Modern Greek Corpus Taxonomy

George Mikros* and George Carayannis*

* Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
{gmikros, gcara}@ilsp.gr

Abstract

The aim of this paper is to explore the way in which different kind of linguistic variables can be used in order to discriminate text type in 240 preclassified press texts. Modern Greek (MG) language due to its past diglossic status exhibits extended variation in written texts across all linguistic levels and can be exploited in text categorization tasks. The research presented used Discriminant Function Analysis (DFA) as a text categorization method and explores the way different variable groups contribute to the text type discrimination.

1. Introduction

Both written and spoken Modern Greek (MG) exhibit extended variation due to the diglossic status of the national language for over a century (Ferguson, 1959). Although, Greece has accepted literary Dhimotiki (D) as official language from the mid 70' a large number of phonological, morphological and syntactic phenomena of Katharevusa (K) have survived in the current language and continue to coexist in the same environment of use. MG has now found a balance between its social dialects and the observed variation is utilized stylistically. The goal of the present research is to comprise an inventory of MG language variation phenomena in order to explore possible intra and extra linguistic factors that may condition their usage and use them as indicators of text type. In addition to the previous mentioned linguistic variables, we will include some other variables related to the word and text organization trying to find the optimum combination of discriminatory parameters to text type classification.

2. Corpus Selection

In order to investigate the variability of written MG we examined 240 press texts of approx. 0.56 M words size. This sample uses the PAROLE project classification scheme (Gavrilidou et al. 1994) and is part of the corpus of MG texts that has been developed by the Institute for Language and Speech Processing (ILSP). The sample we constructed consists of 4 text types (BUS, HUM, LEI, SOC) with 30 texts in each one of them. The only a priori constraint we posed was that the size of all the texts would be at least 1000 words for each one. The distribution of words and texts in the sample according to the text type is given to the Table 1.

The text types labels were kept as they were originally given in the PAROLE project and are described as follows:

BUSiness: Articles related to economy.
HUManities: Articles regarding humanities and culture.
LEIsure: Articles which relate to leisure activities, hobbies, vacations etc.
SOCiety: Articles regarding politics and society issues.

3. Variables Selection

Written variation has been used successfully in order to construct variables for text categorization issues (Biber, 1988; Biber, 1995; Forsyth, 1995). It has been reported (Forsyth, 1997) that in the majority of the stylometric studies the choice of variant forms, which are intended to be used as style indicators are selected subjectively. However, in the written MG corpus we can objectively define linguistic variables exploiting the coexistence of K and D linguistic elements in the same text.

NEWSPAPER	**ELEFTHEROTYPIA**		
TOPIC	**Texts**	**Words**	**%**
BUS	30	43,570	19.7
HUM	30	61,450	27.8
LEI	30	26,983	12.2
SOC	30	89,174	40.3
Total	***120***	***22,177***	***100***
	TO VIMA		
TOPIC	**Texts**	**Words**	**%**
BUS	30	82,221	24.3
HUM	30	106,620	31.4
LEI	30	62,762	18.5
SOC	30	87,431	25.8
Total	***120***	***339,034***	***100***
	Total		
TOPIC	**Texts**	**Words**	**%**
BUS	60	125,791	22.5
HUM	60	168,070	30.0
LEI	60	89,745	16.0
SOC	60	176,605	31.5
Total	***240***	***560,211***	***100***

Table 1: Distribution of words and texts across text type categories.

3.1. Language specific variables

In order to exploit the MG inherent variation related to the distribution of K/D elements we constructed a list of linguistic variables, which traditionally were considered as

markers of K/D linguistic usage[1]. The elements chosen do not form an exhaustive listing of all the possible markers for the previously mentioned dimension. They do, however, comprise a fairly representative sample of characteristics from most linguistic levels (phonology, morphology, syntax) and are frequently even in small texts. The list of the investigated elements is shown in Table 2 and is based mainly on Papatzikou - Cochran (1997).

Linguistic Level	*Variables*
Phonology	**Final –n rule <%n>[2]**
D variant	Deletion of final -n of proclitic words before continuant consonants
K variant	Sustain of final –n in all environments
Morphology	**Genitive of 3rd inflection class feminine nouns <%gen>**
D variant	Endings in -is "-ης"
K variant	Endings in -eos "-εως"
	Adverbial endings <%adv>
D variant	Endings in –a "-α"
K variant	Endings in –os "-ως"
Syntax	**Use of relative pronouns <%pron>**
D variant	pu "που"
K variant	opios, -a, -o "οποίος, -α, -ο"

Table 2: List of investigated language specific variables

Previous studies have confirmed that the above variables exhibit considerable socio-stylistic variation in both written and spoken MG (Mikros & Carayannis, forthcoming; Mikros, 1999; Andriomenos, 1999, Alexiou, 1982). In the present study we calculated the D and K variant occurrences separately and their relative ratio (nD/nK). The frequency of the D and K elements was normalized to a text length of 1000 words in order to have comparable results among texts of different size.

3.2. Language independent variables

Besides language specific variables we calculated for each text a number of language independent variables which are commonly used in text categorization research. Following Karlgren (1999) we have clustered them according to their application domain, that is, word based and text based variables.

3.2.1. Word domain variables

Word based measures have been calculated for all the texts of our corpus. The basic variables that were selected were the following:

a) Word length: For each text we calculated the average word length in characters. The stylometric effect of the word length has been extensively documented in English and other European languages (among others Ziegler, 1998; Wimmer et al., 1994). However, it has not been used in MG stylometry studies yet and we do not know its distribution properties as well as its correlation with other micro and macro textual characteristics.
b) Distribution of word length: We calculated separately for each text the number of one character, two character words (cw) etc. capturing in this way the word length distribution of each text (Uhlirova, 1995). Following this procedure we created 14 variables which represent the number of tokens of word length 1 – 14 characters long.
c) Lexical density: The ratio of content vs. functional words for each text was calculated.

3.2.2. Text domain variables

Text based measures were also calculated. The main set of variables selected for the present research are the following:

a) Number of sentences[3]: The total number of sentences for each text was calculated.
b) Sentence length: The sentence length is also commonly used in stylistic comparisons and is directly related to the dimension of text complexity (Mikk, 1995).
c) Standard deviation of sentence length
d) Number of paragraphs: The use of paragraphs can reveal significant aspects of macrotextual organization.
e) Paragraph length: The average length of the paragraphs for each text.
f) Standard deviation of paragraph length.
g) Number of subordinate clauses: This is also an index that correlates to the sentence structure complexity and broad text type characteristics (Tesitelova, 1992).
h) Number of tense and mood particles: We calculated the number of future tense particles "tha" and the number of the subjunctive particle "na".

4. Methodology

4.1. Investigation of text normalization

Since most of our language dependent variables were carrying a certain amount of ideological load we decided to conduct a preliminary investigation in their percentages of occurrence in the two newspapers. In order to have a clear picture of the equilibrium between variable linguistic elements we had to investigate if the newspapers in which the articles had been publicized had exerted a preliminary normalization towards the one or the other form regarding their language policy. This is very important since we

[1] For a similar approach to MG text categorization using Cluster Analysis see Tambouratzis et al. (2000).

[2] The string inside the angled brackets denotes the code name of the variable. The D type of the variable is denoted by the prefix d or k before the code name e.g. the D variant "pu" of the variable "Use of relative pronouns" is marked as "d<%pron>". When a variable is presented in angled brackets whithout prefix (e.g. <%adv>) it denotes the ratio of D vs. K elements of this variable.

[3] Every variable that involves counting of a specific element is normalized to a text length of 1000 words.

need to know which linguistic markers exhibit real variation in the corpus and which have been normalized in order to conform to specific rules. Furthermore, the inclusion of any normalized variable will introduce bias in the statistical analysis and should not be subtracted from the list of the predictor parameters.

The results showed that final –n occurrences were normalized towards the K variant in the newspaper "TO VIMA" which exhibits zero D variant. The specific variable was not included in the subsequent analysis.

4.2. Newspaper and topic interaction

A second step in our methodology was to investigate the way in which all variables of our study were correlated with the topic and the publisher. Since we had two subcorpora of texts, one for each newspaper, we had to detect which features are used systematically different between the newspapers, between the different topics and between the combinations of them. Our aim was to investigate their behaviour, so we could select a subset of independent variables that is related directly with the research factors "Topic" and the "Topic X Newspaper" and leave out the variables that correlate to the "Newspaper" only. This kind of filtration would improve the performance of the statistical categorization since would include only those variables that contributed significantly to the text type discrimination.

In order to investigate the interaction of newspaper and text topic we used MANOVA analysis. The overall interaction was found statistically significant (Wilks' Lambda 0.18; $p<0.000$). Examining the detailed results of the analysis we constructed the variable matrix which was directly related to the topic, to the newspaper and the interaction of them. The variables that were statistically significant different due to the research factors ("Newspaper", "Topic", "Newspaper X Topic") are marked with an asterisk (*) in the Table 3.

Variables	News-paper	Topic	News-paper X Topic
Text size	*	*	*
Word length		*	*
Number of sentences	*	*	*
Sentence length	*	*	*
sd. Sentence length			*
Number of paragraphs	*		
Paragraph length	*	*	*
sd. Paragraph length	*	*	*
1 cw	*	*	
2 cw	*	*	
3 cw			
4 cw		*	*
5 cw	*	*	*
6 cw		*	
7 cw		*	
8 cw		*	*
9 cw		*	*
10 cw		*	*
11 cw		*	*
12 cw	*	*	
13 cw	*	*	
14+ cw		*	
d<%pron>		*	*
k<%pron>		*	
<%pron>		*	
d<%gen>		*	
k<%gen >	*	*	*
<%gen >	*	*	
d<%adv>	*	*	
k<%adv >	*	*	
<%adv >	*		
d<%n>	*		
k<%n>	*	*	*
<%n>	*	*	
Number of subordinate clauses		*	*
Number of "na" clauses		*	
Number of "tha" clauses	*		
Lexical density	*	*	*

Table 3: Statistical significance of the independent variables for one-way main effects of the factors "Newspaper" and "Topic" and two-way interactions between them.

From the above matrix we selected only the variables that were statistically significant at the "Topic" or the interaction "Newspaper X Topic". The variables that were subtracted following this procedure were: "sd. sentence length", "number of paragraphs", "3cw", "<%adv>", "d<%n>", "tha clauses".

5. Experimental results

5.1. Classification algorithm

A number of classification techniques have been applied to text categorization and involve mainly Linear Prediction Models, Neural Networks, Bayes Belief Network, Nearest Neighbor Classifier, Decision Trees, Rule Learning algorithms and Inductive Learning techniques. The contrastive analysis of their classificatory power (Yang, 1997) has revealed that each technique performs different regarding the corpora involved. Since our corpus consists of press texts of specific newspapers we considered as a first approach to adopt a Linear Prediction Model. As main classificatory technique we used Discriminant Function Analysis (DFA), which is a method of assigning cases in predetermined groups using a set of predictor variables. It has been employed in text categorization with success (Karlgren, 1994) and there are studies (Hull, Pedersen, Schütze, 1996) that claim its superiority over other Linear Predictive Models such as Logistic Regression.

DFA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a priori defined groups.

Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance (Hair et al., 1995). If the dependent variables have more than two categories DFA will calculate C-1 discriminant functions, where C is the number of categories. Each function allows us to compute discriminant scores for each case for each category, by applying the formula:

$D_{jk} = a + W_1X_{1k} + W_2X_{2k} + ... + W_nX_{nk}$

where

D_{jk}= Discriminant score of discriminant function j for object k.

a= intercept

W_i= Discriminant weight for the independent variable i

X_{ik}= Independent variable i for object k

In order to validate the DFA results we used cross-validation procedures so that we can determine the classification accuracy rate. The selected procedure was the *U*-method which is based on the "leave-one-out" principle (Huberty, Wisenbaker, Smith, 1987). Using this method, the discriminant function is fitted to repeatedly drawn samples of the original sample. Estimates k-1 samples, eliminating one observation at a time from a sample of k cases.

5.2. Classification precision

The classification results for the whole corpus are given to the Table 4:

Predicted Group Membership (Whole Corpus)						
	Topic	**BUS**	**HUM**	**LEI**	**SOC**	***Total***
Original Count	**BUS**	49	6	2	3	60
	HUM	0	50	2	7	59
	LEI	1	4	53	2	60
	SOC	9	7	3	41	60
%	**BUS**	81.7	10	3.3	5	100
	HUM	0	84.7	3.4	11.9	100
	LEI	1.7	6.7	88.3	3.3	100
	SOC	15	11.7	5	68.3	100
Cross-validated Count	**BUS**	42	6	7	5	60
	HUM	2	44	4	9	59
	LEI	4	5	48	3	60
	SOC	15	13	4	28	60
%	**BUS**	70	10	11.7	8.3	100
	HUM	3.4	74.6	6.8	15.3	100
	LEI	6.7	8.3	80	5	100
	SOC	25.0	21.7	6.7	46.7	100

Table 4: DFA classification results in the whole corpus

The overall performance was 80.8% of original grouped texts correctly classified (67.8% correct classification in cross-validation). The resulting scatterplot of the first 2 discriminant functions is given to the Figure 1.

In addition we performed a separate DFA for the subcorpus of each newspaper. The classification results are given to the Tables 5 and 6 respectively.

Predicted Group Membership (TO VIMA)						
	Topic	**BUS**	**HUM**	**LEI**	**SOC**	***Total***
Original Count	**BUS**	25	1	1	3	30
	HUM	1	21	4	3	29
	LEI	3	1	17	3	24
	SOC	2	2	3	23	30
%	**BUS**	83.3	3.3	3.3	10	100
	HUM	3.4	72.4	13.8	10.3	100
	LEI	12.5	4.2	70.8	12.5	100
	SOC	6.7	6.7	10	76.7	100
Cross-validated Count	**BUS**	20	1	3	6	30
	HUM	2	18	5	4	29
	LEI	4	2	15	3	24
	SOC	7	8	3	12	30
%	**BUS**	66.7	3.3	10	20	100
	HUM	6.9	62.1	17.2	13.8	100
	LEI	16.7	8.3	62.5	12.5	100
	SOC	23.3	26.7	10	40	100

Table 5: DFA classification results for the texts of the newspaper "TO VIMA"

Predicted Group Membership (ELEFTHEROTYPIA)						
	Topic	**BUS**	**HUM**	**LEI**	**SOC**	***Total***
Original Count	**BUS**	26	3	0	0	29
	HUM	0	26	0	3	29
	LEI	0	1	27	0	28
	SOC	1	1	0	27	29
%	**BUS**	89.7	10.3	0	0	100
	HUM	0	89.7	0	10.3	100
	LEI	0	3.6	96.4	0	100
	SOC	3.4	3.4	0	93.1	100
Cross-validated Count	**BUS**	23	3	2	1	29
	HUM	0	24	0	5	29
	LEI	2	2	24	0	28
	SOC	3	9	0	17	29
%	**BUS**	79.3	10.3	6.9	3.4	100
	HUM	0	82.8	0	17.2	100
	LEI	7.1	7.1	85.7	0	100
	SOC	10.3	31	0	58.6	100

Table 6: DFA classification results for the texts of the newspaper "ELEFTHEROTYPIA"

The percentages of correct classification in the articles of each newspaper separately was:

a) "TO VIMA": 76.1% (57.5% correct classification in cross-validation).

b) "ELEFTHEROTYPIA": 92.2% (76.5% correct classification in cross-validation).

Examining separately the contribution of each variable group to the discrimination results we get the following correct classification ratios (Table 7).

	Whole Corpus		**TO VIMA**		**ELEFTHE-ROTYPIA**	
	%	*%CV*[4]	*%*	*%CV*	*%*	*%CV*
All Variables	**80.8**	**67.8**	**76.1**	**57.5**	**92.2**	**76.5**
Linguistic	60.2	52.8	66.4	52.7	67	51.9

[4] %CV denotes cross-validated results.

Word Domain	64.6	54.2	65.8	50.8	79.2	67.5
Text Domain	65	58.6	54.2	48.3	60.5	49.6

Table 7: Contrastive analysis of the classification results for each variable group when used separately

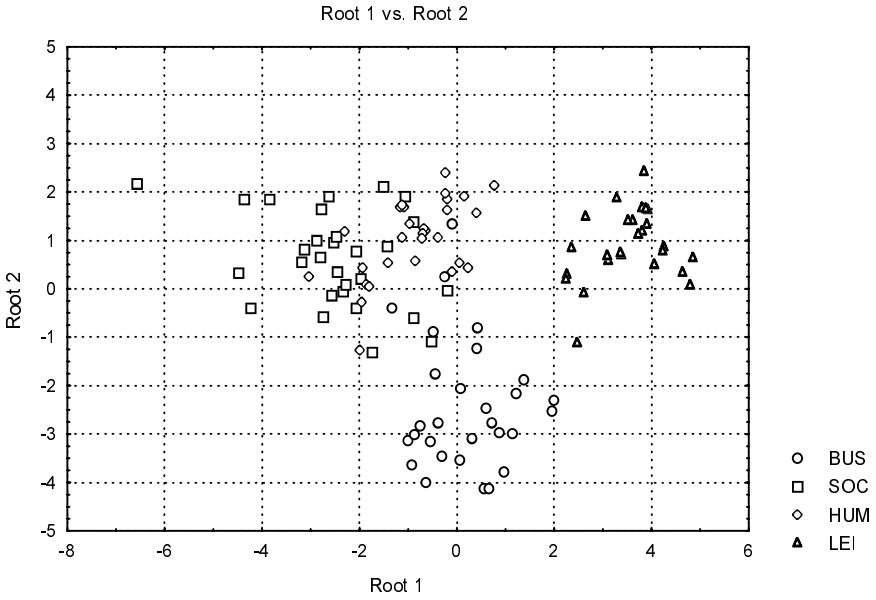

Figure 1: Text categorization scatterplot based on the first two discriminant functions extracted from the analysis.

The discriminatory power of each variable group differs in relation to the newspaper. Only the linguistic variable group performs equally in both newspapers. This fact reveals that K/D elements are productively utilized in stylistic decisions and they correlate with specific text type profiles.

6. Conclusion

The above results show that DFA performs fairly well in broad text type categorization. However, its performance is influenced from the corpus homogeneity. The category "SOC" presented the greater error rates in classification because most variables behaved similarly with the respective variables of the category "HUM". However, the categories "BUS" and "LEI" were classified with high precision and their clusters were kept distant from the other text types.

The significant lower classification results in the newspaper "TO VIMA" can partly be explained by the deviation from the assumption of homogeneity of variance that the majority of the variables exhibited in the specific subcorpus data. In order to investigate the equality of the variances among the two newspapers we used the Levene test. From the whole set of the predictor variables, 16 were violating homogeneity of variance in the data from the newspaper to "TO VIMA" and only 7 from the newspaper "ELEFTHEROTYPIA".

However, it is evident that text categorization in languages that exhibit inherent variation like MG should employ linguistic variables. The contribution of the linguistic variables in the classification was the most stable among the other variable groups and achieved nearly equal classification rates in both newspapers. The combination of variables from different linguistic levels proves to be a fruitful method[5] and constitutes a robust methodology for confronting text type classification.

Further research will be directed toward a more detailed group of linguistic variables and the application of non-linear classification algorithms.

7. Acknowledgements

The authors wish to thank the Greek newspapers "TO VIMA" and "ELEFTHEROTYPIA" for the provision of their articles to the ILSP corpus, a part of which was used in our study.

8. References

Alexiou, M., 1982. Diglossia in Greece. In W. Haas (ed.), *Standard Languages: spoken and written.* Manchester: Manchester University Press.

Andriomenos, G. 1999. Mia keimeniki analisi ipovlitikon anaforon tis astinomias: i anamiksi logion kai dimotikon stoixeion. [A textual analysis of police reports: the merging of K and D elements]. *Greek Linguistics '97, Proceedings of the 3rd international conference on the Greek language,* 675-683.

Baayen, R.H., 1994. Derivational productivity and text typology. *Journal of Quantitative Linguistics*, 1:16-34.

[5] For similar considerations regarding the usefulness of the combination of linguistic variables see Baayen (1994).

Biber, D., 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D., 1995. *Dimensions of register variation.* Cambridge: Cambridge University Press.

Carlgren, J., 1994. Recognizing text genres with simple metrics using discriminant analysis. *COLING94,* 1071-1074.

Ferguson, C.A., 1959. Diglossia. *Word*, 15:325-340.

Gavrilidou, M., P. Labropoulou and A. Christofidou, 1994. Sistima katigoriopiisis grapton keimenon. [Categorisation system of written texts]. *Studies in Greek linguistics 94*, 831-838.

Forsyth, R., 1995. *Stylistic structures: a computational approach to text classification.* Ph.D. thesis, University of Nottingham.

Forsyth, R., 1997. Evolutionary computing and text categorization. Presented in the Workshop of Computationally Intensive Methods in Quantitative Linguistics, Glasgow.

Hair, J., R. Anderson, R. Tatham and W. Black, 1995. *Multivariate data analysis.* New Jersey: Prentice Hall.

Hull, D., J. Pedersen and H. Schütze, 1996. Document routing as statistical classification. AAAI Spring Symposium on Machine Learning in Information Access, Stanford.

Huberty, C., J. Wisenbaker and J. Smith, 1987. Assesing predictive accuracy in discriminant analysis. *Multivariate Behavioural Research*, 22:307-329.

Karlgren, J., 1999. Stylistic experiments in information retrieval. In T. Strzalkowski (ed.), *Natural language information retrieval.* Dordrecht: Kluwer.

Mikk, J., 1995. Methods for determining optimal readability of texts. *Journal of Quantitative Linguistics*, 2:125-132.

Mikros, G. and G. Carayannis, forthcoming. Quantitative analysis of final –n rule usage in Modern Greek texts. *ILSP Working Papers.*

Mikros, G., 1999. *Koinonioglossologiki prosegisi fonologikon provlimaton tis Neas Ellinikis.* [Sociolinguistic approach of the phonological problems of Modern Greek]. Ph.D. thesis, Athens University.

Papatzikou - Cochran, E., 1997. An instance of triglossia ? Codeswitching as evidence for the present state of Greece's "language question". *International Journal of Sociology of Language,* 126:33-62.

Tambouratzis, G., S. Markantonatou, N. Xairetakis, G. Carayannis, 2000. Automatic style categorization of corpora in the Greek language. Paper presented in *LREC 2000*, Athens.

Tesitelova, M., 1992. *Quantitative linguistics.* Amsterdam : John Benjamins.

Uhlirova, L., 1995. On the generality of statistical laws and the individuality of texts. A case of syllables, word forms, their length and frequencies. *Journal of Quantitative Linguistics*, 2:238-247.

Wimmer, G., R. Köhler, R. Grotjahn and G. Altmann, 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1:98-106.

Yang, Y., 1997. An evaluation of statistical approaches to text categorization. School of Computer Science, Carnegie Mellon University, Technical Report No. CMU-CS-97-127.

Ziegler, A., 1998. Word length in Portuguese texts. *Journal of Quantitative Linguistics*, 5:115-120.