Greek in the age of corpora: Challenges and solutions

Dionysis Goutsos
Department of Linguistics, University of Athens, Greece
dgoutsos@phil.uoa.gr

Abstract

The paper offers a state-of-the-art description of corpus research on Greek, focusing on developments in corpus linguistics rather than computational linguistics. It refers to the specific characteristics of Greek that have had an effect on corpus research and outlines the main phases of development for Modern Greek corpora. It also presents the most important findings on the description of the Greek language deriving from corpora, with specific examples and references, and discusses the perspectives of corpus-related work on Greek.

1 Introduction

The hosting of the 10th NooJ conference in Greece offers an excellent opportunity to take stock of the main developments in previous and current corpus research on the Greek language. My perspective is that of the linguist who uses corpora and is interested in what corpora can reveal about language. In this view, I am following Hardie's (2009) distinction between computational linguistics, including language engineering and natural language processing, and corpus linguistics as two distinct fields that may overlap, but do not coincide.

Thus, this paper gives an overview of corpus linguistic work on Greek in order to complement the computational linguistic emphasis of this conference. I will first present the particular features of Greek that have influenced corpus development and analysis and will then refer to the development of Greek corpora for linguistic research. Next, I discuss the ways in which corpus analysis changes our view of the language on the basis of findings from several studies of Greek. Finally, the new challenges of Greek corpus linguistics are outlined with a view to suggesting further developments in the field.

2 Greek: Some peculiarities

A number of idiosyncratic features of Greek have been responsible for the slow rate of development of corpus linguistic research. First of all, Greek is a language with an especially long and complicated history. As Browning puts it, "since [the Homeric poems] Greek has enjoyed a continuous tradition down to the present day. Change there has certainly been. But there has been no break like that between Latin and the Romance languages. Ancient Greek is not a foreign language to the Greek of today as Anglo-Saxon is to the modern Englishman" (1983: vii). As a result, there have been multiple continuities and discontinuities in the history of Greek, leaving their traces in the language's structure and vocabulary. In addition, a complex sociolinguistic situation has emerged, that can broadly be characterized as Greek diglossia, which the language and its study have managed to overcome only in the late 1970s.

Without doubt, this is one of the reasons why Greek linguistics has shown an aversion to empiricism and only limited use of data in linguistic analysis. For instance, there is a gap of some 50 years between the 1940s, when the first fully-fledged descriptions of Modern Greek appeared (Triandafyllidis, 1941; Tzartzanos, 1941-63), and the 1990s, when comparable, modern scientific descriptions were published (Holton et al., 1997; Clairis & Babiniotis, 1998-2004). The same is true for Modern Greek dictionaries, which only made an appearance in the late 1990s (Babiniotis, 1998; Idryma Manoli Triandafyllidi, 1998). Not surprisingly, these reference works were not based on corpus data.

Finally, the peculiarity of the Greek writing system, which differs from the standard Western alphabet, for which most computer applications were initially designed, created obstacles for the computational treatment of the language. Thus, the biggest part of the 1980s and 1990s was taken with the effort of the linguistic community to subvert the rigid ASCII code in order to accommodate the needs of the Greek user. The introduction of Unicode put an end to these problems and made further research redundant, but, meanwhile, a lot of time and resources was wasted on technicalities rather than the analysis of the language.[1]

3 The development of Greek corpora

Renouf (2007) distinguishes five stages in English language corpus evolution:

a) the 1960s-1970s, dominated by the one-million word Small Corpus (e.g. *LOB, Brown corpus*),

b) the 1980s, with the multi-million word Large Corpus or super-corpus (e.g. *Bank of English, BNC*),

c) the 1990s, with the 'Modern Diachronic' Corpus (e.g. *FLOB, Frown*)

d) 1998 onwards, during which the Web as corpus or cyber-corpus is introduced, and

e) 2005 onwards, expected to develop the Grid, i.e. a pathway to distributed corpora.

Because of the peculiarities mentioned in the previous section, among other reasons, Greek has been missing several of these stages. In particular, the first Greek corpora make an appearance in the late 1980s and early 1990s, when literary works are stored and analyzed by computational means (Philippides, 1981; 1986; 1988; Kyriazidis and Kazazis, 1992; Kyriazidis et al., 1992). In 1994 a survey finds that there are 15 small projects of collecting Greek data, but concludes that "corpora, if they are used in linguistic research at all, are not fully exploited" (Goutsos et al., 1994a: 215).

It is the 1990s and, especially, the 2000s which see the development of the two large corpora mainly used in Greek linguistics, the Hellenic National Corpus (HNC) and the Corpus of Greek Texts (CGT). HNC is a development of the Institute for Language and Speech Processing, currently including about 40 million words of mainly journalistic texts (Hatzigeorgiu et al., 2000).[2] CGT is a development of the Universities of Cyprus and Athens, including 30 million words from a wide range of spoken and written texts (Goutsos, 2003).[3] Goutsos (2010) compares the two corpora and argues that CGT fills the need for a representative and authoritative corpus of Greek, since HNC still includes a narrow range of text types, does not contain spoken data, has inadequate classification of texts and offers restricted availability.

[1] For a useful overview of problems and solutions with regard to corpora in the Greek alphabet, see King (1997).

[2] Available at: http://hnc.ilsp.gr/subcorpus.asp#

[3] Available at: http://www.sek.edu.gr

At the same time, a number of specialized corpora have started to make a late public appearance. These include the newspapers and school books corpora of the Greek Language Portal,[4] a biomedical corpus (Pantazara et al., 2007), as well as a Greek learner corpus and a thematic corpus, designed for learners at the University of Athens.[5]

It is clear that Greek is lagging behind other languages, in terms of both the size and the variety of corpora available for the description of the language. In Renouf's (2007) terms, it still lacks full 'super-corpora' and the dynamic, open-ended diachronic corpora available for English. However, research that has been based on the existing corpora has already borne fruit, as will be shown in the following section.

4 Corpus findings on Greek

There are several areas of Greek linguistics in which corpus-related research has produced a number of useful findings. These include the description of grammatical categories, phraseology, language variation, teaching applications, as well as the emergence of language norms and language change. The following presentation reviews the most important work in these areas, always from a corpus linguistic perspective.

4.1 Grammatical categories

Sinclair (1991) has pointed out that data-driven research, by avoiding predetermined linguistic categories, can identify facts about the grammar of a language which had previously been ignored. Thus, the study of corpora has pointed out the occurrence of the so-called shell nouns, general nouns that are used with several textual functions, including the encapsulation and labelling of a stretch of discourse. Koutsoulelou and Mikros (2004-2005) have studied the word *γεγονός* ('fact') in its use as a shell noun in the academic, journalistic and spoken sub-corpora of the CGT and found out its preference for the written mode, its collocations and phraseology, as well as its multiple functions and sub-functions. An extended study of all Greek shell nouns is still necessary in order to uncover similar patterns that will allow us to talk about a new sub-category of nouns in Greek.

Fragaki (2010a; 2010b) is a thorough investigation of Greek adjectives in an opinion articles sub-corpus of the CGT. Although the identification of adjectives follows pre-existing criteria, their classification is extensively corpus-driven, since it starts from evidence in the corpus. Thus, ten adjective categories are identified: classifying, descriptive, evaluative, deictic, relational, specializing, indefinite, colour, verbal and quantitative adjectives. Apart from important quantitative data, concerning e.g. the frequency of adjective categories and the relation between categories and their characteristics, the study also explores the evaluative and ideological role of adjectives in Greek discourse, pointing out that it is only certain adjective sub-categories that can take up these roles.

In all, corpus research has refined our knowledge of two basic grammatical categories of Greek; obviously, much more work is required before we have a full view of Greek grammar through corpora.

[4] Available at: http://www.greek-language.gr/greekLang/modern_greek/index.html

[5] Both available at: http://greekcorpora.isll.uoa.gr/gr/Default.aspx

4.2 Phraseology

Although the study of lexical collocations and phraseology seems to be ideally suited for corpus linguistic research, there have only been sporadic studies of Greek vocabulary (e.g. Goutsos et al., 1994b; Goutsos, 2009a).

An exception to this is the extended study of 3 to 5 word clusters (also known as lexical bundles or n-grams in the bibliography) in four Greek text types, spoken and academic texts, newspapers and fiction (Ferlas, 2011). Four different types of clusters are identified (basic, extended, variant and unique clusters), while the different functions they perform in discourse permits their categorization into the categories of stance, referential, text organizing, title, personal, grammatical and thematic clusters. What especially comes out in this research is the fact that Greek extensively draws on a number of word clusters such as *δεν μπορεί να* ('it cannot'), *δεν πρέπει να* ('it must not'), *θα μπορούσε να* ('it could'), *θα πρέπει να* ('it should') in order to indicate modality in discourse. In addition, the cross-linguistic comparison with English is made possible, pointing to similarities and differences between the two languages.

Again, more research from a corpus linguistic perspective is necessary in this area, in order to complement existing computational studies (e.g. Fragos et al., 2004).

4.3 Language variation

In a series of articles Mikros (1997; 2003; et al. 1996; et al. 2003; 2005, among else) systematically uses corpora in order to identify the parameters of phonological and morphological variation in Greek. This line of research has unearthed a host of interesting material on language variation and thus made possible an objective analysis of phenomena such as word couplets, which are due to the long history of diglossia (see section 2, above). Corpus linguistic methods are here combined with statistical and computational methodology in order to define basic characteristics of Greek texts. The findings of this research can thus be applied to such areas as the automatic identification of authorship, stylistic analysis etc.

Frantzi (2005) also uses statistical techniques in order to identify style features of political discourse. This is another area which is particularly interesting to explore, since it brings together the analysis of stylistic and ideological parameters of language variation in Greek.

Finally, the linguistic construction of gender identity has been studied in a couple of articles (Fragaki and Goutsos, 2005; Goutsos and Fragaki, 2009), which explore the meanings and collocations of gender-related nouns and adjectives in Greek (e.g. *άνδρας* 'man' vs. *γυναίκα* 'woman', *ανδρικός* 'male' vs. *γυναικείος* 'female'). This research has identified the ways in which gender asymmetry prevails in specific text types through patterns of nominal and adjectival use and their ideological implications. It is interesting to note that there have been only a few similar studies on other languages –mainly English– (see Goutsos and Fragaki, 2009: 319) and thus the area is offered for contrastive analysis.

4.4 Teaching applications

The main attempts to apply the findings of corpus linguistics in the teaching of Greek relate to the development of a specialized corpus for teaching Greek as a foreign language and a learner corpus, tagged for errors, both mentioned in section 3 above (see Iakovou et al., 2003). A similar project, aiming at the creation of a Corpus of Academic Greek Texts, is currently being developed at the Aristotle University of Thessaloniki, whereas a PhD dissertation on the basic academic vocabulary of Greek is currently written (Katsalirou, in prep.).

In addition, a first attempt at defining a basic vocabulary for Greek can be found in Goutsos (2006), which presents a number of basic nouns and verbs in Greek for both the CGT as a whole and sub-corpora of different text types, including academic texts, newspaper reports and opinion articles, legal-administrative and spoken texts.

4.5 Language norms and language change

One of the most important contributions of the corpus linguistic approach concerns the identification of language norms that cannot be reached at on the basis of intuition alone.

A case in point concerns the placement of connectives in Greek, which has been extensively studied in Goutsos (2009b). The category of connectives includes particles, discourse markers, sentence adverbials and other elements that are usually placed in the periphery of the clause and can have a crucial role in linking discourse rather than sentence parts. The area is notoriously difficult to divide into neat categories and, as a result, terms, both in Greek and other languages, proliferate, sometimes referring to the same phenomena. Corpus data can be invaluable in identifying frequent patterns and reaching generalizations about the linguistic behaviour of these elements.

In particular, the study of 1 million words of Greek from four sub-corpora of the CGT (academic texts, opinion articles, parliament speeches and TV and radio interviews) suggests that connectives show specific preferences for placement in particular clause positions. Table 1 below presents the figures in percentages for the occurrence of specific connectives at the beginning of the clause in the four sub-corpora.

	Academic	*Opinion articles*	*Parliament speeches*	*Interviews*
αντίθετα	65	▨	56	▨
άρα	60	56	▨	▨
επομένως	45	52	▨	▨
ευτυχώς	66	57	50	58
συνεπώς	55	▨	▨	▨
εντούτοις	▨	-	-	▨
παρ' όλα αυτά	▨	-	-	▨
πρώτα-πρώτα	-	-	▨	▨
συμπερασματικά	▨	▨	-	-

Table 1. Connectives with preferred 1" clause position

As can be seen in Table 1, there are overwhelming tendencies for certain connectives such as adverbials of contrast (*αντίθετα, εντούτοις, παρ' όλα αυτά*) or conclusion (*άρα, επομένως, συνεπώς, συμπερασματικά*) and adverbials of stance (*ευτυχώς*) to occur in first clause position with frequencies that exceed half or even two thirds of their occurrences.

The importance of first clause position for connective elements has been stressed in the literature and has also been observed in several other languages (see Goutsos 2009b, for bibliography). Therefore, it is not surprising as such and can be accounted for on the basis of functional principles. What is more surprising is the tendency of several other Greek connectives to occur in second clause position, i.e. following the first clause constituent, as can be seen from the percentages of occurrence in Table 2.

	Academic	*Opinion articles*	*Parliament speeches*	*Interviews*
ακριβώς	50	45	47	48
άραγε	44			
λοιπόν				65
όμως				60
πράγματι	40	38	52	45

Table 2. Connectives with preferred 2nd clause position

Corpus data suggest that this preference for second position is not accidental, since it concerns extremely frequent connective elements such as *λοιπόν* and *όμως* and holds for two thirds of their occurrences and across spoken and written text types, as can be seen in Table 2. In other words, it seems that second clause position has been conventionalized in Greek as the place for elements that indicate overall connectivity.

Again, several functional principles can be invoked to account for this (e.g. marking thematic position, rhythmic signalling etc.). However, what is most important is to find that connective elements in Greek have specific preferences for placement in the clause and that second clause position is reserved for some of these elements with surprising regularity across genres. These findings suggest that new norms have developed in Greek, about which little can be known without recourse to corpus data.

It is clear that the development of language norms is a predominantly diachronic phenomenon, which cannot be adequately studied in the absence of a diachronic corpus. Indirect evidence for language change can, however, be adduced, among others, through the study of new vocabulary that is introduced in Greek.

As is the case with other languages, computer terminology has been imported into Greek, mainly from English. There are three options for introducing new vocabulary in Greek, at least as far as the written mode is concerned: a) to use the foreign loan wholesale, i.e. in Latin characters (e.g. *computer, internet*), b) to transliterate the foreign loanword in Greek characters (e.g. *κομπιούτερ, ίντερνετ* or *ιντερνέτ*[6]) and c) to use a pre-existing Greek work (e.g. for computer *υπολογιστής* = calculator) or create a neologism by using pre-existing Greek morphemes (e.g. for internet *διαδίκτυο* from*δια*= inter and *δίκτυο*= network).

[6] The difference in transliteration concerns the placement of the accent according to the English or the French preference, respectively.

Although we cannot trace the history of the use of terms without a diachronic corpus (cf. Gorjanc 2006), a large corpus of Greek can offer evidence for the synchronic state of alternative uses. Figure 1 below presents the frequency of the terms used for *computer* in Greek, as found in the CGT. (A fourth option of the abbreviation *H/Y*, that is *ηλεκτρονικόςυπολογιστής* = electronic calculator, which is the full Greek equivalent for computer, has also been included).

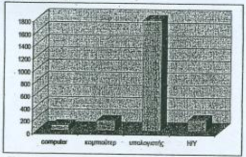

Figure 1. Frequency of alternative terms for 'computer' in CGT

As can be seen in Figure 1, the option that is overwhelmingly preferred in Greek is that of the Greek word, rather than the foreign term, either in Latin characters or transliterated. It is interesting to compare this data to figures from the Web; a Google search (January 2011) shows that the Greek word *υπολογιστής* is more than four times as frequent as the transliterated option κομπιούτερ (959.000 vs. 225.000 pages, respectively).

The numbers for the alternative terms for *internet* are shown in Figure 2.

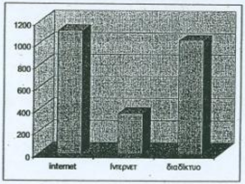

Figure 2. Frequency of alternative terms for 'internet' in CGT

Figure 2 suggests that the non-transliterated option is slightly more frequent than the Greek neologism and both are much more frequent than the transliterated alternative. The respective figures from a Google search favour the neologism *διαδίκτυο*, which occurs almost three times as much as the transliterated option *ίντερνετ* (6.060.000 vs. 2.710.000 pages). This would confirm the trend found in the CGT in favour of the Greek neologism.

Although data from the Web offer updated evidence for current language use, conventional corpora like the CGT are invaluable in studying parameters that cannot be explored in the data offered by the internet. Thus, CGT can be used to study the frequency of the lemmas associated with the two options, as seen in Figure 3.

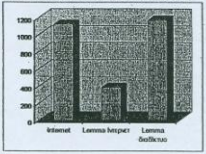

Figure 3. Frequency of alternative terms for the lemma 'internet' in CGT

Figure 3 compares the frequency of the 'non-Greek' *internet* with the frequency of the transliterated *ίντερνετ*, along with its derived nouns and adjectives (e.g. *ιντερνετικός*, *ιντερνετάκι*, even the plural nounι*ντερνέτια* etc) and that of the neologism διαδίκτυο, along with its derived nouns and adjectives (e.g.*διαδικτυακός*, *διαδικτυωμένος*, *διαδικτύωση*, *διαδικτυάκι* etc). It seems then that the lemma of the neologism is slightly higher in frequency than that of the transliterated option. This would suggest that the ease with which derived words can be formed in Greek affects the frequency and adoption of terms: the neologism thus offers an advantage over the other two options in being much easier to form derived words with.

In addition, a reference corpus like CGT is useful in comparing text types and thus identifying possible fields in which neologisms are to be found. In the case of the terms for *internet* in Greek, the frequencies presented above are split in Figure 4 according to text types.

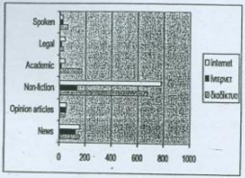

Figure 4. Distribution of frequencies of alternative terms for 'internet' in CGT text types

As shown in Figure 4, the non-transliterated option is more frequent in popularized, non-fiction texts, which is the text type in which all terms are much more frequently used. It also competes with the neologism in news, while the latter is much more frequent in academic texts, as well as spoken texts, although all terms are much less used in this text type.

In other words, non-fiction is the privileged area in which language change of this sort is expected to happen. In addition, while it is expected that academic texts would opt for the adapted Greek term, it is also interesting that spoken data confirm the preference for the neologism. This type of synchronic evidence can be crucial in determining the type and direction of potential language change.

5 Perspectives

The above discussion has made it clear that there are several areas in which corpus linguistic research on Greek is expected to develop in the future. First of all, there is an urgent need for compiling many more and more varied corpora with an emphasis on the diachronic study of Greek. This must include both the longer diachrony and recent language change. With respect to the former, a remaining challenge is to link Modern Greek corpora with corpora or databases for earlier phases of Greek such as ancient or Medieval Greek.[7] With respect to the latter, the challenge is to develop new, dynamic corpora, aimed at covering the decades of the 20th century before the 1990s and expand to the 21st century.

Secondly, the linguistic resources available on the Web can also be fruitfully explored to a larger extent than before, either through existing software such as Sketch Engine or WebCorp or through new methods of compiling corpora from the Web. This may include the compilation of comparable or parallel corpora that can be used in the analysis of Greek in contrast with other languages.

[7] For instance, see the projects available at: http://www.tlg.uci.edu and http://www.mml.cam.ac.uk/greek/grammarofmedievalgreek, respectively.

Thirdly, there is a need for increased interaction between corpus linguistic and computational linguistic methods. To this effect, the availability and standardization of NLP applications such as taggers, parsers etc. have to be improved.

Finally, there is much scope for the improvement of Greek language description with the use of corpus linguistic methods, including the investigation of grammatical categories and specialized phraseology, as well as through the study of new text types and genres of Greek. The final aim would be to produce new grammars and dictionaries[8] that would be based on less intuitive and more accurate empirical data.

Acknowledgments

I would like to thank the organizers of NooJ 2010, and especially Zoe Gavriilidou, for their kind invitation to give this plenary lecture, for their hospitality and for giving me the opportunity to meet the NooJ family. Many thanks to Philip King for his always valuable comments.

References

Babiniotis, G. 1998. *Λεξικό της Νέας Ελληνικής Γλώσσας*. [Dictionary of Modern Greek]. Kentro Lexicologias, Athens.

Browning, R. 1983. *Medieval and Modern Greek*. Cambridge University Press, Cambridge.

Clairis, C. and Babiniotis, G. 1998-2004. *Γραμματική της Νέας Ελληνικής. Δομολειτουργική-Επικοινωνιακή*.[Modern Greek Grammar. Structural-Functional-Communicative]. Ellinika Grammata, Athens.

Ferlas, H. 2011. *Ο προκατασκευασμένος λόγος στα Ελληνικά και Αγγλικά. Μια μελέτη βασισμένη σε σώματα κειμένων με προεκτάσεις στη διδασκαλία της γλώσσας*.[Prefabricated discourse in Greek and English. A corpus-based study with implications for language teaching]. PhD dissertation, University of Athens.

Fragaki, G. 2010a. *Ο αξιολογικός ρόλος του επιθέτου και η χρήση του ως δείκτη ιδεολογίας: Μελέτη βασισμένη σε σώματα κειμένων δημοσιογραφικού λόγου* [The evaluative role of the adjective and its use as a marker of ideology: A study based on journalistic corpora]. PhD dissertation, University of Athens.

Fragaki, G. 2010b. A corpus-based categorization of Greek adjectives. *Proceedings of the 5th Corpus Linguistics Conference. 21-23 July 2009. University of Liverpool.* Available at: http://ucrel.lancs.ac.uk/publications/CL2009/.

Fragaki, G. and Goutsos, D. 2005. Gender adjectives and identity construction in Greek corpora. *Proceedings of the 7th International Conference on Greek Linguistics, University of York, 8-10 September 2005*. Available at: http://83.212.19.218/icgl7/Fragaki-et-al.pdf.

Fragos, K., Maistros, I. and Skourlas, C. 2004. Extracting collocations in Modern Greek language. *Proceedings of the 1^{st} International Conference on Natural Language Understanding and Cognitive Science, Porto, Portugal, 13-14 April 2004.* Available at: http:// glotta.ntua.gr/nlp_lab/Fraggos/files/DiCofinal.pdf

[8] For the potential contribution of corpora to Greek lexicography, see Goutsos & Fragaki (forthcoming).

Frantzi, K. 2005. Γλωσσικά και μη χαρακτηριστικά των προκηρύξεων της 17N. [Linguistic and non-linguistic features of terrorist manifestoes]. *Studies in Greek Linguistics* 25: 639–650.

Gorjanc, V. 2006. Tracking lexical changes in the reference corpus of Slovene texts. In Andrew Wilson, Dawn Archer and Paul Rayson (eds) *Corpus Linguistics Around the World*. Rodopi, Amsterdam, 91-100.

Goutsos, D. 2003. Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση. [Corpus of Greek Texts: Design and implementation]. *Proceedings of the 6^{th} International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Available at the webpage: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm.

Goutsos, D. 2006. Ανάπτυξη λεξιλογίου. Από το βασικό στο προχωρημένο επίπεδο. [Vocabularydevelopmentfromthebasictotheadvancelevel]. *Η ελληνική ως ξένη γλώσσα: Από τις λέξεις στα κείμενα*. Patakis, Athens, 13-92.

Goutsos, D. 2009a. «Λόγος να γίνεται»: Μεταγλωσσική φρασεολογία στο λόγο των πολιτικών του κοινοβουλίου. [Metalinguistic phraseology in Parliament discourse]. In Eleni Karamalengou and Eugenia Makrygianni (eds). *Αντιφίλησις. Studies on Classical, Byzantine and Modern Greek Literature and Culture. In Honour of John-Theophanes A. Papademetriou*. Franz Steiner, Stuttgart, 638-647.

Goutsos, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 5 (1): 29-44.

Goutsos, D. and Fragaki, G. 2009. Lexical choices of gender identity in Greek genres: The view from corpora. *Pragmatics* 19 (3): 317-340.

Goutsos, D. and Fragaki, G. Forthcoming. Λεξικά και σώματα κειμένων. [Dictionaries and corpora]. In Γιώργος Ξυδόπουλος, Άγις Οικονομίδης and Γιώργος Τράπαλης (eds). *Εισαγωγή στη λεξικογραφία*. [Introduction to Lexicography]. Patakis, Athens.

Goutsos, D., King, P. and Hatzidaki, O. 1994b. A corpus-based approach to Modern Greek language research and teaching. In Irene Philippaki-Warburton, Katerina Nicolaidis and Sifianou Maria (eds) *Themes in Greek Linguistics: Papers from the First International Conference on Greek Linguistics. Reading, September 1993.* John Benjamins, Amsterdam/Philadelphia, 507-513.

Goutsos, D., King, P. and Hatzidaki, R. 1994a. Towards a Corpus of Spoken Modern Greek. *Literary and Linguistic Computing* 9 (3): 215-223.

Goutsos, G. 2009b. Μόρια, δείκτες λόγου και κειμενικά επιρρήματα: Η οριοθέτηση των γλωσσικών κατηγοριών με τη χρήση ΗΣΚ. [Particles, discourse markers and text adverbs: The definition of linguistic categories through the use of corpora] *Proceedings of the 8^{th} International Conference on Greek Linguistics, University of Ioannina, 30 August-2 September 2009*, 754-768. Available at: http://www.linguist-uoi.gr/cd_web/ case2.html.

Hardie, A. 2009. Corpus linguistics and the languages of South Asia: Some current research directions. In Paul Baker (ed.) *Contemporary Corpus Linguistics*. Continuum, London, 262-288.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H. and Demiros, I. 2000. Design and implementation of the online ISLP corpus. *Proceedings of the LREC 2000 Conference*, Athens, 1737-1742.

Holton, D., Mackridge, P and Philippaki-Warburton, I.. 1997. *Greek. A Comprehensive Grammar of the Modern Language*. Routledge, London.

Iakovou, M., Markopoulos, M and Mikros, G. 2003. Θεματοποιημένο Βασικό Λεξιλόγιο μέσω ΗΣΚ: Πρακτική εφαρμογή στη διδασκαλία της Νέας Ελληνικής ως ξένης γλώσσας. [Thematic basic vocabulary through corpora. A practical application to the teaching of Modern Greek as a foreign language]. *Proceedings of the 6th International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Availableatthe webpage: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm.

Idryma Manoli Triandafyllidi. 1998. *Λεξικό της κοινής νεοελληνικής*. [Dictionary of Standard Modern Greek]. Institute of Modern Greek Studies, Thessaloniki.

Katsalirou, A. Inpreparation. *Το λεξιλόγιο για γενικούς ακαδημαϊκούς σκοπούς στη διδακτική της νέας ελληνικής ως ξένης γλώσσας*. [Vocabulary for general academic purposes in the teaching of Modern Greek as a foreign language]. PhD dissertation, Aristotle University of Thessaloniki.

King, P. 1997. Creating and processing corpora in Greek and Cyrillic alphabets on the personal computer. In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds). *Teaching and Language Corpora*. Longman, London, 277-291.

Koutsoulelou, S, and Mikros, G. 2004-2005. Το «γεγονός» ως ουσιαστικό κέλυφος. Χρήση και λειτουργία σε ηλεκτρονικά σώματα κειμένων της Ελληνικής. [Γεγονός as a shell noun. Use and function in Greek electronic corpora]. *Glossologia* 16: 65-95.

Kyriazidis, N. and Kazazis, I., K. 1992. *Τα ελληνικά του Μακρυγιάννη με τον υπολογιστή*. [Makriyannis' Greek in the computer]. Papazisis, Athens.

Kyriazidis, N., Kazazis, I., N. and Brehier, J. 1992. *Το λεξιλόγιο του Μακρυγιάννη*. [Makriyannis' vocabulary]. Athens.

Mikros, G. 1997. Radio news and phonetic variation in Modern Greek. *Greek Linguistics '95. Proceedings of the 2nd International Conference on Greek Linguistics 1995*, I, 35-44.

Mikros, G. 2003. Στατιστικές προσεγγίσεις στην αυτόματη κατηγοριοποίηση κειμένων της Νέας Ελληνικής: Μια πιλοτική αξιολόγηση υφομετρικών δεικτών και στατιστικών μεθόδων. [Statistical approaches to the automatic classification of Modern Greek texts. A pilot evalnation of stylometric indexes and statistical methods]. *Proceedings of the 6th International Conference on Greek Linguistics, University of Crete, 18-21 September 2003.* Availableatthe webpage: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm.

Mikros, G., Gavriilidou, M., Lambropoulou, P. and Doukas, D. 1996. Χθες ή χτες; Μια ποσοτική μελέτη φωνητικών και μορφολογικών στοιχείων σε κείμενα της Νέας Ελληνικής. [A quantitative study of phonetic and morphological features in Modern Greek texts]. *Studies in Greek Linguistics* 16: 645-656.

Mikros, G., Hatzigeorgiu, N. and Carayannis, G. 2003. Βασικά ποσοτικά μεγέθη στην γραπτή Νέα Ελληνική γλώσσα: η αξιοποίηση του ΕΘΕΓ στην ελληνική ποσοτική γλωσσολογία. [Basic quantitative measures in written Modern Greek. The exploitation of HNC in Greek quantitative linguistics. *Proceedings of Workshop on Text processing for Modern Greek: From Symbolic to Statistical Approaches", Rethymno, 20 September 2003*, 23-37.

Mikros, G., Hatzigeorgiu, N. and Carayannis, G. 2005. Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics* 12 (2-3): 167-184.

Pantazara, M., Mantzari, E., Vagelatos, A., Kalamara, C. and Iordanidou, A. 2007. Development of a Greek biomedical corpus. Paper given at the 11th Panhellenic Conference on Informatics (PCI 2007), Patras, Greece. Available at: http://www.iatrolexi.gr/vagelat/Iatrolexi-corpus.pdf.

Philippides, D. 1981. Computers and Modern Greek. *Mantatoforos* 17: 5-13.

Philippides, D. 1986. *The Sacrifice of Abraham on the Computer*. Hermes Press, Athens.

Philippides, D. 1988. Literary detection in the *Erotokritos* and *The Sacrifice of Abraham*. *Literary and Linguistic Computing* 3: 1-11.

Renouf, A. 2007. Corpus development 25 years on: from super-corpus to cyber-corpus. In Roberta Facchinetti (ed.) *Corpus Linguistics 25 Years on.* Rodopi, Amsterdam 27-49.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Triandafyllidis, M. 1941. *Νεοελληνική γραμματική (της δημοτικής)* [Modern Greek Grammar (of dimotiki)]. Organismos Ekdoseos Sxolikon Vivlion, Athens.

Tzartzanos, A. 1946-63. *Νεοελληνική σύνταξις (της κοινής δημοτικής)* [Modern Greek Syntax (of dimotiki)]. Organismos Ekdoseos Didaktikon Vivlion, Athens.

Proceedings of the
Nooj 2010
International Conference and Workshop

May 27, 28, 29 2010 Komotini Greece
Democritus University of Thrace

Edited by:

Zoe Gavriilidou
Elina Chadjipapa
Lena Papadopoulou
Max Silberztein