Πρόγραμμα «Αριστεία»
«Διαχρονικό σώμα ελληνικών κειμένων του 20ού αιώνα/
Diachronic corpus of Greek of the 20th century»
Κ.Α. Έρευνας: 70/3/11920
Ακρωνύμιο: Greek Corpus 20
Κωδικός «ΑΡΙΣΤΕΙΑ Ι»: 2396
Επιστημονικός υπεύθυνος: Διονύσης Γούτσος

Παραδοτέο 2.2: Πρακτικά διεθνούς εργαστηρίου

Επισυνάπτεται η έκθεση βασικών αρχών και ορθών πρακτικών όπως διαμορφώθηκε στις συναντήσεις της ημερίδας για τη δημιουργία και συγκρότηση διαχρονικών σωμάτων κειμένων, σε συνεργασία με τη διεθνή συμβουλευτική επιτροπή του προγράμματος.

Report on the design of Greek Corpus 20

This report is based on the discussion in a series of meetings that took place in the frame of the workshop on the compilation and analysis of diachronic corpora that was organized at the University of Athens in June 2013. We would like to thank Mark Davies (Brigham Young University), Susan M. Fitzmaurice (University of Sheffield), Terttu Nevalainen (VARIENG, University of Helsinki) and Sean Wallis (Survey of English Usage, University College London), who participated in the meetings, for their valuable comments and guidance.

1. Selection of material

There are five basic parameters as regards the selection of linguistic material for the compilation of Greek Corpus 20.

a) Time period:

The unit of data selection will be the decade. A basic principle is that texts are to be selected from each year in the decade. If not possible, an attempt will be made to select texts from as many years as possible through the decade. An alternative will be to select texts from the middle of the decade. Fallback solution: 20 years (generation).

The target number of words for Greek Corpus 20 is 20 million words that will be equally divided in the first nine decades of the 20th century, since the last decade of the 20th century is covered by the Corpus of Greek Texts (CGT), to which Greek Corpus 20 will be comparable. For reasons of comparability, CGT structure will be used to provide the basic structure for Greek Corpus 20 (see Appendix). This means that 2,200,000 words will be selected from each decade.

Table 1: Word targets for each decade of Greek Corpus 20

Mode	*Text type*	*Percentage*	*Number of words*
Spoken	News	1 %	22,000
	Interview	2 %	44,000
	Public speech	6 %	132,000
	Conversation	1 %	22,000
Written	Literature	9 %	198,000
	News	17 %	374,000
	Opinion articles	12 %	264,000
	Information items	1 %	22,000
	Academic	15 %	330,000
	Popularized	28 %	616,000
	Law and administration	5 %	110,000
	Private	1 %	22,000
	Procedural	1 %	22,000
	Miscellanea	1 %	22,000
Total			2,200,000

b) Text sampling:

Regarding the question of whether text samples or full texts are to be selected for compilation, previous projects have followed different approaches, including:

- random samples, with one or more samples from a text (*Corpus of English Dialogues*)
- samples which reflect the range of text sections/functions of the complete text as closely as possible (*Corpus of English Religious Prose*)
- full texts as far as they fitted to the predetermined number of words; otherwise, text samples (*Corpus of Early English Medical Writing*)
- full texts and texts of varying length (*The Small Corpus of Political Speeches*).

The data selected for Greek Corpus 20 will be full texts for those text types with relatively small text size, while for longer texts, there will be sampling to an upper word limit, which will depend on the principle of selecting texts from each year of the decade. Thus, smaller texts will be included in full, while larger texts will first be sampled so that an equal number of words will be selected for each year. For instance, in the text type of academic texts for each year in the decade 33,000 words will be sampled (see Table 1). In addition, for larger texts there will be a limit of 5,000 words, with a fallback solution of 10,000 words. This will mean that for academic texts 7 (or alternatively, 4) texts will be sampled for each year.

In addition, there will be upper word limits for the number of texts per author, title of magazine or newspaper or type of context (for spoken texts). In other projects, major literary figures have been excluded (*Lampeter Corpus*), each author appears only once, in order to avoid idiosyncratic language use (*Lampeter Corpus*) or there are never more than three texts selected by the same author (*Corpus of Late Modern English Texts*). The principle of selecting from a wide variety of well-known and less well-known authors will be followed in Greek Corpus 20. In addition, co-authored texts will be avoided, so that comparisons between authors will be easier to make.

A similar issue concerns text circulation and reception: in some projects, texts with a wider circulation and/or greater reception are chosen over such with lesser importance (e.g. number of editions of a text) or only “typical” or “important” texts for the genre are selected (*Corpus of English Religious Prose*). Keeping in mind this preference for typical texts and in order to enhance representativess, in Greek Corpus 20 there will be a limit in the number of popular novels: no more than three will be selected for each decade.

Our basic principle for text sampling is to avoid systematic bias by randomizing variables as much as possible.

c) Selection of editions:

In other projects (e.g. *Corpus of Early English Medical Writing, Corpus of English Dialogues, Lampeter Corpus*) the practice is to generally select first editions. In Greek Corpus 20 first editions will also be selected, with the exception of works extensively reworked and changed by the author to be published at a later date.

d) Sociolinguistic variables:

For spoken texts it is important in principle to sample according to sociolinguistic variables (dialect, gender and social class), where this is possible. For written texts, it is important to sample only works published in Greece. In addition, it will be attempted to sample texts from a large spread of geographical areas of provenance.

e) The question of diglossia:
It is a general principle of Greek Corpus 20 that texts will be selected on the basis of how representative they are of particular text types and *not* of their language variety. It is expected that the principle of sampling from different speakers/authors etc and from a wide variety of text sources will prevent bias towards a certain variety (demotic, katharevousa or a mix thereof). The question of diglossia will not constitute a sampling parameter but will be investigated in the stage of analysis, as an outcome of corpus compilation.

2. Availability and sources of data

The main problem of data availability concerns spoken texts, for which it is difficult to find adequate material from the past, since a) older recordings are scarce, b) recording quality is poor and c) there may be a lack of metadata. In addition, it is easier to find spoken data from official contexts (rehearsed data), while everyday conversations (unrehearsed data) were usually not recorded. Older spoken data can be found in published transcripts (e.g. parliament and court transcripts), which, however, do not provide many speech details. Spoken data can also be found in public TV and radio archives, or other public records including e.g. the text type of commentaries on sports events. Other possible sources are field recordings by historians or anthropologists on local and/or oral history.

Several alternatives can be developed to compensate for the lack of spoken data:

- further develop text types that resemble spoken language, like personal letters (either from private sources or edited sources, e.g. book versions) or film scripts.
- personal recordings or letters, as well as material such as recipes and remedies, can be found through appeals to the public (e.g. through facebook, twitter etc).

Furthermore, it is expected that, because of the peculiarities of Greek history in the 20th century, we may face problems at finding data for particular years or decades (e.g. World War II, Greek dictatorship etc).

Although our goal is to achieve comparability with the CGT, this may not be possible because of non-existent genres in the past (e.g. radio and TV) or new and emergent genres (e.g. electronic texts). A solution to this could be to create *micro-corpora* for text types found only in certain decades. These micro-corpora will be independently linked to the main Corpus and will not follow its sampling principles (e.g. they could be larger).

Another idea would be to expand the current list of text types and include different sub-genres under the same categories: for instance, school books could be added as a sub-genre in the category of popularized non-fiction or academic texts could be selected from individual disciplines. Thus, if data for an academic discipline e.g. biology could be found across the nine decades of our corpus, a micro-corpus of biology texts could be created.

Finally, the text type of literature could also be enlarged from 9% to 20%. The sub-genre of drama could also be enlarged, since it can be also relevant to spoken discourse.

3. Text processing, annotation and visualization

Text processing will involve digitization of written material. In the case of Greek Corpus 20 a major priority will be the development of polytonic OCR that will enable the digitization of older (pre-1980s) texts. A related question concerns spelling and the co-existence of standardized/normalized spelling and spelling variants. There should be one to one correspondence e.g. between monotonic and polytonic forms, while a further issue concerns the treatment of alternative forms e.g. εργαστήριο-εργαστήρι-εργαστήριον.

In terms of annotation, multiple versions, a clean version and an annotated version of the corpus, will be kept. Apart from structural mark-up, lemmatizing and POS tagging will be made at the stage of the pilot corpus (2 million words), while parsing and other annotation can be reserved for later. The standards to be followed will be those of CGT, when annotation for this has been completed.

A basic principle is that the corpus should be searchable from the start; thus the pilot corpus can be made available online.

Finally, a 3 month period of proofreading should be allowed.

With regard to visualization, tools such as open source CQP texts will be explored, while the interface should be dependent on the target audience and the potential corpus users.

APPENDIX

Corpus of Greek Texts: 1990-2010

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 items	100,570	0.36 %
	Academic	3,994,277	14.67 %
	Popularized	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 %

Projection of data for *Greek Corpus 20*

Target: 20 million words for the 9 first decades of the 20th century

2,200,000 words for each decade

Mode	*Text type*	*Percentage*	*Number of words*
Spoken	News	1 %	22,000
	Interview	2 %	44,000
	Public speech	6 %	132,000
	Conversation	1 %	22,000
Written	Literature	9 %	198,000
	News	17 %	374,000
	Opinion articles	12 %	264,000
	Information items	1 %	22,000
	Academic	15 %	330,000
	Popularized	28 %	616,000
	Law and administration	5 %	110,000
	Private	1 %	22,000
	Procedural	1 %	22,000
	Miscellanea	1 %	22,000

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης