



# Dialectal Corpora Building (for oral and written sources)

by  
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

Co-financed by Greece and the European Union



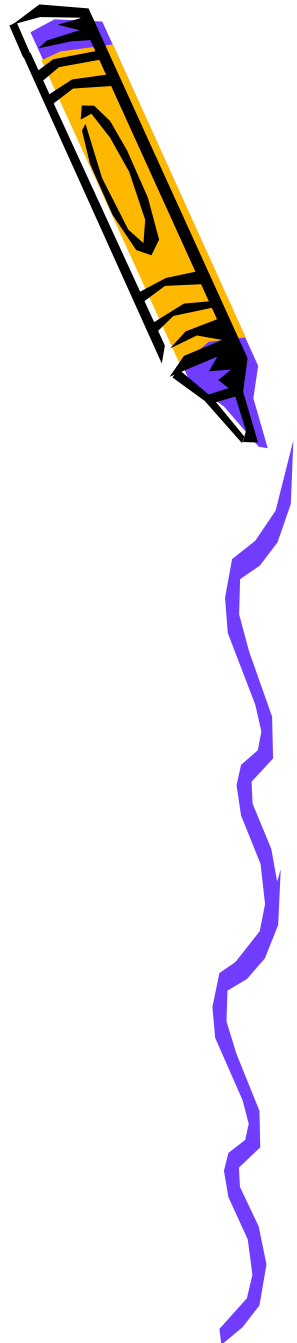
NSRF  
2007-2013  
programme for development  
EUROPEAN SOCIAL FUND

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.



# AMiGre

- Introduction
- Sources
- Applications Overview
- Design Overview



# Introduction

Pontus, Cappadocia, Aivali: In search of Asia Minor Greek

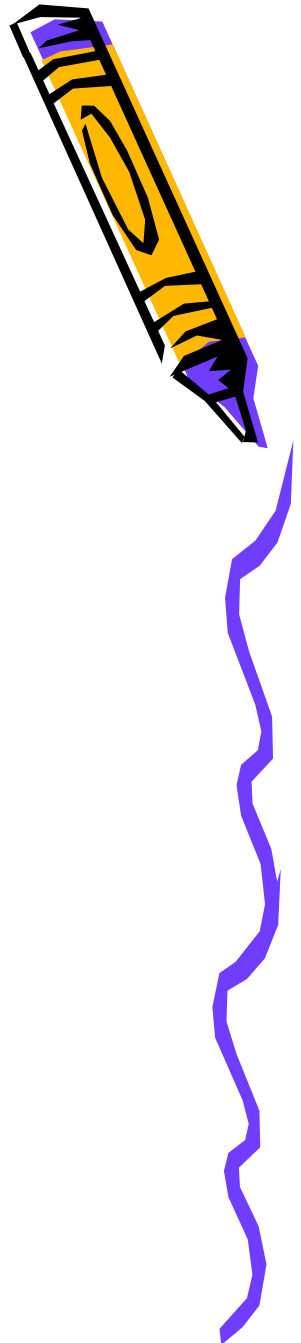
- Cappadocian, Pontic and Aivaliot are three varieties of Greek once spoken in Asia Minor. Following the population exchange between Greece and Turkey enforced by the Lausanne Treaty (1923), speakers of these varieties were relocated in various parts of Greece, leaving behind their lands and material possessions and carrying along with them only their history, mores and customs, traditions, and language.
- Due to their long-lasting contact with neighbouring Turkish and their isolation from the other Greek dialects, the Asia Minor Greek varieties (especially Pontic and Cappadocian) are regarded as ideal case-studies for shedding light on the linguistic evolution of Greek as well as on various language contact phenomena. Crucially, the three dialects channel a rich cultural and linguistic heritage but face a severe danger of extinction: the number of first generation refugees is shrinking rapidly and the next generations are being gradually absorbed by adstratal Modern Greek, both culturally and linguistically. Thus, the necessity of describing and preserving this precious piece of heritage is vital.



Nikitas N. Karanikolas - Dialectal Corpora Building

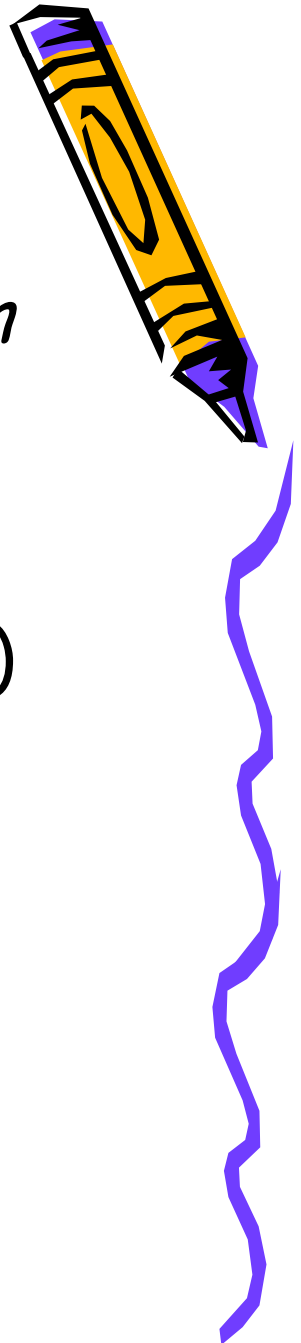
# AMiGre

- Introduction
- **Sources**
- Applications Overview
- Design Overview



# Sources

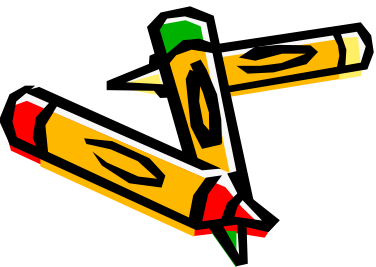
- Lexical data for 3 Asia Minor Greece dialects (see lecture *Dialectal lexicon building: requirements and technical specifications*)
- Digital Audio files (WAV) and Annotations (TextGrid - Praat - files)
- Written sources (digitized)
- their homogenized Transcriptions
- Morphological annotations
- Syntactic and Semantic annotations



# Digital Audio files (WAV) and Annotations (TextGrid, ELAN files)

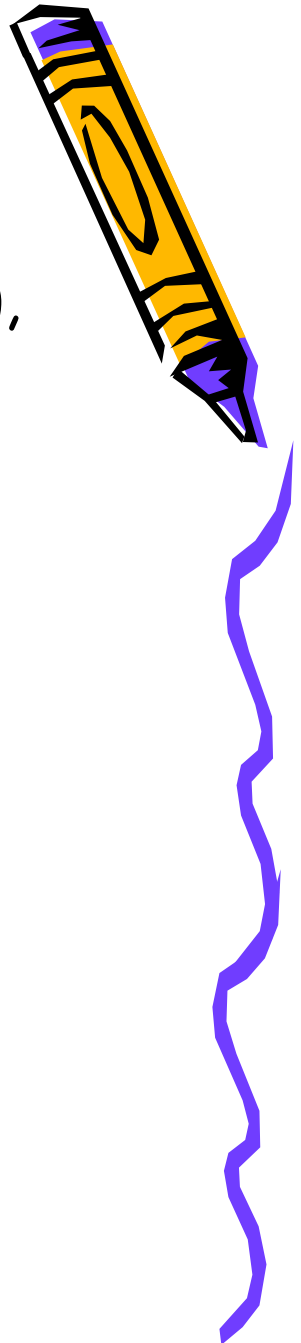


- There are digitized audio files (WAV) of dialectal conversations
- Annotated with: sentences, phrases, syllables, segments, vowels, consonants, tones, phenomena, etc (TextGrid - Praat - files, ELAN files)

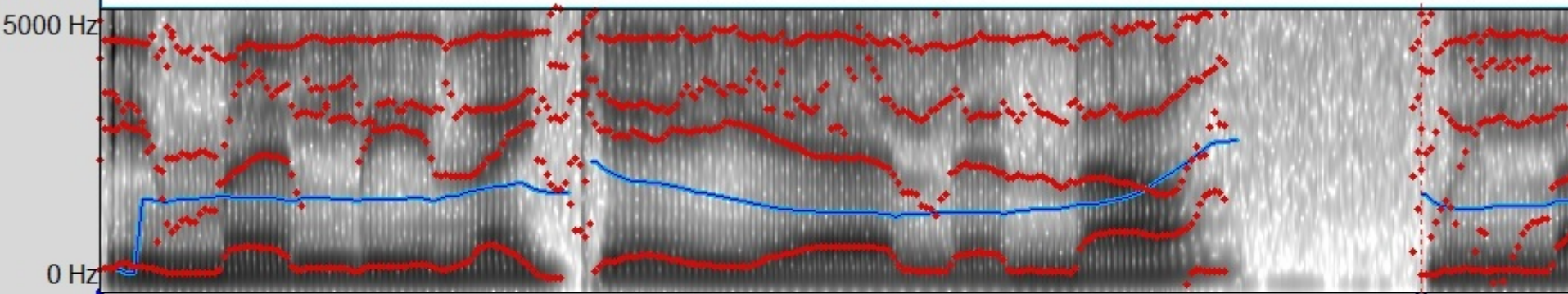


# Some phenomena

- Tone / accent  
(is it a question, an imperative statement, etc),
- Pragmatological phenomena,
- Intonations,
- Morphological words,
- Phonological words,
- Intonation phrases,
- Intonation sentences,
- Phonemes,
- Voices,
- Accent phenomena (stress, unstress).







1 (sentence) Η Μελίνα και η Έλενα μιλάμε με τον Μαν

2 Η Μελίνα και η Έλενα (intonation phrase)

3 Η Μελίνα και η Έλενα (intonation word)

4 i m e l i n a c e i e l e n a (phoneme)

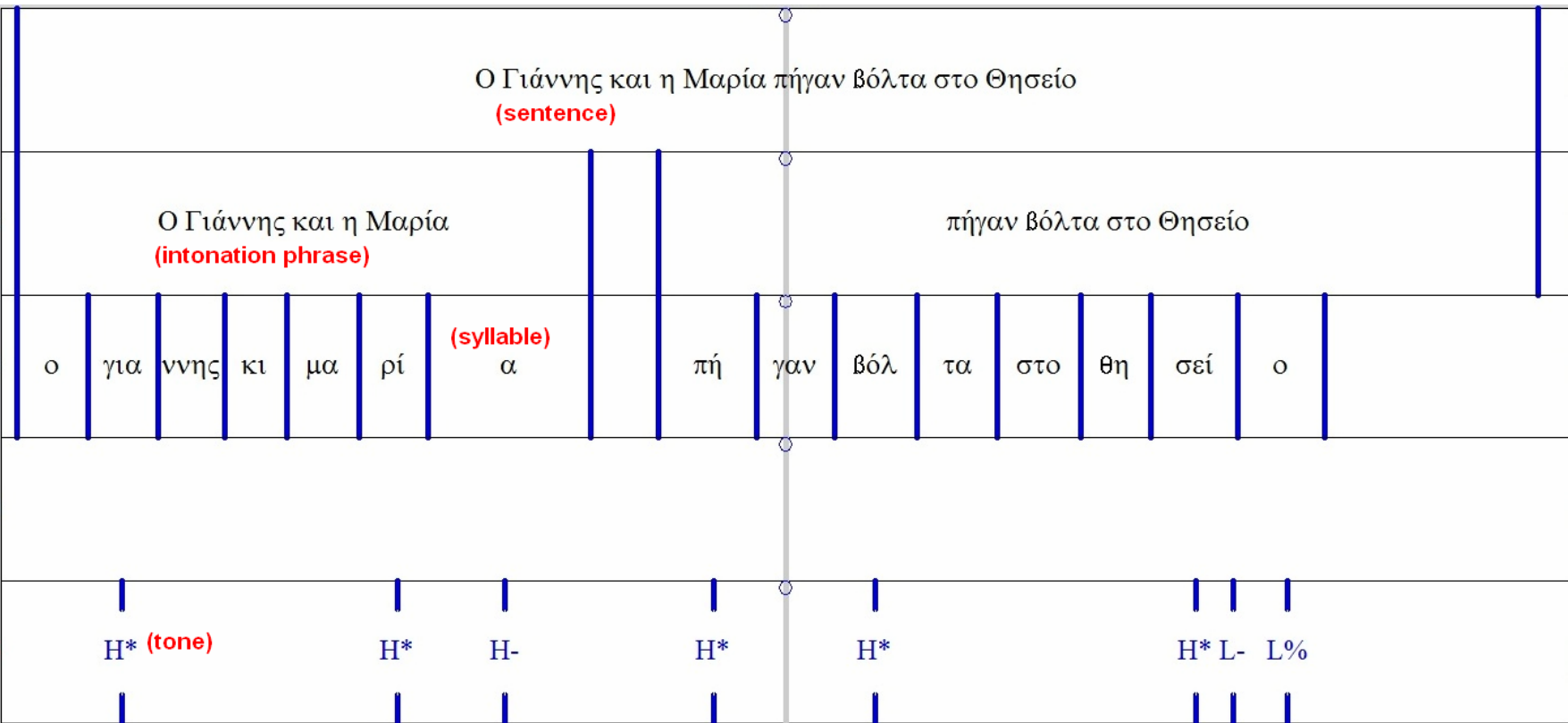
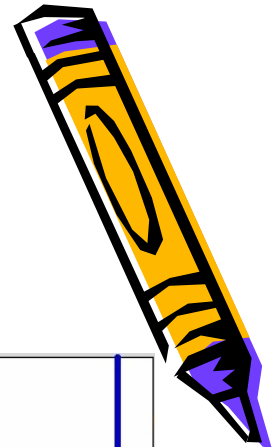
5 L\*+H L\* (tone) H-

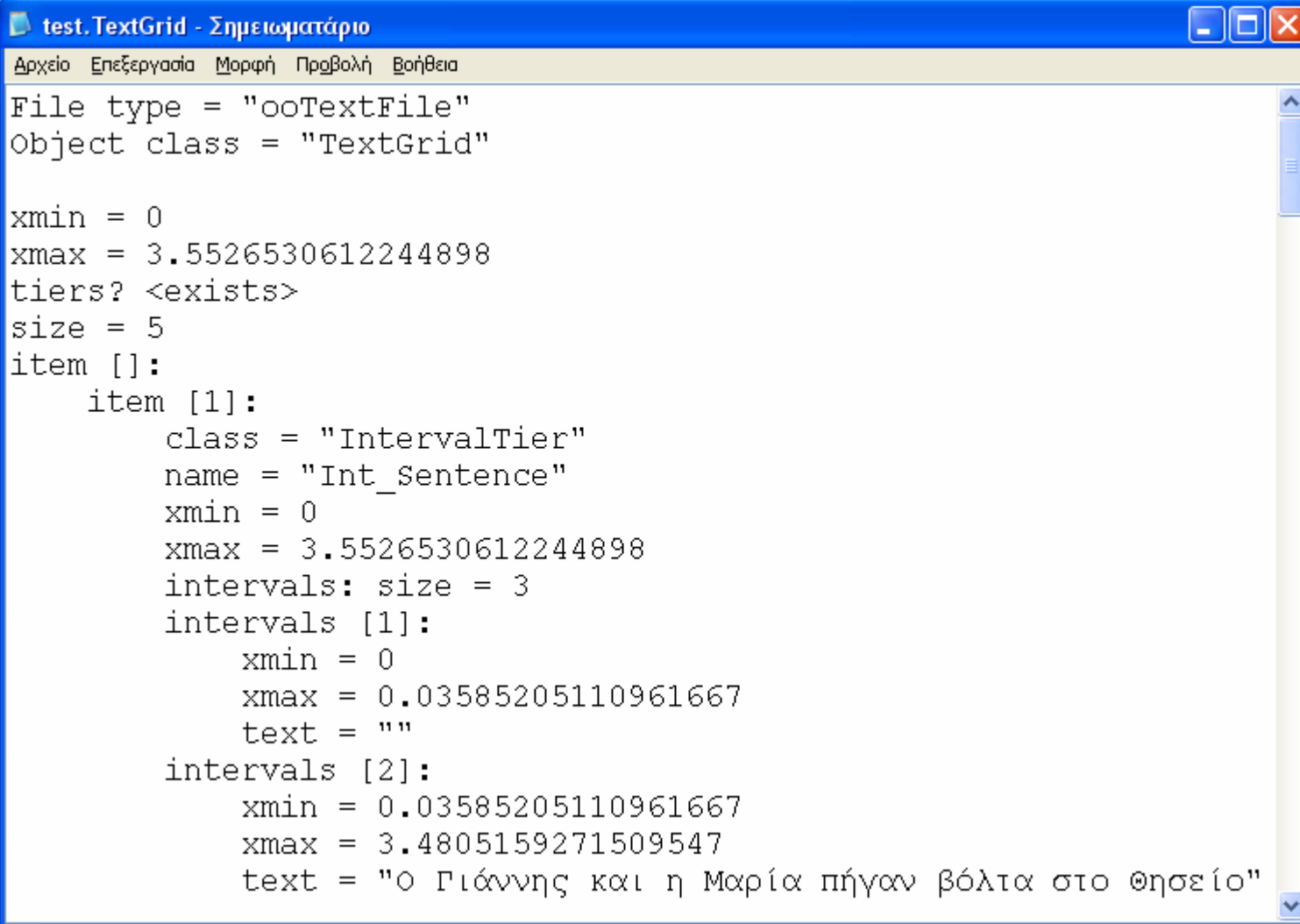
1.717434

14521 0.414521 Visible part 2.958813 seconds

Total duration 24.879728 seconds

# Annotations in various levels with PRAAT

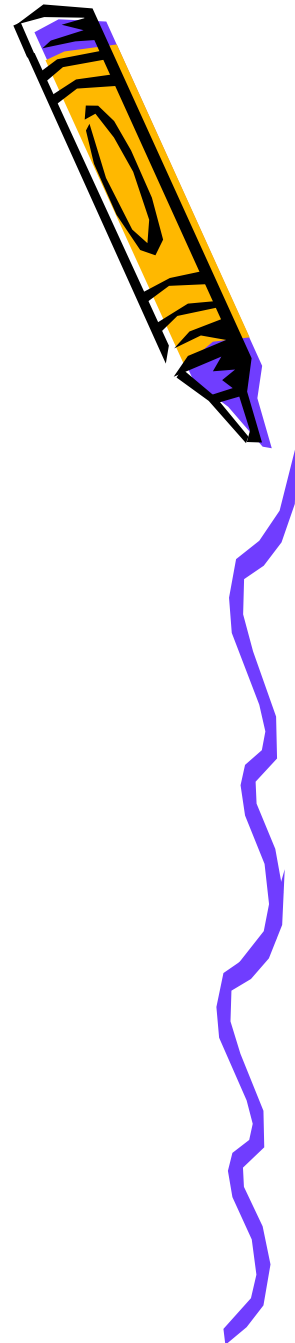




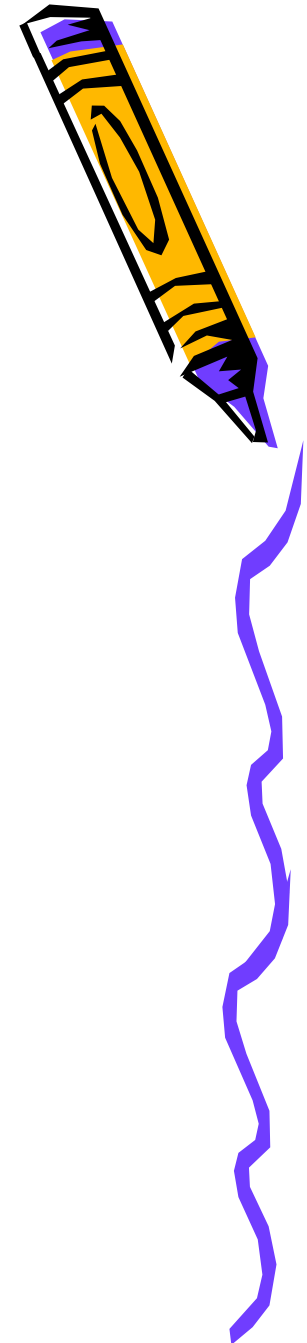
```
test.TextGrid - Σημειωματάριο
Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 3.5526530612244898
tiers? <exists>
size = 5
item []:
  item [1]:
    class = "IntervalTier"
    name = "Int_Sentence"
    xmin = 0
    xmax = 3.5526530612244898
    intervals: size = 3
    intervals [1]:
      xmin = 0
      xmax = 0.03585205110961667
      text = ""
    intervals [2]:
      xmin = 0.03585205110961667
      xmax = 3.4805159271509547
      text = "Ο Γιάννης και η Μαρία πήγαν βόλτα στο Θησείο"
```

## The Sentence tier of Praat (textgrid) file

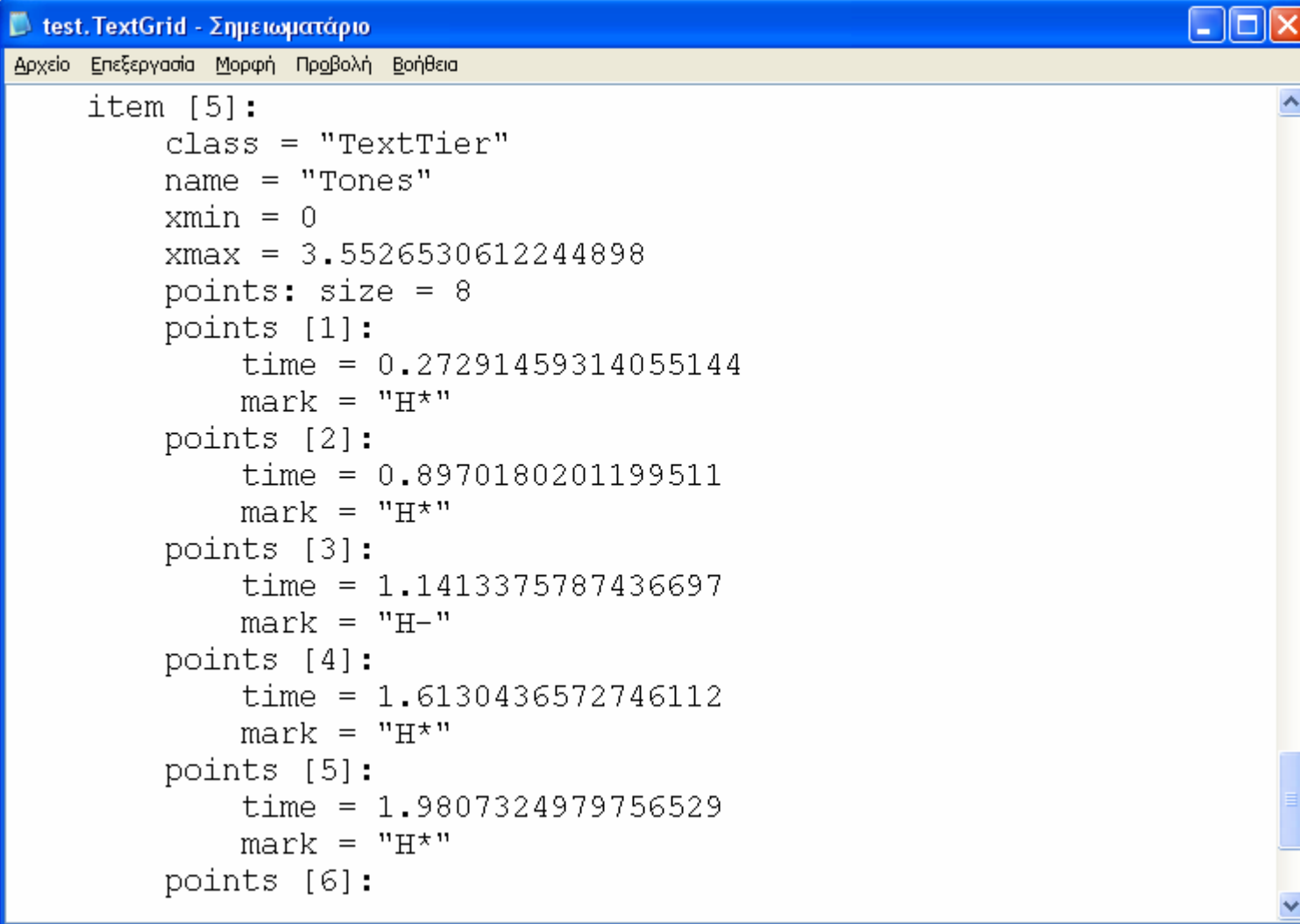


```
test.TextGrid - Σημειωματάριο
Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια
item [2]:
  class = "IntervalTier"
  name = "Int_phrase"
  xmin = 0
  xmax = 3.5526530612244898
  intervals: size = 5
  intervals [1]:
    xmin = 0
    xmax = 0.03585205110961667
    text = ""
  intervals [2]:
    xmin = 0.03585205110961667
    xmax = 1.3348580212179022
    text = "Ο Γιάννης και η Μαρία"
  intervals [3]:
    xmin = 1.3348580212179022
    xmax = 1.489674375197288
    text = ""
  intervals [4]:
    xmin = 1.489674375197288
    xmax = 3.4805159271509547
    text = "πήγαν βόλτα στο Θησείο"
```



The Intonation Phrase tier of Praat (textgrid) file





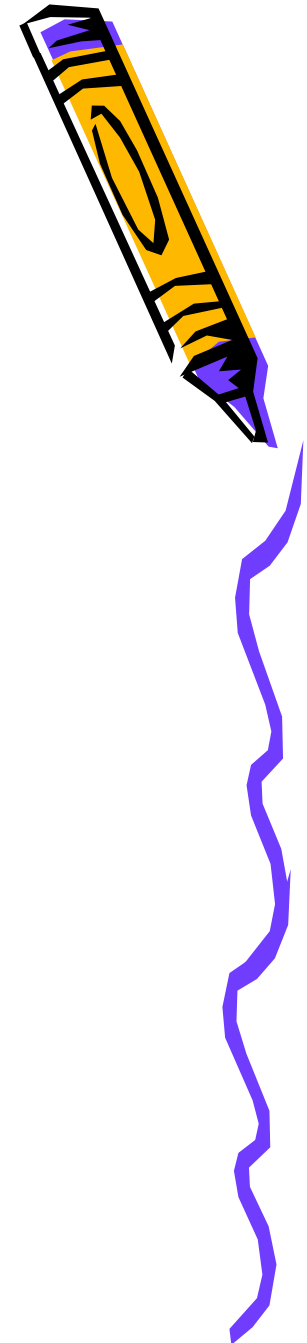
```
test.TextGrid - Σημειωματάριο
Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια

item [5]:
  class = "TextTier"
  name = "Tones"
  xmin = 0
  xmax = 3.5526530612244898
  points: size = 8
  points [1]:
    time = 0.27291459314055144
    mark = "H*"
  points [2]:
    time = 0.8970180201199511
    mark = "H*"
  points [3]:
    time = 1.1413375787436697
    mark = "H-"
  points [4]:
    time = 1.6130436572746112
    mark = "H*"
  points [5]:
    time = 1.9807324979756529
    mark = "H*"
  points [6]:
```

The Tone point tier of Praat (textgrid) file



Nikitas N. Karanikolas - Dialectal Corpora Building



# Written sources (digitized) Pontic (Ποντιακά)



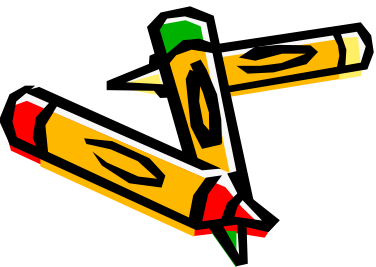
Ἔτον ἕνας πολλά πλούσιος καὶ εἶχεν ἕναν παιδὶν καὶ τὸ παιδὶν ἀτ' ἐπέγ'νεν ἕς σὸ σχολεῖον. Τ' ἄλλ' τὰ παιδία, π' ἐκράτ'ναν μει' ἐκείνον, εἶχαν βαγγέλον κ' ἐκεῖνος ἕκ' εἶχεν. Εἶπεν ἕναν ἡμέραν τῇ μάννῃν ἀχτε «μάννα, τὰ παιδία, ὅλα ποὺ κρατοῦν μετ' ἐμέν, ἔχουν βαγγέλα κ' ἐγὼ ἕκ' ἔχω, καὶ ἕκ' ἰλές τὸν κύρη μ' καὶ παίρ' κ' ἐμέν ἕναν βαγγέλον;». Ἡ μάννα ἕτ' πα εἶπεν ἀτο τὸν κύρ'ν ἀτ' κ' ἐπῆρεν κ' ἐδέκεν ἅ κ' ἐδέστεν. Ἄς σὸ ἐδέστεν κ' ὕστερον, ἔγκεν ἅ ἕς ἕναν κουῖμτῶῃν κ' ἐντῶκεν ἅπάν' ἐκάν πεντακόσα φηριλία κ' ἐδέκεν ἀτο τὸ γιόν ἀτ'. Κι ἀτὸς πάει κ' ἔρται ἕς σὸ σχολεῖον. Ἐρθεν ἕναν ἡμέραν ἕνας καλόγερος ἕς σοῦ πλούσιονος καὶ ἐρώτεσαν ἀτον «ἀπόθεν ἔρχεσαι καὶ ποῦ πάς;» Εἶπεν ἀτ' ἕς κ' ἐκεῖνος «ἅς σ' Ἄγιον Ὀρος ἔρχουμαι καὶ ἕς σὸν Ἄιν Τάφον πάγω». Εἶπεν ἀτον τὸν πουρνόν ὁ Γιαννίτσης τοῦ



# Their homogenized Transcriptions



έτον ένας πολλά πλούσιος και είΣεν έναν παιδίν και το παιδίν ατ' επέγνεν σο  
σχολείον. τ' άλλ' τα παιδιά, π' εκράτναν μετ' εκείνον, είχαν βαγγέΛΟν κι εκείνος  
'κ' είΣεν. είπεν έναν ημέραν την μάναν αχτε «μάνα, τα παιδιά, όλα που κρατούν  
μετ' εμέν, έχουν βαγγέΛΑ κι εγώ 'κ' έχω, κά 'κι λες τον κύρη μ' και παίρ' κι εμέν  
έναν βαγγέΛΟν;». Η μάνα 'τ' πα είπεν ατο τον κύρ'ν ατ κι επήρεν κι εδέκεν α κι  
εδέστεν. ασο εδέστεν κι ύστερον έγκεν ας έναν κουιμτΣήν κι εντώκεν απάν' εκάν  
πεντακόΣα φηριλία κι εδέκεν ατο τον γιόν ατ. κι ατος πάει κι έρται σο σχολείον.  
έρθεν έναν ημέραν ένας καλόγερος σου πλούσιονος και ερώτεσαν ατον «απόθεν  
έρΣεσαι και πού πας;». είπεν ατ'ς κι εκείνος «ας άγιον όρος έρχουμαι και σον  
άιν τάφον πάγω». είπεν ατον τον πουρνόν ο γιαννίτσης του

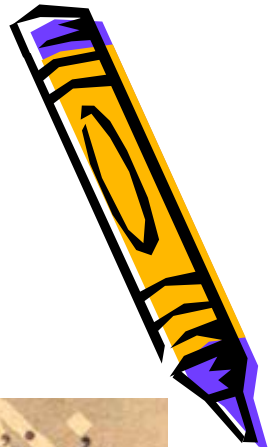


## 2. Ἀτματσᾶς καὶ ἡ ποθίκα<sup>1</sup>.

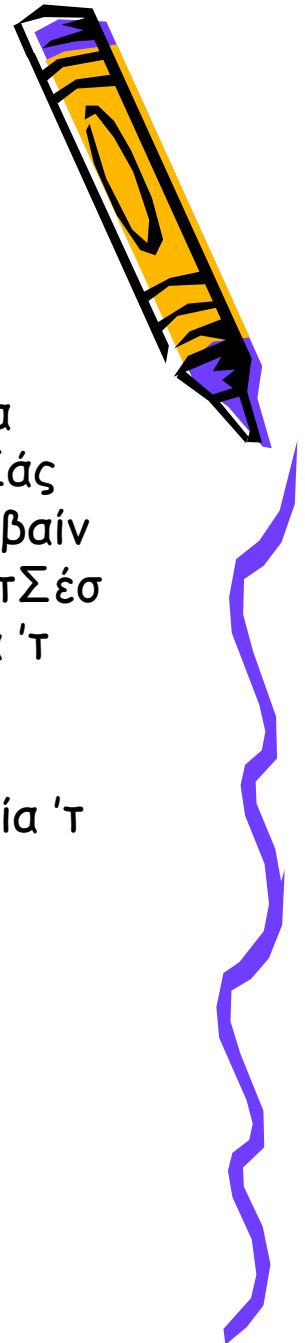
(ᾠθρικ)

Ἀτματσᾶς<sup>2</sup> καὶ ἡ ποθίκα<sup>3</sup> ἐποίκανε  
καρτασλόυκ<sup>4</sup>. Ἀτματσᾶς ἐποῖκε  
πουλία ἐπὶ σὴν πέτρα<sup>5</sup> καὶ ἡ πο-  
θίκα ἐποῖκε ἐπουκά σὴν πέτρα. Ὑ-  
στὲρ ἀτματσᾶς ἐκατέβε καὶ ἔφαιε τσῆ

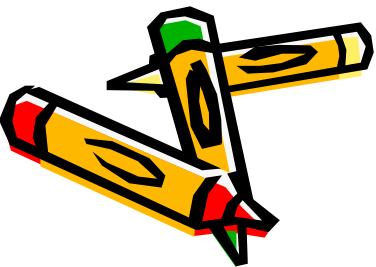
ποθίκα τὰ πουλία. Ἡ ποθίκα οὐκ ἐ-  
πόρευε ν' ἀνιβαίν ἐπὶ σὴν πέτρα  
νὰ τρώει τ' ἀτματσᾶ τὰ πουλία καὶ ἐ-  
πορπάτενε ἐπουκαῖεσ' καὶ ἐκλαιε. Ἐρθε  
βακοίτ<sup>6</sup> καὶ ἀτματσᾶς ἐποίησε νὰ βρῖσκ  
φαῖ γιὰ τὰ πουλία τ' ἐδέαβε σ' ἕνα  
κάμπο μερέα. Ἐκεῖ σὸν κάμπο ἐκά-  
θουσανε ἀργάτ καὶ ἔψηνανε κρέας ἀ-  
πανῖεσ' σ' ἄψιμο. Ἀτματσᾶς παλ  
ἐκατέβε ἐπῆρε ἕναν παρτσᾶ<sup>3</sup> κρέας ἀ-  
σ' ἄψιμο ἀπὸν καὶ ἔφερεν ἀ<sup>4</sup> στή φω-  
λέα τ. Μικερ<sup>5</sup> ἐκρατενε ἐκεῖ ἄψιμο.  
Ἐκολλίε<sup>6</sup> ἡ φωλέαν ἀτ, ἐρώξανε<sup>7</sup> ἐ-  
πουκά τὰ πουλία τ' καὶ ἔφαιεν ἀτα ἡ  
ποθίκα.







ατμασας τσε η ποθίκα εποίκανε καρτασλουκ. ατμασας εποίτσε πουλία  
επάν σην πέτρα τσε η ποθίκα εποίτσε επουκά σην πέτρα. υστέρ ατμασάς  
εκατέβε τσε έφαε τση ποθίκας τα πουλία. η ποθίκα ουτσ επόρενε ν' ανιβαίν  
επάν σην πέτρα να τρώει τ' ατμασά τα πουλία τσε επορπάτενε επουκατσέσ  
τσε έκλαιε. έρθε βακΙτ τσε ατμασάς εποίε να βρίσκ φαΐ για τα πουλία 'τ  
τς' εδΑβε σ' ένα κάμπο μερέα. ετσεί σον κάμπο εκάθουσανε αργάτ και  
έψηνανε κρέας απαντσέσ σ' άψιμο. ατμασάς παλ εκατέβε επήρε έναν  
παρτσά κρέας ας άψιμο. εκολλίε η φωλέαν ατ', ερώξανε επουκά τα πουλία 'τ  
τσε έφασεν ατα η ποθίκα.



# Carpathocian (Καππαδοκικά)



· Τότε πιάσαν βουδαχχήρε να κόψουν το φαβάχ. Κόφτουν το φαβάχ. Δέν πλερούται· πλεμνίσκει λιγόδικο. Το παλτά σακουται. Τότε πιάνουν ένα jadé φαρά· έδωκάν δο ένα πολλά σταφίρες νά τα πλύν. Τα καλά έπέτανέν da, και τα κōτία βαήνεν da. Το κορίσ λέχ το, “Όί ζάεις; τα καλά πετάνεις τα, και τα κōτία στέγγουν.” “Όί να ποίκω; Δέ χιωρῶ.” Σόνγρα πιάνουν ένα βασκά jadé φαρά, και δίνουν δο, να ζυμῶς ζυμάρ. Ζύμωνέν δο μέ το πράϊ τ. Όί ζάεις;” λέχ το κορίσ. “Μέ το πράχ ζυμοῦται ζυμάρ μί;” λέχ. Τότε το κορίσ κατέβη και ζύμωσέν δο. Σόνγρα νανέβη. Δέν δο βάκε· πιάσεν da ἄς τα μαλιά τ. Τότε ήρτε πατισαχιού το παιρί· πήρεν δο. Και σεράνδα μέρες έπκαν γάμος.





τότε πιάσαν μπουδαχτσήρε να κόψουν το ραβάχ. Κόφτουν το ραβάχ. δεν πλερούται. πλεμνίσκει λιγότησικο. το παλτά σακούται. τότε πιάνουν ένα dZadé ραρά. έδωκάν do ένα πολά σταφίρες να τα πλύν. τα καλά επέτανέν da, και τα κΟτία βαήνεν da. το κορίΣ λέχ το, «τΣί ζάεις; τα καλά πετάνεις τα, και τα κΟτία στέγνουν». «τΣί να ποίκω; δε χιωρώ». σόνγρα πιάνουν ένα βασκά dZadé ραρά, και δίνουν do, να ΖυμώΣ Ζυμάρ. Ζύμωνέν do με το πράι τ'. «τΣί ζάεις;» λέχ το κορίΣ. «με το πράχ Ζυμούται Ζυμάρ μι;» λέχ. τότε το κορίΣ κατέβη και Ζύμωσέν do. σόνγρα νανέβη. δεν το βάκε. πιάσεν da ας τα μαλλιά τ. τότε ήρτε πατιΣαχιού το παιρί. πήρεν do. και σεράνδα μέρες έπκαν γάμος.



# Αϊναλιότ (Αϊβαλιώτικα)

Μν'ὰ φουρά η̄δαν ένας βασίλης τσ' εἶχι τς  
τοῦ τσιφάλ' ένα τσιρατέλ' τσι τοῦ εἶχι πουλὸ  
ἀκουφά. "Οποιοὺν βιρβέρο ἐπιονι νὰ τοῦ γουρέψ,  
τοὺν ἔκανι τιβίχ<sup>1)</sup> νὰ μὴ τοῦ λέγ ὄξου. Τώρα  
οὐλ' οἱ βιρβέροδισ δὲν ἰβουροῦσαν νὰ τοῦ βαστά-  
ξιν ἀκουφά· ἵ' ἀφτὸ τς ἔσφαξι.

Πίσου πίσου πῆρι ένα βιρβέρο, τσι σὰ δοῦ  
ἀπουκούριψι, τ εἶπι, νὰ μὴ τοῦ πῆ σὶ κανέναν,  
ποῦς ἔχ τσέρατου, ἵατὶ θὰ πάρ τοῦ τσιφάλ' τ.  
'Ἡ βιρβέρος δὲν ἰβόροσι νὰ βαστάξ, πῆρι, ἔστουψι  
μὲς ένα πγάδ τσι φώναξι μ' οὐλ' τ γαρδιά τ:  
„Ἡ βασίλης ἔχ τσιρατέλ'." Τώρα τοῦ πγάδ ξι-  
ράθτσι, φύτροουσι μέσα μν'ὰ καλαμν'ά. Μιγάλ'νι  
ἢ καλαμν'ά. Πέρα μν'ὰ μέρα ένας δζουβάν'ς,  
ἔκουψι ἄ γαλαμν'ά τσ' ἔκανι μν'ὰ τσαβούνα τσι  
την ἔπιξι. 'Ἡ τσαβούνα ἤλιγι: „Βί! ἰ βασίλης  
ἔχ τσιρατέλ'." Τοῦ ἤκσαν, τοῦ εἶπαν τ βασίλέ.





μια φουρά ήδαν ένας βασιλέσ τσ' είχι στου τσιφάΛ ένα τσιρατέΛ τσι του είχι πουλύ ακρυφά. όποιουν βιρβέρ έπιρني να του γουρέψ, τουν έκανι τιβίχ να μη του λέΓ' όξου. τώρα ούΛ οι βιρβέρδισ δεν ιμπουρούσαν να του βαστάξιν ακρυφά. γί αυτό τς έσφαξι. πίσου πίσου πήρι ένα βιρβέρ, τσι σα δου απουκούριψι, τ' είπι να μη του πη σι κανέναν, πους έχ' τσέρατου, γιατί θα πάρ' του τσιφάΛ τ. ι βιρβέρς δεν ιμπόρσι να βαστάξ, πήγι, έστουψι μες ένα πγάδ τσι φώναξι μ'ούΛ τ' γαρδιά τ': «ι βασιλέσ έχ' τσιρατέΛ». τώρα του πγάδ ξιράθτσι, φύτρουσι μέσα μια καλαμιά. μιγάΛνι η καλαμνιά. πέρνα μια μέρα ένας τΖουβάΝς, έκουψι d' γαλαμιά τσ' ' έκανι μια τσαβούνα τσι τ'ν έπιζι. Η τσαβούνα ήλιγι: «Βί! ι βασιλέσ έχ' τσιρατέΛ». Του ήκσαν, του είπαν τ' βασιλέ.



# Morphological annotations

Morphological categories

Word

Grammatical category

(noun, verb, gerund, particle,  
adjective, pronominal, adverb, ...)

Special characteristics

Loan word

Origin

Archaism

Gender alteration

Other

Simple

Structured

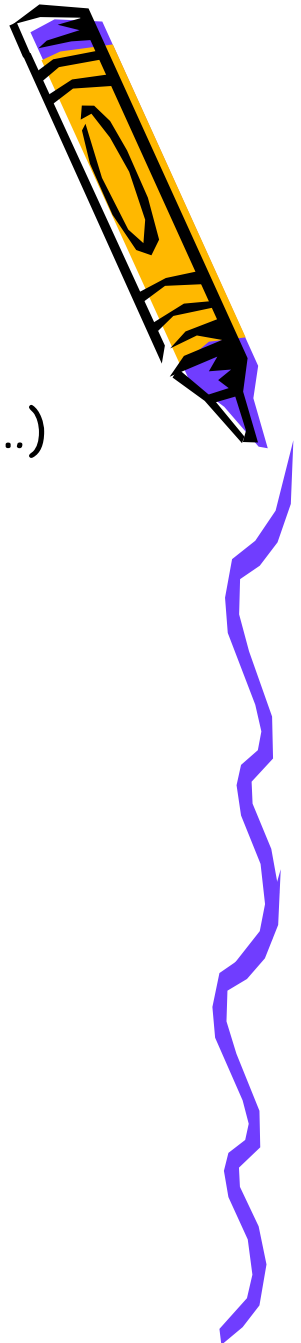
Declinable

Production

Composition

Merging

Other



# More Morphol. annotations

Morphological process

Declension

Noun

Number (αριθμός)

Gender (γένος)

Case (πτώση)

...

Verb

Person (πρόσωπο)

Number (αριθμός)

Tense (χρόνος)

Mood (έγκλιση)

Voice (φωνή)

...

Production

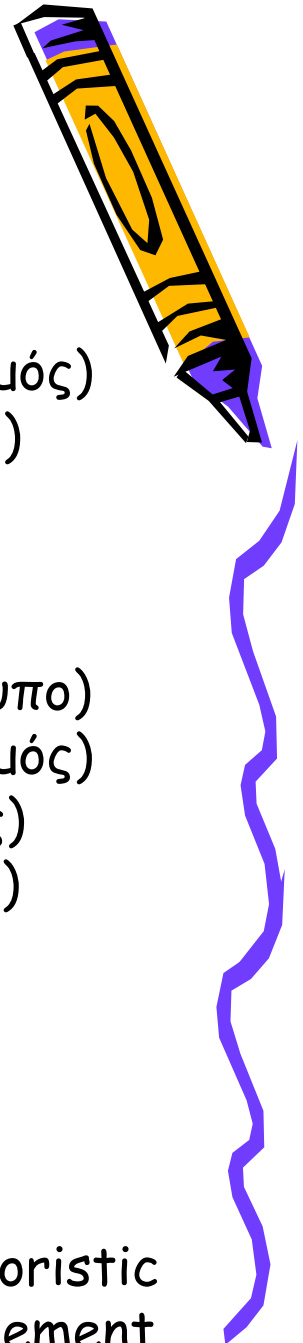
With Postfix

Noun

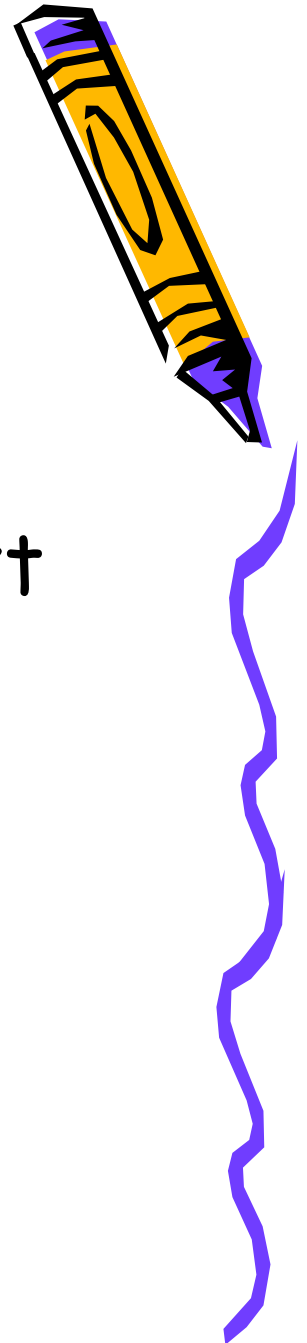
Hypocoristic

Enlargement

Verb



# Syntactic and Semantic annotations



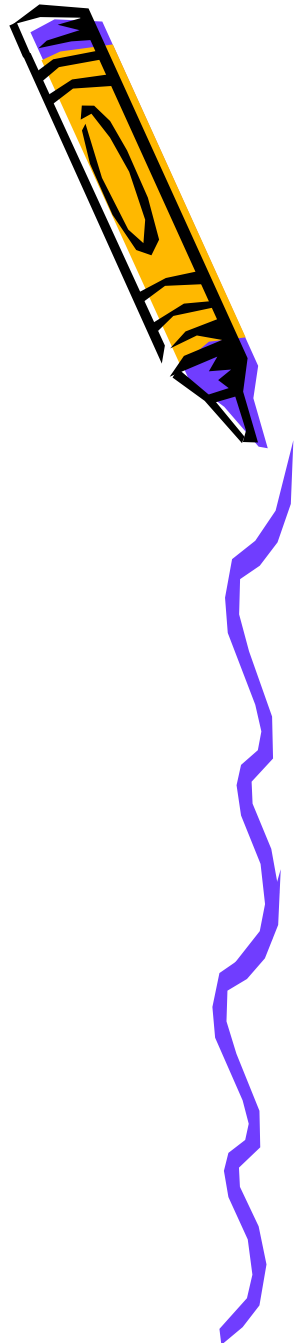
- Intentionally left blank
- It is of less interest in the context of AMiGre
- But, it is implemented





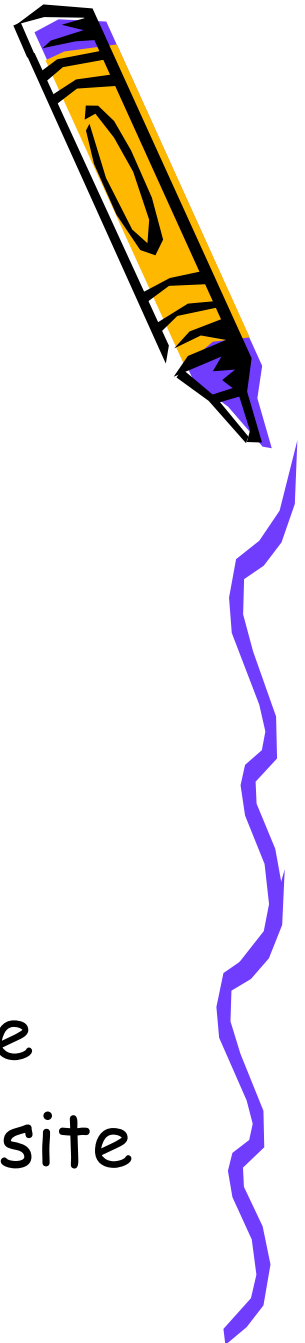
# AMiGre

- Introduction
- Sources
- **Applications Overview**
- Design Overview



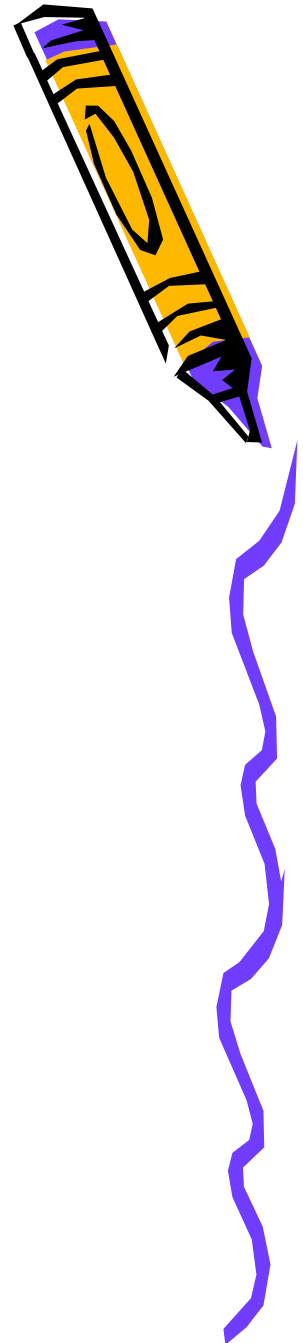
# Applications Overview

- Dialectal Lexicon (3 dialects) GUI
- Dialectal Lexicon Web Interface
- Oral + Written modules
- Written Sources GUI
- Oral Sources GUI
- Oral & Written Sources Retrieval GUI
- Oral & Written Browsing Web Interface
- Dissemination of effort & Results Web site



# Oral + Written - modules

Browse Oral
Browse Written
<b>Import Oral</b>
<b>Metadata for Oral</b>
<b>G. Oral = 3fold oral GUI</b>
<b>G. Written = 3fold written GUI</b>
Metadata for Written
<b>Page (ή Part) Import Written</b>
<b>Morphological Tagging</b>
Syntactic Tagging
Semantic Tagging
<b>Phonological Tagging</b>
<b>Text Imaging</b>
<b>Text Transcription</b>
Massively Import Oral
Massively Import Written
<b>Search &amp; Retrieve</b>



# Written Sources GUI - 3fold

display of annotated document together with attributes of the selected word



Μορφολογική Ανάλυση

Εφαρμογή Εργαλεία

Προβολή σελίδας 1 από 3

Προβολή Εικόνας  Προβολή ορίων λέξεων Προσθήκη νέας σελίδας Στην αρχή του κειμένου Προσθήκη Επεξεργασία Διαγραφή Επιστροφή στη λίστα

έσανε

δύς Σεράντ, είχανε απ' ένα παιδί. έστειλαν α στην κΣενιτεία. ο ένας είπε το παιδίν ατς: «άδεια μη κάσαι, έναν παρά πάλ αν ευρίσκεις, δουλέψο». η άλλη είπε το παιδίν ατς: «ασά είκοσ παράδΑΣ εξ ούκ μη δουλεύεις». επήγανε εδούλεψανε. ε κείνος οπ εδούλευε σ' έναν παρά αργατικό ουκ εχάσε. ο-γι-άλλο ουκ εύρε δουλεία να δουλεύ σα είκοσ παράδΑΣ. απάν σο χρόνο εκλώστανε να πάνε σ' οσπίτ. σο δρόμο είπαν «ας μετρούμε τα παράδΑΣ μουνα». εμέ τρεσανε. ένας είσε ελίγα, ο-γι-άλλο είσε πολλά. εκείνος οπ είσε ελίγα είπε τον άλλονα «αδά σον κόσμο ποίο κυριεύ, η ψευτιά γιόξα η αληθεία;» ο-γι-άλλο είπε «η αληθεία». εκείνος είπε «η ψευτιά». εποίκα νε κάβλ. ο είς είπεν «αν κυριεύ η ψευτιά, εγώ να δίγω σε τα παράδΑΣ». ο-γι-άλλο πάλ είπε «αν κυριεύ η αληθεία, εγώ πάλ να να δίγω σε τα παράδΑΣ». είπανε «ατάρα σάτι πάμε, ό,τινα τσατεύομε, ερωτούμε, να τρερούμε ποίο κυριεύ».

έσανε  
δύς  
Σεράντ  
είχανε  
απ'  
ένα  
παιδί  
έστειλαν  
α  
στην  
κΣενιτεία  
ο  
ένας  
είπε  
το  
παιδίν  
ατς  
άδεια  
μη  
κάσαι  
έναν  
παρά  
πάλ

Αποθήκευση Αναίρεση αλλαγών

Βασικές Πληροφορίες

Λήμμα είμαι

Σημασία

Μορφολογική Διαδικασία Κλίση-Ακλισία

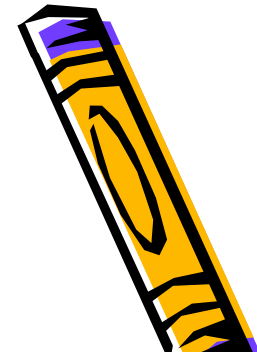
Γραμματική Κατηγορία Ρήμα

Γένος -

Κλιτική Τάξη -

Πρωτότυπη Δάνεια Λέξη

# Display with word borders



Μορφολογική Ανάλυση

Εφαρμογή Εργασία

Προβολή σελίδας 1 από 3

Προβολή Εικόνας  Προβολή ορίων λέξεων Προσθήκη νέας σελίδας Στην αρχή του κειμένου Προσθήκη Επεξεργασία Διαγραφή Επιστροφή στη λίστα

έσανε δύς Σεράντ, είχανε απ' ένα παιδί. έστειλαν α σην κΣεντεία. ο ένας είπε το παιδίν ατς: «άδεια μη κάσαι, έναν παρά πάλ αν ευρίΣκεις, δουλέψο». η άλλοε είπε το παιδίν ατς: «ασά είκοσ παράδΑΣ εξ ούκ μη δουλεύεις». επήγανε εδούλεψανε. εκείνος οπ εδούλευε σ' έναν παρά αργατικό ουκ εχάσε. ο-γι-άλλο ουκ εύρε δουλεία να δουλεύ σα είκοσ παράδΑΣ. απάν σο χρόνο εκλώστανε να πάνε σ' οσπίτ. σο δρόμο είπανε «ας μετρούμε τα παράδΑΣ μουνα». εμέ τρεσανε. ένας είσε ελίγα, ο-γι-άλλο είσε πολλά. εκείνος οπ είσε ελίγα είπε τον άλλονα «αδά σον κόσμο ποίο κυριεύ, η ψευτία γιόξα η αληθεία;» ο-γι-άλλο είπε «η αληθεία». εκείνος είπε «η ψευτία». εποίκα νε κάβλ. ο είς είπεν «αν κυριεύ η ψευτία, εγώ να δίγω σε τα παράδΑΣ». ο-γι-άλλο πάλ είπε «αν κυριεύ η αληθεία, εγώ πάλ να να δίγω σε τα παράδΑΣ». είπανε «ατώρα σάτι πάμε, ό,τινα τσατεύομε, ερωτούμε, ν

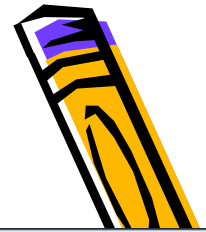
έσανε  
δύς  
Σεράντ  
είχανε  
απ'  
ένα  
παιδί  
έστειλαν  
α  
σην  
κΣεντεία  
ο  
ένας  
είπε  
το  
παιδίν  
ατς  
άδεια  
μη  
κάσαι  
έναν  
παρά  
πάλ

Αποθήκευση Αναίρεση αλλαγών

Βασικές Πληροφορίες

Λήμμα	είμαι
Σημασία	
Μορφολογική Διαδικασία	Κλίση-Ακλισία
Γραμματική Κατηγορία	Ρήμα
Γένος	-
Κλιτική Τάξη	-
Πρωτότυπη Δάνεια Λέξη	

# Annotated text together with the original image scan



Μορφολογική Ανάλυση

Εφαρμογή Εργαλεία

Προβολή σελίδας 1 από 3

Προβολή Λέξεων  Προβολή ορίων λέξεων Προσθήκη νέας σελίδας Στην αρχή του κειμένου Προσθήκη Επεξεργασία Διαγραφή Επιστροφή στη Λ

έσανε δὺς Σεράντ, εἶχανε ἀπ' ἑνα παιδί. ἔστειλαν α σην κΣε νιτειά. ο ἕνας εἶπε το παιδίν ατς: «ἀδεια μη κάσαι, ἕναν παρά πάλ αν ευρίΣκεις, δουλεψο». η ἄλλη εἶπε το παιδίν ατς: «ασά εἴκος παράδΑΣ εξούκ μη δουλεύεις». ἐπήγανε εδούλεψανε. ἐκεῖνο ς οπ εδούλευε σ' ἕναν παρά αργατικό ουκ εχάσε. ο-γι-ἄλλο ουκ ε ὕρε δουλεία να δουλεύ σα εἴκος παράδΑΣ. ἀπάν σο χρόνο εκλώστα νε να πάνε σ' οσπίτ. σο δρόμο εἶπανε «ας μετρούμε τα παράδΑΣ μ ουνα». ἐμέτρεσανε. ἕνας εἶσε ελίγα, ο-γι-ἄλλο εἶσε πολλά. ἐκε ἴνος οπ εἶσε ελίγα εἶπε τον ἄλλονα «αδά σον κόσμο ποῖο κυριεύ , η ψευτιά γιόξα η ἀληθεία;» ο-γι-ἄλλο εἶπε «η ἀληθεία». ἐκεῖ νος εἶπε «η ψευτιά». ἐποίκανε κάβλ. ο εἶς εἶπεν «αν κυριεύ η ψευτιά, ἐγώ να δῖγω σε τα παράδΑΣ». ο-γι-ἄλλο πάλ εἶπε «αν κυ ριεύ η ἀληθεία, ἐγώ πάλ να να δῖγω σε τα παράδΑΣ». εἶπανε «ατ ὠρα σάτι πάμε, ὄ,τινα τσατεύομε, ερωτούμε, να τερούμε ποῖο κυ ριεύ».

ἐπήγανε ετΣάτεψανε ἕναν ποπά νέο. ἐκεῖνος εἶπε «η ψευτιά κ υριεύ». εἶπανε ἐκεῖν «να ρωτούμε δὺς νοματούς κι ἄλλο». ἐρώτε σανε ἕνα μεσοκαιρίτε ποπά. ἐκεῖνος πάλ εἶπεν «η ψευτιά». το υ στερνὸ ἐρώτεσανε ἕνα γέρο ποπά. ἐκεῖνος πάλ εἶπε «η ψευτιά κυ ριεύ». ἐτότε το παιδί εδῶκε τα παράδΑΣ ατ τον ἄλλονα τον τε(μ )πέλ και ατός εκλώστα οπίσ, τΣούγκ ουκ εἶσε παράδΑΣ ν' ἐπέγινε σ' οσπίτ. ο-γι-ἄλλο ἐπῶσε τα παράδΑΣ ατ κυ ἐπερεν ατς τη μύν

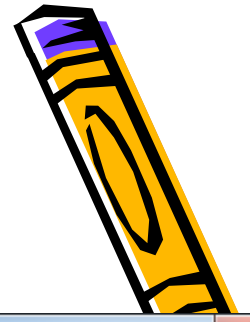
ΠΑΡΑΜΥΘΙΑ ΟΦΕΩΣ

1

Ἔσανε δὺς Σεράντ, εἶχανε ἀπ' ἑνα παιδί. Ἐστειλαν α σην κΣε νιτειά. Ο ἕνας εἶπε τὸ παιδίν ατς: «ἀδεια μὴ κάσαι, ἕναν παρά πάλ αν ευρίσκεῖς, δουλεψο». Η ἄλλη εἶπε τὸ παιδίν ατς: «ασά εἴκος παράδΑΣ εξούκ μη δουλεύεις». ἐπήγανον εδούλεψανον. ἐκεῖνος οπ εδούλευε σ' ἕναν παρά ἀργατικό οὐκ εχάσε. ο-γι-ἄλλο οὐκ ε ὕρε δουλεία να δουλεύ σα εἴκος παράδΑΣ. ἀπάν σο χρόνο εκλώστα νε να πάνε σ' οσπίτ. σο δρόμο εἶπανε «ας μετρούμε τα παράδΑΣ μ ουνα». ἐμέτρεσανε. ἕνας εἶσε ελίγα, ο-γι-ἄλλο εἶσε πολλά. ἐκεῖνος οπ εἶσε ελίγα εἶπε τον ἄλλονα «αδά σον κόσμο ποῖο κυριεύ, ἡ ψευτιά γιόξα ἢ ἀληθεία;» ο-γι-ἄλλο εἶπε «ἢ ἀληθεία». ἐκεῖνος εἶπε «ἢ ψευτιά». ἐποίκανε κάβλ. ο εἶς εἶπεν «αν κυριεύ ἢ ψευτιά, ἐγὼ να δῖγω σε τα παράδΑΣ». ο-γι-ἄλλο πάλ εἶπε «αν κυριεύ ἢ ἀληθεία, ἐγὼ πάλ να να δῖγω σε τα παράδΑΣ». εἶπανε «ἀτώρα σάτι πάμε, ὅτινα τσατεύομε, ἐρωτούμε, να τερούμε ποῖο κυριεύ».

Ἐπήγανον ἐπῶσανον ἕναν ποπά νέο. ἐκεῖνος εἶπε «ἢ ψευτιά κυριεύ». εἶπανον ἐκεῖν «να ρωτούμε δὺς νοματούς κι ἄλλο». ἐρώτεσανον ἕνα μεσοκαιρίτε ποπά. ἐκεῖνος πάλ εἶπεν «ἢ ψευτιά». το ὕστερον ἐρώτεσανον ἕνα γέρο ποπά. ἐκεῖνος πάλ εἶπε «ἢ ψευτιά κυριεύ». ἐτότε τὸ παιδί εδῶκε τα παράδΑΣ ατ τον ἄλλονα τον τε(μ)πέλ και ατός εκλώστα οπίσ, τΣούγκ ουκ εἶσε παράδΑΣ ν' ἐπέγινον σ' οσπίτ. ο-γι-ἄλλο ἐπῶσε τα παράδΑΣ ατ κυ ἐπερεν ατς τὴ μύνη».

# Insertion (selection) of an image for a written source



Προσθήκη Σελίδας

**Βήμα 1: Επιλογή Εικόνας & Εισαγωγή Κειμένου**

Εικόνα Σελίδας:

\* Κείμενο Σελίδας:

τ' ὄρωμαν για την πίταν.

τρει νομάτ συντρόφ εβραδασταν σ' έναν μέρος. έψεσαν έναν πίταν, έφααν τ' ημψόν και για τ' άλλο πα είπαν «θα τρώει ατο ήντισα ν ελέπ πολλά αdζαϊπκον ὄρωμαν. το πουρνόν εκάτσαν και λέγνε τ' ορώματα τουν.

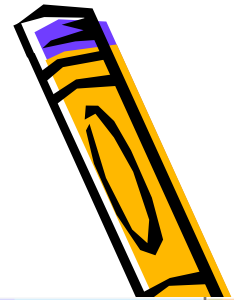
-εγώ, είπεν ο είνας, είδα εξέβα ψηλά ψηλά απάν σ' ουρανού την τΑπ'Αν.

-ατό τιδέν 'κι έν, είπεν άλλος, εγώ είδα εκατήβα αφκά σον κόλον τη ής.

-εγώ, είπεν ο τρίτον, όνταν είδα σας να πάτε ατόσον μακρά, ελογάριασα 'κι θα έρχουζνε και έφαα την πίταν.



# Symbols' selection in order to separate words of a transcription



Προσθήκη Σελίδας

## Βήμα 2: Εξαγωγή Λέξεων

\* Κείμενο Σελίδας:

τ' όρωμαν για την πίταν.

τρεί νομάτ συντρόφ εβραδᾶσαν σ' έναν μέρος. έψεσαν έναν πίταν, έφααν τ' ημψόν και για τ' άλλο πα είπαν «θα τρώει ατο ήντσαν ελέπ πολλά αδΖαίπκον όρωμαν. το πουρνόν εκάτσαν και λέγνε τ' ορώματα τουν.

-εγώ, είπεν ο είνας, είδα εξέβα ψηλά ψηλά απάν σ' ουρανού την τᾶπᾶν.

-ατό τιδέν 'κι έν, είπεν άλλος, εγώ είδα εκατήβα αφκά σον κόλον τη ής.

-εγώ, είπεν ο τρίτον, όνταν είδα σας να πάτε ατόσον μακρά, ελογάριασα 'κι θα έρχουζνε και έφαα την πίταν.

Σύμβολα Διαχωρισμού:

- (Regular Expression: '\s')
- , (Regular Expression: ',')
- . (Regular Expression: '\.')
- (Regular Expression: '\-')
- ? (Regular Expression: '\?')
- ; (Regular Expression: ';')
- ' (Regular Expression: '\')
- ' (Regular Expression: '"')
- ! (Regular Expression: '!')

Regular Expression:

(?!\\V[\\s,\\-;!«»:]

✕ Ακυρο

⏪ Προηγούμενο Βήμα: Εισαγωγή Καμένου & Εικόνας

✔ Επόμενο: Προεπισκόπηση



# Improved 3fold presentation



Μορφολογική Ανάλυση

Εφαρμογή Εργαλεία

Προβολή Εικόνας  Προβολή ορίων λέξεων Προσθήκη νέας σελίδας Στην αρχή του κειμένου Προσθήκη Επεξεργασία Διαγραφή Επιστροφή στη λίστα

Προβολή σελίδας 1 από 2

είς εφτωχός σίτε **έρτον** ασήν χαμελέτεν με το παιδί ν ατ, ενεγκάστεν και εκάτσεν κά ν' αναπάγετον κι ενεστέναξεν και είπεν «ωφ!» ευτύς εξέβεν ασ έναν σπέλον κέσ εις όφισ και είπεν «γιατί κουίεις με;» κι ο φτωχόν είπεν «εγώ εσέν 'κι κουίζω». κι επεκεί είπεν ο όφισ «γιατί 'κι δίς μ' από το παιδί σ', ασ μαθίζ στο τέχνης;» και με το λόγον εκείνον εδέκεν α κεί και ο όφισ επήρεν ατον κι επήγεν αφκά σην νηγήν απέσ σ' έναν σπέλον. εκΑπέσ ο όφισ είξεν έναν κ ουτσήν κι ατέ εγάπεσεν τον παιδάν. ύστερ από κα μπόσον καιρόν η κουτσή είπεν σον παιδάν «ο κύρ η μ' αν λέει σε, έμαθες τέχνην; εσύ πέ, 'κι έμαθα, ήνταν λέει σε, έμαθες; εσύ, 'κι έμαθα, πέ». ύστερ ον ερώτεσεν ο όφισ «έμαθες ακομάν τέχνης;» ο παιδίας είπεν «'κι έμαθα». ατότες ο όφισ εχτύπεσεν στον έναν σιλέν κι εχάταμεν ατον. άμαν ο παιδίας έμαθεν έτον τέχνης. έρθεν σον κύρν ατ και είπεν « τ'ΑΤΑ, εγώ ασ ίνομαι μουλάρ κι εσύ πούλτσον με, άμαν τηνάν πουλείς με το δουκάλι μ' μη δίς στον ». έντον μουλάρ και έρθεν εκείνος ο όφισ ν' αγορ άζ Ατον. ο κύρτς τη παιδά το μουλάρ εδέκεν και το

είς  
εφτωχός  
σίτε  
**έρτον**  
ασήν  
χαμελέτεν  
με  
το  
παιδί  
ν  
ατ  
ενεγκάστεν  
και  
εκάτσεν  
κά  
ν'  
αναπάγετον  
κι  
ενεστέναξεν  
και  
είπεν  
ωφ  
ευτύς  
εξέβεν  
ας  
έναν  
σπέλον  
κέσ  
είς  
όφισ  
και  
είπεν  
γιατί

### Μορφολογική Επισημείωση

ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ

Λήμμα: **έρχομαι**  
Μορφολογική Διαδικασία: **Κλίση-Ακλισία**  
Γραμματική Κατηγορία: **Ρήμα**  
Καταγωγή Λήμματος: **Ελληνική**

ΚΛΙΣΗ

Χρόνος: **Ενεστώτας**  
Αριθμός: **Πληθυντικός**

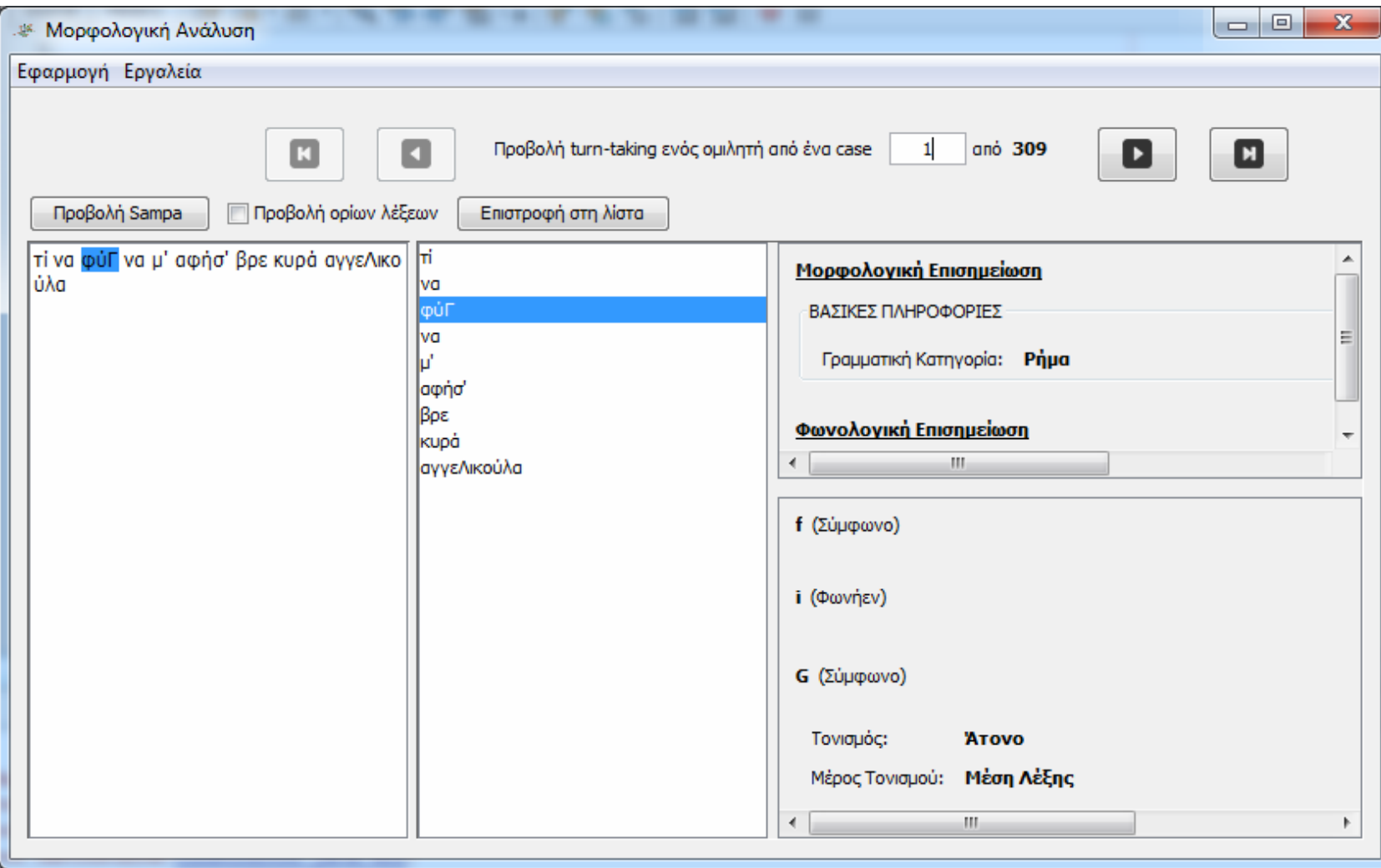
ΜΕΤΑ-ΠΛΗΡΟΦΟΡΙΕΣ

Περιοχή: **ΤΡΑΠΕΖΟΥΝΤΙΑΚΑ (Τραπεζούντα, Κερασούντα, Ριζούντα, Σούρμενα, Όφισ, Λιβερά, Τρίπολις, Ματσούκα)**

### Φωνολογική Επισημείωση

Φαινόμενα: **Ανομοίωση Συμφώνου**

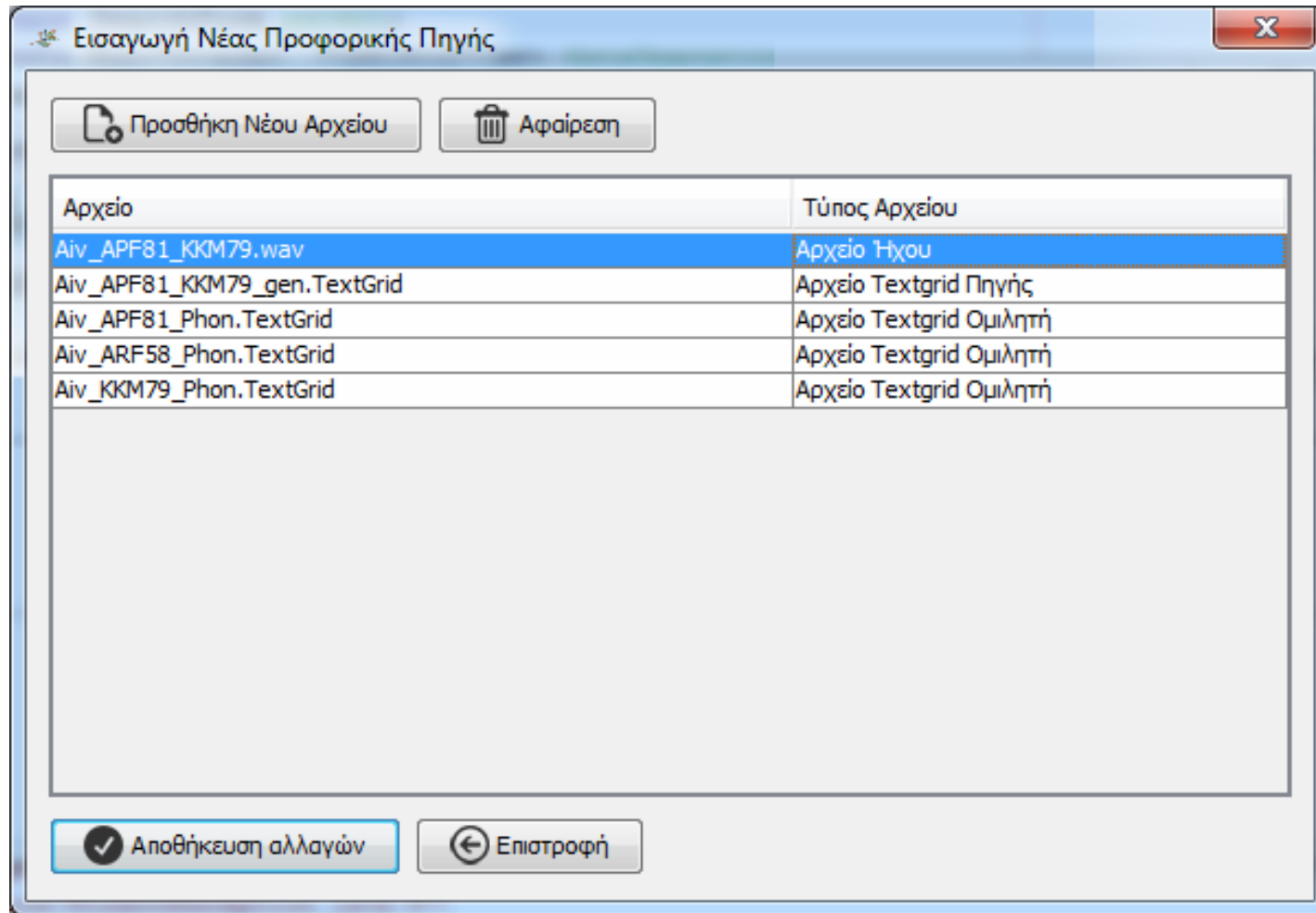
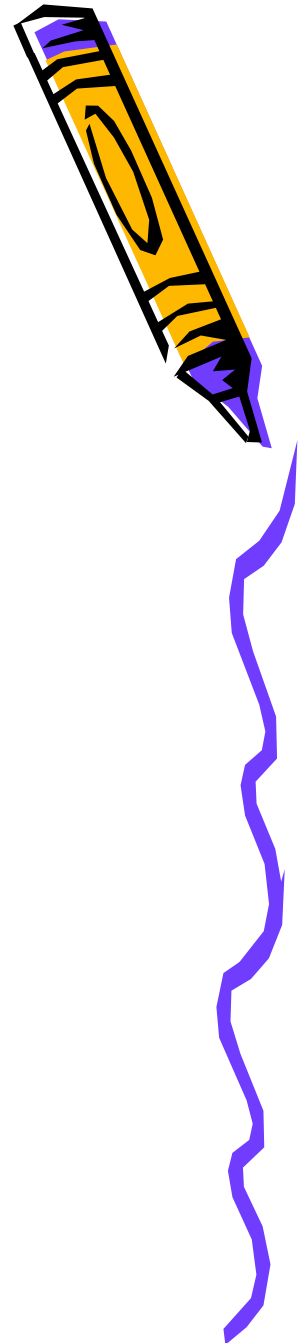
# Oral Sources GUI - 3fold



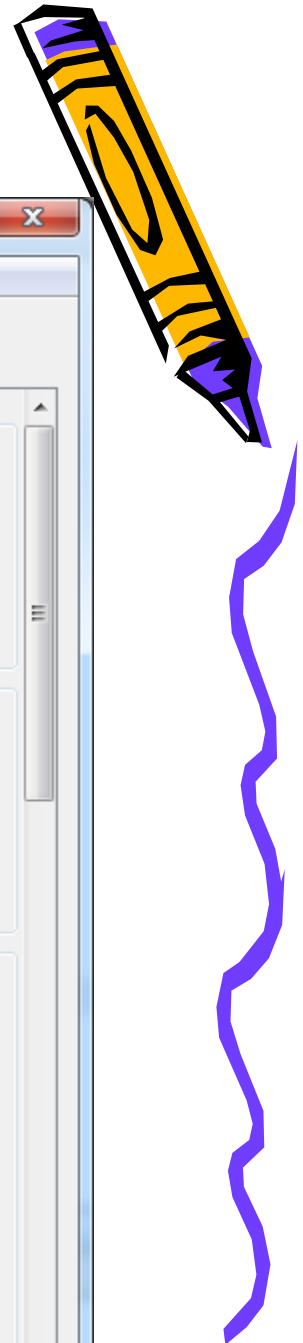
The screenshot shows a software window titled "Μορφολογική Ανάλυση" (Morphological Analysis). The interface includes a menu bar with "Εφαρμογή" and "Εργαλεία", a toolbar with navigation buttons, and a status bar indicating "Προβολή turn-taking ενός ομιλητή από ένα case 1 από 309". Below the toolbar are buttons for "Προβολή Sampra", "Προβολή ορίων λέξεων", and "Επιστροφή στη λίστα". The main area is divided into three panels: a text input field on the left containing "τί να φύγ να μ' αφήσ' βρε κυρά αγγελικο ύλα", a word list in the middle with "φύγ" highlighted, and a detailed analysis panel on the right. The analysis panel shows the morphological classification "Μορφολογική Επισημείωση" with "ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ" (Grammatical Category: **Ρήμα**) and the phonological classification "Φωνολογική Επισημείωση" (Phonetic features: **f** (Σύμφωνο), **i** (Φωνήεν), **G** (Σύμφωνο), Tonicity: **Άτονο**, Part of Tonicity: **Μέση Λέξης**).



# Oral Sources - Import



# Oral Sources - Metadata



Μεταδεδομένα: Επεξεργασία Ιδιοτήτων

Εφαρμογή Εργαλεία

Αποθήκευση αλλαγών  Επιστροφή

**ΣΤΟΙΧΕΙΑ ΑΡΧΕΙΟΥ**

Αύξων Αριθμός Αρχείου: 1

Όνομα Αρχείου: Αίν\_ΚΚΜ80

Θέση του αρχείου:

Ελεύθερο/Κλαδωμένο: Ελεύθερο

**ΔΙΑΛΕΚΤΟΣ**

Όνομα Διαλέκτου: Αιβαλιώτικα

Γεωγραφικός Προσδιορισμός Διαλέκτου: Αιβαλί Μικράς Ασίας

Τόπος Γέννησης:

Τόπος ανατροφής (πού μεγάλωσαν):

**ΕΡΕΥΝΗΤΙΚΟ ΠΡΟΓΡΑΜΜΑ**

Όνομα: AMIGRe

Πηγή Χρηματοδότησης: -

Διάρκεια ερευνητικού προγράμματος:

Επιστημονικός Υπεύθυνος: Αγγελική Ράλλη

Υπεύθυνος Έρευνας Πεδίου: Αγγελική Ράλλη

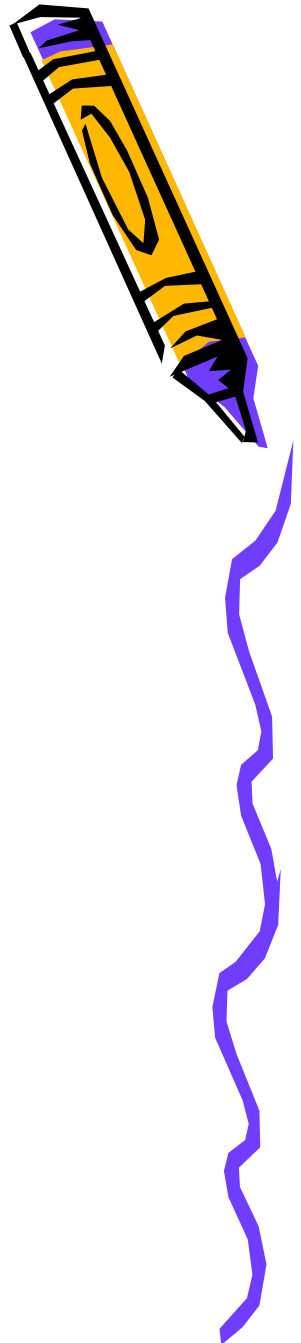
Ερευνητής Πεδίου:

- Αγγελική Ράλλη
- Δημήτρης Παπαζαχαρίου
- Δήμητρα Μελισσαροπούλου



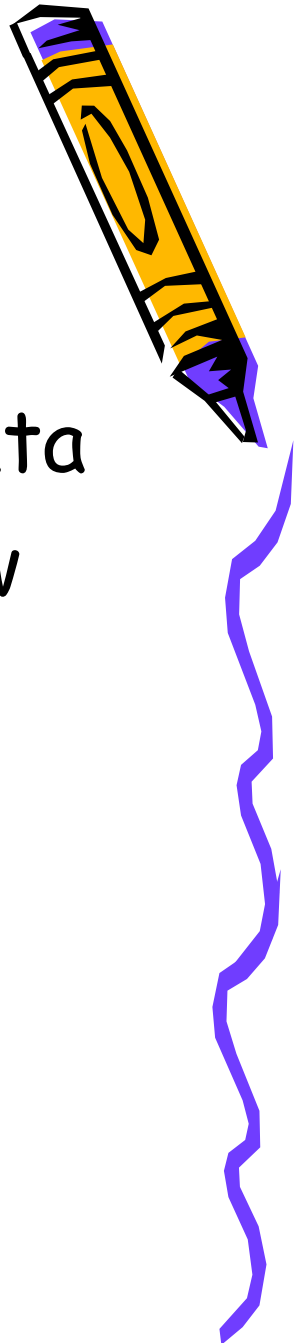
# AMiGre

- Introduction
- Sources
- Applications Overview
- **Design Overview**

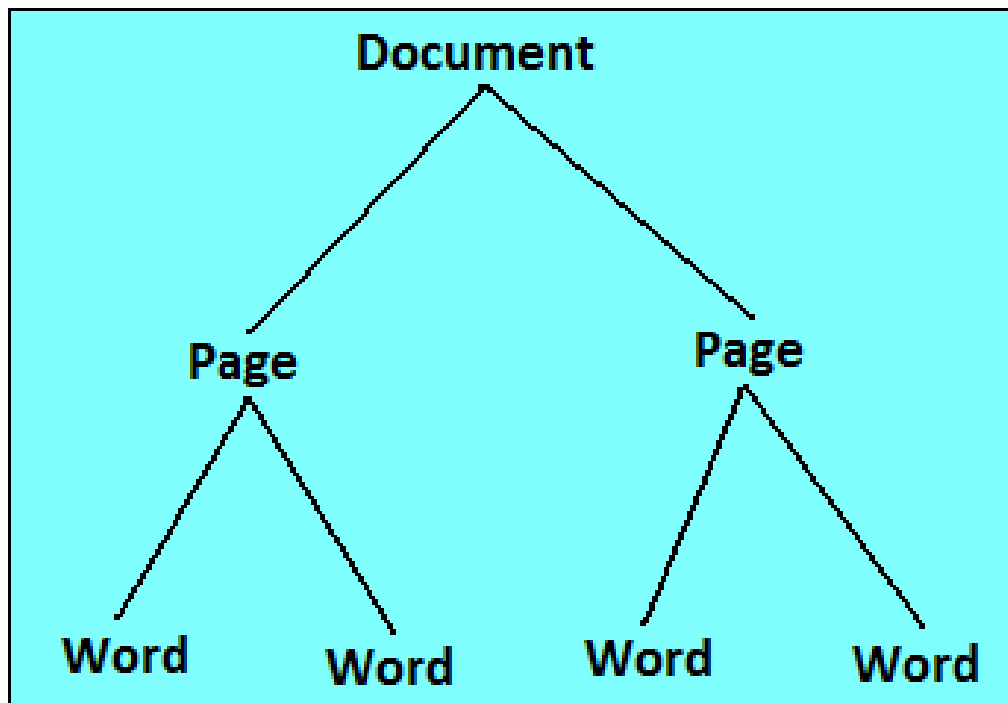


# Design Overview

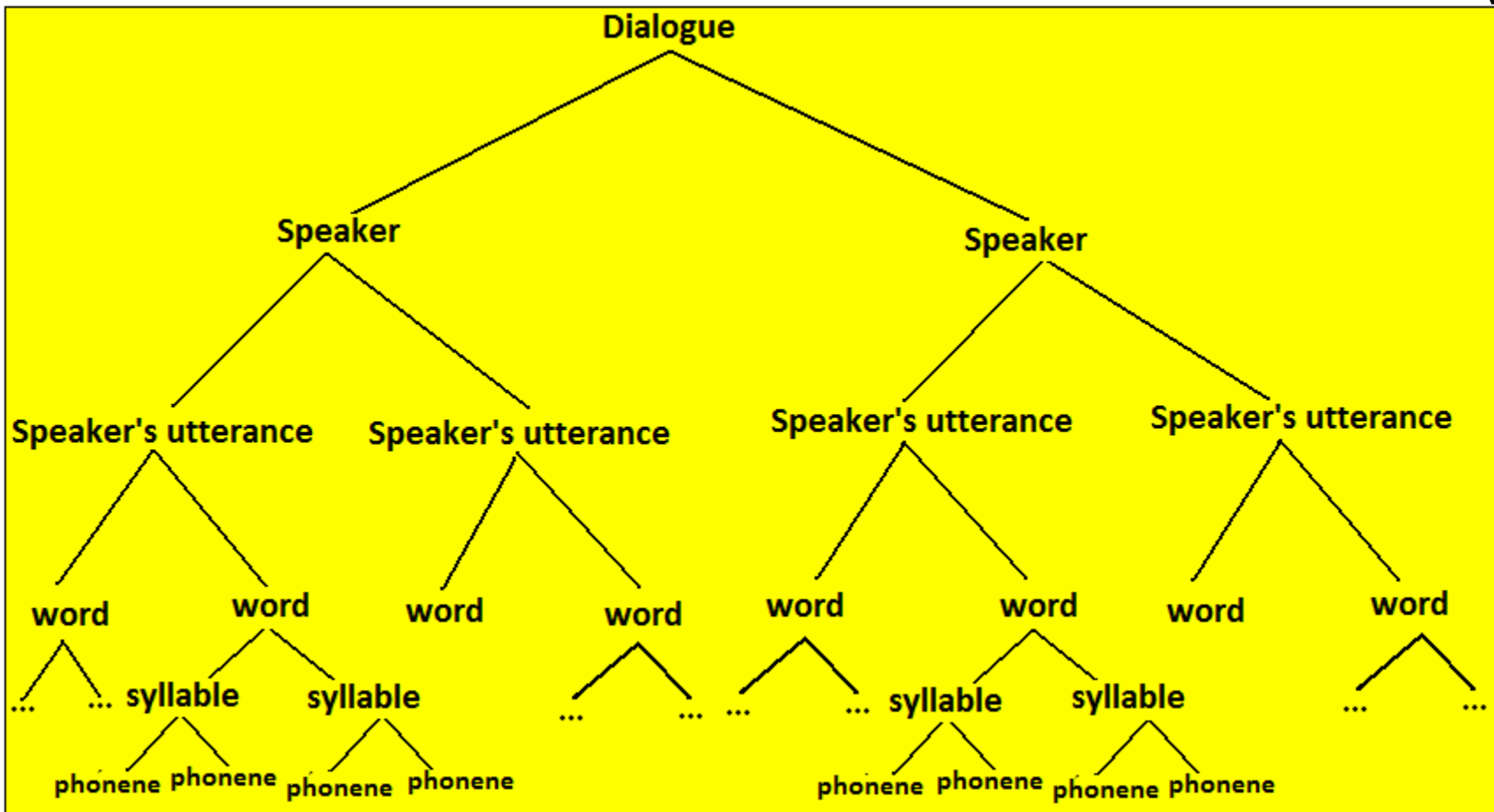
- Alignment of Oral and Written data
- Oral & Written - System overview
- Struct (relational) databases
- EAV data structures



# Alignment of Oral and Written - structure of Written data

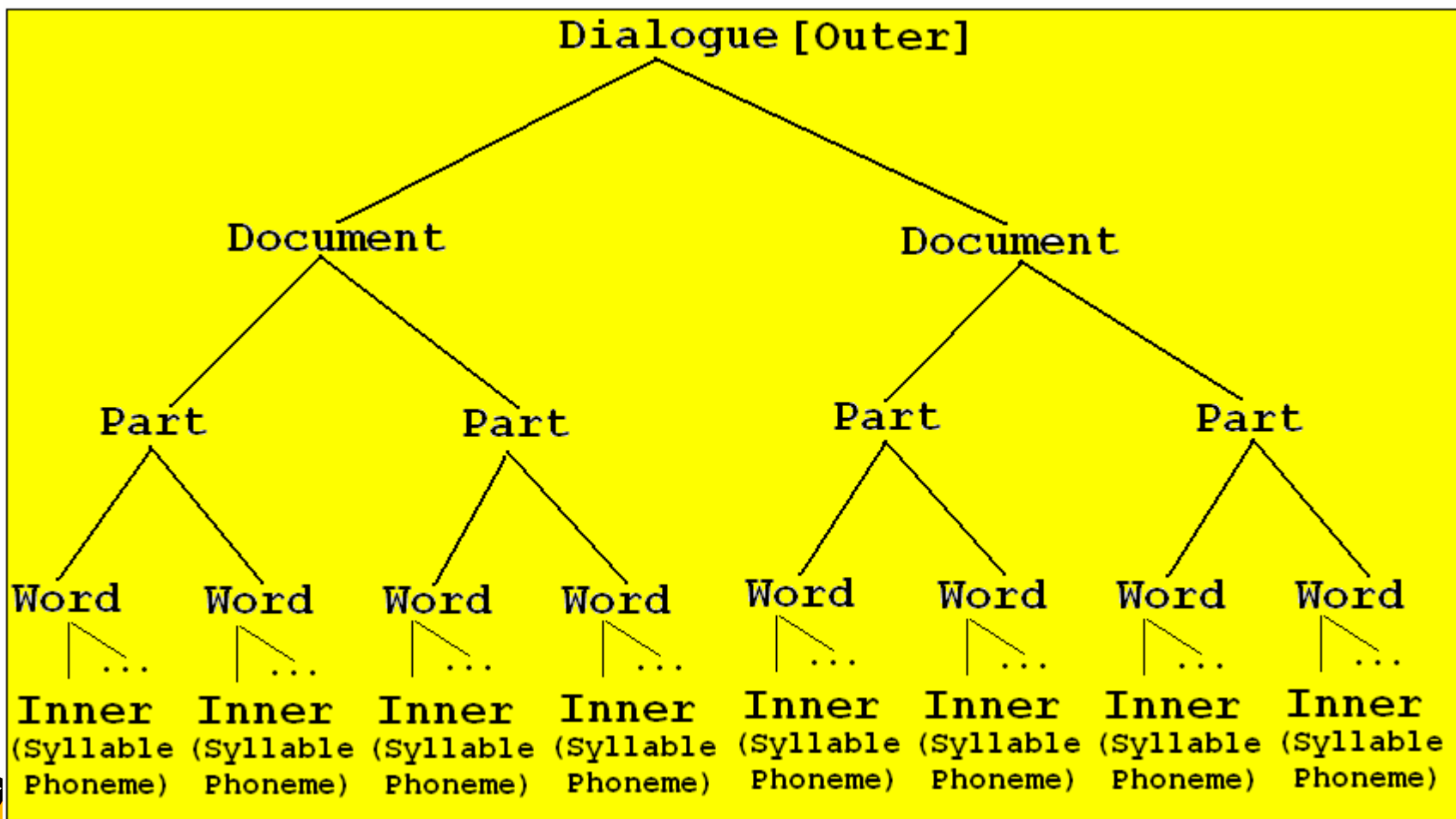


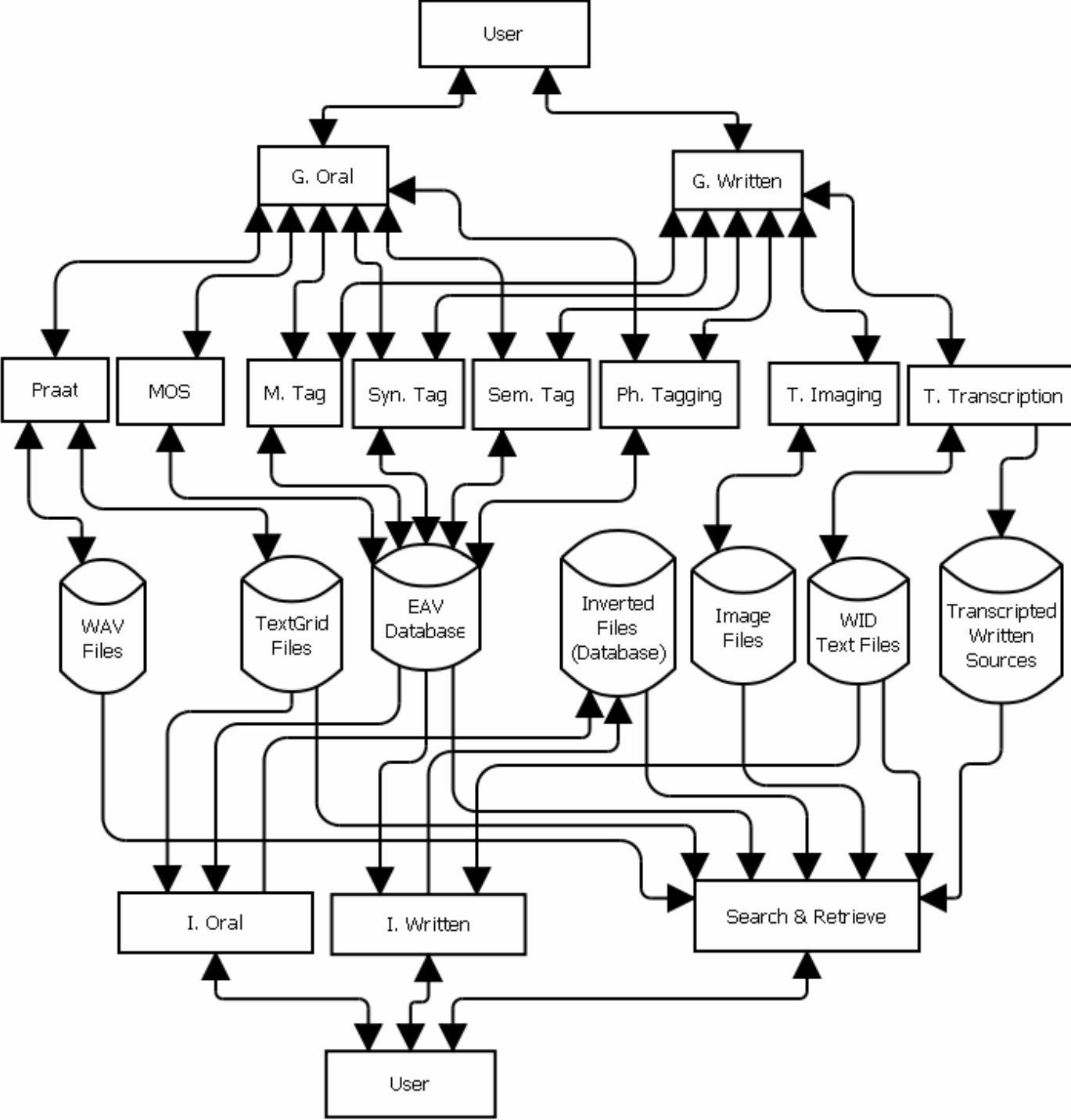
# Alignment of Oral and Written - structure of Oral data



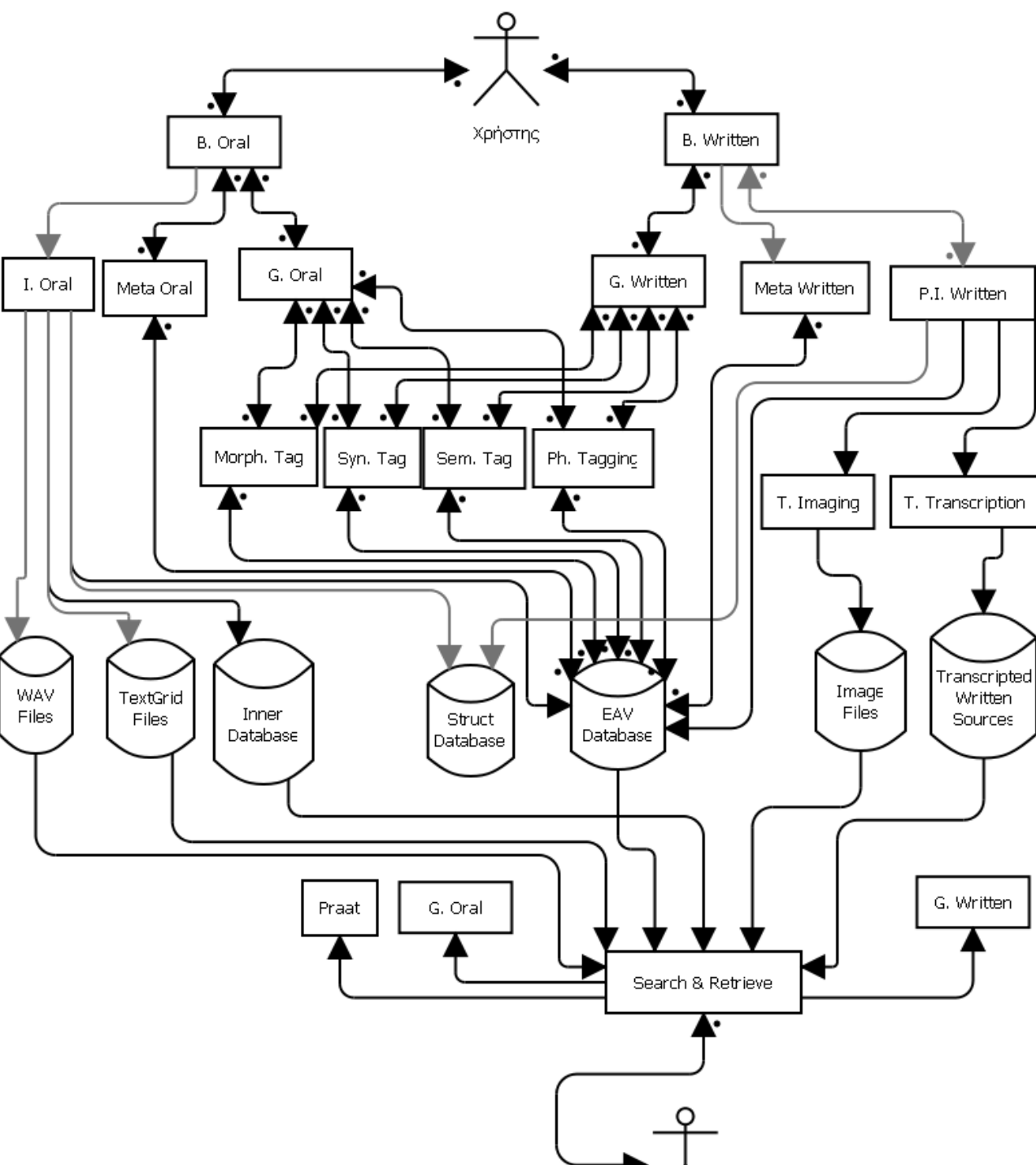


# Alignment of Oral and Written - common structure





Oral &  
Written  
Sources -  
System  
overview  
- OLD



Oral &  
Written  
Sources -  
System  
overview -  
updated

# Struct database

- Within the *Struct* database, the components of the documents are organized on consecutive levels of refinement which will be annotated with the help of the *EAV* database.
- The implementation of the abstract structure of the *Struct* database uses two quasi similar relational schemas.
- The only difference is that the implementation for the oral documents *Struct oral* is composed of all 5 levels, while the implementation for the written documents *Struct written* contains only the 3 intermediate levels.

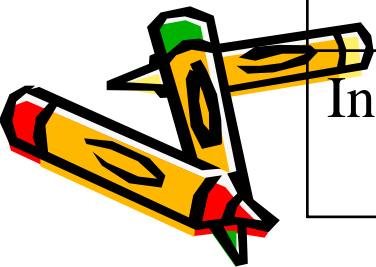


# Alignments of Struct (oral and written) databases



- Next table define the data alignments between the two Struct databases (*Struct oral* and *Struct written*):

Abstract name	Oral	Written
Dialog	Overall oral document	--
Document	Speaker / interlocutor	Overall written document
Part	Speaker's utterance	Page of written document
Word	Morphological word	Morphological word
Inner	Syllables, vowels, consonants etc,	--



# Struct db for oral documents



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)

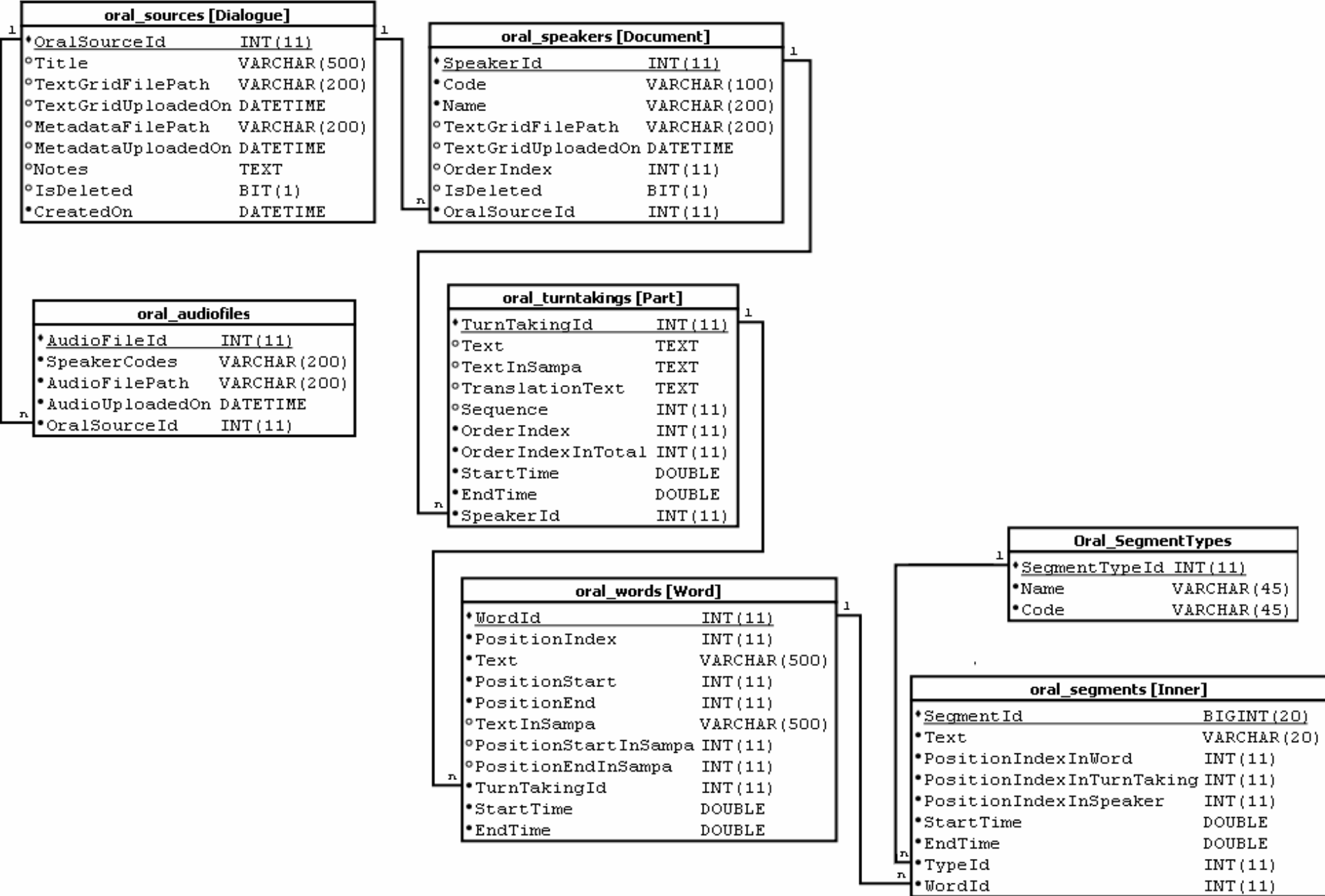
oral_audiofiles	
*AudioFileId	INT(11)
*SpeakerCodes	VARCHAR(200)
*AudioFilePath	VARCHAR(200)
*AudioUploadedOn	DATETIME
*OralSourceId	INT(11)

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)

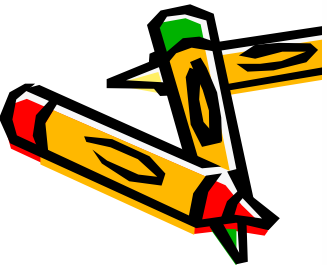
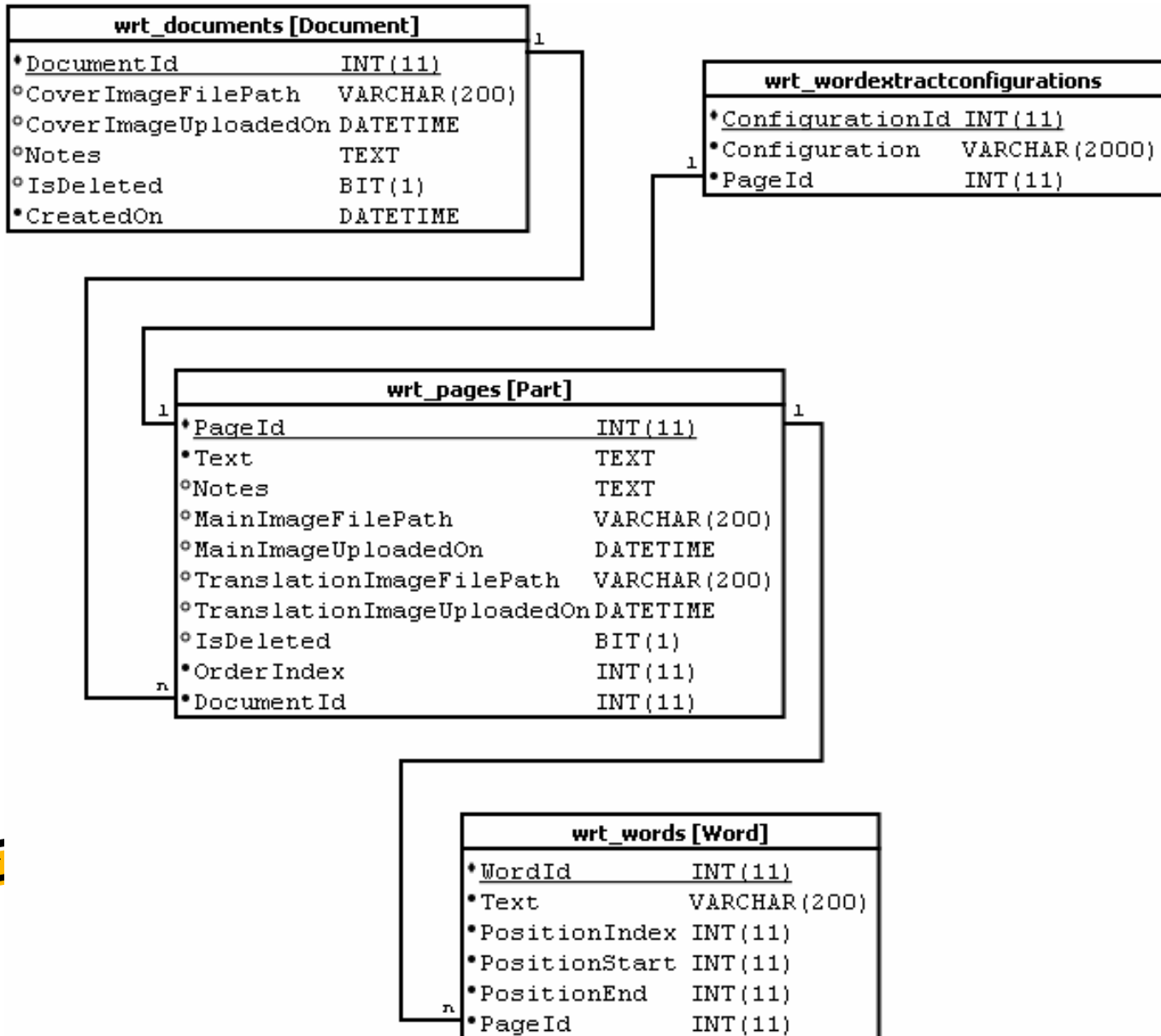
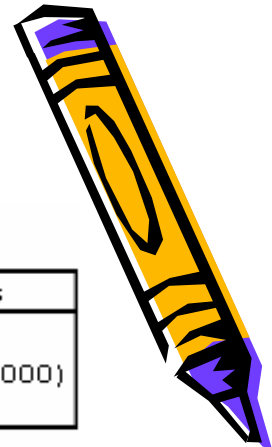
oral_words [Word]	
*WordId	INT(11)
*PositionIndex	INT(11)
*Text	VARCHAR(500)
*PositionStart	INT(11)
*PositionEnd	INT(11)
°TextInSampa	VARCHAR(500)
°PositionStartInSampa	INT(11)
°PositionEndInSampa	INT(11)
*TurnTakingId	INT(11)
*StartTime	DOUBLE
*EndTime	DOUBLE

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

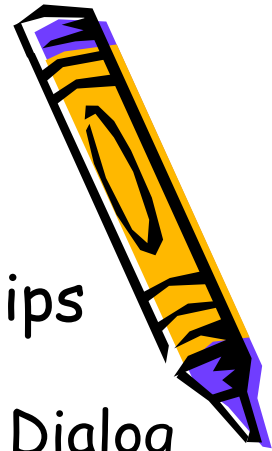
oral_segments [Inner]	
*SegmentId	BIGINT(20)
*Text	VARCHAR(20)
*PositionIndexInWord	INT(11)
*PositionIndexInTurnTaking	INT(11)
*PositionIndexInSpeaker	INT(11)
*StartTime	DOUBLE
*EndTime	DOUBLE
*TypeId	INT(11)
*WordId	INT(11)



# Struct db for written documents



# Struct dbs some explanations



- In both db schemas, one-to-many relationships exist for each pairs of levels. Examples are:
  - many speakers/interlocutors participate in a Dialog (the association between *oral\_sources* and *oral\_speakers* is 1: n)
  - a written document is composed of many pages (the association between *wrt\_documents* and *wrt\_pages* is 1:n).
- A number of auxiliary tables are also included in the diagrams:
  - Table *oral\_audiofiles* is used for storing one or more digital audio files associated to an oral document.
  - Table *oral\_SegmentTypes* contains segments of a morphological word (syllables, phonemes etc).
  - Table *wrt\_wordextractconfigurations* is used for storing the tokenization configuration of a written document page (separators, regular expressions etc)



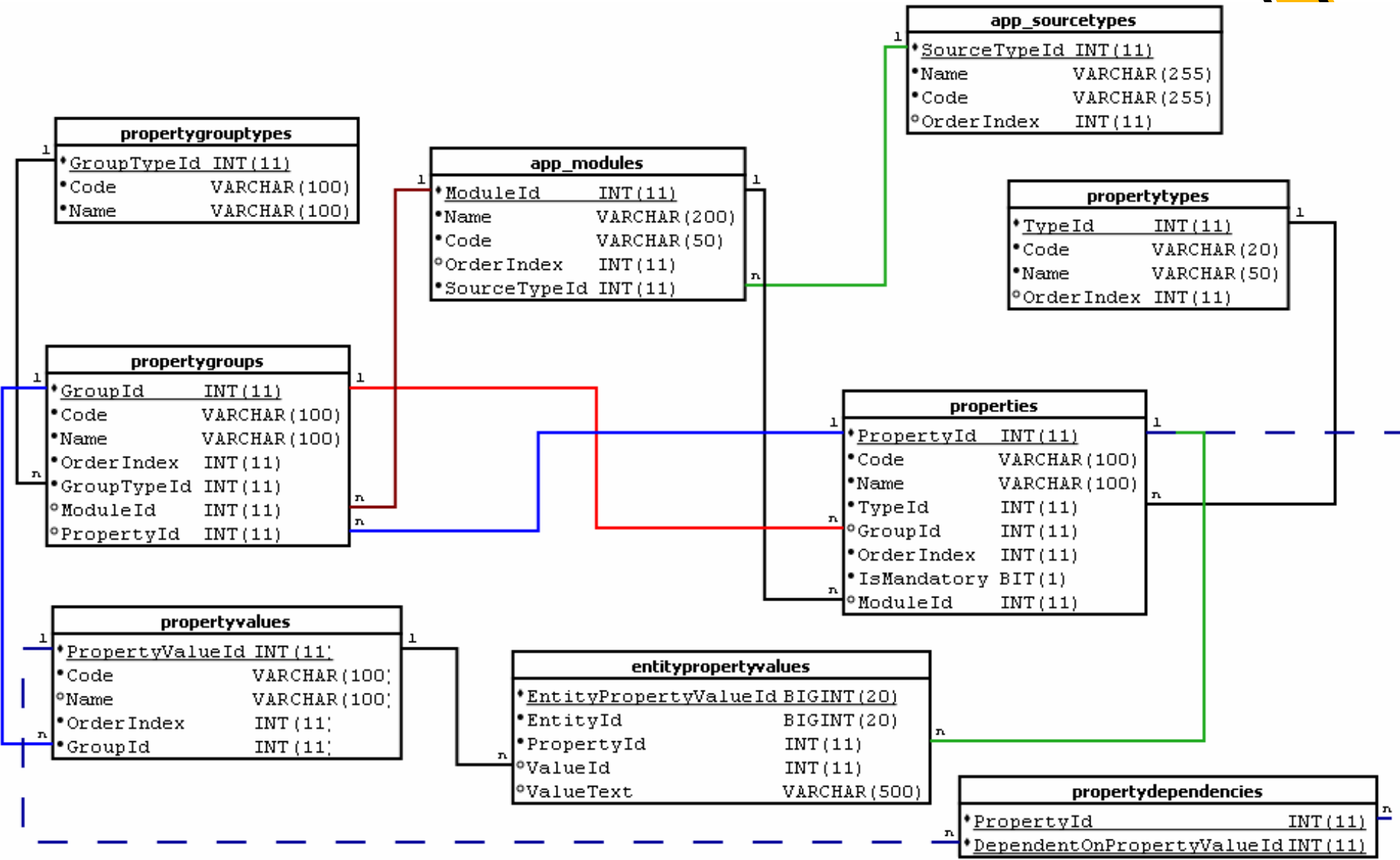
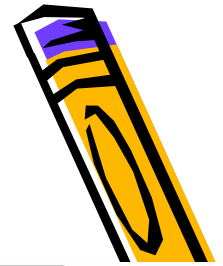


# EAV database

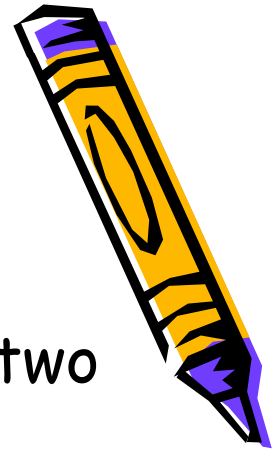
- The EAV database keeps record of all kinds of annotation.
- It is based on the Entity-Attribute-Value representation aiming at the exemption from the continual Schema evolution problem and the extended use of null values.
- Annotations concern entities (tuples) of the database tables *oral\_speakers*, *oral\_turntakings*, *oral\_words*, *oral\_segments*, *wrt\_documents*, *wrt\_pages* and *wrt\_words*.
- In other words, the entity (E of EAV) takes its values from the primary keys of the 7 abovementioned tables.
- As all annotations have the same functionality all annotation modules could be merged into one which would update a different set of attributes.
- In addition, meta-information could be managed by the same module since they share the same functionality but they update different sets of attributes.
- So, all 11 modules, i.e. 9 annotation modules (word part annotation for oral documents, and phonological, morphological, syntactic and semantic for both types of documents) and 2 meta-information modules could be merged into one which is applied to the set of attributes it applies to.



# EAV schema which supports all the requirements



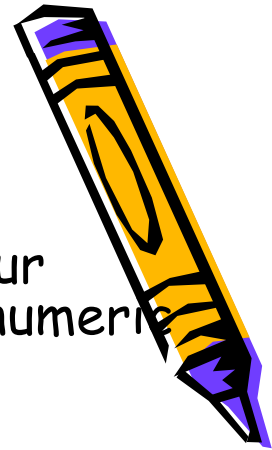
# Brief description of the 9 tables consisting the EAV



1. The *app\_sourcetypes* table contains the two document types (oral / written).
2. The *app\_modules* table defines the 11 modules needed for the processing.
3. The *propertygroups* table is used for two reasons: (a) the definition of thematic subsets of attributes and (b) the definition of predefined values of attributes (lookups).
4. The auxiliary *propertygroupstypes* table contains the two reasons for which the *propertygroups* table is used.
5. The *properties* table contains all properties used by the 11 modules.



# Brief description of the 9 tables consisting the EAV



6. The auxiliary *propertytypes* table contains the four possible types of a property (alphanumeric, alphanumeric with multiple values, predefined value, multiple predefined value).
7. The *propertyvalues* table contains all acceptable values of lookups.
8. The *entitypropertyvalues* table is the main table of the EAV database. The *EntityId* field contains the Entities and takes its values among the primary keys of the 7 primary tables of the Struct database, the *PropertyId* contains the Attributes and takes its values from the primary keys of the properties table. The *ValueId* or the *ValueText* field contains the Values. The value domain of the *ValueId* is the primary key of the *propertyvalues* table. In the case where Attribute defined by the *PropertyId* field is an alphanumeric the *ValueText* is filled instead of the *ValueId*.



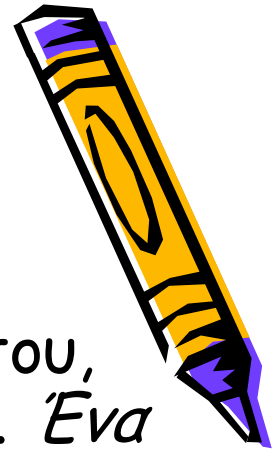
9. The *propertydependencies* table contains the properties which appear under the constraint that another property has a particular value. If an property does not appear in the *propertydependencies* table, then it is constraint free and it always appear in the module it is assigned to.

# See also

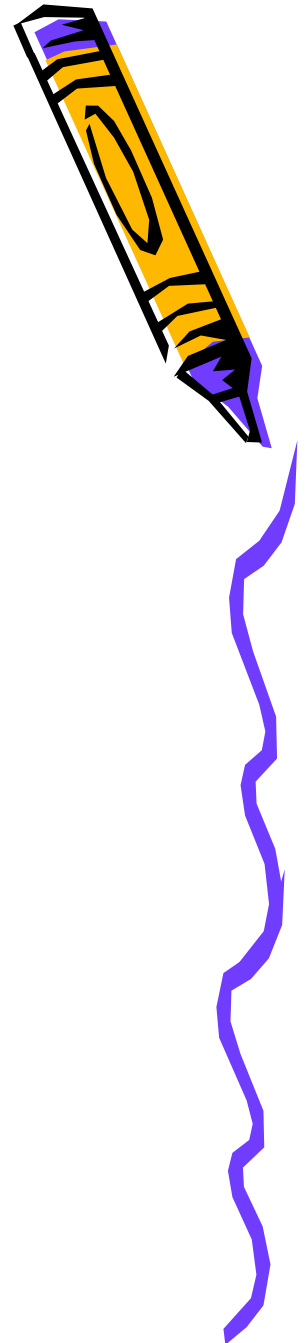
- (in Greek) Νικήτας Ν. Καρανικόλας, Ελένη Γαλιώτου, Κωνσταντίνος Αθανασάκος & Γεώργιος Κορωνάκης. Ένα πολυτροπικό σύστημα αρχειοθέτησης και διαχείρισης γραπτών και προφορικών πηγών μελέτης της γλώσσας και των γλωσσικών ιδιωμάτων. Στο Αγγελική Ράλλη, Πρόγραμμα Θαλής: Πόντος, Κατπαδοκία, Αϊβαλί: Στα Χνάρια της Μικρασιατικής Ελληνικής, ISBN 978-960-99426-2-1.

[http://users.teiath.gr/nnk/papers/C03\\_CR.pdf](http://users.teiath.gr/nnk/papers/C03_CR.pdf)

[http://users.teiath.gr/nnk/papers/C03\\_extended.pdf](http://users.teiath.gr/nnk/papers/C03_extended.pdf)



# Closing



- Thank you for your attention!
- Questions can be asked.
- [nnk@teiath.gr](mailto:nnk@teiath.gr)
- <http://users.teiath.gr/nnk/>

