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

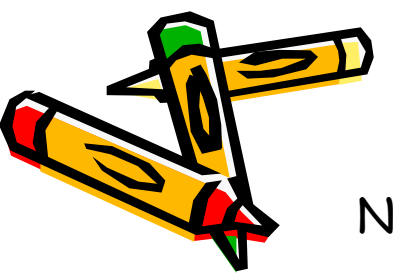
NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

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.

AMiGre

- Introduction
- Sources
- Applications Overview
- Design Overview



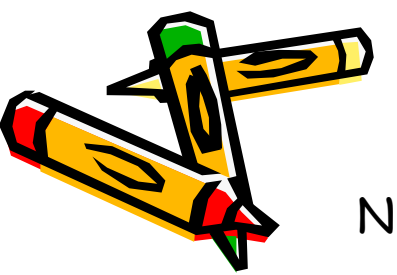
Introduction

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.

AMiGre

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

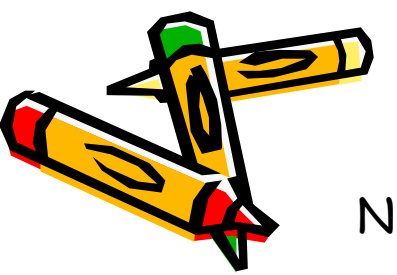


Sources

- 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

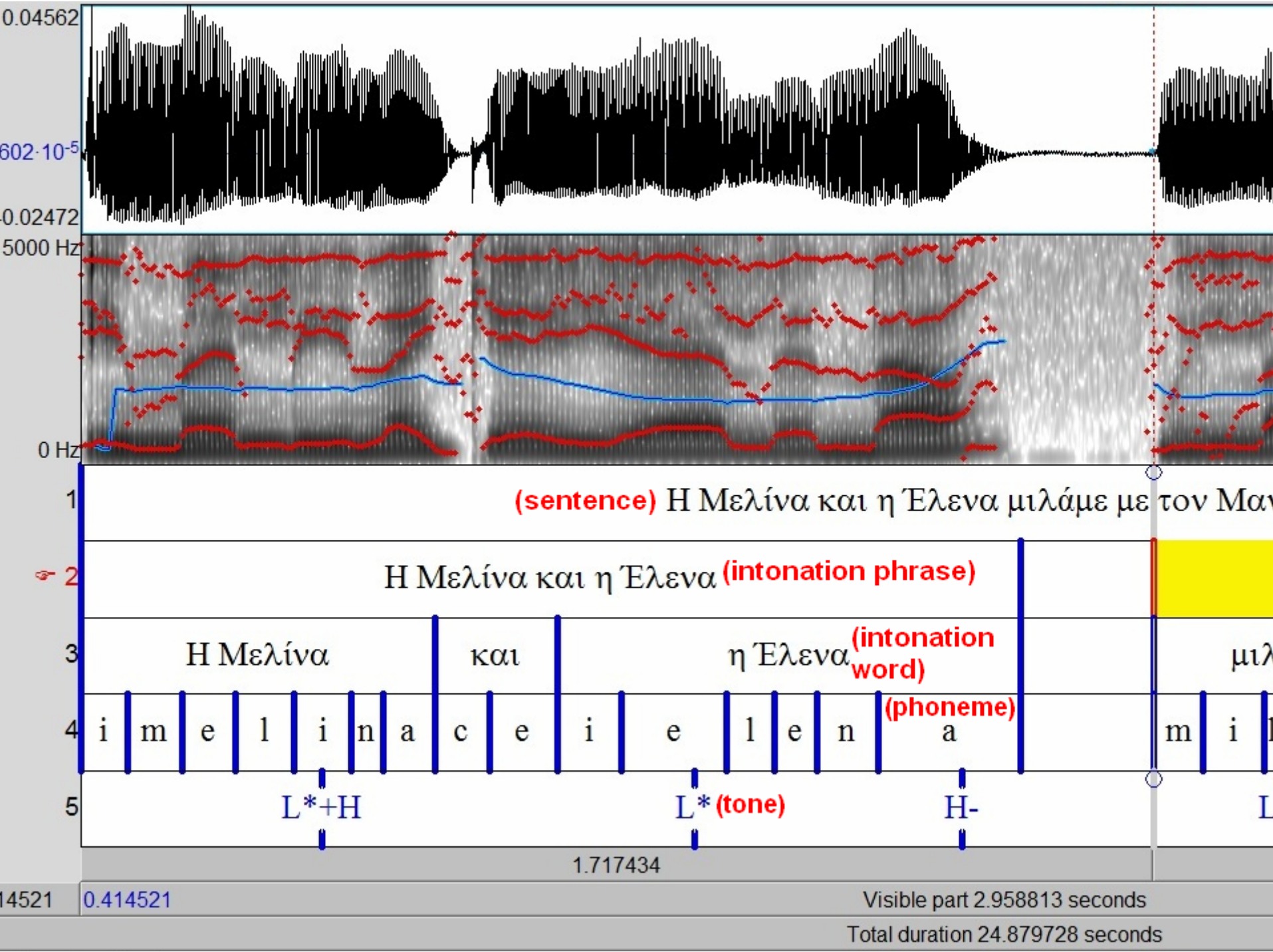
Digital Audio files (WAV) and Annotations (TextGrid, ELAN files)

- 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)

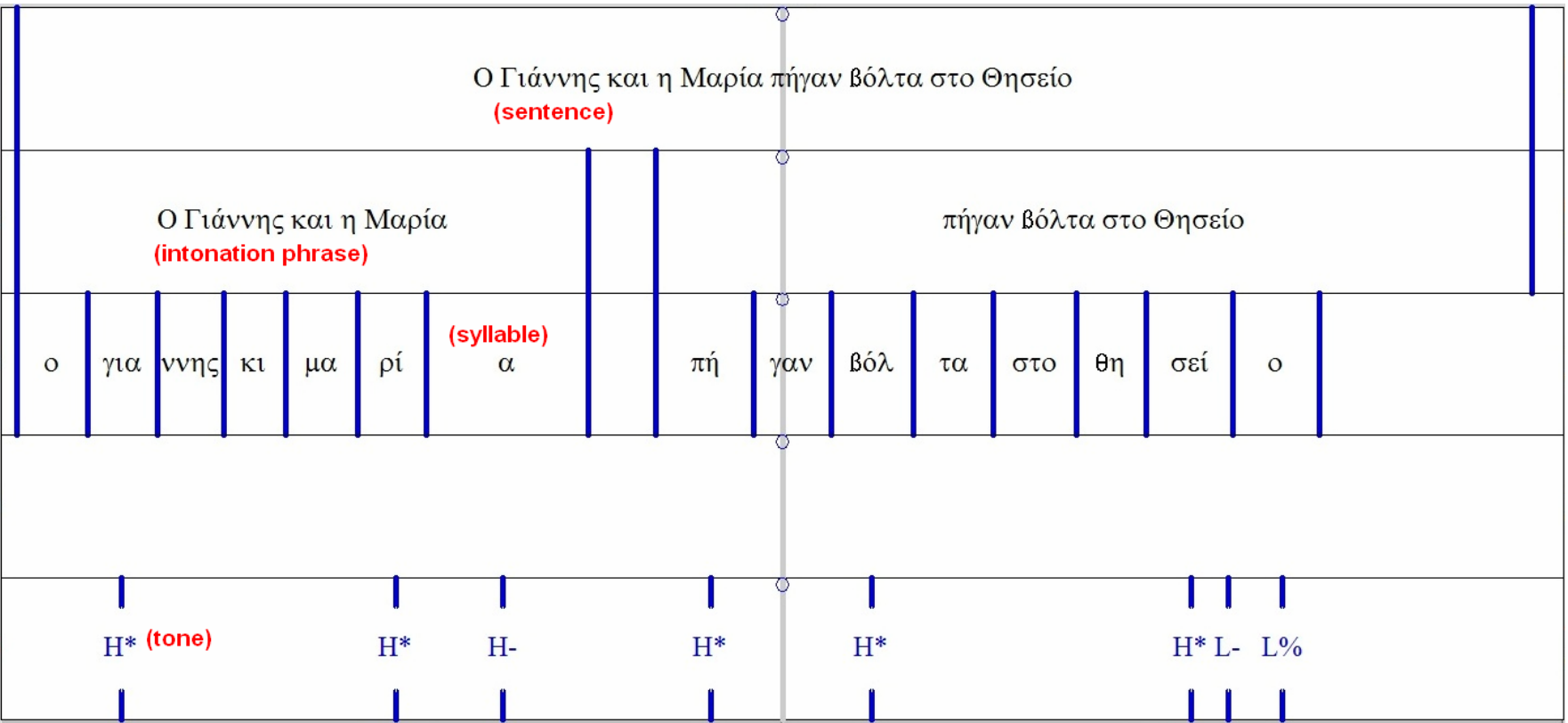


Items to annotate and some phenomena

- 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).


0.04562
602·10-5
-0.02472
5000 Hz
0 Hz
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)
μιλ
4
i m e l i n a c e i e l e n a
(phoneme)
m i
5
L*+H
L* (tone)
H-
L
1.717434
14521
0.414521
Visible part 2.958813 seconds
Total duration 24.879728 seconds

Annotations in various levels with PRAAT



Textgrid file

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 = 0.035852051109616667
            text = ""
        intervals [2]:
            xmin = 0.035852051109616667
            xmax = 3.4805159271509547
            text = "Ο Γιάννης και η Μαρία πήγαν βόλτα στο θησείο"
```

The Sentence tier of Praat (textgrid) file

Textgrid file

test.TextGrid - Σημειωματάριο

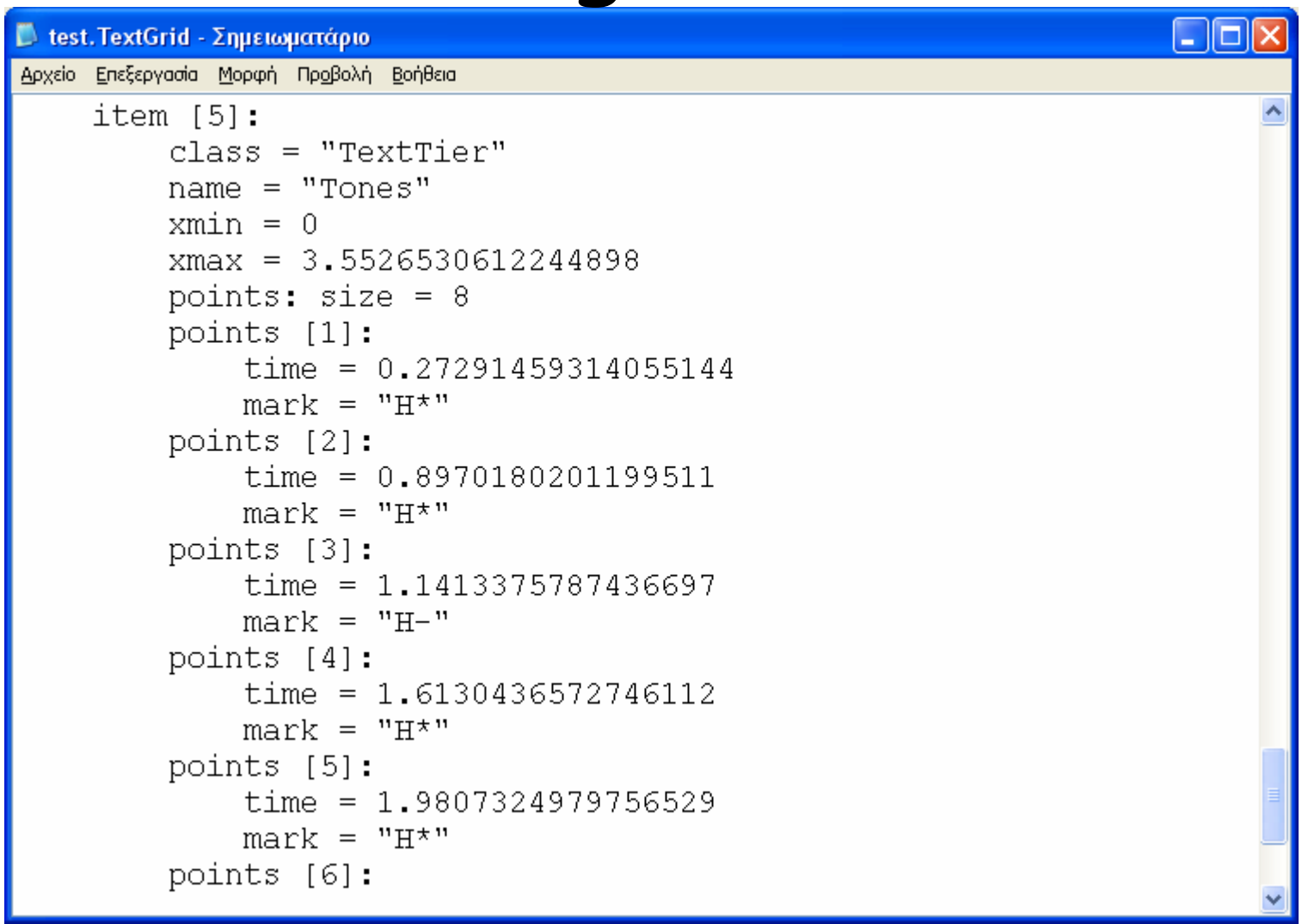
Αρχείο Επεξεργασία Μορφή Προβολή Βοήθεια

```
    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

Nikitas N. Karanikolas - Dialectal Corpora Building – June 2017

Textgrid file



The Tone point tier of Praat (textgrid) file

Written sources (digitized)
Pontic (Ποντιακά)

Ἔτον ἕνας πολλὰ πλούσιος καὶ εἶχεν ἕναν παιδὶν καὶ τὸ παιδίν ἀτ' ἐπέγ'νεν 'ς σὸ σχολεῖον. Τ' ἄλλ' τὰ παιδία, π' ἐκράτ'ναν μετ' ἐκεῖνον, εἶχαν βαγγέλῳν κ' ἐκεῖνος 'κ' εἶχεν. Εἶπεν ἕναν ἡμέραν τὴ μάνναν ἀχτε «μάννα, τὰ παιδία, ὅλα ποὺ κρατοῦν μετ' ἐμέν, ἔχουν βαγγέλᾳ κ' ἐγὼ 'κ' ἔχω, κὰ 'κὶ λές τὸν κύρη μ' καὶ παίρ' κ' ἐμὲν ἕναν βαγγέλῳν;». Ἡ μάννα 'τ' πα εἶπεν ἀτο τὸν κύρ'ν ἀτ' κ' ἐπῆρεν κ' ἐδέκεν ἀ κ' ἐδέστεν. 'Ας σὸ ἐδέστεν κ' ὕστερον, ἔγκεν ἀ 'ς ἕναν κουϊμτᾶῆν κ' ἐντῶκεν ἀπάν' ἐκὰν πεντακόᾶα φιριλία κ' ἐδέκεν ἀτο τὸ γιόν ἀτ'. Κι ἀτὸς πάει κ' ἔρται 'ς σὸ σχολεῖον. Ἔρθεν ἕναν ἡμέραν ἕνας καλόγερος 'ς σοῦ πλούσιονος καὶ ἐρώτεσαν ἀτον «ἀπόθεν ἔρχεσαι καὶ ποῦ πάς;» Εἶπεν ἀτ'ς κ' ἐκεῖνος «ἀς σ' Ἅγιον Ὄρος ἔρχουμαι καὶ 'ς σὸν Ἅιν 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 παιδία, π' εκράτναν μετ' εκείνον, είχαν βαγγέλΟν κι εκείνος 'κ' είΣεν. είπεν έναν ημέραν την μάναν αχτε «μάνα, τα παιδία, 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 φιριλία 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 κὰι ἡ ποθίκα[1].

(Ὄφις)

Ἀτμασᾶς[2] κὰι ἡ ποθίκα[3] ἐποίκανε καρτασλούκ[4]. Ἀτματσᾶς ἐποῖκε πουλία ἐπάν σσὴν πέτρα[5] κὰι ἡ ποθίκα ἐποῖκε ἐπουκα σσὴν πέτρα. Ὑστὲρ ἀτματσᾶς ἐκατέβε κὰι ἔφαε τσῆ ποθίκας τὰ πουλία. Ἡ ποθίκα οὐκ ἐπόρενε ν' ἀνιβαίν ἐπάν σσὴν πέτρα νὰ τρώει τ' ἀτματσᾶ τὰ πουλία κὰι ἐπορπάτενε ἐπουκαικέσ[1] κὰι ἔκλαιε. Ἤρθε βακοίτ[2] κὰι ἀτματσᾶς ἐποῖε νὰ βρίσκ φαὶ γιὰ τὰ πουλία 'τ κ' ἐδεάβε σ' ἕνα κάμπο μερέα. Ἐκεῖ σσὸν κάμπο ἐκάθουσανε ἀργάτ καὶ ἔψηνανε κρέας ἀπανκέσ σσ' ἄψιμο. Ἀτματσᾶς παλ ἐκατέβε ἐπῆρε ἕναν παρτσὰ[3] κρέας ἀσσ' ἄψιμο ἀπάν κὰι ἔφερεν ἀ[4] σσὴ φωλέα τ. Μέκερ[5] ἐκρατενε ἐκεῖ ἄψιμο. Ἐκολλίε[6] ἡ φωλέαν ἀτ, ἐρώξανε[7] ἐπουκά τὰ πουλία 'τ κὰι ἔφαεν ἀτα ἡ ποθίκα.

ατμαΣας τΣε η ποθίκα εποίκανε καρταΣλουκ. ατματΣας εποίτΣε πουλία επάν σην πέτρα τΣε η ποθίκα εποίτΣε επουκά σην πέτρα. υστέρ ατματΣάς εκατέβε τΣε έφαε τση ποθίκας τα πουλία. η ποθίκα ουτΣ επόρενε ν' ανιβαίν επάν σην πέτρα να τρώει τ' ατματΣά τα πουλία τΣε επορπάτενε επουκατΣέσ τΣε έκλαιε. έρθε βακΙτ τΣε ατματΣάς 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. ατματΣάς παλ εκατέβε επήρε έναν παρτΣά κρέας ας άψιμο. εκολλίε η φωλέαν ατ', ερώξανε επουκά τα πουλία 'τ τΣε έφαεν ατα η ποθίκα.

Cappadocian (Καππαδοκικά)

Τότε πιάσαν boυdαχčήρε να κόψουν το qαβάχ. Κόφτουν το qαβάχ. Δέν πλεροῦται· πλεμνίšκει λιγόčικο. Το παλτά σακοῦται. Τότε πιάνουν ένὰ ǰadə́ qαρά· ědωκάν do ένα πολά σταφίρες νά τα πλύν. Τα καλά ἐπέτανέν da, και τα κötία βαήνεν da. Το κορῖς λέχ το, "Čί ζάεις; τα καλά πετάνεις τα, και τα κötία στέγνουν." "Čί να ποίκω; Dέ χιωρῶ." Σόŋgρα πιάνουν ένα baš̆κά ǰadə́ qαρά, και dίνουν do, να ζυμώš̆ ζυμάρ. Žύμωνέν do μέ το πράϊ τ. Čί ζάεις;" λέχ το κορῖς. "Μέ το πράχ ζυμοῦται ζυμάρ μί;" λέχ. Τότε το κορῖς κατέβη και ζύμωσέν do. Σόŋgρα νανέβη. Δέν do βάκε· πιάσεν da ăs τα μαλιά τ. Τότε ἤρτε πατιšαχιοῦ το παιρί· πήρεν do. Και σεράνda μέρες ἔπκαν γάμος.

Cappadocian transcripion

τότε πιάσαν μπουdαχτΣήρε να κόψουν το qαβάχ. Κόφτουν το qαβάχ. dεν πλερούται. πλεμνίΣκει λιγότΣικο. το παλτά σακούται. τότε πιάνουν ένα dZadé qαρά. 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; dε χιωρώ». σόνgρα πιάνουν ένα baΣκά dZadé qαρά, και dίνουν 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. σόνgρα νανέβη. dεν το βάκε. πιάσεν da ας τα μαλλιά τ. τότε ήρτε πατιΣαχιού το παιρί. πήρεν do. και σεράνda μέρες έπκαν γάμος.

Aivaliot (Αϊβαλιώτικα)

Μν'ὰ φουρὰ ἤδαν ἕνας βασλὲς τσ' εἶχι ς τοὺ τσιφάλ' ἕνα τσιρατέλ' τσὶ τοὺ εἶχι πουλὺ ἀκρυφά. Ὅποιουν βιρβέρ ἔπιρνι νὰ τοὺ γουρέψ, τοὺν ἔκανι τιβίχ[1]) νὰ μὴ τοὺ λέ ὄξου. Τώρα οὔλ' οἱ βιρβέρδις δὲν ἰβουροῦσαν νὰ τοὺ βαστάξιν ἀκρυφά· ἴ ἀφτὸ τς ἔσφαξι.

Πίσου πίσου πῆρι ἕνα βιρβέρ, τσὶ σὰ δοὺ ἀπουκούριψι, τ εἶπι, νὰ μὴ τοὺ πῆ σὶ κανέναν, ποῦς ἔχ τσέρατου, γατί θὰ πάρ τοὺ τσιφάλ' τ. Ἰ βιρβέρς δὲν ἰβόρσι νὰ βαστάξ, πῆγι, ἔστσυψι μὲς ἕνα πγάδ τσὶ φώναξι μ' οὔλ' τ γαρδιά τ: „Ἰ βασλὲς ἔχ τσιρατέλ'." Τώρα τοὺ πγάδ ξιράθτσι, φύτρουσι μέσα μν'ά καλαμν'ά. Μιγάλ'νι ἡ καλαμν'ά. Πέρνα μν'ά μέρα ἕνας dξουβάν'ς, ἔκουψι d γαλαμν'ὰ τσ' ἔκανι μν'ά τσαβούνα τσὶ τν ἔτιζι. Ἡ τσαβούνα ἤλιγι: „Βί! ἰ βασλὲς ἔχ τσιρατέλ'" Τοὺ ἤκσαν, τοὺ εἶπαν τ βασλέ.

Aivaliot transcription

μια φουρά ήδαν ένας βασλές τσ' είχι στου τσιφάΛ ένα τσιρατέΛ τσι του είχι πουλύ ακρυφά. όποιουν bιρbέρ έπιρνι να του gουρέψ, τουν έκανι τιbίχ να μη του λέΓ' ὄξου. τώρα ούΛ οι bιρbέρδις δεν ιμπουρούσαν να του βαστάξιν ακρυφά. γι' αυτό τς έσφαξι. πίσου πίσου πήρι ένα bιρbέρ, τσι σα dου απουκούριψι, τ' είπι να μη του πη σι κανέναν, πους έΧ' τσέρατου, γιατί θα πάρ' του τσιφάΛ τ. ι bιρbέρς δεν ιμπόρσι να βαστάξ, πήγι, έστσυψι μες ένα πγάδ τσι φώναξι μ'ούΛ τ' gαρδιά τ': «ι βασλές έΧ' τσιρατέΛ». τώρα του πγάδ ξιράθτσι, φύτρουσι μέσα μια καλαμιά. μιγάΛνι η καλαμνιά. πέρνα μια μέρα ένας τZoubάNς, έκουψι d' gαλαμιά τσ' ΄έκανι μια τσαbούνα τσι τ'ν έπιζι. Η τσαbούνα ήλιγι: «Bί! ι βασλές έΧ' τσιρατέΛ». Του ήκσαν, του είπαν τ' βασλέ.

Morphological annotations

Morphological categories

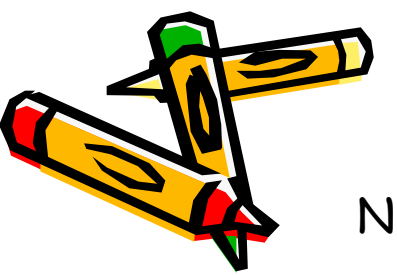
- Word
  - Grammatical category
    - (noun, verb, gerund, particle, adjective, pronominal, adverb, …)
  - Special characteristics
    - Loan word
      - Origin
    - Archaism
    - Gender alteration
    - Other
  - Simple
  - Structured
    - Declinable
    - Production
    - Composition
    - Merging
    - 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

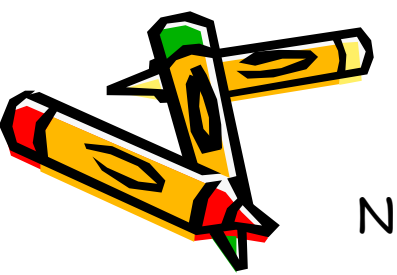
Syntactic and Semantic annotations

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



AMiGre

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



Why our own system

- Initially, advanced software tools such as LaBB-CAT seemed as the perfect solution to the problem of storing and processing of oral and written dialectal resources.
- But, they proved to be unable to meet all the requirements of the AMiGre project:
  - Annotations at many different linguistic levels
  - Combined search at different levels of representation (phonology, morphology, metadata and, potentially, syntax and semantics)
  - Combined search on both oral and written resources
- So we designed and implemented our own system in order to meet all requirements

Applications Overview

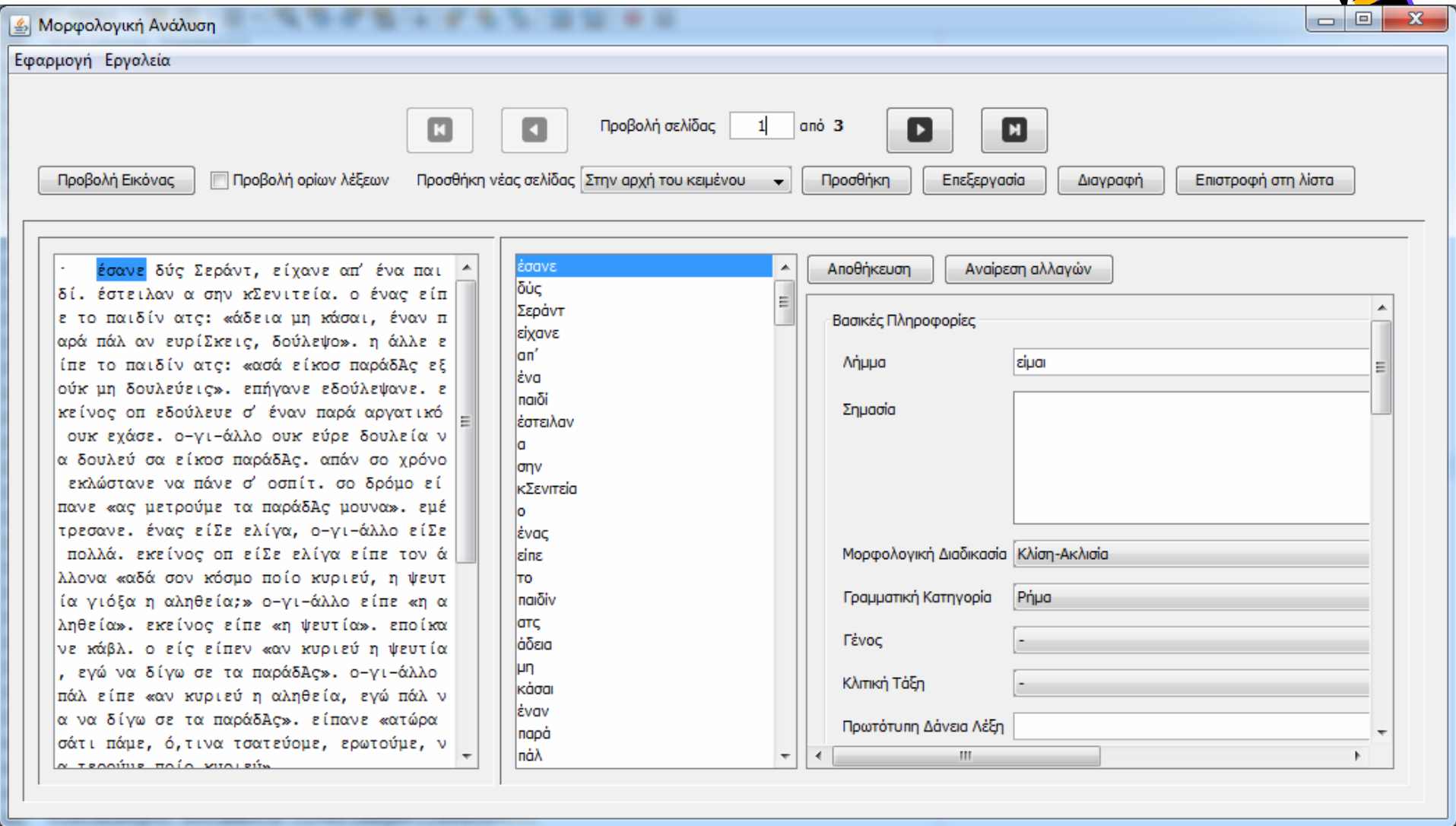
- Oral + Written modules
- Written Sources GUI
- Oral Sources GUI
- Oral & Written Sources Retrieval GUI
- Oral & Written Browsing Web Interface
- Dissemination of effort & Results Web site

Oral + Written - modules

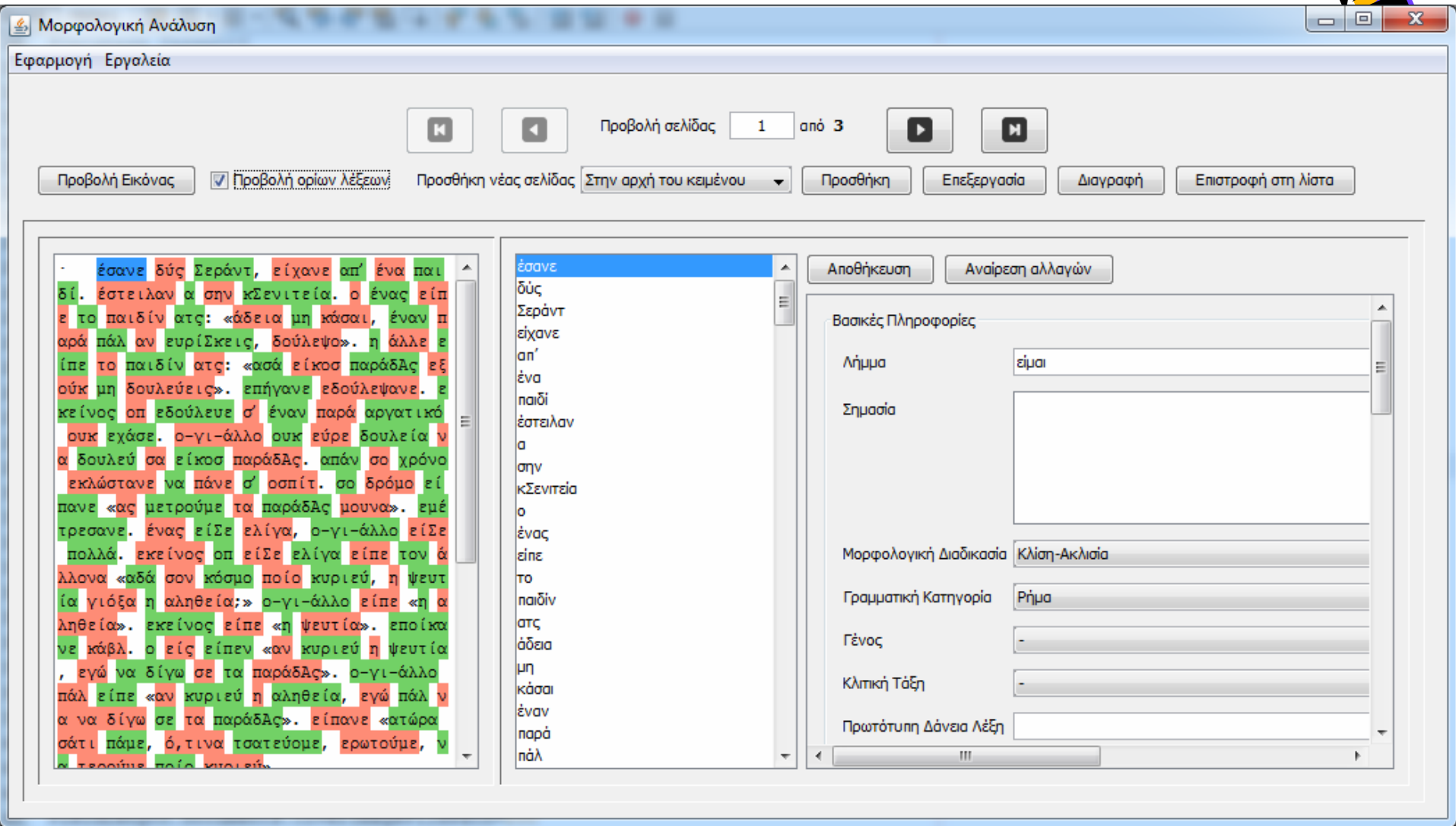
Name	Interpetation	Oral	Written
B. Oral	Browse Oral	X	
B. Written	Browse Written		X
I. Oral	Import Oral	X	
Meta Oral	Metadata for Oral	X	
G. Oral	Oral triptych	X	
G. Written	Written triptych		X
Meta Written	Metadata for Written		X
P.I. Written	Page (or Part) Import Written		X
Morph. Tag	Morphological Tagging	X	X
Syn. Tag	Syntactic Tagging	X	X
Sem. Tag	Semantic Tagging	X	X
Ph. Tagger	Phonological Tagging	X	X
T. Imaging	Text Imaging		X
T. Transcription	Text Transcription		X
M.I. Oral	Massive Import Oral	X	
M.I. Written	Massive Import Written		X

Written Sources GUI – 3fold

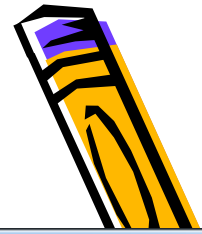
display the transcription of document together with attributes of the selected word



Display transcription with word borders



Transcription together with the original image scan



Μορφολογική Ανάλυση

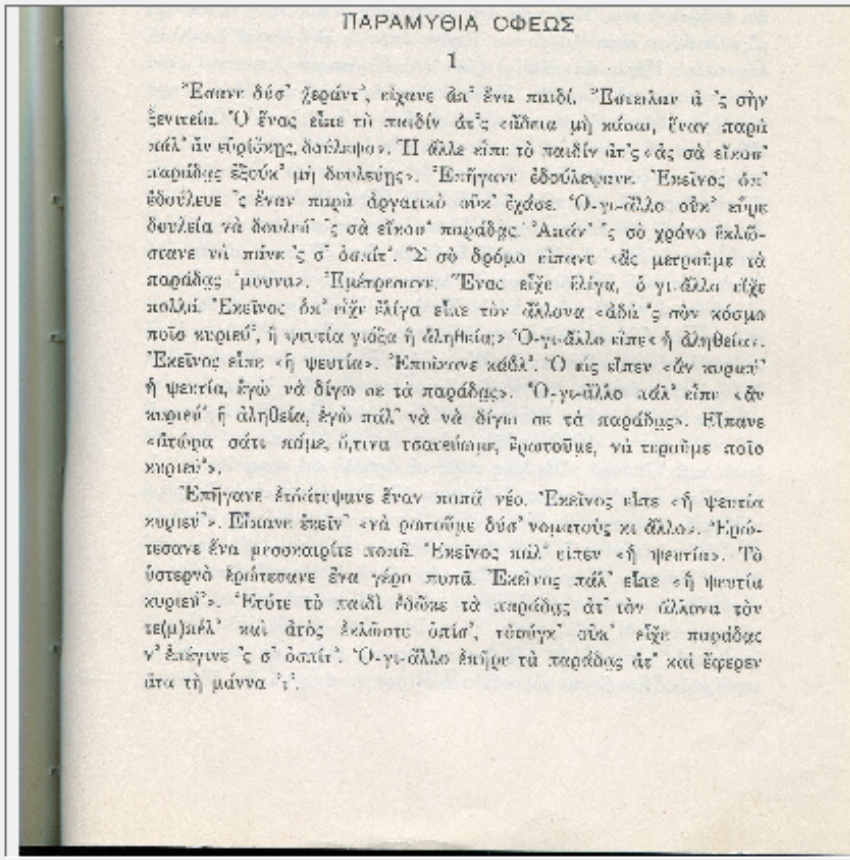
Εφαρμογή Εργαλεία

Προβολή σελίδας 1 από 3

Προβολή Λέξεων | Προβολή ορίων λέξεων | Προσθήκη νέας σελίδας: Στην αρχή του κειμένου | Προσθήκη | Επεξεργασία | Διαγραφή | Επιστροφή στη λ

έσανε δύς Σεράντ, είχανε απ' ένα παιδί. έστειλαν α σην κΣε
νιτεία. ο ένας είπε το παιδίν ατς: «άδεια μη κάσαι, έναν παρά
πάλ αν ευρίΣκεις, δούλεψο». η άλλε είπε το παιδίν ατς: «ασά
είκοσ παράδΑς εξούκ μη δουλεύεις». επήγανε εδούλεψανε. εκείνο
ς οπ εδούλευε σ' έναν παρά αργατικό ουκ εχάσε. ο-γι-άλλο ουκ ε
ύρε δουλεία να δουλεύ σα είκοσ παράδΑς. απάν σο χρόνο εκλώστα
νε να πάνε σ' οσπίτ. σο δρόμο είπανε «ας μετρούμε τα παράδΑς μ
ουνα». εμέτρεσανε. ένας είΣε ελίγα, ο-γι-άλλο είΣε πολλά. εκε
ίνος οπ είΣε ελίγα είπε τον άλλονα «αδά σον κόσμο ποίο κυριεύ
, η ψευτία γιόξα η αληθεία;» ο-γι-άλλο είπε «η αληθεία». εκεί
νος είπε «η ψευτία». εποίκανε κάβλ. ο είς είπεν «αν κυριεύ η
ψευτία, εγώ να δίγω σε τα παράδΑς». ο-γι-άλλο πάλ είπε «αν κυ
ριεύ η αληθεία, εγώ πάλ να να δίγω σε τα παράδΑς». είπανε «ατ
ώρα σάτι πάμε, ό,τινα τσατεύομε, ερωτούμε, να τερούμε ποίο κυ
ριεύ».

επήγανε ετΣάτεψανε έναν ποπά νέο. εκείνος είπε «η ψευτία κ
υριεύ». είπανε εκείν «να ρωτούμε δύσ νοματούς κι άλλο». ερώτε
σανε ένα μεσοκαιρίτε ποπά. εκείνος πάλ είπεν «η ψευτία». το υ
στερνό ερώτεσανε ένα γέρο ποπά. εκείνος πάλ είπε «η ψευτία κυ
ριεύ». ετότε το παιδί εδώκε τα παράδΑς ατ τον άλλονα τον τε(μ
)πέλ και ατός εκλώστε οπίσ, τΣούγκ ουκ είΣε παράδΑς ν' επέγινε
σ' οσπίτ. ο-γι-άλλο επήρε τα παράδΑς ατ και έφερεν ατα τη μάν



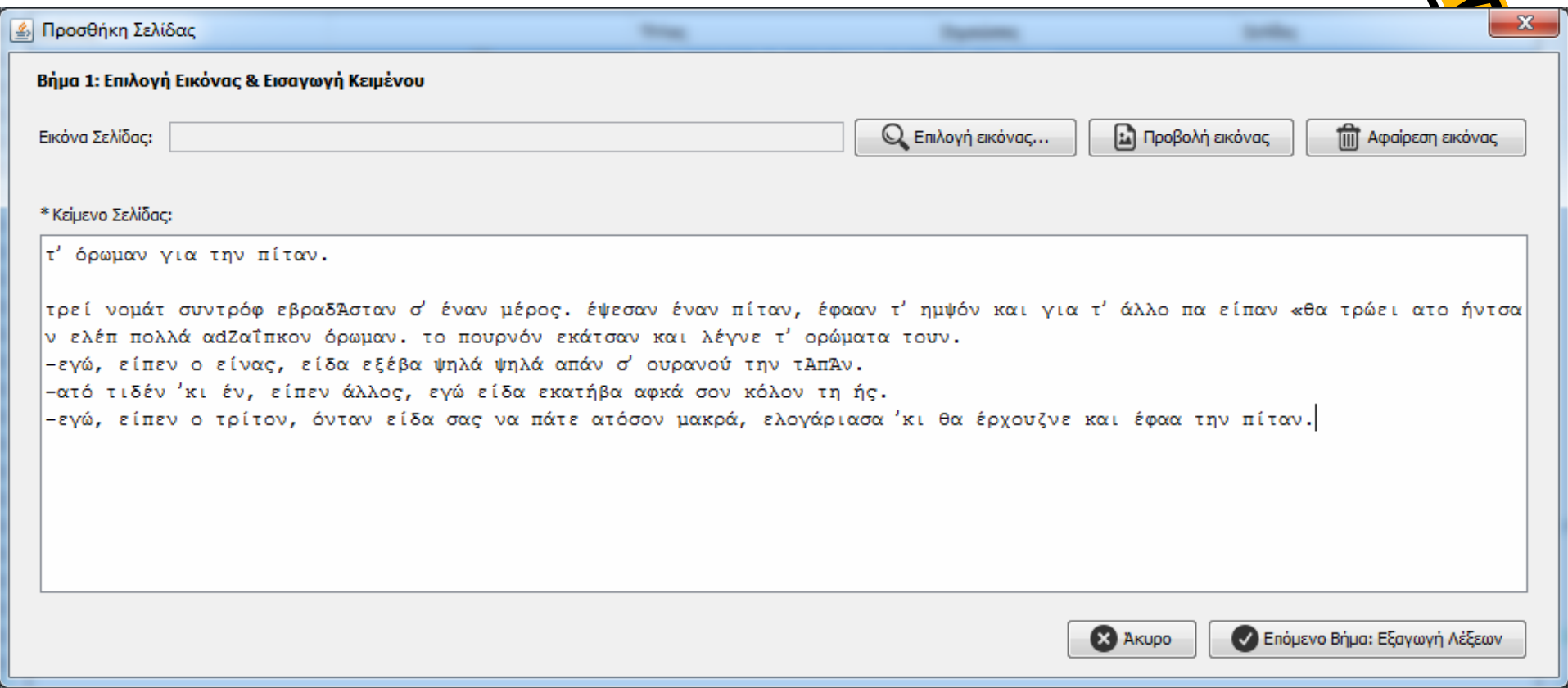
ΠΑΡΑΜΥΘΙΑ ΟΦΕΩΣ

1

Έσανε δύσ' ξεράντ', είχανε άτ' ένα παιδί. Έστειλαν ά 'ς σην ξενιτεία. Ό ένας είπε τὸ παιδίν άτ'ς «άδεια μὴ κάσαι, έναν παρὰ πάλ' άν εὑρίσκης, δούλεψο». Ἡ άλλε είπε τὸ παιδίν άτ'ς «άς σὰ εἶκοσ' παράδες ξεύτ' μὴ δουλεύῃς». Ἐπήγανε εδούλεψανε. Ἐκεῖνος ὅτ' ἐδούλευε 'ς ἕναν παρὰ ἀργατικὸ οὔκ' ἐχάσε. Ὁ-γι-ἄλλο οὐκ' εὗρε δυνλεία νὰ δουλεύ' 'ς σὰ εἶκοσ' παράδες. Ἀπάν' 'ς σὸ χράνο ἐκλώστανε νὰ πάνε 'ς σ' ὁσπίτ'. Σ' σὸ δρόμο εἶπανε «ἂς μετροῦμε τὰ παράδες 'μουνα». Ἐμέτρεσανε. Ἕνας εἶχε ἐλίγα, ὁ γι-ἄλλο εἶχε πολλά. Ἐκεῖνος ὅτ' εἶχε ἐλίγα εἶπε τὸν ἄλλονα «ἀδά 'ς σὸν κόσμο ποῖο κυριεύ', ἡ ψευτία γιόξα ἡ ἀληθεία;» Ὁ-γι-ἄλλο εἶπε «ἡ ἀληθεία». Ἐκεῖνος εἶπε «ἡ ψευτία». Ἐποίκανε κάβλ'. Ὁ κἰς εἶπεν «ἂν κυριεύ' ἡ ψευτία, ἐγὼ νὰ δίγω σε τὰ παράδες». Ὁ-γι-ἄλλο πάλ' εἶπε «ἂν κυριεύ' ἡ ἀληθεία, ἐγὼ πάλ' νὰ νὰ δίγω σε τὰ παράδες». Εἶπανε «ἀτώρα σάτι πάμε, ὅ,τινα τσατεύομε, ἐρωτοῦμε, νὰ τεροῦμε ποῖο κυριεύ'».

Ἐπήγανε ετὰτεψανε ἕναν ποπᾶ νέο. Ἐκεῖνος εἶπε «ἡ ψευτία κυριεύ'». Εἶπανε ἐκεῖν' «νὰ ρωτοῦμε δύσ' νοματούς κι ἄλλο». Ἐρώτεσανε ἕνα μεσοκαιρίτε ποπᾶ. Ἐκεῖνος πάλ' εἶπεν «ἡ ψευτία». Τὸ ὕστερνὸ ἐρώτεσανε ἕνα γέρο ποπᾶ. Ἐκεῖνος πάλ' εἶπε «ἡ ψευτία κυριεύ'». Ἐτότε τὸ παιδὶ ἔδωκε τὰ παράδες ἀτ' τὸν ἄλλονα τὸν τε(μ)πέλ' καὶ ἀτὸς ἐκλώστε ὁπίσ', τσούγκ' οὔκ' εἶχε παράδες ν' ἐπέγινε 'ς σ' ὁσπίτ'. Ὁ-γι-ἄλλο ἐπῆρε τὰ παράδες ἀτ' καὶ ἔφερεν ἀτα τὴ μάννα 'τ'.

Insertion (selection) of an image for a written source



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 εβραδΆσταν σ' έναν μέρος. έψεσαν έναν πίταν, έφαραν τ' ημψόν και για τ' άλλο πα είπαν «θα τρώει
ατο ήντσαν ελέπ πολλά adZaΐπκον όρωμαν. το πουρνόν εκάτσαν και λέγνε τ' ορώματα τουν.
-εγώ, είπεν ο είνας, είδα εξέβα ψηλά ψηλά απάν σ' ουρανού την τΑπΆν.
-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:

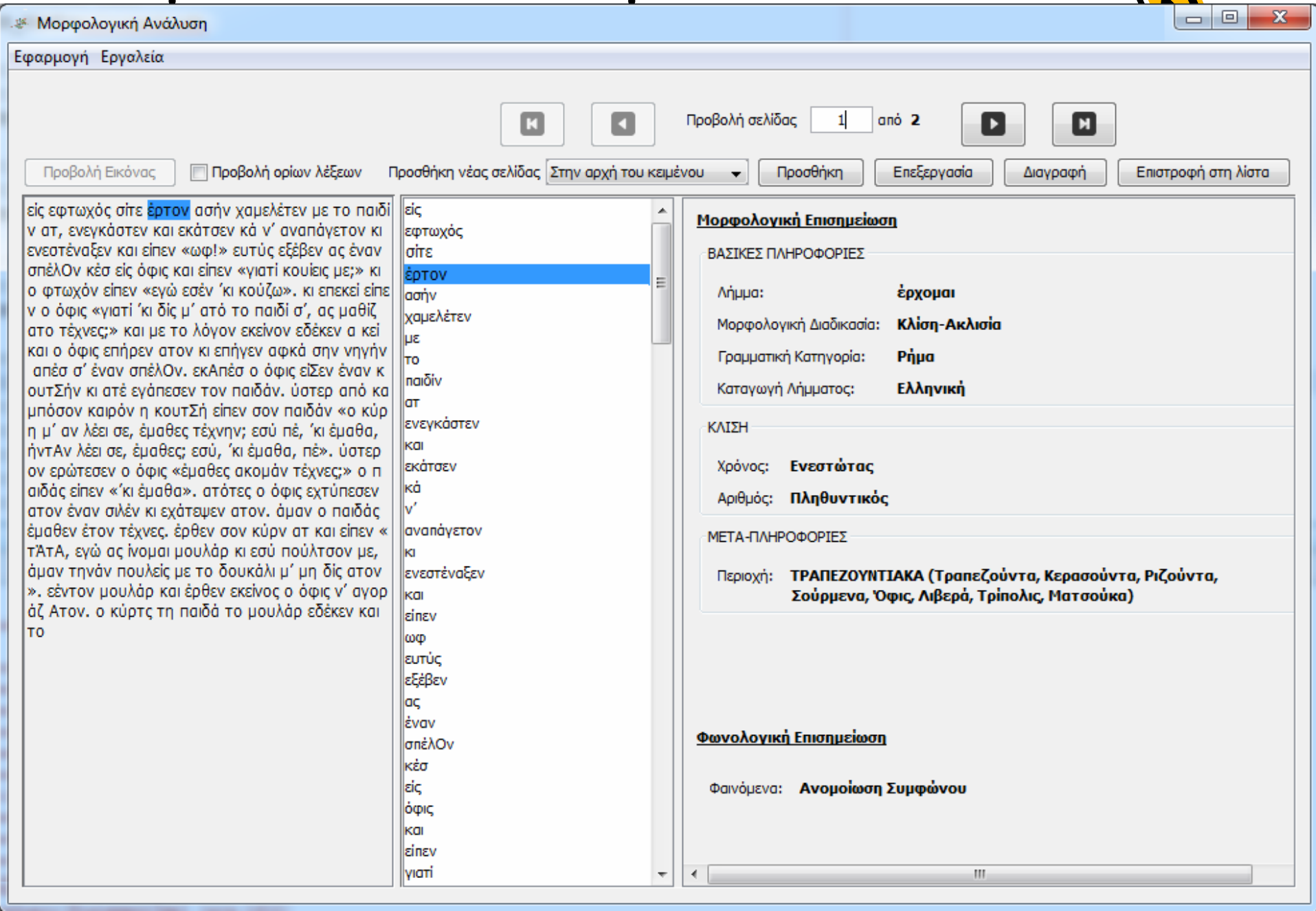
- [x] (Regular Expression: '\s')
- [x] , (Regular Expression: ',')
- [x] . (Regular Expression: '\.')
- [x] - (Regular Expression: '\-')
- [ ] ? (Regular Expression: '\?')
- [x] ; (Regular Expression: ';')
- [ ] ' (Regular Expression: '\'')
- [ ] ’ (Regular Expression: '’')
- [x] ! (Regular Expression: '!')

Regular Expression:

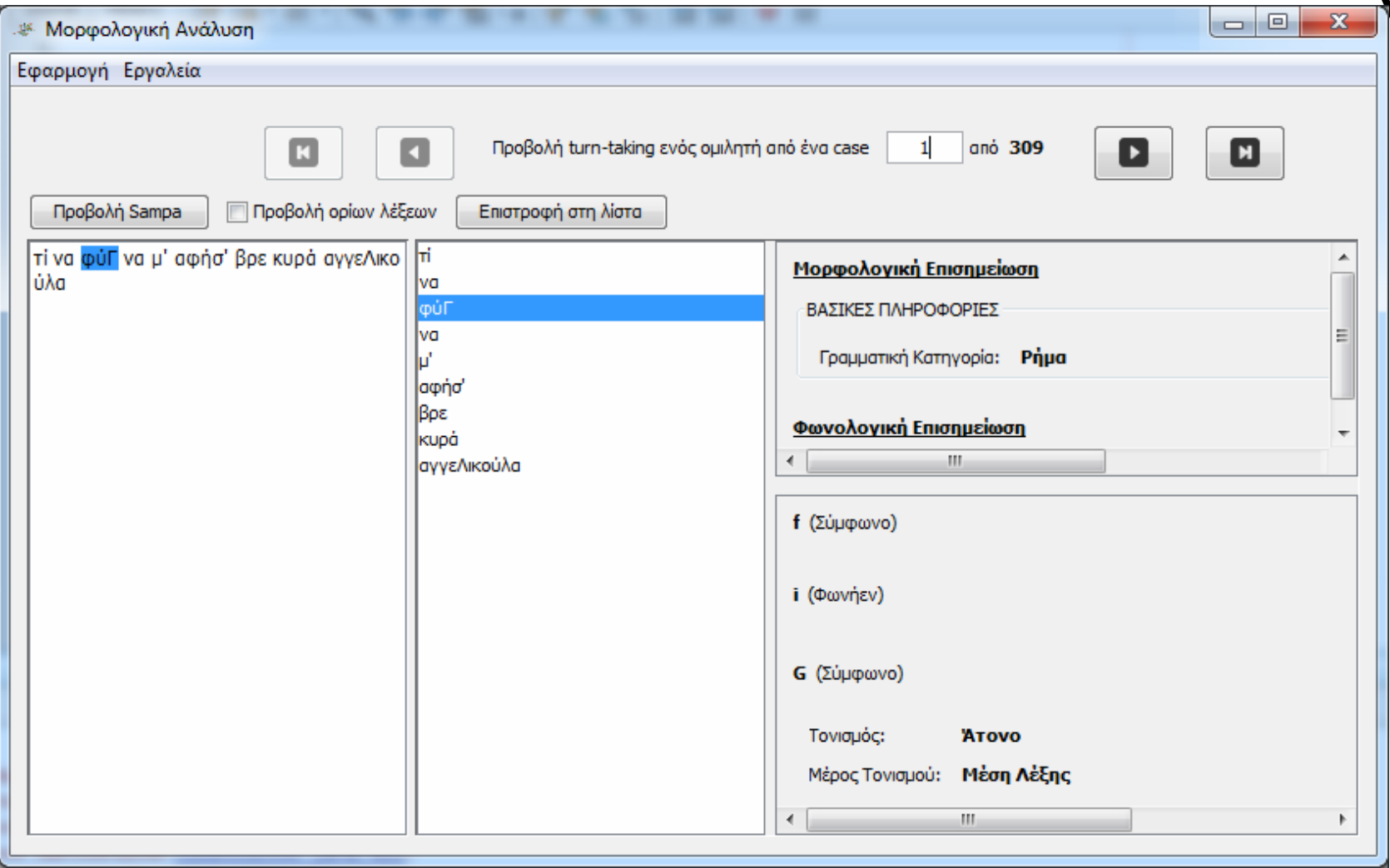
(?<!\\)[\s,\.\-;!«»:]

Άκυρο | Προηγούμενο Βήμα: Εισαγωγή Κειμένου & Εικόνας | Επόμενο: Προεπισκόπηση

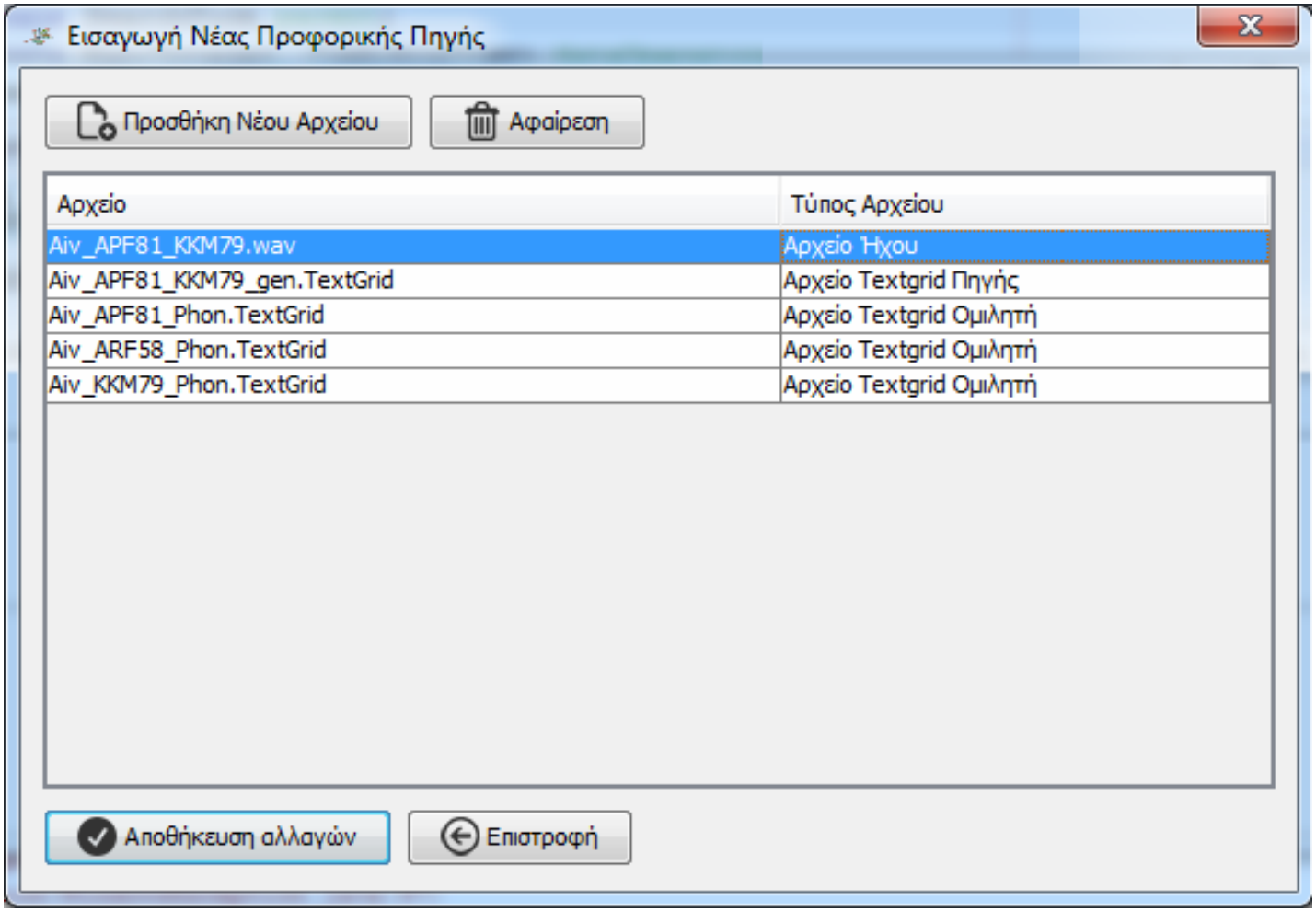
Improved 3fold presentation



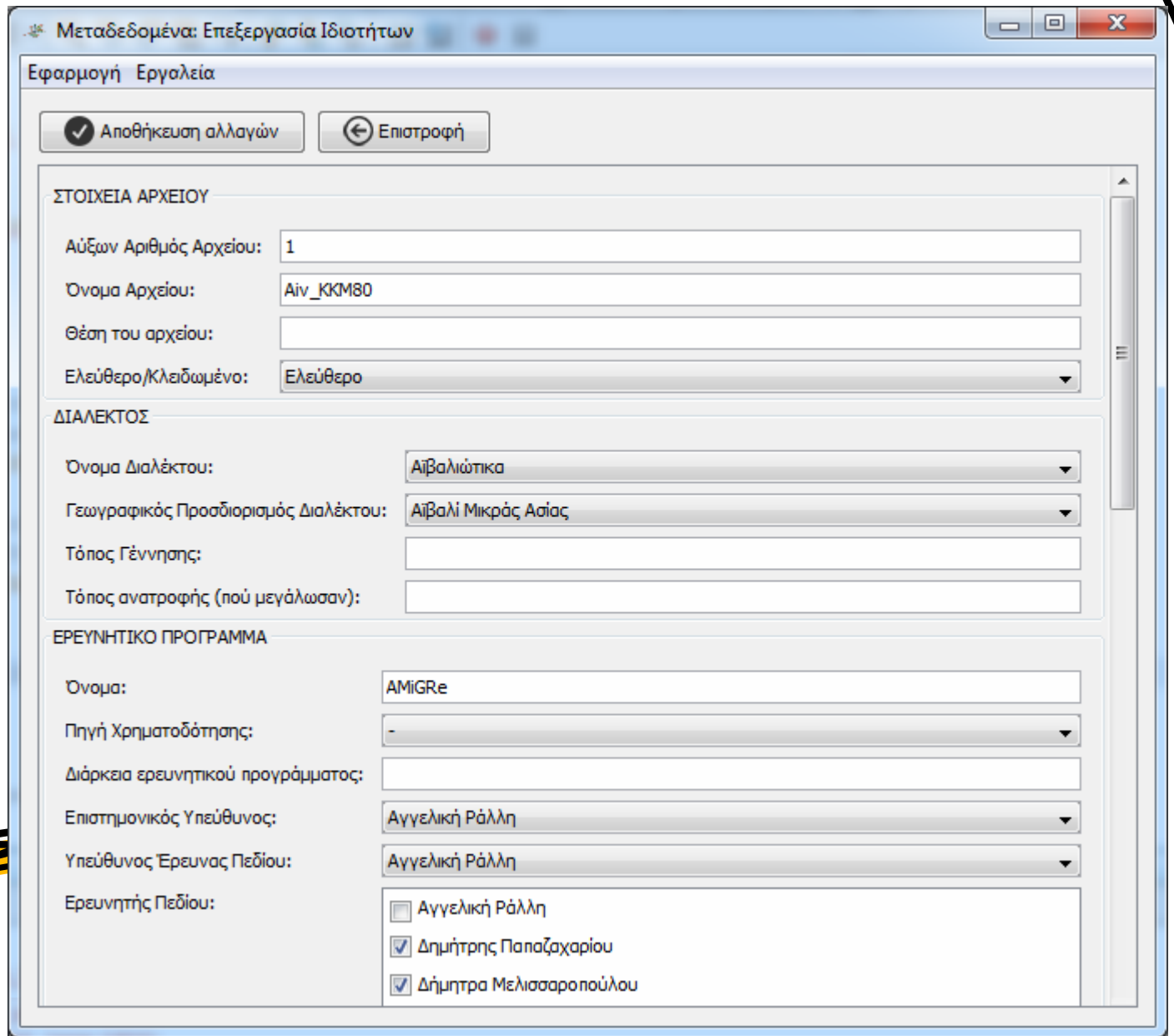
Oral Sources GUI - 3fold



Oral Sources - Import

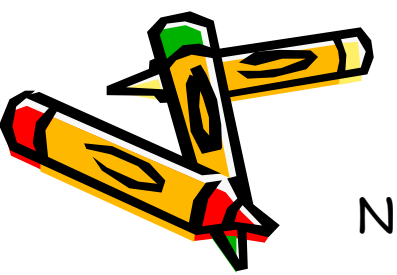


Oral Sources - Metadata



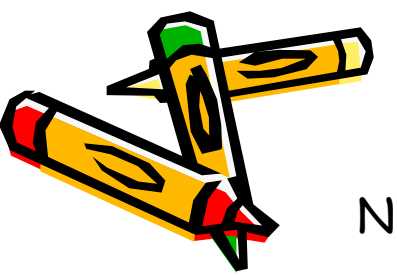
AMiGre

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

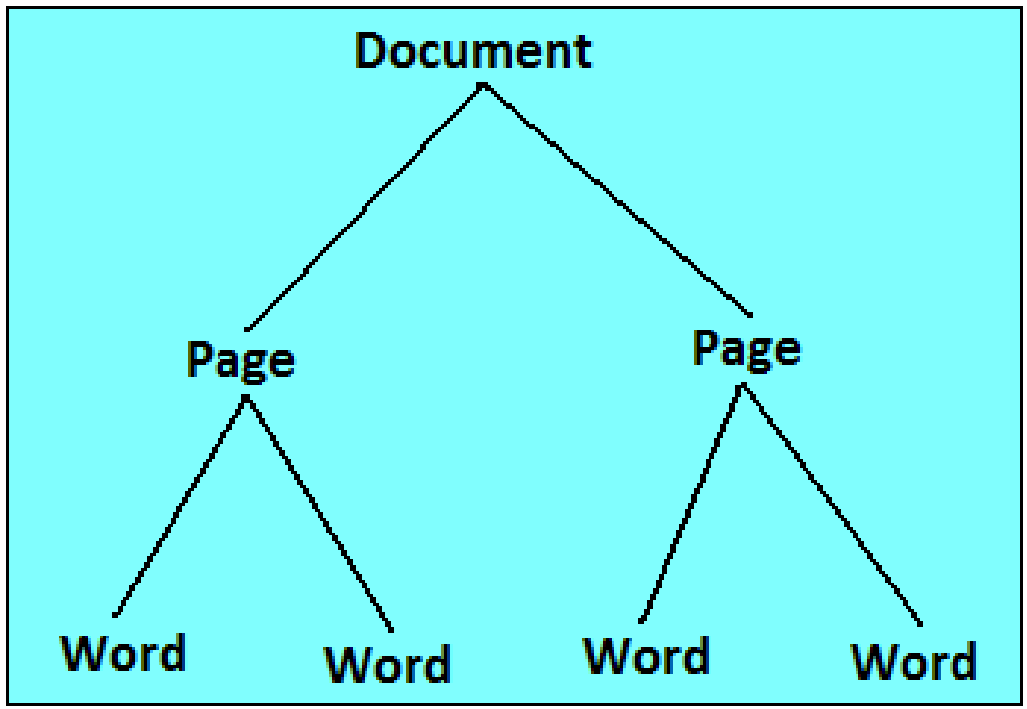


Design Overview

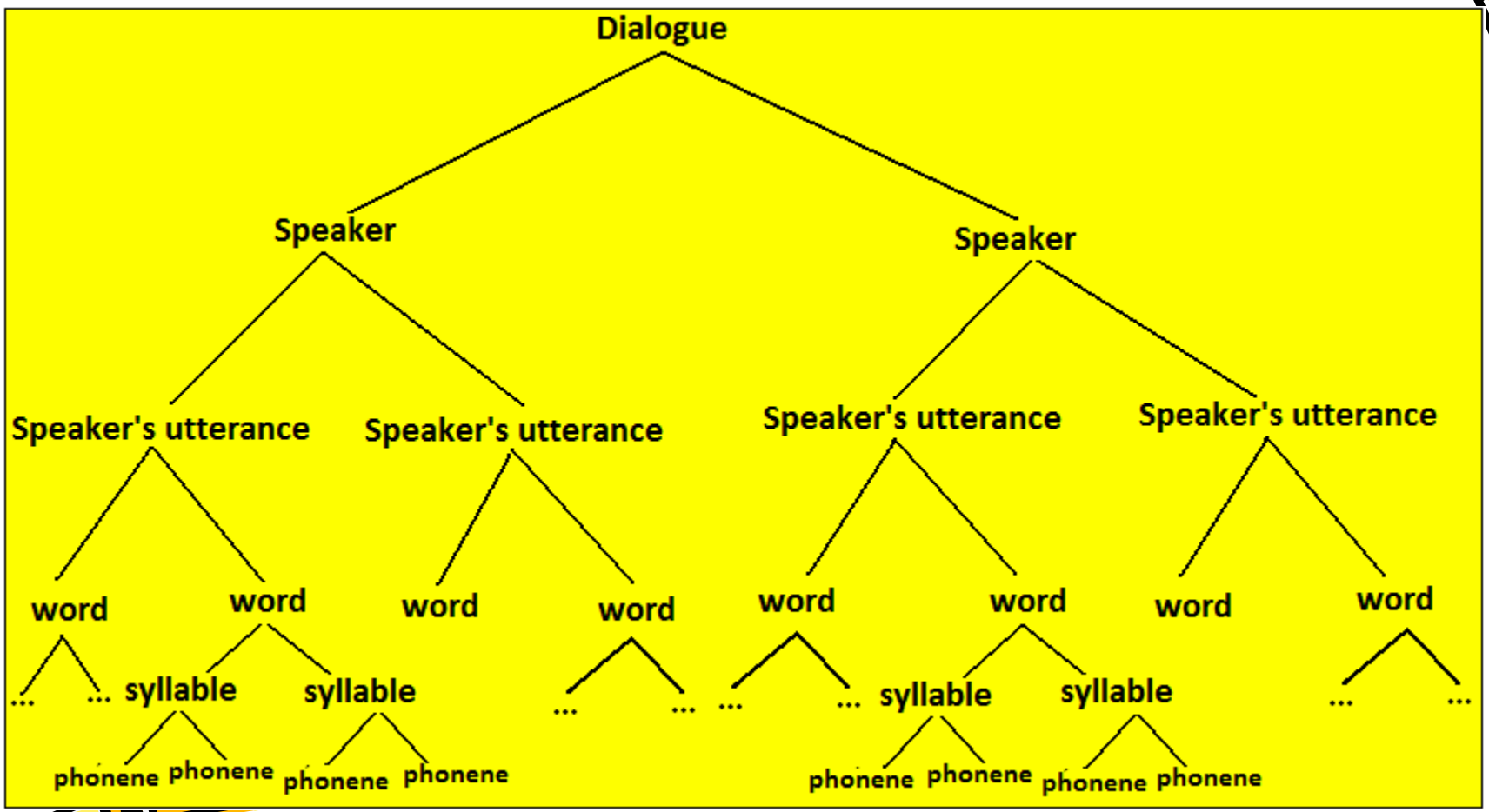
- Alignment of Oral and Written data
- Oral & Written – System overview
- Struct (relational) databases
- EAV data structures



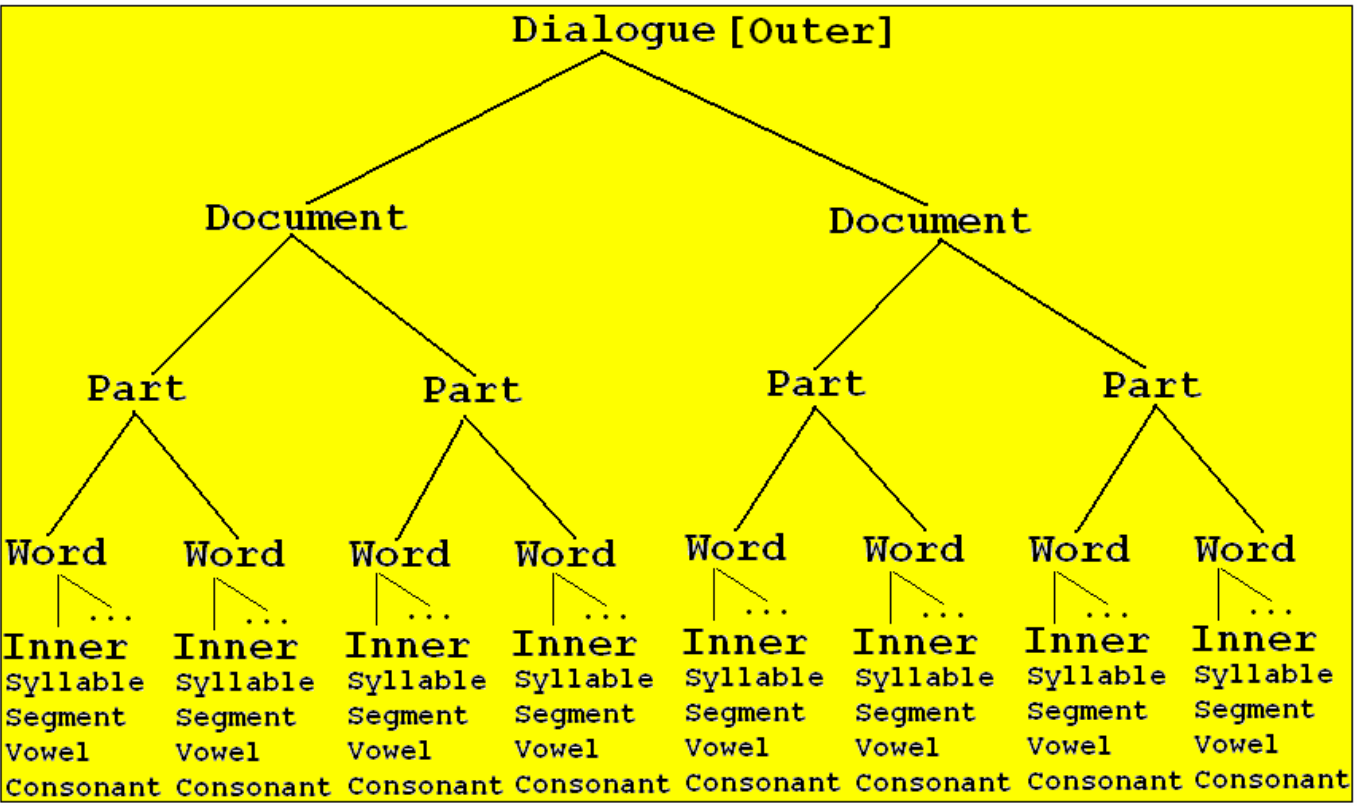
Alignment of Oral and Written – structure of Written data

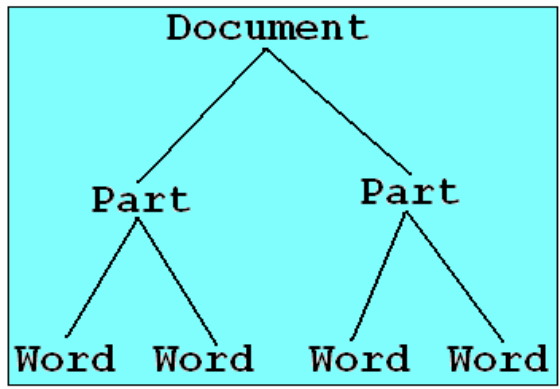


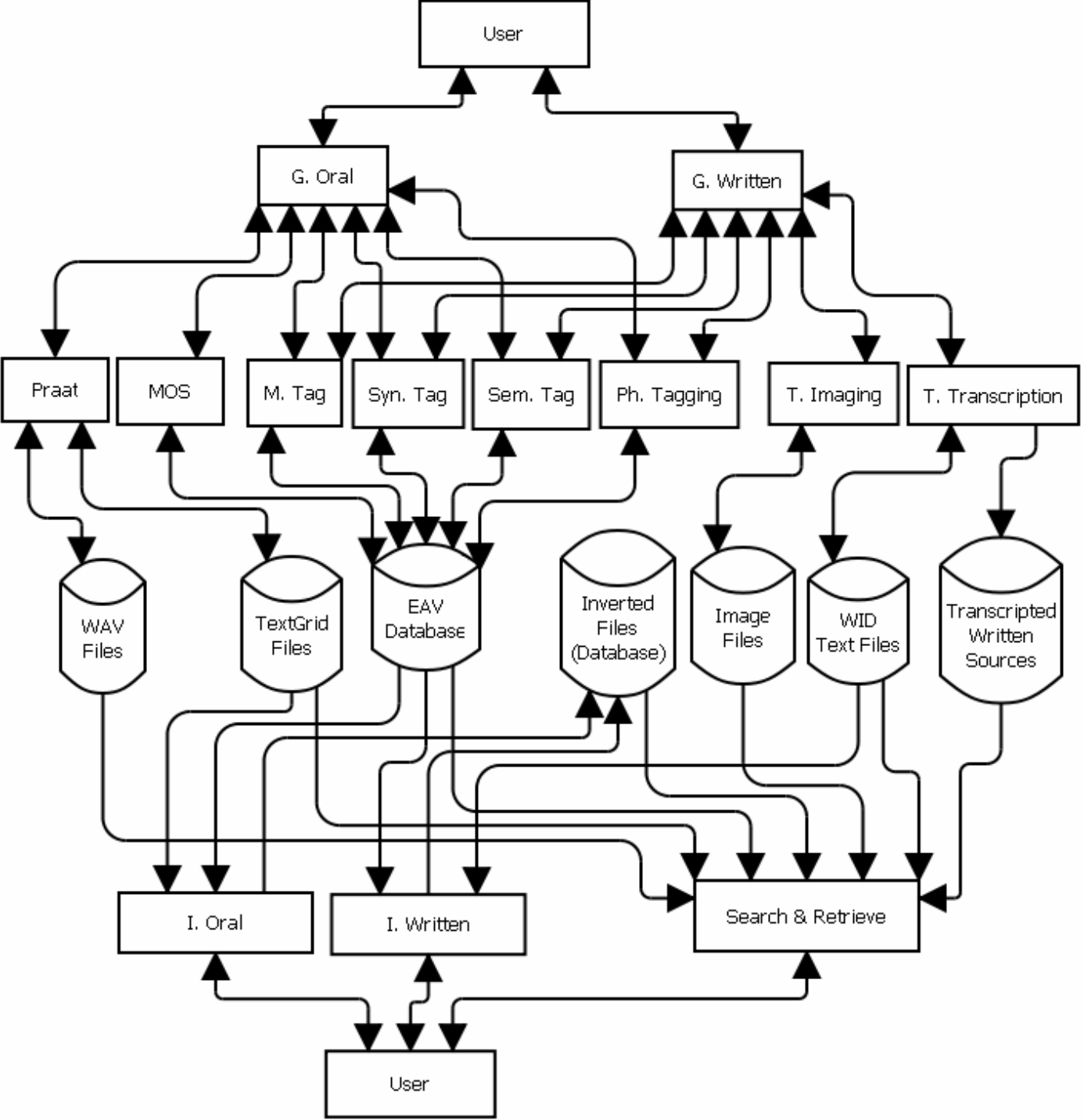
Alignment of Oral and Written – structure of Oral data



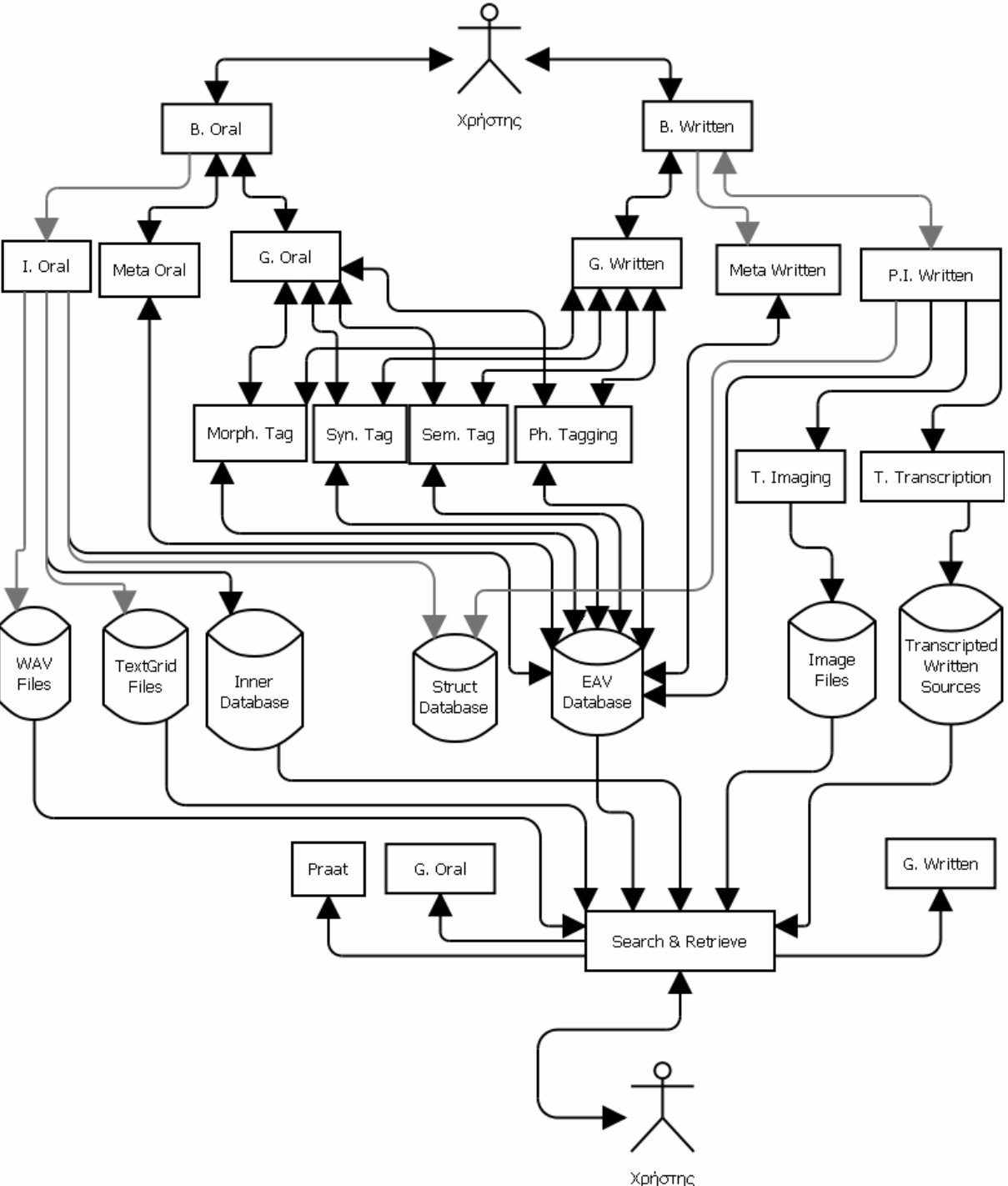
Alignment of Oral and Written – common structure







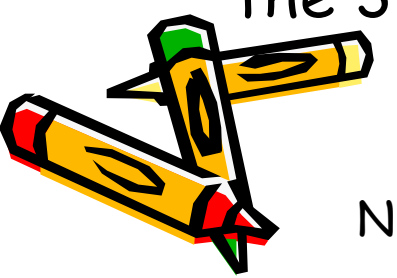
Oral & Written Sources - System overview - OLD



Oral & Written Sources - System overview - updated

Struct database

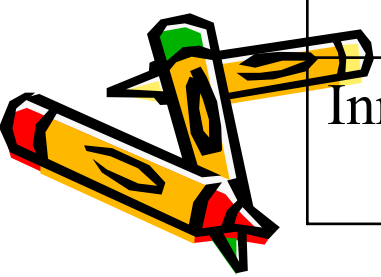
- Within the ***Struct*** database, the components of the documents are organized on consecutive levels of refinement which will be annotated with the help of the *EAV* database.
- The implementation of the abstract structure of the ***Struct*** database uses two quasi similar relational schemas.
- The 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.



Aligments 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_SegmentTypes

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

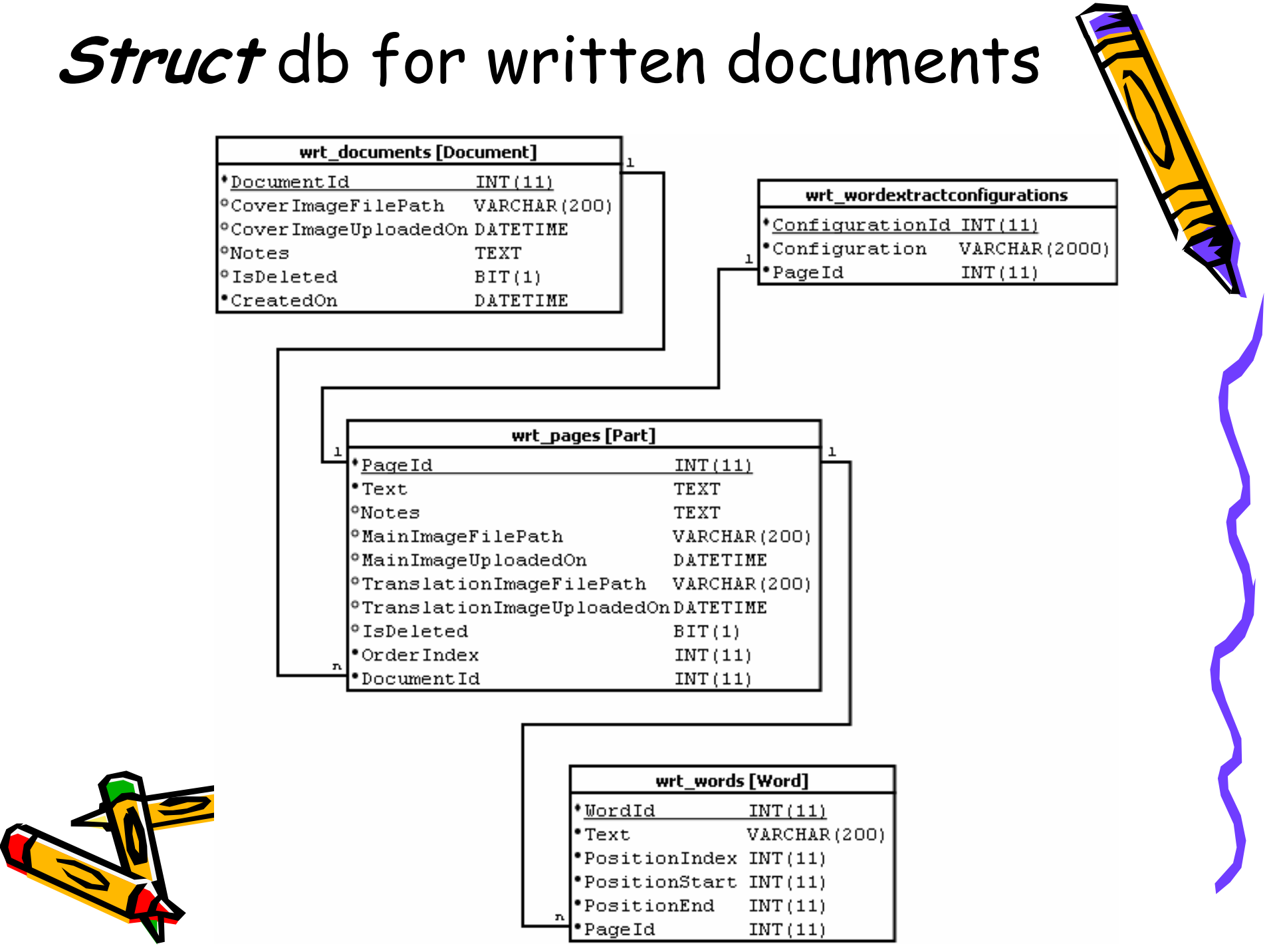
oral_words [Word]

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

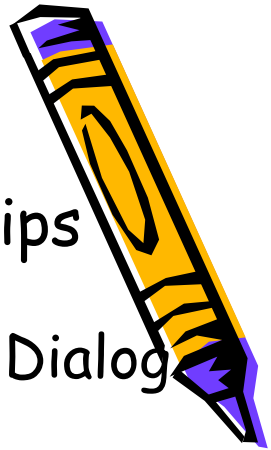
oral_segments [Inner]

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)

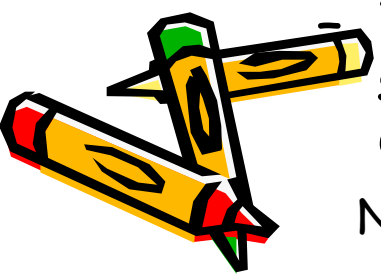
Struct db for written documents



Struct dbs some explanations



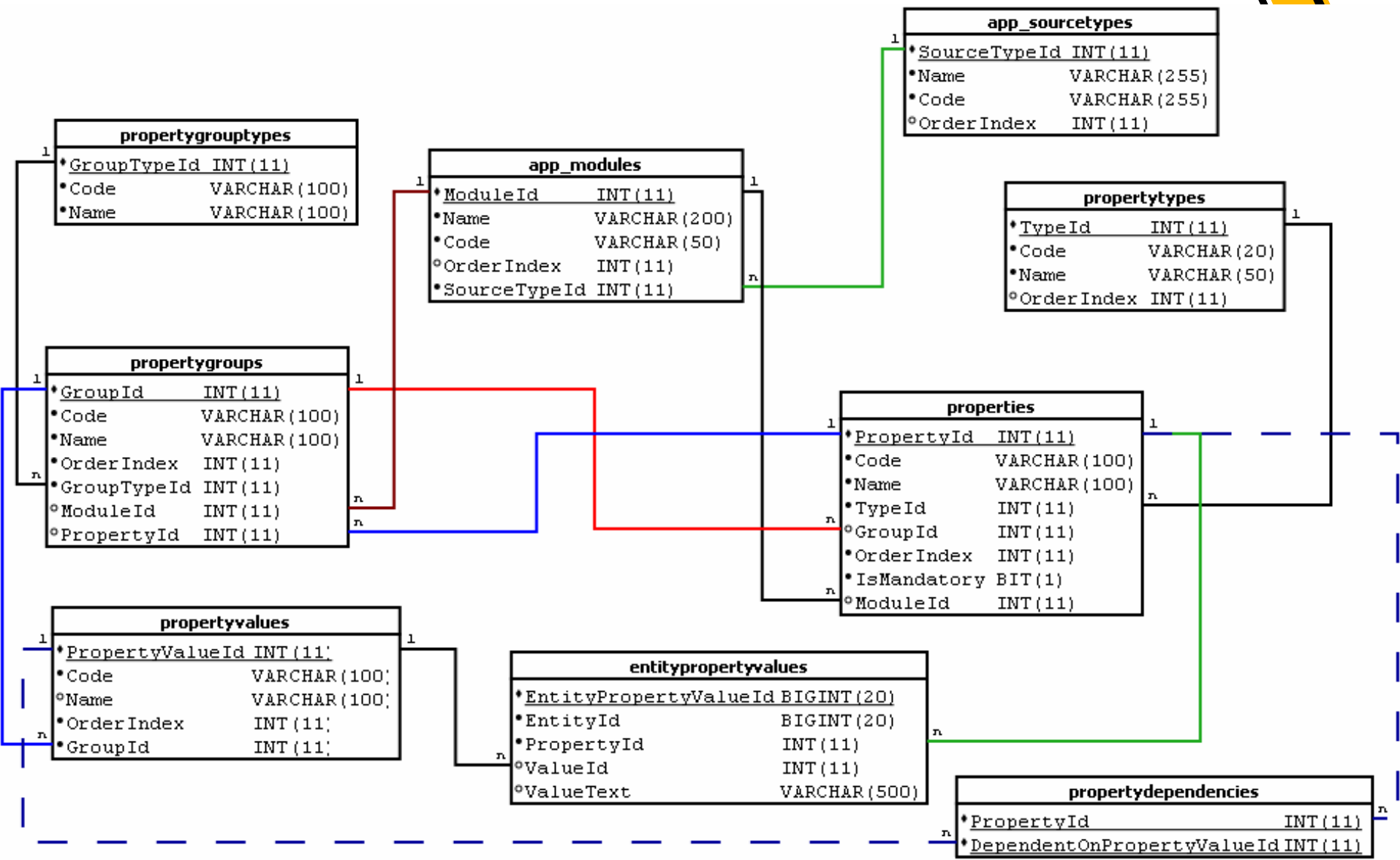
- In both db schemas, one-to-many relationships exist for each pairs of levels. Examples are:
  - many speakers/interlocutors participate in a Dialog (the association between *oral_sources* and *oral_speakers* is 1: n)
  - a written document is composed of many pages (the association between *wrt_documents* 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)



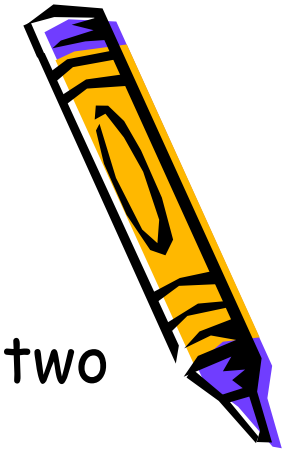
EAV database

- 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

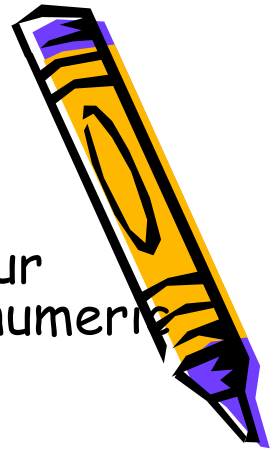


Brief description of the 9 tables consisting the EAV



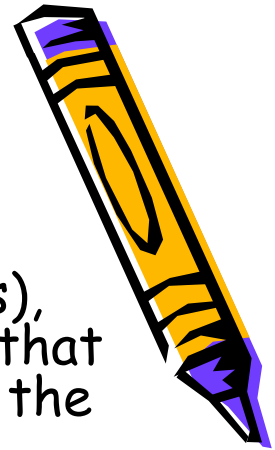
1. The *app_sourcetypes* table contains the two document types (oral / written).
2. The *app_modules* table defines the 11 modules needed for the processing.
3. The *propertygroups* table is used for two reasons: (a) the definition of thematic subsets of attributes and (b) the definition of predefined values of attributes (lookups).
4. The auxiliary *propertygrouptypes* 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 Strcuct 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 insetad 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.

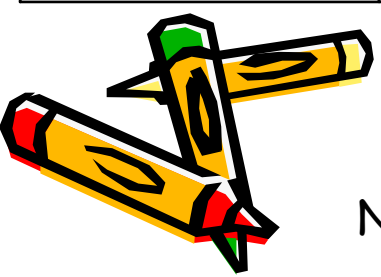
Need for a relation for the Inner level



- 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.

Tentative 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

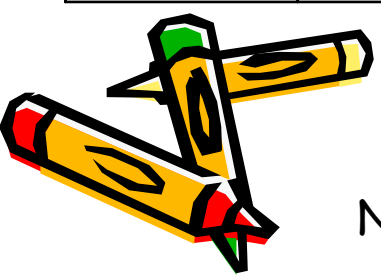
- All retrieved intervals of the same tier can be ordered based on the property **interval_no** which emanates from the original TextGrid file.
- In this way, we can formulate queries concerning the distance between segments at a certain level.
- The values of the attributes phenomenon, start and stop 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

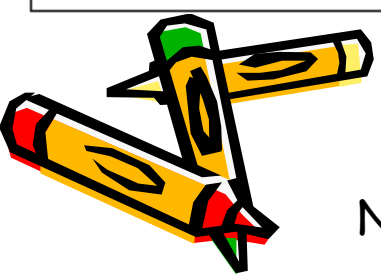
Word/ token /phenomenon			**Location**				
<Value>	{ Between, And, Or, -- }	<Value>	<DB>	<At-tribute >	<Part distances>	<Word distances>	<Interval_no distances >



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	-

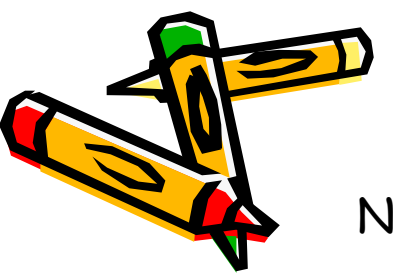
Output	Part	-



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

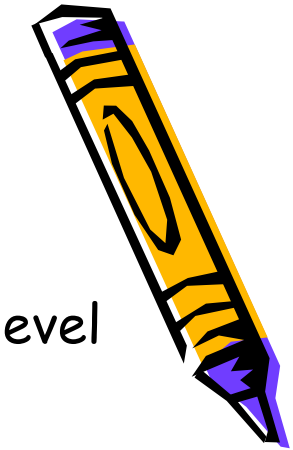


search based on metadata of participants

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	-	-	-

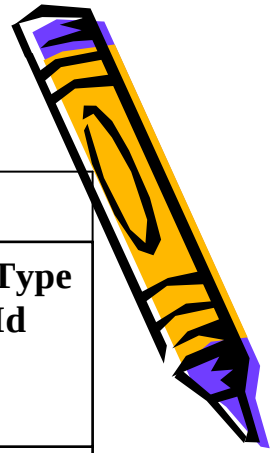
Output	Document	-

EAV implementation of Inner level annotations



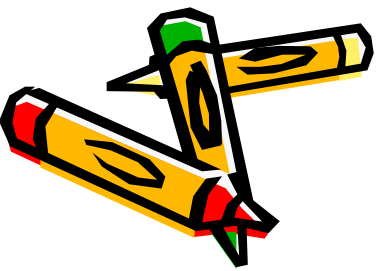
- The tentative relational implementation for the inner level of annotations has constraints as for the expansion of annotations caused by the schema evolution problem.
- In order to solve this problem, we have decided to store information on the stress of the vowels in the EAV database.
- Instead of using the previous table (the Inner level Relation), we have placed the *oral_segments* table in the *Struct* database and also created the *oral_SegmentTypes* auxiliary table.
- Actually, the oral_segments table contains only the transciptions in the lower levels (syllable, phoneme, etc) while, the eventual lower level element attributes are stored in the EAV database where the *Properties* and *PropertyValues* tables are updated with information on stress and stress position.

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

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, Καππαδοκία, 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_extended.pdf

Closing

- Thank you for your attention!
- Questions can be asked.

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