EMPLOYING STATISTICAL METHODS FOR OBTAINING DISCRIMINANT STYLE MARKERS WITHIN A SPECIFIC REGISTER

Tambouratzis G. INSTITUTE FOR LANGUAGE & SPEECH PROCESSING

Markantonatou S. INSTITUTE FOR LANGUAGE & SPEECH PROCESSING

Vassiliou M. INSTITUTE FOR LANGUAGE & SPEECH PROCESSING

Tambouratzis D. AGRICULTURAL UNIVERSITY OF ATHENS

Περίληψη

Στην παρούσα εργασία παρουσιάζεται μία στατιστική μέθοδος για τη διάκριση μεταξύ συγγραφέων σε ένα συγκεκριμένο, καλά καθορισμένο είδος λόγου της Νέας Ελληνικής. Ο στόχος είναι να προσδιορισθεί (α) κατά πόσον η στατιστική ανάλυση μπορεί να αναδείξει τις υφολογικές διαφορές μεταξύ συγγραφέων και (β) ποια είναι εκείνα τα υφολογικά χαρακτηριστικά που επιτρέπουν έναν αξιόπιστο διαχωρισμό. Για το σκοπό αυτό, μελετήθηκε ένα είδος λόγου για το οποίο συγκεντρώθηκε ένα σώμα με περισσότερα από 1000 κείμενα. Για τη στατιστική επεξεργασία των κειμένων, κάθε κείμενο μετατρέπεται σε ένα διάνυσμα χαρακτηριστικών, τα οποία καλύπτουν μία ποικιλία γλωσσικών φαινομένων. Τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος διαχωρίζει με επιτυχία τα κείμενα βάσει του ύφους των συγγραφέων τους. Μία συστηματική μελέτη των αποτελεσμάτων δείχνει ότι τα χαρακτηριστικά που αναφέρονται στη συχνότητα μερών του λόγου, δομών και λημμάτων είναι εκείνα που παρέχουν τη μεγαλύτερη πληροφορία και οδηγούν σε μία ακρίβεια κατηγοριοποίησης των κειμένων ανά συγγραφέα που ξεπερνά το 90%.

Λέξεις - κλειδιά

υφολογία (Stylometry), αναγνώριση συγγραφέων (author recognition), γλωσσικά χαρακτηριστικά (linguistic features), διακριτική ανάλυση (discriminant analysis)

1. Introduction

The identification of the language style characterising the constituent parts of a corpus is very important to several applications. A simple example encompasses information retrieval applications, where large corpora need to be accessed in order to retrieve documents that are of interest. In such cases, information regarding the linguistic style inherent in each text can be used to improve the accuracy of the search according to the user requirements. The large, constantly-increasing number of texts prohibits the use of manual labelling techniques. Thus, the requirement exists to perform style categorisation in an automated manner. Style categorisation may be at the level of register or at the level of author or both.

A major task of stylometry studies has been the attribution of authorship of mainly literary manuscripts and historical texts. The linguistic features studied exhibit a considerable variety. For instance, Mosteller et al. (1984) have relied on word frequency-of-occurrence histograms. Gurney et al. (1998) have followed a broadly similar approach, using lemma-based features to perform authorship studies of the Scriptores Historiae Augustae corpus. Elliot et al. (1997) and Foster (1999), who – among other researchers – have addressed the issue of authorship of Shakespeare′s plays, have introduced idiosyncratic variables such as sets of rare words or enclitic and proclitic micro-phrases.

Stylometric studies have employed (i) traditional linguistic techniques, (ii) statistics-based techniques and, more recently, (iii) techniques inspired from the artificial intelligence paradigm (Holmes, 1998). The present article focuses on the use of statistical techniques to perform author style identification tasks, because statistical methods are based on a theoretical foundation and can provide a reliable estimate of the classification certainty.

An extensive stylometric study for texts written in Modern Greek was performed (Tambouratzis et al. 2000), based on statistical techniques. This research was conducted in two successive steps: (1) using cluster analysis in order to perform register discrimination and (2) using discriminant analysis in order to perform author discrimination for texts within one given register. In this study, features encompassed a variety of linguistic properties, such as frequency-of-occurrence of Parts-of-Speech (PoS) and macro-structural characteristics reflecting the length of sentences and words. Information reflecting the diglossia situation in Modern Greek, as expressed by the contrast between *Katharevousa* and *Demotiki* at the level of verbal morphology, provided the most important features for the register discrimination task, while PoS information was less important.

From the register discrimination experiments it turned out that the Minutes of the Hellenic Parliament form a well-defined register of Modern Greek. Consequently, we exploited texts from this register to perform our author discrimination experiments. It was found that diglossia information was inappropriate for author discrimination within the Parliament Minutes, because the minutes were edited before being published, resulting in, at least morphologically, homogeneous texts. In this context, other linguistic features, including PoS information and certain indicators of discourse preferences, provided the highest discrimination accuracy. Hence, our focus has shifted towards (i) using a larger set of linguistic features that enhance the author discrimination accuracy and (ii) determining the importance of the various features for the author discrimination task.

2. An Overview of Stylistics Studies in the Greek language

Up to date, only few stylometric studies have been performed for texts written in Modern Greek. Mikros et al. (2000) employed discriminant analysis to perform register discrimination on corpora consisting of newspaper articles on selected topic areas. The features chosen quantify over the diglossia phenomenon of Modern Greek and also reflect macro-structural properties. Stamatatos et al. (2001) studied the discrimination of ten regular columnists of a specific Greek newspaper, who specialise in different topics, employing features that quantify over information extracted automatically via an NLP tool performing morphological and shallow-syntactic analysis, coupled with word frequency information and macro-structural information.

These stylometric experiments involved a limited number of texts of a relatively homogeneous size, spanning more than one registers. The set-up chosen here is substantially different. The specific register is characterised by a well-defined sub-language, while all texts within this register and produced within a given time period have been extracted for use.

Consequently, the text size ranges from less than 300 words to over 5000 words. In our view, this is a more realistic set-up, which moves away from the confines of a controlled laboratory experiment and towards a real-world application.

3. Distinguishing Personal Author Styles within One Register

Register separation experiments (Tambouratzis et al. 2000) have shown that the Parliament register consists of texts, which exhibit a very high degree of similarity in the pattern space. Our experience with (i) register and (ii) author discrimination experiments has shown that the latter task is undoubtedly more difficult. Therefore, we assembled a considerably expanded set of linguistic features in comparison with the set of features used in our earlier experiments, in an effort to achieve high recognition accuracy and to investigate the contribution of each category of linguistic features.

3.1. Composition of Corpora from the Hellenic Parliament Register

The Parliament register, which has been used as the material for a series of experiments, possesses a set of desirable properties. It is well-defined in terms of content and is indicative of a sub-language of Modern Greek. The texts within this register/ corpus have been meticulously prepared by the Parliament Secretariat, so that the amount of inherent errors is very low. The Parliament register is represented by a corpus of sufficient size and contains texts by several hundred authors, posing a considerable challenge in the area of stylistics. In the experiments reported here, the source domain has been set to the period 1996-2000, between two consecutive general national elections. During this period, the Parliament composition remained unchanged and five political parties had a parliamentary representation. One speaker from each political party was selected for this study. By extracting all speeches for these speakers, a corpus has been formed containing in excess of 1,000 texts. This corpus is designated as Corpus IV, this designation indicating its relation to Corpus II and Corpus III, which were used in earlier experiments (Tambouratzis et al. 2000). The composition of Corpus IV is summarised in Table 1.[i]

Table 1.Composition of Corpus IV, containing all speeches from the period 1996-2000

Corpus IV	*1996-1997*	*1998*	*1999*	*2000*	*1996-2000*	*Size (in words)*
Speaker A	161	107	125	25	418	463,680
Speaker B	36	28	18	3	85	177,853
Speaker C	81	71	67	26	245	241,882
Speaker D	72	49	30	3	154	217,305
Speaker E	16	42	38	7	103	190,601
Total	**366**	**297**	**278**	**64**	**1005**	**1,291,321**

3.2. Set of Variables Studied

A core set of 46 variables, reflecting selected linguistic features, was used for style identification (Tambouratzis et al. 2000). These variables can be grouped in seven classes:

1. Seventeen morphological and lexical variables expressing the Katharevousa / Demotiki contrast, comprising variables:
   - ranging over verbal endings, obtained from Clairis et al. (1999);
   - ranging over infixes in the past tense forms;
   - measuring the distribution of negative words of Katharevousa (*ud′is* – *"ουδείς"*, *aneu* – *"ἀνευ"*);
   - measuring the relative distribution of the anaphoric pronouns *op′ios* – *"οποίος"* (Katharevousa) and *p′u* - *"που"* (Demotiki)).
2. Six morphological variables expressing discourse tendencies. These range over the verbal grammatical features *person* and *number* (for the indicative and the subjunctive mood) and indicate the manner in which speakers address their audience.
3. Three lexical variables quantifying the distribution of negation particles.[ii]
4. Six variables, which quantify over macro-structural properties, including average sentence length, average word length, as well as average number of specific punctuation marks.
5. Eleven variables expressing the frequency of occurrence of the Part-of-Speech categories.
6. Two (micro-structural) variables quantifying over the use of the genitive case for nouns and adjectives.[iii]
7. One variable reflecting the order of the given speech in the daily schedule, i.e. whether it was the first speech of the particular speaker that day (hereafter denoted as "protoloyia"). After the first round of "protoloyiai", the discussion normally continues with a second round of speeches. As a rule, "protoloyiai" are prepared in advance while, as the order of speech increases, the corresponding speech becomes more spontaneous.

This set of 46 variables has been expanded by introducing the following variables:

8. Three variables belonging to the set of negative words of Modern Greek reflecting the Katharevousa / Demotiki contrast and resulting in eight negation-related variables in total.
9. Two variables encoding discourse tendencies (see point 2 above), again retrieved from the set of verbal endings, this time quantifying over the imperative mood.
10. Eighteen variables, which provide more accurate measurements regarding the sentence/word length and thus a more detailed representation of the text properties. These include:
    - 5 variables describing the distribution of word length in letters;
    - a single variable counting the overall amount of punctuation marks; and
    - 12 variables describing the distribution of sentence length in words.
11. Seventeen variables counting the occurrences of specific lemmata.

The aforementioned lemmata were selected with an algorithmic procedure, which determined words characteristic of a dichotomy of speakers. Although this measure might seem text-dependent, the majority of words chosen belong to the set of the so-called "functional words", which have been shown to be style indicators cross-linguistically (Mosteller et al. 1984). To select the lemmata, initially, a small number of texts of known authorship, with a minimum size of 1,000 words, are chosen from the corpus to ensure that the relative frequencies of lemmata are representative. The texts are lemmatised and the lemmata ordered according to their frequency-of-occurrence. Then, lemmata are searched for that have a consistently low frequency for (at least) one author, while having a consistently high frequency for (at least) another author. The number of lemmata chosen is limited to 20% of the total number of variables, to ensure that the fraction of lemma-based variables is bounded with respect to the entire set of variables.

4. Experimental Results Using Discriminant Analysis

Initial experiments (Tambouratzis et al. 2000) on smaller corpora used the 46-variable vector to generate discriminant functions that identify the author of a given text. Two different approaches were followed, one involving a full discriminant model comprising all 46 variables, the other involving a reduced model following both forward and backward stepwise analysis. In the cases of the forward and backward models, the values of the F parameter used to both introduce and remove a variable from the discriminant model were set to 4. Out of the 46 variables available in the data vector, ten variables are retained in both reduced models (whose sizes are 14 and 13 variables respectively), illustrating the importance of the specific variables in the discrimination task. These common variables comprise PoS counts, punctuation marks, verbal discourse features and the average number of letters per word. On the contrary, diglossia-reflecting variables were not successful in discriminating the author styles.

The experimental results indicated that the performance of this model is improved if:

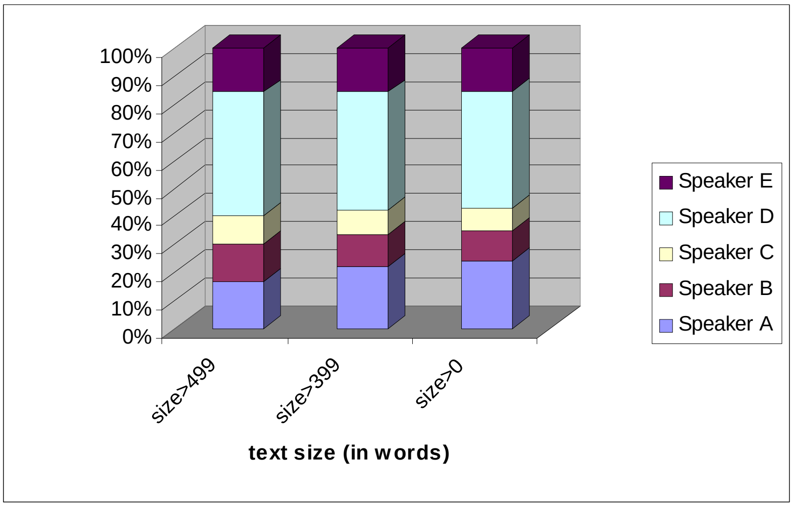
1.The order of speech delivery is taken into account.

2.The corpus comprises only sessions where more than one speaker have delivered speeches.

In the experiments described here, the aim has been to study how these results scale up, when a more extensive corpus is used. Furthermore, by introducing additional linguistic features in the representation of each text and performing a variety of experiments, a more thorough investigation of the role of variables is possible.

The order of each speech is used as a filter that determines subsets of the Corpus, thus leading to a series of experiments, where either only "protoloyiai" or, alternatively, speeches of any order were selected. These subsets were then processed using discriminant analysis with the SPSS statistical package. The distribution of speeches per speaker varied as the minimum allowable text size was modified, as shown in Figure 1.

Figure 1 - Distribution of speeches of all orders per speaker for different minimum text sizes.



Initially, a forward stepwise discriminant analysis was performed using the 85-variable vectors. This analysis suggested that 25 linguistic variables (which are listed in Table 2) were sufficient to describe the data of Corpus IV. Then, more extensive classification experiments were carried out using this set of 25 variables, in order to determine the classification accuracy of the model. More specifically, three different cases were studied:

(i) when only texts containing more than 500 words were processed, shorter texts being discarded;

(ii) when only texts containing more than 400 words were processed;

(iii) when all texts were used, irrespective of their size.

Table 2. Variables retained in the reduced model for Corpus IV when using a stepwise discriminant analysis (85 variables), using F-to-enter and F-to-remove values of 4.

Variable	Order of introduction	Meaning
ADVERBS	1	Number of Adverbials
ENAS	2	Occurrences of Lemma “ἐνας” (a, one)
VB03PL	3	Number of Verb Endings, 3rd person plural
LETRPERWO	4	Average Number of Letters per Word
MOU	5	Occurrences of Lemma “μου” (me, mine)
CONJUNCT	6	Number of Conjunctions
KYRIOS	7	Occurrences of Lemma “κὐριος” (mister)
REST	8	Number of Tokens Unidentified by Tagger
VERBS	9	Number of Verbs
ARTICLES	10	Number of Articles
W1_3LPER	11	Number of Words with less than 4 letters
VB01PL	12	Number of Verb Endings, 1st person plural

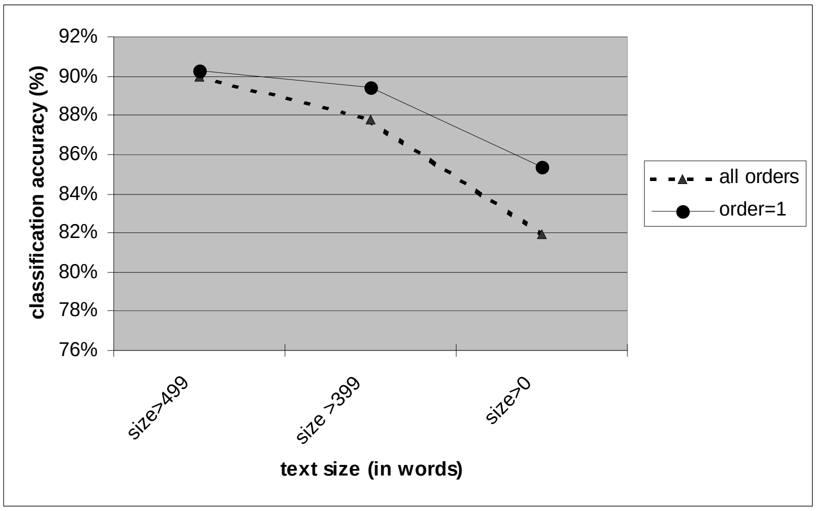
VB03SG	13	Number of Verb Endings, 3rd person singular
VB01SG	14	Number of Verb Endings, 1st person singular
VB02PL	15	Number of Verb Endings,2nd person plural
S76_100W	16	Number of Sentences with 76 =<Words=<100
SE2_5WP	17	Number of Sentences with 2=<Words=<5
YPARXW	18	Occurrences of Lemma "υπάρχω" (to exist)
DASHES	19	Number of Dashes
THELW	20	Occurrences of Lemma "θέλω" (to want)
PRONOUN	21	Number of Pronouns
EGW	22	Occurrences of Lemma "εγώ" (I)
ORDER	23	Order of speech delivery
LAOS	24	Occurrences of Lemma "λαός" (people)
ELLHNAS	25	Occurrences of Lemma "Έλληνας" (Greek)

A further variation involved studying whether a significant difference was observable, when speeches of all orders were processed in comparison with the study of speeches of order 1, for each of the three aforementioned cases. These datasets are compared and contrasted in Table 3.

Table 3. Number of speeches for different datasets generated from Corpus IV, as a function of the order of speech and the minimum text length.

		Minimum allowable text length in words		
		500 words	400 words	unconstrained
Order of speech	order set to 1	372	433	541
	Unconstrained order	579	698	1005

Figure 2 - Classification accuracy for (i) "protoloyiai" and for (ii) speeches of all orders, for different minimum text sizes.



4.1. Experimental Result Assessment

The classification accuracy obtained for each of the aforementioned datasets, when a leave-one-out strategy is adopted, is shown in Figure 2. The classification accuracy is higher when speeches of order 1 are exclusively used than when speeches of all orders are processed. When the minimum text size is restricted to 500 words, if only texts of order 1 are processed as compared to processing texts of all orders, the improvement obtained is marginal (in both cases the accuracy being very close to 90.0%). However, as the minimum text length becomes smaller, the margin becomes wider. Thus, in general, speeches of order 1 are classified more accurately than speeches of lower orders. Furthermore, the text size seems to have a substantial impact on the classification accuracy, which is considerably improved when the text length is constrained to a minimum of 500 words. However, even when the text size is unconstrained, a classification accuracy exceeding 80% is obtained, indicating the effectiveness of the proposed method.

In order to quantify the system performance, some indicative cases are presented in more detail. These are cases with a comparatively high error rate, so as to determine the speakers that are more difficult to classify. In Table 4, the confusion matrix is displayed when all ″protoloyiai″ are processed irrespective of their text length, the total accuracy exceeding 85.4%. As can be seen, speakers C and E are the most difficult to classify correctly. To confirm these results, a 10-fold evaluation method is employed where 90% of the dataset is used to generate a model, which is then applied to classify accurately the remaining 10% of the dataset texts (that are unlabeled as far as the discriminant analysis was concerned). The result obtained for the aforementioned setup is 84.7%, indicating a deviation in accuracy from the leave-one-out approach of less than 1%, thus confirming the accuracy of the discriminant analysis model.

Table 4 - Confusion matrix for ″protoloyiai″ from the period 1996-2000, irrespective of the text size, using the 85-variable data vector

		Speaker identity predicted by the model				
		A	B	C	D	E
Actual speaker identity	A	92.3%	3.3%	3.3%	1.1%	0.0%
	B	3.6%	89.2%	3.6%	3.6%	0.0%
	C	2.1%	10.3%	81.3%	6.3%	0.0%
	D	2.7%	5.1%	5.1%	85.5%	1.6%
	E	2.2%	14.0%	3.3%	2.2%	78.3%

4.2. The Dominant Features

The results of the forward step-wise model provide an insight to the variables that are more useful when performing speaker discrimination studies (shown in Table 2). As summarised in Table 5, in terms of type, the majority of these variables are:

- lemmata,
- Part-of-Speech frequencies and

- verbal features referring to the person/ number features rather than the Demotiki/Katharevousa contrast.

Thus, for the Hellenic Parliament register, individual speaker styles may not be determined on the basis of morphological features expressing the contrast Katharevousa/ Demotiki. This must be attributed to the fact that the Parliament texts undergo intensive editing towards a well-established sub-language, which homogenises the morphological profile of the texts.

Macro-structural properties such as the average word length and the frequency of punctuation marks turn out to be important. The same holds for PoS counts, such as the frequency of use of articles, conjunctions, adjunctions and - especially - verbs. The distribution of verb forms into persons and numbers seems to be important, though the exact verb-related variables selected differ depending on the exact set of speeches used (these variables are, of course, complementary). Among the different indices of information concerning style, the verb distribution is the most explicit and reflects the choices made by the speakers in addressing their audience. For instance, it shows whether a certain speaker adopts a more personal and direct style using predominantly a plural/ second person attitude to address his/ her colleagues or whether he/ she talks in an impersonal and more indirect style, preferring a singular/ third person attitude. Part-of-speech information, on the other hand, is less transparent: for instance, the classes of PoS that seem to play a role, possibly indicate that the distinction is influenced by the complexity of structures, that is, whether the speakers extensively exploit subordinate clauses or not. Macro-structural properties offer effective quantitative information, which does not, however, represent the linguistic properties of the texts at hand in a transparent manner.

One of the most interesting findings of this research is that it is significant whether the speaker delivers a ″protoloyia″ or not. For a large corpus, ″protoloyiai″ can be classified at an average rate exceeding 90%, while mixed deliveries result in a lower rate, as low as 82%. This may be caused by two factors:

1. ″Protoloyiai″, representing longer stretches of text, are more characteristic of a given speaker.
2. Speakers prepare meticulously for their ″protoloyiai″, while their other speeches are more spontaneous, tending to contain patterns of speech shared by all (or most of) the parliament members.

Other complementary statistical approaches to discriminant analysis have been investigated with respect to the author recognition task, these involving the combination of (i) factor analysis and discriminant analysis or (ii) analysis of variance and discriminant analysis. The aim of these experiments has been to determine which variables are the most salient within the 85-variable data vector. All methods result in a data vector comprising 25 variables. In each case, the data vector comprises a similar number of variables from each variable group, while the actual variables retained in the final model are in most cases the same.

Table 5. Distribution of the 25 variables used in the reduced discriminant analysis model to identify the 5 speakers.

	Variable type	Original number of variables in class	Variables retained following discriminant analysis
1	Negation	8	0
2	Lemmata	17	8
3	Micro-structural	3	0
4	Macro-structural	24	2
5	Verbal	22	8
6	Part-of-speech	11	7
	Total	**85**	**25**

Lemmata represent the most important group of variables in the final model. Also, PoS and verbal features contribute strongly. Other micro–structural properties and the macro-structural properties are not of importance. Finally, the negation features are not used at all, illustrating that negation is not useful for author discrimination tasks in the Greek language, at least for the specific register. This result does not agree with the observations of Labbe (1983) on the French political speech register, where negation seems to be a prominent style feature.

Table 5 indicates whether the salient information residing in each group may be compressed, so that it is expressed by only a handful of variables. As can be seen, in particular the PoS and lemmata groups require a proportionately higher amount of variables to convey the existing information. On the contrary, the macro-structural and verbal groups require a proportionately smaller number of variables.

5. Conclusions

In this article, we have reported on ongoing research on the issue of author style categorisation in written Modern Greek. The study has focused on the author identification task for documents belonging to a specific register, namely the political speech register, with texts being obtained from the Minutes of the Hellenic Parliament. The experimental results indicate that consistent author discrimination with a high degree of accuracy is feasible. A study of the sub-language used in the Minutes of the Hellenic Parliament sessions has revealed that the resulting corpus is homogenised to a large extent, both as a result of the formal language employed by members of Parliament and the extensive post-editing performed on the texts. Compared to earlier experiments, an expanded set of linguistic features has been used, incorporating both detailed structural features and lemmata that have been algorithmically derived from a small set of sample texts.

The experimental results show that diglossia-specific features are not effective in the author discrimination task. Using discriminant analysis, author identification is achieved with an accuracy often exceeding the level of 90% for five speakers. A systematic evaluation process indicates that the PoS, macro-structural and lemma variables - determined with an algorithmic procedure on a very small corpus of texts - are the most important variables for author discrimination.

Alternative approaches involving the combination of statistical techniques have been studied in order to determine the classification accuracy of the variable sets. The results have confirmed the effectiveness of structural, PoS and lemma-based features, indicating their suitability in discriminating between authors for the given register.

Acknowledgements

The authors wish to acknowledge the assistance of the Secretariat of the Hellenic Parliament in obtaining the session transcripts. Additionally, they thank Dr. Harris Papageorgiou and Mr. Prokopis Prokopidis of ILSP for generating the lemmatised versions of the parliament transcripts. Last but not least, the authors wish to acknowledge the considerable contribution of Mr. N. Hairetakis in creating the computer programs and providing the effort required to transform the texts into feature vectors for further processing.

Notes

[i] Corpus II contained all speeches from the same five speakers over a 16-month period spanning years 1999 and 2000. Corpus III referred to the same period, but only contained "protoloyiai" (see section 3.2 for an explanation of the term) from sessions, when at least two of the speakers presented their views.

[ii] Negation has been shown to be a style marker in political speech, at least in the French language (Labbe, 1983).

[iii] These variables indirectly quantify over the frequency of use of structures headed by a predicative noun or adjective.

References

Clairis, C. & Babiniotis, G. 1999. "Grammar of Modern Greek – II Verbs". *Ellinika Grammata*, Athens (in Greek).

Elliott, W.E. & Valenza, R.J. 1997. "Glass Slippers and Seven-League Boots: C-Prompted Doubts about Ascribing a Funeral Elegy and A Lover's Complaint to Shakespeare". *Shakespeare Quarterly* 48. 177-207.

Foster, D.W. 1999. "The Claremont Shakespeare Authorship Clinic: How Severe are the Problems?" *Computers and the Humanities* 32. 491-510.

Gurney, P.J. & Gurney, L.W. 1998. "Authorship Attribution of the Scriptores Historiae Augustae". *Literary and Linguistic Computing* 13. 119-131.

Holmes, D.I. 1998. "The Evolution of Stylometry in Humanities Scholarship". *Literary and Linguistic Computing* 3. 111-117.

Labbe, D. 1983. "Francois Mitterrand. Essai sur le discourse". La Pensee Sauvage, Grenoble.

Mikros, G. & Carayannis, G. 2000. "Modern Greek Corpus Taxonomy". *Proceedings of the LREC-2000 Conference*, Athens, Greece, 31 May - 2 June, Vol. 3. 129-134.

Mosteller, F. & Wallace, D.L. 1984. "*Applied Bayesian and Classical Inference. The Case of The Federalist Papers*". 2nd edition. Springer-Verlag, New York.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. 2001. "Computer-Based Authorship Attribution without Lexical Measures". *Computers and the Humanities* 35. 193-214.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D. & Carayannis, G. 2000. "Discriminating the Registers and Styles in the Modern Greek Language". *Proceedings of the Workshop on Comparing Corpora* (held in conjunction with the 38th ACL Meeting), Hong Kong, China, 7 October. 35-42.