Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments

George Tambouratzis and Marina Vassiliou
Institute for Language and Speech Processing, Greece

Abstract

This article reports on experiments performed with a large corpus, aiming at separating texts according to the author style. The study initially focusses on whether the classification accuracy regarding the author identity may be improved, if the text topic is known in advance. The experimental results indicate that this kind of information contributes to more accurate author recognition. Furthermore, as the diversity of a topic set increases, the classification accuracy is reduced. In general, the experimental results indicate that taking into account knowledge regarding the text topic can lead to the construction of specialized models for each author with higher classification accuracy. For example, by focussing on a specific topic, the accuracy with which the author identity is determined increases, the exact amount depending on the specific topic. This also applies when the topic of the text is more broadly determined, as a set of topic categories.

In an associated task, the most salient parameters within an 85-parameter vector are studied, for a number of subsets of the corpus, where each subset contains speeches from a single topic. These studies indicate that the salient parameters are the same for the different subsets. Two fixed data vectors have been defined, using 16 and 25 parameters, respectively. The classification accuracy obtained, even with the smallest data vector, is only 5% less than with the complete vector. This indicates that the parameters retained in the reduced vectors bear a large amount of discriminatory information and suffice for an accurate classification of the corpus.

Correspondence:
George Tambouratzis,
Institute for Language and Speech Processing,
Paradissos Amaroussiou,
15125 Athens, Greece.
E-mail: giorg_t@ilsp.gr

1 Introduction

During the last decades, the organization of information in readily accessible format has become an essential task, given the rapid growth of electronic text collections. The amount of constantly increasing textual information makes it imperative to handle texts in an efficient manner for various purposes: information retrieval, text search applications, document management, text versioning, stylometry and so on. These purposes are to a considerable extent intertwined. For instance, stylometry (i.e. information regarding author style) can be employed for classifying more accurately a collection of texts, enhancing the effectiveness of information retrieval applications.

doi:10.1093/llc/fqm003 Advance Access Published on 19 March 2007

Within the domain of stylometry two basic lines of research have developed, which rely on the employment of various stylistic markers as indicators of speech type or personal manner of expression. The two lines of research that have been developed are namely (1) register discrimination (see among others Biber, 1993; Karlgren and Cutting, 1994) and (2) author discrimination (see among others Mosteller and Wallace, 1984; Holmes, 1994).

Given the large number of digitized texts as well as the high rate with which new texts are introduced in electronic text collections, the requirement exists to automate the task of stylistic categorization. To achieve that, the use of computers is combined with methods from the paradigms of traditional stylometry and statistics (Rudman, 2000). Furthermore, instead of statistical methods, the use of pattern recognition techniques and neural networks has also been proposed (Holmes, 1998).

In the present article, we study the effectiveness of specific stylistic markers in discriminating between the personal styles of different authors. These markers have been determined following a series of studies on texts in the Greek language, though most of them are language-independent, as they have also been utilized by researchers studying texts in other languages (Holmes, 1985, 1994). The aim here is to study the effect of specific variables in author discrimination experiments, when using a variety of subcorpora from a well-defined register, which correspond to different topics. To that end, we have used the political speech register as this is defined in the Minutes of the Greek Parliament.

In Section 2, we present an overview of previous research in the area of stylistics, investigating both the methodologies employed and the linguistic markers used. In Section 3, we present the methodology chosen to perform the author identification task. The linguistic markers used to extract the author style characteristics from the documents are described in Section 4. In Section 5, we give an account of the process followed to define the topics and divide the corpus into sets of texts, each set belonging to a single topic. Section 6 contains the experiments that aim at recognizing author identities, when the texts used originate from a single topic. In Section 7, the emphasis shifts towards performing author identification tasks for text corpora that cover sets of several related topics. Finally, the conclusions derived from this piece of work are presented in Section 8.

2 Overview of Previous Research

Stylometric studies have predominantly focussed on the attribution of authorship of literary manuscripts and historical texts. Mosteller and Wallace (1984) and Collins *et al*. (2004) have attempted to determine the author of a disputed subset of the Federalist papers using classical statistical methods, while Gurney and Gurney (1998) have applied broadly similar methods to the analysis of the Scriptores Historiae Augustae corpus. The authorship of poems and plays attributed to Shakespeare has also formed the subject of numerous debates and studies (see among others Matthews and Merriam, 1993; Elliott and Valenza, 1999; Foster, 1999). Other areas of application of authorship attribution tasks have included the issue of single versus multiple authorship of Mormon scripture (Holmes, 1992), the authorship of Biblical texts (Holmes, 1994) as well as more theoretical aspects of theology (Bartholomew, 1988).

In the past few decades, stylometric studies have been mainly based on statistical methods in order to reach solid conclusions regarding the authorship of a given text. In their seminal study, Mosteller and Wallace (1984) have investigated the frequency-of-occurrence of words using a Bayesian-type analysis to fit the experimental data to different models in order to distinguish between the various potential authors. Baayen *et al*. (1996) have employed Principal Component Analysis to differentiate between the writing styles of two different authors. Khmelev and Tweedie (2002) have proposed the use of Markov chains for identifying between the texts of different authors. To accomplish the same task, Burrows (2002) has introduced the ‘Delta’ procedure that measures the divergence between the frequencies of occurrence of a set of words over different authors.

Regarding the selection of variables, stylometric studies were initially based on studying the frequency-of-occurrence of isolated words (see among others Mosteller and Wallace, 1984). Stylometric studies have been enriched by introducing additional style markers, such as word-length and sentence-length distribution (Sichel, 1974). Furthermore, the distribution of parts-of-speech and vocabulary histograms has been studied (see Holmes, 1985, 1994 for an extensive review). Following a letter-based approach, Khmelev and Tweedie (2002) have proposed using Markov chains of single letters to recognize the works of specific authors.

Several studies within the topic of stylistics have focussed on examining the effectiveness of even more complex stylistic markers, which are indicators of the manner in which a given author expresses himself. For instance, Collins *et al.* (2004) have tried to determine linguistic variables that map the representational choices of authors. The frequencies of specific word sequences as stylistic markers have been studied by Hoover (2002). Baayen *et al.* (1996) have proposed the study of syntactic structures to enhance the accuracy of authorship attribution. In a similar vein, Stamatatos *et al.* (2001) study the styles of ten columnists of a Greek newspaper, relying on syntactic variables extracted automatically with a custom NLP tool.

Alternative approaches to stylometric studies have also been proposed, these relying on techniques that fall within the paradigm of pattern recognition, yet are inspired by artificial intelligence algorithms rather than classical statistical methods. A prominent example is that of neural networks, which are intended to exploit highly parallel computational structures consisting of simple processing elements, in order to mimic the structure of biological neural networks and perform complex computational tasks. Such models include the supervized multi-layer perceptron, which has been applied to distinguish between the works of Fletcher and Shakespeare (Matthews and Merriam, 1993) and to attribute the Federalist papers to one of their candidate authors (Tweedie *et al.*, 1996). More recently, supervised network models, not relying on a fixed structure but allowing the evolution of the structure to suit the task at hand, have been proposed (Waugh *et al.*, 2000). It should be noted that architectures based on the self-organizing map model have also been proposed for grouping large collections of texts on the basis of their content, in order to support information retrieval applications (Merkl, 1999; Kohonen *et al.*, 2000; Dittenbach et al., 2001).

Our studies in stylistics have focussed mainly on clustering documents written in the Greek language according to their register (Tambouratzis *et al.*, 2004a) as well as their author style (Tambouratzis et al., 2004b). In these studies, we have employed agglomerative statistical clustering methods and statistical discriminant analysis. To achieve a high accuracy, the set of linguistic variables studied encompasses a wide variety of phenomena. It has been found that both register classification and author style recognition may be achieved with the desired degree of accuracy.

Amongst the two aforementioned experimental directions for stylistic studies, the second direction represents a more complex task, and therefore calls for determining and studying a larger set of linguistic variables. It should be noted that neural network models such as the self-organizing map have also been used for separating authors in the case of the Greek Parliament Minutes (Tambouratzis *et al.*, 2003; Tambouratzis 2004). Though in several cases neural networks have generated promising results, the present article focusses on the use of statistical methods. This is due to the fact that statistical methods possess a robust mathematical foundation, and thus can be expected to yield more accurate and consistent results.

3 Methodology

3.1 Discriminant analysis principles and application in the given task

As noted previously, the aim is to determine whether, in author identification tasks, knowledge regarding the topic of the text (henceforth also termed as document) can assist in the creation of more accurate discriminant models. Let us assume that a corpus C is defined, containing n documents in total, which are denoted as d_i, *where* $1 \leq i \leq n$. The documents in corpus C have been created by one of k possible authors. If the number of documents from a given author j is n_j, then:

$$n = \sum_{j=1}^{k} n_j \tag{1}$$

For each of these documents, d_i, the author identity is stored in variable a_i, whose values range between *1* and k, where generally $k > 2$. Information regarding the topic of the specific document is defined and stored in a dedicated variable, t_i, whose value ranges between 1 and M.

Initially, each document is transformed into a vector of p real-valued variables x_i, each of which represents the frequency of occurrence of a linguistic phenomenon (as detailed in Section 4) and forms a potential predictor variable for the model:

$$X = [x_1, x_2, .., x_p] \tag{2}$$

As expressed by Equation (2), via the extraction of the linguistic variables the collection of documents is mapped onto a p-dimensional pattern space. Within this pattern space, the aim is to determine the linguistic variables that reliably predict the authorship of a given document d_i. To that end, a multi-class discriminant analysis is performed, to generate a segmentation of the p-dimensional pattern space into regions, each of which corresponds to one of the k authors. Assuming initially that the discriminant analysis is performed on the entire corpus of documents, C, a set of $k-1$ axes shall be determined (since $k \ll p$), which reliably separate the documents into k groups, minimizing the cumulative classification error. This is achieved by maximizing the inter-group variance relative to the intra-group variance. To that end, the matrix T representing the sums-of-squares of cross-products is calculated for the classification scores over all documents:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{x})(X_{ij} - \overline{x})' \tag{3}$$

Then, for the i-th group, the matrix W_i of intra-group sums-of-squares and cross-products is expressed as:

$$W_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{x_i})(X_{ij} - \overline{x_i})' \tag{4}$$

The cumulative intra-group sum-of-squares is:

$$W = W_1 + W_2 + \cdots + W_k \tag{5}$$

The matrix B, which represents the sums-of-squares and cross-products between groups, is then expressed as:

$$B = T - W \tag{6}$$

If the linear compound Y is defined as $Y = \hat{b}' \cdot X$, where the '^' sign indicates sample-based estimates, the between-groups sum-of-squares is equal to $\hat{b}' \cdot B \cdot \hat{b}$, while the within-groups sum-of-squares is equal to $\hat{b}' \cdot W \cdot \hat{b}$. Then, the ratio of between-groups to within-groups that we wish to maximize is equal to:

$$\hat{\lambda} = \frac{\hat{b}' \cdot B \cdot \hat{b}}{\hat{b}' \cdot W \cdot \hat{b}} \tag{7}$$

By rearranging the terms:

$$(B - \hat{\lambda} \cdot W) \cdot \hat{b} = 0 \tag{8}$$

According to Equation (8), the maximum value of $\hat{\lambda}$ is the largest eigenvalue of the matrix $W^{-1} \cdot B$. The corresponding eigenvector corresponds to the optimal discriminating surface. Hence, the set of the q highest eigenvalues of matrix $W^{-1} \cdot B$ defines the required discriminant model. The eigenvectors of $W^{-1} \cdot B$ correspond to the axes that discriminate most effectively between groups, while the effectiveness of each of the axes is reflected by the relative magnitude of the corresponding eigenvalue.

3.2 Combining multiple discriminant analyses to draw conclusions regarding the saliency of variables

The treatment of Section 3.1 describes the main principles of discriminant analysis (see also Dillon and Goldstein, 1984). Two distinct variants of discriminant analysis exist:

(1) The full discriminant analysis, which generates a set of discriminator functions, combining all predictor variables with suitable coefficients to separate the pattern space into the desired classes.

(2) The step-wise discriminant model, which generates a set of discriminator functions, using only selected predictor variables, which are inserted in the functions, provided that they convey a significant amount of discrimination capability (as defined by a pair of *F*-parameters).

Discriminant analysis has four main objectives (Dillon and Goldstein, 1984):

(a) To find linear composites of the predictor variables, which maximize the ratio of between-groups to within-groups variability.
(b) To determine whether the group centroids are statistically different.
(c) To assign new observations to one of the groups based on the linear composites.
(d) To determine which of the predictor variables contributes more to discriminating among the groups.

In the line of research described here, a number of variables, covering a variety of linguistic phenomena, have been collected to process the document corpus. One of the aims of the work presented in this article is to determine from the set of studied linguistic variables the ones which convey the most discriminatory information regarding the author identity. To that end, the step-wise analysis variant (2) is most useful, since it provides information regarding the most significant linguistic variables (these being the ones that are inserted in the step-wise model). Additionally, it is possible to determine which linguistic variables are consistently used, when the discriminant analysis operates on different subsets of the corpus, extracted with specific criteria. To that end, reduced discriminant models have been created by using different document subcorpora $C_1, C_2, .., C_d$, which are subsets of the main corpus C, and have been determined by defining different topics $h_1, h_2, .., h_d$. Every effort has been made to select topics that are independent to each other (so that the resulting corpora are non-overlapping). For every one of these subcorpora, a discriminant model has been generated using step-wise discriminant analysis. Each of these models comprises a limited number of retained linguistic variables, multiplied by their corresponding factors:

$$D_i = \sum_{j=1}^{p} a_{i,j} \cdot x_j \tag{9}$$

Hence, in the *i*-th model, for the retained variables x_j, a total of L factors $a_{i,j}$ shall be non-zero, the remaining factors being equal to 0. According to this notation, the factors over all models may be assembled as follows:

$$\begin{bmatrix} corpus\_1/theme\_h_1 : [a_{1,1}, a_{1,2}, .., a_{1,p}] \\ corpus\_2/theme\_h_2 : [a_{2,1}, a_{2,2}, .., a_{2,p}] \\ . \\ . \\ corpus\_d/theme\_h_d : [a_{d,1}, a_{d,2}, .., a_{d,p}] \end{bmatrix} \tag{10}$$

Then, the information conveyed by the *d* step-wise discriminant models of Equation (10) is combined to determine the most frequent linguistic variables within the set used in Equation (2). More specifically, for each linguistic variable x_j, the number of corresponding non-zero factors over all step-wise discriminant models is calculated (and denoted as f_j). Then, the linguistic variables are ordered on the basis of their frequencies of occurrence f_j, to create a list of the most salient variables according to the *d* independent discriminant analyses.

4 Description of Variables

For the experiments performed we have employed a substantial amount of variables (for a detailed presentation of the variables employed see Tambouratzis *et al.*, 2004a,b), reflecting a range of linguistic and structural information, mainly in terms of frequency of occurrence, which may be classified into eight classes:

(1) Lemmas: these are variables based on the frequency of occurrence of specific lemmas.
(2) Macro-structural information: these are variables related to the length of words and sentences.

(3) Negation: this category comprises variables based on the frequency of occurrence of negative words.
(4) PoS: these are variables based on the frequency of occurrence of parts of speech.
(5) Punctuation: this category comprises variables reflecting the frequency of occurrence of punctuation marks.
(6) Syntax: variables which fall under this category are related to the occurrence of nominals in specific cases.
(7) Verb_diglossia: these variables exploit verbal information on the differences between distinct phases of the Greek language.
(8) Verb_discourse: these variables describe verbal information on discourse tendencies.

5 Defining the Topics Within the Parliament Register

In this set of experiments, the effect of the topic on the classification of the text corpus is investigated. In particular, the study focusses on whether the classification accuracy with respect to the author identity may be improved if the topic of a given speech is known in advance. Therefore, only speeches of a single topic are studied in each classification task.

To perform the experiments, as a first step the corpus of speeches has been annotated on the basis of the topic of each speech (Table 2). For this classification we have relied on the archival summary incorporated within the Parliament Minutes for each speech/session. In total, twenty-seven distinct topics were defined, which range from foreign issues to educational issues and agricultural policy. It is possible, though, that one speech spans more than one topic; in such cases the predominant topic of the Parliamentary session was chosen as the primary topic. Furthermore, a given speech may cover several secondary topics; nevertheless, it has been decided to record at the most two topics (one primary and—optionally—one secondary) for each speech.

When defining the speech topics, one special case involves the inaugural speeches delivered at the first Parliament session following a general election. In these inaugural speeches, a comprehensive review of the proposed government policy is performed, and thus the speeches cover various topics. Since such speeches have a particular flavour, they have been grouped into one distinct class (topic 28). This allows the separation of inaugural speeches, which, if included in all topics, might bias the experimental results, at least at a first stage. The alternative would have been to split each speech within this class into subspeeches, each characterized by a single topic. However, it has been decided to only process entire speeches without any alterations (apart from correcting any existing spelling errors), and thus such a splitting into subspeeches would be futile.

The distribution of texts with respect to their topic is illustrated in the diagrams of Fig. 1. Fig. 1a indicates the distribution of speeches according to their primary topic, while Fig. 1b takes into account both the primary and secondary topics. According to Fig. 1a, most topics span a very limited number of texts, with only a handful of topics having a relatively large number of texts associated with them. When being restricted to primary topics, the largest topics are 'Foreign Issues' (topic 11) and 'Finance' (topic 19), each of which comprises approximately 200 speeches. The next largest categories are 'Education' (topic 20) and 'Internal Issues' (topic 14), which comprise seventy and eighty speeches, respectively. When both primary and secondary topics are simultaneously taken into account, the largest topics are found to be 'Internal Issues' (topic 14) with over 450 texts, 'Finance' (topic 19) with more than 250 texts and 'Foreign Issues' (topic 11) with approximately 230 texts.

6 Studying Isolated Topics Within the Parliament Register

6.1 Selecting the appropriate topics

In order to perform a series of experiments that lead to reliable conclusions, it is essential to select topics that comprise a large number of speeches from different speakers. If, for example, a relatively small category with few members were chosen, then most likely that amount of training data would be insufficient for the successful extrapolation of the class characteristics for each author as well as the precise evaluation of the classification accuracy.

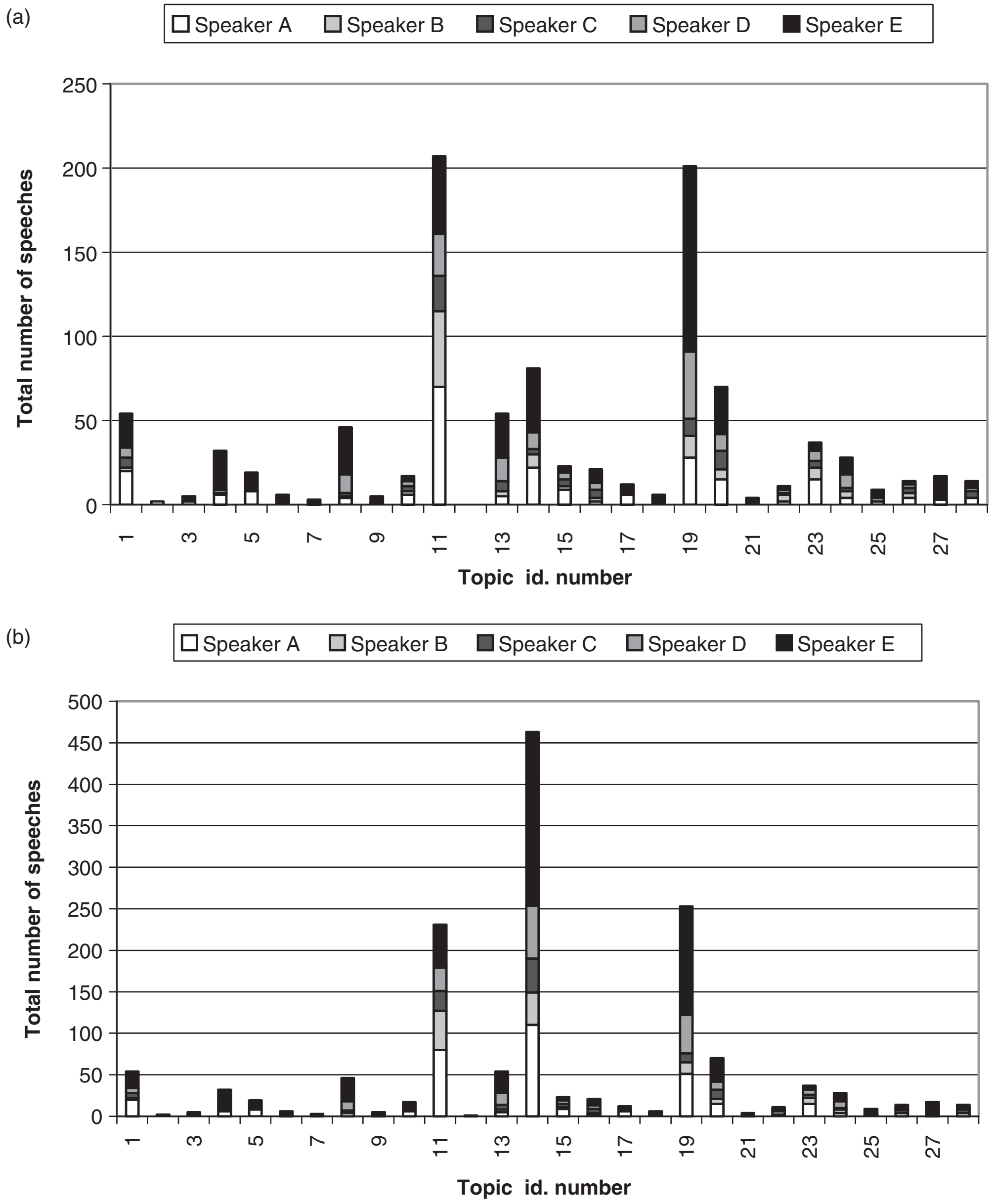

Fig. 1 Histogram of texts over the twenty-eight different topics (indicating via a different shading the number of speeches corresponding to each speaker), when (a) only the primary topic is taken into account and (b) both the primary and secondary topics are taken into account

Though it would be desirable to have categories for which the number of texts per speaker is equal, it is highly unlikely that such a constraint may be enforced in a real-world data set such as the one studied here, where the rate of occurrence of speeches per speaker varies by as much as 1:4. On the other hand, it is essential to have within the data set a sufficient number of training samples for each speaker. To that end, the topics selected for our initial experiments contain at least twenty speeches

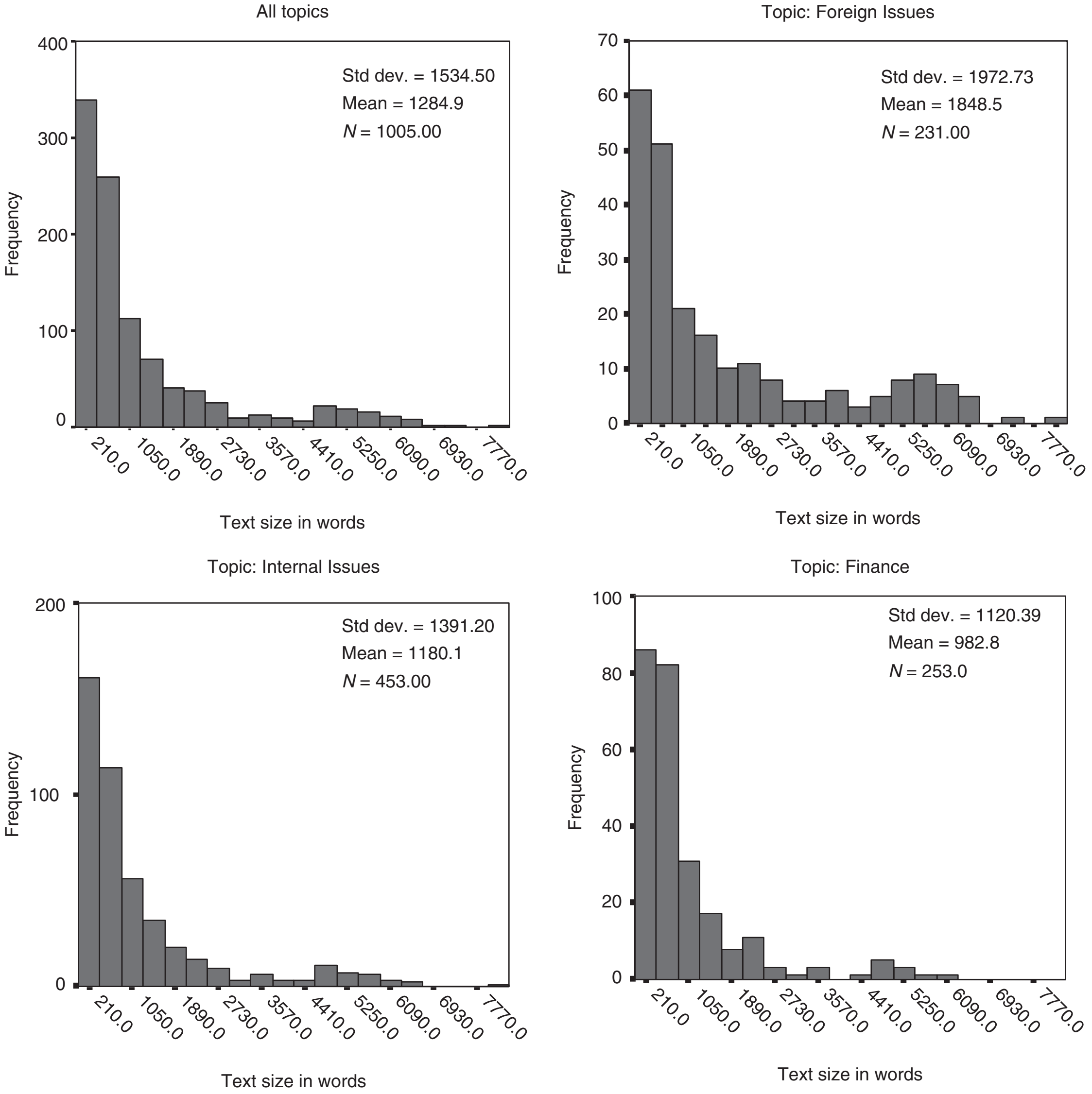

Fig. 2 Distribution of texts within the four corpora used in our experiments

per speaker, when taking into account both primary and secondary topics. These are the following:

(1) 'Internal Issues' (category 14),
(2) 'Finance' (category 19) and
(3) 'Foreign Issues' (category 11)

The statistics of each of these categories, together with the histograms of the text length, are shown in Fig. 2, with the entire set of speeches included for ease of comparison. As can be seen, the histograms for the four data sets are fairly similar, with a relatively high frequency for smaller-sized texts and a lower frequency of occurrence, as the text size increases. Still, some differences become apparent. Probably the most important difference concerns the group of texts belonging to the topic 'Finance', where only a very limited number of large speeches are included (where speeches termed as large are these with more than 4,000 words). This is contrasted by the situation, e.g. the topic ‘Foreign Issues’, where the frequency of large speeches is substantially higher.

6.2 Experiments with fixed *F*-to-enter and *F*-to-remove values

The first experiment involves studying the accuracy of classification of the texts belonging in each of the three aforementioned categories. To that end, the discriminant analysis method has been used to generate a model (this analysis being implemented via the SPSS software package). In order to determine the effect of different linguistic variables on the system performance, for each experiment three runs are performed, in each run evaluating the behaviour of each one of the topics. In addition to the subcorpora corresponding to the three topics, the corpus comprising all speeches (irrespective of topic) is used as a baseline for comparisons.

Further variations involve studying for each of the three topics the accuracy with which the author/speaker of each text may be determined as a function of (1) the speech order as well as (2) primary topic information or both primary and secondary topics. Hence, for each topic there exist four distinct data set variants, corresponding to:

(a) texts/speeches of any order, whose primary topic belongs to the desired category;
(b) texts/speeches of order 1 (i.e. opening speeches only), whose primary topic belongs to the desired category;
(c) texts/speeches of any order, whose primary or secondary topic belongs to the desired category;
(d) texts/speeches of order 1 (i.e. opening speeches only), whose primary or secondary topic is of the desired category.

When the corpus comprising all texts—irrespective of their topics—is used, then only two variants are distinct, since in this case data set variants (a) and (c) are identical, and data set variants (b) and (d) are also identical. Hence, a total of fourteen variants of the discriminant analysis are generated, based on the selection of texts.

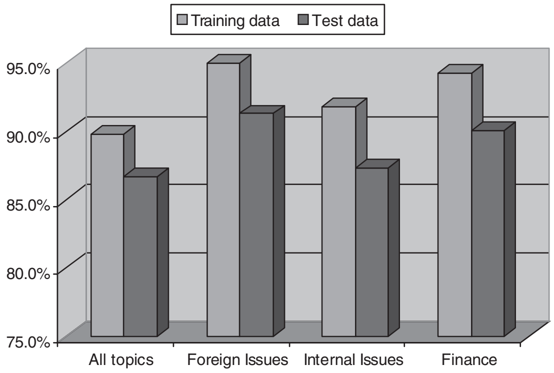

Fig. 3 Classification accuracy of the training data and the testing data (using the cross-validation technique), for the models generated via discriminant analysis using values of *F*-to-enter and *F*-to-remove equal to 4.00 and 3.99, respectively, for speeches of order 1 corresponding to different topic categories

In the first step of the analysis, the entire set of eighty-five variables is used in the input vector, as reported in Section 4. A step-wise forward discriminant analysis has been performed, with the values of the *F* variable, in order to enter/remove a variable in the discriminant model, being set to 4/3.99, respectively. The classification accuracy obtained using speeches of order 1 for the three selected topic categories is depicted in Fig. 3. For comparison purposes, the results when using all texts irrespective of their topic are also presented, under the label ‘All Topics’.

The experimental results indicate that the classification accuracy is improved, when the corpus is restricted on the basis of the text topic as compared to lack of any topic information. The improvement is more substantial in the case of the topic ‘Foreign Issues’, the improvements being more limited in the case of ‘Finance’ and even smaller in the case of ‘Internal Issues’.

In a nutshell, the general trend observed is one of substantial improvement, when speeches of order 1 are studied and when speeches are selected on the basis of only their primary topic, rather than their primary and secondary topic.

One point that should be noted is that using a constant *F*-value to enter/remove variables from the discriminant model, the models generated contain a varying number of variables. In particular, in the case of the topic 'Internal Issues' the number of variables within the model ranges from five to six, resulting in a 'lighter' (in terms of variables) model, which is more tractable to both calculate and explain. For the other two topics (Finance and Foreign Issues), the model size ranges from fifteen to twenty variables. The ability to generate a lighter model is an advantage, especially with respect to the computational effort required to process texts and to determine the model representation for each text, even if the classification accuracy is slightly lower. Naturally, there does exist a correlation between the number of texts in each data set and the number of variables in the corresponding discriminant model, with smaller text corpora requiring fewer variables to be described.

6.3 Comparing models of a similar size in terms of variables

The second experiment with topics involves using the same data sets and varying the *F*-values used to enter/remove variables. More specifically, the *F*-values are selected in order to generate discriminant models with approximately twenty variables in each run. This specific value for the model size has been chosen, as it corresponds to the majority of discriminant models generated in the previous set of experiments, where a fixed *F*-value has been used. The experimental results indicate that the 20-variable models have a behaviour similar to the models obtained with fixed *F*-values. As an example, the comparative classification accuracy, when using 20-variable models for the different topic categories, is shown in Fig. 4. As can be seen, the topic 'Foreign Issues' gives the best classification accuracy, while both 'Finance' and 'Internal Issues' are characterized by a classification accuracy higher than the one obtained by the entire Parliament corpus (which covers all topics). Finally, as shown in Fig. 4, when using 20-variable models and restricting the topic only to the primary one, the classification accuracy is either unchanged (in the case of 'Finance') or improved by 2–3% (this being the case for 'Internal Issues' and 'Foreign Issues'). However, if one compares the actual classification performance, when a constant *F* is used (resulting in a varying number of variables for different models) and when a constant number of variables are used in the model, a broadly similar classification rate is generated.

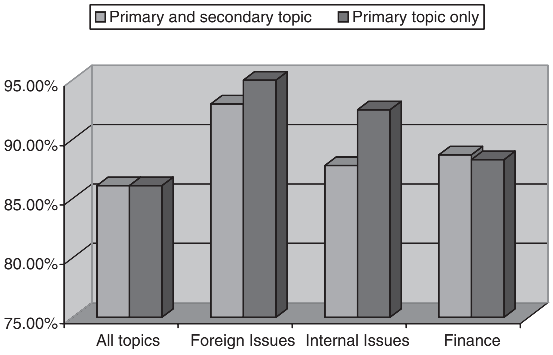

Fig. 4 Comparative classification accuracy of the test data from each data set, for each of the four topic data sets used, when the number of variables per discriminant model approximates twenty. In the majority of cases, when the selection of texts is based solely on the primary topic, as opposed to both primary and secondary topics, the classification accuracy is increased

6.4 Determining the most frequently used variables in the discriminant models

An issue that deserves further study concerns the set of variables that are used to form the discriminant functions. For instance, within the 85-variable vector used in our initial experiments (and furthermore, within the set of variable classes) it is of interest to determine the variables that are most frequently employed in the models, since these specific variables are expected to (1) be the most salient ones and (2) have the highest discriminatory power.

As noted in Section 6.2, a total of fourteen discriminant models have been generated for each experiment. Then, for each of the eighty-five variables, the number of models that retain the specific variable is calculated.

In Table 1, the twenty-five most frequently used variables (these being the variables employed in at least four models) are depicted, together with the class they belong to. In this table, the frequency of use of each variable is depicted collectively over all fourteen discriminant models as well as in

Table 1 The twenty-five most frequently used variables (which are used in at least four out of the fourteen discriminant models created with *F*-values to enter and to remove variables of 4/3.99), the total occurrences of each variable being reported, as well as partial occurrences per specific topic

	Variable description	Variable category	Cumulative number of occurrences	Occurrences per topic		
				Foreign Issues	Internal Issues	Finance
1	Average number of letters per word	macrostructural	14	4	4	4
2	Verb, 3rd person plural	verb_discourse	13	3	4	4
3	Frequency of lemma ‘enas’	lemma	12	2	4	4
4	Frequency of lemma ‘mou’	lemma	11	3	3	3
5	Verb, 1st person plural	verb_discourse	11	3	2	4
6	Frequency of adverbs	PoS	10	3	3	2
7	Verb, 3rd person singular	verb_discourse	10	2	2	4
8	Verb, 2nd person plural	verb_discourse	10	3	2	3
9	Frequency of articles	PoS	9	4	0	3
10	Frequency of verbs	PoS	9	2	2	3
11	Frequency of lemma ‘laos’	lemma	9	4	1	2
12	Frequency of lemma ‘kyrios’	lemma	8	4	2	0
13	Frequency of lemma ‘yparxw’	lemma	8	3	0	3
14	Frequency of question mark	punctuation	6	0	3	1
15	Frequency of rest	PoS	6	4	1	1
16	Verb, 1st person singular	verb_discourse	6	2	2	0
17	Frequency of words containing 1–3 letters	macrostructural	5	3	0	0
18	Frequency of lemma ‘synennohsh’	lemma	5	1	2	0
19	Frequency of dashes	punctuation	5	2	1	0
20	Frequency of conjunctions	PoS	5	1	2	0
21	Frequency of sentences with 76–100 words	macrostructural	5	3	0	0
22	Frequency of lemma ‘den’	lemma	4	1	0	2
23	Frequency of adjectives	PoS	4	2	1	0
24	Frequency of lemma ‘egw’	lemma	4	0	0	2
25	Frequency of pronouns	PoS	4	0	2	0

detail, for each specific topic. It can be seen that these twenty-five variables fall mainly under three categories:

- lemmas (a total of seven lemma-related variables are retained, of which only one refers to negation);
- parts-of-speech (a total of seven variables) and
- verb-discourse variables (a total of five variables).

Additionally, three macrostructural variables are retained (of which the most frequently-used variable concerns the number of letters per word) and two punctuation-related variables (though both have very low frequencies, equal to five and six, respectively, over the fourteen models studied). Finally, a single negation variable is included. Notably, no variables are retained from the syntax and verb-diglossia categories.

A further comment concerns the difference in the frequency of use of specific variables over the different topics. As shown in Table 1, this applies in particular to the case of lemmas, where for example:

- The lemma ‘yparxw’ (translated as ‘exist’) is not used in any of the models generated for the ‘Internal Issues’ category, though it is frequently used in the case of the other two topics.
- The lemma ‘kyrios’ (translated as ‘mister/sir’) is not used in any of the models corresponding to the ‘Finance’ category, though it is frequently used in the models of the other two categories.
- The lemma ‘egw’ (translated as ‘I’) is used solely in the models of the ‘Finance’ category.

In the case of lemmas, the high frequency within models indicates a discriminatory capability due to the different usage of lemmas by each speaker. It could be argued that the frequency of use of specific words depends on the topic being studied. Hence, a given word(s) may be used by one speaker and thus may characterize his speeches in contrast to those of other speakers within the same topic.

This would be of benefit if the topic of a speech were known in advance. For, for speeches whose topic is not known (or for topics for which a model has not been generated), the effectiveness of lemma-type variables would probably be more limited. However, most of the lemmas, whose frequencies are used as variables, correspond to general-purpose words. For example, lemmas 'egw' and 'kyrios' are not used with the same frequency in all categories, yet they do not correspond to a specialized terminology that may be expected to occur more frequently within a specific topic.

Apart from lemmas, several other variables are also used with varying frequencies over different topic categories, as depicted in Table 1. For instance, the frequency of articles is found to have no importance when studying the 'Internal Issues' topic, though it is widely used in the other two topics. The variable recording the frequency of words with a length of between 1 and 3 letters is only used in models corresponding to the topic 'Foreign Issues'. Finally, the variable representing the frequency of very long sentences (with 76–100 words) is only used in the models referring to the topic 'Foreign Issues'.

Table 2 The twenty-eight topics included within the Parliament register

Topics	Topic codes
Agriculture	1
Sports	2
National Defence	3
Development	4
Social Security	5
Forestry	6
Public Security	7
Procedural Issues (Internal Operating Processes)	8
Judiciary Issues	9
National Issues	10
Foreign Issues	11
Salary Supplements	12
Employment Issues	13
Internal Issues	14
Social Issues	15
Transport Issues	16
Press & Mass Media	17
Mercantile Marine	18
Finance	19
Education	20
Environmental Policy	21
Constitutional Issues	22
General Political Issues	23
Commemorative speeches	24
Culture	25
Parliamentary Control/Educational System Review	26
Health Issues	27
Inaugural Speeches	28

6.5 The effect of reducing the number of variables in the data vector

Following the results of the aforementioned series of discriminant analyses, it is evident that only a limited subset of the variables contains a sufficiently large amount of salient information, so as to be consistently utilized in the models. Thus, a comparative study has been performed, the full vector of eighty-five variables being restricted to (1) the twenty-five most salient variables and then (2) the sixteen most salient variables (as determined by the measurements of Table 1, for threshold values of 4 and 6 appearances within the fourteen discriminant models, respectively).

The classification accuracies for the training and the test set are shown in Fig. 5a and b. According to these figures, the classification is more successful, when using the data vector of all eighty-five variables. However, the decrease of the data vector size has only a limited impact on the classification accuracy. For instance, the degradation in performance, when using twenty-five rather than eighty-five variables in the discriminant analysis, is <2%. Furthermore, even when the number of variables in the data vector is reduced from twenty-five to sixteen (substantially reducing the amount of information in the data set), the resulting reduction in the classification accuracy is equal to 5% or less.

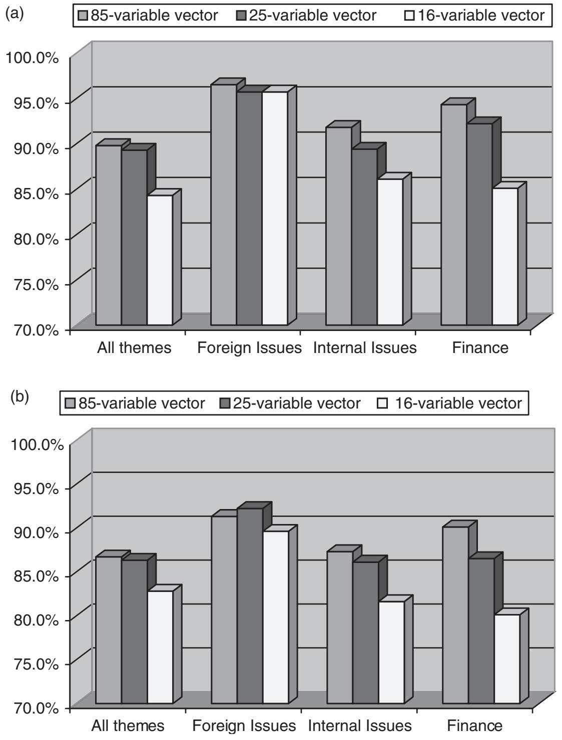

Fig. 5 Classification accuracy of (a) the training data and (b) the testing data (using the cross-validation technique), for the models generated via discriminant analysis. The *F*-to-enter and *F*-to-remove values are equal to 4.00 and 3.99, respectively, for speeches of order 1 corresponding to different topic categories

7 Merging of Related Topics in Order to Create Sets of Topics

7.1 Set-up of the experiments with topic sets

When classifying texts according to their author style, using information about the speech topic to select a subset of the corpus within a register (in our case, the political speech register), the resulting accuracy is improved. What would be of interest is whether the definition of subcorpora corresponding to sets of topics (i.e. groups of related topic categories) would result in a higher classification accuracy than the one obtained, when working with the entire corpus of speeches, irrespective of topic. To that end, the twenty-eight topic categories of our corpus have been studied, with the aim of merging the topics into larger categories of thematically related topics (hereafter termed **topic sets**). It has turned out that merging into larger sets may not be applicable to all topic categories. Thus, we have created four topic sets, as described in Table 3:

{1} **Political Issues**: This topic set comprises six categories concerning (a) parliamentary inspections, (b) inaugural speeches, (c) commemorative speeches related to political events, (d) constitutional issues, (e) procedural issues, and (f) political issues regarding the electoral system.

{2} **Foreign Relations**: This topic set comprises speeches, which refer to the relationship of Greece to other countries, and includes the following topics: (a) Foreign Issues, (b) National Defence Issues, and (c) National Issues (this term indicating specific matters of prime national importance with respect to the international political scene).

{3} **Labour Issues**: This topic set comprises speeches referring to (a) the policy on labour and employment issues, (b) the social security system, and (c) health issues.

{4} **Education and Culture**: This topic set consists of speeches referring to (a) general educational issues, (b) the review of the educational system, and (c) cultural matters.

As can be seen from the aforementioned combinations, certain topic sets are more comprehensive than others. For example, topic set {1} consists of six categories and is the most diverse one in terms of the number of categories, while the remaining three topic sets contain only three categories each. Among these three sets, topic set {2} is almost equivalent to the topic ‘Foreign Issues’, since the vast majority of speeches are characterized as ‘Foreign Issues’ according to either their primary or their secondary topic. In the case of the remaining three topic sets (which include speeches with only a primary categorization but no secondary topics), the distribution of speeches to the constituent categories is much more balanced than in topic set {2}. In these cases, the diversity within a topic set is substantially higher than within a given topic category. Hence, it is expected that the

Table 3 The four topic sets created following the merging of related topic categories

Topic sets	Constituent topics	Topic codes	Number of speeches in each topic	Total number of speeches in category
Set {1} Political Issues	Electoral System	(22)	11	155
	Inaugural Speeches	(28)	14	
	Commemorative Speeches	(24)	28	
	Constitutional Issues	(23)	42	
	Procedural Issues	(8)	46	
	Parliamentary Control	(26)	14	
Set {2} Foreign Relations	National Defence	(3)	5	234
	National Issues	(10)	17	
	Foreign Issues	(11)	231	
Set {3} Labour Issues	Employment Issues	(13)	54	90
	Social Security	(5)	19	
	Health Issues	(27)	17	
Set {4} Education & Culture	Education	(20)	70	93
	Educational System Review	(26)	14	
	Culture	(25)	9	

For each of the four selected topic sets, the constituent topics as well as their codes (in accordance to the notation of Fig. 1) are depicted. Additionally, the number of files per topic is shown, together with the total number of files for each topic set (the deviation in the values being attributed to the existence of files, which would be inserted twice in a topic set according to their primary and their secondary topics).

classification accuracy will be reduced when studying the topic sets. Still, apart from topic set {2}, all other sets comprise only speeches due to their categorization in terms of primary topic. Therefore, the speeches are expected to exhibit certain similar characteristics and thus should be classified with a relatively high accuracy.

7.2 Experimental results using topic sets

The classification accuracy for the different topic sets has been investigated using the three variable vectors introduced in Section 6, namely the 85-variable vector, the 25-variable vector and the 16-variable vector. The classification accuracies obtained for the different sets are depicted in Fig. 6a and b, for the training and test sets, respectively. As can be seen, the set classified with the highest accuracy is set {2}, while the remaining sets are classified with a lower accuracy, both with respect to the training and the test data. This is probably attributable to the fact that set {2} consists of 234 speeches, of which the vast majority (231 speeches) belong to the same topic category ('Foreign Issues'). On the contrary, a lower classification accuracy is obtained for the remaining three topic sets, which have a more balanced

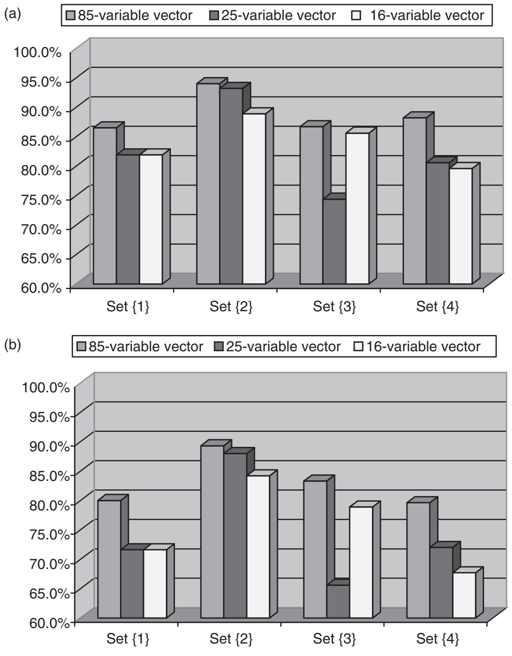

Fig. 6 Classification accuracy of (a) the training data and (b) the testing data (using the cross-validation technique), for the models generated via discriminant analysis for each of the four topic sets. The *F*-to-enter and *F*-to-remove values are equal to 4.00 and 3.99, respectively, for speeches of all orders corresponding to different topic categories

Table 4 Detailed author recognition results for the topic set {4}; the number of speeches from each constituent topic category and each speaker

	Speaker A	Speaker B	Speaker C	Speaker D	Speaker E	All speakers
Topic 19	15	6	11	28	10	70
Topic 24	2	0	0	5	2	9
Topic 26	4	3	3	2	2	14
All topics	21	9	14	35	14	93

Table 5 Detailed author recognition results for the topic set {4}; classification accuracy per speaker for both 25- and 16-variable vectors

	Data vector	Speaker A	Speaker B	Speaker C	Speaker D	Speaker E
Speaker classification accuracy	25-variables	71%	78%	71%	74%	64%
	16-variables	67%	78%	64%	71%	57%

representation of speeches from their constituent topics, and thus a greater variability in the data set.

According to Fig. 6, as the variable vector size is decreased from eighty-five down to twenty-five and finally down to sixteen variables, the classification accuracy is gradually reduced. For the test data, this reduction is substantial, ranging from 5% (in the case of set {2}) up to 12% (in the case of set {3}). This is justified by the fact that the reduced variable vectors have been defined using specific topic categories, which are not identical to the topic sets used in this set of experiments. When the classification task includes these topics (for example in set {2}), the reduced variable set is adequate to distinguish between the patterns with a sufficient degree of accuracy.

On the contrary, when speeches from other topics are classified with this set of variables, the use of the reduced variable vectors results in lower accuracy levels. This raises the question of whether the choice of variables, when defining a reduced variable model, is to a great extent dependent on the topic category and may not be directly generalizable over categories. To elaborate on this issue, a detailed inspection of the classification performance for each topic set was performed. Indicative results are shown in Tables 4–6 for topic set {4}. As can be seen from Table 4, for set {4} the number of speeches per topic category and per speaker varies considerably. The corresponding classification accuracy per speaker is depicted in Table 5. The lowest classification accuracy occurs for speakers C and E, who have relatively few speeches in the topic set. A similar observation applies to the topic classification accuracy, which is presented in Table 6. In this case, the lowest classification accuracy, which is <60%, is observed for topic 24. This is the topic with the lowest number of speeches (only nine out of ninety-three speeches, which correspond to only three out of the five speakers). Such observations are typical of the situation observed in all four topic sets, namely that lower classification rates are reported for topics for which the number of training patterns is low. This indicates that the low classification is due to the limited number of patterns from which to extrapolate an accurate model of the given speaker and topic combination rather than from the use of a non-representative set of variables, when operating on the 25-variable and 16-variable data vectors.

Table 6 Detailed author recognition results for the topic set {4}; recognition results per topic for 25- and 16-variable vectors

		25-variable vector	16-variable vector
Topic classification accuracy	Topic 19	73%	67%
	Topic 24	56%	56%
	Topic 26	79%	79%

A final note regarding the variable vectors involves set {3}, for which the 16-variable vector

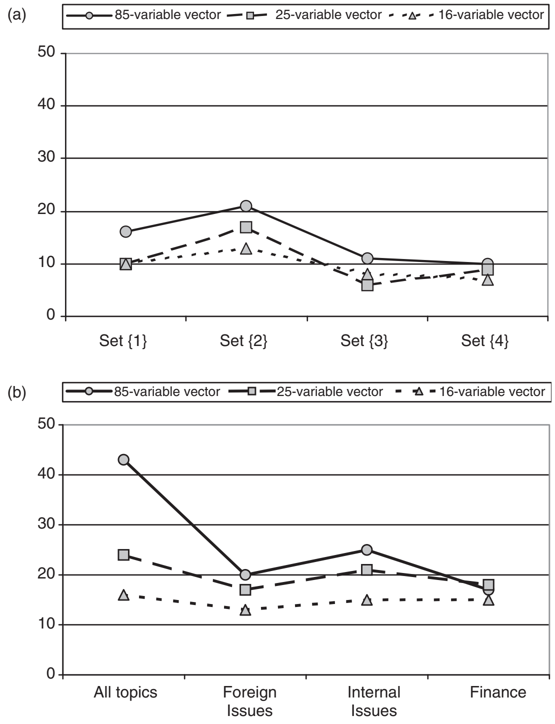

Fig. 7 (a) Number of variables used for the discriminant models generated for the 85-, 25-, and 16-variable vectors (using an *F*-to-enter value of 4.00 and an *F*-to-remove value of 3.99), over the four topic sets. For comparison purposes, the number of variables, when using the topic categories in the experiment described in Fig. 5, is shown in (b)

seems to generate a higher classification accuracy than the 25-variable vector (which would be counter-intuitive, since the 16-variable vector is a subset of the 25-variable vector). However, one should examine the number of variables used in each model. This is depicted in Fig. 7a, for the different combinations of variable vectors and sets. As can be seen, in the case of set {3}, the number of variables in the discriminant model is less for the 25-variable vector, as compared to the 16-variable vector. This indicates that the discriminant analysis process results for the 25-variable vector in a lighter, more tractable model, with fewer variables, which however achieve a sufficient representation of the corpus. The use of fewer variables leads to a reduced overall classification accuracy.

An issue of particular interest is the classification accuracy of the speeches for each particular speaker. A study of the accuracy over the different topic sets indicates that in general most speakers are characterized by similar classification accuracies. However, there exists a tendency, in particular for speakers C and E, to be recognized with a lower accuracy by the discriminant analysis model, for most topic sets.

8 Conclusions

This article has focussed on the issue of recognizing the style of specific authors, within a well-defined register, namely the political speech register. The main topic studied has been to determine whether the definition of subcorpora on the basis of the speech topic has an impact on the accuracy of speaker recognition.

The study initially focusses on whether the classification accuracy with respect to the author identity may be improved, if the topic of a given speech is known in advance, so that the comparison to author models may be restricted to models specialized for a given subject. Different text corpora are studied, corresponding to a variety of topics, in order to determine the effect of topic specialization on the accuracy of classification. The experimental results indicate that knowledge of the topic of a given speech contributes to the more accurate recognition of the author.

In an associated task, the most salient variables are studied within the vector of variables used, for subsets of the corpus, generated so that each subset contains speeches from a single topic. These studies indicate that the salient variables within the variable vector are similar for the different subsets. Two reduced data vectors have been algorithmically defined, containing sixteen and twenty-five variables, respectively. The classification accuracy obtained, even with the smallest data vector, is only 5% less than with the complete vector of eighty-five variables for a given topic category. This indicates that the variables retained in the reduced vectors contain most of the discriminatory information and thus suffice for an accurate classification of text corpora.

In a final series of experiments, the effect of grouping associated topics is studied. It is found that, as the subcorpora from several topics are joined together, the classification accuracy is reduced, as a rule. Additionally, the reduced variable vectors are found to be substantially less effective than the full vector. A factor affecting the classification accuracy has turned out to be the low number of training patterns, from which to extrapolate sufficiently accurate models for the given classification tasks. The best recognition results may be obtained, when using extensive topic categories, with substantial numbers of speeches/texts from each speaker, to allow the successful extrapolation of a representative model for each combination of speaker and topic. This allows the improvement of the classification accuracy over the given text corpus, by taking into account the available topic information in the analysis.

References

Baayen, R. H., van Halteren, H., and Tweedie, F. J. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**(3): 121–31.

Bartholomew, D. (1988). Probability, statistics and theology. *Journal of the Royal Statistical Society – Series A*, **151**(Part 1): 137–78.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2): 219–41.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Collins, J., Kaufer, D., Vlachos, P., Butler, B., and Ishizaki, S. (2004). Detecting collaborations in text. Comparing the Authors' Rhetorical Language Choices in The Federalist Papers. *Computers and the Humanities*, **38**(1): 15–36.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods & Applications*. New York: John Wiley & sons.

Dittenbach, M., Rauber, A., and Merkl, D. (2001). Recent advances with the growing hierarchical self-organising map. In Allinson, N., Yin, H., Allinson, L., and Slack, J. (eds), *Advances in Self-Organising Maps*. London, UK: Springer-Verlag, pp. 140–5.

Elliott, W. E. and Valenza, R. J. (1999). The Professor Doth Protest too Much, Methinks: problems with the "Foster": "Response". *Computers and the Humanities*, **32**(6): 425–90.

Foster, D. W. (1999). The Claremont Shakespeare Authorship Clinic: how severe are the Problems? *Computers and the Humanities*, **32**(6): 491–510.

Gurney, P. J. and Gurney, L. W. (1998). Subsets and homogeneity: authorship attribution in the Scriptores Historiae Augustae. *Literary and Linguistic Computing*, **13**(3): 133–40.

Holmes, D. I. (1985). The analysis of literary style – a review. *Journal of the Royal Statistical Society – Series A*, **148**(Part 4): 328–41.

Holmes, D. I. (1992). A stylometric analysis of Mormon Scripture and related texts. *Journal of the Royal Statistical Society – Series A*, **155**(Part 1): 91–120.

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

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, **13**(3): 111–7.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, **17**(2): 157–79.

Karlgren, J. and Cutting, D. (1994). *Recognising Text Genres with Simple Metrics Using Discriminant Analysis, Proceedings of the COLING'94 Conference*, August 5–9, Kyoto, Japan, pp. 1071–5.

Khmelev, D. V. and Tweedie, F. J. (2002). Using Markov Chains for identification of authors. *Literary and Linguistic Computing*, **16**(4): 299–307.

Kohonen, T., Kaski, S., Lagus, K. *et al.* (2000). Self-organisation of a massive document collection. *IEEE Transactions on Neural Networks*, **11**(3): 574–85.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry I: an application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, **8**(4): 203–9.

Merkl, D. (1999). Document classification with self-organising maps. In Oja, E. and Kaski, S. (eds), *Kohonen Maps*. Amsterdam, The Netherlands: Elsevier, pp. 183–97.

Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference. The Case of The Federalist Papers* 2nd edn. New York: Springer-Verlag.

Rudman, J. (2000). Non-traditional authorship attribution studies: Ignis Fatuus or Rosetta Stone?

Bibliographical Society of Australia and New Zealand Bulletin, **24**(3): 163–76.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society – Series A*, **137**(Part 1): 25–34.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Automatic text categorisation in terms of genre and author. *Computational Linguistics*, **26**(4): 471–95.

Tambouratzis, G. (2006). Assessing the effectiveness of feature groups in author recognition tasks with the eSOM model. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, **36**(2): 249–259.

Tambouratzis, G., Hairetakis, N., Markantonatou, S., and Carayannis, G. (2003). Applying the SOM model to text classification according to register and stylistic content. *International Journal of Neural Systems*, **13**(1): 1–11.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D., and Carayannis, G. (2004a). Discriminating the registers and styles in the modern greek language – Part 1: diglossia in stylistics analysis. *Literary and Linguistic Computing*, **19**(2): 197–220.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D., and Carayannis, G. (2004b). Discriminating the registers and styles in the modern Greek language – Part 2: extending the feature vector to optimise author discrimination. *Literary and Linguistic Computing*, **19**(2): 221–42.

Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural networks in stylometry: the Federalist papers. *Computers and the Humanities*, **30**(1): 1–10.

Waugh, S., Adams, A., and Tweedie, F. (2000). Computational stylistics using artificial neural networks. *Literary and Linguistic Computing*, **15**(2): 187–98.