Automatic Style Categorisation of Corpora in the Greek Language.

George Tambouratzis, Stella Markantonatou, Nikolaos Hairetakis, George Carayannis

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

Abstract

In this article, a system is proposed for the automatic style categorisation of text corpora in the Greek language. This categorisation is based to a large extent on the type of language used in the text, for example whether the language used is representative of formal Greek or not. To arrive to this categorisation, the highly inflectional nature of the Greek language is exploited. For each text, a vector of both structural and morphological characteristics is assembled. Categorisation is achieved by comparing this vector to given archetypes using a statistical-based method. Experimental results reported in this article indicate an accuracy exceeding 98% in the categorisation of a corpus of texts spanning different registers.

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, in order to improve the accuracy of the search (Karlgren, 1999). Language style can thus serve as a factor increasing the accuracy of the search itself. 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.

Modern Greek presents itself as an interesting case study in style issues. As a whole, the Greek language has evolved continuously during the past 3000 years. One of the results of this evolution is that the correspondence between a given set of grammatical features, say, *Verb+Past+Imperfect+3rd+Plural* and the surface morphological markers is one-to-many, e.g. for the set of grammatical features given above and the verb root: *ντυν-* [din-] (to dress someone) the following forms are in widespread use (in the following phonetic transcriptions, a stress following a vowel identifies the intonation point in the word.:

(1) *έντυναν* [e'dinan], *ντύναν* [di'nan], *ντύνανε* [di'nane] (=dressed).

Politipia is the Greek term reserved for this phenomenon. In addition, it is often the case that more than one words are available for the same sense (e.g. *ύδωρ* [i'*δ*or] / *νερό* [nero'] (=water)). Syntax also presents cases of politipia.

Politipia in Modern Greek was endorsed by the so-called *Katharevousa* which was used as the official language of the Greek state (and the secondary and tertiary eduction) until 1979 – when it was replaced by the so-called *Demotiki*. *Katharevousa* was a somehow haphazard mix of colloquial and ancient Greek (Holton, Mackridge & Philippaki-Warburton, 1997). Modern Greek, while being closer to *Demotiki*, contains a considerable number of features which are easily identified as belonging to *Katharevousa*. Such features occur in varying proportions in the different registers (Biber, 1993; Biber, Conrad & Reppen, 1998). Typically, registers employing a more "formal" linguistic style, such as academic prose and the language of the law, contain more *Katharevousa* features than, say, fiction and spoken language.

It is important to stress here that several of the grammatically equivalent word forms may be encountered relatively frequently in Modern Greek. However, as some of them normally characterise formal versus informal speech, it is the case that the choice of a particular word form signifies a particular linguistic style.

In this paper, a system is proposed that allows for the automatic categorisation of the texts contained in a corpus on the basis of the style employed by the respective author. The proposed system draws on morphological and structural features of the texts. In the case of Modern Greek, the system takes advantage of the phenomenon of politipia and draws heavily on the inflectional nature of this language. Information regarding the type of endings used is collected using automated tools. The system also employs information regarding structural features such as the size and structure of words and sentences. The morphological and structural features are assembled to form a data vector representing that text. This vector is then compared with the use of statistical methods to archetype vectors of the different styles taught to the system and is classified as belonging to the class with the highest likelihood.

In the next section, the problem of text-style categorisation is presented in more detail and related work is reviewed. In section 3, the morphological characteristics this study has relied on are presented. In section 4, the structural characteristics that were searched for are detailed. In section 5, the methods used to extract the structural and morphological information are presented. The corpora used in our experiments and the experimental results are described in section 6. Finally, section 7 concludes this paper with a review and discussion of the system structure and performance.

2. Categorisation of text style.

The recent dramatic developments in information technology assist in the integrated storage and processing of linguistic resources. For example, textual data is generated in an electronic format and therefore can be readily transferred and replicated by electronic means. Additionally, as a result of the improvement in the processing power and storage capacity of computer systems, the storage of very large corpora of texts has become possible. By virtue of the increased quality and capabilities of telecommunications it is possible to search through a number of distributed databases in order to retrieve the required information. As the cost of such computer systems falls, their use becomes more widespread, giving rise to the need for advanced information retrieval systems.

It has been argued (Karlgren, 1999) that in information retrieval applications, the language style used by the authors of the different texts is important to the accuracy of the search, this being determined objectively as the number of desired texts over the number of actually retrieved texts. Evidently, in such systems, this accuracy depends strongly on the selection of the appropriate feature set. This feature set, in turn, is dependent on the language used in the texts, though the system principles should hold over different languages, by selecting the appropriate characteristics. In the following two sections, the characteristics selected for the style classification of texts in the Greek language shall be presented.

3. Morphological Characteristics.

As noted earlier, politipia is a prevalent phenomenon in Modern Greek. Politipia is the phenomenon whereby a set of grammatical features is mapped onto more than one surface forms. Politipia is observed mainly in the verb system of Modern Greek as noted earlier in example (1). To a lesser degree, politipia can be observed in the noun and adverb systems of Modern Greek (see also Mikros & Carayannis, 2000). Adverbs, however, are not as frequent as verbs. In addition, in the case of adverbs, morphological issues interfere with semantic issues rendering automated processing intractable. On the other hand, politipia in the noun system is restricted to a very small number of forms. In the light of these facts, the research described in this article is confined to the study of verbal politipia.

Verbal politipia in Greek is mainly due to suffixes and, to a lesser degree, to prefixes (1). That is, verbal politipia in Greek is mainly due to the multitude of endings available for the same cluster of grammatical features. Some of the grammatically equivalent endings do not reflect a difference between formal (*Katharevousa*) and informal (*Demotiki*) speech but others do. This work focuses on the distinction between formal and informal speech.

Alternative endings are due to active evolutional tendencies of the language. Such is the tendency of *Demotiki* to have words ending with an ‘open’ syllable. So, *3rd Plural* endings in *–n* are augmented to *–ne* (2).

(2) *έλεγαν* [e’leγan] *(Kath)* / *λέγανε* [le’γane] *(Dem)* (=they said)

A second active tendency of *Demotiki* is to convert *Katharevousa’s* consonant clusters consisting of two fricatives or two plosives into clusters consisting of one fricative and one plosive. In clusters containing an /s/, the non-strident partner converts to a plosive. In clusters with voiceless fricatives not containing an /s/, the first of the consonants is a fricative and the second one is a plosive (Holton, Mackridge & Philippaki-Warburton, 1997, pp. 14) as shown in examples (3)-(5).

(3) *πεισθώ* [pis*θ*o’] /*πειστώ* [pisto’] (=to be convinced)

(4) *καλυφθώ* [kalif*θ*o’] /*καλυφτώ* [kalifto’] (=to be covered / protected)

(5) *απαλλάχθηκα* [apala’x*θ*ika] / *απαλλάχτηκα* [apala’xtika] (=got rid of)

Third, in Modern Greek there exist classes of verbs which may either follow the inflectional paradigm of ancient verbs having an /a/ (6) or an /o/ (7) or an /e/ (8) as a thematic vowel or choose the set of endings offered by *Demotiki* (Clairis and Babiniotis, 1999).

(6) *εξαρτάται* [eksarta’te] *(Kath)* / *εξαρτιέται* [eksartiete] *(Dem)* (=depends)

(7) *αξιούμε* [aksiu’me] *(Kath)* / *αξιώνουμε* [aksio’nume] *(Dem)* (=demand)

(8) *θεωρείτο* [*θ*eori’to] *(Kath)* / *θεωρούνταν* [*θ*eoru’dan] *(Dem)* (= was considered)

It is sometimes the case that *Demotiki* uses a verbal root which is similar but not identical to the *Katharevousa* one (9).

(9) *δεικνύω* [*δ*ikni’o] *(Kath)* / *δείχνω* [*δ*i’xno] *(Dem)* (=to show, to point)

Finally, there is a bulk of verbs (and their compounds) which are clearly inherited from *Katharevousa* but either they have not been replaced by some word of *Demotiki* (10) or, the existing *Demotiki* equivalent, is marked as colloquial (11) (Clairis and Babiniotis, 1999).

(10) *προίσταμαι* [proi’stame] (=supervise)

(11) *διάκειμαι* [δia’kime] *(Kath)* / *νιώθω* [nio’*θ*o] *(Dem)* (=I am disposed)

These morphological contrasts run through the inflectional paradigms of Greek verbs and affect hundreds of verbal wordforms. By studying the quantitative distribution of these forms, it shall be shown that morphological features play an important role in the identification of linguistic style (Biber, Konrad & Reppen, 1998). A total of approximately 230 endings which are characteristic of *Katharevousa* or *Demotiki* have been selected to determine the morphological information within a given text.

4. Structural Characteristics.

In the experiments, several different structural characteristics were measured in each text contained in the corpus:

a) Word count and sentence count.
b) Count of commas and other punctuation signs (parentheses, dashes).

c) Average word and sentence length.

d) Frequency of words with length *x* [where *x* varies from 1 to 30 letters].

e) Frequency of sentences composed by *x* words [where *x* varies from 1 to 150 words].

Apart from the obvious need to measure the characteristics mentioned in (a,c,d,e), characteristics (b) were also introduced since punctuation signs such as brackets, dashes and commas frequently indicate the existence of sub-sentences inside a sentence. Thus, the frequency-of-occurrence of these punctuation marks provides an indication of the stylistic variation in a text. It should be noted that in our measurements digits were counted as words.

Broadly similar factors were used in the experiments of Karlgren (1999), who counted characteristics (a, c and e) with the addition of long words and digits. Long words have not been used in this piece of research as they are not expected to be indicative of a given style in the Greek language, in contrast to other languages.

5. Extraction of Characteristics.

As noted earlier, the classification of text styles relies on two main categories of characteristics, morphological and structural. To automate the extraction of these characteristics, specialised programs are employed.

5.1. Extraction of Structural Characteristics

The structural characteristics mentioned above were measured via a custom-built program running under Linux. This program calculated all structural metrics for each text in a single pass and the results were processed with the help of a spreadsheet software package.

5.2. Extraction of the Morphological Characteristics

The extraction of morphological characteristics is performed via the Automated Morphological Processor (AMP). The AMP has been designed to take advantage of the highly inflectional nature of the Greek language and allow the generation of morphological lexica in an automated manner (Tambouratzis & Carayannis, 1999). The generation of morphological lexica is a labour-intensive task, which requires several man-years to be completed. To encompass the language evolution throughout the ages, one would need to generate manually a large number of different morphological lexica, each one covering a specific **Language Evolution Sample** (hereafter denoted as LES) corresponding to a particular time-period in the history of the Greek language and/or a geographical area. The use of the AMP allows the automated generation of morphological lexica with a large degree of accuracy. This is of particular importance in text retrieval and information extraction purposes from texts in the Greek language, where in order to represent fully the inflectional evolution of the language through time it is necessary to use two or three different morphological dictionaries.

The AMP is based on a rule-based iterative masking-and-matching principle, which relies on matching parts of different patterns while ignoring the remainders of these patterns. For example, if two patterns (here words) x_1x_2 and y_1y_2 are to be compared, the technique might focus on the possible similarity between x_1 and y_1 (attempting to match these parts) while ignoring x_2 and y_2 (temporarily masking off these parts).

This principle may be augmented by a modest amount of information regarding the specific LES being used, in order to optimise the accuracy of the morphological processing. Indeed, different approaches (involving various levels of a priori knowledge) have been followed to investigate the system behaviour (Tambouratzis & Carayannis, 1999). These indicate that the inclusion of a modest amount of LES-specific information (for example, a handful of a priori endings) can improve substantially the segmentation accuracy. The optimal morphological segmentation accuracy achieved by the AMP is equal to 96%, measured against the contents of the morphological lexicon, which has been constructed manually at the ILSP.

As noted, the AMP is able to operate on texts belonging to several different LESs. This is ensured by providing the AMP with a number of interchangeable modules, which may be selected according to the current LES requirements, while the core of the system remains the same. The interchangeable modules are of a declarative nature and may incorporate a priori knowledge.

For the purposes of text-style characterisation, it is necessary to calculate the frequencies-of-occurrence of specific endings, as detailed in section 3. Therefore the AMP is constrained to use the selected endings which are representative of style as a priori knowledge. For each of these endings, the frequency-of-occurrence as well as the corresponding root and its frequency-of-occurrence is recorded by the AMP. The results obtained for the given corpus are detailed in the following section.

6. Experiments.

6.1. Corpus Composition and Pre-processing

To study the performance of the proposed style-characterising system, a corpus has been created. This consists of three different types of text:

(i) excerpts from novels, which represent the Fiction register. These novels have been composed during the period 1988-1997 and are contained in the ILSP corpus of texts (Gavrilidou et al., 1998).

(ii) a collection of essays concerning the history of the Greek province of Macedonia, which represent the academic register. This collection of essays was created in 1992 and forms part of the ILSP corpus of texts. This collection is hereafter denoted as 'History'.

(iii) official transcripts from randomly-chosen sessions of the Greek Parliament, held during the period January 1999 - May 1999. This set of transcripts is hereafter denoted collectively as ‘Parliament’.

To perform the experiments detailed in this article, all texts were spell-checked. A more extensive pre-processing was performed in the case of the Parliament register. The aim of this piece of research has been to process relatively large portions of text which provide a sufficiently clear view of the style of speech used at the Greek Parliament. Therefore, the texts contained in the parliament sub-corpus were processed and very short pieces of dialogue (which typically were related to the Parliament regulations) were removed from the transcripts. Such pre-processing need not be performed for the other two registers. The size and characteristics of the different registers of texts in the corpus are shown in table 1.

register	sample texts	size (in words)
Fiction	24	363,783
History	32	360,993
Parliament	12	509,917
Total	**68**	**1,234,693**

Table 1: Corpus composition in terms of register.

The tools detailed in section 5 were used to process each text contained in the corpus, generating a vector of characteristics. Since the texts are of varying sizes, it was decided to normalise them in order to remove possible sources of variability prior to performing the statistical analysis. Thus, all characteristics were normalised so that the values reported corresponded to occurrences per 100,000 words of text. The values of these vectors are summarised in tables 2, 3 and 4, by providing for each register a set of average values and of the corresponding standard deviations for all the characteristics studied. In table 2, the frequencies of characteristic *Katharevousa* verb endings are detailed, while table 3 contains a similar table for *Demotiki* verb endings. It should be noted that these frequencies do not represent all verb endings but are restricted to the 230 endings that are grammatically representative of *Katharevousa* and *Demotiki*, respectively, as determined in section 4. Furthermore, as the use of a vector with 230 characteristics would be computationally intractable, in both tables 2 and 3, the endings are displayed collectively for each combination of the grammatical categories ‘person’ and ‘number’. Tables 2, 3 and 4 provide a macroscopic view of the vectors of characteristics used to describe each of the texts contained in the corpus.

According to these collective results for the three registers (while being restricted to the set of 230 style-characteristic endings), certain patterns seem to emerge:

- ❖ In the Parliament register (tables 2 and 3), the verb endings corresponding to *Katharevousa* are as a whole more frequent than these of *Demotiki*, in comparison to the other registers.
- ❖ Interestingly, though, in the case of the 2nd person (both plural and singular) in the Parliament register the frequency of *Demotiki* endings is comparatively higher than that of *Katharevousa.* This is more marked in the 2nd person singular, where the frequency-of-occurrence of *Demotiki*-type endings is 50 times higher than that of the *Katharevousa*-type endings. On the contrary, the situation is reversed for the 1st and 3rd persons (both singular and plural), where *Katharevousa* endings are as a rule more frequent.
- ❖ In History, the majority of verb endings correspond to the third person (singular or plural), indicating a predominantly narrative style.
- ❖ The third person is also most frequently used in the Fiction register, though to a lesser extent than in the History register.
- ❖ Texts from the Fiction register have a considerably larger number of sentences and sub-sentences (if commas, brackets and dashes are collectively examined) as compared to the other two registers, which have a relatively high degree of similarity (see table 4).
- ❖ Collectively, the Fiction register shows a higher degree of variance over all characteristics than the other registers. The History register has a considerably lower variance of its characteristics while the Parliament register has the lowest variance, indicating possibly more tightly clustered representative vectors for its members. This, in conjunction with the previous observations, indicates the possible existence of a sub-language specific to politicians.

In the following paragraph, a more formal analysis of the data shall be provided using statistical methods. It should be noted that a selected subset of characteristics were used in this analysis, the aim of the vector being to study whether the three registers may be separated from each other with a sufficient degree of confidence. This is the subject of the following section.

	1st singular	2nd singular	3rd singular	1st plural	2nd plural	3rd plural
Fiction	157.2 *(108.5)*	33.5 *(41.3)*	55.7 *(56.6)*	103.8 *(182.0)*	2.0 *(9.1)*	123.3 *(67.3)*
History	0.7 *(2.9)*	0.2 *(1.1)*	208.3 *(129.8)*	43.8 *(51.0)*	0.4 *(1.9)*	343.9 *(140.2)*
Parliament	166.0 *(51.9)*	1.0 *(1.6)*	247.3 *(62.1)*	308.0 *(38.0)*	51.8 *(20.6)*	472.7 *(100.8)*

Table 2: Average Frequency (in normal typescript) and standard deviation (in italics) of *Katharevousa* verb endings over 100,000 words, for each of the three text registers.

	1st singular	2nd singular	3rd singular	1st plural	2nd plural	3rd plural
Fiction	69.0 *(57.2)*	85.9 *(65.8)*	330.8 *(154.2)*	49.3 *(74.3)*	37.3 *(64.2)*	67.2 *(45.6)*
History	0.0 *(0.0)*	37.7 *(53.3)*	228.3 *(102.0)*	8.9 *(16.1)*	0.0 *(0.0)*	153.0 *(72.6)*
Parliament	50.4 *(32.9)*	67.0 *(16.0)*	180.2 *(32.4)*	82.2 *(28.9)*	48.6 *(26.9)*	77.9 *(16.6)*

Table 3: Average Frequency (in normal typescript) and standard deviation (in italics) of *Demotiki* verb endings over 100,000 words, for each of the three text registers.

	sentences	commas	brackets	dashes	Katharevousa verb endings	Demotiki verb endings
Fiction	8649.0 *(2975.0)*	7728.7 *(1435.8)*	40.7 *(95.4)*	1793.8 *(1386.2)*	475.5 *(294.43)*	639.5 *(223.5)*
History	4678.2 *(1164.1)*	6834.9 *(1422.4)*	849.1 *(463.1)*	495.2 *(254.5)*	597.37 *(221.5)*	427.93 *(161.1)*
Parliament	5722.4 *(463.5)*	6726.0 *(442.0)*	56.2 *(49.5)*	570.1 *(69.8)*	1252.5 *(137.7)*	506.3 *(49.5)*

Table 4: Average frequency (in normal typescript) and standard deviation (in italics) of macroscopic structural and morphological characteristics over 100,000 words, for each of the three text registers. The final two columns represent cumulative results for the measurements summarised in tables 2 and 3 respectively and are included here for reasons of clarity, though they do not represent independent measurements and thus are not used as vector elements.

6.2. Experimental Results.

To compare and contrast the three registers, a cluster analysis approach was selected. During the experiments, normalisation was performed over all vector parameters prior to the clustering operation, so that they occupied the same ranges of values. Additionally, different criteria (nearest, furthest, median, centroid and average) were studied when using non-seeded clustering experiments. It was found that all these clustering methods generated similar results and thus they shall not be distinguished henceforth, the results reported being those obtained by the majority of criteria.

Initially a non-seeded clustering approach was studied, to indicate the natural clusters existing in the corpus. When 6 clusters were used, five out of these each contained one or at most two texts from the Fiction register, while the sixth cluster contained the remaining texts from the Fiction register together with all texts from the History and Parliament registers.

Following that, 10 clusters were used to cluster the text vectors. This experiment showed that the register with the highest variability was the Fiction register, whose members occupied 9 out of the 10 available clusters. Out of the three registers, only the Parliament register was fully separated from the other registers, its members occupying a single cluster, which contained no texts from the other registers. Finally, all members of the History register occupied a single cluster, though this also contained texts belonging to the Fiction register.

These preliminary results indicated that the Parliament register was the one whose members were most closely spaced on in the pattern space defined by the characteristics vector, confirming the conclusions of sub-section 6.1.. The History register was also relatively tightly-spaced, while the Fiction register contained several "outlier" members which were at a large distance from the majority of Fiction members. To evaluate these observations, a cluster analysis of the elements of each register was performed. This confirmed that the Parliament and History registers were the most tightly coupled. On the contrary, the Fiction register contained 5 elements, which were significantly different to the others.

Following this analysis, a seeded clustering approach was chosen. Initially, a cluster number of 6 was used, the clusters being equally distributed between the 3 registers, resulting in two clusters per register. Following that, the number of clusters was reduced to 4 (and then 3), in both cases one cluster being devoted to Parliament and History and two (and then one) being devoted to the Fiction register. These experiments were carried out using vectors consisting of:

❖ the endings frequencies for all persons in *Demotiki* and *Katharevousa* as well as the number of sentences, commas, parentheses and dashes (giving a 16-element vector);

❖ the endings frequencies for all persons in *Demotiki* and *Katharevousa* as well as the number of sentences and commas (giving a 14-element vector);

❖ the endings frequencies for all persons in *Demotiki* and *Katharevousa* (giving a 12-element vector consisting solely of morphological characteristics).

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 5 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).

	16-elem.	14-elem.	12-elem.
6 clust.	98.5%	92.7%	95.6%
4 clust.	98.5%	94.1%	97.1%
3 clust.	97.1%	92.7%	95.6%

Table 5: Clustering accuracy as a function of the number of clusters used and the size of the characteristic vector used.

This table indicates that the optimal clustering performance is obtained when the number of characteristics in the vector is equal to 16, that is when both structural and morphological information is considered. On the contrary, if part of the morphological information is suppressed (using a 14-element vector), the recognition rate is reduced (though it still remains around 93-94%). Indeed, the experimental results indicate that it is then probably beneficial to remove all structural information, recognition rate being higher in the case of 12-element vectors as opposed to 14-element vectors.

According to table 5, even when using 3 clusters the clustering result accurately represents the 3 registers, with very few misclassifications being performed. Indeed, the optimal results for each vector configuration are obtained using a 4 cluster classifier. This indicates that removing redundant clusters from the well-defined classes (that is, the Parliament and History registers) increases the recognition rate.

7. Discussion and Conclusions.

In this article, a system has been proposed for the automatic style categorisation of corpora in the Greek language. This categorisation is based on the type of language used in the text. To arrive to this categorisation, the highly inflectional nature of the Greek language is used. Characteristics reflecting this inflectional nature are combined with characteristics based on the structure of sentences to provide automatic style categorisation of texts.

The study of the corpus indicates that the three registers used here have different characteristics. The texts belonging to the Parliament register form a well-defined class in the pattern space. Similarly, historical texts also form a well-defined class, though to a slightly smaller extent than the Parliament register. On the contrary, texts belonging to the Fiction register form a less tight class. Thus, to successfully recognise this register, a comparatively large number of seeds is required to sufficiently cover the register space. However, as a whole, the seeded clustering approach using both 4 and 6 clusters achieves a recognition accuracy exceeding 98%. Thus, the presented method can be used to accurately define the register of a given text. Of course, these results have been obtained with a relatively constrained set of registers, but the results are of such a high accuracy as to support the conclusion that this line of work may lead to an accurate style-characterising system.

Future work on this area is planned to focus on confirming these results with larger-scale experiments. Additionally, there exist a number of parameters that may also be introduced in an effort to more accurately define the register of texts (such as negation parameters measured in (Markantonatou & Tambouratzis, 2000)). In this respect, the results obtained for the Fiction register are indicative of the existing scope for improvement. Indeed, techniques ranging from alternative statistical techniques to other pattern recognition methods (such as neural networks) are being considered for the recogniser part of this system.

8. Acknowledgements

The authors wish to acknowledge the assistance of the Secretariat of the Hellenic Parliament in obtaining the session transcripts studied in this piece of research.

9. References

Biber, D., 1993. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, Vol. 19, No. 2, pp. 219-241.

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

Clairis, C. & G. Babiniotis. 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., P. Mackridge and I. Philippaki-Warburton. 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.

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

Mikros, G. & Carayannis, G. (2000) Modern Greek Corpus Taxonomy. *Proceedings of the 2nd Conference on Language Resources and Evaluation*, Athens, Greece, 31 May-2 June (in print).

Tambouratzis, G. & Carayannis, G., 1999. Automated Construction of Morphological Lexica Possessing Terminology Wealth on the Basis of Term-Intensive Documents. *Proceedings of the Second Conference of Hellenic Language and Terminology*, Athens, 21-23 October 1999, pp. 149-155. ISBN 960-86069-1-8 (in Greek).