Towards a Unified Exploitation of Electronic Dialectal Corpora: Problems and Perspectives*

Nikitas N. Karanikolas[1], Eleni Galiotou[1], and Angela Ralli[2]

[1] Department of Informatics, Technological Educational Institute of Athens, GR-12210 Aigaleo, Athens, Greece
{nnk,egali}@teiath.gr

[2] Department of Philology, University of Patras, GR-26504 Rio, Patras, Greece
ralli@upatras.gr

Abstract. In this paper, we deal with the problem of storing and retrieving dialectal data in a unified framework. In particular, we discuss issues concerning the design and implementation of a multimedia database which will contain written and oral data from three Greek dialects in Asia Minor. At first, we describe the overall architecture of a system aiming at providing the user with the possibility to store audio recordings, text transcripts, and other annotations. Then we discuss the possibilities and limitations of a retrieval module aiming at combining different linguistic levels for a unified exploitation of oral and written corpora.

Keywords: Computational Dialectology, Electronic Corpora, Modern Greek Dialects, Multimedia databases

1 Introduction

The use of computational techniques and the possibility to store oral and written dialectal data on electronic media has greatly contributed to the advancement of research in dialectal change and language contact. For example, the on-line tool for Dutch dialect syntax research DynaSAND (the Dynamic Syntactic Atlas of Dutch Dialects) provides a database, a search engine, a cartographic component, and a bibliography concerning the syntactic variation in dialects located in the Netherlands, Belgium and France [2]. Another interesting approach is that of LAMSAS (Linguistic Atlas of Middle and South Atlantic States) [18] where dialectal material from the Atlantic coast of the United States is comprised. As for Greek dialectal data, no results of a computational processing were reported until very recently, with the exception of an electronic dictionary of Cypriot Greek [21]. Greek dialects of Asia Minor constitute a quite interesting case in the scientific field of dialectology and language contact; although they are of the same Indo-European origin (Greek), they have gradually diverged from one another partly under the influence of an Altaic language (Turkish) to such an extent that they

* This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European SocialFund.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 257–266, 2014.

are considered as different dialects. Moreover, Greek and Turkish belong to different typological groups (fusional vs. agglutinative). Therefore, a systematic study of Asia Minor Greek dialects would give useful insights as for the nature of language change within the domain of dialectal variation. The aforementioned task would be greatly facilitated by the availability of dialectal data on electronic media and the development of computational tools for their processing. Recently, an attempt to combine Informatics and Theoretical Linguistics in order to describe and analyze dialectal phenomena of three Greek dialects in Asia Minor has been undertaken in the course of the "Thalis" program: "Pontus, Cappadocia, Aivali: in search of Asia Minor Greek" (AMiGre). The aim of his project is twofold: (a) to provide a systematic and comprehensive study of Pontic, Cappadocian and Aivaliot, three Greek dialects of Asia Minor of common origin and of parallel evolution that are faced with the threat of extinction; (b) to digitize, archive and process a wide range of oral and written data thus contributing to the sustainability and awareness of this longwinded cultural heritage.

The computational component of the project comprises activities such as [8]: (a) design and development of a multimedia tri-dialectal dictionary which contains lemmata from the three dialects (Pontic, Cappadocian, Aivaliot) in a comparative way [11]; (b) design and development of a multi-media software and database for the archiving and processing of oral and written dialectical data.

In this paper, we deal the problem of storing and retrieving dialectal corpora on electronic media. In particular, we present the design of a multimedia software and database for archiving and processing of oral and written data from the three Asia Minor Greek dialects. In Section 2, we present the design principles and the modules of the multimedia software. Next we discuss the technical aspects and the structure of the database. In Section 4, we present the search and retrieve module. Finally, in Section 5, we draw conclusions and point to future work.

2 A Multimedia Software for Oral and Written Dialectal Resources

2.1 Design principles

The nature of the data. The oral corpus of the AMiGre project in its current state consists of approx. 180 hours (i.e. 60 hours/dialect) of recorded raw data and was compiled in the Laboratory of Modern Greek Dialects of the University of Patras [19]. The raw data were annotated, abstracted and analyzed according to the 3A (annotation, abstraction, analysis) model [22]. Some 45 hours (15hours/dialect) of raw data were further processed resulting in a multimodal sub-corpus which combines raw data with transcription, translation, annotation and metadata. This multimodal sub-corpus is processed using ELAN, a software package for the creation of audio and video resources ([6]; [20]) while spoken data are further analyzed using Praat, a scientific software package for the analysis of Speech in Phonetics ([3,4]). The spoken data are annotated according to the speaker's turn-takings, utterances and intonation phrases. Phonological words, syllables and phonemes are also annotated using the IPA (International Phonetic Alphabet) symbols [10]. Explicit representations of vowels, diphthongs, consonants and

consonant clusters appear on different layers of representation (tiers). So, the output files of the processing with ELAN and Praat act as input files for our system. The written corpus consists of 1,000,000 words of digitized dialectal texts from primary written sources of the 19th and early 20th centuries [12]. The inclusion of a text in the corpus was based upon criteria such as representativeness (according to a dialect, a local sub-dialect, chronological period) and quality (closeness to the actual spoken language, consistent linguistic terminology or transcription system). A sub-corpus consisting of 200,000 words was transcribed using a custom-made transcription system based on SAMPA (Speech Assessment Methods Phonetic Alphabet) [23]. The use of the particular transcription system aims at: (a) facilitating further electronic elaboration by allowing transcription without the use of special diacritics on keyboard configurations; (b) representing all the special sounds included in the phonetic inventory of the dialects under investigation; (c) unify the disparate and inconsistent notation of the original written sources [15]. Finally, some 50,000 words of the transcribed sub-corpus were annotated with the use of a special tool developed for the purposes of this project. The annotation describes the levels of phonology, morphology and the lexicon. Annotation at the morphological level follows the principles described in [13]. Note that, the linguistic annotation of the written corpus follows the same principles and categories as the oral one, in order to enable unified searches across the whole available dialectal corpus.

Design. The whole dialectal corpus (oral and written) will be stored in a multimedia database for further exploitation. The variety of linguistic information and annotation types would ask for an advanced software tool such as Labb-CAT ([7], [14]) which provides the user with the possibility to store audio or video recordings, text transcripts and other annotations. Yet, the system in question could not deal with our basic requirements, i.e. (a) Annotations at many different linguistic levels and, (b) Combined search at both the oral and written corpus. Consequently, we opted for the design and implementation of a software which would be tailored to our needs. The architecture of the proposed system is depicted in Figure 1. As shown in the schema, the two subsystems ”G. Oral” (Graphical User Interface for Oral resources) and ”G. Written” (GUI for Written resources) invoke a number of web-like applications related to the processing of oral and written resources respectively. The system also comprises two indexing modules (I. Oral= Indexing module for oral resources, I. Written= Indexing module for written resources). Finally, the ”Search and Retrieve” module invokes all the web-like applications for a combined research in both oral and written data.

2.2 The applications

The aforementioned system comprises 8 web-like modules for the processing of oral and written resources:

Phon Tagger: The phonological tagging application is used on both oral and written resources. Annotation on the written resources is performed at the word level. In order to achieve a unified treatment, information on morphological word boundaries will be added to oral resources.

Morph Tagger: The morphological tagging application is also used on both oral and written resources. In all the resources annotation is performed at the word level. For

each morphological word, information on part of speech, grammatical properties, and morphological phenomena such as derivation and compounding, is provided.

Synt Tagger: The syntactic tagger is also used on both oral and written resources. In the current state of the system, annotation is performed at the word level; each word is associated to at most one syntactic structure. The application provides also the possibility to perform annotations on a phrase or sentence level.

Sem Tagger: The semantic tagging application is also used on both oral and written resources. Annotation is performed on a sequence of words with values such as "loan", "idiomatic phrase" etc.

Text Imaging: This application aims at the preview of pages of written resources.

Text Transcription: The application of transcription of a written resource provides two panels. The left panel contains the image of the page under processing and the right one contains the transcription of the page in the form of processable text.

MOS (Oral Metadata). This application provides the possibility to store and update the metadata of oral resources comprising information as age, sex, cultural background of the speaker. Note that, such information is not available for written resources.

PRAAT: It invokes the Praat software for phonetic analysis [4].

3 The Structure of the Database

3.1 Transcribed Written Documents source

The Transcribed Written Documents source is a collection of text files encoded with UTF-8 (Unicode) encoding. Each file contains the transcription of a written document. The written documents are manuscripts, typescripts or printed material that is related with records of the Asia Minor Greek Dialects. The transcriptions are based on lowercase letters of the Greek alphabet and some other symbols (uppercase Greek letters, Latin letters and letter-couples). This is an endeavor to homogenize the various symbols used by different researchers of the Asia Minor Greek Dialects, in order to indicate the same phonological phenomenon [15].

3.2 Image Files source

The Image Files source is a collection of digitized versions of the original documents (manuscripts, typescripts or printed materials that are related with records of the Asia Minor Greek Dialects). Each image file is usually a digitized version of a single page of one written document. The alternative solution to keep a single file (e.g. a multi-page tiff file) for each written document is considered inexpedient. This is because the system runs on the web and loading the pages can cause great delays when the user moves from one document to another. The overhead of the selected solution is to associate a list of Image File Paths (one path for each page) with any single written document (the primary record used for the written document). It is obvious that the values of the Image File Paths list should be accessible by the Text Imaging module.

3.3 WID Text Files source

Each WID Text File is a collection of words that emanate from the tokenization of a Transcribed Written Document. Each word (token) is characterized by:

- WordID (an identifier of the word/token which is unique for the whole WID Text Files source),
- Word (the word as is in the transcribed text),
- PositionIndex (sequence number of the word/token in the transcribed written document),
- Location (starting byte and ending byte of the word/token in the transcribed written document),
- PartID (an identifier of the page containing the word/token, it is unique for the whole set of parts/pages constituting the written documents).

From the PartID we can deduce the source (the primary record used for the written document) and the page of the document (the sequence number of page into the document). A set of words/tokens (identified by a set of WordIDs coming from the same document) constitute a single WID Text File. The set of WID Text Files constitute the WID Text Files source.

3.4 TextGrid Files source

This is a collection of files produced by the Praat software package for phonetic analysis [4]. Praat uses different tiers defined by the user to annotate an audio file. Annotations can vary and can comprise turn takings, transcription, intonation phrases, intonation words, syllables, phonemes, etc. It also supports point tiers that can be used for pitch (e.g. pitch accent, phrase accent, boundary tone), etc. In order to capture all necessary levels of annotation, we have also introduced a ”morphological words” tier. In our system, each TextGrid file is associated with a relevant audio file (discussed in Section 3.5). We can make the abstraction that tokens/words in a WID Text File are the equivalents to the ”morphological words” in a TextGrid file. The association of a TextGrid file with the relevant oral document (the primary record) imposes to keep the TexGrid File Path inside the primary record of the oral document.

3.5 WAV Files source

This is a collection of audio files in the WAV (Wave Audio File) format. The purpose of an audio file has been discussed in Section 3.4. The association of an audio file with the relevant oral document (the primary record) is achieved by including a WAV File Path field into the primary record of the oral document.

3.6 EAV database

Actually the EAV (Entity Attribute Value) database consists of five sub-schemas, namely:

- EAV Phon (for storing phonological phenomena of single words),
- EAV Syn (for storing syntactic phenomena regarding sequences of words),
- EAV Sem (for storing semantic phenomena regarding sequences of words),
- EAV Morpho (for storing morphological attributes characterizing single words),
- EAV Meta (for storing metadata for oral documents).

With the term “single words” we refer to tokens from the transcribed written documents or morphological words from the TextGrid files. With the term “sequences of words” we refer to sequences of tokens from the transcribed written documents or sequences of morphological words from the TextGrid files. The EAV Morpho sub-schema is the most complicated one because we have to store a sparse set of morphological attribute - value pairs for each morphological word. In addition to the sparse nature of morphological attributes, we have to deal with subschema evolution (the attributes are not fixed from the beginning) and also with hierarchical values in the attribute - value pair. A similar sparse nature and need for subschema evolution we have to cope with, is the EAV Meta subschema. These are the main reasons for adopting the EAV (Entity - Attribute - Value) technology ([1,9,16,17]) to design the above five sub schemata. An example of the EAV Morpho subschema is depicted in Tables 1, 2 and 3. The ”Repetition” field in the ”EAV Morpho” table indicates the number of morphological phenomena associated to a single word while, the ”TypeId” field in the ”EAV Morpho Properties” table indicates whether the value of a property supports single or multiple values and whether the values come from a predefined list or they are text values.

Table 1. EAV Morpho

WordId	PropertyId	Repetition	PropertyValueId
385	4	1	177
385	26	1	268
387	4	1	176
387	5	1	182
390	4	1	175
391	4	1	176
391	5	1	182

3.7 Inverted Files database

Inverted Files [5] do not have the simple form used in Text Information Retrieval (couples of a word/lemma and a list of occurrences of the word/lemma in documents). In our case (an Electronic Multimedia Dialectal Corpus) there is a need for a form such as :

- a word/token/phenomenon,
- a list of quadruples, where each quadruple specifies the location of the occurrence of a word/token/phenomenon.

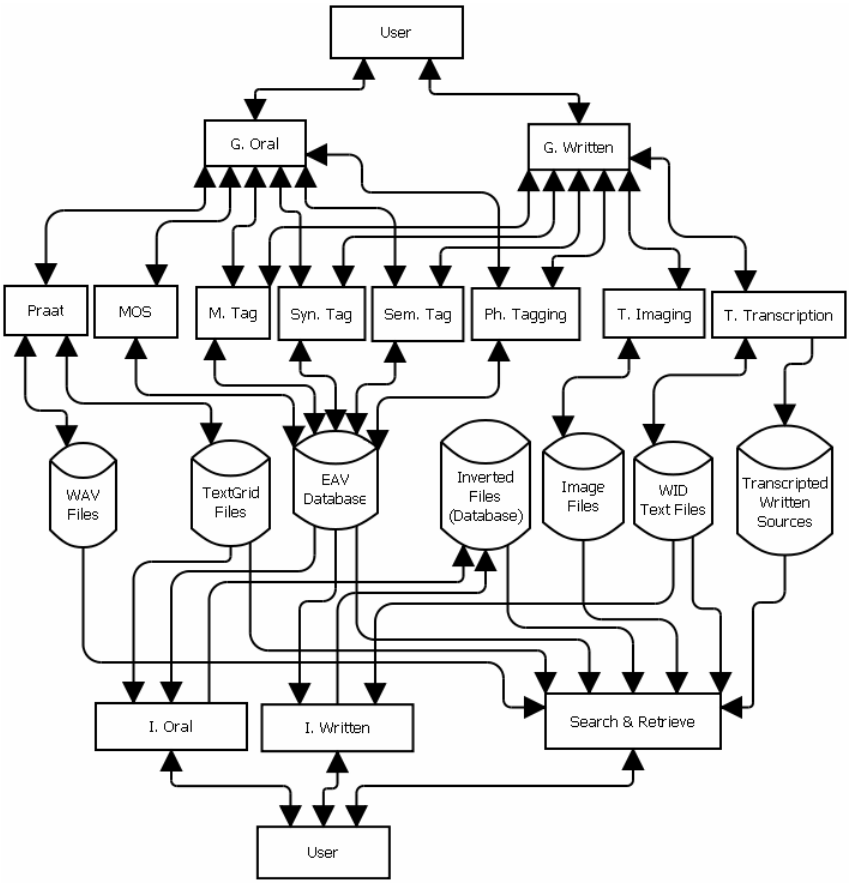

Fig. 1. The Architecture of the system

Each quadruple (location) contains:

- The identifier of the relevant database or source (type of information), a value from the set EAV Phon, EAV Syn, EAV Sem, EAV Morpho, EAV Meta, Text-Grid File, Transcribed Written Document,
- The identifier (the primary key or other unique name) that specifies a concrete instance among the collection of instances in the relevant database (file collection),
- The attribute (only in the case of inversion of EAV Morpho and EAV Meta data),
- The word/token sequence (a couple of the form (StartWordId, EndWordId)).

4 Search and Retrieve and Other Advanced Modules

The Search & Retrieve module invokes the relevant application (among the ones explained in Section 2.2) using OLE (Object Linking and Embedding) Automation

Table 2. EAV Morpho Properties

PropertyId	Name	TypeId
4	'PART OF SPEECH'	4
5	'GENDER'	4
6	'INFL. CLASS'	4
...	...	...
24	'ORIGIN OF BASE FORM'	4
25	'ORIGINAL LOAN WORD'	1
26	'PART OF SPEECH OF LOAN WORD'	4

Table 3. EAV Morpho Lookup

PropertyValueId	Description	PropertyId
175	'Adjective'	4
176	'Noun'	4
177	'Verb'	4
...	...	...
182	'Neuter'	5
183	'3-gender'	5
184	'Masculin'	5
185	'Feminin'	5
...	...	...
225	'Turkish'	15
226	'Italian'	15
227	'Roman Dialects'	15
228	'Greek'	15
...	...	...

or other equivalent technology. The selection of the relevant application, among the available ones, is automatically decided, depending on the types of information that match the user defined criteria. It can also be defined by the user. The Search & Retrieve module provides a query builder to the user who can add and combine criteria. For each criterion, the user defines the requested value (word/token/phenomenon) and the location (a completely or partially defined quadruple) where the value should occur. Table 4 depicts a simple example of a query defined by the Search and Retrieve query builder:

Table 4. Query on the transcriptions of written documents where both phenomena of synaeresis and palatalization appear into a word/token window of size 10.

Word/token/phenomenon	Location (Quadruple)			
synaeresis	EAV Phon	-	-	(X, X+10)
palatalization	EAV Phon	-	-	(X, X+10)
Output	Transcribed Written Document			

The G. Oral (GUI for Oral sources) and G. Written (GUI for Written sources) are typical administrative applications that permit Add, Update, Delete and Browse facilities. The I. Oral (Indexing Module for Oral sources) and the I. Written (Indexing Module for Written sources) are mainly responsible for performing inversions.

5 Conclusions and Future Work

In this paper, we have presented the design of a software application aiming at storing and exploiting oral and written dialectal corpora in unified way. Our system takes into account alternative sources of information using the EAV technology and performing a combined inversion which is adapted to the complexity of linguistic dialectal data. The software is currently under development and the first implementation results are expected to contribute to our understanding and awareness of dialectal corpora processing at both linguistic and computational levels.

References

1. Anhoj, J. (2003). Generic Design of Web-Based Clinical Databases. Journal Medical Internet Research, 4.
2. Barbiers, S. et al (2006). Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute. http://www.meertens.knaw.nl/sand/
3. Boersma, P. (2012). The use of Praat in corpus research, in: Jacques Durand, J. Gut, U. Kristofferson, G. (eds.): Handbook of corpus phonology. Oxford: OUP
4. Boersma, P., Weenink, D. (2013). Praat: Doing phonetics by computer. http://www.praat.org
5. Buttcher, S., Clarke, C., Cormack, G. (2010). Information Retrieval: Implementing and Evaluating Search Engines. Cambridge, Massachusetts: MIT Press.
6. ELAN: http://tla.mpi.nl/tools/tla-tools/elan/, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands
7. Fromont, R. , Hay, J. (2008). ONZE Miner: the development of a browser-based research tool, Corpora, 3(2), 173–193
8. Galiotou, E., Karanikolas, N., Manolessou, I., Pantelidis, N., Papazachariou, D., Ralli, A., Xydopoulos, G. "Asia Minor Greek: Towards a Computational Processing", Procedia: Social and Behavioral Science, Elsevier (forthcoming, 2014)
9. Johnson S. B., Chatziantoniou D. (1999). Extended SQL for manipulating clinical warehouse data. In AMIA 1999, pp. 819–823.
10. IPA chart : http://www.langsci.ucl.ac.uk/ipa/ipachart.html
11. Karanikolas, N., Galiotou, E., Xydopoulos, G., Ralli, A., Athanasakos, K., Koronakis, G. (2013)- Structuring a Multimedia tridialectal dictionary, Proc. 16th Int. Conf. on Text, Speech and Dialogue (TSD 2013), September 1 - 5 2013, Plzen, CZ, LNCS vol. 8082, 509–518, Springer.
12. Koliopoulou, M., Markopoulos Th., Pantelidis, N. (2013). Pontus, Cappadocia, Aivali: Challenges of a digital corpus of written material (in Greek), The 11th International Conference of Greek Linguistics, Rhodes, Sep.2013
13. Koutsoukos, N. , A. Ralli (2012). 'From derivation to inflection: a process of grammaticalization'. Morphology Meeting 2012, (Leiden, the Netherlands, 08-09-2012)
14. LaBB-CAT (formerly ONZE Miner). http://onzeminer.sourceforge.net/

15. Manolessou, I., Beis, S., Bassea-Bezantakou (2012). The phonetic transcription of Modern Greek dialects (in Greek), Lexicographicon Deltion 26, 161–222.
16. Nadkarni P. (2000). Clinical Patient Record Systems Architecture: An Overview. Journal of Postgraduate Medicine 46 (3), 199–204.
17. Nadkarni P. (2002). An introduction to entity-attribute-value design for generic clinical study data management systems. Presentation in: National GCRC Meeting. Baltimore, MD.
18. Nerbonne, J. and Kleiweg, P. (2003). Lexical distance in LAMSAS, Computers and the Humanities 37 (3), 339–357
19. Ralli, A., Papazachariou, D. Karasimos, A. (2010). Laboratory of Modern Greek Dialects and the project GreeD. In Ralli, A. et al. (eds.), Proc. 4th Int. Conf. of Modern Greek Dialects and Linguistic Theory
20. Sloetjes, H., Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)
21. Themistocleous, C., Katsogiannou, M., Armosti, S., Christodoulou, K. (2012). Cypriot Greek Lexicography: An Online Lexical Database, Proceedings of Euralex 2012, 889–891.
22. Wallis, S. , Nelson, G. (2001). Knowledge discovery in grammatically analyzed corpora. Data Mining & Knowledge Discovery, 5, 305:335
23. Wells, J.C. (1997). 'SAMPA computer readable phonetic alphabet'. in Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.