Diachronic Corpus of Greek of the 20th century: Design and implementation

Dionysis Goutsos, Georgia Fragaki, Irene Florou, Vasiliki Kakousi & Paraskevi Savvidou
National and Kapodistrian University of Athens

The project
This poster presents the research project “Diachronic Corpus of Greek of the 20th century”, involving the compilation and analysis of a corpus of Greek of the 20th century (Greek Corpus 20). The project is funded in the frame of the action ARISTEA I.

International advisory board:
• 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
• Sean Wallis, University College London

Aims of the project
a) to examine the issues involved in the compilation of a diachronic corpus of Greek of the 20th century, including:
   - the availability of data across decades,
   - the availability and continuity of text types
   - the issue of representativeness;
b) on the basis of exploration of data sources, to collect data for a diachronic corpus of Greek of the 20th century;
c) to analyze the corpus with a view to drawing basic conclusions on linguistic change across the decades of the 20th century.

Background
Three types of corpora used to study recent language change in other languages:
• diachronic corpora e.g. Corpus of Historical American English (COHA) (data from 1810 to 2009),
• corpus “families” e.g. Lancaster-Oslo-Bergen Corpus, including BLOB-1931 (1928-1934), LOB (1961) and F-LOB (1991) with data from every 30 years in the 20th century
• synchronic, monitor corpora e.g. BNC (1960-now)

Corpora of different size:
• Very big (Corpus Diacrónico del Español: 125 m),
• Big (O corpus do Português: 45 m),
• Medium-sized (Diachronic Czech National Corpus: 2 m, Helsinki Corpus of English Texts: 1.5 m),
• Small: (Sheffield Corpus of Chinese: 18.000).

Phenomena studied:
• vocabulary changes (Baker 2011)
• grammatical change (Leech et al. 2009)
• morphology (Baayen & Renouf 1996, Fischer 1998, Duguid 2010)
• phraseology (Davies 2012)
• cultural changes (Baker 2010, Marchi 2010, Partington 2012).

Issues of compilation
• Availability of data/text types:
  - spoken data > development of micro-corpora
  (cf. Culpeper & Kyto 2010):
speech-like: private letters
speech-based: Parliament speeches, newsreels
speech-purposed: theatrical plays, film scripts
• written data > development of polytonic OCR for Greek 20th century texts

Representativeness:
• Based on the structure of CGT (1990-2010)
• texts are selected on the basis of what representative they are of particular text types and not of their language variety. 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.

Greek Corpus 20

Aims of the project
a) to examine the issues involved in the compilation of a diachronic corpus of Greek of the 20th century, including:
   - the availability of data across decades,
   - the availability and continuity of text types
   - the issue of representativeness;
b) on the basis of exploration of data sources, to collect data for a diachronic corpus of Greek of the 20th century;
c) to analyze the corpus with a view to drawing basic conclusions on linguistic change across the decades of the 20th century.

Background
Three types of corpora used to study recent language change in other languages:
• diachronic corpora e.g. Corpus of Historical American English (COHA) (data from 1810 to 2009),
• corpus “families” e.g. Lancaster-Oslo-Bergen Corpus, including BLOB-1931 (1928-1934), LOB (1961) and F-LOB (1991) with data from every 30 years in the 20th century
• synchronic, monitor corpora e.g. BNC (1960-now)

Corpora of different size:
• Very big (Corpus Diacrónico del Español: 125 m),
• Big (O corpus do Português: 45 m),
• Medium-sized (Diachronic Czech National Corpus: 2 m, Helsinki Corpus of English Texts: 1.5 m),
• Small: (Sheffield Corpus of Chinese: 18.000).

Phenomena studied:
• vocabulary changes (Baker 2011)
• grammatical change (Leech et al. 2009)
• morphology (Baayen & Renouf 1996, Fischer 1998, Duguid 2010)
• phraseology (Davies 2012)
• cultural changes (Baker 2010, Marchi 2010, Partington 2012).

Issues of compilation
• Availability of data/text types:
  - spoken data > development of micro-corpora
  (cf. Culpeper & Kyto 2010):
speech-like: private letters
speech-based: Parliament speeches, newsreels
speech-purposed: theatrical plays, film scripts
• written data > development of polytonic OCR for Greek 20th century texts

Representativeness:
• Based on the structure of CGT (1990-2010)
• texts are selected on the basis of what representative they are of particular text types and not of their language variety. 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.

Aims of the project
a) to examine the issues involved in the compilation of a diachronic corpus of Greek of the 20th century, including:
   - the availability of data across decades,
   - the availability and continuity of text types
   - the issue of representativeness;
b) on the basis of exploration of data sources, to collect data for a diachronic corpus of Greek of the 20th century;
c) to analyze the corpus with a view to drawing basic conclusions on linguistic change across the decades of the 20th century.

Background
Three types of corpora used to study recent language change in other languages:
• diachronic corpora e.g. Corpus of Historical American English (COHA) (data from 1810 to 2009),
• corpus “families” e.g. Lancaster-Oslo-Bergen Corpus, including BLOB-1931 (1928-1934), LOB (1961) and F-LOB (1991) with data from every 30 years in the 20th century
• synchronic, monitor corpora e.g. BNC (1960-now)

Corpora of different size:
• Very big (Corpus Diacrónico del Español: 125 m),
• Big (O corpus do Português: 45 m),
• Medium-sized (Diachronic Czech National Corpus: 2 m, Helsinki Corpus of English Texts: 1.5 m),
• Small: (Sheffield Corpus of Chinese: 18.000).

Phenomena studied:
• vocabulary changes (Baker 2011)
• grammatical change (Leech et al. 2009)
• morphology (Baayen & Renouf 1996, Fischer 1998, Duguid 2010)
• phraseology (Davies 2012)
• cultural changes (Baker 2010, Marchi 2010, Partington 2012).

Issues of compilation
• Availability of data/text types:
  - spoken data > development of micro-corpora
  (cf. Culpeper & Kyto 2010):
speech-like: private letters
speech-based: Parliament speeches, newsreels
speech-purposed: theatrical plays, film scripts
• written data > development of polytonic OCR for Greek 20th century texts

Representativeness:
• Based on the structure of CGT (1990-2010)
• texts are selected on the basis of what representative they are of particular text types and not of their language variety. 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.

Deliverables
• Workshop on the compilation and analysis of diachronic corpora
  http://www.blod.gr/lectures/Pages/viewevent.aspx?EventID=259
• Bibliography on diachronic corpora
  Research projects of diachronic corpora
  http://en.greekcorpus20.phil.uoa.gr/publications

Publications

Papers

Contact
Dionysis Goutsos
Department of Linguistics, University of Athens
e-mail: dgoutsos@phil.uoa.gr

Project webpage
http://greekcorpus20.phil.uoa.gr/