgreen and Cutting 1994; Kessler, Nunberg, and Schütze 1997) and have been applied to many works in text genre detection. Their calculation is simple, but nontrivial 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" (p. 28). Furthermore, the results of such studies are highly language dependent. Michos et al. (1996) introduce the idea of grouping certain words in categories, such as idiomatic expressions, scientific terminology, formal words, and so on. 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 of common high-frequency words (typically 30 or 50 words) has been applied mainly to authorship attribution studies (Burrows 1987).

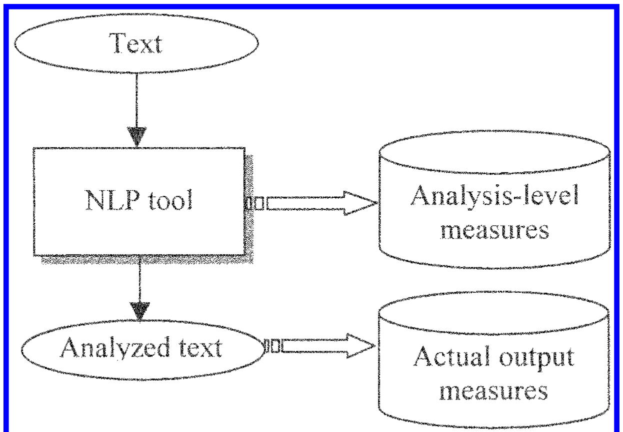

Figure 1
The proposed method.

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 computationally inexpensive. 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.), aimed at improving its performance. For a fully automated system, such restrictions require robust and accurate NLP tools.

3. The Proposed Method

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

- measures relevant to the actual output of the NLP tool (i.e., usually tagged or parsed text), and
- measures 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 a black box. Therefore, full access to its source code is required in order 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, researchers using this approach can define analysis-level measures based on their deep understanding of a particular NLP tool even if they are not familiar with the natural language to which the methodology is to be applied.

To illustrate the proposed method, we apply it to Modern Greek using the SCBD, an 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 the 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, the 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 to represent the structure of the sentences that compose the text. However, it is set to "skip" or surrender attempts to parse clauses after reaching a time-out 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.

4.1 Description of the SCBD

The 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 potential sentence boundaries: period, exclamation point, question mark, and ellipsis. A set of automatically acquired disambiguation rules (Stamatatos, Fakotakis, and Kokkinakis 1999) is 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:** 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 prepositions) 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),

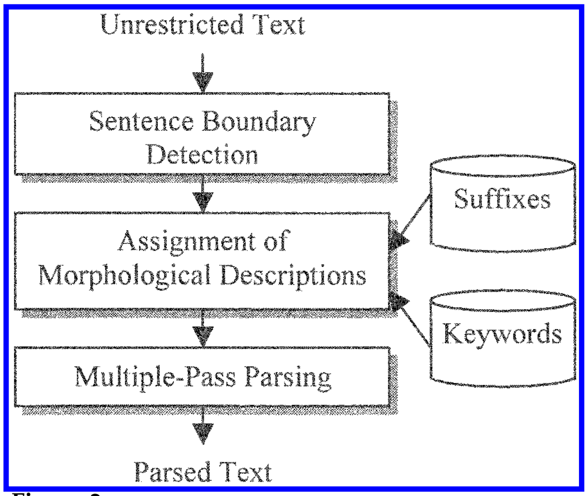

Figure 2
The SCBD structure.

prepositional phrases (PPs), verb phrases (VPs), and adverbial phrases (APs). In addition, two chunks are usually connected by a sequence of conjunctions (CONs).

The SCBD is able to cope rapidly with any piece of text, even ill-formed text, and its performance is comparable to more sophisticated systems that require more complicated resources. Figure 2 gives an overview of the SCBD. An example of its output for a sample text, together with a rough English translation (included in parentheses), 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 opinion)].

It is worth noting that we did not modify the structure of the SCBD in order to calculate style markers, 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 the SCBD 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:** Measures that represent the way in which the sample text has been analyzed by the particular methodology of the SCBD are included here:

Code	Description
M14	detected keywords/words
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 M15 are valuable markers that indicate of 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 to 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 large part of a syntactically complicated text would not be analyzed by them (e.g., high values for M18, M19, and M20 in conjunction with low values for M21 and M22). Similarly, a syntactically simple text would be characterized by low values for M18, M19, and M20.

1 We consider words as word tokens.

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 automatically classifying this vector into one group (either genre or author) various techniques are available, which stem 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 on 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 10) 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 used 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. Thus, they can be easily incorporated into a real-time application.

5.1 Multiple Regression

Multiple regression predicts values of 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:

$$y_i = b_0 + z_1 b_{1i} + z_2 b_{2i} + \cdots + z_r b_{ri} + e_i$$

where y_i is the response for the ith category (i.e., text genre), $z_1, z_2, \ldots, z_r$ are the predictor variables (i.e., in our case $r = 22$), $b_0, b_{1i}, b_{2i}, \ldots, b_{ri}$, 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 semi-partial 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 ith 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 jth 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.

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 into 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) raises several problems since such categories may not be stylistically homogeneous (Kessler, Nunberg, and Schütze 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 thus may be ill-formed.

Table 1
The genre-based corpus.

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*

- *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.

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. Although the complete set of text genres may differ significantly among two languages (Biber 1995), 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 must also be pointed out 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 part 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 described in Section 7.

6.2 Setting the Baseline

To evaluate 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 computationally inexpensive.

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

$$
\begin{aligned}
K &= \frac{10^4 \left(\sum_{i=1}^{\infty} i^2 V_i - N\right)}{N^2} \\
R &= \frac{(100 \log N)}{\left(1 - \left(\frac{V_1}{V}\right)\right)} \\
W &= N^{V^{-\alpha}} \\
S &= \frac{V_2}{V} \\
D &= \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)}
\end{aligned}
$$

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, Van Halteren, and Tweedie 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 (which stands for vocabulary richness).

The second method, which is lexically based, uses the frequencies of occurrence of the most frequent words of the training corpus as style markers. Typically, sets of 30 or 50 most frequent words are used (Baayen, Van Halteren, and Tweedie 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. These two approaches will be called CWF-30 and CWF-50 for common word frequencies and the number of the high-frequency words.

6.3 Results

The entire corpus described in the previous section was analyzed by the SCBD, which automatically provided a vector of 22 parameters for each text. 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 and for the CWF-30 and CWF-50 methods.

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-based corpus (Tweedie and Baayen 1998). Moreover, our approach is more accurate than the CWF-30 and the CWF-50. The identification error rate of our approach using both multiple regression and discriminant analysis is given in Table 2. Although the average error rate 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 rate is more normally distributed using discriminant analysis. Moreover, approximately 60% of the identification errors using multiple re-

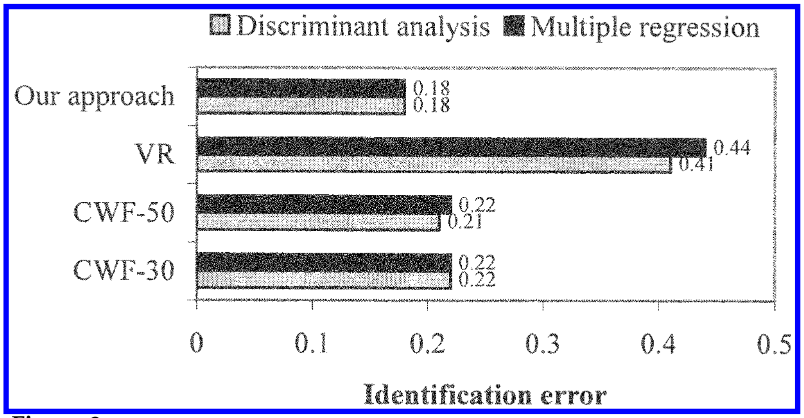

Figure 3
Comparative results for text genre detection.

Table 2
The text genre detection results.

	Identification Error	
Code	**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

gression were caused by G01 and G07, while 65% of the identification errors using discriminant analysis were caused by G01, G05, and G07. On the other hand, G04, G06, G08, and G10 are stylistically homogeneous to a great extent in both cases.

The complete identification results of our method using discriminant analysis are presented in a confusion matrix in Table 3. Each row represents a text genre being tested 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, which is published on a weekly basis. In many cases, therefore, the reportage documents review a whole week and present some comments by the author.
- *curricula vitae* → *official documents*. Both are usually characterized by an abstract style.
- *literature* → *interviews and planned speeches*. These two text genres of the spoken language usually involve narration.

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

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

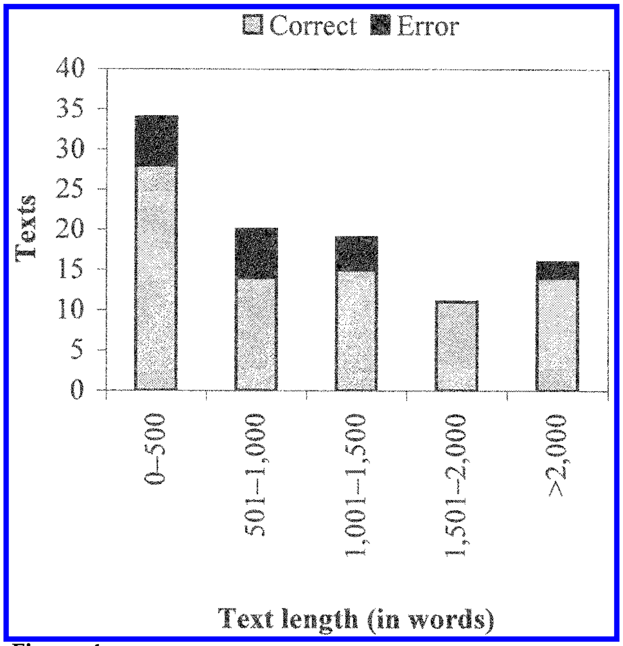

Figure 4
Text length related to accuracy for the text genre detection experiment.

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.

Table 4
The structure of the Modern Greek weekly newspaper *TO BHMA*.

Section Code	Title (Translation)	Description
A	*TO BHMA* (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

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 frequently publish their writings in the press, making the collection of a considerable number of texts for several authors 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:

1. **Group A:** 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 to conform to a predefined style, thus washing out specific characteristics of the authors, which complicates the task of attributing authorship.
2. **Group B:** Ten randomly selected authors whose writings are frequently found in section B. This supplement comprises essays on science, culture, history, and so on, 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 journalists.

Analytical information on the author-based corpus is in Table 5. All the downloaded texts were taken from issues published during 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. This information was not taken into account during the construction of the corpus. The author-based

2 The Web address is: http://tovima.dolnet.gr

Table 5
The author-based corpus.

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

corpus was divided into a training part and a test part of equal size (i.e., 10 texts per author for training and 10 texts per author for test).

7.2 Author Identification

As for the text genre detection experiment, the entire corpus was first analyzed automatically by the SCBD. We then used the stylistic vectors of the training corpus to train the classification model for each group separately, based on multiple regression and discriminant analysis. We cross-validated the acquired models 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 rate for 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 than in group A. In both groups, our approach achieved the best performance.

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).

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

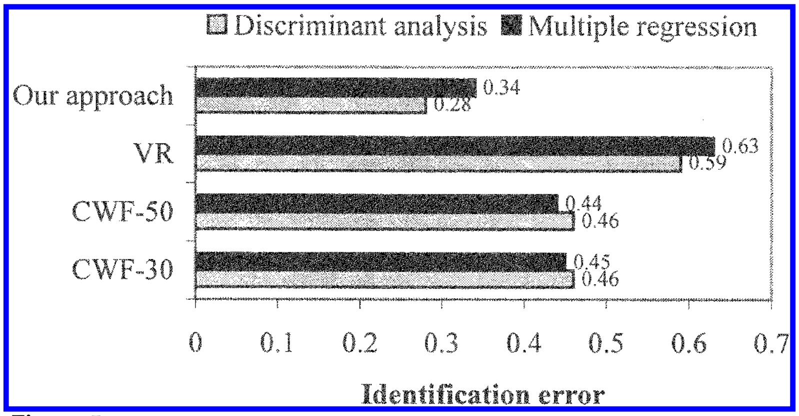

Figure 5
Comparative results for authorship identification in group A.

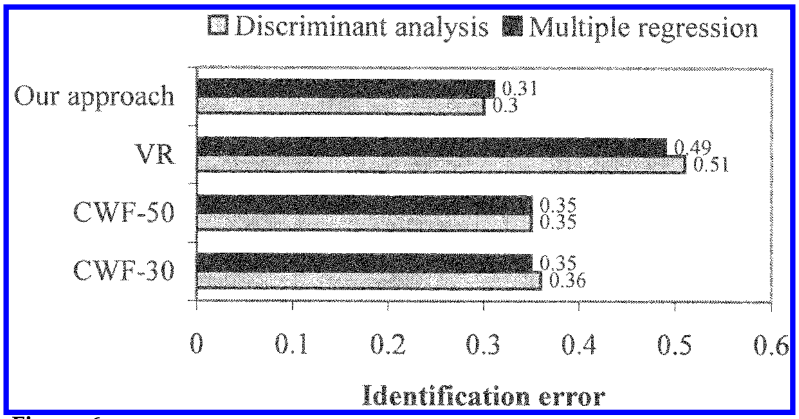

Figure 6
Comparative results for authorship identification in group B.

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

	Identification Error			Identification Error	
Code	Multiple Regression	Discriminant Analysis	Code	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

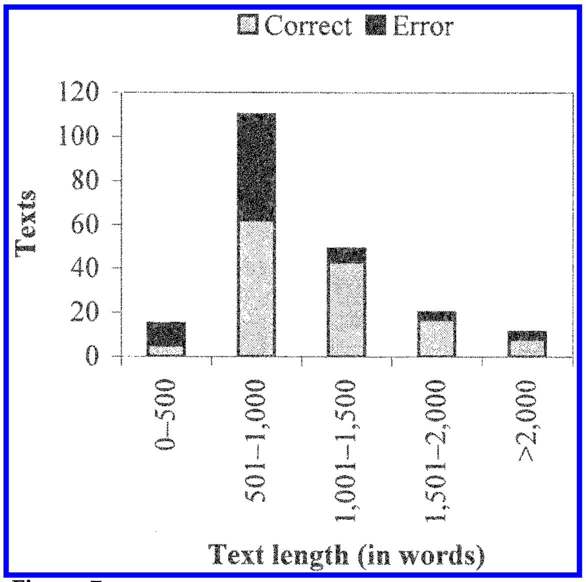

Figure 7
Text length related to accuracy for the author identification experiments.

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 when 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.

7.3 Author Verification

Instead of trying to select the most likely author of a given text from among a given group of authors (i.e., the author identification problem), many applications require the confirmation (or rejection) of the hypothesis that 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 there are only two possible answers: *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.

Implementing an automatic author verification system requires:

- The development of a **response function** for a given author. For a given 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 that of the threshold is accepted as written by the author in question. Otherwise, it is rejected.

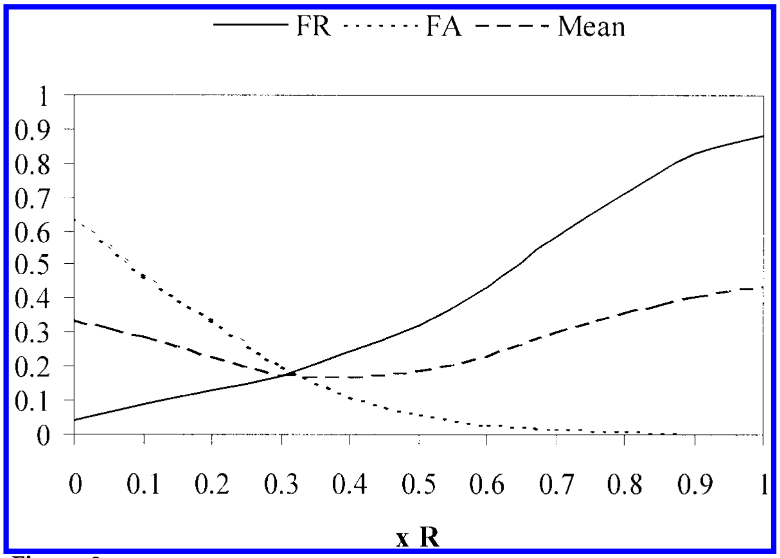

Figure 8
FR, FA, and mean error for group A related to threshold values expressed as subdivisions of R.

Additionally, for measuring the accuracy of the author verification method for a given author, False Rejection (FR) and False Acceptance (FA) can be used. These measures are commonly used in the area of speaker verification in speech processing (Fakotakis, Tsopanoglou, and Kokkinakis 1993) and are defined as follows:

$$FR = \textit{rejected texts of the author/total texts of the author}$$
$$FA = \textit{accepted texts of other authors/total texts of other authors}$$

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

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 to 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 evaluation shown used texts within the same group of authors for testing (i.e., closed-set evaluation). Low threshold values correspond to minimal FR, while high threshold values correspond to minimal FA. The minimal mean error corresponds to threshold values between $0.4R$ and $0.5R$ for both groups. 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.

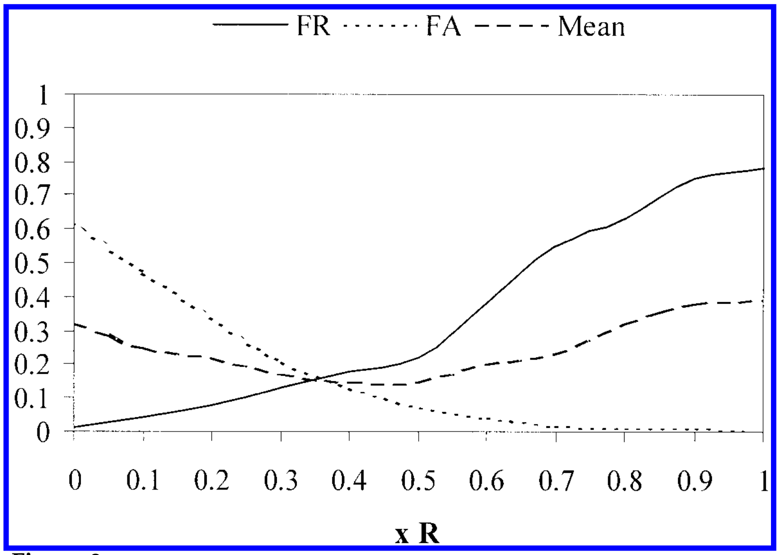

Figure 9
FR, FA, and mean error for group B related to threshold values expressed as subdivisions of *R*.

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

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

8. Performance Issues

8.1 Training Set Size

Our study makes use of 10 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. 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 (as few as 10 texts), but we were interested in exploring the way in which the identification error rate is affected by increasing the training data. 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 rate related to the training set size are shown in Figure 10.

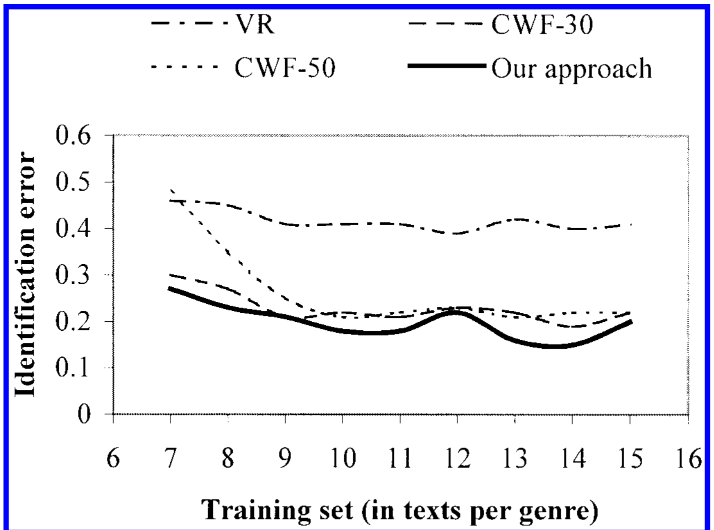

Figure 10
The identification error rate of the text genre detection experiment related to the training set size.

The performance of VR is not significantly affected by increasing the training set size. On the other hand, the identification error rate 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 as compared to CWF-50 but is lower than that of 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 rate does not continuously decrease from 11 to 15 training texts; the identification error rate using 12 as well as 15 training texts for each category is greater than the rate attained by using 10 texts. Thus, it is clear that satisfactory accuracy can be achieved with only 10 training texts.

8.2 Significance of Style Markers

The proposed set of style markers is divided into three levels—token level, syntax level, and analysis level. 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 contributes 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).

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

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

Stylometric Level	Style Marker	Absolute t Values (Average): Text Genre Detection	Absolute t Values (Average): Authorship Attribution
	M01	1.06	1.80
Token level	M02	2.52	1.85
	M03	1.43	1.98
	Level average	1.67	1.88
	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
Syntax level	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
	M14	1.03	1.30
	M15	2.11	1.05
	M16	1.45	0.79
	M17	1.08	1.42
Analysis level	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

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

- in the sentence boundary detection procedure, only periods were considered to denote a potential sentence boundary, and
- the fifth parsing pass was excluded from the chunk boundary detection procedure.

These changes significantly decreased the accuracy of the output of the SCBD. We performed the text genre experiment again using multiple regression based on the defective data. The average identification error rate was increased approximately 25% (i.e., new identification error = 0.23). As expected, the accuracy of the text categorization methodology strongly depends on the accuracy of the SCBD. Note that the contribution of the stylometric levels to the final response has also 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 contributors to the response, the disproportion between them and both the analysis-level and the syntax-level measures has considerably decreased.

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

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

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 and verification experiments are strongly encouraging; this methodology outperforms existing lexically based methods. Since the stylistic differences are clearer among text genres, the results achieved in text genre detection are considerably better than those of the 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 rate.

As 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, it is not always possible to reach this lower bound. The corpora used in all the experiments presented here 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.

Our experiments have shown that our method can be applied to a randomly selected group of stylistically homogeneous categories without any manual adaptation restrictions. A training corpus consisting of 10 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 to the final response than measures related to the actual output of the tool on the syntactic level (see Table 8).

Much work remains to be done on the stylistic interpretation of the acquired results and the automatic extraction of stylistic conclusions related to both the text itself and its author. Such stylistic conclusions could explain the differences and 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 in the detection of certain sections of the input text where the useful information is more likely to be found. We believe that such tasks can be performed using a set of style markers similar to the one we proposed. Finally, the combination of our approach with lexically based methods, such as CWF-30, can result in a very reliable text categorization system in terms of stylistically homogeneous categories.

Acknowledgment
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, Hans Van Halteren, and Fiona Tweedie. 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*. W. H. Freeman, San Francisco, CA.

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

Fakotakis, Nikos, Anastasios Tsopanoglou, and George Kokkinakis. 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 Richard S. Forsyth. 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. 1999. Stylistic experiments in information retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer Academic Publishers, pages 147–166.

Karlgren, Jussi and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, pages 1,071–1,075.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of 35th Annual Meeting*, pages 32–38. Association for Computational Linguistics.

Michos, Stefanos, Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 1996. An empirical text categorizing computational model based on stylistic aspects. In *Proceedings of the 8th Conference on Tools with Artificial Intelligence (ICTAI'96)*, pages 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 David Wallace. 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*. University of South Carolina Press, Columbia.

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, Nikos Fakotakis, and George Kokkinakis. 1999. Automatic extraction of rules for sentence boundary disambiguation. In *Proceedings of the Workshop in Machine Learning in Human Language Technology, Advance Course on Artificial Intelligence (ACAI'99)*, pages 88–92.
Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. A practical chunker for unrestricted text. In *Proceedings of the 2nd International Conference on Natural Language Processing*, pages 139–150.
Strzalkowski, Tomek. 1994. Robust text processing in automated information retrieval. In *Proceedings of the 4th Conference On Applied Natural Language Processing (ANLP'94)*, pages 168–173.
Tabachnick, Barbara G. and Linda S. Fidell. 1996. *Using Multivariate Statistics*. Third edition. HarperCollins College Publishers.
Tweedie, Fiona and Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.
Tweedie, Fiona, Sameer Singh, and David I. Holmes. 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30(1):1–10.
Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, 1(1):69–90.
Yang, Yiming and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 42–49.
Yule, George U. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

This article has been cited by:

1. Kumiko Tanaka-Ishii, Shunsuke Aihara. 2015. Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy. *Computational Linguistics* **41**:3, 481-502. [Abstract] [Full Text] [PDF] [PDF Plus]
2. George Papadakis, George Giannakopoulos, Georgios Paliouras. 2015. Graph vs. bag representation models for the topic classification of web documents. *World Wide Web* . [CrossRef]
3. Duc-Thuan Vo, Cheol-Young Ock. 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications* **42**, 1684-1698. [CrossRef]
4. Tie-Yun Qian, Bing Liu, Qing Li, Jianfeng Si. 2015. Review Authorship Attribution in a Similarity Space. *Journal of Computer Science and Technology* **30**, 200-213. [CrossRef]
5. Peng Tang, Tommy W.S. Chow. 2014. Mining language variation using word using and collocation characteristics. *Expert Systems with Applications* **41**, 7805-7819. [CrossRef]
6. Alaa Saleh Altheneyan, Mohamed El Bachir Menai. 2014. Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University - Computer and Information Sciences* . [CrossRef]
7. Snehasish Banerjee, Alton Y.K. Chua. 2014. A theoretical framework to identify authentic online reviews. *Online Information Review* **38**, 634-649. [CrossRef]
8. Renkui Hou, Jiang Yang, Minghu Jiang. 2014. A Study on Chinese Quantitative Stylistic Features and Relation Among Different Styles Based on Text Clustering. *Journal of Quantitative Linguistics* **21**, 246-280. [CrossRef]
9. Peng Tang, Mingbo Zhao, Tommy W.S. Chow. 2014. Text style analysis using trace ratio criterion patch alignment embedding. *Neurocomputing* **136**, 201-212. [CrossRef]
10. T. Clement, D. Tcheng, L. Auvil, B. Capitanu, J. Barbosa. 2013. Distant Listening to Gertrude Stein's 'Melanctha': Using Similarity Analysis in a Discovery Paradigm to Analyze Prosody and Author Influence. *Literary and Linguistic Computing* **28**, 582-602. [CrossRef]
11. Miroslav Kubát, Jiří Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* **20**, 339-349. [CrossRef]
12. Peng Tang, Tommy W.S. Chow. 2013. Recognition of word collocation habits using frequency rank ratio and inter-term intimacy. *Expert Systems with Applications* **40**, 4301-4314. [CrossRef]
13. Hemlata Pande, H.S. Dhami. 2013. Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language #. *Journal of Quantitative Linguistics* **20**, 1-12. [CrossRef]
14. Jahna Otterbacher. 2012. Gender, writing and ranking in review forums: a case study of the IMDb. *Knowledge and Information Systems* . [CrossRef]
15. L. Pearl, M. Steyvers. 2012. Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and Linguistic Computing* **27**, 183-196. [CrossRef]
16. Mike Kestemont, Kim Luyckx, Walter Daelemans, Thomas Crombez. 2012. Cross-Genre Authorship Verification Using Unmasking. *English Studies* **93**, 340-356. [CrossRef]
17. Murat Yasdi, Banu DiriAuthor recognition by Abstract Feature Extraction 1-4. [CrossRef]
18. Dae-Neung Sohn, Jung-Tae Lee, Kyoung-Soo Han, Hae-Chang Rim. 2011. Content-based Mobile Spam Classification Using Stylistically Motivated Features. *Pattern Recognition Letters* . [CrossRef]
19. Senja Pollak, Roel Coesemans, Walter Daelemans, Nada Lavrač. 2011. Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. *Pragmatics* **21**:10.1075/prag.21.4, 647-683. [CrossRef]
20. Tao Li, Mitsunori Ogihara, Sheng Ma. 2010. On combining multiple clusterings: an overview and a new perspective. *Applied Intelligence* **33**, 207-219. [CrossRef]
21. Fazli Can, Jon Patton. 2010. Change of Word Characteristics in 20th-Century Turkish Literature: A Statistical Analysis. *Journal of Quantitative Linguistics* **17**, 167-190. [CrossRef]
22. Dae-Neung Sohn, Jung-Tae Lee, Seung-Wook Lee, Joong-Hwi Shin, Hae-Chang Rim. 2010. Korean Mobile Spam Filtering System Considering Characteristics of Text Messages. *Journal of the Korea Academia-Industrial cooperation Society* **11**, 2595-2602. [CrossRef]
23. Guilherme T. de Assis, Alberto H. F. Laender, Marcos André Gonçalves, Altigran S. da Silva. 2009. A Genre-Aware Approach to Focused Crawling. *World Wide Web* **12**, 285-319. [CrossRef]
24. Peter Garrard. 2009. Cognitive archaeology: Uses, methods, and results. *Journal of Neurolinguistics* **22**, 250-265. [CrossRef]

25. Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* **60**:10.1002/asi.v60:3, 538-556. [CrossRef]
26. Blake Stephen Howald. 2009. Authorship Attribution under the Rules of Evidence: Empirical Approaches in a Layperson's Legal System. *International Journal of Speech Language and the Law* **15**. . [CrossRef]
27. Moshe Koppel, Jonathan Schler, Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* **60**:10.1002/asi.v60:1, 9-26. [CrossRef]
28. Jun Xu, Yuxin Ding, Xiaolong Wang, Yonghui WuGenre identification of Chinese finance text using machine learning method 455-459. [CrossRef]
29. F. Alias, X. Sevillano, J.C. Socoro, X. Gonzalvo. 2008. Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification. *IEEE Transactions on Audio, Speech, and Language Processing* **16**, 1340-1354. [CrossRef]
30. T KUCUKYILMAZ, B CAMBAZOGLU, C AYKANAT, F CAN. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management* **44**, 1448-1466. [CrossRef]
31. Shlomo Argamon, Jeff Dodick, Paul Chase. 2008. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics* **75**, 203-238. [CrossRef]
32. Zafer Kaban, Banu DiriGenre and author detection in Turkish texts using artificial immune recognition systems 1-4. [CrossRef]
33. Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management* **44**, 790-799. [CrossRef]
34. Lei Dong, Carolyn Watters, Jack Duffy, Michael ShepherdAn Examination of Genre Attributes for Web Page Classification 133-133. [CrossRef]
35. Chung-Hong Lee, Chih-Hong Wu, Hsin-Chang YangA Platform Framework for Cross-Lingual Text Relatedness Evaluation and Plagiarism Detection 303-303. [CrossRef]
36. Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* **58**:10.1002/asi.v58:6, 802-822. [CrossRef]
37. Vedrana Vidulin. 2007. Training the Genre Classifier for Automatic Classification of Web Pages. *Journal of Computing and Information Technology* . [CrossRef]
38. EFSTATHIOS STAMATATOS. 2006. AUTHORSHIP ATTRIBUTION BASED ON FEATURE SET SUBSPACING ENSEMBLES. *International Journal on Artificial Intelligence Tools* **15**, 823-838. [CrossRef]
39. Tao Li, M. Ogihara. 2006. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia* **8**, 564-574. [CrossRef]
40. F. SebastianiClassification of Text, Automatic 457-462. [CrossRef]
41. Chul Su Lim, Kong Joo Lee, Gil Chang Kim. 2005. Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management* **41**, 1263-1276. [CrossRef]
42. DAVID BISANT. 2005. AN APPLICATION OF NEURAL NETWORKS TO SEQUENCE ANALYSIS AND GENRE IDENTIFICATION. *International Journal of Pattern Recognition and Artificial Intelligence* **19**, 199-215. [CrossRef]
43. 2004. Automatic Classification of Web documents According to their Styles. *The KIPS Transactions:PartB* **11B**, 555-562. [CrossRef]
44. S Park. 2004. Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. *Information Processing & Management* **40**, 421-439. [CrossRef]
45. Scott JarvisLexical Challenges in the Intersection of Applied Linguistics and ANLP 50-72. [CrossRef]
46. Georgia Frantzeskou, Stephen G. MacDonell, Efstathios StamatatosSource Code Authorship Analysis For Supporting the Cybercrime Investigation Process 470-495. [CrossRef]