CGT: BUILDING A REFERENCE CORPUS OF GREEK

Dionysis GOUTSOS, Pavlos PAVLOU
University of Athens, University of Cyprus

This paper documents the design and implementation of a new reference corpus for Modern Greek, the Corpus of Greek Texts (henceforth CGT). This corpus has been initially developed as a product of the co-operation between the University of Athens and the University of Cyprus and is now at the phase of implementation at the University of Athens.[1] CGT represents a new, extensive and representative reference corpus of Greek, collecting a substantial amount of data (30 million words) to be used as a basis for linguistic research and a resource for teaching applications.

There has been only one major corpus of Greek so far, the ILSP Corpus, now developed to constitute the Hellenic National Corpus (HNC). This Greek corpus was compiled in the early 1990s but followed the sampling procedures of 'first generation' corpora, by including fragments of texts. In addition, it has not involved a systematic collection of varied text types but has mainly focused on journalistic texts, which happened to be easily available at the time. Its overall design thus seems oriented towards computational applications rather than linguistic research, while accessibility has been problematic, since, at least in the online version, no details are given about the overall structure of the corpus.[2] CGT has been developed in this context with the aim of restoring the balance with other European languages and by giving emphasis on the uses of electronic corpora for the needs of linguistic research. In the rest of this paper we will present the aims and structure of CGT, with particular reference to questions of design and implementation, followed by a discussion of future applications and prospects.

CGT has been envisaged as a core collection of Modern Greek texts, stored in electronic format and representative of basic genres in the language, to be used for linguistic analysis and pedagogical applications. Its main characteristics are the following:

- It represents a well-defined collection of texts from a variety of genres that are central in Greek contexts of communication and important 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, reflecting for the first time the geographical variation of Greek;

- It is designed as a basis for larger (e.g. monitor) corpora of the future;

- It will be available to researchers and learners through user-friendly applications.

---

1 The first phase of implementation was financed by the University of Cyprus (project: «Basic corpus of Greek texts») and the current phase is supported by the research project Pythagoras of the University of Athens. The project's webpage can also be found at the following URL address: www.ucy.ac.cy/sek.

2 Some details are given in the relevant webpage www.hnc.ilsp.gr.

In sum, CGT has been designed as:

- a general or *reference* corpus,
- a *monolingual* corpus, including a major geographical variety (Cyprus Greek),
- a *mixed* corpus, including both spoken and written material, and
- a *synchronic* corpus, collecting data from 1990 to 2005.

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

- Identification of data resources/ Development
- Data collection
- Transcription (for spoken data)
- Data clean-up and storage
- Standardization
- Coding
- 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 CGT. This is followed by data clean-up, involving getting rid of redundant, non-verbal elements that are incompatible with CGT's format (e.g. pictures, blank spaces or lines etc). Standardization includes basic annotation in terms of paragraphs, sections, titles, speakers etc., where relevant. Information about text structure is thus included in the files. Finally, an independent database stores the identity features of each text, including author, date of production, title, first words, number of words etc., as well as detailed classification information.

- As hinted at above, CGT classification is multiple and involves the following aspects:
- Mode: written-spoken
- Medium: radio, TV, live, book, telephone, newspaper, magazine, electronic, other
- Class: information–non-information
- Type: academic, popularized, law-administration, private, literature, news, opinion articles, interview, public speech, conversation, miscellanea
- Sub-type: 01-99
- Geographical variety: standard-Cyprus
- Keywords

Flexibility, a feature pointed out above, arises from the multiple ways of access to the above categories. In other words, classification is not binding in order to invoke a group of texts but allows for a varied composition of sub-texts according to research needs and priorities. For instance, users who do not agree with or are in no need of the coding of Class (*information vs. non-information texts*) can leave this out and select material by using other categories. The same goes for the written vs. spoken distinction, which could be argued to lie in a different position than that predicted in 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 CGT.

The design of the corpus matches closely 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 John Sinclair's definition of corpus,[3] 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.

Geoff Barnbrook,[4] for instance, points out that a linguistic sample should have similar features with those of the linguistic population it aims at representing in the analysis of a language. In fact, sampling, especially in cases of reference corpora, aiming at representing 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. These are two central examples of different ways of sampling in current practice based on statistical evidence and text taxonomy.

Following the discussion in Karel Kučera with respect to the Czech National Corpus,[5] we can consider that representativeness refers to three dimensions in each corpus: size, authenticity and proportionality (that is relative balance between the various text types contained in it). In terms of size, CGT cannot claim full representativeness of the language in this phase of its development, although, as noted above, large-scale linguistic applications have been achieved with corpora of a similar size. In addition, CGT to a large extent satisfies the other two dimensions. In particular, sampling is based on a variety of textual criteria such as text type, subject, thematic area, medium etc., aiming at an identification of a broad spectrum of Greek genres, intuitively recognized by the linguistic community in question. Its identity makes sure that only texts that satisfy certain criteria and only whole texts (where this is possible) are excluded. 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 characterized as authentic.

Furthermore, CGT aims at giving special emphasis on types of data that have been neglected in Greek research, namely spoken data[6] and data from the Cyprus geographical variety (not the dialect as such), contributing thus to a more comprehensive view of the language. In this way, representativeness is dependent on the aims of CGT, which point to a general picture of Greek with the widest applications possible. Finally, the proportionality of text types was based on reception studies, especially concerning reading, according to data

3 SINCLAIR John, 1996, "Preliminary recommendations on corpus typology. EAGLES document", available at www.ilc.pi.cnr.it/EAGLES/corpustyp/corpustyp.html.

4 BARNBROOK Geoff, 1996, *Language and computers*, Edinburgh, Edinburgh University Press.

5 KUČERA Karel, 2002, "The Czech national corpus: Principles, design and results", *Literary and Linguistic Computing* 17(2), p. 245-257.

6 GOUTSOS Dionysis, HATZIDAKI Ourania & KING Philip 1994, "Towards a corpus of spoken Modern Greek", *Literary and Linguistic Computing* 9(3), p. 215-223.

from the National Book Centre of Greece.[7] 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 CGT has been the detailed and systematic coding of identity features for each text that is included so as to offer immediate access to the specifics of its origin and thus allow monitoring the textual classification used. This aspect drastically improves on existing Greek corpora, whose composition and structure, as mentioned above, cannot be sufficiently monitored.

Finally, one of the most important characteristics of 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 creating different sub-corpora. In this sense, word targets for each category can be regarded as a tentative option, which 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 use in a later phase of development.

The particularities of CGT, involving a less widely spoken language such as Greek, are clearly expected to offer useful insight in 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, respectively, for redressing the balance in favour of text types that have been comparatively neglected in Greek linguistic research and because of the provisional nature of each text taxonomy. Since we aim at offering the possibility of research into the totality of the Greek language (however this may be conceived), 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, TV interviews, academic lectures etc., as well as data from a wide geographical spectrum. Giving access to language variation thus becomes one of the major tasks in corpus compilation and research.

Our goal in 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 noted above, a major implication of our project involves the process of re-designing the corpus as a means of incorporating feedback from implementation.[8] In other words, future plans include evaluation of our compilation practices and results, which will feed back into CGT's structure. We are positive that the further development of CGT will radically change the current picture we have of the Greek language, providing evidence for a more comprehensive, accurate and authentic description of the language.

7 Some details of this data are given in http://book.culture.gr.

8 Cf. BIBER Douglas, 1993, "Representativeness in corpus design", *Literary and Linguistic Computing* 8, p. 1-15.