THE DIACHRONIC CORPUS OF GREEK OF THE 20TH CENTURY:
DESIGN AND COMPILATION
Dionysis Goutsos, Georgia Fragaki, Irene Florou,
Vasiliki Kakousi & Paraskevi Savvidou
National and Kapodistrian University of Athens
dgoutsos@phil.uoa.gr, efraga@phil.uoa.gr, eirini.florou@gmail.com, vaswk1202@hotmail.com & psavvidou@phil.uoa.gr

Περίληψη

Στο άρθρο παρουσιάζεται το Διαχρονικό Σώμα Ελληνικών Κειμένων του 20ού αιώνα, το πρώτο διαχρονικό σώμα κειμένων της ελληνικής, που έχει σχεδιαστεί για τη μελέτη της πρόσφατης γλωσσικής αλλαγής στα ελληνικά. Ειδικότερα, παρουσιάζονται ζητήματα που αφορούν τη συλλογή γλωσσικών δεδομένων του εικοστού αιώνα στα ελληνικά, η σύνθεση του σώματος κειμένων (γένη, είδη, αριθμός λέξεων κ.λπ.), τα παραδοτέα του ερευνητικού προγράμματος που οδήγησε στη δημιουργία του σώματος κειμένων, καθώς και ορισμένα προκαταρκτικά ευρήματα από την ανάλυσή του.

Keywords: corpus design and compilation, diachronic corpus, recent language change

1. Diachronic corpora and Greek

Corpus linguistics has considerably improved the description of languages by allowing access to large bodies of authentic texts, as well as by contributing to a broad range of applications in lexicography, the writing of grammars, lexical semantics, language teaching, the study of language and ideology, translation, media studies etc. (see, among else, Hunston 2002: 13-14, Meyer 2002: 1-29, Baker et al. 2006). Unlike other
languages, Greek has only benefited to a small extent by the development of this field, mainly because of the lack of large Greek corpora, with the exception of the Hellenic National Corpus (HNC, 47 million words, texts published from 1976 to 2007) and the Corpus of Greek Texts (CGT, 30 million words, texts from 1990 to 2010). Both can be characterized as synchronic corpora, in the sense that they offer a view of a specific period of the Greek language.

This paper presents the design and compilation of the Diachronic Corpus of Greek of the 20th century ((Greek Corpus 20 or GC20)), the first diachronic corpus of Greek, developed with a view of studying recent language change. Its goal is to gather 20 million words from Greek texts coming from the first nine decades of the 20th century, to be integrated with the existing 30 million word CGT, which includes texts from the 1990s onwards.

Historical or diachronic corpora have been compiled or are under preparation for other languages or language varieties like the Helsinki Corpus of English Texts, which covers Old, Middle and Early Modern English, the Corpus of Historical English Registers (ARCHER), which contains British and American English texts from 1650 to the present, the four corpora including Brown and Frown, LOB and FLOB, which can together supply evidence for change in the two varieties of English between 1961 and 1991-1992, DiaCoris for Italian etc. In practice, three types of corpora have been used to study recent language change in most languages:

a) diachronic corpora, e.g. the Corpus of Historical American English (COHA, with data from 1810 to 2009),

b) corpus families, e.g. the Lancaster-Oslo-Bergen Corpus, with data from every 30 years in the 20th century, including BLOB-1931 (1928-1934), LOB (1961) and F-LOB (1991),

c) synchronic, monitor corpora, e.g. the British National Corpus (BNC), including data from 1960 up to now and thus offering a large time span of linguistic material.

For more details, see Hatzigeorgiu et al. (2001) for HNC and Goutsos (2010) for CGT.
2 For a definition of recent language change, see Mair (2009: 1120), Davies (2011, 2012).
3 For more details on existing diachronic corpora, see Onelli et al. (2006), Beal et al. (2007), Mair (2009), Baker (2010: 57 ff.), Partington (2010), Aarts et al. (2013).
Diachronic corpora are also of different sizes, from the very big (e.g. *Corpus Diacrónico del Español* with 125 million words), to big (e.g. *O corpus do Português*, with 45 million words), medium-sized (e.g. *Diachronic Czech National Corpus* with 2 million words, *Helsinki Corpus of English Texts* with 1.5 million words) and small (e.g. *Sheffield Corpus of Chinese*, with 18,000 words).

It is also important to notice that a wide range of linguistic phenomena have been studied in diachronic corpora, including vocabulary changes (Baker 2011), grammatical change (Leech et al. 2009), diachronic morphological processes (Baayen & Renouf 1996, Fischer 1998, Duguid 2010), development of phraseology (Davies 2012) and cultural changes (Baker 2010, Marchi 2010, Partington 2012).

Greek has not had a similar diachronic corpus for a number of reasons, among which, as we will discuss below, the difficulty of collecting data is surely prominent. Extra-linguistic factors, such as the socio-historical background in Greece of the 20th century, can account for the lack of data or the occurrence of minimal data for several periods. In addition, linguistic factors such as the persisting diglossia, which is related with important socio-historical events throughout the 20th century, complicate issues of data collection and analysis. For this reason, the development of a diachronic corpus for Greek of the 20th century has been more than imperative.

The research project for the development of GC20 has had the following aims:

4 The project was funded in the frame of the action “Aristeia I” (“Excellence I”) by the European Cohesion Fund and the Greek government (General Secretariat for Research and Technology). We are grateful to the project’s international advisory board, namely Claudia Claridge (University of Duisburg-Essen), Mark Davies (Brigham Young University), Hendrik De Smet (KU Leuven), Susan M. Fitzmaurice (University of Sheffield), Marianne Hundt (University of Zurich), Christian Mair (University of Freiburg), Terttu Nevalainen (University of Helsinki), Fabio Tamburini (Università di Bologna) and Sean Wallis (University College London), for their help in the various stages of the project.
With respect to these aims, in what follows we first outline the most important issues concerning the compilation of the corpus, we then present the data collected in the corpus and, finally, we discuss the project’s deliverables and some preliminary findings.

2. 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. A major problem concerning the collection of Greek data of the 20th century concerns, first of all, the lack of fully functioning OCR processing facilities for polytonic Greek, the script in which Greek was written for most of the 20th century (specifically, up to 1982). We have developed our own tools by training the open source OCR engine Tesseract\(^5\) 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.

Furthermore, the lack of freely available archives for many Greek genres has been a serious obstacle in data collection. Specifically, Greek TV archives, after a two year period of sudden closure (2013-2015) have become publicly available again, but do not keep news data. In addition, public radio archives are not publicly available. Parliament proceedings are only available online at the site of the Hellenic Parliament for 1900-1935 and from the end of 1989 onwards, leaving thus out five decades of the 20th century. Newspaper archives, especially those of major newspapers that were published for most of the 20th century (e.g. *Kathimerini, Vima*) have limited or no 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 notorious example of this practice concerns the online archives of the influential 20th century journal *Nea Estia*, which cannot be processed by any means, but can only be leafed through like a hard copy. Another example concerns

---

\(^5\) The software is available at: https://github.com/tesseract-ocr/tesseract
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 mostly affects genres of journalistic texts, which constitute a large part of modern Greek synchronic corpora (Goutsos 2010), as well as public records of spoken material, which are sadly underdeveloped for Greek.

A third major issue has to do with 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) and particularly affects Greek 20th century data in major genres like popularized non-fiction texts. For instance, although there have been several literary journals in the 20th century, no magazines on other subjects seem to be easily accessible for the whole of this 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 (Karaoglou 2005).

At the same time, electronic media-related text types emerged 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 a pilot version, we 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 and including them in the future. Specifically, for full processing it was decided to focus 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. Data were collected but have not been fully processed and integrated for the genres of Interview, as regards the spoken mode, and News, Opinion Articles, Information Items and Procedural, as regards the written mode. This decision accords well with the trend noted by Nevalainen & Raumolin-Brunberg (2003: 27) to move

---

6 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.
“from textually balanced multi-purpose corpora towards larger single-genre corpora”, although in our case it is based on the idea of developing micro-corpora for text types found only in certain decades as part of the initial corpus design. It must also be noted that the text types that were fully processed and integrated in the final corpus 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.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Genres</th>
<th>Text types</th>
<th>Codes</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>Spoken news</td>
<td>Newsreels</td>
<td>SRF01</td>
<td>78,441</td>
</tr>
<tr>
<td></td>
<td>Public speeches</td>
<td>Parliament</td>
<td>STL16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Academic</td>
<td>SAL06</td>
<td>339,194</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>SOL16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conversation</td>
<td>Film scripts</td>
<td>SFF19</td>
<td>208,207</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>Novels</td>
<td>WFB08</td>
<td>1,355,629</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Short stories</td>
<td>WFB09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poetry</td>
<td>WFC11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Drama</td>
<td>WFB12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Academic</td>
<td>Humanities</td>
<td>WAB13</td>
<td>1,044,200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Social/Finance</td>
<td>WAB14</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Science</td>
<td>WAB15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Popularized</td>
<td>Non-fiction</td>
<td>WLB13</td>
<td>285,252</td>
</tr>
<tr>
<td></td>
<td>Non-fiction</td>
<td>Letters</td>
<td>WLB14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Law and</td>
<td></td>
<td>WDC34</td>
<td>265,924</td>
</tr>
<tr>
<td></td>
<td>administration</td>
<td></td>
<td>WDC35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>Letters</td>
<td>WPO26</td>
<td>188,856</td>
</tr>
<tr>
<td></td>
<td>Miscellanea</td>
<td></td>
<td>WMO99</td>
<td>1,136</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>3,766,839</strong></td>
</tr>
</tbody>
</table>

Table 1 | Composition of Greek Corpus 20 (data integrated in May 2016)
3. Corpus composition

Table 1 presents in detail the number of words integrated so far for each genre and text type in all decades covered in GC20, as of May 2016. The total number of words integrated in the corpus so far is 3,766,839, which roughly corresponds to 20% of the target for GC20. It is estimated that the data collected for the genres that have not been integrated in the final corpus amount to more than 15 million words, a figure which covers the remaining percentage of the corpus target, although it is hard to be accurate with non-OCR processed texts. In all, the divergence from the projected target is indicative of the problems related with corpus compilation pointed out in the previous section.

Figure 1 presents the distribution of the data that have been integrated in GC20 across the nine decades of the 20th century. 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 reflect the availability of existing data, as most projects collected data have concentrated on the 19th and the beginning of the 20th century, in particular with respect to literary texts and journals, which have been thought to be of special value. (Copyright restrictions also account for the less easy access to more recent material). It is also clear that more effort is needed to collect data from the 1920s, 1930s and the 1950s.

![Figure 1](image-url)
4. Deliverables and preliminary findings

The project’s deliverables include, among else, the compilation of a bibliography on diachronic corpora and an inventory of research projects of diachronic corpora. Both are extensive and offer an updated picture of research conducted in a large variety of languages on issues of language change with the help of diachronic corpora. This is necessary background material for anyone attempting a state-of-the-art description of diachronic corpus research. They also include the oral presentations of an international workshop on the compilation and analysis of diachronic corpora, available online, the various reports, evaluations and publications that were made in the successive stages of the corpus compilation and the project’s webpage, which gives access to the corpus itself.7

The corpus webpage gives access to all data that has been fully processed, while more texts are constantly being added. The total number of words (tokens) and of their unique occurrence (types) included in the corpus appears in each search. In search one can type in the word or phrase (up to 4 words) they are interested in, using any tone or other diacritic, in order to take all possible versions of the word form occurring in the corpus. For instance, all relevant word forms will appear if you type in ημέρα or ημέρα or even ημέρα. One can also search for part of a word (but not phrase) using wildcards; for instance, the search ημ*ρα will give all word forms for ημέρα and ημέτερα and the search ημ?ρα all word forms for ημέρα. Results can be sorted according to the node word or phrase, the previous or the next word, the text type in which word forms appear and the data of the texts in which they occur, in ascending or descending order.

Figure 2 presents a screen from a corpus query result, whereas Figure 3 that follows on the second next page illustrates further search statistics provided on the webpage about the frequency development of the words found for the query throughout the nine decades, within a decade etc.

Our preliminary findings from the analysis of the corpus have been reported in Goutsos & Fragaki (2014) and Fragaki & Goutsos (2015) and suggest exciting prospects for the analysis of recent language change in Greek. Thus, an analysis of grammatical words of Greek at the top of the corpus word frequency list, such as διά vs. γιά/για ‘for’ or εἰς vs. σέ/σε ‘in/at’ can be revealing of the complex patterns of Greek diglossia in the 20th century. Data analysis supports a variationist view of language

7 The corpus is freely available at: http://greekcorpus20.sek.edu.gr/
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. A full-scale investigation of variants like these is expected to contribute to an informed view on standardisation and a better understanding of what happened in the Greek of the 20th century.

Secondly, corpus data support the general principle that recent language change in Greek largely depends on genre (see e.g. Taavitsainen et al. 2015). For example, 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. This latter finding underlines the importance of collecting and analyzing private letters in understanding recent language change (cf. Dossena & Del Lungo Camiciotti 2012).

Figure 3 | Greek Corpus 20 query statistics
It is clear that the study of sociolinguistic phenomena such as the Greek diglossia will be greatly helped by diachronic corpora such as GC20, 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 the Diachronic Corpus of Greek of the 20th century 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.

More generally, it is expected that GC20 will offer an invaluable resource for the study of Greek language and culture, providing a point of reference for diachronic research in the still limited spectrum of Greek corpora. Since GC20 was designed to complement the synchronic CGT (Goutsos 2010), the range of available material for Greek is greatly extended. Future perspectives include both the integration of further genres and texts and the morphosyntactic annotation of the corpus, something that has not been attempted before for polytonic Modern Greek.

References


Basingstoke: Palgrave Macmillan.


Thanasis Georgakopoulos, Theodossia-Soula Pavlidou, Miltos Pechlivanos, Artemis Alexiadou, Jannis Androutsopoulos, Alexis Kalokairinos, Stavros Skopeteas, Katerina Stathi (Eds.)

PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON GREEK LINGUISTICS

ΠΡΑΚΤΙΚΑ ΤΟΥ 12ΟΥ ΣΥΝΕΔΡΙΟΥ ΕΛΛΗΝΙΚΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

VOL. 1