Towards a Corpus of Spoken Modern Greek

DIONYSIS GOUTSOS, OURANIA HATZIDAKI, and PHILIP KING
School of English, University of Birmingham, UK

Abstract

Our aim in this paper is twofold: to provide a state-of-the-art description of corpora of Modern Greek and, on the basis of this, to argue for the development of a spoken corpus of Modern Greek. To this end, we discuss the main attempts at electronic text collection in the context of empirical research in Greek linguistics, present a survey of existing corpora we have conducted, and provide an assessment and an identification of current needs at the beginning of the nineties. In the second part, we put forward a linguistic and pedagogical rationale for a corpus of Modern Greek spoken texts and delineate the basic features of its design.

1. Introduction

The need and usefulness of Corpus linguistic research on Modern Greek can be argued on a number of grounds. First, because of the distinctively long history of Greek, there is the potential for systematic historical work linked with the existing well-established projects on Classical Greek, such as the *Thesaurus Linguae Grecae* (University of California at Irvine) and the *Perseus Project* (Harvard University). (A computer search at the Georgetown Center for Text and Technology database of projects requested by us in 1992 has indicated 10 major corpora which store Classical and New Testament Greek texts).

From a synchronic point of view, the fact that a large number of the twelve to thirteen million Greek speakers live in various countries of the 'diaspora' can be both an advantage and a disadvantage in the development of Modern Greek corpora: at present, there is little communication and dissemination of information between researchers around the world, but as our survey has shown, projects have been undertaken in places as far apart as Australia and Sweden. At the same time, the status of Greek as a European Community language has increased the opportunities for research and the demand for teaching Greek as a Foreign Language.

The significant structural differences between Greek and Germanic and Romance languages have also played a crucial role in the progress of research. The use of a different alphabet has made necessary the development of special software (and hardware), while the elaborate systems of noun declension and verb conjugation present a constant challenge to parsing and tagging.

Finally, Greek is interesting from a sociolinguistic point of view due to the (related although independent) issues of diglossia and the predominance of orality in Greek society. The use of corpora may prove to be undoubtedly crucial not only in providing us with the 'facts' of the language but also by recasting the issues on a more fruitful basis.

Correspondence: Dionysis Goutsos, School of English, University of Birmingham, Edgbaston B15 2TT, UK. E-mail: dionysis@collins.co.uk

2. Empirical Research and Corpora in Modern Greek Linguistics

In view of this, the relative absence of any systematic use of corpora in linguistic research would seem surprising. However, it can be explained if we consider the fact that there has hardly been a tradition of empirical linguistic research on Modern Greek similar to that found in other countries (see Francis, 1992; Stubbs, 1993). In the pre-electronic age, traditional grammars have tended to use a variety of sources without any reasoned criteria (e.g. Mirambel, 1978 [1959] quotes examples from literature, newspapers, magazines, essays and translated works, from 1824 to 1957 and Seiler, 1952 refers additionally to dialectal works and dictionaries). The emphasis was predominantly on literary works and especially folk songs with the concomitant undervaluing of everyday speech material.

Tzartzanos's *Syntax* (1946–63) is illustrative of this tendency: in the preface he states that he 'first and foremost' consulted folk songs and literary works (10–11: τα δημοτικά τραγούδια κατά πρώτον και κυρίως υπόψιν — νεοελληνικά λογοτεχνήματα ... [first and foremost, folk songs, then Modern Greek literary words]) and only mentions 'everyday speech' at the end (11: Τέλος ελάβομεν προ οφθαλμών αυτόν τούτον τον καθ ημέραν λόγον όπως ομίτος κινείται εις τα κυριότερα αστικά κέντρα της Ελλάδος [Finally, we considered everyday speech itself as this is current in the main urban centres of Greece]).

Modern grammars are based on a more systematic collection of data [although Joseph and Warburton (1987) employ mainly data from other accounts of aspects of Greek grammar or constructed data]. Mackridge criticizes the shortcomings of traditional grammars and reverses the emphasis on literature at the expense of other sources (1985; p. vii). Other linguistic research has used corpora only to a limited extent. As an exception, one may note Laskaratou (1989) (whose data are drawn randomly from newspapers, dating back into the early eighties) and Rydå (1988) (a corpus consisting of 1,200,000 words from twenty-two newspapers). Spoken data figure prominently in the work of Hedin (1987) (who also included material excerpted from newspapers, literary works, periodicals etc.—the spoken language data come from ten hours of tape recordings from radio and television broadcasts) and Makri-Tsilipakou (1983) (who also uses radio and TV as sources along with private conversations, dating from 1982–3). All in all, corpora, if they are used at all in linguistic research, are not fully exploited.

Biber and Finegan (1991, p. 203) point out that there is an irony surrounding the prevalence of introspection over corpus-based analysis, since modern linguistic theorizing received its modern impetus from historical linguistics which was rooted in the analysis of corpus-

based data. In Greek linguistics, this irony is all the greater, given the longer predominance of historical linguistics as well as the long tradition of dialectology and collection of folk texts. It is especially paradoxical, therefore, that arguments on linguistic issues are still based on an extremely small amount of data; thus Kavoukopoulos (1989) makes statistical observations based on a corpus consisting of half-an-hour of spoken data and one newspaper article and Holton (1990) argues about diglossia on the basis of two newspaper articles, thirty pages of an essay and twenty-four pages from a novel. Philippaki-Warburton aptly summarizes the problem when she points out the need for 'up to date descriptions under the guidance of the linguist' (1990, p. 64) and the need for 'objective linguistic and sociolinguistic appraisal of MG' (1990, p. 65).

3. The Development of Electronic Corpora

In 1986, Tsitsopoulos referred to a 'double vacuum' in the computational research in Modern Greek language, constituted by 'the lack of a substantial body of theoretical work on the Greek language inspired by contemporary linguistic paradigms, and the total absence of ongoing programmes, academic or otherwise, in any branch of computational linguistics, however vaguely this field is defined' (1986, p. 149). This observation, although fairly accurate in general terms, is probably an overstatement, given that at that time the first two major concordances on Greek texts had already appeared, namely Kazazis *et al.*'s concordance to the *Opera Omnia* of Makriyannis (first published in 1983, see Kyriazidis and Kazazis, 1992) and Philippides' work on *Erotokritos* and *The Sacrifice of Abraham* (Philippides, 1986, 1988).

These early attempts at a computational analysis of Modern Greek texts dealt successfully with the linguistic and technical problems posed by Greek. However, their scope was limited: they each analysed literary landmarks of a single author from earlier stages of Modern Greek (seventeenth and nineteenth centuries). They were designed for purposes of literary analysis and have a linguistic relevance only from a diachronic point of view.

In short, it would seem that in electronic corpora, as in their non-electronic predecessors, literary texts have tended to be favoured, perhaps because literary style has been held to be the prestige variety and a model for all to copy (see *inter alia* Mackridge, 1985, p. 338ff).

In the last two decades, there has not been much general awareness of work on Modern Greek electronic corpora. Philippides (1981) was not aware of any literary projects in Modern Greek carried out with 'extensive mechanical assistance', and appealed for information. Burke (1992) found himself starting from scratch, and the MGSA Bulletin (1993) appeals for a listing of projects to be compiled as there is a growing number of reports, but information is still piecemeal.

4. The State-of-the-Art in Greek Corpora

In order to try to get a complete and up-to-date picture of current 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 projects there might be. The questionnaire was based on the example of the NERC Textual Data Survey, the Georgetown University Center for Text and Technology project, as well as Taylor *et al.* (1991) (see Appendix 1).

The response over a period of six months was quite encouraging—although we cannot guarantee that our findings are necessarily comprehensive. These are summarized in Table 1 below (see Appendix 2 for explanation of abbreviations and more detail).

We have gathered detailed information about fourteen projects—for one (MLDA) we have not managed to find out many details—which can be arranged in four major categories:

(i) Literary Corpora: Oxford, King's, Burke—these follow the tradition of the 1970s and 1980s (see Section 3) by concentrating on (major) literary works and aiming at literary research;
(ii) Specialized Corpora: IBM, Alexa, Tennis, Stephany, ECI—consisting of a series of electronic databases, ranging from dictionaries (IBM) to Greek data within larger collections (ECI), with a specialized interest;
(iii) General Written Corpora: GREVOC, CTI, ILSP, WLC—large collections of written texts, mainly from newspapers, including journals, official documents (EC, court cases, wills), academic papers, and literature;
(iv) General Spoken Corpora: only two projects concentrating on spoken data are reported: King, Georgakopoulou—with emphasis on linguistic analysis.

The period covered by the texts is broad: the emphasis is on the last three decades, but literary corpora go as far back as Medieval Greek (from the twelfth century onwards) and WLC includes some Ancient Greek texts from around 800 BC. CTI focused on a limited period (July–September, 1990).

For all these corpora, the texts were mainly entered by keyboarding, but a lot were also acquired from publishers in machine-readable form. Corpora with spoken data used personal recordings from observation (Stephany, Georgakopoulou) or from the radio (King). Literary corpora used standard printed editions, along with keyboarding from microfilm and original manuscripts.

With regard to the size, three categories can be established:

(i) small corpora: less than 1 Mb—because of the nature of the data or the purpose for which it was assembled: Alexa, Tennis, ECI, Georgakopoulou, King, Burke (800 Kb);
(ii) medium-sized corpora: with more than 1 Mb: Stephany, Oxford, King's, IBM;
(iii) large corpora: over 3 Mb: CTI (3.5Mb), ILSP (5+5.2 million words), WLC (8–10 million words), GREVOC (11.5Mb).

Table 1 Current projects on Modern Greek electronic corpora

NAME	CONTENT	SIZE	H/WARE & S/WARE	PERIOD	USE
Oxford	Literary	3Mb	VAX, PC X-Writer	C19-C20	literary teaching
King's College	Literary	25000 ls	VAX, Mac, concord, lemmat	C12-C16	lexicogr literary
Burke	Literary	15000w 800 Kb	Mac, Word	62-72	linguist literary
IBM GD	NLP Lexicon	2.4Mb compac	AIX,PC,VMl emmat		NLP
Stephany	Child Speech	1.4 Mb	IBM,CLAN (Childes)	70-72	psycho-ling
Alexa	Ads	48 Kb	UNIX,PC	91-92	NLP
ECI	Document	24000w			multilin
Tennis	Terms	8000 w	PC, concord		biling
P.King	Spoken	46000w 300 Kb	PC, Locoscrpt	80-	linguist
Georgak.	Spoken	80 Kb	PC, WordCraft	91-	linguist
ILSP	Press Docs	10.2m w	UNIX,PC, concord colloc	75-87-92	general lexicogr linguist
WCL	Press Technic	8-10m w	VAX,PC, tagger parser	- 1993	general lexicogr NLP
GREVOC	Press	11.5Mb	UNIX, PC, Mac, concord	89-92	general lexicogr linguist
CTI	Press Docs	3.5Mb	PC, statistic	90	linguist NLP
MLDA				76-88	

4. Assessment and Current Needs

The survey details given show that computer-based projects in Greek have begun a respectable development. From an international perspective it may be instructive to compare the amount of work done on Modern Greek with that done on Swedish, with a roughly comparable number of native speakers. For Swedish, Gellerstam (1992) reports eighteen current projects undertaken in the 1990s alone, of which four involve spoken data with a combined corpus size of just over 1,000,000 words.

There are two major projects aiming at the design of a general corpus of MG (CTI and ILSP, currently ten million words each, with plans for expansion). In both cases, the bulk of the corpus is made up of journalism, official documents and technical texts. Some literature and personal prose is also included. The smaller corpora tend to be more restricted in terms of content range. They are mainly single genre collections which concentrate on literature, specialized terminology, newspaper texts or official documents.

The research purposes of corpus compilation vary according to the content. The major areas of interest relate to the text types collected: literary and linguistic analysis are predominant, and NLP, multilingual studies and lexicography (an area lagging behind in Modern Greek studies) also feature centrally.

In the technical sphere a variety of systems are used. Hardware facilities mainly include IBM machines and compatibles, Apple Macs and Unix machines. One of the central issues of MG computerized linguistic research is the Greek character set. In certain cases a latinized transliteration is used (ECI, Alexa. cf. Philippides' project). Most researchers however have been able to adhere to Greek characters, as most computer and software manufacturers now offer adapted hardware as well as applications packages able to handle Greek. It is becoming progressively easier to import and export Greek text files while preserving orthographic and accentual features. Processing of machine-readable modern Greek text is therefore a reducing problem in technical terms. This is in part due to the advances made by projects working on Ancient and Medieval Greek which share the same alphabet and have thus had to face the same issues. However, the area of optical character recognition is still lagging far behind, although progress is being made.

Corpus processing software is in most cases customized. Programs available include monolingual and parallel concordancers, part-of-speech taggers, lemmatizers, programs to remove typesetting commands from data in machine-readable form, statistical analysis packages, etc. The development of parsers and collocation programs is under way. The general tendency in the technical area is for the design of more varied and sophisticated text-processing tools; however, lack of communication and sharing of expertise has resulted in duplication of effort in some cases and has seriously delayed technical advancements. In many cases corpus processing tasks are carried out manually or semi-automatically when at the same time the relevant software already exists. There is also lack of awareness of the fact that certain basic programs such as concordancers need not be designed from scratch as they are already available in the form of pre-written packages.

Moreover, issues of standardization have only poorly been addressed. Even though in most cases corpus data are made available by researchers to other interested parties, compatibility is not guaranteed. This is a crucial point with regard to the diachronic study of MG. One way of facilitating interchange would be if MG text collections were compatible with the substantial Ancient and Medieval Greek corpora which already exist (see Section 1), but the problem is still some way from solution.

With regard to plans for further development (see Section 6), researchers' comments make it clear that rapid promotion of corpus studies is a necessity and in many cases a set goal. This involves enlargement of corpora by the addition of more texts, broadening of the variety of text types or elimination of particular genres in order to achieve content balance. Systematization of the principles for text sampling is also considered.

The question of balance in the corpus has been a concern for a few respondents. ILSP have a long-term goal of establishing a balanced corpus with a view to developing a monitor corpus along the lines of the Bank of English Corpus (Sinclair, 1987). Given this and similar objectives, the most glaring omission from such corpora so far is any spoken data.

This is a general problem with existing corpora in all languages; in Sinclair's words 'most corpora keep well away from the problems of spoken language ... and this is most unfortunate' (1991, p. 15). Zampolli's recent survey of European language corpora (excluding English) indicates a gulf between quantities of written and spoken material: only 5% of the total 365 million words already collected are from spoken texts (quoted in Leech, 1991, p. 21). The gulf reported above between proportion of spoken and written texts for Swedish is bad enough; the overall position for Greek is even less well balanced.

5. Towards a Spoken Modern Greek Corpus

5.1 *The Necessity for a Spoken Corpus*

As noted above, there is a gulf between spoken and written corpora. The earliest known computer-held corpus of spoken English dates from as recently as 1963 (Sinclair *et al.*, 1970). Spoken corpora have always been more difficult to assemble than written ones, because (i) the gathering of the raw data is more complex, and (ii) they have to be converted to some kind of written form before any processing can take place (see Svartvik, 1991, p. 556, for the problems). They are likely in the future to lag even further behind corpora of the written language simply because more and more written text is now being produced directly in machine-readable form, while conversion of spoken data is no quicker than before.

To compound the problem, it may not be easy to agree on how the transcript should be encoded (Edwards, 1992) in the first place, and decisions taken at this stage may well pre-empt the kind of analysis that can be done later. The argument **for** a spoken corpus in any language rests on the belief that spoken and written forms of the language differ in certain significant ways (Biber, 1988; Halliday, 1989) and that there is a linguistic or pedagogical payoff (Svartvik, 1991, p. 560). The usefulness of spoken corpora has been extensively argued on the basis that 'the spoken form of the language is a better guide to the fundamental organization of the language than the written form' (Sinclair, 1991, p. 16). It should be reflected that while most language use is in the form of unscripted dialogue, much of our knowledge of language is based on prepared material.

Practical applications of spoken corpora to language teaching and language pedagogy are beginning to be developed. One obvious value is in the large number of learners of English who could ultimately benefit from an accessible description of the spoken language, though Svartvik noted as recently as the 1991 Georgetown conference that the London–Lund corpus had served as a basis for descriptive rather than pedagogical research. It may be felt that this reflects a natural sequence of operations, but in fact it is possible to address pedagogical concerns usefully with relatively little processing [as is implicit throughout Johns and King (1991), although most papers in this collection report on the basis of written corpora]. There is no doubt either that the range and scale of the work on English means that there are the resources and skills to undertake relatively many analyses in distinct areas of grammar, discourse studies, intonation and so on.

For Greek, the basic arguments are the same. However, because the range of resources that can be brought to bear is several orders of magnitude smaller, the question of prioritizing which way the corpus construction and exploitation goes is much more important. The argument for developing a spoken corpus for Modern Greek is moreover particularly strong if we consider the almost total lack of linguistic work done so far on the spoken language at any level beyond segmental phonology. A corpus of spoken Greek would provide much-needed information which is not already available and would emphasize spoken text as a body of knowledge and an object of study in its own right.

There are two specific points to be made for Greek, however, one sociolinguistic and the other pedagogical. The sociolinguistic argument is that by contrast with most English-speaking countries, 'the culture of Modern Greece is still to a larger extent an oral one ... The advent of the telephone, radio and television in this century has only served to consolidate the oral basis of the culture. Their conversational style is likely to be very different from their written.' (Mackridge, 1985, p. 338). Even where spoken language occurs in the same situation, Greek may serve a different function, as Sifianou's (1989) investigation of telephone behaviour in Greece and England points out.

Pedagogically, the number of learners is certainly smaller, but as against this, it needs pointing out that there is currently no database which could provide unmediated evidence as to the nature of the spoken language. Svartvik (1992) argues (with regard to English) that real spoken data provides evidence of difference in information structure in speech, in particular that the tone unit rather than the (much longer) sentence is the basic information unit, and that one useful activity among others for language learners is to study minimally marked-up transcripts as a way of getting at the make-up of spoken English from the point of view of real-time performance. In particular the recognition of preassembled chunks of language plays an important part in raising awareness.

5.2 *Developing a Spoken Corpus*

The nature of a spoken corpus is very much a question of what it is feasible to gather. Questions of design are raised by Du Bois (1991), Edwards (1992) and others. Considerations include: what you need to collect, what it is feasible to collect (and any gap between these), and questions of how much and what detail of phonological features need to be transcribed. In addition, for Greek there is the issue of standardization of encoding conventions.

With regard to the design, a broad range of spoken text-types should be covered. Categories identified in the Bank of English (COBUILD) Corpus (Renouf, 1984) or the London–Lund corpus (Svartvik, 1992) will be applicable to Greek in general terms. Some types are expected to be more central and others less well established.

The proposed general policy is to follow the Bank of

Table 2 Text-types in Philip King's Corpus

RADIO RECORDINGS: TEXT TYPES	
SPONTANEOUS	WRITTEN TO BE READ
discussion programs	news
phone-ins	documentary
interviews	lecture
dj-chat	public service announcements
field reports	
sports commentary	

English example of 'clean' text (Sinclair, 1991), so that in the first instance tagging and parsing would not constitute a problem. The first step is a broad 'orthographically normalized' transcription. This follows from the general purposes of the corpus and has the initial advantages of minimizing the cost and time of transcription and maximizing its accessibility and readability (Atkins *et al.*, 1992). However, provision should be made for comparability and adaptability; thus the first layer of a broad transcription should be designed so as to be easily adapted to a two-layered system like the one used in the Corpus of Spoken American English (Du Bois, 1991).

These principles have guided the design of the corpus which one of us (Philip King) has begun to develop. At present it consists almost entirely of radio broadcasts, because these have been easiest to gather. Within this text type, there has been an attempt to maximize variety. The range is summarized in Table 2.

The basic distinction is between language which has been written-to-be-spoken, and spontaneous speech (cf. references above). The categories are set up for convenience, but any radio program may move between them. For instance, a scripted news bulletin may contain a recorded or live clip from a correspondent in the field; a documentary may contain an unscripted interview as an insert, and so on. While radio can capture a variety of speech behaviour, it is not sufficient to be able to claim representativeness of all types of speech. In particular, informal conversation is going to be poorly represented. However this requires much more elaborate preparation to obtain, and on our resources is not feasible. The main drag on developments is doing the transcription.

5.3 *Applications and Implications for Further Research*

Linguistically and sociolinguistically, a spoken corpus will enable a beginning to be made on investigating the nature of the orality claimed for Greek. It would thus be expected to contribute to the enhancement and improvement of empirical research in Modern Greek.

A second area which a spoken corpus can feed into is the teaching of Greek both as a first and as a foreign language. First, it will cover the needs of lexicography—an area still lagging dramatically behind in Greek—by providing texts which reflect current language usage and facilitating the development of a descriptive dictionary of Modern Greek. Second, it will provide a substantial resource for extracting and studying authentic language data either for presentation in the classroom or for the preparation of teaching materials.

Of the two main types of text focused on above, clearly the scripted texts do not serve as a model for real-time production. They do, however, represent valid listening-practice material. The spontaneous types will have clear lessons for fluency and interaction activities [see Johns and King (1991) for examples].

In general, we may also expect that, as has been the case with corpora already developed for other languages, a spoken corpus will provide a valuable resource for computational linguists, translators, and sociolinguists. It can be expected to steer analysis towards an empirical approach, and stimulate research on issues of intonation, transcription, quantitative analysis, and the writing-speech differences. It would also yield detailed data on recent and current usage that would facilitate discussion on the present state of diglossia in Greek.

6. A View to the Future

In summing up, Greek corpus development has come some way, although it has been fragmented and with little contact between corpus-based researchers. The need for a spoken corpus has become more obvious as this picture has begun to develop. Perspectives for the future envisaged by respondents to our survey include the following plans for development: (i) the extension of existing corpora with the encoding of further data (Stephany, Georgakopoulou, Alexa, Oxford, MGD, King); (ii) the broadening of the types of data collected to include literary texts in General corpora (GREVOC), non-literary texts for Literary corpora (Burke), a wider variety of media (CTI), aiming towards representativeness (WCL); (iii) changes in the design (ISLP); (iv) development of software for automatic tagging, converting characters, parsing, concordancing and the addition of TEI markup.

The idea of establishing a General Corpus attracted the interest of all respondents and there appears to be a distinct willingness to participate in such a project. Areas of usefulness of such a General Corpus which they stressed are, first, lexicography, followed by linguistic analysis, literary analysis and speech synthesis (WCL), while teaching also received great emphasis (King, GREVOC). Multilingual corpora were also favoured (WCL, CTI, Alexa), although not by all (GREVOC, Burke).

A most important common theme in the responses to the questionnaire was the need for cooperation and sharing of information and software. It is hoped that our paper by giving the current state of play across a broad range of corpora will serve to bring researchers into closer contact with each other and will thus be seen as a contribution towards greater cooperation and dissemination of information.

Appendix 1: The questionnaire used for the survey

A. Corpus Profile
A1. By what name is the corpus known?
A2. Who compiled the corpus?
A3. Where was it compiled? (Institution)
A4. Contact Address (Telephone, Fax, E-mail)

A5. When did the compilation start?
A6. What was the incentive for starting the compilation?
B. Computer Facilities and Software
B1. How are texts entered? (word-processor, text-editor, typesetting tapes, optical scanning, other)
B2. How is the corpus stored and in what format?
B2.1. What computer facilities do you use? (IBM Personal Computer or compatible, Apple Macintosh—workstation—mainframe)
B2.2. What software do you use for corpus processing? (please specify item and function: word frequency, concordancing of selected items etc.)
B2.3. Do you use ready-made or customized software?
B2.4. If you use your own software, which programming language do you use?
B3. How do you handle the special problem of Greek characters?
- in input processing
- in screen output
- in printing

B4. Do you have software for linguistic annotation (tagging, parsing, lemmatization)?
If yes, specify
C. Text Details
C1. How was the text acquired?
C2. How is the corpus organized?
C3. Can you give some details of the content?
C3.1. Written texts:
C3.1.1. What genres are included in your collection?
C3.1.2. What are the media of the original texts? (printed book, periodical, manuscript, ephemera, other)
C3.1.3. Do you encode typographic and layout information? If so, specify
C3.2. Spoken texts (transcriptions):
C3.2.1. What genres are included in your collection?
C3.2.2. What is the medium of the original source? (TV, radio, telephone, direct: talk, conversation, other)
C3.2.3. Is the material spontaneous or not, surreptitious or not?
C3.2.4. Do you encode information about speakers (e.g. age, sex) or about the recording?
C3.2.5. What transcription system do you use? (phonetic, phonological, enhanced orthographical, orthographical)
C4. What period do the texts in the corpus represent?
from _____ to _____
C5. What is the total amount of data stored in your collection?
- in bytes
- in words
- in minutes of spoken text recording

C6. What use is made of the corpus? (specify, where appropriate)
- to build up a multifunctional linguistic corpus
- for lexicographic purposes
- for literary research
- for stylistic research
- for preparation of a scholarly edition
- for research in linguistics
- for research in language learning/teaching
- for commercial applications
- for natural language processing applications
- other

C7. Is it available to other interested parties?
If so, under what conditions?
D. Views and Perspectives:
D1. Do you plan any changes in the composition of your corpus?
D2. Are you planning to develop new text-handling software?
D3. Are there any specialized areas of Modern Greek for which a corpus approach would be particularly useful?
D4.1. What are your views on the development of a general corpus of Modern Greek (such as the Brown Corpus of English or the Birmingham English Corpus)?
D4.2. What would you consider to be the optimal size of it?
D5. Do you prefer a 'clean text' strategy (i.e. plain orthographic files)
as opposed to annotated, phonologically coded, parsed etc. text?
D6. Do you think that multilingual corpora or corpora containing 'parallel texts' are needed?
D7. Do you have any other views on the development of Modern Greek corpora and software for processing them?
E. Publications:
Please list any publications that you are aware of that were based on the electronic text you describe

Appendix 2: Current projects on Modern Greek Corpora:

Greekads.txt
Compiled by: Melina Alexa
Compiled at: UMIST, CCL
Sampling period: 1991–
How transcribed: W/P
Storage details: PC & workstation
Software tools: MTAS; TACT
Greek characters: transliterated alphabet
Details of material: job classified ads
Organization: one ad per text unit
Language variety: Non-fiction
Period of texts: 1991–1992
Size: 48K; 11 025 tokens (6328 annotated)
Use of corpus: NLP applications
Availability: not yet

J. Burke's Corpus
Compiled by: J.B. Burke
Compiled at: University of Melbourne
Sampling period: 1990–
How transcribed: OCR (much hand-edited), Microsoft Word
Storage details: Apple Mac
Software tools: C, customized
Greek characters: difficulty in OCR; Mac characters
Details of material: Tachtsis *To Trito Stefani*, Ioannou *I Sarkofagos*
Organization: in chunks of books (200K each)
Language variety: Literature
Period of texts: 1960–1972
Size: 800K; 15,000 words
Use of corpus: linguistic and literary analysis
Availability: only for research purposes
Other: RTF encoding

CTI INTRALEX Project (Patras)
Compiled by: CTI Team (Dr D. Christodoulakis)
Compiled at: CTI Patras
Sampling period: July 1990–
How transcribed: text-editor
Storage details: IBM
Software tools: word frequency; trigram/digram analysis; C, C++
Greek characters: Greek page of MS-DOS (437); postscript for printing
Details of material: newspaper articles (incl. interviews, want ads, advertisement), documents of court cases, wills
Organization: thematic
Language variety: Non-fiction

Period of texts: July 1990–September 1990
Size: 3.5 Mb
Use of corpus: linguistic research, teaching, NLP applications
Availability: unconditional
Other: developing automatic tagging software

ECI Corpus
Compiled by: ECI (D. McKelvie)
Compiled at: ECI (European Corpus Initiative)
Sampling period: current
Greek characters: latin characters
Details of material: the Greek version of the EC ESPRIT call for research proposals
Language variety: Non-fiction
Size: 24 000 w
Use of corpus: multilingual corpus for scientific research
Availability: public domain

Georgakopoulou's Corpus
Compiled by: Georgakopoulou, Alexandra
Compiled at: University of Edinburgh
Sampling period: 1991–
How transcribed: word-processor
Storage details: normal orthography
Software tools: WordCraft
Details of material: Intra-conversational narratives
Language variety: Spoken
Period of texts: 1991–
Size: 80Kb
Use of corpus: linguistic; narrative analysis

GREVOC
Compiled by: Dr Bo-Lennart Eklund
Compiled at: University of Gothenburg
Sampling period: 1990/91–
How transcribed: from typesetting files
Storage details: UNIX, accessible on-line in 3 formats: international ASCII, IBM, Kadmos (AppleMac)
Software tools: 'Conc' 1.70 beta, 'Transcribe' (for Mac); customized program in Turbo–Pascal (for IBM)
Greek characters: Hercules (PC); Kadmos (Mac)
Details of material: newspaper ('Eleftherotypia', 'To Vima tis Kyriakis') and journal ('Diavazo') texts
Organization: integrated text concordance index
Language variety: Non-fiction
Period of texts: Dec 1989–Aug, 1992
Size: approx. 11.5 Mb (1.8 mil words)
Use of corpus: basis for general corpus, lexicography, linguistic research
Availability: accessible on-line

IBM Greek Dictionary
Compiled by: IBM Greece
Compiled at: IBM Greece, IBM Bathesda
Sampling period: 1987–
How transcribed: XEDIT editor on VM-OS
Storage details: Compacted & encrypted with IBM architecture; accessible on DOS, Windows, OS/2, AIX, OS/400, VM, MVS
Software tools: customized (in C)
Greek characters: IBM customized software; keyboard remapper
Details of material: Basic word list with hyphenation, morphological and synonym data and algorithmic hyphenation rules
Organization: N/A
Language variety: NLP lexicon
Size: 2.4Mb (compacted), 650,000 words (headwords)
Use of corpus: NLP applications
Availability: by licencing agreement as part of NLP service (contact Ian Hersey)

ILSP Corpus
Compiled by: ILSP (plus partners)
Compiled at: ILSP, Athens
Sampling period: 1992–
How transcribed: typesetting tapes, OCR; cartridges
Storage details: various; ASCII files—IBM; UNIX
Software tools: concordancer, word-frequency count, collocator–customized software
Greek characters: ELOT, ISO standards—special drivers for keyboard
Details of material: printed books, periodicals, newspapers, ephemera (LOGOS); CELEX database, official documents (TRANSLEARN)
Organization: text type, language type, media (LOGOS); text-tupe, sublanguage (TRANSLEARN)
Language variety: Non-fiction
Period of texts: 1975-(LOGOS); 1987-92 (TRANSLEARN)
Size: 5 million words (LOGOS); 5.2 million words (TRANSLEARN)
Use of corpus: general corpus, linguistic, teaching, NLP applications, MT applications
Availability: only for research purposes (LOGOS), no (TRANSLEARN)
Other: TRANSLEARN is part of the construction of parallel corpora in the framework of TRANSLEARN/LRE

Philip King's Corpus
Compiled by: Philip King
Compiled at: University of Birmingham
Sampling period: 1980–
How transcribed: Locoscript; normal orthography
Storage details: textfiles; ASCII files
Software tools: Locoscript; MicroConcord (OUP)
Greek characters: Locoscript
Details of material: Radio scripted (news; documentaries) and unscripted (discussions; phone-ins)
Organization: individual textfiles of each program
Language variety: Spoken
Period of texts: 1980–
Size: 23,000 words (growing); 150Kb
Use of corpus: linguistic analysis (by concordancing)
Availability: to be arranged

Medieval Greek Database
Compiled by: Dr R. Beaton
Compiled at: King's College, London
Sampling period: 1989–
How transcribed: WP
Storage details: Apple Mac
Software tools: OCP in VAX mainframe
Greek characters: SuperGreek with Word 4 convertible to ASCII
Details of material: Digenes Akrites (2 versions), Livistros & Rodamni (5 versions)
Organization: plain text with line nos, lemmatized concordances
Language variety: Literature
Period of texts: 12th–16th C.
Size: approx 25,000 lines
Use of corpus: lexicography, literary and stylistic research, preparation of scholarly edition
Availability: not yet (contact R. Beaton)
Other: two more verse romances to be added

MLDA Greek Corpus
Compiled by: Professor W. Paprotté

Compiled at: University of Münster
Sampling period: 1984–
How transcribed: W/P, typesetting files
Software tools: WordCruncher, customized software
Greek characters: ISO 8879; ETL 16/Gresun character sets (for screen)
Period of texts: 1976–1988

Oxford Modern Greek Text Project
Compiled by: Dr Peter Mackridge
Compiled at: University of Oxford
Sampling period: 1992–
How transcribed: W/P
Storage details: IBM–PC & VAX
Software tools: Micro-OCP & OCP; Chiwriter (adapted) for text entering
Greek characters: Chiwriter fonts convertible to ASCII
Details of material: Poems of: D. Solomos, C.P. Cavafy, G. Seferis (standard editions)
Organization: As in printed volumes (each on one disk)
Language variety: Literature
Period of texts: 19th and 20th C. (1820–1972)
Size: approx. 3Mb
Use of corpus: literary research, undergraduate teaching
Availability: perhaps in the future

Spontaneous Greek Child Speech
Compiled by: Ursula Stephany
Compiled at: Institut für Sprachwissenschaft, Universität zu Köln
Sampling period: 1970
How transcribed: Microsoft Word 4.0
Storage details: IBM compatible
Software tools: CLAN programs of the CHILDES project
Greek characters: Latin chars
Details of material: Spontaneous children's speech
Organization: according to the CHAT format of the CHILDES project
Language variety: Conversation
Period of texts: 1970–1972
Size: 1,044,000 bytes (unformatted)
Use of corpus: language acquisition research
Availability: not yet
Other: Publication: Stephany, U. (in prep.). The Acquisition of Greek. To appear in D.I. Slobin (ed.), *The Crosslinguistic Study of Language Acquisition*. Erlbaum, Hillsdale, NJ. Vol. 4.

Tennis Corpus
Compiled by: Sclavounou Elsa (supervision Maurice Gross)
Compiled at: Aristotle University, Greece; Centre d' Etude et de Recherche en Informatique Linguistique, Paris
Sampling period: May–July, 1992; Nov, 1992–Jan, 1993
How transcribed: DBase IV Filemaker; Excel 3.0 Windows
Storage details: IBM
Software tools: concordances
Greek characters: customized DOS program; Universal Greek Math font (accents still a problem); AppleMac printer
Details of material: Greek tennis compound nouns with French and English equivalents from press, television, tennis manuals
Organization: records
Language variety: Terminology
Size: 8000 lexical entries, 4000 records
Use of corpus: lexicography, linguistic research
Availability: to LADL, CERIL and researchers interested in parallel corpora

WCL Corpus of Modern Greek
Compiled by: WCL Language & Speech Research team
Compiled at: Wire Communications Lab, University of Patras
Sampling period: 1987–
How transcribed: word-processor
Storage details: VAX, PC
Software tools: customized (Fortran, C)
Greek characters: adapted by manufacturers (high ASCII)
Details of material: newspaper, EC official docs, literature, academic (technical)
Organization: major categories; to be thematically structured by source and subject
Language variety: Non-Fiction; Prose
Period of texts: (800 BC)–1993
Size: 8–10 million words
Use of corpus: to build general corpus, for lexicographic and NLP applications (speech synthesis, tagging etc.)
Availability: access on site after permission
Other: working towards representative corpus of MG

Note

This paper is a revised and updated version of one delivered at the ACH–ALLC 1993 Georgetown University Conference.

References

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus Design Criteria, *Literary and Linguistic Computing*, 7: 1–16.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

—— and Finegan, E. (1991). On the Exploitation of Computerized Corpora in Variation Studies. In K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics*, Longman, London, pp. 204–20.

Burke, J. (1992). Research Tools for Modern Greek, *ALLC–ACH 1992 Conference Abstracts*.

Du Bois, J. (1991). Transcription Design Principles for Spoken Discourse Research, *Pragmatics*, 1.1: 71–106.

Edwards, J. (1992). Design Principles in the Transcription of Spoken Discourse. In J. Svartvik (ed.), *Directions in Corpus Linguistics*. Mouton de Gruyter, Berlin/New York, pp. 129–44.

Francis, W. N. (1992). Language Corpora BC. In J. Svartvik (ed.), *Directions in Corpus Linguistics*. Mouton de Gruyter, Berlin/New York, pp. 17–34.

Gellerstam, M. (1992). Modern Swedish Text Corpora. In J. Svartvik (ed.), *Directions in Corpus Linguistics*. Mouton de Gruyter, Berlin. [= Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991] pp. 149–63.

Halliday, M. A. K. (1989). *Spoken and Written Language*. Oxford University Press, Oxford.

Hedin, E. (1987). *On the Use of the Perfect and Pluperfect in Modern Greek*. Almqvist and Wiksell, Stockholm [PhD: Studia Graeca Stockholmiensia: 6].

Holton, D. (1990). Modern Greek Today—One Grammar or Two. In M. Roussou and S. Panteli (eds), *Greek Outside Greece II*. Diaspora Books, Athens, pp. 23–33.

Johns, T. and King, P. (ed.) (1991). *Classroom Concordancing*. University of Birmingham, Birmingham [= ELR Journal 4].

Joseph, B. and Philippaki-Warburton, I. (1987). *Modern Greek*. Croom Helm, London.

Kavoukopoulos, F. (1989). Η δυναμική της γενική Νεοελληνική' [The dynamics of genitive in Modern Greek], Μελέτες για την Ελληνική γλώσσα 1989, Konstandinidis, Thessaloniki, pp. 265-84.

Kyriazidis, N.I. and Kazazis, I.N. (1992). Τα ελληνικά του Μακρυγιάννη με τον υπολογιστή [Makriyannis' Greek on the computer] A tagged concordance of and other indices to the Opera Omnia of Makriyannis. Papazisis, Athens.

Laskaratou, C. (1989). *A Functional Approach to Constituent Order with Particular Reference to Modern Greek.* Journal 'Paroussia', Athens.

Leech, G. (1991). The State of the Art in Corpus Linguistics. In K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics.* Longman, London, pp. 8–29.

Mackridge, P. (1985). *The Modern Greek Language.* Clarendon Press, Oxford.

Makri-Tsilipakou, M. (1983). Απόπειρα περιγραφής της νεοελληνικής προοφώνησης' [An attempt to describe Modern Greek forms of address], Μελέτες για την Ελληνική γλώσσα 1983 Konstandinidis, Thessaloniki, pp. 219–239. *MGSA Bulletin* (1993) 25(1), Spring, 1993.

Mirambel, A. (1978) [1959]. Η Νέα Ελληνική Γλώσσα. [Modern Greek Language] Aristotle University of Thessaloniki, Thessaloniki.

Philippaki-Warburton, I. (1990). Linguistic Theory and MG. In M. Roussou and S. Panteli (ed.), *Greek Outside Greece II.* Diaspora Books, Athens, pp. 53–66.

Philippides, D. (1981). Computers and Modern Greek, *Mantatoforos*, 17: 5–13.

—— (1986). *The Sacrifice of Abraham on the Computer.* Hermes Press, Athens.

—— (1988). Literary Detection in the *Erotokritos* and *The Sacrifice of Abraham, Literary and Linguistic Computing*, 3: 1-11.

Renouf, A. (1984). Corpus Development at Birmingham University. In J. Aarts and W. Mejis (eds), *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English.* Rodopi, Amsterdam, pp. 3–39.

Rydå, S. (1988). *Present and Aorist Participles in Contemporary Greek Newspapers.* Almkvist and Wiksell, Stockholm.

Seiler, H. (1952). *L' Aspect et le Temps dans le Verbe Néogrec.* Les Belles Lettres, Paris.

Sifianou, M. (1989). On the Telephone Again! Differences in Telephone Behaviour: England versus Greece, *Language in Society*, 18: 527–44.

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

—— (1991). *Corpus, Concordance, Collocation.* Oxford University Press, Oxford.

——, Jones, S. and Daley, R. (1970). *English Lexical Studies.* Office for Scientific and Technical Information, University of Birmingham.

Stubbs, M. (1993). British Traditions in Text Analysis—From Firth to Sinclair. In M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, Benjamins, Amsterdam, pp. 1–33.

Svartvik, J. (1991). What Can Real Spoken Data Teach Teachers of English, *Proceedings of Georgetown University Round Table on Languages and Linguistics*, pp. 555–66.

—— (1992). Corpus Linguistics Comes of Age. In J. Svartvik (ed.), *Directions in Corpus Linguistics.* Mouton de Gruyter, Berlin. [= Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991] pp. 7–13.

Taylor, L., Leech, G. and Fligelstone, S. (1991). A Survey of English Machine-readable Corpora. In S. Johansson and A.-B. Stenström (eds), *English Computer Corpora: Selected Papers and Research Guide.* Mouton de Gruyter, Berlin pp. 319–54.

Tsitsopoulos, S. (1986). Linguistic Research in the Greek Group, *Multilingua: Journal of Interlanguage Communication*, 5: 149–51.

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