Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 147 (2014) 458 – 466

Procedia
Social and Behavioral Sciences

ICININFO

Asia Minor Greek: Towards a Computational Processing

Eleni Galiotou[a*], Nikitas Karanikolas[a], Ioanna Manolessou[b], Nikolaos Pantelidis[b], Dimitris Papazachariou[c], Angela Ralli[c], George Xydopoulos[c]

[a]*Dept. of Informatics, Technological Educational Institute of Athens, Ag. Spyridonos, Egaleo Athens GR-122 10, Greece*
[b]*Dept. of Philology, University of Athens, Panepistimiopolis, Ilisia Athens GR-116 35, Greece*
[c]*Dept. Of Philology, University of Patras, Panepistimiopolis, Rio Patras GR-265 04, Greece*

Abstract

In this paper, we discuss issues concerning the computational aspect of an on-going research project which aims at providing a systematic study of three Greek dialects of Asia Minor ("Pontus, Cappadocia, Aivali: In search of Asia Minor Greek"-AmiGre) In fact, the project constitutes the first attempt to describe dialectal phenomena at a phonological, morphological, and structural level. Furthermore, it also constitutes the first attempt in Greece to combine Informatics and Theoretical Linguistics in order to facilitate the above-mentioned task. The aim here is to provide the design principles of the computational component of the project namely, an electronic dictionary and a multimedia database which would provide an innovative mechanism of storing, processing and retrieving oral and written dialectal data.

© 2014 Elsevier Ltd. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/3.0/).

Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
Keywords: Computational Dialectology; Asia Minor Greek; Modern Greek Dialects; Electronic Dictionaries; Multimedia Databases

1. Introduction

In recent years, research in dialectal change and language contact has come to interesting conclusions especially in the case of Germanic and Romance languages (Thomason, 2001; Matras, 2009 ; Stolz, Bakker, & Palomo (eds.), 2008). Asia Minor Greek dialects constitute a particularly interesting case in the scientific fields of dialectology and contact linguistics; although they genetically share a common Indo-European origin (Greek), they have diverged from one another partly under the influence of an Altaic language (Turkish) to such an extent that they constitute different dialects.

* Corresponding author. Tel.: +30-210-538-5824; fax: +30-210-591-0975.
E-mail address: egali@teiath.gr

1877-0428 © 2014 Elsevier Ltd. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/3.0/).
Selection and peer-review under responsibility of the 3rd International Conference on Integrated Information.
doi:10.1016/j.sbspro.2014.07.138

It is to be noted that Greek and Turkish not only belong to different language families but to different typological groups as well (fusional vs. agglutinative). Of significant importance for the documentation of Asia Minor Greek is the contribution of the Centre for Asia Minor Studies with archival and bibliographic material (Giannakopoulos, 2003). Yet, little interest has been shown in the dialects in question with the exception of certain mentions to Cappadocian such as in Thomason (2001) and Thomason & Terrence Kaufman (1988). Therefore, an analysis of Asia Minor Greek dialects would give useful insights as for the nature and mechanism of language change within the domain of dialectal variation. On another matter, the availability of dialectal data on electronic media and the development of computational tools has greatly contributed to the advancement of research in dialectology; such is the case of the Dynamic Syntactic Atlas of Dutch Dialects (DynaSAND) (Barbiers et al., 2006) an on-line tool for dialect syntax research. It consists of a database, a search engine, a cartographic component and a bibliography concerning syntactic variation found in varieties found in the Netherlands, Belgium and France. A convincing argumentation in favour of Computational Linguistics techniques in Dialectology is reported in Nerbonne, J. (2003) while in Nerbonne, J. and Kleiweg, P. (2003) the treatment of lexical variation in LAMSAS (Linguistic Atlas of the Middle and South Atlantic States) is presented. As far as Greek dialects are concerned, no results of a computational processing of dialects are reported so far, with the exception of the electronic dictionary of Cypriot Greek (Themistocleous et al., 2012). Therefore, a computational approach to the problem is a challenge, which would contribute to the development of innovative mechanisms of processing dialectal data. Up to now, a number of linguistic studies have focused on the collection and analysis of dialectal data. Yet, none had attempted to represent both raw and processed material in the digital space. The first attempt in Greece to combine Theoretical Linguistics and Informatics for a scientific presentation of dialectal data to the academia is the THALIS program "Pontus, Cappadocia, Aivali: In search of Asia Minor Greek" (AmiGre) which aims at:

- providing 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;
- digitizing, archiving and processing a wide range of oral and written data thus contributing to the sustainability and awareness of this longwinded cultural heritage.

Computational activities comprise:

- The design and development of a multimedia tri-dialectal dictionary of three Greek dialects of Asia Minor (Pontic, Cappadocian, Aivaliot). The dictionary will contain lemmata from three dialects in a comparative way.
- The design and development of a multimedia database for the archiving and processing of oral and written dialectal data
- The construction of the web site of the project where the aims, the progress and the final results will be published. (amigre.cs.teiath.gr).

In this paper, we present the design principles and the current state of development of the computational component of the project. In section 2, we present the design and implementation of the multimedia 3-dialectal dictionary. In section 3, we discuss issues concerning the design of the multimedia database. Finally, in section 4 we draw the necessary conclusions and point to future work.

2. The 3-dialectal dictionary

2.1. Linguistic principles

One of the aims of the project is to provide a multimedia tri-dialectal dictionary which will greatly improve the documentation of the three Asia Minor Greek dialects which are already in a way of extinction (Karanikolas et al. 2013). With the exception of Papadopoulos' historical dictionary of Pontic (Papadopoulos,. 1958), there are no dictionaries of these dialects – only glossaries where lemmata are stored in a very unsystematic way and crucial information such as pronunciation and usage is missing. The proposed dictionary is based on a sound linguistic analysis and in addition to other aspects provides the users the possibility to access a graphic representation of each lemma in a conventionally - adopted character set, the pronunciation, the meaning, different usages and related lemmata.

Although dialectal dictionaries are usually treated as monolingual synchronic dictionaries due the limitation of their macrostructure (Landau, 2001), our dictionary was designed as a tri-lingual one since its macrostructure is in a different system from that of microstructure (Three Asia Minor dialects vs. Standard Greek) (Béjoint, 2000; Xydopoulos & Ralli, 2013). As for the geographical and time scope, our dictionary contains information from different areas and time periods therefore, it is regarded as a local / microareal dialectal dictionary of a non-synchronic nature.

The projected macrostructure of the dictionary includes approx. 7,500 entries (approx. 2,500 entries from each of the three dialects). Vocabulary in common with Standard Greek (unless differently used) is not included in this listing, which is based on an alphabetical organization. The information of the dictionary microstructure comprises pronunciation (phonetic form), grammar (categorial and morphological information), etymology, meaning (descriptive definition and/or synonyms), usage (thematic and register labels). Moreover, a linkage to multimedia information resources is provided in order to enrich the semantics and pragmatics of the lemmata (Barbato& Varvaro, 2004; Rys. & Van Keymeulen, 2009; Xydopoulos & Ralli, 2013). In order to avoid different and arbitrary spelling codes for the same dialect, Headwords appear in a capitalized orthographic form instead of a "semi-phonetic" form (Durkin, 2010; Xydopoulos, 2011). Cross-reference to other entries is achieved through derivational processes or through semantic relations. The entries contain also authentic examples of use, which are encoded in non-standard spelling, reflecting as close as possible the pronunciation with the use of diacritics (Rys & Van Keymeulen, 2009).

2.2. Design

Following the linguistic principles in 2.1, we have designed a dialectal dictionary hosting different realizations of lemmata depending on the geographical area where the dialects are spoken. For each lemma we have taken into account the following information: headword, dialect, morphological information/process, etymology, realizations, meanings, usage examples, related lemmata. The abovementioned information is related to the following assertions:

- the primary information defining a lemma is composed of: the headword, the dialect (e.g. the dialectal region), the morphological information and the etymology.
- each different realization of a lemma is characterized by a slightly different phonetic realization depending on the microdialectal region it originates from.
- usage examples are considered essential information
- a lemma can be polysemous or homonymous to other lemmata

As for the relationship between lemmata and meanings we have proposed the following relations:

- cross-reference ("see also" link): connects lemmata which are etymologically / morphologically / semantically / pragmatically related
- synonymy / antonymy ("thesaurus" link): restricted between lemmata of the same dialect. The distinction between synonyms and antonyms is realized as an attribute of the "thesaurus" link.
- same meaning between lemmata of different dialects ("other dialect" link): Contrary to previous relations, the relation "other dialect" is an asymmetrical one.

The principles defined so far are presented in detail in Karanikolas et al. (2013).

2.3 Implementation

The dictionary was implemented in a relational database the relation schema of which contains 13 tables. Four tables are the relational equivalents of the conceptual entities "lemma", "meaning", "realization types" and "usage examples". Three of them are the relational equivalents of the conceptual relations "see also", "thesaurus", "other dialect" and the other six tables are just look-up tables. A detailed discussion on the design and implementation of the tri-dialectal dictionary appears in Karanikolas et al. (2013). The dictionary and a friendly user interface for data entry are implemented in Java and MySQL. An example of a dictionary entry is depicted in figures 1 and 2.

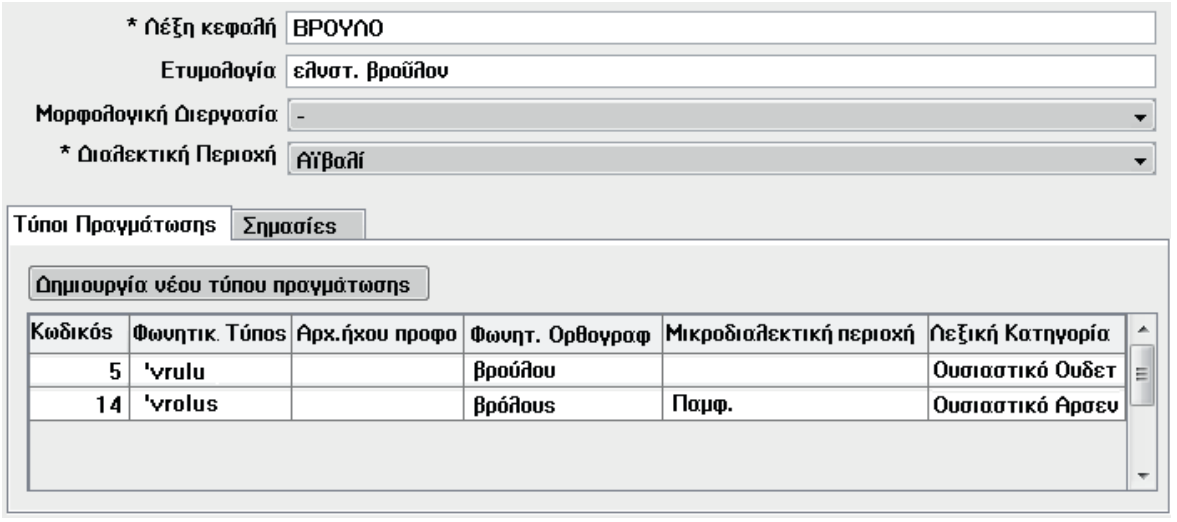

Fig. 1. Realization types of the lemma ΒΡΟΥΛΟ ('vrulo): The first one ('vrulu) is noun-neuter and the second one ('vrolus) is noun-masculine

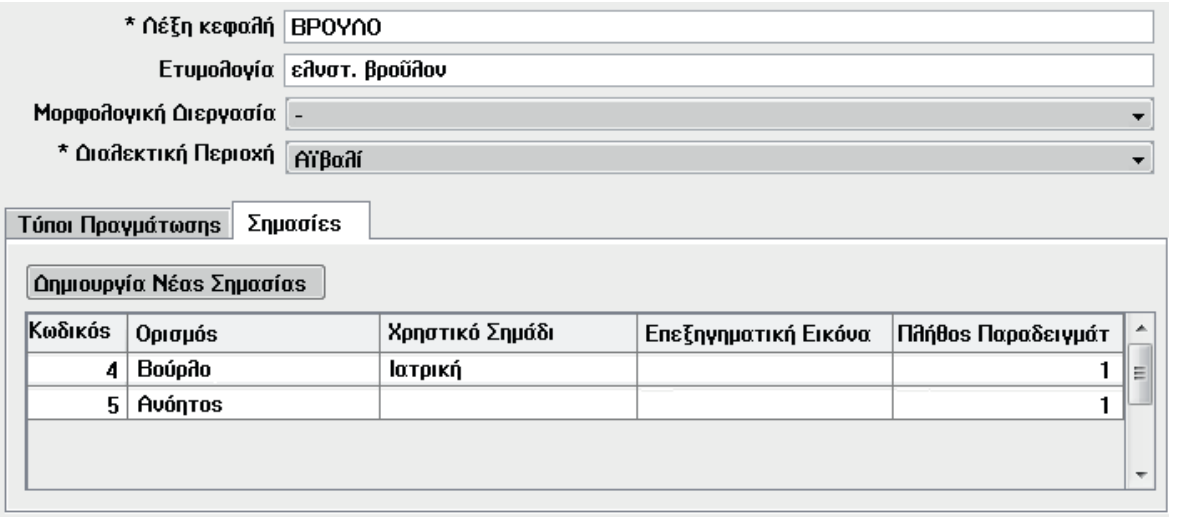

Fig. 2. Meanings of the lemma ΒΡΟΥΛΟ ('vrulo): The first one is "βούρλο (bulrush)" and the second one is ανόητος (silly)"

3. The multimedia software & tools

3.1. The oral corpus

A corpus composed of approximately 180 hours (60 hours /dialect) of recorded raw oral data is compiled. The raw data are processed according to the 3A model as proposed by Wallis & Nelson (2001) according to which data are (a)annotated, (b)abstracted, (c) analyzed. More specifically, our oral corpus is composed of:

i. Raw data accompanied by relevant metadata
ii. A multimodal corpus of approx.. 45 hours (15 hours /dialect) combining raw data with transcription, translation, annotation and metadata.

The multimodal corpus is processed with the use of the ELAN software for multimodal annotation. (ELAN, Sloetjes & Wittenburg, 2008). Further phonetic analysis of spoken data is performed with the use of Praat (Boersma 2012; Boersma & Weenink, 2013). The spoken data are annotated in relation to speakers' turn-takings. The next step is the transcription and translation of utterances within each speaker's turn. On the phonological level, we first annotate utterances. Within each utterance, intonation phrases are the next phonological unit for annotation, where tones (according to ToBI annotation system) are indicated. Phonological words, syllables and phonemes are also annotated, using IPA symbols in order to annotate each element in these three tiers. Our phonological annotation includes two more pieces of information, i.e. the phonological phenomenon that is taking place either in the phonological word, or between phonological words, and the existence of phonological variables. On the phonetic level, we annotate each one of the segments. We further use different tiers for vowels, diphthongs, consonants and consonant clusters.

3.2. The written corpus

The written corpus of the Amigre project (Koliopoulou, et al. 2013) consists of 1.000.000 words of digitized dialectal texts (in the form of jpeg images) from primary written sources of the 19th and early 20th century.: grammatical descriptions, glossaries, folklore collections of tales, proverbs, songs et sim., manuscript memoirs, word-lists etc. Inclusion of a text in the corpus was determined on the basis of criteria such as representativeness

(according to dialect, local sub-dialect and chronological period) and quality (closeness to actual spoken language, consistent linguistic terminology or transcription system). Part of this corpus (200.000 words) was further transcribed using a custom- made transcription system based on the SAMPA symbols, especially developed with the aim to:

(a) allow transcription without the use of special diacritics or keyboard configurations, thus facilitating further electronic elaboration, searches etc.

(b) represent all the special sounds included in the phonetic inventories of the dialects under investigation and

(c) unify the disparate, inconsistent and impressionistic phonetic notation of the original written sources (Manolessou, Beis & Bassea – Bezantakou, 2012).

As a final step, part of the transcribed corpus (50.000 words) is annotated with the use of a special tool developed for the purposes of this project. Linguistic annotation concerns the levels of phonology (primary division into consonants vs. vowels, sub-division into anaptyxis, deletion, change and coalescence of sounds), morphology (primary division: inflectional vs. derivational morphology, sub-division into grammatical properties like gender, number, case, person, tense etc.) and the lexicon (etymological origin of lexical items, basically native vs. borrowed, archaic vs. innovative). The linguistic annotation of the written corpus follows the same principles and categories as the annotation of the oral corpus (naturally excluding linguistic information unavailable for older written sources, such as intonation, precise phonetic realization, metadata information on the age or social class of the informant etc.), thus offering the future users of the database the possibility of unified searches across the whole available dialectal corpus. Each section of transcribed/annotated text is connected within the database with its digital image, thus allowing verification and clarification of the information provided.

3.3. The multimedia software

The aforementioned oral and written corpora will be integrated in a multimedia database for further analysis and evaluation. To this end, we have investigated the possibilities offered by well-known software such as LaBB-CAT (Fromont & Hay, 2008, LaBB-CAT), which provides the user the possibility to store audio or video recordings, text transcripts, and other annotations. Yet, such a software would not comply fully with our needs which are:

- The combination of oral and written levels
- The possibility for annotations at many different linguistic levels

Therefore, we opted for the creation of a software tailored to our needs, the architecture of which appears in figure 3. The acronyms used in fig. 3 are defined as follows:

- G. Oral = GUI for Oral sources
- G. Written = GUI for Written sources
- M. Tag = Morphological Tagger
- Syn. Tag = Syntax Tagger
- Sem. Tag = Semantics Tagger
- T. Imaging = Text Imaging
- T. Transcription = Text Transcription
- Ph. Tagging = Phonological Tagging
- EAV Database = Entity Attribute Value Database
- WID Text Files = Word Identified Text Files (a list of 2-uples: word - identifier)
- Oral = Indexing Module for Oral sources
- Written = Indexing Module for Written sources

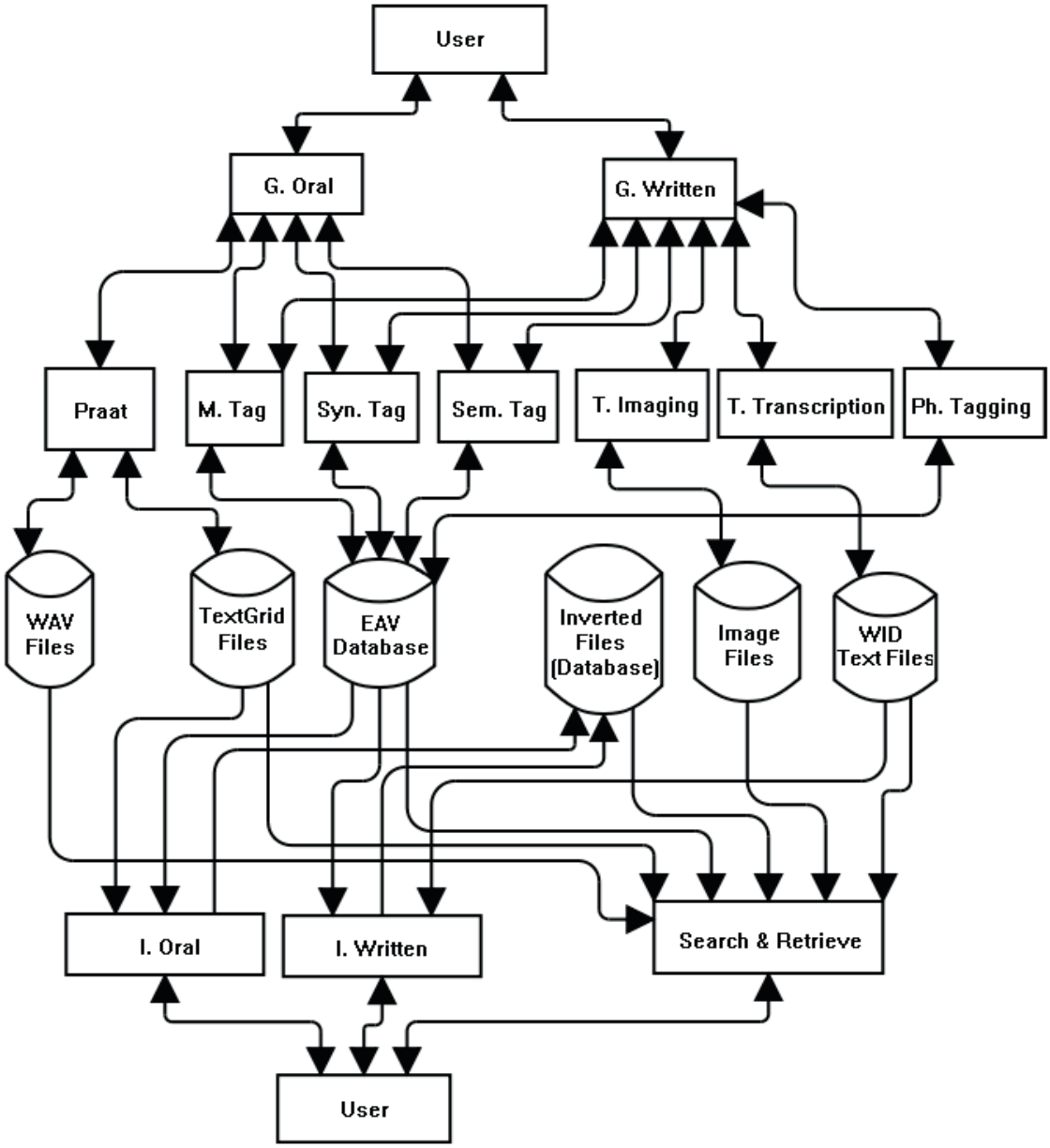

Fig. 3. The architecture of the multimedia software

The Search & Retrieve module of the software invokes the relevant application (one of "Praat", "G.Oral", "G.Written", "M.Tag", "Syn.Tag", "Sem.Tag", "T.Imaging", "T.Transcription", "Ph.Tagging") using OLE Automation (Object Linking and Embedding Automation) or other equivalent technology. The selection of the relevant application, between the available ones, is automatically determined by the Search & Retrieve module de-

pending on the types of information (WAV, TextGrid, EAV, Image, WID Text) that correspond to the criteria defined by the user. In our case, "Inverted Files" are composed of the following information:

1. a word/lemma,
2. a list of 2-uples, in the form:

- The identifier of the relevant database (or collection of files), a value from the set {WAV, TextGrid, EAV, Image, WID Text},
- The value of the primary key (or another unique name) that specifies a concrete instance (a tuple or a file) among the collection of instances in the relevant database (file collection).

4. Conclusions and future work

To this end, we have described the current state of the implementation of a multimedia 3-dialectal dictionary and the design principles of a multimedia software for the exploitation of oral and written corpora. Future activities comprise the development of an advanced retrieval component of the 3-dialectal dictionary as well as the full implementation of the multimedia software. Complementary to the aforementioned multimedia software, a number of tools that will further facilitate the linguistic tasks of the project, such as a text segmentation software or a morphological analyzer for the dialects in question, are under consideration. Such tools are of particular interest since they will rely on linguistic investigation that is still in progress.

Acknowledgements

This research is 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.

References

Barbato, M. & Varvaro A. (2004). Dialect dictionaries, Int. Journal of Lexicography, 17, 4, 429-439

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

Béjoint, H. (2000). Modern lexicography: An introduction. Oxford: Oxford University Press.

Boersma, P. (2012). The use of Praat in corpus research, in: Jacques Durand, J. Gut, U. & Kristofferson, G. (eds.): Handbook of corpus phonology. Oxford: Oxford University Press

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer. http://www.praat.org

Durkin, P. (2010). Assessing non-standard writing in lexicography. In *R. Hickey (ed.)* Varieties of English in writing: The written word as linguistic evidence, 43-60. Amsterdam: John Benjamins.

ELAN: http://tla.mpi.nl/tools/tla-tools/elan/, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

Fromont, R. & Hay, J. (2008). ONZE Miner: the development of a browser-based research tool, Corpora, 3(2), 173-193

Giannakopoulos, G. (2003). The contribution of the Centre for Minor Asia Studies to the study of Greek dialects in the Minor Asia region: archival material and bibliographic sources, in *Proc. 4th Int. Conf. on Modern Greek Dialectology*, Athens 2003, pp. 95-106. [in Greek].

Karanikolas, N., Galiotou, E., Xydopoulos, G. Ralli, A., Athanasakos, K., Koronakis G. (2013)- Structuring a Multimedia tri-dialectal dictionary, Proceedings of the 16th International Conference on Text, Speech and Dialogue (TSD 2013), September 1 – 5 2013, Plzeň, CZ, LNCS vol. 8082, 509-518, Springer.

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 [= Κολιοπούλου, Μ., Μαρκόπουλος, Θ., & Παντελίδης, Ν.(2013) : Πόντος, Καππαδοκία, Αϊβαλί: Προκλήσεις ενός ψηφιακού σώματος γραπτού υλικού].

LaBB-CAT (formerly known as ONZE Miner). http://onzeminer.sourceforge.net/

Landau, S.I. (2001) Dictionaries: The art and craft of lexicography (2nd edition).

Manolessou, I., Beis, S., & Bassea-Bezantakou (2012). The phonetic transcription of Modern Greek dialects (in Greek), Lexicographicon Deltion 26, 161-222. [= Μανωλέσσου, Ι., Μπέης, Σ., & Μπασσέα-Μπεζαντάκου, Χ. (2012). Η φωνητική απόδοση των νεοελληνικών διαλέκτων και ιδιωμάτων]

Matras, Y., (2009). Language contact. Cambridge University Press.

Nerbonne, J. (2003). Linguistic Variation and Computation, *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics*, April, 2003. 3-10.

Nerbonne, J. and Kleiweg, P. (2003). Lexical distance in LAMSAS, Computers and the Humanities 37 (3), 339-357.

Papadopoulos, A. 1958. Historical Dictionary of the Pontic Dialect. Athens: Epitropi Pontiakon Meleton. (Annex 3).

Ralli, A., Papazachariou, D. & Karasimos, A. (2010). Laboratory of Modern Greek Dialects and the project GreeD. In Ralli, A. et al. (eds.) Proceedings of the 4th International Conference of Modern Greek Dialects and Linguistic Theory

Rys, K. & J. Van Keymeulen. (2009). 'Intersystemic correspondence rules and headwords in Dutch dialect lexicography'. (2009). In International Journal of Lexicography 22, 129-150.

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)

Stolz, T., Bakker, D., & Palomo, S. (eds.), (2008). Aspects of language contact. New theoretical, methodological and empirical findings with special focus on Romanisation processes. Mouton de Gruyter.

Themistocleous, C., Katsogiannou, M., Armosti, S., Christodoulou, K. (2012). Cypriot Greek Lexicography: An Online Lexical Database. Proceedings of Euralex 2012, 889-891.

Thomason, Sarah G. (2001). Language Contact. An Introduction. Edinburgh: Edinburgh University Press.

Thomason, Sarah G. & Terrence Kaufman. (1988). Language Contact, Creolization, and Genetic Linguistics. Berkeley: University of California Press.

Wallis, S. & Nelson, G. (2001). Knowledge discovery in grammatically analyzed corpora. Data Mining & Knowledge Discovery, 5, 305:335

Xydopoulos, G. J. (2011). Metalexikografikes paratiriseis sta leksika Benardi ke Syrkou [Metalexicographic comments to Benardi and Syrkou [dialectal] dictionaries]. In Patras Working Papers in Linguistics 2.1 2011, 96-113.

Xydopoulos, G. & Ralli, A. (2013). 'Greek Dialects in Asia Minor. Setting Lexicographic Principles for a Tridialectal Dictionary'. In M. Janse, B. Joseph and A. Ralli (eds.) Proceedings of the 5th International Conference on Modern Greek Dialects and Linguistic Theory (Ghent, 20-22 September 2012).