FACHLITERATUR
EDITION ROMIOSINI
ΕΠΙΣΤΗΜΗ

ICGL12

12th INTERNATIONAL CONFERENC
ON GREEK LINGUISTICS
16 – 19 SEPTEMBER 2015
FREIE UNIVERSITÄT BERLIN, CEMOG

Proceedings of the ICGL12

vol. 1

The International Conference on Greek Linguistics is a biennial meeting on the study and analysis of Greek (Ancient, Medieval and Modern), placing particular emphasis on the later stages of the language.

PROCEEDINGS OF THE ICGL12
ΠΡΑΚΤΙΚΑ ΤΟΥ ICGL12

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

EDITION ROMIOSINI

Vertrieb und Gesamtherstellung: Epubli (www.epubli.de)
Satz und Layout: Rea Papamichail / Center für Digitale Systeme, Freie Universität Berlin
Gesetzt aus Minion Pro
Umschlaggestaltung: Thanasis Georgiou, Yorgos Konstantinou
Umschlagillustration: Yorgos Konstantinou

ISBN 978-3-946142-34-8
Printed in Germany

Online-Bibliothek der Edition Romiosini:
www.edition-romiosini.de

Στη μνήμη του Gaberell Drachman (†10.9.2014)
και της Αγγελικής Μαλικούτη-Drachman (†4.5.2015)
για την τεράστια προσφορά τους στην ελληνική γλωσσολογία
και την αγάπη τους για την ελληνική γλώσσα

ΣΗΜΕΙΩΜΑ ΕΚΔΟΤΩΝ

Το 12ο Διεθνές Συνέδριο Ελληνικής Γλωσσολογίας (International Conference on Greek Linguistics/ICGL12) πραγματοποιήθηκε στο Κέντρο Νέου Ελληνισμού του Ελεύθερου Πανεπιστημίου του Βερολίνου (Centrum Modernes Griechenland, Freie Universität Berlin) στις 16-19 Σεπτεμβρίου 2015 με τη συμμετοχή περίπου τετρακοσίων συνέδρων απ᾽ όλον τον κόσμο.

Την Επιστημονική Επιτροπή του ICGL12 στελέχωσαν οι Θανάσης Γεωργακόπουλος, Θεοδοσία-Σούλα Παυλίδου, Μίλτος Πεχλιβάνος, Άρτεμις Αλεξιάδου, Δώρα Αλεξοπούλου, Γιάννης Ανδρουτσόπουλος, Αμαλία Αρβανίτη, Σταύρος Ασημακόπουλος, Αλεξάνδρα Γεωργακοπούλου, Κλεάνθης Γκρώμαν, Σαβίνα Ιατρίδου, Mark Janse, Brian Joseph, Αλέξης Καλοκαιρινός, Ναπολέων Κάτσος, Ευαγγελία Κορδώνη, Αμαλία Μόζερ, Ελένη Μπουτουλούση, Κική Νικηφορίδου, Αγγελική Ράλλη, Άννα Ρούσσου, Αθηνά Σιούπη, Σταύρος Σκοπετέας, Κατερίνα Στάθη, Μελίτα Σταύρου, Αρχόντω Τερζή, Νίνα Τοπιντζή, Ιάνθη Τσιμπλή και Σταυρούλα Τσιπλάκου.

Την Οργανωτική Επιτροπή του ICGL12 στελέχωσαν οι Θανάσης Γεωργακόπουλος, Αλέξης Καλοκαιρινός, Κώστας Κοσμάς, Θεοδοσία-Σούλα Παυλίδου και Μίλτος Πεχλιβάνος.

Οι δύο τόμοι των πρακτικών του συνεδρίου είναι προϊόν της εργασίας της Εκδοτικής Επιτροπής στην οποία συμμετείχαν οι Θανάσης Γεωργακόπουλος, Θεοδοσία-Σούλα Παυλίδου, Μίλτος Πεχλιβάνος, Άρτεμις Αλεξιάδου, Γιάννης Ανδρουτσόπουλος, Αλέξης Καλοκαιρινός, Σταύρος Σκοπετέας και Κατερίνα Στάθη.

Παρότι στο συνέδριο οι ανακοινώσεις είχαν ταξινομηθεί σύμφωνα με θεματικούς άξονες, τα κείμενα των ανακοινώσεων παρατίθενται σε αλφαβητική σειρά, σύμφωνα με το λατινικό αλφάβητο· εξαίρεση αποτελούν οι εναρκτήριες ομιλίες, οι οποίες βρίσκονται στην αρχή του πρώτου τόμου.

Η Οργανωτική Επιτροπή του ICGL12

ΠΕΡΙΕΧΟΜΕΝΑ

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).[1] 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.[2] 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.[3] 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.

1 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]

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, and 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.

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.[6] 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.

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

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.

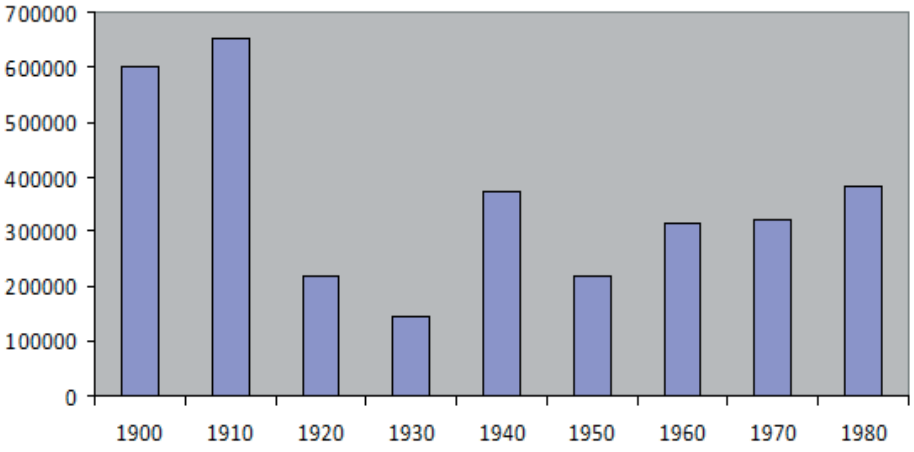

Figure 1 | *Current distribution of data in Greek Corpus 20*

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 ἡμέρα or even ἠμέρα. 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 ήμέρα and ήμετέρα and the search ημ?ρα all word forms for ήμέρα. 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/

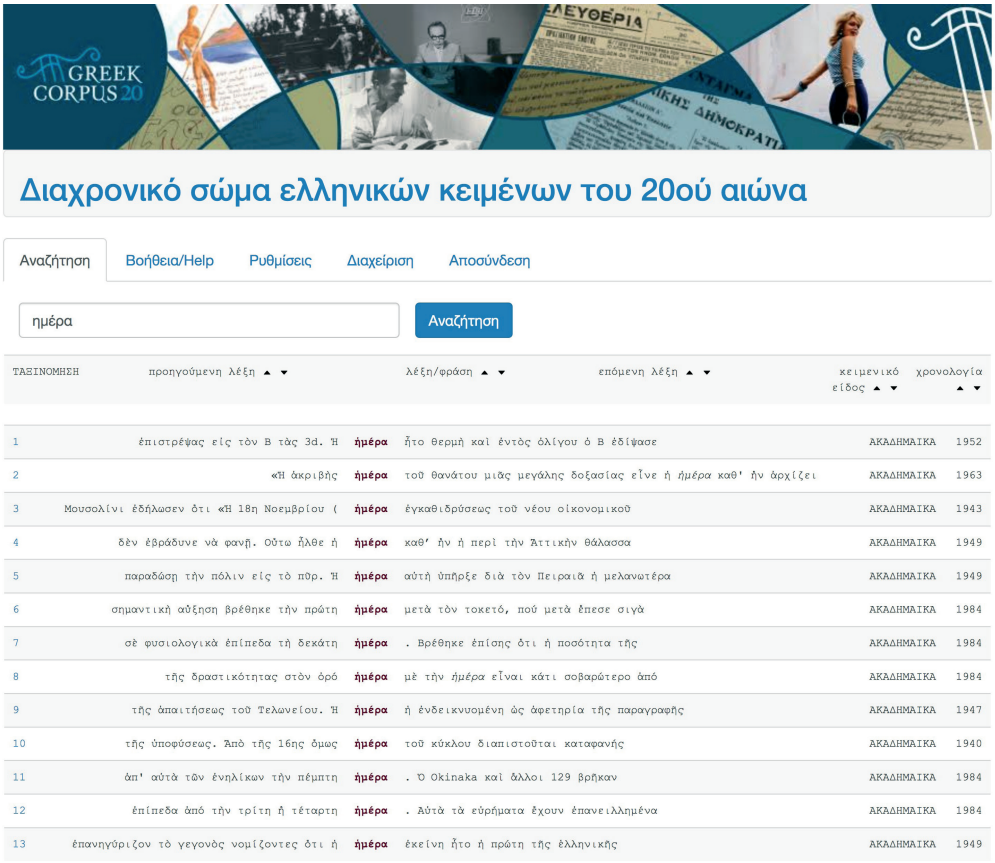

Figure 2 | *Greek Corpus 20 query result*

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

Aarts, Bas, Close, Joanne, Leech, Geoffrey, and Sean Wallis, eds. 2013. *The Verb Phrase in English: Investigating Recent Language Change with Corpora*. Cambridge: Cambridge University Press.

Baayen, Harald R. and Antoinette Renouf 1996. "Chronicling the *Times*: Productive lexical innovations in an English newspaper." *Language* 72 (1):69–96.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Baker, Paul. 2011. "Times may change, but we will always have money: Diachronic variation in recent British English." Journal of English Linguistics 39 (1):65–88.

Baker, Paul, Hardie, Andrew, and Tony McEnery. 2006. *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Beal, Joan C., Corrigan, Karen P., and Hermann L. Moisl, eds. 2007. *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases.*

Basingstoke: Palgrave Macmillan.
Culpeper, Jonathan, and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.
Davies, Mark. 2011. "Synchronic and diachronic uses of corpora". In *Perspectives on Corpus Linguistics*, edited by Vander Viana, Sonia Zyngier, and Geoff Barnbrook, 63–80. Amsterdam/Philadelphia: John Benjamins.
Davies, Mark. 2012. "Examining recent changes in English: Some methodological issues". In *The Oxford Handbook of the History of English*, edited by Terttu Nevalainen, and Elizabeth Closs Traugott, 263–287. Oxford: Oxford University Press.
Dossena, Marina, and Gabriella Del Lungo Camiciotti. 2012. *Letter Writing in Late Modern Europe*. Amsterdam/Philadelphia: Benjamins.
Duguid, Alison. 2010. "Investigating anti and some reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)." *Corpora* 5 (2):191–220.
Fischer, Roswitha. 1998. *Lexical Change in Present-Day English. A Corpus Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms.* Tübingen: Gunter Narr.
Fragaki, Georgia, and Dionysis Goutsos. 2015. "Greek diglossia in the 20th century: A historical corpus linguistics approach". Presented at the 12th International Conference on Greek Linguistics (ICGL12), Berlin.
Goutsos, Dionysis. 2010. "The Corpus of Greek Texts: A reference corpus for Modern Greek". *Corpora* 5 (1):29–44.
Goutsos, Dionysis, and Georgia Fragaki. 2014. "Prosfati glossiki allagi sta ellinika: Sxediasmos tou Diaxronikou Somatos Ellinikon Kimenon tou 20ου aiona." In *Selected Papers of the 11th International Conference on Greek Linguistics*, edited by G. Kotzoglou, K. Nikolou, E. Karantzola, K. Frantzi, I. Galantomos, M. Georgalidou, V. Kourti-Kazoullis, Ch. Papadopoulou, and E. Vlachou, 318–29. Rhodes: University of the Aegean.
Hatzigeorgiu, Nikolaos, Spiliotopoulou, Sophia, Vakalopoulou, Anna, Papakostopoulou, Anna, Piperidis, Stelios, Gavriilidou, Maria, and George Carayannis. 2001. "Ethnikos Thisavros Ellinikon Kimenon (ETHEG): Soma kimenon tis Neas Ellinikis sto Diadiktyo." *Studies in Greek Linguistics* 21:812–21.
Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Karaoglou, Ch. L. 2005. "Ta idika periodika kai i taksinomisi tous". In *La presse Grecque de 1784 à nos jours. Approches historiques et théoriques. Actes du Colloque International, Athènes, 23-25 mai 2002*, edited by Loukia Droulia, 263–75. Athens: Institute of Modern Greek Studies.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study.* Cambridge: Cambridge University Press.

Mair, Christian. 2009. "Corpora and the study of recent change in language". In *Corpus Linguistics. An International Handbook*. Vol. 2., edited by Anke Lüdeling, and Merja Kytö, 1109–25. Berlin/New York: Walter de Gruyter.

Marchi, Anna. 2010. "'The moral in the story': A diachronic investigation of lexicalised morality in the UK press." *Corpora* 5 (2):161–89.

Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Routledge.

Onelli, C. Proietti, D. Seidenari C., and F. Tamburini. 2006. The DiaCORIS project: A diachronic corpus of written Italian." In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa*. Available at: http://hnk.ffzg.hr/bibl/lrec2006/pdf/611_pdf.pdf.Partington, Alan. 2010. "Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project." *Corpora* 5 (2):83–108.

Partington, Alan. 2012. "The changing discourses on antisemitism in the UK press from 1993 to 2009: A modern diachronic corpus-assisted discourse study." *Journal of Language and Politics* 11 (1):51–76.

Schneider, Edgar W. 2004. "Investigating variation and change in written documents." In *The Handbook of Language Variation and Change*, edited by J. K. Chambers, Peter Trudgill, and Natalie, Schilling-Estes, 67–96. Oxford: Blackwell.

Taavitsainen, Irma, Kytö, Merja, Claridge, Claudia, and Jeremy Smith 2015. *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press.