The Corpus of Greek Texts: a reference corpus for Modern Greek

Dionysis Goutsos[1]

Abstract

This paper reports on the construction of a reference corpus for Modern Greek, the Corpus of Greek Texts (CGT), that is currently being developed at the University of Athens. In particular, it points out the need for an authoritative corpus of Greek in view of the limitations of existing attempts to compile corpora for the language. It also presents the aims and identity of CGT with particular reference to its structure (composition of data and text classification). Questions of corpus design, which are particularly important with respect to available resources for Greek, are considered in relation to the issue of representativeness in material selection. The phases of implementation of CGT compilation are presented in detail. Finally, the larger implications of the project are detailed and applications, as well as prospects for further development, are outlined. Special mention is made of linguistic research papers on aspects of Greek that have used CGT data.

1. Modern Greek corpora

This paper documents the design and implementation of a new reference corpus for Modern Greek, the Corpus of Greek Texts (henceforth, CGT). This corpus was developed initially as a joint project between the University of Athens and the University of Cyprus, and it is now in the final phase of implementation at the University of Athens.[2] The CGT has been designed as

[1] Department of Linguistics, Faculty of Philology, School of Philosophy, University of Athens, 157 84 Zografou, Athens, Greece
Correspondence to: Dionysis Goutsos, *e-mail:* dgoutsos@phil.uoa.gr

[2] The first phase of implementation was financed by the University of Cyprus (entitled, 'Basic Corpus of Greek Texts') and the second phase was supported by the research project, 'Pythagoras', at the University of Athens. Earlier documentation of the project can be found under Goutsos (2003a) and Goutsos and Pavlou (forthcoming), in Greek and English, respectively. The webpages for the two phases of implementation are found at www.ucy.ac.cy/sek and www.greekcorpora.org. Research for this paper and the final phase of

Corpora 2010 Vol. 5 (1): 29-44
DOI: 10.3366/E1749503210000353
© Edinburgh University Press
www.eupjournals.com/cor

a representative reference corpus of Greek, consisting of a substantial amount of data (30 million words) that is to be used as a basis for linguistic research and as a resource for teaching applications. It is now available and freely accessible online.[3]

Goutsos *et al.* (1994) provided an early summary of the scant resources for Modern Greek available at the time and pointed out the need for a reference corpus of the language, along the lines of Kennedy (1998: 291), who remarks that the most important need of current linguistic research is 'a systematic and comprehensive programme of research on the structure and use of particular languages [which will make its] results easily accessible'. Since then, there has been only one major attempt to establish a reference corpus of Greek, the ILSP Corpus, now developed to constitute the Hellenic National Corpus (HNC).[4]

This Greek corpus was compiled in the early 1990s and has since been revised and expanded. It contains texts published from 1976 (Hatzigeorgiou *et al,* 2001: 813) to 2007 and has followed the sampling procedures of earlier English corpora, by including fragments of texts rather than entire texts (see Renouf, 2007: 28). In addition, despite its considerable size (47 million words at present), it has not involved a systematic collection of varied text types but has mainly focussed on journalistic texts, which happened to be more easily available when the project started. A further complication is that no details are given in the relevant publications about the overall structure of the corpus or the classification scheme that is used. According to the information gleaned from the corpus website (see Appendix 1), it can be surmised that 61.29 percent of the texts included come from newspapers, while 23.08 percent are unclassified with respect to medium. Similarly, 51 percent of the texts included belong to the 'Informative' text-type and 38.25 percent belong to the "Opinion" text-type, whereas within text-type categories the overwhelming majority of texts are left unclassified. It seems, thus, that its overall design has been rather opportunistic - dictated by the needs of developing computational tools rather than representing the state of the language. Finally, a major problem with the Hellenic National Corpus concerns accessibility: the online version gives free access to five concordance lines, while unlimited access is only available to subscribers.

By contrast, the only other large-scale project involving Modern Greek corpora since the early 1990s offers free and well-designed access to corpora of somewhat limited text types. These are the corpora available at the Portal for the Greek language (Πύλη για την ελληνική γλώσσα) from the

implementation, including its online publication, were supported by the University of Athens research programme, 'Kapodistrias', code 70/4/7607.

[3] See: www.sek.edu.gr

[4] Documentation of the ILSP corpus includes Hatzigeorgiu *et al.* (2000) and in Greek, Gavriilidou *et al.* (1993), and Hatzigeorgiu *et al.* (2001). The related webpage can be found at: hnc.ilsp.gr

Centre for the Greek Language, and consists of data from two newspapers and school handbooks.[5]

We can conclude, then, that research into Greek has suffered from a lack of linguistic projects that would combine the features mentioned in Kennedy's quote above, (i.e., systematicity, comprehensiveness and accessibility), in a context where most European languages are now turning from super-corpora to cyber-corpora, to use Renouf's (2007) terms.[6] It is this situation that the creation of the CGT seeks to remedy, by giving emphasis to the needs of linguistic research. Its design explicitly addresses issues of comprehensiveness and representativeness, while special care has been taken to make the outcome of the project as freely available as possible. In this sense, it can be claimed that the CGT aims to create 'a body of text which could be claimed to be an authoritative object of study' (Renouf, 2007: 32) in the fashion of large, general corpora that already exist for other languages. The remainder of this paper presents the more specific aims and the structure of the CGT, with particular reference to issues of design and implementation, followed by a discussion of future applications and prospects.

2. Aims and identity of the Corpus of Greek Texts

The Corpus of Greek Texts was envisaged as a core collection of Modern Greek texts, stored in an electronic format and representative of basic genres in the language, to be used for linguistic analysis and pedagogical applications. Its main characteristics are as follows:

- it represents a well-defined collection of texts from a variety of genres that are central in Greek contexts of communication and useful for the teaching of Greek as a first/second language;
- it contains a substantial percentage of spoken data, constituting the biggest existing collection of spoken Greek;
- it contains a substantial percentage of data from Cyprus, offering for the first time a valuable resource for the study of Greek geographical varieties;
- it is designed as a basis for larger (e.g., monitor) corpora of the future; and,
- it is freely available online to researchers and learners.

To summarise, the CGT has been designed as:

- a general or reference corpus;
- a monolingual corpus, including a major geographical variety (Cypriot Greek);

[5] The webpage for these corpora is: http://www.greek-language.gr/greekLang /modern_greek/index.html

[6] Goutsos *et al.* (1994a) draw a comparison with languages like Swedish, which is of comparable size in terms of speakers, as well as other European languages.

— a mixed corpus, including both spoken and written material; and,
— a synchronic corpus, collecting data from two decades (1990 to 2010).

In terms of the size of the CGT, the aim of the project is to collect 30 million words in total. Although this would seem a rather small corpus by current standards, it should be viewed in the context of existing projects in Greek. The case of the HNC indicates that the major priority for Greek corpus compilation is not to increase the size of the corpus but to enhance the range of text types covered, avoiding, at the same time, producing a collection of genres that is biased. In addition, it must be noted that this target number of words should be sufficient for major applications; to take a prominent example, Cobuild 1 was based on a 20 million corpus (Sinclair, 1987). Finally, this amount is projected to cover the needs of linguistic research for the next decade with a view to expanding the CGT into a monitor corpus of Greek, in which new material could be constantly added and old data would be removed.

3. **Corpus design and the question of representativeness**

The design of the corpus closely matches the explicit aims of the project presented above. The selection of written and spoken texts, and the scope and type of text types for compilation, are inextricably linked with the question of representativeness, since, according to Sinclair's (1996: 4) definition of corpus ('a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language'), the selection and arrangement of language material follows specific linguistic criteria, which make this material a representative sample of the language in question. Of course, what 'representative' means has been a vexed issue in corpus linguistics and researchers have taken opposite views as to criteria of representativeness. Barnbrook (1996: 24), for instance, points out that a linguistic sample should have features similar to those of the linguistic population it aims to represent in the analysis of a language. In fact, sampling, especially in the case of reference corpora, aiming to represent general use, can take different forms. Thus, corpora like the BNC have been compiled on the basis of a strict classification of genres, based on statistical sampling for spoken data, whereas the Bank of English has developed into a monitor corpus - a huge database of material that is constantly renewed to the extent that questions of representativeness become moot (Barnbrook, 1996: 25). These are two central examples of different ways of sampling in current practice based on statistical evidence and text taxonomy (see Biber *et al,* 1998: 248).

Following the discussion in Kučera (2002) with respect to the Czech National Corpus, we can understand representativeness as referring to three dimensions in each corpus: size, authenticity and proportionality. In terms of size, the CGT is still far behind other corpora in this phase of its development, even though, as noted above, large-scale linguistic applications have been achieved with corpora of a similar size. In addition, the CGT to a large extent satisfies the dimensions of authenticity and proportionality or relative balance between the various text types it contains. In particular, sampling is based on a variety of textual criteria, such as text type, subject, thematic area, medium, ***etc.,*** aiming to identify a broad spectrum of Greek genres, that are intuitively recognised by the linguistic community in question. Its identity ensures that only texts that satisfy certain criteria and only whole texts (where this is possible) are included. These texts come from contexts of communication that are of central importance in Greek and have been naturally created (that is, they were not produced under experimental conditions) so that they can be characterised as authentic. Translated texts, on the other hand, have been excluded (to the extent that this is possible) when collecting text from a wide variety and quality of sources.

Furthermore, the CGT aims to give special emphasis to types of data that have been neglected in Greek research, namely spoken data (see Goutsos *et al.,* 1994) and data from the Cyprus geographical variety (not the dialect as such), contributing, thus, to a more comprehensive view of the language.[7] In this way, representativeness is dependent on the aims of the CGT, which give emphasis to expanding existing resources on varieties of the Greek language. Finally, the proportionality of text types was also informed by evidence from reception studies, especially concerning reading, according to data from the National Book Centre of Greece.[8] Obviously, this concerns written data, whereas for spoken data similar studies do not seem to be viable or even useful.

It has to be pointed out that a main concern in designing the CGT has been the detailed and systematic coding of metadata for each text that is included so as to offer immediate access to the specifics of its origin, and thus allow monitoring of the textual classification used. This aspect drastically improves on existing Greek corpora, whose composition and structure, as discussed above, cannot be sufficiently checked.

Finally, one of the most important characteristics of the CGT is its in-built potential to be used as an archive of language resources. In other words, its architecture is flexible enough to allow for a broad range of combinations in selecting material and, thus, in creating different sub-corpora. In this sense, category descriptions and targets for the number of words in each category

[7] The contribution of spoken data to the overall composition of the CGT (10 percent or 3 million words), although at first glance seems to be small, corresponds to standard practice (cf. the BNC), due to the extremely demanding requirements for recording and transcription of spoken data.

[8] Details of related studies on book production and reception are given in http://www.ekebi.gr. These studies point out, among other things, the increasing importance of non-fiction (academic and popularised) books in Greek reading habits, something which is reflected in the corpus structure.

can be regarded as tentative, and can be replaced at any time, according to the needs of the user. Moreover, texts which cannot be used now in the compilation because they exceed word targets for their respective category are stored for later use.

4. From design to implementation

The implementation of the designed structure involves a series of procedures that have been standardised according to the needs of the project.[9] The main procedures of compilation are the following:

- Identification of data resources/development;
- Data collection;
- Transcription (for spoken data);
- Data clean-up and storage;
- Standardisation;
- Coding; and,
- Data annotation (to be developed).

In particular, a large part of the project has been taken up with the search for data sources and the development of linguistic resources relevant to the CGT. Data collection includes sound- or video-recording for spoken data, and scanning, typing or Internet searches for written material. Transcription of spoken material is broad orthographic, marking basic features of spontaneous discourse such as overlap, pauses, interruption, lengthening, *etc.* A digital copy of all spoken texts allows for more detailed transcriptions when the need arises in the future.

This is followed by data clean-up, such as the removal of non-verbal elements that are incompatible with the CGT's format and cannot be handled at present (e.g., pictures, blank spaces or lines). Standardisation includes basic annotation in terms of paragraphs, sections, titles, identity of speakers, *etc.,* where relevant. This information about text structure is included in the files. Finally, an independent database stores the metadata for each text, including author, date of production, title, first words, number of words, *etc.,* as well as detailed classificatory information.

As suggested above, classification in the CGT is multiple and involves the following:

- Mode: written-spoken;
- Medium: radio, television, live, book, telephone, newspaper, magazine, electronic, other;

[9] For the identification, acquisition and design of data, see Renouf (1987).

- Class: spontaneous *versus* planned (for spoken texts), information *versus* non-information (for written texts);
- Type: academic, popularised, law-administration, private, literature, news, opinion articles, interview, public speech, conversation, miscellaneous;
- Sub-type: 01-99, referring to specific sub-genres within each type; for example, one-to-one, one-to-many in conversation, humanities, social sciences or science in academic texts, socio-political, economic or leisure in news and opinion articles, and so on;
- Geographical variety: Standard Modern Greek *versus* Cypriot Greek; and,
- Keywords: words relating to the text's main topic taken from a list of themes, where applicable.

Flexibility, a feature pointed out above, arises from the multiple means of access to the above categories so that variation in the composition of the sub-corpora is possible; naturally, this is determined by research needs and priorities. For instance, users who do not agree with, or are in no need of, the coding of Class can disregard this category and select material from other categories. The same goes for the written *versus* spoken distinction, which, it could be argued, lies in a different position along the continuum from that predicted in the CGT. The multiple and detailed coding allows, thus, for a broad range of choice in selecting material, ensuring at the same time detailed identification of each text included in the CGT.

The implementation of the project consisted of four phases. The first phase involved the collection of spoken and written material, the transcription of part of the spoken data, the setting up of a webpage with information on the project and the preliminary design of applications. The second phase involved finishing compilation, the design of developments that may be made in future, improving the webpage, *etc.* The third phase involved the tentative online publication of the corpus, while in the fourth phase the CGT was made available to the public through a web interface. At the same time, the compilation continues with the remaining 2.5 million words to be added in the following year.

5. **CGT structure**

According to the aims and the design principles discussed above, a rough outline of the CGT's structure is given in Table 1.

Table 1 shows that the CGT combines linguistically relevant criteria with genre distinctions relating to the Greek society. For instance, distinctions such as novels *versus* short stories, or academic *versus* popularised non-fiction texts reflect common genre distinctions made in Greek, which are also found in several other corpora (e.g., the British National Corpus, the International Corpus of English in English), whereas labels such as written

Spoken	Spoken planned	News	Current affairs
			Entertainment
		Interview	One-to-one
			One-to-many
		Public speech	Academic
			Non-academic
	Spoken spontaneous	Conversation	One-to-one One-to-many
			Other
Written	Written non-information	Literature	Novels
			Short stories
			Biography
			Poetry
			Drama
			Fairy tales
			Lyrics
		Miscellaneous	Anecdotes
			Other
	Written information	News	
		Opinion articles	
		Information items	
		Academic	Humanities
			Social Sciences
			Science
		Popularised non-fiction	Humanities
			Social Sciences
			Science
		Law and administration	
		Private letters	
		Electronic texts	E-mail
			E-chat
		Diary	
		Ephemera	
		Procedural texts	
		Other	

Table 1 CGT structure according to medium

Medium	*Number of words*	*Percentage*
Book	6,190,045	22.73
Newspaper	8,054,039	29.58
Magazine	5,999,059	22
Electronic	1,598,291	5.87
Live	2,150,674	7.9
Radio	105,121	0.38
Television	675,485	2.5
Other	2,451,061	9
Total	27,223,775	100

Table 2: Classification of CGT texts according to medium

Mode	*Text type*	*Number of words*	*Percentage*
Spoken	News	291,382	1
	Interview	592,584	2
	Public speech	1,839,766	6.75
	Conversation	207,548	0.76
Written	Literature	2,455,080	9
	News	4,764,337	17.5
	Opinion articles	3,189,132	11.7
	Information item	100,570	0.36
	Academic	3,994,277	14.67
	Popularised	7,648,513	28
	Law and administration	1,472,700	5.4
	Private	186,210	0.68
	Procedural	145,770	0.53
	Miscellanea	335,906	1.65

Table 3: Classification of CGT texts according to text type

versus spoken, or planned *versus* spontaneous, reflect linguistic decisions in classification.

Tables 2 and 3 present the current status of the corpus with regard to the medium of texts and the basic text types included. The figures given correspond to the number of words currently included (January 2010).

The data under Table 2 suggest that, although one-third of the material currently included comes from newspapers, there is a wide variety of media from which texts are selected, including a substantial 8 percent of texts from spontaneous face-to-face (live) communication. In addition, spoken material currently accounts for more than 10 percent of the number of words collected, as has been the original provision in the design of the corpus.

6. Implications, applications and prospects

The particularities of the CGT, involving a less widely spoken language, such as Greek, are clearly expected to offer useful insights with respect to corpus design and compilation in various ways. Our experience has indicated the need for increased emphasis on both the widest collection of genres possible and greater flexibility in accessing these genres. This emphasis is necessary for redressing the balance in favour of text types such as conversation or electronic communication that have been comparatively neglected in Greek linguistic research and also because of the provisional nature of each text taxonomy, respectively. Since we aim to offer the possibility of research into a body of Greek texts which could be claimed to be authoritative, we have to develop an increased awareness of genres that are important for communication in Greek communities, including material such as e-mail, e-chat, television interviews, academic lectures, *etc.,* as well as data from a wide geographical spectrum. Giving access to linguistic varieties thus becomes one of the major tasks in corpus compilation and research for Greek.

Keeping these basic principles in view, the main applications of the CGT have already been envisaged in the following areas:

(a) *Linguistic research.* The CGT can offer invaluable data for corpus-based research on the lexis and syntax of Greek, the description of discourse and stylistic phenomena, the study of the spoken language and register variation, as well as sociolinguistic research on norms and dialectal phenomena (see Chafe *et al.,* 1991: 64-6). The CGT has already been used in the various phases of its development as a source for linguistic studies on a variety of aspects of Greek grammar and lexis, including discourse markers (Georgakopoulou and Goutsos, 1998), place adverbials (Goutsos, 2007), shell nouns (Koutsoulelou and Mikros, 2004-2005), the classification of Greek adjectives (Fragaki, 2009), male and female lexical noun and adjective pairs (Fragaki and Goutsos, forthcoming; and Goutsos and Fragaki, 2009), aspects of discourse and pragmatics (Goutsos, 2002, 2010), *etc.* It is also currently being used in Ph.D. research on lexical clusters (Ferlas, 2008) and academic vocabulary (Katsalirou, forthcoming).

(b) *Pedagogical applications.* The CGT has already been used for the development of pedagogical applications (Goutsos *et al.* 1994b; Goutsos, 2003b, 2006; and Goutsos and Koutsoulelou-Michou, 2009). The CGT will also enable the development of software applications for material design or use in the classroom (see Wichmann *et al.,* 1997). These can include unmediated

learner access to the corpus for self-learning purposes, Internet-based tools for distance learning, as well as specifically designed exercises to supplement classroom teaching, in the form of workbooks or CDs on the basis of authentic language material.

(c) *Interface with other projects.* The CGT is expected to contribute to: (*i*) translation studies by connecting with corpora in other languages as well as multilingual and parallel corpora, (*ii*) the development of Greek lexicography by interfacing with existing dictionaries of Greek, and *(iii)* historical studies, by relating Modern Greek to earlier phases of the language (linking with, for example, the Thesaurus Linguae Graecae, the Perseus Project, the Grammar of Medieval Greek or the Thesaurus of the Cypriot Language[10]). And,

(d) *Development of computational tools and applications.* The CGT is expected to offer material for the development of parsers, taggers and other computational tools. Although at this stage data coding does not include detailed annotation, the future development of the project includes both part-of-speech tagging and prosodic annotation of the linguistic material included in the CGT.

The goal of this paper has been to delineate the basic issues and problems arising with respect to the compilation of a reference corpus of Greek, as a case-study of a language with distinctive linguistic resources. As hinted at above, a major implication of this project concerns the process of re-designing the corpus as a means of incorporating feedback from implementation in the way illustrated by Biber (1993: 256). To this end, future plans include evaluation of CGT compilation practices and outcomes, which will feed back into the CGT's structure. It is certain that the further development of the CGT will radically change the current picture we have of the Greek language, providing evidence for a more comprehensive, accurate and authoritative description of the language.

References

Barnbrook, G. 1996. Language and Computers. Edinburgh: Edinburgh University Press.

Biber, D. 1993. 'Representativeness in corpus design', Literary and Linguistic Computing 8, pp. 1-15.

[10] The webpages for these projects can be found at: http://www.tlg.uci.edu, http://www.perseus.tufts.edu/hopper/collection.jsp?collection=Perseus:collection:Greco-Roman, http://www.mml.cam.ac.uk/greek/grammarofmedievalgreek and http://www.imkykkou.com.cy/politistiko_idryma_arxangelou.shtml, respectively.

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

Chafe, W.L., J. W. Du Bois and S. A. Thompson. 1991. 'Towards a new corpus of spoken American English' in K. Aijmer and B. Altenberg (eds) English Corpus Linguistics, pp. 64-82. London: Longman.

Ferlas, E. 2008. 'Lexical clusters in the Corpus of Greek Texts'. Paper given at the Formulaic Language Research Network (FLaRN) conference, 19 June 2008, University of Nottingham.

Fragaki, G. 2009. The evaluative role of the adjective and its use as a marker of ideology. (In Greek.) Ph.D. thesis, University of Athens.

Fragaki, G. and D. Goutsos. Forthcoming. 'Gender adjectives and identity construction in Greek corpora'. Proceedings of the seventh International Conference on Greek Linguistics. University of York. 8-10 September 2005. Internet publication.

Gavriilidou, M., P. Lambropoulou and S. Ronioti 1993. 'Design and annotation of a Greek corpus', Studies in Greek Linguistics 14, pp. 308-22. (In Greek.)

Georgakopoulou, A. and D. Goutsos. 1998. 'Conjunctions versus discourse markers in Greek: the interaction of frequency, positions and functions in context', Linguistics 36 (5), pp. 887-917.

Goutsos, D. 2002. 'The use of electronic corpora in discourse analysis' in C. Clairis (ed.) Recherches en linguistique grecque. Proceedings of the fifth International Conference on Greek Linguistics. 13-15 September 2001, pp. 219-22. Paris: L'Harmattan. (In Greek.)

Goutsos, D. 2003a. 'Corpus of Greek Texts: design and implementation' in Proceedings of the sixth International Conference on Greek Linguistics, University of Crete, 18-21 September 2003. CD-ROM publication. (In Greek.) Also available at: http://www.philology.uoc.gr/conferences/6thICGL/gr.htm

Goutsos, D. 2003b. 'The use of electronic corpora in the teaching of Modern Greek vocabulary' in Proceedings of the first International Conference on Teaching Greek as a Foreign Language, Athens. 25-26 September 2000, pp. 259-67. (In Greek.) Athens: University of Athens.

Goutsos, D. 2006. 'Vocabulary development. From the basic to the advanced level' in D. Goutsos, M. Sifianou and A. Georgakopoulou (eds) Greek as a Foreign Language: From Words to Texts, pp. 13-92. (In Greek.) Athens: Patakis.

Goutsos, D. 2007. 'Basic adverbs of space in corpora: preliminary remarks' in Department of Linguistics, University of Athens (ed.) Studies Dedicated to Dimitra Theophanopoulou-Kontou, pp. 36—46. (In Greek.) Athens: Kardamitsa.

Goutsos, D. 2010. 'Analysing speech acts with the Corpus of Greek Texts: implications for a theory of language' in M. Mahlberg, V. Gonzalez-Diaz and C. Smith (eds) Proceedings of the Corpus Linguistics Conference, CL 2009 University of Liverpool, 20-23 July 2009. Available online at: http://ucrel.lancs.ac.uk/publications/CL2009/

Goutsos, D. and G. Fragaki. 2009. 'Lexical choices of gender identity in Greek genres: the view from corpora', Pragmatics 19 (3), pp. 317—40.

Goutsos, D., P. King and R. Hatzidaki. 1994. 'Towards a corpus of spoken Modern Greek', Literary and Linguistic Computing 9 (3), pp. 215-23.

Goutsos, D., R. Hatzidaki and P. King. 1994. Ά corpus-based approach to Modern Greek language research and teaching' in I. Philippaki-Warburton, K. Nicolaidis and M. Sifianou (eds) Themes in Greek Linguistics: Papers from the First International Conference on Greek Linguistics. Reading, UK. September 1993. Amsterdam /Philadelphia: John Benjamins, 507-13. Reprinted in W. Teubert and R. Krishnamurthy (eds) Corpus Linguistics: Critical Concepts in Linguistics (Volume 6), pp. 150-6. London and New York: Routledge.

Goutsos, D. and S. Koutsoulelou-Michou. 2009. 'The teaching of academic vocabulary in Greek with the use of corpora' in Proceedings of the third International Conference on Teaching Greek as a Foreign Language. Athens. 22-23 October 2004. (In Greek.)

Goutsos, D. and P. Pavlou. Forthcoming. 'CGT: Building a reference corpus of Greek' in Proceedings of the thirtieth International Conference on Functional Linguistics, University of Cyprus, 18-21 October 2006.

Hatzigeorgiu, N., S. Spiliotopoulou, A. Vakalopoulou, A. Papakostopoulou, S. Piperidis, M. Gavriilidou and G. Karayannis. 2001. 'National thesaurus of Greek Texts: a corpus of Modem Greek on the internet', Studies in Greek Linguistics 21, pp. 812-21. (In Greek.)

Hatzigeorgiu N., M. Gavrilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou and I. Demiros. 2000. Design and implementation of the online ISLP corpus, in Proceedings of the LREC 2000 Conference, pp. 1737-42. Athens.

Katsalirou, A. Forthcoming. General Academic Vocabulary in the Teaching of Greek as a Foreign Language. (In Greek.) Ph.D. thesis, Aristotle University of Thessaloniki.

Kennedy, G. 1998. An Introduction to Corpus Linguistics. London: Longman.

Koutsoulelou, S. and G. Mikros, 2004-2005. 'The word γεγονός as a shell noun: use and function in Greek corpora', Glossologia 16, pp. 65-95. (In Greek.)

Kučera, K. 2002. 'The Czech National Corpus: principles, design and results', Literary and Linguistic Computing 17 (2), pp. 245-57.

Renouf, A. 1987. 'Corpus development' in J. Sinclair (ed.) Looking Up, pp. 1-40. London and Glasgow: Collins ELT.

Renouf, A. 2007. 'Corpus development 25 years on: from super-corpus to cyber-corpus' in R. Facchinetti (ed.) Corpus Linguistics 25 Years on, pp. 27-49. Amsterdam: Rodopi.

Sinclair, J. (ed.) 1987. Looking Up. London and Glasgow: Collins ELT.

Sinclair, J. 1996. 'Preliminary recommendations on corpus typology'. EAGLES document. Available online at: www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpus typ.ps.gz

Wichmann, A., S. Fligelstone, T. McEnery and G. Knowles (eds). 1997. Teaching and Language Corpora. London: Longman.

Appendix 1: HNC structure[11]

Classification according to text-type

Text-type	*Percentage*	*Sub-type*	*Number of texts*	
Opinion	38.25	General interest article	756	19,443
		Comment article	2,527	
		Other	16,160	
Informative	51	Announcement	15	25,913
		Press release	56	
		Essay	1	
		News article	218	
		Bulletin	7	
		Chronicle	8	
		Other	25,608	
Official texts	1.60	Legal document	72	809
		Proceedings	268	
		Other	469	
Scientific/Education	0.1	Thesis	8	55
		Essay	6	
		Research paper	14	
		Reference work	1	
		Study	23	
		Textbook	1	
		School book	2	
Private texts	0.02			10
Literature	0.8	Biography	211	420
		Short story	19	
		Novel	22	
		Novella	3	
		Children's novel	12	
		Other	153	
Conversation	8	Letters to the press	8	4,067
		Public talk/ lecture	4	
		Interview	395	
		Other	3,660	
Miscellaneous	0.33			107
TOTAL	100			50,824

[11] Source: http://hnc.ilsp.gr/subcorpus.asp# (last accessed: 26 February 2009)

Appendix 1: *{continued):* HNC structure

Classification according to medium

Medium	*Percentage*
Book	9.41
Internet	0.32
Newspapers	61.29
Magazine	5.89
Other	23.08