GREED: Cataloguing and Encoding Modern Greek Dialectal Oral Corpora

Athanasios Karasimos
University of Patras
akarasimos@upatras.gr

Dimitra Melissaropoulou
University of Patras
dmelissa@upatras.gr

Angela Ralli
University of Patras
ralli@upatras.gr

Dimitrios Papazahariou
University of Patras
papaz@upatras.gr

Dimitrios Asimakopoulos
University of Patras
asimakop@ceid.upatras.gr

Abstract

Greece contains a rich variety of dialects (cf. Kontosopoulos 2006), some of which are used systematically in every-day speech, while others, are restricted to specific groups of elders and are in danger of extinction. Although little studied, remarkable variation appears among Modern Greek dialects not only on the level of phonology and vocabulary, but also on other levels of linguistic analysis, e.g. morphology, syntax, and semantics. Few research centres and associations are interested in written and spoken dialectal data, and publish relevant work; see, among others, the Historical Archive of Refugees (Thessaloniki), the Research Centre for Modern Greek Dialects of the Academy of Athens, the Centre of Asia Minor Studies (Athens). However, no attempt has been made to digitalize, catalogue and encode dialectal data until very recently, when the Laboratory of Modern Greek Dialects of the University of Patras started developing the GREED Database, with the aim to facilitate access to dialectal data, both linguistic and metalinguistic, in order to share dialectal information with the international linguistic community and ultimately preserve a significant linguistic heritage.

More particularly, GREED contains recordings from almost all dialectal areas of Greece (e.g. Grico, Cappadocian, Asia-Minor Dialects, Lesbian, Eptanisian, etc.); they have resulted from research programs and/or Master or Ph.D. dissertations. It has been built on established standards, such as the TEI Header, OLAC and IMDI, and its main goal is the enhancement of the material, the exchange of data and the support of academic research activities. It is not dependent on specific operational systems (OS) and commercial software, and a principal characteristic is the three-dimensional structuring concerning (i) the *dialectal data*, (ii) the *metalinguistic data* and (iii) the *combined browsing* of both.

The database is web-based, and its main settings are:

- Language Selection (Greek and English version)
- Dynamic development and use
- Combined use of sound and text
- Conforming to established standards and suggested proposals by the international linguistic community.
- Fast and easy processing of huge amount of the stored information (250 GB of audio and text files)
- User-friendly environment
- Different access levels for users and versioning

Since the basic aim is to develop a tool, which will be exploitable, fruitful and open to the linguistic community in the long term, an inventory has been built (bearing in mind the ISCC features, Dipper et al. 2007) in order to interact with different encoding systems, i.e. Praat (http://www.fon.hum.uva.nl/praat/), ELAN (http://www.lat-mpi.eu/tools/elan/),

and Toolbox (http://www.sil.org/computing/toolbox).

This well-established and widely-used software fulfills specific requirements and standards, such as:

- Open-source software and freeware in their use
- Wide range of design parameters
- Compatibility among various kinds of software that are used as annotation tools in different linguistic levels
- Possible add-ons and plug-ins
- Up-to-date support by the developers/ programmers
- Multilinguality and use of Unicode protocol

Although the work is still in progress, GREED contains 400 hours of natural dialectal speech, which is accompanied by metalinguistic information, and 40 hours of material are already annotated and transcribed.

Among the future benefits of the database will be to help future dialectal research, in categorizing and organizing various phonological and morphological phenomena, which are found cross-dialectally, and make easier the publication of dictionaries and grammars of the various Modern Greek dialects.

1. Introduction

Modern Greek is rich in dialectal varieties which are used in everyday speech while others, are restricted to specific groups of elders and are in danger of extinction (Trudgill 1998, Kontosopoulos 2001).

Although little studied, dialects display a remarkable variation on every level of linguistic analysis. This dialectal mosaic is due, to a large extent, to the specific historical, political and social circumstances that marked the History of the Modern Greek State, which was liberated from the Ottoman Empire at the beginning of the 19th century, and at that time consisted of the geographic areas of Peloponnese, Sterea Ellada and few islands. Since then, other parts of today's Greece got integrated into the Modern State (e.g. Crete, Macedonia, and the Dodecanese islands), a considerable number of Greek speaking refugees moved from Turkey (former Asia Minor) to Greece, after the end of the war between the two countries in 1922.

Today, Standard Modern Greek is mainly based on the Peloponnesian dialect, while the dialects of the other geographic areas constitute a multi-colored linguistic mosaic, which merits to be described, analyzed and thus, preserved, before it gets extinct.

Nevertheless, not much had been accomplished towards this direction. There exists in Greece, one research centre at the Academy of Athens, which is interested in written and spoken dialectal data, but its data collection is not digitized, most of it unpublished, and thus, difficult to be accessed. Non-digitized dialectal data can also be found in certain associations of Asia Minor refugees, as for instance, the Historical Archive of Minor Asia Greeks in Thessaloniki, and at the Centre of Asia Minor Studies, but they have been collected for historical purposes and are not systematically classified.

The first attempt to digitize, catalogue and encode dialectal data is made at the Laboratory of Modern Greek Dialects of the University of Patras, under the direction of Prof. Angela Ralli, with the development of a Database, GREED, which contains dialectal linguistic and meta-linguistic corpora. These data have been collected during our field work, where we register natural dialectal speech in order to provide a representative picture of the linguistic situation in particular geographical and social areas of Greece. At the same time, we collect dialectal manuscripts and various texts in order to create a digitized text corpus as well, the latter being though a long term goal. Our ambition is that GREED will assist our future research in categorizing and organizing various linguistic phenomena – phonological, morphological, sociolinguistic, etc. – which are found cross – dialectally. Therefore, it will make easier the publication of dictionaries and grammars of the various Modern Greek dialects.

2. GREED Corpus and Data Collection

The cornerstone for the development of the GREED were the following research projects:

- “*Grico: Dialect spoken in the area of Salento, South Italy*” (Interreg II, European Union, total of 55 hours, Directed by Angela Ralli).
- *“Dialectal varieties of Eastern Lesbos. Comparison with Asia Minor Dialect of Kydonies and Moschonisia”* (Ministry of Education, total of 45 hours, directed by Angela Ralli).
- *“Asia Minor Dialect of Kydonies and Moschonisia”* (Ministry of the Aegean and Ministry of Education, total 112 hours, directed by Angela Ralli).
- “*Cappadocian”*. Endangered Languages and Documentation Programme. University of London SOAS, total of 40 hours, directed by Mark Janse, Angela Ralli and Dimitris Papazachariou)
- “*Dialectal variation of Patras*” (University of Patras, total of 100 hours, directed by Dimitris Papazachariou).

However, other small individual or team fieldwork projects leading to M.A. or Ph.D theses are not to be neglected, since their contribution to the development of the database is also significant.

The following figure gives a representative picture of our material:

Dialectal Areas	*Hours*	*Percentage*	*Speakers*	*Percentage*
Cappadocian	41 Hours	8%	82 Speakers	12,77%
Asia Minor	105Hours	21,00%	78 Speakers	12,14%
Cyprus	2,5 Hours	0,50%	12 Speakers	1,89%
Dodecanese	9,5 Hours	2%	13 Speakers	2,02%
Epirus	12 Hours	2,20%	17 Speakers	2,60%
Ionian Islands	15 Hours	3,00%	33 Speakers	5,10%
Macedonia	9 Hours	1,60%	16 Speakers	2,50%
Lesbos	128 Hours	25,30%	80 Speakers	12,46%
South Italy	55 Hours	11,00%	68 Speakers	10,60%
Sterea Ellada	12 Hours	2,20%	21 Speakers	3,27%
Thessalia	8 Hours	2%	16 Speakers	2,50%
Thrace	8 Hours	2%	6 Speakers	1%
Peloponnese	100 Hours	20%	200 Speakers	31,15%
Summary	505 Hours	100,0 %	642 Speakers	100,0%

Speakers	*Person*	*percentage*
a. Age		
έως 20	16	2%
20-30	78	10%
30-40	37	4,7%
40-50	18	2,3%
50-60	184	23,5%
60-70	162	20,7%
70-80	92	11,7%
80-90	108	13,8%
90 -	60	7,7%
Unknown	28	3,6%
b. Genre		
Female	474	59,3%
Male	325	40,7%
Summary	799	100,0%

Recording date	*Number*	*Percentage*
2000 year	54	9,3%
2001 year	18	3,1%
2002 year	156	26,9%
2003 year	169	29,2%
2004 year	80	13,8%
2005 year	12	2%
2006 year	12	2%
2007 year	58	9,7%
2008 year	40	6,9%
Summary	579	100,0%

There are basically two different types of oral dialectal material a) Recordings of spontaneous dialectal live speech in audio footage and b) interviews targeting specific linguistic information in audio footage as well. It is important to stress that our data contain as well a significant number of oral history collections. The seek for oral histories was an intentional choice made in the case of our second research project, which used ethnographic methods and aimed at the preservation of cultural heritage. To be more specific, the recordings were made by field workers who shared existing social ties with the speech community and the informants in particular, or with the help and the physical presence of an intermediary, a local member who shared existing social ties with the informants (friend of a friend, relative of a relative). For example, talking about the informants' experiences, was the more efficient way to make Asia Minor refugees from Kydonies (Aivali) and Mochonisia to feel comfortable and open up – both emotionally and linguistically and –, to be able to get the interview schema off their minds and function as in their everyday life. This method proved precious as well for the recording of the living history of their life experiences as such.

In GREED, the dialects are distributed into geographical divisions (this will help us to build a dialectal atlas), and information concerning metadata is structured in seven major categories: File properties, Dialect, Research program, Technical Information, Communicative settings, Informants, Linguistic data. These major categories have many subcategories that provide us many options for an advanced search system. Two examples of the available subcategories of the basic metadata sets are the following:

(2 ΣΧΗΜΑΤΑ)

Although the work is still in progress, GREED contains 400 hours of natural dialectal speech, accompanied by metalinguistic information as shown above, and 40 hours of material already annotated and transcribed.

3. Formats (audio files and transcriptions)

Data recordings were held with the use of digital recorders. However, we also had source recordings from a number of consumer formats as CD, minidisc and computer files; additionally we were given some analog tape recordings from individuals or associations, who collaborated with us.

The informants were usually recorded in pairs, and the average recording is approximately 45 minutes. In the last three projects, the recordings were made with digital recorders (Marantz devices), which record in uncompressed .wav format, and minimize the time and effort spent for the digitization of the audio files. Furthermore, there were autonomous recordings in each audio channel – left and right – using two table microphones – one for each channel/informant. In this way, we managed to reduce the noise (around 40 db), to achieve a very good sound quality, and at the same time to reduce the effect of overlapping, when two interlocutors spoke at the same time or interrupted each other.

Note that the recording files are imported in the computer with no degradation of quality and stored into a NAS-Storage System for higher security. They are also edited before their integration in the corpus. Typical tasks include proper code-filename, separating channels, trimming the segments, boosting the low-volume recordings, reducing noise and clearing background microphone–artifacts.

So, our files types are:

- Audio footage: recording of live speech.
- Audio transcription: transcription of speech (turn takings, orthographical, phonetic [IPA-based], morphological).
- Texts/ books: documents originally written in a Greek dialect.

Apart from the digitized audio files, a proper data-base should include transcriptions of its audio files. There is a great deal of discussion as to what is the most relevant type of transcription (phonetic, phonemic or orthographic transcription). We agree with Durand and Eriksson, Anderwald and Wagner (2007:42-43) about the drawbacks of phonetic and phonemic transcription, and have, therefore, opted for orthographic transcription. Our choice was also influenced by the prospect of using morphological parsers on this transcribed material for further morphological analysis. We follow the conventions of Conversation Analysis, indicating turn-takings, interruptions, overlaps, pauses, lengthening, quick and slow tempo of speech, loudness and quiet speech, that is, characteristics of speech which can possibly influence post-lexical phonological phenomena.

Orthographic transcriptions make possible the investigation of morphosyntactic features and sociolinguistic phenomena, but there are obvious problems of legibility and numerous technical issues, as for instance, how to represent phonemic transcription on programs, such as Praat and E-Lan).

Finally, only orthographic transcription of the data would meet the other requirements of our corpus: the finished corpus should be machine-readable, enabling easy access to a variety of searches with various tools and most importantly, comparability with other corpora/ projects and simplicity in their usage. Moreover, orthographic transcription would allow us to compare our data with other written and spoken collections and enable us to make comparisons between different speakers, different dialects, different dialectical areas and different corpora.

Although we made our interviews accessible in the form of digitized audio files [the interested researcher may go to the original files to conduct some analysis], the absence of phonological transcription prevents quick and large-scale phonological analysis without using the audio files. All of our already transcribed audio files are documented and, thus, in many cases, phonological phenomena may be traced from the transcription without having to return to sound files. We hope that in the future the desirable alignment of sound and transcription, as in Necte (see Allen *et al.* Volume 2) and ONZE (see Gordon *et al.* Volume 2) will be succeeded in our database; for now this alignment is achieved only via ELan and Praat.

4. Administration and Website

Our requirements were for a system that could give access to the entire spoken data set from one interface. Tight controls of validation and other rules regarding the integrity of the data must be possible. Since the primary goal is to develop a tool, which will be exploitable, fruitful and open to the linguistic community in the long run, an inventory has been built (bearing in mind the ISCC features, Dipper et al. 2007) in order to interact with different encoding systems, i.e. Praat. Our system is capable to support Greek and Latin character set. The user interface we provide to the researchers is easy to use; therefore the training time is reduced.

Our database structure is based on four objects. All our objects (*Metadata, Metadatadetails, mdListValues* [pre-inserted values] and *FileAttribs* [table with all the files]) are connected to each other with a 'one for many' relationship, for example the value 'dialect name' of Metadatadetails is linked with 'Pontic', 'Lesvian', 'Cypriot' among others values of mdListValues. Our systems is built on a client-server architecture (apache server), which is associated with our relational database MySQL. All the pages are constructed using templates and processed using PHP scripts. The users access our database with PHP interface via HTML protocol. An important reason we chose a client/ server network is because it allows many users to access the base at the same time and the files to be stored at the server.

Our internet-based system follows, as we mentioned, the client/ server model as regards the information delivery. Based on this model a client pc unit is connected to the server unit, which contains the information and of course the client depends on the server to obtain the information. As web technology, it is open to any computing platform with a web browser. For now, to ensure the stability of the system, the users can upload files, but the metadata values must be entered by the admin after a user's request. In the presence, we work to compliant TEI (Text Encoding Initiative) to our data sets. Additionally, the system generates log reports for every change and problem, so we can detect the problem, for example when a contributor fails to update the necessary metadata forms within a specific period of time (30 seconds).

5. Tools

The software fulfills specific requirements and standards, such as:

- Open-source software and freeware in their use
- Wide range of design parameters
- Compatibility among various kinds of software that are used as annotation tools in different linguistic levels
- Possible add-ons and plug-ins
- Up-to-date support by the developers/ programmers
- Multilinguality and use of Unicode protocol

5.1. Why use ELAN and Praat?

ELAN has been successfully used for seven years and is equipped with a continuous institutional support base by Max Planck Institute; with ELAN a user can add an unlimited number of annotations to audio and/or video streams. They provide huge detailed guidelines for annotation (the ELAN manual) and an important set of software, a converter between popular programs (EConv), an IMDI metadata descriptions creator, an annotation exploration tool (Annex) among others. This program can be obtained free from the web. A number of researchers associated with psycholinguistics, laboratory phonetics/ phonology and dialectology had already used ELAN.

The advantages of ELAN are quite strong. An annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media. Annotations can be created on multiple tiers, which allow us to enter or extract tiers from/ to Praat. The textual content of annotations is always in Unicode and the transcription is stored in an XML format; it is common ground that XML is a markup metalanguage of World Wide Web with powerful tools for analyzing data and it was designed to be read by computer rather than a human. ELAN provides several different views on the annotations; each view is connected and synchronized to the media playhead. ELAN is written in the Java programming language and the sources are available for non-commercial use. It runs on Windows, Mac OS X and Linux.

Main features of the program:

- navigate through the media with different step sizes
  - easy navigation through existing annotations
  - waveform visualization of .wav files
  - support for template documents
  - input methods for a variety of script systems
  - multi-tier regular expression search, within a single document or in a selection of annotation documents
  - support for user definable Controlled Vocabularies
  - import and export of Shoebox/Toolbox, CHAT, Transcriber (import only), Praat and csv/tab-delimited text files
  - export to interlinear text, html, smil and subtitles text
  - printing of the annotations
  - multiple undo/redo

On the other hand, Praat, developed by the phoneticians Paul Boersma and David Weenink (University of Amsterdam) is a powerful tool with sixteen years of history; it is a computer program via which you can analyze, synthesize, and manipulate speech, and create high-quality pictures. Although Praat is primarily intended as a tool for phonetic research, it can be conveniently used for transcription purposes. However, due to some Unicode font problems concerning particular Greek characters, we use Praat for turn-taking and phonemic transcription and multi-layer transcription in ELan. Praat and ELan enable the transcriber to line up the speech signal with the transcription directly, such that researching through the transcriptions makes it possible to find the corresponding sound fragment. Another advantage of Praat is that it allows the user to divide the transcription into different tiers, which makes it possible to separate the contributions of the speech participants.

Main features of the program:

- Speech analysis, speech synthesis and listening experiments
- Labeling and segmentation
- Speech manipulation
- Learning algorithms and statistics
- Programmability, portability and configurability.

5.2 Why use Toolbox?

Finally, we also use the Field Linguist's Toolbox (HISTORY AND WHY) computer program, which is a data management and analysis tool for field linguists. It is especially useful for maintaining lexical data, and for morphological parsing and interlinearizing text, but it can be used to manage virtually any kind of data.

Toolbox is a text-oriented database management system with added functionality designed to meet the needs of a field linguist. But for ease of use, the Toolbox package includes prepared database definitions for a typical dictionary and text corpus.

The Toolbox database management system offers powerful functionality like customized sorting, multiple views of the same database, browse view to show data in tabular form, and filtering to show subsets of a database. It can handle any number of scripts in the same database. Each script has its own font and sorting characteristics. While Unicode is preferred, Toolbox can handle scripts in most legacy encoding systems. We have already worked with the design group of Toolbox (xxxxx) and we are glad to let you know that a Greek version of Toolbox is available.

One of Toolbox main qualities is its powerful linguistic functionality. It includes a morphological parser that can handle almost all types of morphophonemic processes. It has a user-definable interlinear text generation system which uses the morphological parser, lexicon and word formula components to generate annotated text. Interlinear text can be exported in a form suitable for use in linguistic papers. Toolbox has export capabilities that can be used to produce a publishable dictionary from a dictionary database.

Although Toolbox is very powerful, it is designed to be easy to learn. The user can start with a simple standard setup and gradually add the use of more powerful features as desired. The Toolbox downloads include a training package that is usable for self-paced individual learning as well as for classroom teaching of Toolbox. Of course, this package has been translated in Greek.

Toolbox may be freely distributed and hopefully serious bugs are usually fixed as soon as possible. The regular updates and the continuous support influenced us in favor of using Toolbox.

6. Preservation and backup

All our spoken data (and the written) are kept, either in digital form (computer files and DAT) or other storage form (minidisks, CDs, analog tapes). We kept the original copies of our recordings and books; for rare audio documents and documents we keep a copy, since we owe to return them to their owners. All the original copies are held in secure storage and computer files are accessible only by project staff, because our project processing is still in progress. To borrow any of the original copies the researchers contact project manager to be allowed to access to this material that should remain inside our lab.

All computer files relating to the corpus are stored on network drive administered by our lab technician; also we do a twice-a-week manual backup with archives stored at NAS Storage system. When conducting major systems works, major updates on the server we create two system images; nevertheless, if we encounter serious problems, we can draw back to a previous version of our system.

7. Future plans

GREED is a work in progress. It is in our intentions to provide a completed form, which can be open to the public for free academic use. With respect to our future plans, the following points need to be stressed:

a. [Tech] During the current phase of the project, we are developing an easy-to-use web interface, where no specific software is needed for the user. We have a comprehensive sociolinguistic metadata available, as well as details about the audio files. However, we should also provide information for our documents, which have not been yet properly catalogued. Our version is still under a testing mode, but has already proved to be quite fast and user-friendly.

b. [Tech] We are building now a more advanced system for metadata search. We plan to make our database faster, bug-free and provide code stability.

c. [Tech] Continue the annotation and transcription of the available dialectal data.

d. [Tech] Start a phonological transcription along with the orthographic one, and the morphological interlinearization (analysis).

e. [Tech] Beta-Evaluation by persons who have already worked with a database.

f. [Ling] Start exploiting the dialectal corpus by using a morphological analyzer, i.e. TOOLBOX, as a tool to our morphological research.

g. [Ling] Enrich the dialectal material, both oral and written, by organizing new fieldwork projects and digitalizing written dialectal corpora that we have already in our possession.

h. [Research] After the edition of the forth-coming Lexicon and Grammar of the Asia Minor Dialect of Kydonies and Moschonisia, we plan to provide lexica, dictionaries and grammars for the rest of major-collected dialects.

i. [Research]Apply for grants in order to gain the necessary financial support to carry on with the further development of GREED .

j. [Research] Use the database as a helpful tool for future dialectal research, in categorizing and organizing various phonological and morphological phenomena, which are found cross-dialectally, and make easier the publication of articles and monographs on the study of various Modern Greek dialects.

k. Communicate with the international linguistic community, facilitate access to Greek dialectal data, and ultimately preserve a significant linguistic heritage.

Selected References

✓ Academy of Athens. (1933-). *Ιστορικόν Λεξικόν της Νέας Ελληνικής Γλώσσης, της τε Κοινώς Ομιλουμένως και των Ιδιωμάτων*. [Historical Lexicon of Modern Greek Language, of Spoken Koine and of its Dialects]. Athens.

✓ Dipper, S., Goetze, M., Skopeteas, S. (2007). *Information Structure in Cross-linguistic corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. ISIS Working papers of the SFB 632.

✓ Historical Archive of Greek Refugees (http://www.kalamaria.gr/index.php?option=com_content&task=view&id=85&Itemid=599)

✓ Kontosopoulos, N. (2006)[4]. *Διάλεκτοι και ιδιώματα της Νέας Ελληνικής*. [Dialects and Idioms of Modern Greek]. Athens: Grigoris.

✓ Minas, K. (2003), *Η γλώσσα των Δημοσιευμένων Μεσαιωνικών ελληνικών εγγράφων της Κάτω Ιταλίας και της Σικελίας*. [The language of the published Greek Medieval documents of South Italy and Sicily] (reprinted by I.Λ.N.E.) Athens.

✓ Karanastasis, A. (1986-1992). *Ιστορικόν Λεξικόν των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας*. [Historical Lexicon of the Greek Dialects of South Italy]. Books A' - E'. Athens.

✓ Karanastasis, A. (1992). *Γραμματική των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας*. [Grammar of the Greek Dialects of South Italy]. Athens.

✓ Kostakis, Th. (1986-1987). *Λεξικό της Τσακωνικής Διαλέκτου*, [Lexicon of the Tsakonian Dialect]. Books A'-C', Athens.

✓ Ralli, Angela. 2006. Syntactic and Morphosyntactic Phenomena in Modern Greek Dialects: The State of the Art. 2007. *Journal of Greek Linguistics* 2006: 121-159.

✓ Ralli, Angela. To appear. *Λεξικό των Ιδιωμάτων Κυδωνιών, Μοσχονησίων και Ανατολικής Λέσβου* [A Dictionary of the Asia Minor Dialects of Kydonies, Moschonisia and the Dialect of East Lesbos]. University of Patras

✓ Manolis Triandafillidis Fountation (http://ins.web.auth.gr/english.htm)

✓ Academy of Athens (http://www.academyofathens.gr/echome.asp?lang=2)

✓ Anderson, J., Beavan, D., Kay, C. (2007). *SCOTS: Scottish Corpus of Texts and Speech*, Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication, pp. 17-34.

✓ Anderwald, L., Wagner, S. (1997). *FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data.* Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

✓ Barbiers, S., Cornips, L., Kunst, J.-P. (2007). *The Syntactic Atlas of the Dutch Dialects (SAND): A Corpus of Elicited Speech as an On-line Dynamic Atlas.* Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

✓ MacWhinney, B. (2007). *The Talkbank Project*. Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

✓ Allen, W., Beal, J., Corrigan, K., Maguire, Moisl, H. (2007). *A Linguistic 'Time Capsule': The Newcastle Electronic Corpus of Tyneside English*. Creating and digitalizing Language Corpora Vol.2 (edited by Beal J. *et al.*). Palgrave McMillan Publication, pp. 16-48.

✓ Gordon, E., Maclagan, M., Hay, J. (2007). *The ONZE Corpus*. Creating and digitalizing Language Corpora Vol.2 (edited by Beal J. *et al.*). Palgrave McMillan Publication, pp. 82-104.