The general outcome of the investigation was presented at the Modern Greek Seminar, University of Birmingham, U.K. on Dec. 12, 1992 (Eklund 1992). The overall distribution of sporting terms in this material is:

(16)

	%	#
1. Mixed origin	28	44
2. English	44	71
3. French	16	26
4. Italian	7	10
5. Other languages	5	8
6. Unknown origin	[< 1]	1
	100	160

During the last year of the project, 1993-94, we intend to conduct comprehensive studies in vocabulary, morphology and syntax, based on the entire corpus. Towards the end of the academic year the corpus should also be accessible online via the Faculty Gopher.

REFERENCES

Cederholm, Yvonne & Marleen van Stam Rydehell. 1989. "Lexikaliska Databaser. Logiska Modeller och Användarmodeller". Examensarbete på Datalingvistlinjen. Göteborgs Universitet. (Lexical Databases. Logical Models and User Models). BA thesis in Computational Linguistics, University of Göteborg.

Eklund, Bo-Lennart. 1992. "A Selection of Sport Terms of Foreign Origin in the *Eleftherotipia*". Göteborg: University of Göteborg.

Fabricius, Cajus & Daniel Ridings. 1989. *A Concordance to Gregory of Nyssa.* Göteborg: Studia Graeca et Latina Gothoburgensia L.

Kyriazidis, N. I., I. N. Kazazis & J. Bréhier. 1983. *To Lexilogio tou Makriyanni i Pos Milousan i Ellines Proton Viasti i Glossa mas apo tin Katharevousa* (Makriyannis' Vocabulary or How the Greeks Spoke Before Our Language Was Violated by Katharevousa). Athens.

Philippides, Dia M.L. 1986. *The Sacrifice of Abraham on the Computer* (I Thisia tou Avraam ston Ipologisti). Athens.

Rosengren, Per & David Mighetto. 1986. *'ONE71' - ONE71-PROJEKTET. Språkvetenskaplig Forskning Rorande Spansk Skönlitteratur 1951-1971.* Göteborg: Göteborgs Universitet. ('ONE 71' - The ONE 71 Project. Linguistic Research on Spanish Fiction 1951-1971). University of Göteborg.

Sjögreen, Christian 1988. "Creating a Dictionary from a Lexical Database". *Studies in Computer-Aided Lexicology.* 299-338.

A CORPUS-BASED APPROACH TO MODERN GREEK LANGUAGE RESEARCH AND TEACHING

DIONYSIS GOUTSOS, OURANIA HATZIDAKI
& PHILIP KING
University of Birmingham

1. *Empirical research and corpora in Modern Greek linguistics*

Modern Greek (MG) does not have a strong tradition of empirical linguistic research such as many other European languages have. A major lack has been in the use of corpora. In the pre-electronic age, traditional grammars used a variety of sources without any systematic criteria, and relied heavily on introspection. Modern grammars are based on a more systematic collection of data but, in general, where corpora are used in linguistic research, they are not fully exploited and generalizations on sociolinguistic issues (cf., Kavoukopoulos 1989, Holton 1990), for example, are claimed on the basis of relatively small amounts of data. Philippaki-Warburton aptly summarizes the problem when she points out the need for "up to date descriptions under the guidance of the linguist" (1990:64) and the need for "objective linguistic and sociolinguistic appraisal of MG" (1990:65). It is our view that corpora are an excellent way of meeting this need.

In the last two decades, the small amount of work that has been done on Modern Greek corpora has remained little known outside the group of specialists for whom it was intended. In order to improve the availability of information on current and recent work, we designed and distributed a questionnaire to obtain information on the specifications (nature and size of corpus, hardware and software, processing tools and intended users and applications) of any corpus projects there might be. The response over a period of 6 months was quite encouraging and the findings are summarized in Appendix 1 here, although we cannot guarantee their comprehensiveness. They are discussed in detail in Goutsos, Hatzidaki and King (1993).

The survey findings indicate a respectable beginning to computer-based projects. There are two major projects aiming at the design of a general corpus of MG (CTI and ILSP, currently 10 million words each, with plans for expansion). Apart from general corpora, there are also literary corpora, specialized corpora and spoken corpora.

2. *The linguistic relevance of electronic corpora*

An electronic corpus can be defined as a collection of texts, of the written or spoken word, which is stored and processed on computer (Sinclair 1987). Corpora contain authentic language data, that is, instances of language produced in a real communicative situation (in the broad sense, including literature). This renders them particularly suitable for linguistic analysis; a major strength is that they consist of genuine examples of language use and not intuitive, invented sentences or text fragments. The intention of the producer has not been to produce samples of language but to achieve something through the use of language.

There is in principle no technical limit to the size of a corpus; as a rule of thumb, the bigger it is, the more reliable findings from it are likely to be, and the stronger the claims that what it shows may be characteristic of the language at large.

The standard tool of the corpus linguist is the KWIC (Key Word In Context) concordance (see Appendix 2). This is "a collection of the occurrences of a word-form, each in its own textual environment" (Sinclair 1991:32). The major advantage of this technique is that, by directly juxtaposing all usages of a word, it brings to light linguistic phenomena which normally pass unnoticed in plain running text. Essentially, it visually merges the two Saussurean dimensions of language, the single instance (*parole*) and the collective knowledge (*langue*) in a single set of axes (Tognini-Bonelli 1993). A further corpus-processing technique which is of particular relevance to Modern Greek lexicography is lemmatization. This automatically conflates the inflectional variants of a word into one lemma, a process which is extremely tedious and time-consuming if done manually. Finally, quantitative processing (frequency counts, statistical analysis, etc.) is an essential component of the corpus approach now that the amount of language data available is rapidly increasing.

Research with corpora has not only manifested their usefulness as field sites in which to explore traditional linguistic questions. It also, and most crucially, points to the ways in which the analysis of large amounts of empirical data can re-define traditional questions and effectively redraw the map of linguistic analysis. In this context, the relevance of electronic corpora for Modern Greek linguistics would seem obvious. Here, we would like to offer some suggestions on areas which are likely to be amenable to investigation by means of electronic corpora. Our examples are drawn from a (necessarily) small corpus (Philip King's Spoken Corpus) and, therefore, should not be taken as anything but first hints at likely-to-be-useful hypotheses.

First, issues of morphology and lemmatization are fruitful areas for exploration. A glance at any corpus data shows that not all inflected forms of a word may occur or not all occur with the same frequency. For instance, in our data the form *arχisi* is more frequent than *arχizo* and *arχizis*. Research can be

expected to show significant patterns of distribution and frequency; these may be associated with different meanings. Our data on the lemma *ora* shows varying frequency of occurrence of the different case forms, and also enables us to see that, in this corpus at any rate, the singular and plural forms are associated with different meanings: in the plural (*ores*), the commonest meaning is 'hours'; in the singular, the commonest meaning is 'time' (cf., also the lemma *o anθropos*: human being/somebody vs. *i anθropi*: people(s)/some persons). Greek corpora should be a rich field for examining the frequency and distribution of fossilised phrases, idioms, patterns and routines (such as: *en to metaksi*, etc.) as well as the productive range of morphemes such as: *-pio*, etc. or augmentative and diminutive suffixes.

At the level of lexis, computer analysis can be expected to shed light on the otherwise notoriously difficult area of collocation. Concordances can offer graphic representation of meaningful distinctions and patterns of association which would not be apparent otherwise. Data analysis suggests that meaning is associated with or even dependent on the pattern of co-occurrence of a word. A concordance of the lemma *tropos* in our data (see appendix 2) shows accusative singular as the commonest case form. Most of these occur after the preposition *kata*, and most of these in turn occur in the fixed expression *kata kapjo(n) tropo*. The nominative is the next most frequent form, but has the different sense of mode or manner. These findings clearly could not have been reached by simple recourse to intuition. Notice further that our statements about the language can no longer be couched in the 'all-or-nothing' terms of the grammar rule or the dictionary, but instead are expressed in terms of frequency or probability.

Corpus analysis can also provide insights which may help us to distinguish free variation from functional variation. Is the distinction between *na* and *ja na*, *os* and *eos*, *san* and *os*, *δioti*, and *jati* meaningful or do these constitute pairs of stylistic variants, as would seem to be the case for *prama* and *pragma*? What about the different forms of the perfect (*eχo kurasti - ime kurazmenos*) or of the relative pronouns (*pu - o opios*)? The examination of their distribution and relative frequency by means of concordances is more likely to reveal any contrastive function than the traditional juxtaposition of invented examples. The Greek examples may thus justify Sinclair's claim that "each meaning can be associated with a distinct formal patterning. [...] There is ultimately no distinction between form and meaning" (1991:6-7).

As becomes apparent, in a corpus perspective the levels of syntax, semantics and lexis interact and overlap in significant ways. However, issues which typically fall under the domain of syntax such as word order may also be investigated with the help of corpora. Furthermore, discourse questions such as the patterning of discourse markers (like *δilaδi*, *as pume* , etc.) or the patterning of pronouns seem worth investigating. For *lipon*, concordances from our data suggest second position in the sentence as the most frequent.

Considering the lack of work on spoken discourse, Greek linguistics can be profited by the raised sensitivity to the importance of collecting natural discourse brought about by corpus work. A corpus perspective may be instrumental in the area of text-typology (Biber 1988) and in the exploration of sociolinguistic questions such as orality and diglossia. In Goutsos *et al.* (1993) we argue for the necessity of a spoken corpus of Greek on the basis of the predominance of oral modes in Greek society. Diachronic corpora recording language change can also be assumed to be especially significant for a language like Greek.

3. *Pedagogical applications of corpora*

There is a vast area of applications in which Corpus Linguistics has proved its worth with the predominant example of the ground-breaking COBUILD dictionary (Sinclair *et al.* 1987). Corpus-based approaches have begun to come into use in the pedagogy of English as a Foreign Language (Johns and King 1991) and are beginning for other languages. A corpus of 1.75 million words of English takes up about 10Mb, and with the software required for processing, now available commercially, can easily be installed on a PC. For Greek, there are a few complications, but these are diminishing as time goes on.

Methodologically, there are many ways of using the fruits of corpus analysis in language learning. The first possibility is to cut out the teacher entirely, and let learners have direct access to the corpus, formulating their own questions about the language, and exploring the data to answer them. This provides another mode of private study and a valuable supplement to existing ways of learning. It puts learners firmly in charge of taking their own decisions about what to study and how. However, there can still be a role for the teacher. In the first place, teachers can access corpora to answer their own questions about the language and its patterning. This can be of great help in teacher preparation for example. Secondly, it is possible to use concordance searches or output to prepare class exercises (see Tribble and Jones 1990, Johns and King 1991 for examples of how this can be done for English). Software to enable the construction of exercises from concordance data is now under development (Johns in preparation).

For Greek as a Foreign Language, some of these possibilities are more important than they are for English. There may be no teachers of Greek in some areas, and anyone wanting to learn the language therefore has to make use of radio or TV or distance study materials. Greek is less well supplied with teaching materials; once a corpus is available, processing is easy, and the permutations of what can be discovered are endless.

In mother-tongue teaching too, an important role can be argued for this approach. It would enable students to make detailed studies of how words are used, and help them to develop their own 'word power', make informed stylistic judgements, and improve their own writing skills. Corpora consisting of school

textbooks for different ages would also make it possible to see at a glance what patterns of language and usages students are being exposed to in their studies. As far as Greek is concerned, all these applications are still in the future, but there is no doubt that they will open up flexible and beneficial methods of language study, including the preparation of a pedagogical grammar as Babiniotis (1994) suggests. In this way the computer will be brought into use as an integral part of the broader curriculum.

REFERENCES

Babiniotis, George. 1994. *Contemporary Linguistics and the Teaching of Modern Greek.* This volume.

Biber, Douglas. 1988. *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Goutsos, Dionysis, Ourania Hatzidaki & Philip King. 1993. *Towards a Corpus of Spoken Modern Greek.* Paper presented at the ACH-ALLC93 Georgetown University Conference, June 16-19, Georgetown University.

Holton, David. 1990. "Modern Greek Today - One Grammar or Two". *Greek Outside Greece II* ed. by Maria Roussou & Stavros Panteli, 23-33. Athens: Diaspora Books.

Johns, Tim. In preparation. *Contexts: A Program for Generating Concordance-based Exercises.* Birmingham: University of Birmingham (EOSU).

Johns, Tim & Philip King. (eds). 1991. *Classroom Concordancing.* Birmingham: University of Birmingham ELR Series.

Kavoukopoulos, Fotis. 1989. "I Dinamiki tis Genikis sti Neoeliniki" (The Dynamics of the Genitive in Modern Greek). *Studies in Greek Linguistics. Proceedings of the 10th Annual Meeting of the Department of Linguistics,* 265-284. University of Thessaloniki.

Philippaki-Warburton, Irene. 1990. "Linguistic Theory and MG". *Greek Outside Greece II* ed. by Maria Roussou & Stavros Panteli, 53-66. Athens: Diaspora Books.

Sinclair, John McH. (ed). 1987. *Looking Up.* London: Collins.

-----. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

----- *et al.* 1987. *Collins Cobuild English Language Dictionary.* London and Glasgow: Collins.

Tognini-Bonelli, Elena. 1993. "Interpretative Nodes in Discourse: Actual and Actually". *Text and Technology. In Honour of John Sinclair* ed. by Mona Baker, Gill Francis & Elena Tognini-Bonelli, 193-212. Amsterdam & Philadelphia: John Benjamins.

Tribble, Chris & Glyn Jones. 1990. *Concordances in the Classroom.* Harlow: Longman.

Appendix 1

Summary of current projects on Modern Greek electronic corpora

NAME	*CONTENT*	*SIZE*	*H/WARE & S/WARE*	*PERIOD*	*USE*
Oxford University	Literary	3Mb	VAX, PC X-Writer	C19-C20	literary teaching
King's College	Literary	25000 lines	VAX, Mac, concord.	C12-C16	lexicogr. literary
J. Burke Melbourne	Literary	15000w 800 Kb	Mac, Word	62-72	linguistic literary
IBM GD	NLP Lexicon	2.4Mb compact	AIX,PC,VM lemmat		NLP
Stephany	Child Speech	1.4 Mb	IBM,CLAN (Childes)	70-72	psycholing
Alexa	Ads	48 Kb	UNIX,PC	91-92	NLP
ECI	Document	24000w			multiling
Sklavounou	Terms	8000 w	PC, concord		bilingual
P.King (B'ham)	Spoken	36000w 190 Kb	PC, Locoscript	80-	linguistic
Georgakop. Edinburgh	Spoken	80 Kb	PC, WordCraft	91-	linguistic
ILSP (Athens)	Press Docs	10.2m w	UNIX,PC, concord	75- 87-92	lexicogr linguistic
WCL (Patras)	Press Technical	8-10m w	VAX,PC,tagger, parser	-1993	lexicogr. NLP
GREVOC (Sweden)	Press	11.5Mb	UNIX,PC, Mac, conc.	89-92	lexicogr linguistic
CTI (Patras)	Press Docs	3.5Mb	PC, statistic	90	linguistic NLP

Appendix 2

CONCORDANCE LINES FOR τρόπο
From Philip King's spoken corpus

άλος κίνδυνος: υπήρχαν ορισμένοι τρόποι - όπως η αριθμητική, ή γραμματική
ιμοποιήσουμε με κάπως διαφορετικό τρόπο. M.P. Αυτή η πληροφορική, η ινφορμ
ξέρει αριθμητική. Κατά τον ίδιο τρόπο, υπάρχουν ήδη πολλά συστήματα που π
ναι των υπολογιστών. Με τον ίδιο τρόπο θα μπορούσε λοιπόν κανείς να πει, δ
ς χειρότερους. Εμείς κάτα κάποιο τρόπο στέλνουμε πολύ κόσμο στο εξωτερικό.
την επιστήμη και έχω κάτα κάποιο τρόπο μια αρχαιότητα - νόμισα ότι ήταν χρ
Ελλάδα είναι να γίνει κατά κάποιο τρόπο κάτι, να γίνουν άλλοι καλύτεροι, να
την Αγγλία, οι οποίοι κατά κάποιο τρόπο έχουνε βοηθήσει εις την ανάπτυξη τη
εί να υπήρχανε, αλλά, κατά κάποιο τρόπο οι ξένοι ανταγωνισται μας, κι οι Αγ
έρχονται. Δηλαδή πως κατά κάποιο τρόπο, όλη η κλασική γνώση που ξέραμε, εί
ξεπερνάει. Και πως, κατά κάποιο τρόπο, έχει σηκώσει τα χέρια ψηλά... M.A.
αρύγιο στην τέχνη και κατά κάποιο τρόπο πικραίνομαι - χωρίς να μετανιώνω -
η και έπρεπε να φύγει κατά κάποιο τρόπο για την ολιγαρχία και για τους Αμερ
υπολογισμούς. Είναι κάτα κάποιο τρόπο ένα διανοητικό εργαλείο. Δηλαδή απ
ίδιοι να αναπτύξουμε κάτα κάποιο τρόπο τη δική μας την τεχνολογία ... M.P.
της τεχνολογίας, που κατά κάποιο τρόπο μπορούμε να αγοράσουμε απ' έξω, παρ
εφάλων ... Δ.Τ. Ναι. κάτα κάποιο τρόπο, ο επαναπατρισμός των εγκεφάλων. Κ
κή του πρωτοβουλία. Κατά κάποιον τρόπο, αποκτήσαμε ένα Μπικήνα. M.P. Κύρι
πωνες είναι πολύ πιο δεμένοι στον τρόπο που προγραμματίζουν το πού θα πάνε,
ιστήμης με κάποιο πιο συστηματικό τρόπο. Δηλαδή, η κάθε Ευρωπαϊκή χώρα προ
να ψέλνει. Εψελνε με το δικό της τρόπο διάφορα τροπάρια. Πολλές δε φορές
. Και φοβάται μήπως μ' αυτόν τον τρόπο αδικήσει αυτός το μαθητή του. Με α
αι ο λόγος είναι ότι μ' αυτόν τον τρόπο μπορεί αυτές τις πληροφορίες να τις
ε κι εγώ ν' απαντήσω μ' αυτόν τον τρόπο. M.P. Αλλά θεωρείς ότι σε σπατάλησ
έει ή Ελλάδα τώρα λέει. Βέβαια ο τρόπος που διασκεδάζουν οι εγγλέζοι είναι
σης η διακριτικότητα και αθόρυβος τρόπος με τον οποίο ασκούσαν την επιχειρη
ς ακούνε την όπερα, ποιος είναι ο τρόπος που την ακούνε. Εσύ πώς ξεκίνησει
τικό εργαλείο, πώς θα βρεθεί ένας τρόπος να χρησιμοποιήσει κανείς άλλες θεω
ιν: Να μπορούν να βρουν αυτομάτως τρόπους για να κάνουν γρήγορα υπολογισμού
ς ερώτηση, πέστε, αναπτύξτε, τους τρόπους που οι υπολογιστές μπορούν να χρη
και μάλιστα με πολύ διαφορετικούς τρόπους. Δηλαδή, όχι μόνο να είναι με γρ

AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE

General Editor
E. F. KONRAD KOERNER
(University of Ottawa)

Series IV - CURRENT ISSUES IN LINGUISTIC THEORY

Volume 117

Irene Philippaki-Warburton, Katerina Nicolaidis & Maria Sifianou (eds)

Themes in Greek Linguistics

THEMES IN GREEK LINGUISTICS

PAPERS FROM THE FIRST INTERNATIONAL CONFERENCE ON GREEK LINGUISTICS, READING, SEPTEMBER 1993

Edited by

IRENE PHILIPPAKI-WARBURTON
University of Reading

KATERINA NICOLAIDIS
University of Reading

MARIA SIFIANOU
University of Athens

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM/PHILADELPHIA