



Towards a computational processing of oral dialectal data

Nikitas N. Karanikolas
Eleni Galiotou
Dimitris Papazachariou
Konstantinos Athanasakos
George Koronakis
Angela Ralli





European Union
European Social Fund



OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society

MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS, CULTURE & SPORTS
M A N A G I N G A U T H O R I T Y

Co-financed by Greece and the European Union



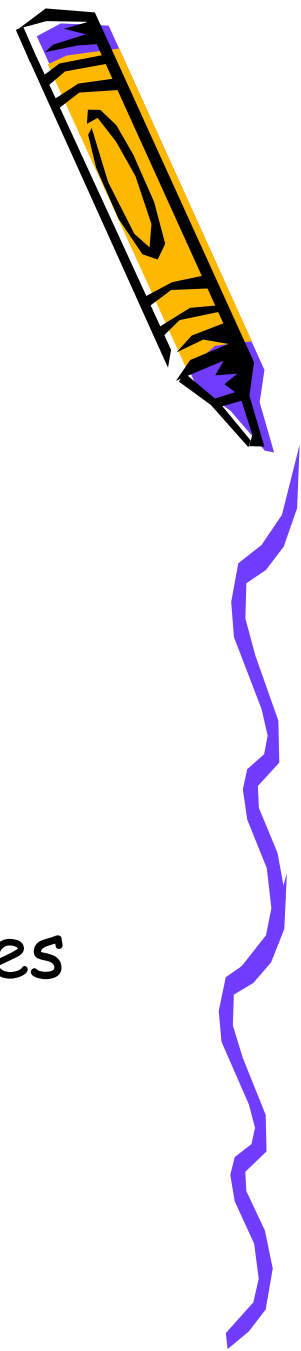
NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND

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 Social Fund.



INTRODUCTION

- We discuss issues concerning the computational processing of oral data in a unified framework for the exploitation of oral and written dialectal corpora.
- We describe the analysis and design of a multimedia database and software for storing and retrieving dialectal data, focusing on the subsystem of oral resources from three Greek dialects in Asia Minor.



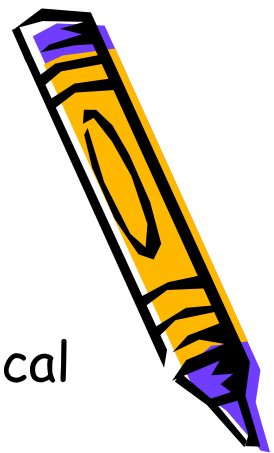
The oral corpus of AMiGre

- It was compiled in the Laboratory of Modern Greek Dialects of the University of Patras.
- It consists of approximately 180 hours of recorded raw data accompanied by metadata.
- The duration of the recordings are more or less equally distributed between the three dialects, i.e. approx. 60 hours/dialect.
- The raw data were processed according to: annotation, abstraction and analysis.
- A multimodal sub-corpus of approx. 45 hours (15 hours /dialect) was created combining raw data with transcription, translation, annotation and metadata.
- This multimodal sub-corpus was processed using the ELAN software for multimodal annotation.
- Praat software is used for phonetic analysis of spoken data which are annotated in relation to intonation phrases where tones were indicated.
- Explicit representations of vowels, diphthongs, consonants and consonant clusters appear on different layers of representation (tiers).



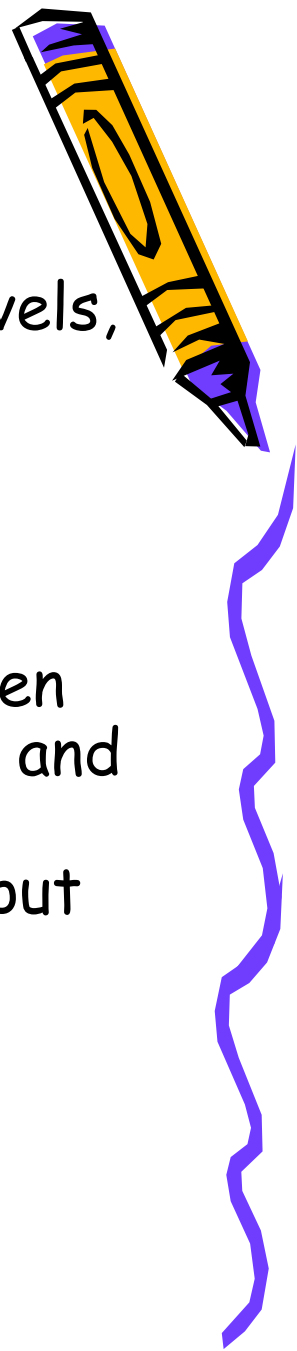
The oral corpus of AMiGre

- In order to incorporate our oral corpus into the unified framework, we added one more tier, that of the morphological representation.
- Morphological words and syllables were annotated using the SAMPA phonetic alphabet.
- IPA symbols were used in order to annotate segments and consonants.
- Vowels were encoded as triplets (v, s, p) where: $v \in \{a, e, i, o, u\}$, $s \in \{s(\text{stressed}), u(\text{unstressed}), a(\text{accented})\}$, $p \in \{b(\text{beginning of word}), m(\text{middle of word}), e(\text{end of word}), f(\text{end of phrase})\}$.
- Initially, advanced software tools such as Labb-CAT which provide the user with the possibility to store audio or video recordings, text transcripts and other annotations seemed to be adequate for the archiving and processing of this variety of linguistic information and annotation types. Yet, they could not deal with our basic requirements:



our basic requirements

- (a) Annotations at many different linguistic levels,
- (b) Combined search at different levels of representation (phonological, morphological, metadata and, eventually, syntactic and/or semantic) and,
- (c) Combined search in both the oral and written corpora. Consequently, we opted for the design and implementation of a software which would be tailored to our needs and would accept the output files of the processing with ELAN and Praat as input files.



MOTIVATION



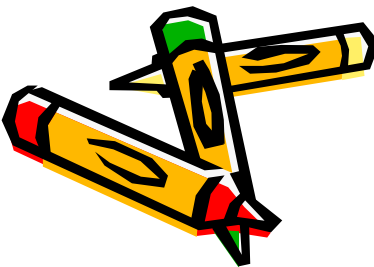
- Our motivation is to implement a Query Builder able to support a system handling original dialectical data (scanned book / text transcripts, audio files, etc), and annotations at many different linguistic levels (phonological, morphological, metadata, etc).
- Our *Search interface* should combine criteria at different levels of representation (phonological, morphological, metadata and, eventually, syntactic and/or semantic).
- Obviously, it should support combined search for both the oral and written corpora .
- We describe the analysis and design of a multimedia database and software for storing and retrieving dialectal data, focusing on the subsystem of oral resources from three Greek dialects in Asia Minor.



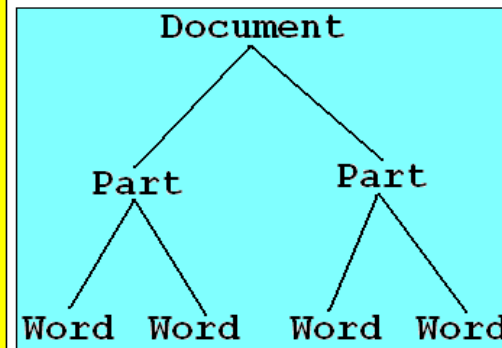
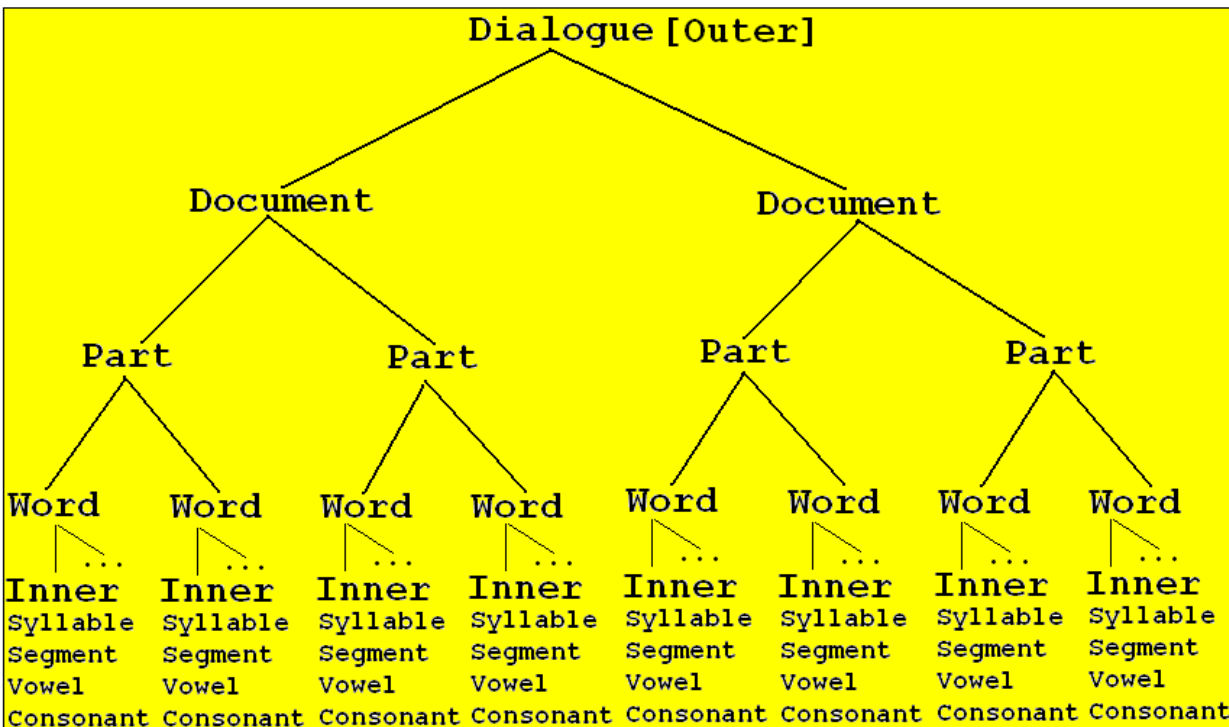
challenge to cope with a uniform structure



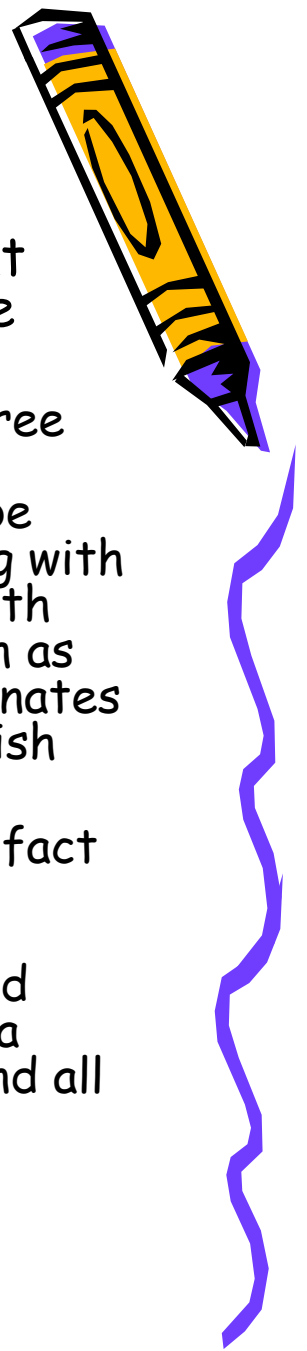
- Since our corpus is based on two different collections, we faced the challenge to cope with a uniform structure.
- Our written resources contain books, transcripts and articles which are subdivided into pages and, in turn, pages are subdivided into morphological words.
- The annotations are performed at all three above levels.
- Our oral resources contain sound recordings of single or few dialect speakers. They are subdivided into intonation phrases and the latter are subdivided into words (either intonation words, either morphological words). Moreover, words are subdivided into syllables, and segments (phonemes). Phonemes are also subdivided into vowels and consonants.
- To our convenience, we maintained a 5 level conceptual hierarchy of data (see Figure 1).



Oral (yellow) and Written sources (cyan)



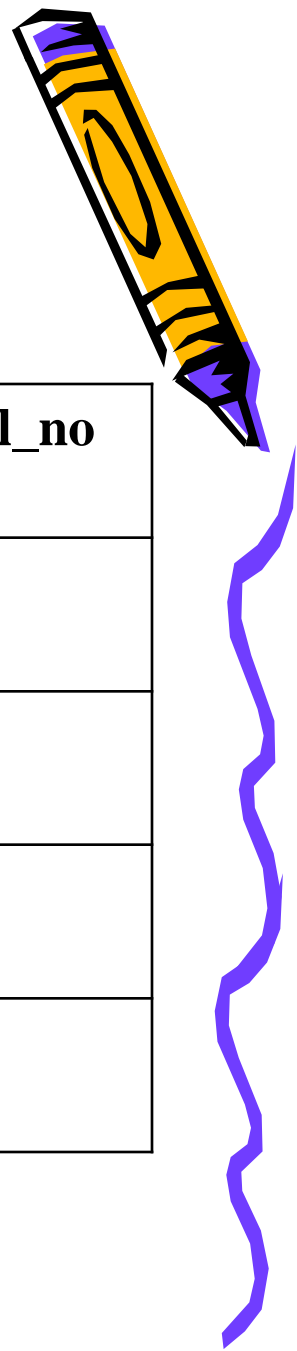
Need for Inner database



- Annotations defining Syllables, Phonemes (Segments), Vowels and Consonants are the results of a process that imports TextGrid (Praat output) files. The way that the imported data are encapsulated should aim at:
 - a) Defining criteria for retrieving items at the three main levels (document, part, word) and the inner level (syllables, phonemes, etc). For example, we would like to be able to formulate a criterion such as seeking words ending with a stressed [u]. Obviously, this criterion should combine with other criteria (for example, metadata-based criteria such as that the speaker should be at least 75 years old and originates from Trabzon (Greek "Τραπεζούντα", [trape`zunta], Turkish "Trabzon" [`trabzon]).
 - b) We should be able to create (on the fly) an artifact TextGrid (praat-like output) file with all the relevant annotations, from the information extracted from inner database. In the previous example (seeking words that end with a stressed [u]), our system should be able to create a Textgrid (praat-like output) file representing the word and all of its annotations, i.e. word, syllables, segments, vowels, consonants.



basic relation in our inner database schema

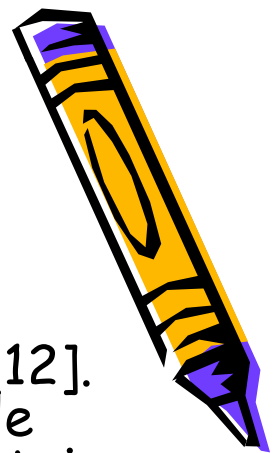


Phenomenon	WID	start	stop	Level	interval_no
u_s_e	30852	4.1234	4.2345	Vowel (11)	22
t	30852	3.9876	4.1233	Consonant (12)	27
u	30852	4.1234	4.2345	Segment (10)	28
tu	30852	3.9876	4.2345	Syllable (6)	12



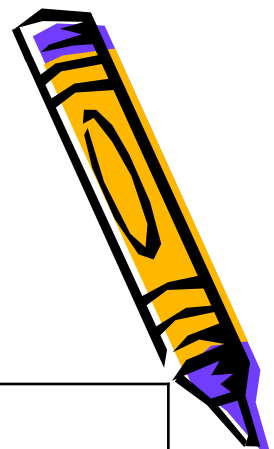
Interface requirements

- Intuitive usage,
- Support Multi valued fields. As a consequence, the "And" operator is introduced for the values of a single criterion [12]. That means that a demand for two or more values in a single record (item of a level of the data hierarchy) should be met, in addition to classical data demands (Exact, Range, Disjunction),
- 2 kinds of criteria (*main criteria* and *distance criteria*),
- Conjunction between main criteria (implicit use of *And* between rows of conditions) [10],
- Expression of Retrieval requirements for: actual data, data aggregations, artifacts (on the fly created data),
- Expression of distance conditions (distance criteria) between items which are compatible with the main criteria. Therefore, the hierarchy supports three different item levels (*Part, Word, Inner*).



Interface template

Word/ token /phenomenon			Location				
<Value>	{ Between, And, Or, -- }	<Value>	<DB>	<At- tribute >	<Part distanc es>	<Word distances>	<Interval_no distances >



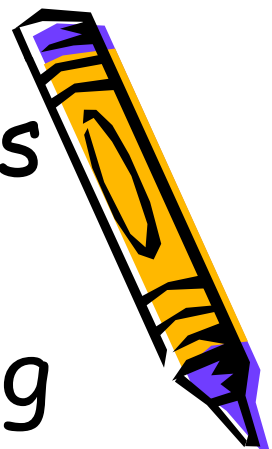
search of parts (pages for written resources) that contain the phenomenon of *Vowel Archaism*, followed by an adjective which is a loan word with a *Noun Part of Speech* and a *Masculin Gender*



Word/ token /phenomenon		Location				
vowel archaism	--	EAV Phon	↓	-	X	-
Adjective	--	EAV Morpho	PART OF SPEECH	-	Y in (X+1, X+10)	-
Noun	--	EAV Morpho	PART OF SPEECH OF	-	Y	-
Masculin	--	EAV Morpho	GENDER OF LOAN WORD	-	Y	-
Output		Part		-		



search of parts (intonation phrases
in case of oral resources) ending
with an unstressed vowel, appearing
in the (oral) collection



Word/ token /phenomenon			Location				
?_u_f	--		detailed Phon	Vowel	-	-	-
Output			Document		count_part		



search based on metadata of participants



Word/ token /phenomenon			Location					
<u>Ifigenia Zisi</u>	Or	Mary <u>Karra</u>	EAV (O)	Meta	Annotator	-	-	-
Male	--		EAV (O)	Meta	Inf. Sex	-	-	-
75	Between	100	EAV (O)	Meta	Inf. Age	-	-	-
<u>cappadocians</u>	--		EAV (O)	Meta	Inf. Origin	-	-	-
Output			Document			-		



Questions

- Thank you for attending the presentation
- nnk@teiath.gr
- <http://users.teiath.gr/nnk/>

