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 10.2: Συγκεντρωτική τεχνική αναφορά

Επισυνάπτεται η τελική έκθεση για το σώμα κειμένων, όπως προβλέπεται στο Τεχνικό Παράρτημα του έργου.

Diachronic corpus of Greek of the 20th century (Greek Corpus 20)

Report on the final corpus

According to the initial design of the project *Diachronic corpus of Greek of the 20th century*, its aims have been:

1) to examine the issues involved in the compilation of a diachronic corpus of Greek of the 20th century,

2) on the basis of exploration of data sources, to collect data for a diachronic corpus of Greek of the 20th century that will be available to researchers, and

3) to analyze the corpus with a view to drawing basic conclusions on linguistic change across the decades of the 20th century.

This final report describes our findings with respect to these aims, on the basis of:

a) the description of the Greek Corpus 20 (see 1. Greek corpus 20 – Short presentation),

b) the preliminary report on the design of the corpus (see 2. Report on the design of Greek Corpus 20),

c) the pilot corpus data, (see description in 3. Pilot corpus data),

d) the report on the pilot corpus (see 3. Report on the pilot corpus),

e) the evaluation of the pilot corpus (see 4. Evaluation of the pilot corpus), and

f) the final corpus data (see description in 5. Final corpus data). (All related documents are attached).

1. Issues regarding the compilation of the corpus

The project, first, investigated the availability of data in different text types, the feasibility of collecting particular data categories and the possibility of collecting as much data as possible. The major problems concerning the collection of Greek data of the 20th century are:

a) the lack of fully functioning OCR processing facilities for polytonic Greek, the script that Greek was written in for most of the 20th century (up to 1982). We have developed our own tools by training the open source OCR engine Tesseract with Greek polytonic data and have created a platform, which will be freely available to researchers after the end of the project. However, extensive training is still needed for a fully satisfactory processing of polytonic texts and thus post-editing for several genres has been time-consuming with the effect that it was not possible to process more data. It is expected that, once this platform is available, further training on Greek polytonic data will be easier.

b) the lack of freely available archives. Specifically:

- Greek TV archives are publicly available but do not keep news data.
- Public radio archives are not publicly available.
- Parliament proceedings are only available online for 1900-1935 and the end of 1989 at the site of the Hellenic Parliament.
- Newspaper archives, especially those of the major newspapers that were published for most of the 20th century (e.g. *Kathimerini, Vima* etc.) have closed access and, despite our efforts to gain access, no progress has been made.
- Most importantly, archives that were made open access in the 1990s and 2000s mainly keep image rather than OCR-processed records with the effect that further processing is needed. A particularly bad example of this practice concerns the archives of the influential 20th century journal Nea Estia, which cannot be processed by any means, but can only be leafed through as if it were a hard copy. Another example concerns the newspaper archives of the National Library of Greece, which have been processed by a shallow OCR engine but for which fully OCR-processed files are missing.

This problem, obviously, mostly affects the major genres of journalistic texts, which constitute a large part of all modern Greek synchronic corpora (see Goutsos 2010). It, additionally, affects public records of spoken material, which is sadly underdeveloped for Greek.

c) the continuity of text types, i.e. the fact that several text types may only be found in certain decades. This is a well-known problem in the diachronic corpora literature (see e.g. Nevalainen & Raumolin-Brunberg 2003: 28), but particularly affects Greek 20th century data in major genres like popularized non-fiction. For instance, although there were several literary journals, no magazines on other subjects seem to be easily accessible for the whole of the 20th century. This is partly an effect of digitization policies, which have exclusively focused on literary journals, especially for the 19th century and the beginning of the 20th century (e.g. the University of Patras collections), but also reflects the fluid limits of general interest magazines for the first half of the 20th century, which mostly included literary contributions (see Karaoglou 2005).

At the same time, as expected, electronic media-related text types emerge quite late in Greece, with sound films and radio stations appearing in the 1930s and public TV in the 1960s. Full operation of these media was further delayed because of the effects of the Second World War in the 1940s and the military dictatorship of the 1970s.

Taking into account these problems and based on our experience from the pilot corpus, we have decided to follow a double strategy consisting in concentrating on a subset of genres to be fully processed and integrated in the final corpus data, while for the other genres it was decided to collect as much data as possible with a view to processing them and including them in the corpus in the future.

Specifically, it was decided to concentrate on the genres of *Spoken News, Public Speech* and *Conversation*, as regards the spoken mode, and *Literature, Academic, Popularized Non-Fiction* and *Private*, as regards the written mode, for full processing.

Data were collected but has not been fully processed and has not been integrated in the final corpus for the genres of *Interview*, as regards the spoken mode, and *News, Opinion Articles, Information Items* and *Procedural*, as regards the written mode. (Sources for these data include, among else, the National Library of Greece, with which there has been an agreement for data sharing, the Greek Parliament Library collections, mainly for newspapers up to the 1930s and other private collections).

This decision accords well with the trend noted by Nevalainen & Raumolin-Brunberg (2003: 27) "from textually balanced multi-purpose corpora towards larger single-genre corpora", although in our case it also follows the idea of developing *micro-corpora* for text types found only in certain decades, developed in the design phase of the project. It must also be noted that the genres that were fully processed and integrated in the final corpus include text types that give emphasis on *speech-like* (private letters), *speech-based* (public speeches) and *speech-purposed* (films, drama, newsreels) text types (cf. Culpeper & Kytö 2010). In this sense, the final corpus is oriented towards data that are more likely to reveal actual speaking patterns of the past.

Further details on the particularities of genres and the processes of corpus compilation are given in the Report on the pilot corpus.

2. Data collection for Greek Corpus 20

Detailed figures for the number of words collected for each text types and decade are given in the file on Final corpus data. The following table summarizes these data. (Please note that data for *Law and administration* have not been finally integrated and for this reason they do not appear in the file on Final corpus data).

Genre	*Text types*	*Text types codes*	*Number of words integrated*
Spoken news	Newsreels	SRF01	78,441
Public speeches	Parliament	STL16	339,194
	Academic	SAL06	
	Other	SOL16	
Conversation	Film scripts	SFF19	208,207
Literature	Novels	WFB08	1,086,687
	Short stories	WFB09	
	Poetry	WFC11	
	Drama	WFB12	
Academic	Humanities	WAB13	1,044,200
	Social/Finance	WAB14	
	Science	WAB15	
Popularized Non-fiction		WLB13 WLB14	285,252
Law and administration		WDC34 WDC35	120,000
Private	Letters	WPO26	178,720
Miscellanea		WMO99	1,136
Total			3,341,837

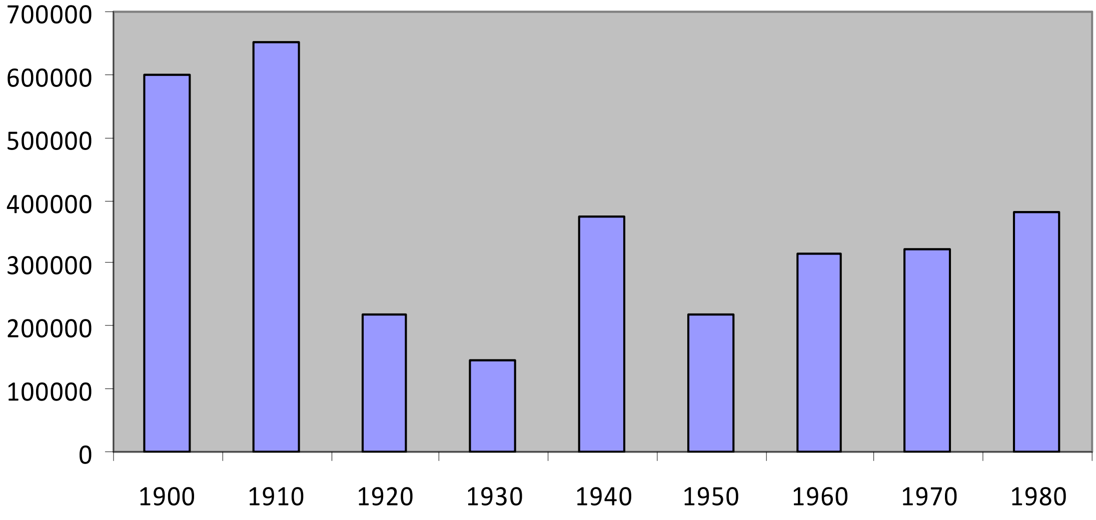
Current contents of Greek Corpus 20

The total number of words comprising the pilot corpus is **3,341,837**, which roughly corresponds to 20% of the target for Greek Corpus 20.

It is expected that approx. 500,000 more words will be further integrated in existing categories in the near future, including the text type of *Lyrics* in the genre of *Literature*, for which data has been gathered for all the nine decades of the 20th century concerned. This will bring the corpus closer to the 25% of the initial target.

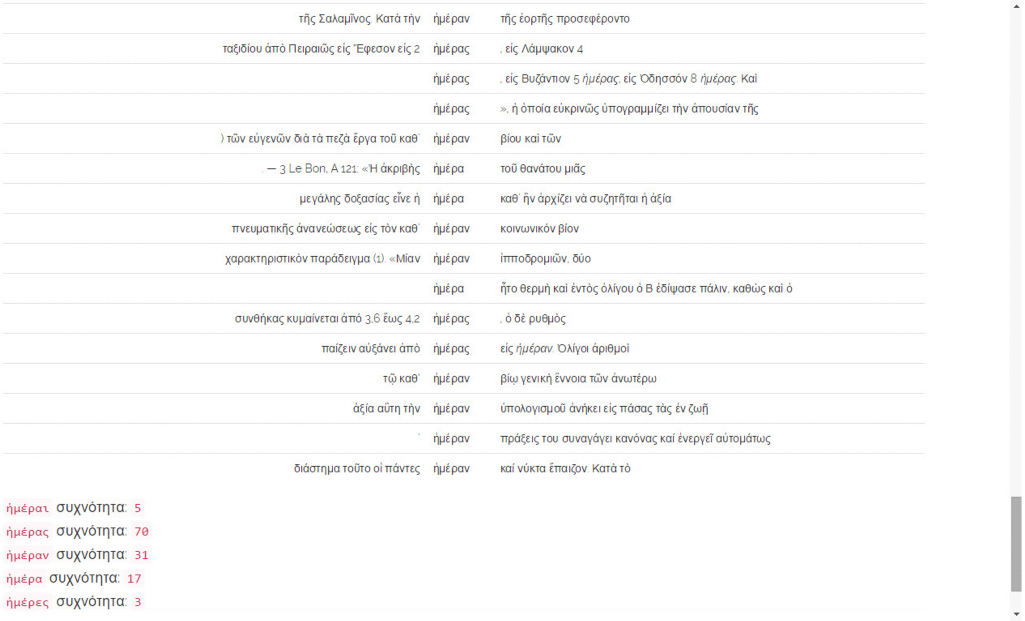
It is also estimated that the data collected for the genres that have not been integrated in the final corpus amount to more than 15 million words, although it is hard to be accurate with non-OCR processed texts.

The following figure presents the distribution of the data that has been integrated across the nine decades of the 20th century. As can be seen in the figure surprisingly enough, more data have been integrated for the first two decades of the 20th century, while there is a slight increase from the 1950s onward. This may well reflect the availability of data and brings attention to the fact that more effort is needed to collect data from the 1930s and the 1950s.



Current distribution of data in Greek Corpus 20

As regards access to data, the web interface is currently being completed. The following is a snapshot from a search, showing full functionality with polytonic Greek:



τῆς Σαλαμῖνος. Κατὰ τὴν	ἡμέραν	τῆς ἑορτῆς προσεφέροντο
ταξιδίου ἀπὸ Πειραιῶς εἰς Ἔφεσον εἰς 2	ἡμέρας	, εἰς Λάμψακον 4
	ἡμέρας	, εἰς Βυζάντιον 5 *ἡμέρας*, εἰς Ὀδησσόν 8 *ἡμέρας*. Καὶ
	ἡμέρας	», ἡ ὁποία εὐκρινῶς ὑπογραμμίζει τὴν ἀπουσίαν τῆς
) τῶν εὐγενῶν διὰ τὰ πεζὰ ἔργα τοῦ καθ᾽	ἡμέραν	βίου καὶ τῶν
. — 3 Le Bon, A 121: «Ἡ ἀκριβὴς	ἡμέρα	τοῦ θανάτου μιᾶς
μεγάλης δοξασίας εἶνε ἡ	ἡμέρα	καθ᾽ ἣν ἀρχίζει νὰ συζητῆται ἡ ἀξία
πνευματικῆς ἀνανεώσεως εἰς τὸν καθ᾽	ἡμέραν	κοινωνικὸν βίον
χαρακτηριστικὸν παράδειγμα (1). «Μίαν	ἡμέραν	ἱπποδρομιῶν, δύο
	ἡμέρα	ἦτο θερμὴ καὶ ἐντὸς ὀλίγου ὁ Β ἐδίψασε πάλιν, καθὼς καὶ ὁ
συνθήκας κυμαίνεται ἀπὸ 3,6 ἕως 4,2	ἡμέρας	, ὁ δὲ ρυθμὸς
παίζειν αὐξάνει ἀπὸ	ἡμέρας	εἰς *ἡμέραν*. Ὀλίγοι ἀριθμοὶ
τῷ καθ᾽	ἡμέραν	βίῳ γενικὴ ἔννοια τῶν ἀνωτέρω
ἀξία αὕτη τὴν	ἡμέραν	ὑπολογισμοῦ ἀνήκει εἰς πάσας τὰς ἐν ζωῇ
᾽	ἡμέραν	πράξεις του συναγάγει κανόνας καί ἐνεργεῖ αὐτομάτως
διάστημα τοῦτο οἱ πάντες	ἡμέραν	καί νύκτα ἔπαιζον. Κατὰ τὸ

ἡμέραι συχνότητα: 5
ἡμέρας συχνότητα: 70
ἡμέραν συχνότητα: 31
ἡμέρα συχνότητα: 17
ἡμέρες συχνότητα: 3

The project's deliverables also include:

- the project's webpage
- the compilation of a bibliography on diachronic corpora
- the compilation of an inventory of research projects of diachronic corpora
- the proceedings of a workshop on the compilation and analysis of diachronic corpora.

3. Findings on linguistic change in Greek of the 20th century

Our preliminary findings have been reported in Goutsos & Fragaki (2014) and Fragaki & Goutsos (2015). (The Powerpoint presentation of the latter is attached to this document). These findings can be summarized in the following.

First, data analysis supports a variationist view of language change on the basis of the thoroughly attested role of frequency (Schneider 2004); specifically, demotic (or Low) variants in Greek diglossia show a U-curve, rather than the expected S-curve of sociolinguistic variation, whereas katharevousa (or High) variants show a “roller-coaster” pattern that is indicative of their stereotypical (in Labov’s sense) or emblematic use.

Secondly, recent language change in Greek largely depends on genre; specifically, in film scripts and literature there is steady preference for Low variants across the century. By contrast, in academic texts and public speeches High variants are preferred in most decades before the 1960s, when there is a sudden rise of Low variants. Newsreels show a haphazard pattern, conforming to the expected rise of Low variants only after the 1960s, whereas private letters are the only genre in which the expected gradual rise of Low variants across all decades is found (cf. Dossena & Del Lungo Camiciotti 2012 on private letters and Taavitsainen et al. 2015 on the role of genre).

Third, the study of sociolinguistic phenomena such as the Greek diglossia is greatly helped by diachronic corpora which give access to evidence about what actual people said and wrote (language use) rather than what they believed (language attitudes). The analysis of data from Greek Corpus 20 can provide secure indications about the questions surrounding Greek diglossia, by clarifying e.g. whether it is related to the spoken vs. written dichotomy, by identifying when changes took place and by establishing how public attitudes influence the private use of language.

Finally, a full-scale investigation of variants is also expected to contribute to an informed view on standardisation and a better understanding of the range of registers and language varieties involved in Greek of the 20th century.

References

Culpeper, J. & Kytö, M. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Dossena, M. & Del Lungo Camiciotti. G. 2012. *Letter Writing in Late Modern Europe*. Amsterdam/Philadelphia: Benjamins.

Fragaki, G. & Goutsos, D. 2015. Greek diglossia in the 20th century: A historical corpus linguistics approach. Presented at the 12th International Conference on Greek Linguistics (ICGL12), Berlin.

Goutsos, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 5 (1), 29-44.

Goutsos, D. & Fragaki, G. 2014. Recent language change in Greek: Design of the Diachronic Corpus of Greek of the 20th century. In Kotzoglou, G. et al. (eds) *Selected Papers of the 11th International Conference on Greek Linguistics*. Rhodes: University of the Aegean, 318-329. [In Greek]

Karaoglou 2005. Special magazines and their classification. In Droulia, L. (ed.) *La presse Grecque de 1784 à nos jours. Approches historiques et théoriques. Actes du Colloque International, Athènes, 23-25 mai 2002*. Athens: Insitute of Modern Greek Studies, 263-275. [In Greek]

Nevalainen, T. & Raumolin-Brunberg, H. 2003. Historical Sociolinguistics: Language Change in Tudor and Stuart England. London: Routledge.

Schneider, E. W. (2004). Investigating variation and change in written documents. In Chambers, J. K., Trudgill, P. & Schilling-Estes, N. (eds) *The Handbook of Language Variation and Change*. Oxford: Blackwell, 67–96.

Taavitsainen, I., Kytö, M., Claridge, C. & Smith, J. (2015). *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press.

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