AMiGre: A unified framework for archiving and processing oral and written dialectal data

Eleni Galiotou[1], Angela Ralli[2]
[1] Technological Educational Institute of Athens, Greece – egali@teiath.gr
[2] University of Patras, Greece – ralli@upatras.gr

ABSTRACT

Asia Minor Greek constitutes an ideal case study for historical linguistics and language contact. It offers testimonies for tracing the evolution of the Greek language and displays numerous borrowings, due to its long-lasting contact with Turkish. Its dialects belong to a rich cultural and linguistic heritage but are faced with the threat of extinction. Consequently, the description and preservation of this invaluable heritage is of crucial importance, the sustainability and awareness of which can be done in a decisive way with the use of modern Information Technology. In this paper, we deal with the problem of archiving and processing digitized corpora of oral and written data of three dialects of Asia Minor Greek, Pontic, Cappadocian, and Aivaliot, that have been compiled in the frame of the research program AMiGre (THALIS, 2012-2015), co-financed by the European Union and Greek national funds. The project aimed at: (a) providing a systematic and comprehensive study of Pontic, Cappadocian and Aivaliot, three Asia Minor dialects of common origin and parallel evolution, which are also in danger of extinction; (b) digitizing, archiving and processing a wide range of oral and written data, thus contributing to the sustainability and awareness of this longwinded cultural heritage. The project resulted in the scientific presentation of data from Asia Minor Greek to the academia, under the form of multimedia corpora, an electronic tri-dialectal dictionary and a number of publications. The corpora of oral and written data are stored in a multimedia database which allows the parallel visualization of raw and processed dialectal data as well as the coding of a large number of metadata. The search and retrieve subsystem provides a combined search at two levels of linguistic representation, phonological and morphological, access in metadata and combined search in both oral and written resources.

KEYWORDS

Asia Minor Greek, dialects, multimedia databases, electronic corpora

1. INTRODUCTION

In recent years, there has been a growing interest in the use of Information Technology for the study of language contact and dialectal change. In fact, the availability of dialectal data on electronic media and the development of computational tools has contributed in a decisive way to the sustainability and awareness of this invaluable cultural heritage. To this end, several attempts to build dialectal corpora are reported. Indicatively, a multimedia database was used in order to create a voice language map of Japanese dialects [12], and the Linguistic Atlas of Middle and South Atlantic States (LAMSAS) was created in order to archive and process dialectal data from the Atlantic coast of the United States [10]. As far as European languages are concerned, several attempts are also reported. Indicatively, the Scottish Corpus of Text and Speech (SCOTS) is the outcome of a project aiming at constructing a large-scale electronic corpus of oral and written texts of the languages of Scotland [1]. A set of Natural Language Processing tools for the processing of Swiss German oral dialectal data is described in [11]. Similarly, the Corpus Oral Dialectal (COD) – a corpus of contemporary Catalan – is provided in [13]. The DynaSAND (Dynamic Syntactic Atlas of Dutch Dialects), an on-line tool for the processing of Dutch syntactic variation ([3]) is enriched and enhanced with a web service interface to the DynaSAND corpus, so that the data from the corpus can be used for other applications as well [9]. The first attempt in Greece to combine Information Technology and Theoretical Linguistics in order to present both raw and processed dialectal material on digital space, with the use of the most up-to-date software programs was the “THALIS” project “Pontus, Cappadocia, Aivali: in search of Asia Minor Greek (AMiGre[1])”. The project resulted in the scientific presentation of Greek dialectal data to the academia under the form of a multimedia corpus, an

[1] This research was co-financed by the European Union (European Social Fund–ESF) and Greek national funds through the Operational Program "Education and Life-long Learning" of the National Strategic Reference framework (NSRF) - Research Funding Program: “THALIS. Investing in knowledge society” through the European Social Fund. The project was directed by Prof. Angela Ralli and hosted at the *Laboratory of Modern Greek Dialects* of the University of Patras.

electronic tri-dialectal dictionary [5] and a number of publications. In fact, the aim of this project was 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.

This paper, focuses on the issue of storing and retrieving dialectal corpora on electronic media. In particular, we present the design and development of a multimedia software and database for archiving and processing oral and written dialectal data which were compiled in the course of the AMiGre project [4].

2. THE ORAL AND WRITTEN CORPORA

The corpora of the AMiGre project consist of both written and oral data, which have been collected through archival search and field work, respectively. The collection was granted the written permission of the owners of written material or that of the informants, depending on the case, and had the approval of the Ethics committee of the University of Patras. They contain 180 hours of recorded raw data (narratives) [8], and digitized texts of 1,000,000 words, compiled from primary written sources which were detected in various archives of the 19th and early 20th centuries [8]. The corpora constitute the Asia Minor Archive (AMiGre, URL: http://amigre.upatras.gr), which is stored at the server of the *Laboratory of Modern Greek Dialects* (LMGD) of the University of Patras (htpp://lmgd.philology.upatras.gr), the personnel of which assures its long-term maintenance and preservation.[2] From these corpora, all written texts and a subset of 15 hours (5 hours per dialect) of the oral data are stored in an archive accessed electronically at the address http://amigredb.philology.upatras.gr. This material was annotated as for some metadata, for instance, information is given about the source, the informant and the conditions of collection. Note that from the written material, a subset of 50.000 words and the 15 hours of the oral material have been further annotated, translated and linguistically analyzed as far as phonetics/phonology and morphology are concerned. All processed written and oral data, together with those of the rest of raw oral material are accessible only to the researchers of LMGD, through a specific application. It is worth mentioning that the corpora were processed by using ELAN[3] for multimodal annotation and phonetically analyzed with the use of Praat,[4] resulting in the representation of vowels, diphthongs, consonants and consonant clusters appearing on different layers (tiers). One more tier was added, that of morphological representation, consisting of word internal morphological segmentation, constituent recognition and categorization. Segments and consonants were transcribed with the IPA[5] symbols, while morphological words and syllables with the use of the SAMPA alphabet [14].

An advanced software tool such as Labb-CAT[6] which provides the user with the possibility to store audio or video recordings, text transcripts and other annotations should be able to deal with the variety of linguistic information and annotation types. Yet, the system in question could not deal with our basic requirements, that is, (a) annotations at different linguistic levels, and (b) combined search at both the oral and written corpora. Consequently, we opted for the design and implementation of a software which would be tailored to our needs [6, 7].

3. THE MULTIMEDIA DATABASE

Our system comprises four collections of data files: digital recordings of oral data (WAV files), initial annotations of oral data (TextGrid files – output of Praat), digitized pages from the original texts (Image files), and transcriptions of pages of written resources (Transcripted Written Sources). Editable elements are stored in three databases: (a) “Struct” database: implements the abstract hierarchical structure common to both oral and written resources; (b) “EAV” database: it follows the “Entity-Attribute-Value” schema [2], since it comprises an extendable set of entities and attributes; (c) “Inner” database: it also

[2] One of the products, the tri-dialectal dictionary (LEPOKAM) consisting of 7,500 dialectal lemmas of Pontic, Aivaliot and Cappadocian is already available to researchers, upon request and the permission of the director of the project, Prof. Angela Ralli (http://lepokam.philology.upatras.gr).

[3] http://tla.mpi.nl/tools/tla-tools/elan/

[4] http://www.praat.org

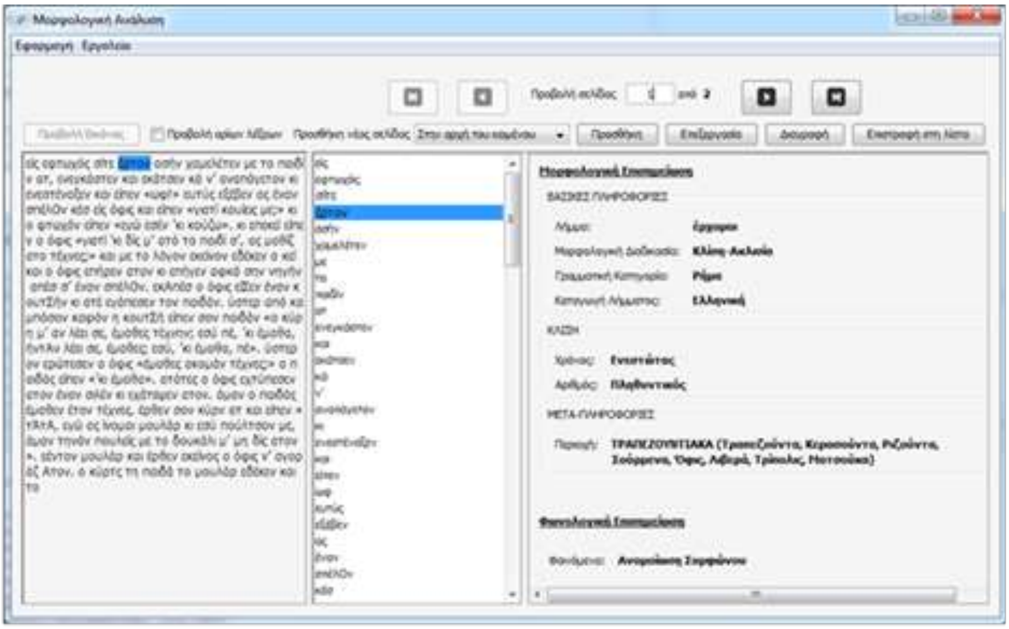
[5] http://www.langsci.ucl.ac.uk/ipa/ipachart.html

[6] http://onzeminer.sourceforge.net/

follows the EAV schema and contains word fragments and their annotations. Our system is built around two basic subsystems: "G. Oral" (Graphical User Interface for Oral resources) and "G. Written" (GUI for Written resources), which invoke a number of web-like applications related to the processing of oral and written resources respectively. These applications comprise Taggers at different levels (Phonological, Morphological, Syntactic, Semantic) -although only the first two were used for the AMiGre purposes- Preview of Image resources, Transcription and Annotation modules. It also comprises a module for storing and updating metadata. An example of an overview of written data is depicted in Figure 1. The interface is in the form of a triptych; the left panel depicts a transcript page of the document, the middle one shows the list of morphological words and the right one contains the annotations of the highlighted word. A similar approach is followed when oral data are treated. Finally, the "Search and Retrieve" module invokes all the web-like applications for a combined research in both oral and written data. It allows the user to combine multiple criteria, define restrictions (value, distance) and focus on a certain level (document, part, etc.). The form of the query follows the template which is depicted in Figure 2.

4. CONCLUSIONS

In this paper, we have presented a system of archiving and processing oral and written dialectal corpora in a unified framework. We have described the design and development of a multimedia software and database for the preservation and processing of both oral and written data from three Greek dialects of Asia Minor, Pontic, Cappadocian and Aivaliot. Our system enables the parallel visualization of raw and annotated data at different levels, as well as the coding of a large number of metadata. The search and retrieve subsystem provides (a) a combined search at two levels of linguistic representation (phonological, morphological), (b) access in metadata and (c) combined search in both oral and written resources.



Μορφολογική Επισημείωση

ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ

Λήμμα: έρχομαι

Μορφολογική Διαδικασία: Κλίση-Ακλιτο

Γραμματική Κατηγορία: Ρήμα

Κατηγορία Λήμματος: Ελληνική

ΚΛΙΣΗ

Χρόνος: Ενεστώτας

Αριθμός: Πληθυντικός

ΜΕΤΑ-ΠΛΗΡΟΦΟΡΙΕΣ

Περιοχή: ΤΡΑΠΕΖΟΥΝΤΙΑΚΑ (Τραπεζούντα, Κερασούντα, Ριζούντα, Σούρμενα, Όφις, Αδεμά, Τρίπολις, Ματσούκα)

Φωνολογική Επισημείωση

Φαινόμενα: Ανομοίωση Συμφώνων

Figure 1. Triptych of written documents

Word/ token /phenomenon			Location				
<Value>	{ Between, And, Or, Exact }	<Value>	<EAV subschema>*	<Attribute>	<Part distances>	<Word distances>	<Interval_no distances>

Figure 2. Query template

5. ACKNOWLEDGEMENTS

We wish to thank the following members of the research team of the AMiGre project, Prof. Nikitas Karanikolas (Department of Informatics, TEI of Athens), Assoc. Prof. Dimitris Papazachariou and Ass. Prof. Thodoris Markopoulos (Department of Philology, University of Patras), Prof. Nikolaos Pantelidis (University of Athens), Dr. Io Manolessou (Academy of Athens) and Dr. Nikos Koutsoukos (post-doctoral researcher at LMGD) who have helped significantly into setting the requirements for the design of the multimedia database.

6. REFERENCES

[1] Anderson, J., Beavan, D., Kay, C. 2007. SCOTS: Scottish Corpus of Texts and Speech. In *Creating and digitalizing Language Corpora* Vol.1, Beal J. (ed.). Palgrave McMillan Publication, pp. 17-34.

[2] Anhøj, J. (2003), "Generic Design of Web-Based Clinical Databases", Journal Medical Internet Research, 4. DOI= http://dx.doi.org/10.2196/jmir.5.4.e27

[3] Barbiers, S. et al. 2006. Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute . http://www.meertens.knaw.nl/sand/

[4] Galiotou, E., Karanikolas, N.N., Manolessou, I., Pantelidis, N., Papazachariou, D., Ralli, A., Xydopoulos, G. 2014. Asia Minor Greek: Towards a Computational Processing, Procedia: Social and Behavioral Sciences, 147, Elsevier, 458-466. DOI=http://dx.doi.org/10.1016/j.sbspro.2014.07.138

[5] Karanikolas, N.N., Galiotou, E., Xydopoulos, G., Ralli, A., Athanasakos, K., Koronakis, G. 2013. Structuring a Multimedia tri-dialectal dictionary, In Proceedings of the 16th Int. Conference on Text, Speech and Dialogue TSD 2013 (Plzeň, Sept. 1–5 2013) LNCS vol. 8082, Springer, 509-518. DOI= http://dx.doi.org/10.1007/978-3-642-40585-3_64

[6] Karanikolas, N.N., Galiotou, E., Ralli, A. 2014. Towards a unified exploitation of electronic dialectal corpora: Problems and perspectives, In Proceedings of the 17th Int. Conference TSD 2014 (Brno, Sept. 8-12 2014), LNAI vol. 8 655, Springer, 257-266. DOI=http://dx.doi.org/10.1007/978-3-319-10816-2

[7] Karanikolas, N.N., Galiotou, E., Papazachariou,D,, Athanasakos, K., Koronakis, G., and Ralli, A., Towards a computational processing of oral dialectal data. PCI 2015, October 01 - 03, 2015, Athens, Greece. ACM 978-1-4503-3551-5, DOI=http://dx.doi.org/10.1145/2801948.2801966

[8] Karasimos A., Galiotou E., Karanikolas, N., Koronakis, G., Athanasakos, K., Papazachariou, D., and Ralli, A., Challenges of Annotating a Multi-Dialect, Multi-Level Corpus of Spoken and Written Modern Greek Dialects. MGDLT6: 6th International Conference on Modern Greek Dialects & Linguistic Theory, 25-28 Sept. 2014, Patras, Greece.

[9] Kunst, J.P., and Wesseling, F. 2010. Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools. In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (Nov. 4-7, 2010), Tohoku University, 759-767. http://www.aclweb.org/anthology/Y10-1088

[10] Nerbonne, J., Kleiweg, P.: Lexical distance in LAMSAS. *Computers and the Humanities*, 37(3), 339–357 (2003)

[11] Scherrer, Y. 2010. Natural Language Processing for Swiss German Dialects. In: the 55th Annual Conference of the International Linguistic Association, New Paltz, NY (USA), April 2010. http://archive-ouverte.unige.ch/unige:22810

[12] Ubul, A., Kake, H., Sakoguchi, Y., Kishie, S. 2015. Research on Oral Map in Regional Dialect Using Google Map. Int. Jour. Comp. ch. 2 (2), 31-35. http://ijcat.org/IJCAT-2015/2-2/Research-on-Oral-Map-in-Regional-Dialect-Using-Google-Map.pdf

[13] Valls, E., Nerbonne, J., Prokić, J., Wieling, M., Clua, E. and Lloret, M-R. 2012. Applying Levenshtein Distance to Catalan Dialects. A Brief Comparison of Two Dialectometric Approaches. *Verba* 39, 35-61.

[14] 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 & New York: Mouton de Gruyter. Part IV, section B.