Discriminating the registers and styles in the Modern Greek language

George Tambouratzis*, Stella Markantonatou*, Nikolaos Hairetakis*, Marina Vassiliou*, Dimitrios Tambouratzis^, George Carayannis*

* Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
{giorg_t, marks, nhaire, mvas, gkara} @ilsp.gr

^ Agricultural University of Athens,
Iera Odos 75, 118 55 , Athens, Greece.
{dtamb@aua.gr}

Abstract

This article investigates (a) whether register discrimination can successfully exploit linguistic information reflecting the evolution of a language (such as the diglossia phenomenon of the Modern Greek language) and (b) what kind of linguistic information and which statistical techniques may be employed to distinguish among individual styles within one register. Using clustering techniques and features reflecting the diglossia phenomenon, we have successfully discriminated registers in Modern Greek. However, diglossia information has not been shown sufficient to distinguish among individual styles within one register. Instead, a large number of linguistic features need to be studied with methods such as discriminant analysis in order to obtain a high degree of discrimination accuracy.

1 Introduction

The identification of the language style characterising the constituent parts of a corpus is very important to several applications. For example, in information retrieval applications, where large corpora of texts need to be searched efficiently, it is useful to have information about the language style used in each text, to improve the accuracy of the search (Karlgren, 1999). In fact, the criteria regarding language style may differ for each search and therefore – due to the large number of texts – there is a requirement to perform style categorisation in an automated manner. Such systems normally use statistical methods to evaluate the properties of given texts. The complexity of the studied properties varies. Kilgarriff (1996) employs mainly the frequency-of-occurrence of words while Karlgren (1999) applies statistical methods primarily on structural and part-of-speech information.

Baayen et al. (1996), who study the topic of author identification, apply statistical measures and methods on syntactic rewrite rules resulting by processing a given set of texts. They report that the accuracy thus obtained is higher than when applying the same statistical measures to the original text. On the other hand, Biber (1995) uses Multidimensional Analysis coupled with a large number of linguistic features to distinguish among registers. The underlying idea is that, rather than being distinguished on the basis of a set of linguistic features, registers are distinguished on the basis of combinations of weighted linguistic features, the so-called "dimensions".

This article reports on the discrimination of texts in written Modern Greek. The ongoing research described here has followed two distinct directions. First, we have tried to distinguish among registers of written Modern Greek. In a second phase, our research has focused on distinguishing among individual styles within one register and, more specifically, among speakers of the Greek Parliament. To achieve that, structural, morphological and part-of-speech information is employed. Initially (in section 2) emphasis is placed on distinguishing among the different registers used. In section 3, the task of author identification is tested with selected statistical methods. In both sections, we describe the set of linguistic features measured, we argue for the statistical method employed and we comment on the results. Section 4 contains a description of future plans for extending this line of research while in section 5 the conclusions of this article are provided.

2 Distinguishing Registers

To distinguish among registers, we successfully exploited a particular feature of Modern Greek,

namely the contrast between Katharevousa and Demotiki. These are variations of Modern Greek which correspond (if only roughly) to formal and informal speaking. Katharevousa was the official language of the Greek State until 1979 when it was replaced by Demotiki. By that time, Demotiki was the established language of literature while, in times, it had been the language of elementary education. Compared to Demotiki, Katharevousa bears an important resemblance to Ancient Greek manifested explicitly on the morphological level and the use of the lexicon. At a second step, we dropped the Katharevousa-Demotiki approach and relied on part-of-speech information, which is often exploited in text categorisation experiments (for instance, see Biber et al. 1998). Again, we obtained satisfactory results.

2.1 Method of work

The variables used to distinguish among registers may be grouped into the following categories:

1. *Morphological variables*: These were verbal endings quantifying the contrast Katharevousa / Demotiki. Although the morphological differences between these two variations of Greek are not limited to the verb paradigm, we focused on the latter since it better highlights the contrast under consideration (Tambouratzis et al., 2000). A total of 230 verbal endings were selected, split into 145 Demotiki and 85 Katharevousa endings (see also the Appendix). These 230 frequencies-of-occurrence were grouped into 12 variables for use in the statistical analysis.
2. *Lexical variables*: Certain negation particles (*ουδείς, ουδέποτε, ουδαμού, άνευ*) clearly signify a preference for Katharevousa while others (*δίχως, μήτε, χωρίς*) are clear indicators of Demotiki. However, the most frequently used negation particles (*όχι, μην, δεν*) are not characteristic of either of the two variations.
3. *Structural macro-features*: average sentence length, number of commas, dashes and brackets (total of 4 variables).
4. After the completion of the experiments with variables of type 1-3 (Tambouratzis et al., 2000), Part-of-Speech (PoS) counts were introduced. The PoS categories were adjectives, adjunctions, adverbs, articles, conjunctions, nouns, pronouns, numerals, particles, verbs and a hold-all category (for non-classifiable entries), resulting in 11 variables expressed as percentages.

These variables are more similar to the characteristics used by Karlgren (1999), and differ considerably from those used by Kilgarriff (1996) and Baayen et al. (1996). For the metrics of the first and third categories, a custom-built program was used running under Linux. This program calculated all structural and morphological metrics for each text in a single pass and the results were processed with the help of a spreadsheet package. The metrics of the second category were calculated using a custom-built program in the C programming language. PoS counts were obtained using the ILSP tagger (Papageorgiou et al., 2000) coupled with a number of custom-built programs to determine the actual frequencies-of-occurrence from the tagged texts. Finally, the STATGRAPHICS package was used for the statistical analysis.

The dataset selected consisted of examples from three registers:

(i) fiction (364 Kwords - 24 texts),
(ii) texts of academic prose referring to historical issues, also referred to as the history register (361 Kwords – 32 texts) and
(iii) political speeches obtained from the proceedings of the Greek parliament sessions, also referred to as the parliament register (509 Kwords – 12 texts).

The texts of registers (I) and (II) were retrieved from the ILSP corpus (Gavrilidou et al., 1998), all of them dating from the period 1991-1999. The texts of register (III) were transcripts of the Greek Parliament sessions held during the first half of 1999.

This dataset was processed using both seeded and unseeded clustering techniques with between 3 and 6 clusters. The unseeded approach confirmed the existence of distinct natural classes, which correspond to the three registers. The seeded approach confirmed the ability to accurately separate these three registers and to cluster their elements together. Initially, a 'short' data vector containing only the 12 morphological variables quantifying the Demotiki/Katharevousa contrast was used (Tambouratzis et al. 2000), as well as a 16-element vector combining structural and morphological characteristics. The seeds for the Parliament and History registers were chosen randomly. The seeds for the Fiction register were chosen so that at least one of them would not be

an "outlier" of the Fiction register. Representative results are shown in Table 1 for the different vectors and numbers of clusters. In each case, the classification rate quoted corresponds to the number of text elements correctly classified (according to the register of the respective seed).

	12-elem.	16-elem.
6 clust.	95.6%	98.5%
4 clust.	97.1%	98.5%
3 clust.	95.6%	97.1%

Table 1 - Seeded clustering accuracy as a function of the cluster number and vector size.

The vector size was augmented with PoS information, resulting in a 27-element data vector. A new set of clustering experiments were performed using Ward's method with the squared Euclidean distance measure to cluster the data in an unseeded manner. Finally, a 15-element data vector was used with PoS and structural information but without any morphological information. The results obtained (Table 2) show that PoS information improves the clustering performance.

2.2 Comments on the Results

Our results strongly suggest that registers of written Modern Greek can be discriminated accurately on the basis of the contrast Katharevousa / Demotiki manifested with morphological variation. Languages with a different history may not be suited to such a categorisation method. This is evident in Biber's work (1995) for the English language, where a variety of grammatical and macro-structural linguistic features but no morphological variation features were employed. It seems then that corpora of languages which are characterised by the phenomenon of diglossia, may be successfully categorisable on the basis of morphological information (or other reflexes of diglossia). Such a discrimination method may give results as satisfactory as approaches which are closer to the Biber (1995) spirit and rely on PoS and structural measures (see Tables 1 and 2).

Tables 1 and 2 show that the accuracy of clustering reaches approximately 99% while the seeded clustering approach had a high degree of accuracy, reaching 100% when using 5 clusters. For the 27-element vector with both morphological and PoS information, perfect clustering has been achieved even with 4 clusters. On the other hand, a successful clustering (albeit with a lower level of accuracy) is achieved using only structural and PoS information.

It should be noted that the lexical variables used, that is the negation particles, did not contribute at all (Markantonatou et al., 2000). Furthermore, the system performed almost as well with and without macro-structure features, the difference in accuracy being less than 5%.

The parliament texts can be claimed to form a register whose patterns are closely positioned in the pattern space. Of the three registers, the literature one presented the highest degree of variance, with more than one sub-clusters existing as well as outlier elements. This may be explained by the fact that the parliament proceedings, contrary to literature, undergo intensive editing by a small group of specialised public servants.

3 Distinguishing Styles within One Register

In this section, we report on our efforts to distinguish among individual styles within one register. In particular, we intend to distinguish among speakers of the Parliament by studying the transcripts of the speeches of five parliament members over the period 1997-2000. Each of these speakers belongs to one of the five political parties that were represented in the Greek parliament over that period. Up to date, the experiments have been limited to the period 1999-2000.

3.1 Method of work

The number of variables (46 in total) calculated for each of the five speakers can be grouped as follows:

1. *Morphological variables* (20 variables):
   - Verbal endings expressing the Katharevousa / Demotiki contrast giving rise to 12 variables.

	12-elem.	16-elem.	27-elem.	15-elem
6 clust.	95.5%	100.0%	100.0%	100.0%
5 clust.	95.5%	100.0%	100.0%	89.6%
4 clust.	94.1%	98.5%	100.0%	83.4%
3 clust.	94.1%	98.5%	98.5%	83.4%

Table 2 - Unseeded clustering accuracy as a function of the cluster number and vector size used.

- the use of infixes (2 variables) in the past tense forms.
- the person and number of the verb form (6 variables).

The last two types of variable are expressed as percentages normalised over the number of verb forms.

2. *Lexical variables* (6 variables):
   - Negation particles (*όχι, δεν, μην*).
   - Negative words of Katharevousa (*ουδείς, άνευ*).
   - Other words which also express the contrast Katharevousa / Demotiki (the anaphoric pronouns '*οποίος*' (Kath) and '*που*' (Dem)), currently resulting in a single variable.
3. *Structural macro-features*: average sentence and word length, number of commas, question marks, dashes and brackets, resulting in a total of 6 variables.
4. *Structural micro-features* (other than lexical):
   - Part-of-Speech counts (10 variables).
   - Use of grammatical categories such as the genitive case with nouns and adjectives (2 variables).
5. The year when the speech was presented in the Parliament and the order of the speech in the daily schedule, that is whether it was the first speech of the speaker that day (hereafter denoted as "protoloyia") or the second, third etc. (resulting in a total of 2 variables).
6. The identity of the speaker, denoted as the speaker Signature (1 variable), which was used to determine the desired classification.

Similarly to the clustering experiments, a set of C programs was used to extract automatically the values of the aforementioned variables from the transcripts. Most of these programs rely on measuring the occurrence of di-grams, and more generally n-grams, for letters, words and tagsets, thus being straight-forward. In the case of speaker identification, Discriminant Analysis was used, as the clustering approach did not give very good results, indicating that the distinction among personal styles is weaker than that among registers. Even when only 2 speakers were used, the clusters formed involved patterns from both speaker classes.

We experimented with two corpora, Corpus I and Corpus II, as described in Table 3. Corpus II is a subset of Corpus I. Each of the speeches included in Corpus II was delivered as an opening speech ("protoloyia") at a parliament session when at least two of the studied speakers delivered speeches.

An important issue is whether the selected variables are strongly correlated. If indeed strong correlations do exist, these might be used to reduce the dimensionality of the pattern space. For the purposes of this analysis, the 46 independent variables were used (45 in the case of Corpus II where only "protoloyiai" exist, since then the order variable is constantly equal to 1). The number of correlations of all variable pairs exceeding given thresholds is depicted in Figure 1, for both Corpus I and Corpus II. According to this study, in Corpus II, the percentage of variable pairs with an absolute value of correlation exceeding 0.5 is approximately 3%, indicating a low correlation between the parameters. Additionally, out of 990 pairs of Corpus II, only a single one has a correlation exceeding 0.8. The correlations for the same parameter pairs over the two corpora are similar, though as a rule the correlation for Corpus I is less that that for Corpus II, reflecting the larger variability of texts in Corpus I. The correlation study indicated that most of the parameters are not strongly correlated. Thus, a factor analysis step is not necessary and the application of the discriminant analysis directly on the original variables is justified.

Initially, Corpus I (see Table 3) was processed. The 46 aforementioned variables were used to generate discriminant functions accurately recognising the identity of the speaker. To that end, three different approaches were used:

(i) the full model: all variables were used to determine the discriminant functions;

(ii) the forward model: starting from an empty model, variables were introduced in order to create a reduced model, with a small number of variables;
(iii) the backward model: starting from the full model, variables were eliminated to create a reduced model.

In the cases of the forward and backward models, the values of the F parameter to both enter and delete a variable were set to 4 while the maximum number of steps to generate the model was set to 50.

Year	1999-2000	
Speaker	Corpus I	Corpus II
A	92	30
B	45	24
C	33	21
D	21	16
E	150	36

Table 3 – Comparative composition of Corpus I and Corpus II.

The performance of this model is improved if:
1. the order in which each particular speech was delivered is taken into account: the subset of "protoloyiai" is well-defined and presents a low variance while the speeches of second or lower order have a higher variance.
2. the corpus comprises only sessions where more than one speaker has delivered speeches. Thus, the more balanced Corpus II (Table 3) presents an improved discrimination performance.

For these two corpora, the results of the discriminant analysis are shown in Table 4. The discrimination rate obtained with Corpus II is much higher than that for Corpus I. In addition, smaller models, with 8 variables, may be created that correctly classify at least 75% of Corpus II. An example of the factors generated and the manner in which they separate the pattern space is shown in the diagrams of Figure 2.

3.2 Comments on the Results

Though this research is continuing, certain facts can be reported with confidence.

Within the Greek Parliament Proceedings register, individual styles can not be classified on the basis of morphological features expressing the contrast Katharevousa/Demotiki. This may be explained by the fact that these texts undergo intensive editing towards a well-established sub-language. This editing homogenises the morphological profile of the texts but, of course, does not go as far as homogenising the lexical preferences of the various speakers. That is why, contrary to the register-clustering experiments, lexical variables expressing the particular contrast seem to play a role in discriminating between speakers and why the use of Katharevousa-oriented negative particles, which was not important in register discrimination, seems to be of some importance in style discrimination. The observation that negative words play a role in style identification is in agreement with the observations of Labbé (1983) on the French political speech.

Structural features have turned out to be important: the average word length, the use of punctuation and question marks and the use of certain parts-of-speech such as articles, conjunctions, adjunctions and - especially - verbs. Furthermore, the distribution of verbs into persons and numbers seems to be important, though the exact variables selected differ depending on the exact set of speeches used (these variables are of course complementary).

One of the most interesting findings of this research is that it is important whether the speaker delivers a "protoloyia" or not. "Protoloyiai" can be classified at a rate of 95% while mixed deliveries result in a lower rate, as low as 75%. This may be caused by two factors:
1. "Protoloyiai" represent longer stretches of text, which are more characteristic of a given speaker.
2. Speakers prepare meticulously for their "protoloyiai" while their other deliveries represent a more spontaneous type of speech, which tends to contain patterns shared by all the parliament members.

Finally, certain additional patterns are emerging for each of the speakers. Certain speakers (e.g. speaker A) are more consistently recognised than others (e.g. speaker B) while speaker B is similar to speaker C and speaker D is similar to speaker E. This indicates that additional variables may be required to improve the classification accuracy for all speakers.

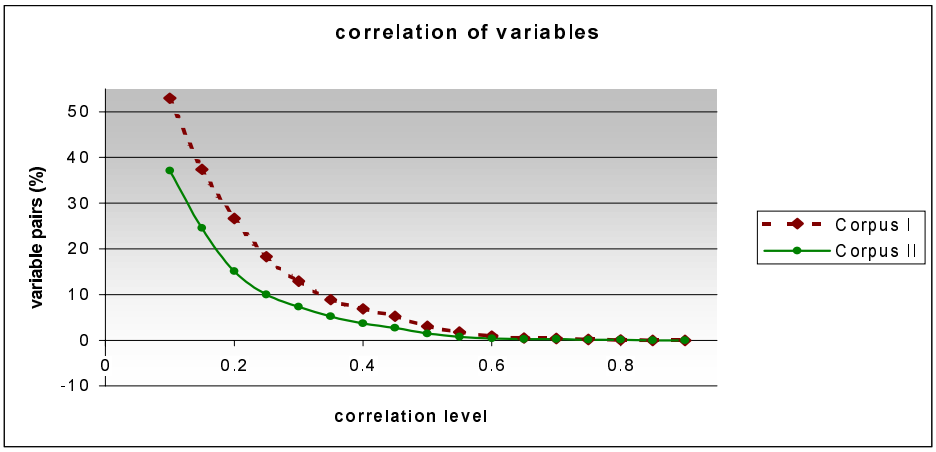

Figure 1 – Percentage of variable pairs exceeding a given level of absolute correlation.

Dataset	Model			observations
	full	forward	backward	
Corpus I	93.79 % *(46)*	75.37 % *(46)*	78.30 % *(46)*	341
Corpus II	97.64 % *(45)*	94.49 % *(13)*	92.91 % *(20)*	127
Corpus II (reduced model)	97.64 % *(45)*	87.40 % *(8)*	79.53 % *(8)*	127

Table 4 – Discrimination rate (the corresponding model size is shown in italics).

4 Future Plans

As a next step, frequency of use of certain lemmata shall be introduced since visual inspection indicates that they may provide good discriminatory features. We also plan to substitute average lengths (of both words and sentences) with the distribution of lengths. Furthermore, we intend to introduce certain structural measurements such as repetition of structures, chains of nominals and the occurrence of negation within NP phrasal constituents. Another possible extension involves the inclusion of the speech topic. As certain speakers' characteristics seem to change through time, we plan to process the entire corpus of speeches for the target period 1997-2000. Finally, an important issue is the comparison of the results obtained in our experiments to these generated by alternative techniques proposed by other researchers. This will allow the deduction of more accurate conclusions regarding the strengths and the weaknesses of the research strategies.

5 Conclusions

In this article, ongoing research on register and individual style categorisation of written Modern Greek has been reported. A system has been proposed for the automatic register categorisation of corpora in Modern Greek exploiting the highly inflectional nature of the language. The results have been obtained with a relatively constrained set of registers; however their recognition accuracy is remarkably high, exceeding 98% with an unseeded clustering approach using between 3 and 6 clusters.

On the front of individual style categorisation, a discrimination rate of over 80% was achieved for five speakers within the Greek Parliament register. Morphological variables were shown to be of less importance to this task, while lexical and structural variables seemed to take over. We are planning to introduce several new lexical and structural variables in order to achieve better discrimination rates and to determine discriminating features of the different styles.

Acknowledgements

The authors wish to thank the Dept. of Language Technology Applications and specifically Dr. Harris Papageorgiou and Mr. Prokopis Prokopidis in obtaining the lemmatised versions of the parliament transcripts. Additionally, the authors wish to acknowledge the assistance of the Secretariat of the Hellenic Parliament in obtaining the session transcripts.

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*, Vol. 11, No. 3, pp. 121-131.

Biber, D. (1995) *Dimensions of Register Variation: A cross-linguistic comparison.* Cambridge University Press.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press.

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

Gavrilidou, M., Labropoulou P., Papakostopoulou N., Spiliotopoulou S., Nassos N. (1998) Greek Corpus Documentation, Parole LE2-4017/10369, WP2.9-WP-ATH-1.

Holton D., Mackridge, P. & Philippaki-Warburton, I. (1997) *Greek: A Comprehensive Grammar of the Modern Language.* Routledge, London and New York.

Karlgren, J., (1999) Stylistic Experiments in Information Retrieval. In T. Strzalkowski (ed.), *Natural Language Information Retrieval,* pp. 147-166. Dordrecht: Kluwer.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition,* Sussex University, April , pp. 33-40.

Labbé, D. (1983). *François Mitterrand. Essai sur le discours*. La Pensée Sauvage, Grenoble.

Markantonatou, S. & Tambouratzis, G. (2000) Some quantitative observations regarding the use of grammatical negation in Modern Greek. *Proceedings of the 21st Annual Meeting of the Department of Linguistics*, Faculty of Philosophy of the Aristotelian University of Thessaloniki, May 2000 (in print/in Greek).

Papageorgiou, H., Prokopidis, P., Giouli, V. & Piperidis, S. (2000) A Unified PoS Tagging Architecture and its application to Greek. *Proceedings of the 2nd International Conference on Language Resources and Evaluations*, Athens, Greece, 31 May - 2 June, Vol. 3, pp. 1455-1462.

Tambouratzis, G., Markantonatou, S., Hairetakis, N. & Carayannis, G. (2000) Automatic Style Categorisation of Corpora in the Greek Language. *Proceedings of the 2nd International Conference on Language Resources and Evaluations*, Athens, Greece, 31 May - 2 June, Vol. 1, pp. 135-140.

APPENDIX
Characteristics of Katharevousa and Demotiki

Diglossia in Modern Greek is due to the contrast between *Katharevousa* and *Demotiki* and is well-manifested on the morphological level. Here we concentrate on verb morphology.

Demotiki tends to have words ending with an 'open' syllable. So, *3rd Plural* verbal endings in *–n* (1) are augmented to *–ne* (2).

(1) *έλεγαν* [e'leγan] *(Kath)* (=they said)

(2) *λέγανε* [le'γane] *(Dem)* (=they said)

In *Demotiki*, *Katharevousa's* consonant clusters of two fricatives or two plosives are converted into clusters of one fricative and one plosive (3) – (4) (Holton et al., 1997, pp. 14).

(3) *πεισθώ* [pisθo'] */πειστώ* [pisto'] (=to be convinced)

(4) *καλυφθώ* [kalifθo'] */καλυφτώ* [kalifto'] (=to be covered)

Certain verb classes exhibit thematic vowel alternations either following the inflectional paradigm of Ancient Greek or Demotiki (5) (Clairis and Babiniotis, 1999).

(5) *εξαρτάται* [eksarta'te] *(Kath) / εξαρτιέται* [eksartiete] *(Dem)* (=depends)

Sometimes *Demotiki* uses a verbal root, which is similar though not identical to the *Katharevousa* one (6).

(6) *λύω* [li'o] *(Kath) / λύνω* [li'no] *(Dem)* (=to solve)

Finally, many verbs inherited from *Katharevousa* survive in *Demotiki*, either having an equivalent – mainly colloquial- (7) or not (8) (Clairis and Babiniotis, 1999).

(7) *προτίθεμαι* [proti'θeme] *(Kath) / σκοπεύω* [skope'vo] *(Dem)* (=I intend to)

(8) *προΐσταμαι* [proi'stame] (=supervise)

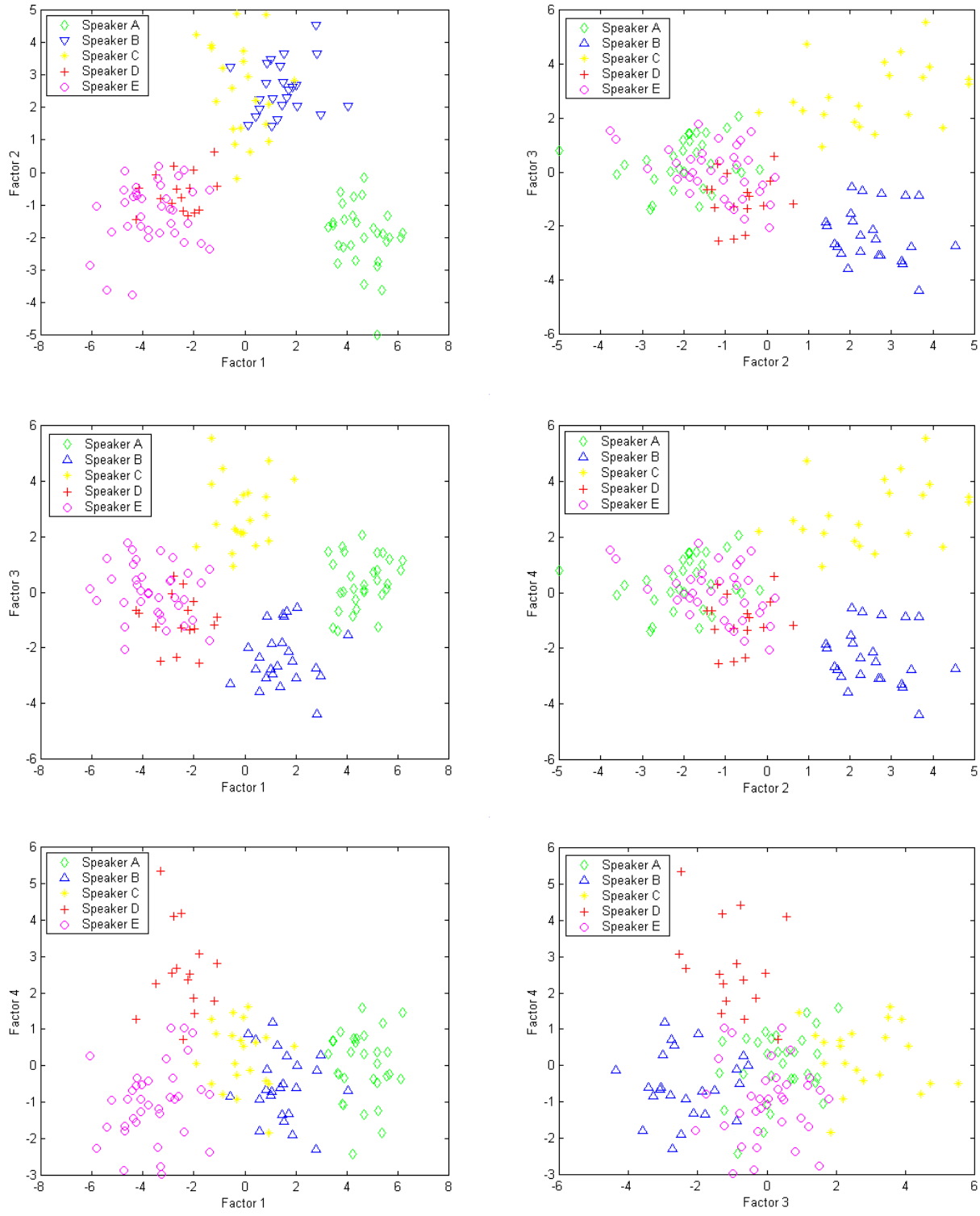

Figure 2 – Discriminant factors plotted against the patterns for corpus II.