a Query Builder for dialectical corpora

oral and witten dialectical data and annotations at many different linguistic levels (phonological, morphological, metadata, etc)

Nikitas N. Karanikolas

An International Lecture, Based on the AMiGre project, Thalis framework.

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society

MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS, CULTURE & SPORTS
MANAGING AUTHORITY

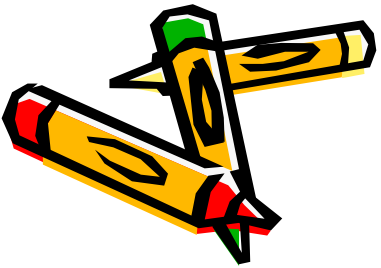
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

Co-financed by Greece and the European Union

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.

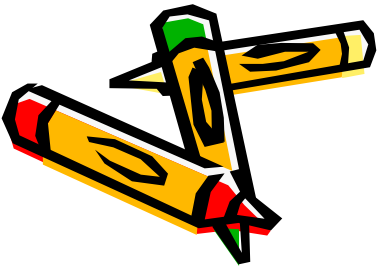
INTRODUCTION

- We discuss issues concerning the computational processing of annotated oral and written data in a unified framework for the exploitation of oral and written dialectal corpora (from three Greek dialects in Asia Minor).
- We are focusing on the oral data and the subsystem for expressing the needs for retrieval (the query builder).
- For written data see lecture Dialectal Corpora Building.



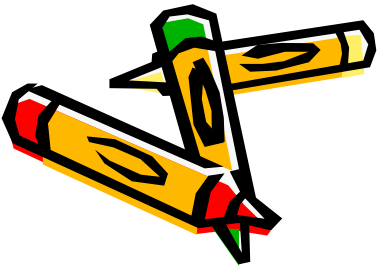
The oral corpus of AMiGre

- 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).

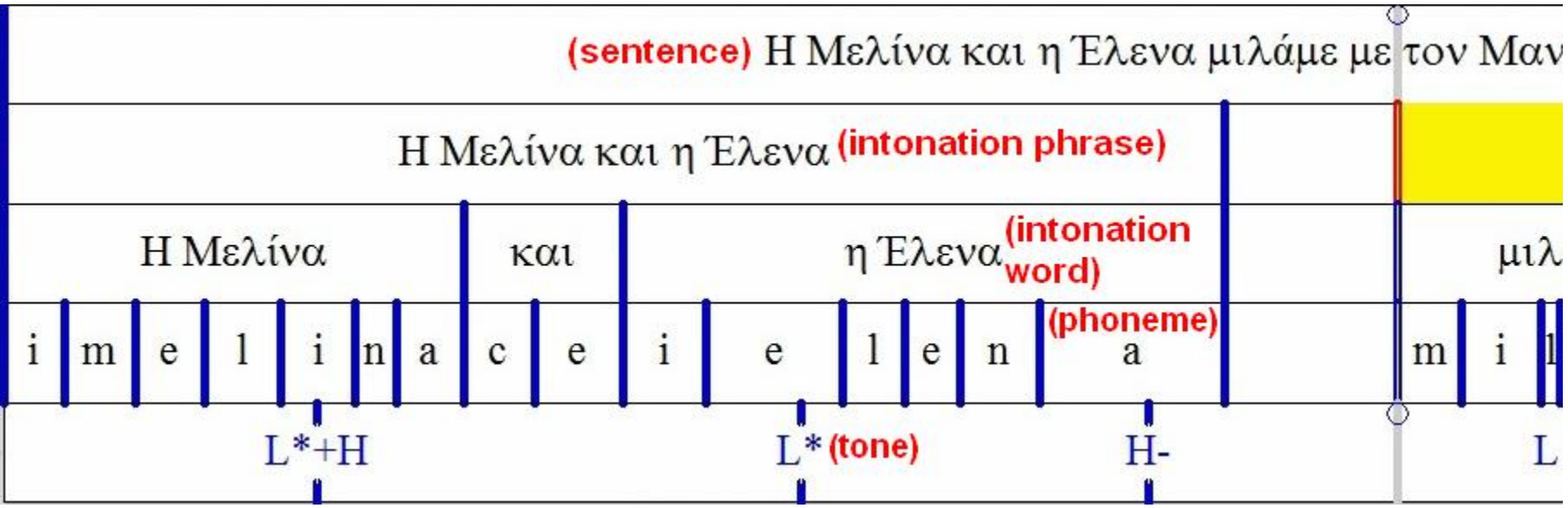


The oral corpus of AMiGre

- 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(stressed), u(unstressed), a(accented)}, p $\in$ {b(beginning of word), m(middle of word), e(end of word), f(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:



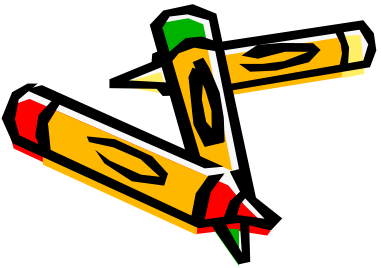
Oral data (digitized records) of dialectical dialogues annotated with Praat software



Query Builder for dialectical corpora, Nikitas N. Karanikolas

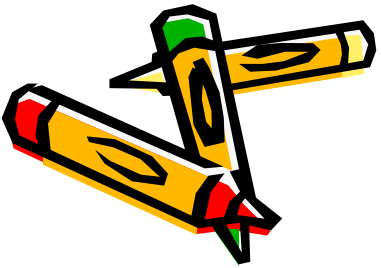
our basic requirements

- (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.



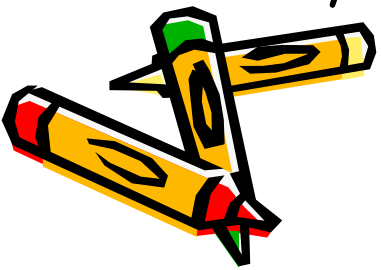
MOTIVATION

- 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.

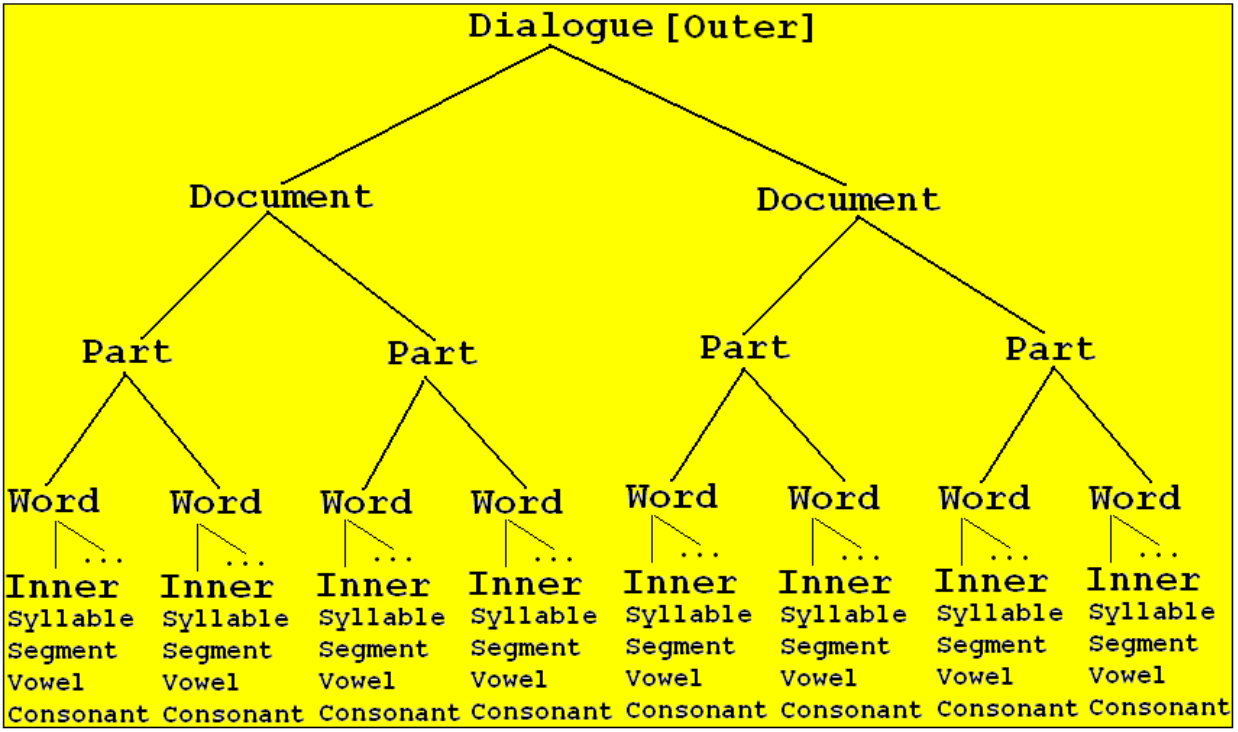


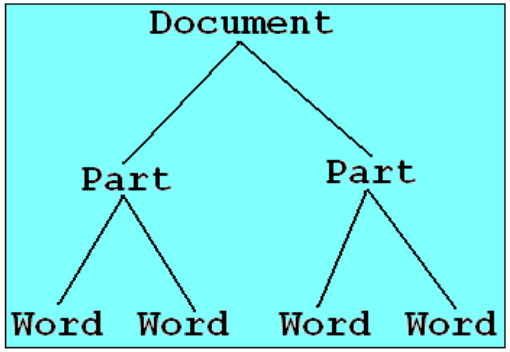
challenge to cope with a uniform structure

- 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.



Oral (yellow) and Written sources (cyan)





Struct DB for oral data

oral_sources [Dialogue]

OralSourceId	INT(11)
Title	VARCHAR(500)
TextGridFilePath	VARCHAR(200)
TextGridUploadedOn	DATETIME
MetadataFilePath	VARCHAR(200)
MetadataUploadedOn	DATETIME
Notes	TEXT
IsDeleted	BIT(1)
CreatedOn	DATETIME

oral_speakers [Document]

SpeakerId	INT(11)
Code	VARCHAR(100)
Name	VARCHAR(200)
TextGridFilePath	VARCHAR(200)
TextGridUploadedOn	DATETIME
OrderIndex	INT(11)
IsDeleted	BIT(1)
OralSourceId	INT(11)

oral_audiofiles

AudioFileId	INT(11)
SpeakerCodes	VARCHAR(200)
AudioFilePath	VARCHAR(200)
AudioUploadedOn	DATETIME
OralSourceId	INT(11)

oral_turntakings [Part]

TurnTakingId	INT(11)
Text	TEXT
TextInSampa	TEXT
TranslationText	TEXT
Sequence	INT(11)
OrderIndex	INT(11)
OrderIndexInTotal	INT(11)
StartTime	DOUBLE
EndTime	DOUBLE
SpeakerId	INT(11)

Oral_SegmentTypes

SegmentTypeId	INT(11)
Name	VARCHAR(45)
Code	VARCHAR(45)

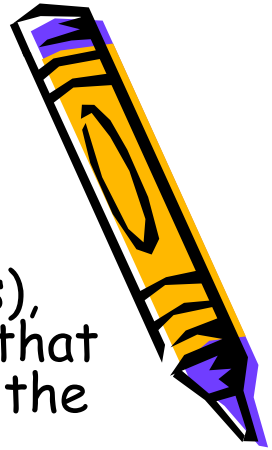
oral_words [Word]

WordId	INT(11)
PositionIndex	INT(11)
Text	VARCHAR(500)
PositionStart	INT(11)
PositionEnd	INT(11)
TextInSampa	VARCHAR(500)
PositionStartInSampa	INT(11)
PositionEndInSampa	INT(11)
TurnTakingId	INT(11)
StartTime	DOUBLE
EndTime	DOUBLE

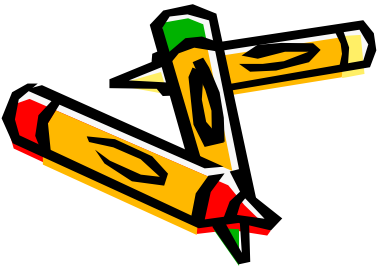
oral_segments [Inner]

SegmentId	BIGINT(20)
Text	VARCHAR(20)
PositionIndexInWord	INT(11)
PositionIndexInTurnTaking	INT(11)
PositionIndexInSpeaker	INT(11)
StartTime	DOUBLE
EndTime	DOUBLE
TypeId	INT(11)
WordId	INT(11)

Need for a relation for the Inner level

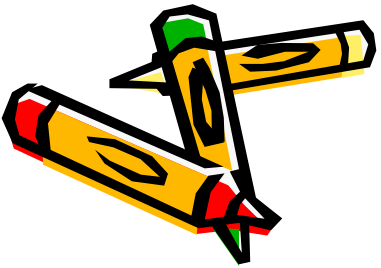


- Annotations defining Syllables, Phonemes (Segments), Vowels and Consonants are the results of a process that imports TextGrid (Praat output) files. The way that the imported data are encapsulated should aim at:
  - a) Defining criteria for retrieving items at the three main levels (document, part, 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 form 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.



Structure of the Inner level relation

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



This is our first approach. We have switched to EAV. However we keep it because it is easy for explanation.

inner level relation explanations

- All retrieved intervals of the same tier can be ordered based on the property **interval_no** which emanates from the original TextGrid file.
- In this way, we can formulate queries concerning the distance between segments at a certain level.
- The values of the attributes phenomenon, start and stop also emanate from the original TextGrid file.
- This structure enables the hierarchical reproduction of the data from the lower levels of fig. 1 and, at the same time, the serial access to the elements of the lower level.

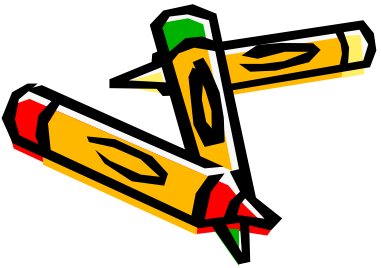
Interface requirements

- Intuitive usage,
- Support Multi valued fields. As a consequence, the "*And*" operator is introduced for the values of a single criterion. 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),
- 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 interface should support three different distance conditions (*Part, Word, Inner*).

Interface template

Word/ token /phenomenon			**Location**				

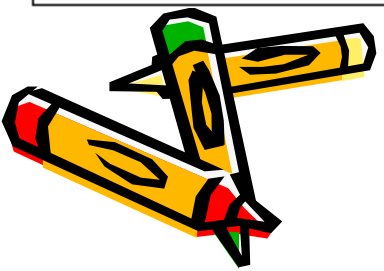
<Value>	{
	Between,
	And, Or,
	--
	}
		<Value>	<DB>	<At-tribute >	<Part distances>	<Word distances>	<Interval_no distances >



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 LOAN WORD	-	Y	-
Masculin	--		EAV Morpho	GENDER OF LOAN WORD	-	Y	-

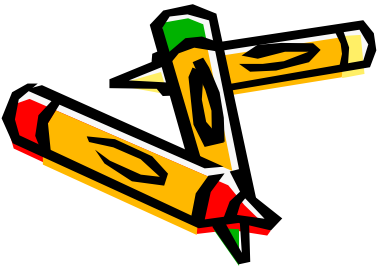
Output	Part	-



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	-	-	-

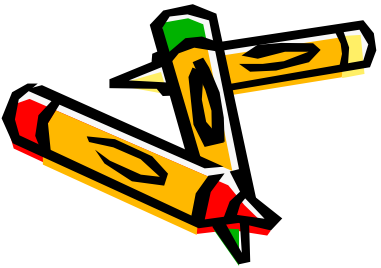
Output	Document	count_part



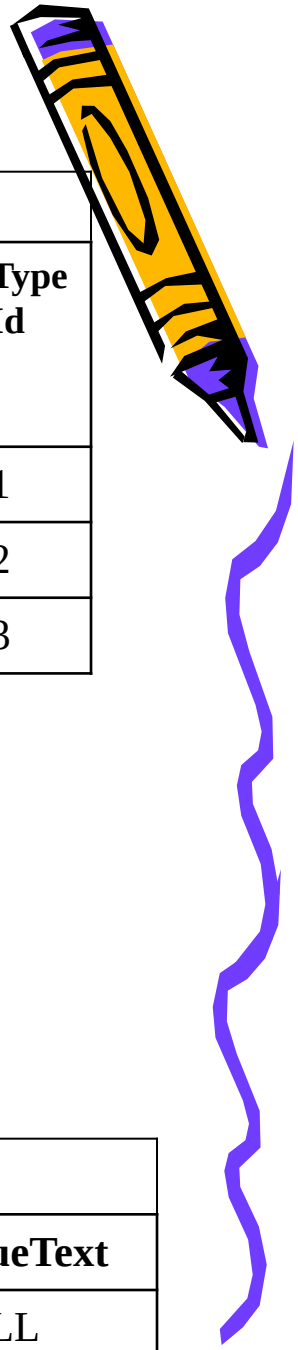
search based on metadata of participants

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

Output	Document	-

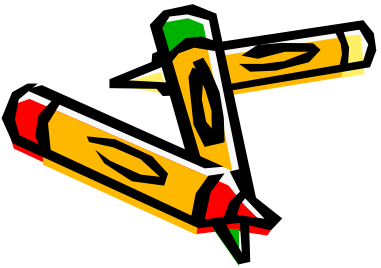


Inner Annotations in EAV



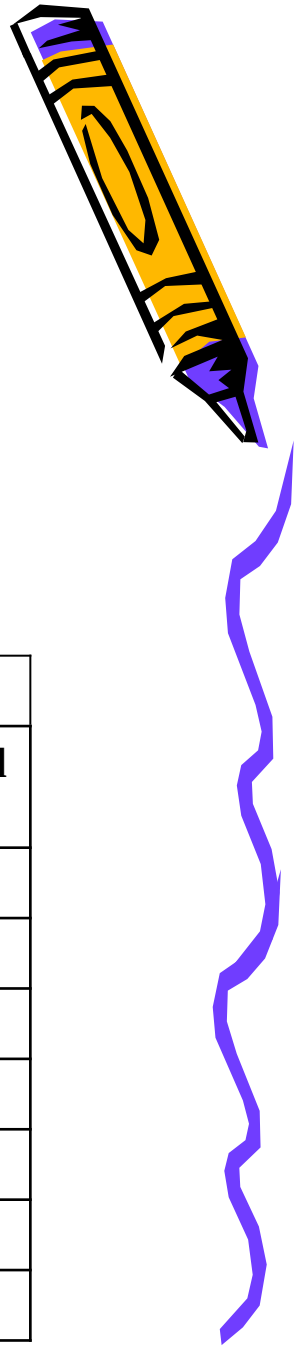
Oral_Segments								
Segment Id	**Text**	**Word Id**	**Position Index In Word**	**Position Index In Turn Taking**	**Positio Index In Speaker**	**Start Time**	**End Time**	**Type Id**
8501	u	30852	3	13	22	4.1234	4.2345	1
8502	t	30852	4	14	27	3.9876	4.1233	2
8503	tu	30852	1	5	12	3.9876	4.2345	3

oral_SegmentTypes		
SegmentTypeId	**Name**	**Code**
1	Φωνήεν	VOWEL
2	Σύμφωνο	CONSONANT
3	Συλλαβή	SYLLABLE



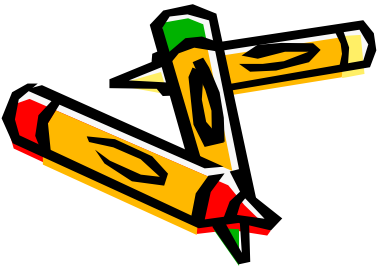
EntityPropertyValues				
EAV_Id	**EntityId**	**Property Id**	**Value Id**	**ValueText**
11240	8501	148	1446	NULL
11241	8501	149	1450	NULL

Definitions of attributes and values in EAV

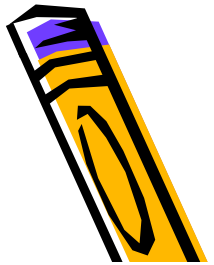


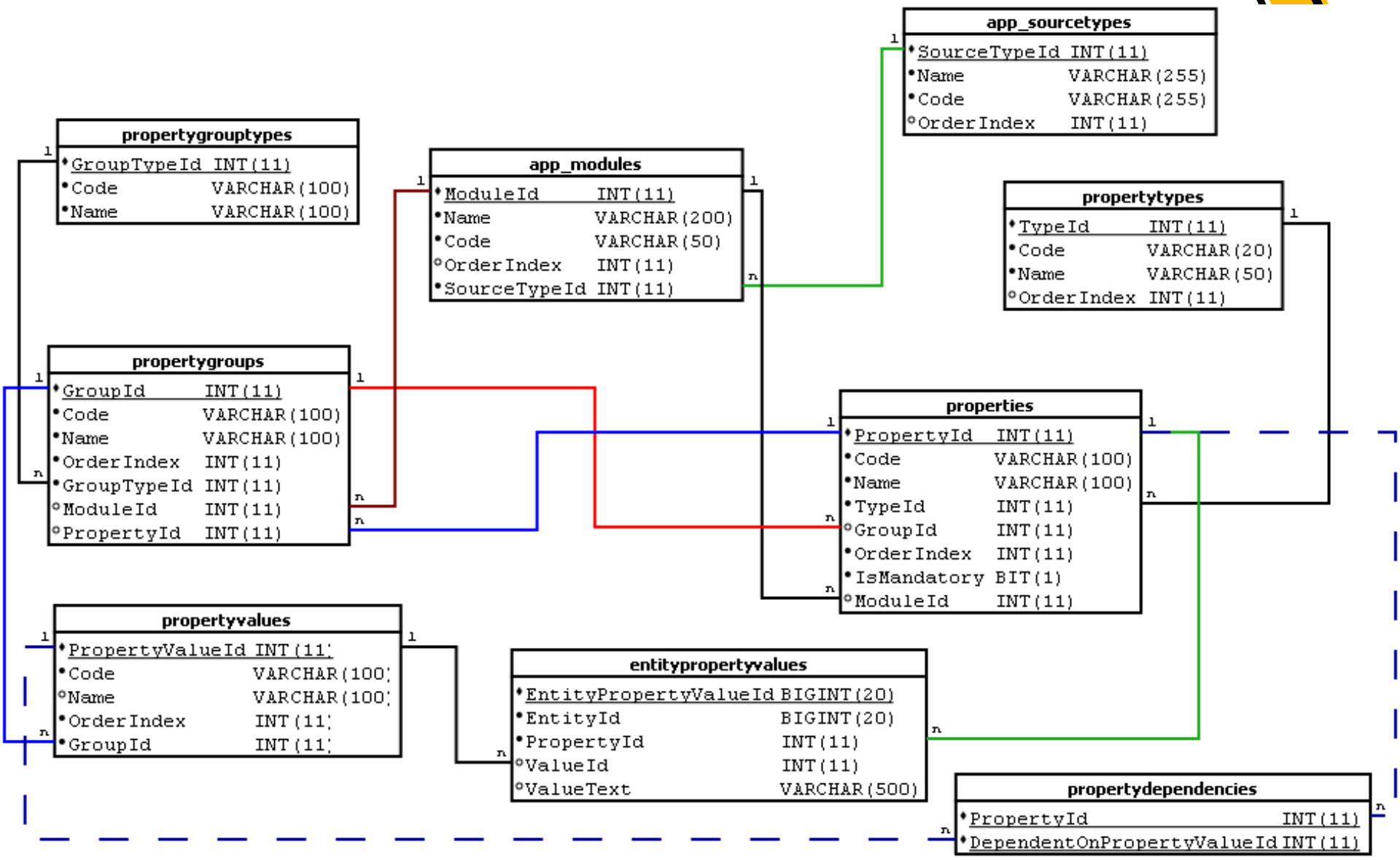
Properties				
Property Id	**Name**	**TypeId**	**Is Mandatory**	**ModuleId**
148	accent	3	0	11
149	Accent location	3	0	11

PropertyValues			
PropertyValueId	**Code**	**Name**	**Property Id**
1445	Unstressed	Άτονο	148
1446	Stressed	Τονισμένο	148
1447	Accented	Εστιασμένο	148
1448	Beginning of word	Αρχή Λέξης	149
1449	Middle of word	Μέση Λέξης	149
1450	End of word	Τέλος Λέξης	149
1451	End of phrase	Τέλος Φράσης	149



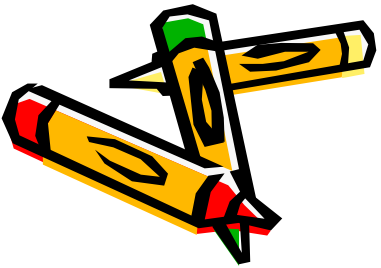
EAV schema




app_sourcetypes
SourceTypeId INT(11)
Name VARCHAR(255)
Code VARCHAR(255)
OrderIndex INT(11)
propertygrouptypes
GroupTypeId INT(11)
Code VARCHAR(100)
Name VARCHAR(100)
app_modules
ModuleId INT(11)
Name VARCHAR(200)
Code VARCHAR(50)
OrderIndex INT(11)
SourceTypeId INT(11)
propertytypes
TypeId INT(11)
Code VARCHAR(20)
Name VARCHAR(50)
OrderIndex INT(11)
propertygroups
GroupId INT(11)
Code VARCHAR(100)
Name VARCHAR(100)
OrderIndex INT(11)
GroupTypeId INT(11)
ModuleId INT(11)
PropertyId INT(11)
properties
PropertyId INT(11)
Code VARCHAR(100)
Name VARCHAR(100)
TypeId INT(11)
GroupId INT(11)
OrderIndex INT(11)
IsMandatory BIT(1)
ModuleId INT(11)
propertyvalues
PropertyValueId INT(11)
Code VARCHAR(100)
Name VARCHAR(100)
OrderIndex INT(11)
GroupId INT(11)
entitypropertyvalues
EntityPropertyValueId BIGINT(20)
EntityId BIGINT(20)
PropertyId INT(11)
ValueId INT(11)
ValueText VARCHAR(500)
propertydependencies
PropertyId INT(11)
DependentOnPropertyValueId INT(11)

See also

- **Nikitas N. Karanikolas**, Eleni Galiotou, Dimitris Papazachariou, Konstantinos Athanasakos, George Koronakis and Angela Ralli, "Towards a computational processing of oral dialectal data". PCI 2015, October 01 - 03, 2015, Athens, Greece. ACM 978-1-4503-3551-5, doi:10.1145/2801948.2801966. http://doi.acm.org/10.1145/2801948.2801966



Questions

- Thank you for attending the presentation

- nnk@teiath.gr
- http://users.teiath.gr/nnk/