Multi-dialectal Lexicon Building: Lessons Learned

Nikitas N. Karanikolas
nnk@teiath.gr

Eleni Galiotou
egali@teiath.gr

*Technological Educational
Institute of Athens
Department of Informatics*

ABSTRACT: In this paper, we discuss the lessons learned through the lifecycle of a dialectal electronic lexicon. Our approach is innovative because our lexicon is designed and built as a multi-dialectal (trilingual) dictionary (three dialects vs. one target language) instead of three monolingual dialectal dictionaries. Our system offers features that could not be possible with three monolingual dialectal dictionaries. Moreover, during the system's lifecycle we have got very specific demands for improvements (new requirements) that users were not able to express during the analysis phase. The lessons learned and the solutions invented for the system's ultimation (to respect the new requirements) can be helpful for other research or project with similar purposes.

KEYWORDS: Computational Dialectology, Dialectal Lexicography, Electronic Dictionaries, Lexical Resources, Modern Greek Dialects, Asia Minor Greek.

1 Introduction

In a previous work (Karanikolas et. al., 2013), we have presented the design and implementation of a multimedia electronic dictionary of three Greek dialects in Asia Minor (Pontic, Cappadocian, Aivaliot). We had presented the linguistic and lexicographic approach adopted, as well as the principles for designing the macro/microstructure of the dictionary. We also had presented the conceptual model of the tri–dialectal dictionary and the equivalent relational schema. According to the above analysis a system has

been implemented that hosts lemmas and relevant lexicographic information from three Asia Minor Greek dialects.

However, during the lifecycle of the system, and because of the highly–qualified users, we have got very specific demands for improvements and we have caught the ultimate goal (an excellent system). In this paper, we report the lessons learned through this system's lifecycle and we present the improved design and the extended facilities of the 3–dialectal dictionary. We claim that our system can be used for other Greek dialects and that our extended design can be the base for multi–dialectal dictionaries for other languages. Our ultimate system can be used for multi–dialectal dictionaries of other languages so long as other virtual keyboards can be appended to it.

The paper is organized as follows. Section 2 presents the motivation for building the 3–dialectal lexicon and relevant work is presented in section 3. Sections 4 and 5 describe the initial requirements and the design according to the requirements set. Section 6 gives some details from the first implemented version of the system. Section 7 presents the demands for improvements and the relevant implementations. The result of the improvements is an excellent system and the design of this system is the topic of section 8 while conclusions are drawn in section 9.

2 Motivation

Pontic, Cappadocian and Aivaliot are three Greek dialects in Asia Minor which are not sufficiently documented and they are on the way to extinction. Until now, little interest has been shown in the dialects in question. The most interesting exception is the Papadopoulos' historical dictionary of Pontic (Papadopoulos, 1958). We can also find mentions to Cappadocian in some other works (Thomason, 2001; Thomason and Kaufman, 1988). There are also some glossaries for the Asia Minor Greek dialects containing words and idiomatic phrases accompanied by their meaning in Standard Modern Greek. However, in most of these glossaries, lemmas are stored in a very unsystematic way and crucial information, such as pronunciation or usages, is missing. Moreover, some verbs are listed in their past tense form while others appear in the present tense. Therefore, a sound linguistic analysis of Asia Minor Greek dialects is indispensable and gives insights as for the nature and mechanism of language change within the domain of dialectal variation.

This and other relevant social speculations (syllogisms) motivated us for the initiation of the AMiGre project, within the framework of THALIS pro-

gram. The project acronym (AMiGre) comes from the project's title: "Pontus, Cappadocia, Aivali: in search of Asia Minor Greek". One of the deliverables of the AMiGre project was the design and implementation of a multimedia tri–dialectal dictionary for three Greek dialects in Asia Minor (Pontic, Cappadocian, Aivaliot), which we discuss in this paper.

Dialectal dictionaries are usually treated as monolingual synchronic dictionaries. In our case (AMiGre), instead of creating three monolingual dialectal dictionaries, we have decided to treat and design a trilingual dictionary (three Asia Minor dialects vs. Standard Modern Greek). This is the most interesting technical motivation. It is also an interesting innovation because, it is permitting cross–reference links from lemma to lemma (of the same or different dialect) and equivalence links between meanings of lemmas from different dialects. These could not happen with three monolingual dialectal dictionaries.

3   Relevant Work

Electronic lexicography for Modern Greek was not concerned with the creation of dialectal dictionaries until very recently. The online dictionaries developed at the Portal for the Greek Language (Online, 2016) comprise the computerized versions of Georgacas' Greek–English Dictionary, Triandafyllides' Dictionary of Standard Modern Greek and Anastasiadi–Symeonidi's Reverse Dictionary. In addition, the Portal provides access to the computerised version of Kriaras' Concise Dictionary of Medieval Vulgar Greek Literature. The Institute for Language and Speech Processing has developed online bilingual dictionaries (Greek–English, Greek–German, Greek–Russian, Greek–Turkish, and Greek–Arabic). The dictionaries are under continuous development and enhancement and they are available from (ILSP, 2016). In addition, NLP tools for supporting lexicographic applications have been developed. Indicatively, in (Tsalidis et. al., 2010) infrastructure tools which are used for encoding morphological, syntactic and semantic information are reported as well as proofing tools such as a spelling checker, a hyphenator etc. As far as Greek dialects are concerned, the only computerized dictionary to our knowledge is the online lexical database of Cypriot Greek (Themistocleous, 2012). The online dictionary environment provides an enhanced searching mechanism as well as text to speech features for the pronunciation of Cypriot Greek words.

4 Initial Requirements

Dialectal dictionaries are usually treated as monolingual synchronic dictionaries (Béjoint, 2000; Geeraerts, 1989), due to limits in macrostructure (overall organizational scheme of lemmas) (Landau, 2001; Zgusta, 1971). Given that our purpose was to design and build an online dictionary, its macrostructure will not be restricted by physical constraints (limitations existing for print dictionaries), and could offer (virtually) "multiple macrostructures" mirroring the various searching options that we could build (Burke, 2003). Therefore, since there were no limits in macrostructure, we have decided to design and build a trilingual dictionary (Three Asia Minor dialects vs. Standard Modern Greek) (Xydopoulos and Ralli, 2012), instead of three monolingual dialectal dictionaries. The dictionary is named TDGDAM (Tri–Dialectal Greek Dictionary of Asia Minor) and it aims to be a linguistically–sound tri–dialectal dictionary in electronic form. One basic requirement of TDGDAM was that users should have access to a graphic (form based) representation of each lemma permitting them to handle pronunciation, meaning, usages and relations with other lemmas. The representation should be editable and for this to be possible, conventionally–adopted character sets should be used. Among other things, each lemma should contain the dialectal area and the source from which the lemma has been extracted. This type of dictionary constitutes an innovation not only for the Greek language and its dialects, but also for the international standards, as will be explained below.

Regarding its geographic and time scope, TDGDAM was designed to be a local/ microareal dialectal dictionary of non–synchronic nature that should include entries from different areas and time periods (Penhallurik, 2009). As it was decided from the beginning, the lemmas of TDGDAM should be drawn (directly or indirectly) from oral speech and written material of the particular dialectal varieties (Keymeulen, 2010).

Regarding TDGDAM's microstructure, our aim is to include formal information about pronunciation (phonetic form), grammar (categorial and morphological information), origin (etymology), meaning (synonymic and/or descriptive definitions), usage (thematic and register labels) and to provide linked multimedia resources (internal or external to TDGDAM) to enrich the semantics and pragmatics of lemmas (Barbato and Varvaro, 2004; Rys and Keymeulen, 2009; Xydopoulos and Ralli, 2012). To avoid different and arbitrary spelling codes for the same dialect (Durkin, 2010; Xydopoulos, 2012), headwords do not appear in a "semi–phonetic" transcription but in

(capitalized) orthographic form. In particular, the capitalized orthographic form departs from the spelling form in the standard dialect; it does not prescribe spelling rules in the dialect and allows for any alternative orthographic forms to appear in microstructure (Markus and Heuberger, 2007; Xydopoulos, 2012). Finally, authentic examples of use were considered as essential constituent information in entries which will appear in non–standard spelling, reflecting pronunciation as closely as possible with the use of diacritics, but avoiding a "semi–phonetic" transcription (Rys and Keymeulen, 2009).

Regarding the abilities for cross linking between items of the TDG-DAM, we have defined 3 necessities: Cross–reference to other entries, related either through derivational processes or through semantic relations; Equivalence links between meanings of lemmas from different dialects; Synonymic/Antonymic relations.

The following 3 figures (figures 1, 2 and 3) present draft structural depictions of an equivalent number of lemmas that TDGDAM should contain. Based on these and other similar draft structural depictions we designed the TDGDAM system.

The terms synonymy (Synonym, 2016) and antonymy (Opposite, 2016) used previously and the terms homonymy (Homonym, 2016) and polysemy (Polysemy, 2016) that will be used later are very well defined. Their definitions are available on the internet.

5 Design

Based on the analysis presented in the requirements section and the draft structural depictions (see figures 1, 2 and 3) the following structure of lemmas is the result:

- – Headword, dialect (dialectal region), morphological information/process and etymology are primary information with single values that together define and are dependent on the lemma.
- – Each lemma can have many different realizations and each one of them is characterized by a slightly different phonetic realization dependent on the micro–dialectal region it originates from (the specific area within the wider dialectal region where the lemma's realization occurs).
- – Each lemma can possibly have different meanings (i.e. polysemy), or be homonymous with other, semantically distinct, lemmas.

ΠΕΔΙΟ (FIELD)	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (LEMMA VALUES)
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΒΡΟΥΛΟ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ (lexical category)	(O. ουδ.) (noun, neuter)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ (phonetic type)	['vrulu]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ (digital record)	WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ (alternative types)	Βρόλους ['vrolus] (Παμφ. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ ΣΥΝΔΕΣΗ) (Pamphylia microdialectal region)
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (dialectal region)	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (microdialect)	
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ (morphological process)	-
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ (usage)	1. ΦΥΤΟΛΟΓΙΑ (phytology)
10. ΟΡΙΣΜΟΣ (definition)	Βούρλο (bulrush – reedy plant)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ (usage example)	«Έκουψα καμπόσα βρούλα κι πέρασα αρμαθιά τα ψάρια πό πιασα σήμιρα» (a dialectal sentence using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ (modern Greek translation)	('Εκοψα μερικά βούρλα και ....) (the same sentence in Modern Greek)
14. ΘΗΣΑΥΡΟΣ (thesaurus)	
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (etymology)	[ΕΤΥΜ ελνστ. βρούλον] (originates from the Hellenistic βροῦλον)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	2. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ	Ανόητος (fatuous)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	«Ντιπ για ντιπ βρούλου τούτου του πιδί». (a dialectal sentence using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Τελείως ανόητο αυτό το παιδί) (the same sentence in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	---

Figure 1. Draft structural depiction of lemma ΒΡΟΥΛΟ

ΠΕΔΙΟ (FIELD)	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΛΙΩΣΤΡΑ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ (lexical category)	O. Θηλ. (noun, feminine)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ (phonetic type)	['λostra]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ (digital record)	αρχείο WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ (alternative types)	
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (dialectal region)	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ (microdialectal region)	
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ (morphological process)	Παραγωγή (derivation)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ (usage)	1. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ (definition)	γυναίκα που περιφέρεται εδώ κι εκεί (woman who strolls)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ (picture)	-
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ (usage example)	«Ξιπόρτσι πάλ'-η λ'ώστρα» (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ (Modern Greek translation)	(Ξεπόρτισσε πάλι η γυρίστρα) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ (thesaurus)	Συν: *αλλουγυρίστρα, σόρτα, τακιού* (Synonymy: *αλλουγυρίστρα, σόρτα, τακιού*)
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ (etymology)	[<*λιέμι* με -ωστρα ίσως από επιδρ. άλλων θηλυκών σε -ωστρα] (*λιέμι* with affix ώστρα)
16. ΔΙΑΠΑΡΑΠΟΜΠΕΣ (see also)	

Figure 2. Draft structural depiction of lemma ΛΙΩΣΤΡΑ

ΠΕΔΙΟ	ΛΗΜΜΑΤΙΚΗ ΠΛΗΡΟΦΟΡΙΑ
1. ΛΕΞΗ-ΚΕΦΑΛΗ / HEADWORD	ΑΛΛΟΥΓΥΡΙΣΤΡΑ
2. ΛΕΞΙΚΗ ΚΑΤΗΓΟΡΙΑ	(Ον. Θηλ.) (noun, feminine)
3. ΦΩΝΗΤΙΚΟΣ ΤΥΠΟΣ	[aluji'ristra]
4. ΑΡΧΕΙΟ ΗΧΟΥ ΠΡΟΦΟΡΑΣ	WAV
5. ΕΝΑΛΛΑΚΤΙΚΟΙ ΤΥΠΟΙ	---
6. ΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ	Αϊβαλί (Aivali)
7. ΜΙΚΡΟΔΙΑΛΕΚΤΙΚΗ ΠΕΡΙΟΧΗ	---
8. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΕΡΓΑΣΙΑ	Σύνθετο (composite)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	1. ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)
10. ΟΡΙΣΜΟΣ	γυναίκα που περιφέρεται εδώ κι εκεί (woman who strolls)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	---
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	"Ξιπόρτσι πάλ'ι η γ'αλλουγυρίστρα" (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Πάλι βγήκε η αλλουγυρίστρα.) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	ΣΥΝ *λυώστρα, σόρτα* (Synonymy: *λυώστρα, σόρτα*)
9. ΧΡΗΣΤΙΚΟ ΣΗΜΑΔΙ	2. ΙΑΤΡΙΚΗ (medicine)
10. ΟΡΙΣΜΟΣ	πόνος με πρήξιμο γύρω από το νύχι (pain & turgescence around nail)
11. ΑΡΧΕΙΟ ΕΠΕΞΗΓΗΜΑΤΙΚΗΣ ΕΙΚΟΝΑΣ	JPG
12. ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ	"Έχου μνιαν αλλουγυρίστρα στου δαχτύλ'ιμ τσι πουν'εί" (dialectal phrase using the lemma)
13. ΜΕΤΑΦΡΑΣΗ ΠΑΡΑΔΕΙΓΜΑΤΟΣ ΣΤΗΝ ΚΝΕ	(Έχω μια αλλουγυρίστρα στο δάχτυλο και με πονάει.) (the same in Modern Greek)
14. ΘΗΣΑΥΡΟΣ	---
15. ΕΤΥΜΟΛΟΓΙΚΗ ΠΛΗΡΟΦΟΡΙΑ	[ΕΤΥΜ από το ρ. *αλλουγυρίζω*] (from the verb *αλλουγυρίζω*)
16. ΔΙΑΠΑΡΑΠΟΜΠΕΣ	---

Figure 3. Draft structural depiction of lemma ΑΛΛΟΥΓΥΡΙΣΤΡΑ

– For each meaning, different usage examples are essential.

Regarding the relations (between lemmas and meanings of lemmas) we concluded the following:

– Cross reference ("See also") links can be available for connecting lemmas that are semantically / pragmatically / morphologically / etymologically related to each other.
– Synonyms and Antonyms are two semantic relations that apply between lemmas. Both relations relate a lemma meaning with a lemma (the referenced one). Synonym and Antonym links are restricted between a lemma meaning and a lemma from the same dialect.
– There are meanings of different lemmas from different dialects that share the same definition. This relation is labeled "Other Dialect". In contrast with the rest of the relations, "Other Dialect" is a symmetrical relation.

The overall idea (lemma structure and relations) is strictly defined as it is depicted with the Entity Relation Diagram of figure 4.

The following four data dictionaries (tables 1, 2, 3, 4) explain the four sections (sub–schemas) of the overall ERD.
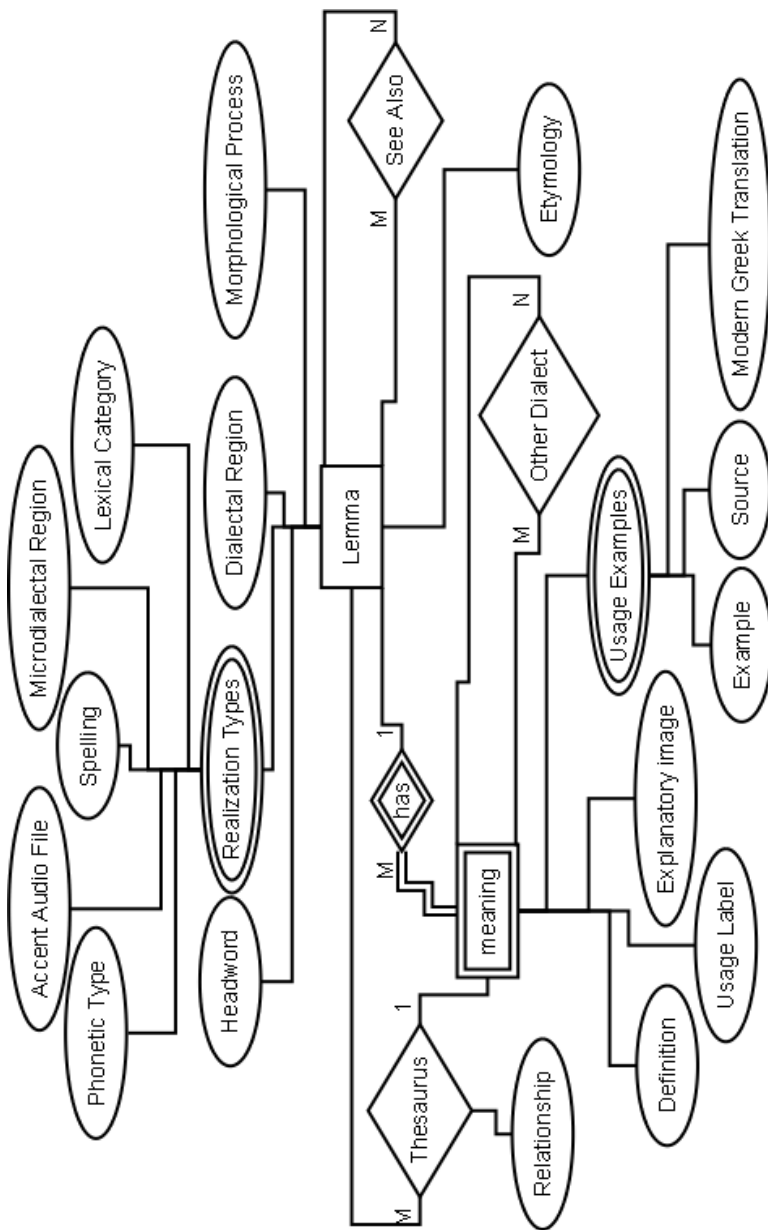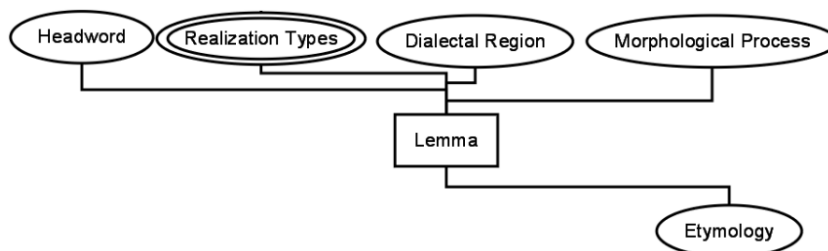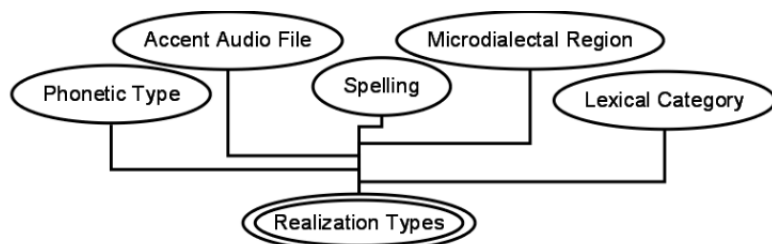
Figure 4. Entity Relation Diagram

Attribute	Definition	Data format	Example
Headword	The canonical form of the word	String containing only capital letters of the Greek alphabet	ΑΛΛΟΥΓΥΡΙΣΤΡΑ
Etymology	Basic information about the origin of the word.	String written in Greek with accents (polytonal)	Από το ρήμα αλλουγυρίζου (from the verb aluji'rizu)
Morphological Process	Different processes involved in word–formation.	A value from a predefined list of morphological processes	Σύνθετο (Compound noun)
Dialectal Region	The region/dialect in which the lemma is found	A value from a predefined list of Dialects	Αϊβαλί (AIVALI)

Table 1. Data dictionary for "Lemma"

One possible implementation of the conceptual model (ERD) using a relational database is depicted in the relational schema of figure 5 which contains thirteen tables. However, only seven tables are important. The other six tables are lookup tables (listing the set of available values existing) related to some fields of the important tables. The important tables are highlighted (in figure 5) with thicker border and larger font in their title. Four out of seven important tables are the relational equivalents of the main conceptual entity ("Lemma"), the weak entity ("meaning") and the two multiple-valued com-

sub Attribute	Definition	Data format	Example
Phonetic Type	Phonetic transcription of (the examined) pronunciation of the word.	String containing letters of the International Phonetic Alphabet (IPA).	aluji'ristra
Accent Audio	Audio file of the authentic pronunciation of the word	String containing a file path	http://amigre. gr/xyzR1.wav
Spelling	Non–standard graphic representation of pronunciation according to the orthographic rules of Standard Greek, combined with diacritics to annotate any phonological alternations.	String containing the letters of the Greek alphabet and other diacritic symbols (accent, hyphens, parentheses and apostrophes)	αλλουγυρίστρα
Microdialectical Region	Name of a specific area within the wider dialectal region of the lemma in which the realization form is found	Value from a predefined list of microdialectical regions	
Lexical Category	Part of Speech & Gender	Value from a predefined list of lexical categories	Ουσιαστικό Θηλυκό (noun feminine)

Table 2. Data dictionary for "Realization Types"

posite attributes ("Realization Types" and "Usage Examples"). The remaining three important tables are the relational equivalents of the conceptual relations ("See Also", "Thesaurus" and "Other Dialect").
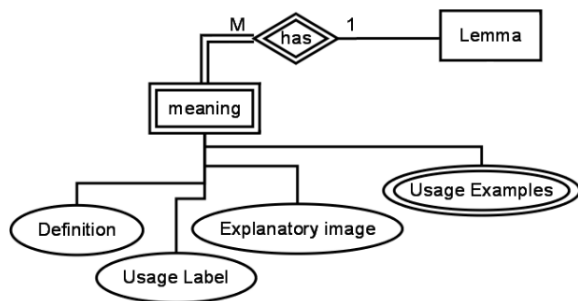
Attribute	Definition	Data format	Example
Definition	Short description of the meaning of a lemma	String in StandardModern Greek	Γυναίκα που περιφέρεται εδώ κι εκεί ("woman who goes around")
Explanatory Image File	Image illustrating the meaning of a lemma.	String containing a file path	http://amigre.gr/xyzM1.png
Usage Label	Formal indication of the context (stylistic/register/other) in which the lemma is used.	A value from a predefined list of domains	ΥΠΟΤΙΜΗΤΙΚΟ (pejorative)

Table 3. Data dictionary for "Meaning"

Only the table MeaningSets (the implementation of the conceptual relation "Other Dialect") needs more explanation. This relation is symmetrical by nature, i.e. whenever a meaning of a certain lemma from one dialect is declared as being the equivalent of the meaning of another lemma from a different dialect, then the reverse is implied. It is the structure of table MeaningSets and the application's logic that assures this symmetry. The other two relations ("See Also" and "Thesaurus") are not symmetrical by nature. This is

reflected in the relational schema (and the application logic). Consequently, the user must define the relation in both directions, in case an instance of them (the "See Also" or the "Thesaurus" relation) is symmetrical.

sub Attribute	Definition	Data format	Example
Usage example	Example (phrase or sentence) demonstrating the usage of the lemma under one specific meaning, in the original dialect	The whole example (the whole string) is written with the letters of the Greek alphabet and other diacritic symbols (accent, hyphens, parentheses and apostrophes)	Ξιπόρτσι πάλ'–η– γ'–αλλουγυρίστρα
Standard Modern Greek Translation	Translation of the usage example into Standard Modern Greek	String in Standard Modern Greek	Πάλι βγήκε η αλλουγυρίστρα
Source	Reference to the source from which the usage example was extracted	String (can be a book, a URL, etc)	

Table 4. Data dictionary for "Usage Examples"

The International Phonetic Alphabet (IPA) which is used in Table 2 – sub–schema for "Realization Types" – is an alphabetic system of phonetic notation based primarily on the Latin alphabet. It was devised by the International Phonetic Association as a standardized representation of the sounds of spoken language. The IPA is used by lexicographers, foreign language students and teachers, linguists, speech–language pathologists, singers, actors,

constructed language creators, and translators. Figures 6 and 7 present the most useful IPA charts.

Another approach to phonetic notation is SAMPA (Speech Assessment Methods Phonetic Alphabet) and it is a machine–readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987–89. It applied first to Danish, Dutch, English, French, German, and Italian (1989). Later, it applied to Norwegian and Swedish (1992). Subsequently it applied to Greek, Portuguese, and Spanish (1993). It has now been extended to Bulgarian, Estonian, Hungarian, Polish, and Romanian (1996).

6   Implementation

The GUI version of the system is based on two forms: "main form" and "meaning form". Figure 8 presents the main form for the lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ". The main form is divided into 3 sections. The upper section provides information on the headword, etymology, morphological process and dialect. The middle section is a two–card panel. The first card in the panel is used for displaying and editing realizations, while the second one is used for providing the meanings list of lemmas. The lower section of form is a panel for hosting the "see also" reference list. A more detailed description of main form's middle section is provided in figure 9 which depicts the second card (meanings list) for the same lemma.

The "meaning form" of a lemma is invoked by an action button once the user selects an item from the "meanings list" of "main form". Figure 10 depicts the "meaning form" presenting certain meanings of the lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ". The meaning form is divided into 3 sections. The upper section provides the definition of the meaning, optionally a picture and the usage label. The middle section of the form is a panel for hosting the "usage examples" list. The lower section of the form is a two-card panel. The first card in the panel is used for displaying and editing synonymic/antonymic relations (thesaurus), while the second one is used for handling the equivalents in other dialects.
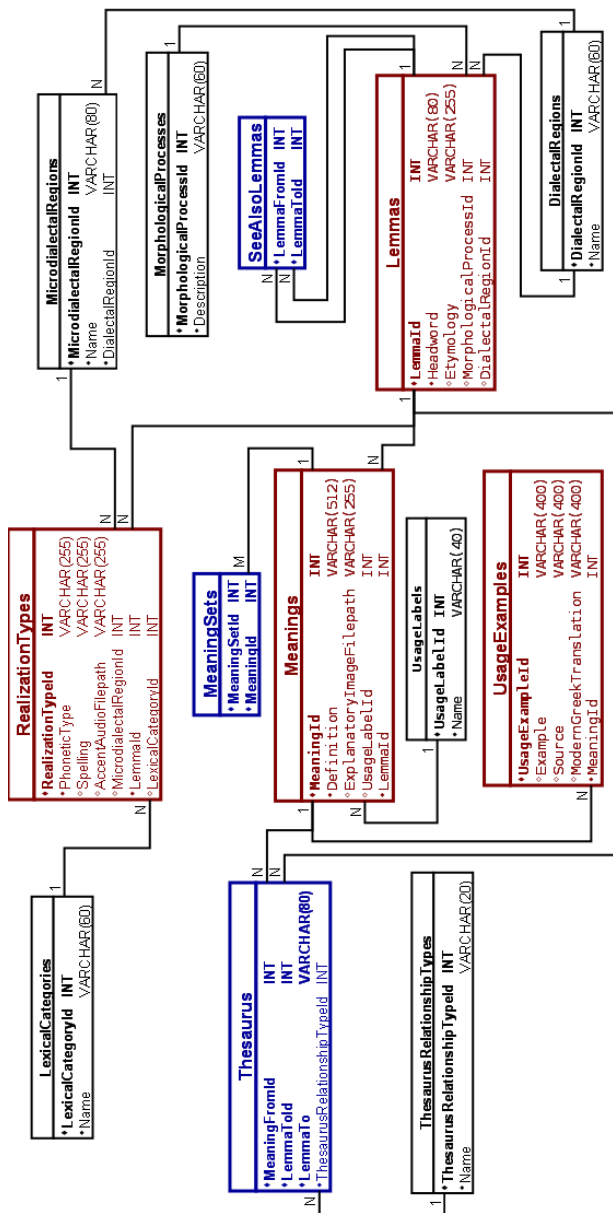
Figure 5. Logical Relational Schema

CONSONANTS (PULMONIC)   © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	ʙ			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	L			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 6. International Phonetic Alphabet (rev. 2005) – Consonants (Pulmonic)

VOWELS
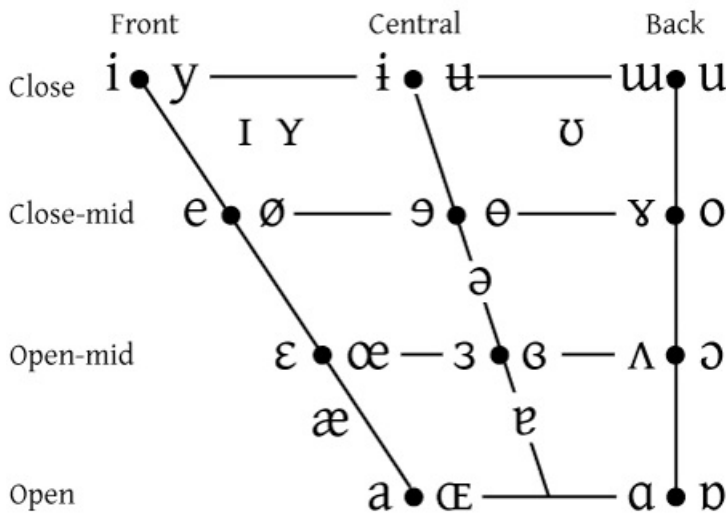
Where symbols appear in pairs, the one to the right represents a rounded vowel

Figure 7. International Phonetic Alphabet – Vowels

Figure 8. Main form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Realizations card in front

Figure 9. Main form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Meanings card
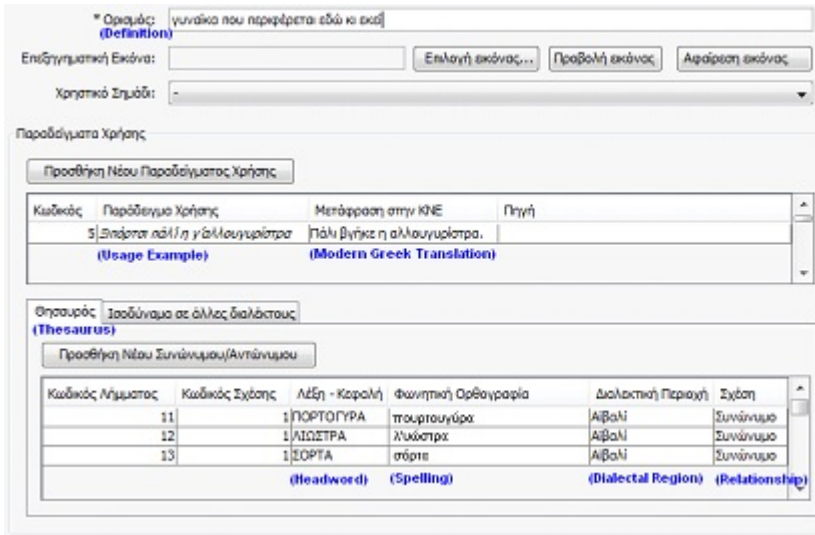
Figure 10. Meaning form of lemma "ΑΛΛΟΥΓΥΡΙΣΤΡΑ" – Some meaning with its thesaurus

According to Analysis (Requirements) and Design sections, the implemented system provides the user with the following character sets for editing the relevant fields:

- Etymology
    - Greek Polytonal
    - Loan characters from other alphabets in case of loan words (e.g. characters from the Turkish alphabet)
- Phonetic type
    - IPA
- Spelling (Phonetic Orthography)
    - Modern Greek
    - Accents
    - Hyphen, parentheses, apostrophe.

7   Ultimation

1. As it is well known, users express most of their arguments during the final stage of lifecycle of the initial development of the system (**acceptance,**

installation, deployment) and during the maintenance stage. There was no exception to this rule in the electronic lexicon. In our case users explained that the origin (etymology) attribute of a lemma, should be denoted with respect to the sources and the conventions of the originating language (from where the lemma comes). Therefore, in the case of a multi–dialectal dictionary, the etymology attribute can contain words from any of the languages that have influenced the dialect. Consequently, we came up with a solution which was to provide (in a latter development) visual keyboards for any affecting language. In the case of the 3 dialects of AMiGre the influencing alphabets are Greek, Ancient Greek and Turkish. Figure 11 presents a 3–card virtual keyboard with cards for lowercase ancient Greek, uppercase Ancient Greek and Turkish. These, together with Modern Greek (provided by the physical keyboard), permitted users to enter the required etymology of each lemma without any restriction.
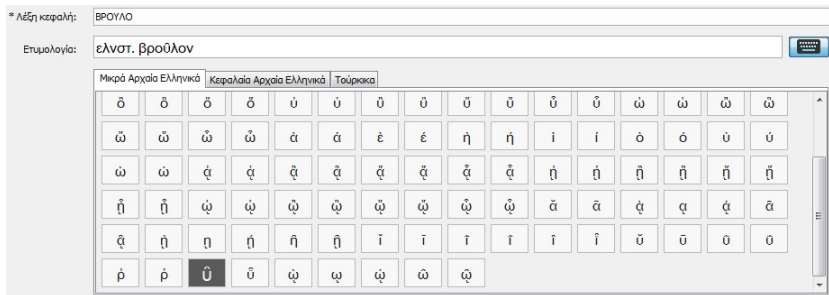
Figure 11. Virtual keyboard for Ancient Greek and Turkish

2. Regarding the phonetic type attribute and the intonation of vowels there are two options available. The first option is to place an accent before stressed syllable. The second option is to use the stressed version of the vowel. Both options are semantically equivalent. However, the first option is easier to use and implement in a system because there is no need to support both versions (with and without accent) of the IPA

symbols used for the vowels. Thus, in the initial development, our system permitted only the vertical accent before the syllable for denoting the intonated vowel. Therefore, in figure 8 the phonetic type is denoted with "aluji'ristra". After the initial development and during the maintenance stage a need to support accented versions of vowels came up. To cope with this demand, we have extended the system with virtual keyboards for easily inserting any IPA symbol, with and without accents. Figures 12, 13, 14 and 15 present the virtual keyboards used for the phonetic type attribute.

Figure 12. Alphabetic (IPA) symbols

The combinations of IPA symbols (fig. 12) with some of the Diacritic symbols (fig. 13) produce the accented IPA symbols. Figure 16 present another lemma that has phonetic type with accented IPA symbols.

Figure 13. Diacritic symbols

3. According to the linguistic analysis, the "spelling" attribute represents a non–standard graphic representation of pronunciation according to the orthographic rules of Modern Greek (target language), combined with diacritics to annotate any phonological alternations. Therefore, we concluded that the domain of values for the spelling attribute can be strings containing the letters of the Greek alphabet and definite diacritic symbols (accent, hyphens, parentheses and apostrophes). This conclusion was followed for the first implemented version of the system. Since the system got into the production stage, we became recipients of very specific demands for improvements in the spelling attribute. The users pointed out that some words, since they originate from other languages, have vowels that do not exist in the standard spelling system of the target language (Modern Greek). Since the studied dialects have a prominent number of loan words originated from other languages, this remark could not be ignored. So, we had to provide some way for the users of system to be able to understand how to pronounce the dialectal words, without engaging them to read phonetic (IPA) symbols. The simplest way was to allow the insertion of the grapheme representations (letters) used in the originating language of the loan words for the representation of vowels that do not exist in the target language (Modern Greek). The outcome was a small number of vowels with grapheme representations having umlauts (u, e and α, with umlaut). These letters were included in one virtual keyboard added for editing the spelling (phonetic transcription) attribute. Figure 17 represents this virtual keyboard.

Figure 14. Compound symbols

4. As we have already described, each lemma can have many different realizations and each one of them is characterized by a slightly different phonetic realization dependent on the micro–dialectal region it originates.

This principle drove our design and we built a system where each realization is characterized by one micro–dialectal area (sub–area of the wider dialectal region). However, in the production stage of the system, users pointed out that a single realization can exist in more than one micro–dialectal regions (of the dialectal region). To comply with this lately defined requirement we modified the data schema design and the application. The "Microdialectal region" attribute was changed to become multi–valued and the corresponding GUI item changed to hold a list of values (the domain of each value is the set of micro–dialectal regions existing for the dialect of lemma). Figure 18 presents the realizations of the Pontic lemma "OMMATOTZATZI" where the third realization (third line) has 3 micro–dialectal regions (Τραπεζούντα, Χαλδία, Σάντα).
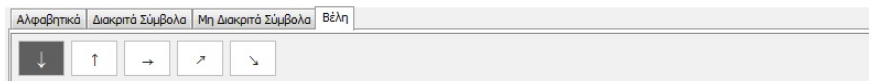
Figure 15. Arrow symbols

5. Another feature of the system which was not defined in the analysis phase but emerged during the development phase was the content of a "see also" reference list item for a destination lemma having more than one meaning and/or having more than one realization type. The solution we decided to follow was to represent the definition of each meaning for the referenced lemma and the spelling of each realization for the referenced lemma in the "see also" list item. The lower section of figure 16 represents the empty reference list of a lemma but we can see the plural number used in titles of the relevant columns (Spellings and Meanings).

6. Another worth–mentioning feature of the system is that Synonymic/Antonymic links can refer to another lemma of the same dialect which may or may not be present in the system. Figure 19 represents a meaning of lemma "ΛΙΩΣΤΡΑ" with its "thesaurus" (table providing

Figure 16. Main form of lemma "ΛΙΩΣΤΡΑ" with accented IPA symbols in phonetic type

the Synonymy/Antonymy links). In this figure, we can see 3 links (referring lemmas ΣΟΡΤΑ, ΤΑΚΙΟΥ and ΑΛΛΟΥΓΥΡΙΣΤΡΑ) but only the third one is present in the system. A comparison of figures 10 and 19 denotes that in the thesaurus table of the newer version (lower section of figure 19) we have replaced Spelling with Etymology. Note that we have also added a new column (Source type) in the usage examples table (middle section of figure 19).

Figure 17. Virtual keyboard for graphemes representing additional vowels

7. So far, we haven't seen any example of equivalent in another dialect. Figure 20 is the main form of the Cappadocian lemma "ANTET" which has a single meaning. In figure 21 we provide the lower section of the meaning form displaying the "equivalent in other dialects" card. As depicted in figure 21 the only available meaning of the Cappadocian lemma "ANTET" has an equivalent meaning in the lemma Aivaliot lemma "ANTETI".

Figure 18. Realizations of Pontic lemma "OMMATOTZATZI"

8. During the production stage of the system's lifecycle we have noticed that users diverged from the regulations for writing the usage examples of lemmas. Usually users exploited the copy/paste feature of the operating system in order to enter characters not provided directly by the applications for the "usage example" attribute. For example, the value "Αςλαγεύω το μήλον" was entered in the usage example attribute of the lemma "ΑΣΛΑΕΥΩ". This value has Greek characters that are according to the regulations but also contains a Turkish character (the second character in the string value). The phrase "Αςλαγεύω το μήλον" as value in the usage example, together with the value "τουρκ. AŞlamak" (i.e. "from Turkish aŞlamak") in the etymology attribute of the same lemma can be an indication of how native speakers of the dialect could possibly write the dialectal word in their documents ("Αςλαγεύω"). However, this indication is hidden inside one of the meanings of a lemma. We suppose that it could be better to provide another attribute (named "indicative

Figure 19. Meaning form of lemma "ΛΙΩΣΤΡΑ" – A meaning and its thesaurus

writing") in each realization of the lemma. In this way, the "indicative writing" would be directly available in the main form of lemmas and moreover it would be differentiated in each realization (micro–dialectal regions). This is the only feature that is not implemented in the system's ultimation because it is denoted very late, but we consider it very valuable for next multi–dialectal lexicons.

8 Extended – Improved design

The data schema (ERD) for supporting the ultimate system is given in figure 22.

9 Conclusions

TDGDAM's projected macrostructure includes ca. 2,500 entries from each of the three dialects of Asia Minor Greek (a total of ca.7,500 entries). These entries are drawn from collected vocabulary solely from the three dialects concerned and exclude all vocabulary found in Standard Greek (unless

differently used). Their listing is based on alphabetical, and not onomasiological, organization, accessed via dynamic searching options (Xydopoulos, 2012).

Figure 20. Main form of Cappadocian lemma "ANTET"

Figure 21. Meaning form of lemma "ANTET" – Equivalents in other dialects card in front
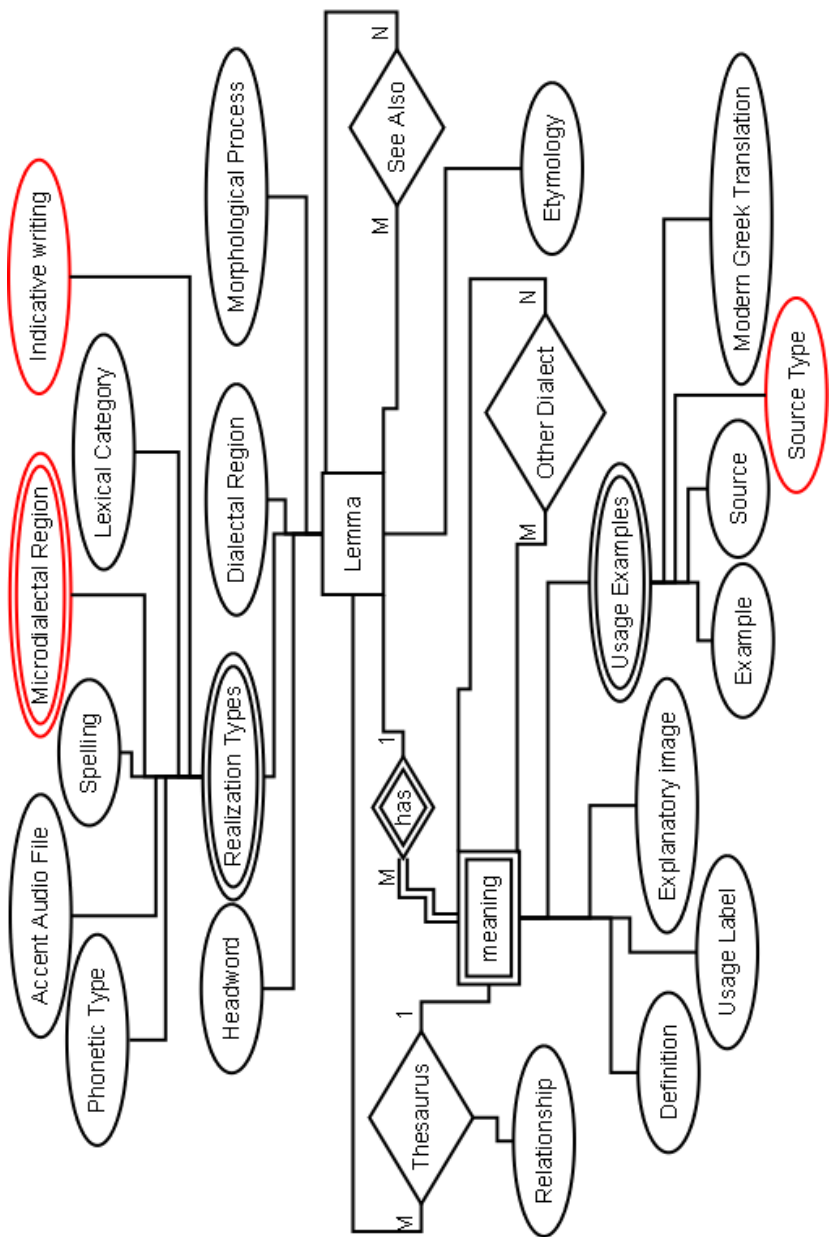
Acknowledgments

Figure 22. The improved ERD

Ralli, Professor of Linguistics at the Department of Philology of the University of Patras, who is the Coordinator of the whole AMiGre project. We also thank George J. Xydopoulos, Associate Professor of Linguistics at the Department of Philology of the University of Patras, who set the linguistic requirements for the dialectal electronic lexicon.

References

"The Dictionaries of ILSP". Institute for Language and Speech Processing. Accessed December 22, 2016. http://www.xanthi.ilsp.gr/dictionaries/

"Homonym". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Homonym

"Online Dictionaries". Portal for the Greek Language. Приступљено 22.12.2016. http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/index.html

"Opposite (semantics)". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Opposite_(semantics)

"Polysemy". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Polysemy

"Synonym". Wikipedia. Accessed December 22, 2016. https://en.wikipedia.org/wiki/Synonym

Barbato, Marcello and Alberto Varvaro. "Dialect dictionaries". *International Journal of Lexicography* Vol. 17 no. 4 (2004): 429–439

Béjoint, Henri. *Modern lexicography: An introduction.* Oxford: Oxford University Press, 2000

Burke, Sean. M. "The Design of Online Lexicons". In *A practical guide to lexicography*, Piet van Sterkenburg, 240–249. Amsterdam: John Benjamins, 2003.

Durkin, Philip. "Assessing Non-standard Writing in Lexicography". In *Varieties of English in writing: The written word as linguistic evidence*, Raymond Hickey, 43–60. Amsterdam: John Benjamins, 2010.

Geeraerts, D. "Principles in Monolingual Lexicography". In *Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, Franz Josef Hausmann et al., 287–296. Berlin: Walter de Gruyter, 1989.

Karanikolas, Nikitas N., Eleni Galiotou, George J. Xydopoulos, Angela Ralli, Konstantinos Athanasakos et. al.. "Structuring a Multimedia Tri-dialectal

Dictionary". In *LNAI 8082: International Conference on Text, Speech and Dialogue*, Ivan Habernal and Vaclav Matousek, 509–518. Berlin Heidelberg: Springer, 2013.

Keymeulen, J. Van. "Pilot project: A Dictionary of the Dutch Dialects". Paper presented at the 14th EURALEX International Congress, Fryske Akademy, Ljouwert/Leeuwarden, 6–10 July 2010

Landau, Sidney. *The art and craft of lexicography (2nd edition)*. Cambridge: Cambridge University Press, 2001

Markus, Manfred and Reinhard Heuberger. "The architecture of Joseph Wright's English Dialect Dictionary: preparing the computerized version". *International Journal of Lexicography* Vol. 20 no. 4 (2007): 355–368

Papadopoulos, A. *Historical Dictionary of the Pontic Dialect*. Athens: Epitropi Pontiakon Meleton, 1958

Penhallurik, Rob. "Dialect Dictionaries". In *The Oxford History of English Lexicography Volume II*, A. P. Cowie, 290–313. Oxford: Oxford University Press, 2009.

Rys, Kathy and Jacques Van Keymeulen. "Intersystemic correspondence rules and headwords in Dutch dialect lexicography". *International Journal of Lexicography* Vol. 22 no. 2 (2009): 129–150.

Themistocleous, C. "Cypriot Greek Lexicography: An Online Lexical Database". Paper presented at the 15th EURALEX International Congress, University of Oslo, Oslo, 7-11 August 2012.

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

Thomason, Sarah G. and Terrence Kaufman. *Language Contact, Creolization, and Genetic Linguistic*. Berkeley: University of California Press, 1988.

Tsalidis, Christos, Mavina Pantazara, Panagiotis Minos and Elena Mantzari. "NLP Tools for Lexicographic Applications in Modern Greek". In *eLexicography in the 21st Century: New Challenges, New Applications (Proceedings of eLex2009)*, S. Granger and M. Paquot, 457–462. Louvain-la-Neuve: UCL Presses, 2010.

Xydopoulos, George J. "Metalexikografikes paratiriseis sta leksika Benardi ke Syrkou (Metalexicographic comments to Benardi and Syrkou dialectal dictionaries)". *Patras Working Papers in Linguistics* Vol. 2 no. 1 (2012): 96–113

Xydopoulos, George. J. and Angela Ralli. "Greek dialects in Asia Minor: Setting lexicographic principles for a tridialectal dictionary". Paper presented at the 5th MGDLT Conference, Ghent, Belgium, September 2012

Zgusta, Ladislav. *Manual of lexicography*. The Hague: Mouton, 1971