

Αθήνα, Παρασκευή 1 Οκτωβρίου 2010

***GREED CORPUS: ΟΡΓΑΝΩΣΗ ΚΑΙ
ΕΡΓΑΛΕΙΑ ΓΛΩΣΣΙΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ ΨΗΦΙΑΚΗΣ ΒΑΣΗΣ ΕΛΛΗΝΙΚΩΝ
ΔΙΑΛΕΚΤΩΝ ΤΟΥ ΕΡΓΑΣΤΗΡΙΟΥ
ΝΕΟΕΛΛΗΝΙΚΩΝ ΔΙΑΛΕΚΤΩΝ***

Δημήτρης Παπαζαχαρίου

Εργαστήριο Νεοελληνικών Διαλέκτων
Τμήμα Φιλολογίας, Πανεπιστήμιο Πατρών

Ψηφιακή Βάση Ελληνικών Διαλέκτων

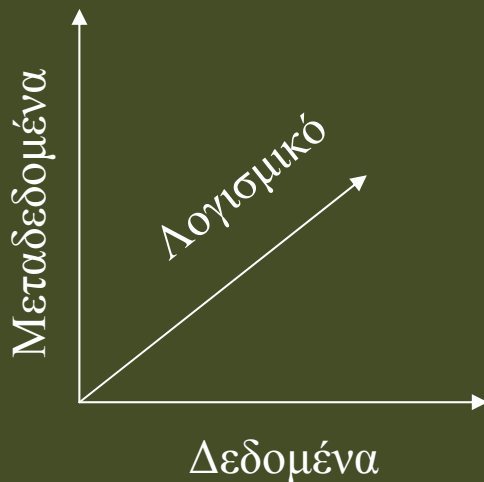
- Χαρακτηριστικά: α) Ανεξαρτησία από λειτουργικά συστήματα και εμπορικά λογισμικά

β) Τρισδιάστατη οργάνωση βάσης ως προς:

i) τα *Δεδομένα*

ii) τα *Μεταδεδομένα* που χαρακτηρίζουν τα δεδομένα

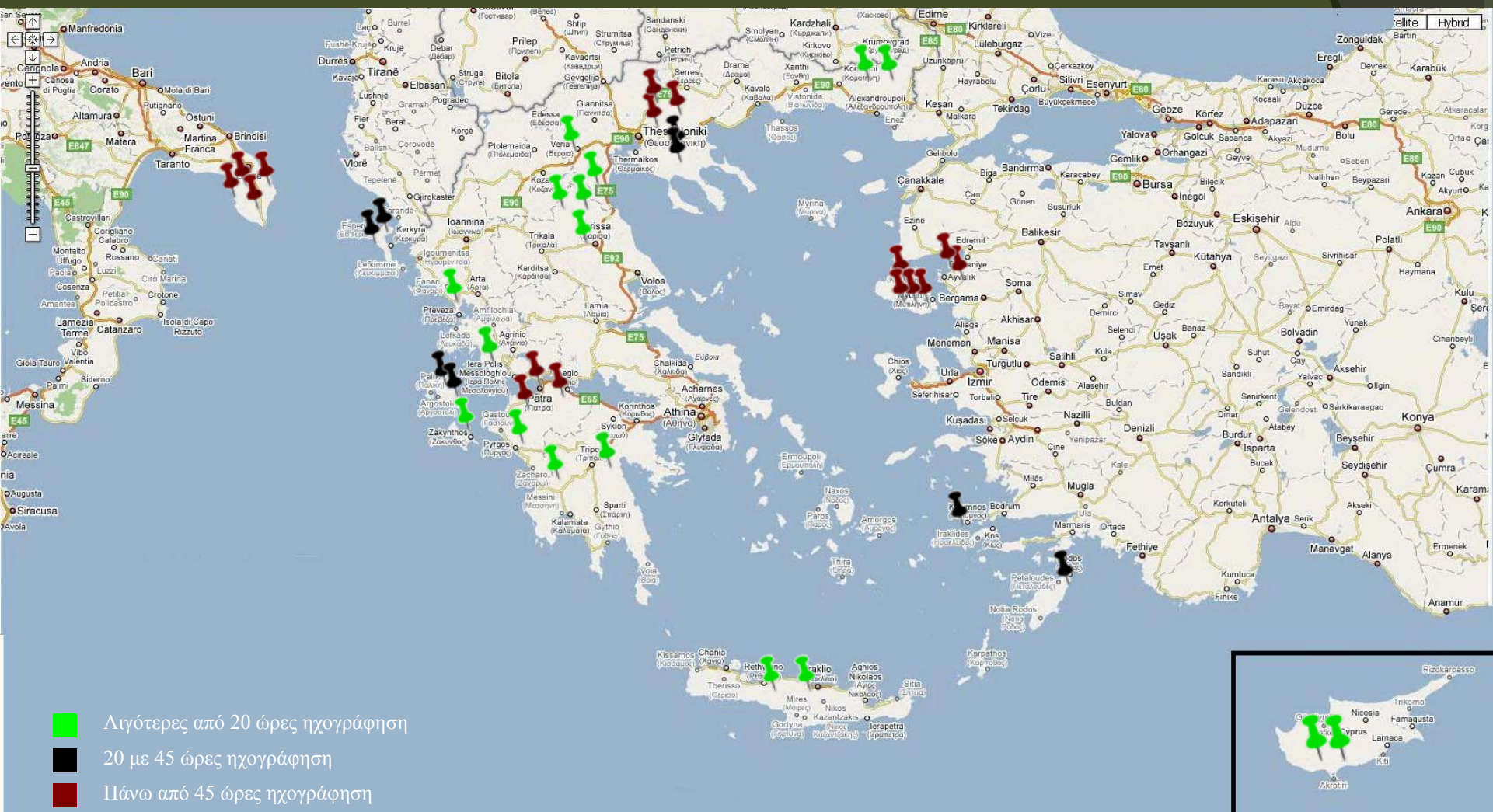
iii) το *Λογισμικό*, που μας επιτρέπει την α) διαχείριση των δεδομένων και μεταδεδομένων, και β) αναζήτηση των δεδομένων σε σχέση με τα μεταδεδομένα που τα περιγράφουν



A) Δεδομένα (πρωτογενή)

- Ερευνητικά προγράμματα + Διπλωματικές εργασίες προπτυχιακών και μεταπτυχιακών φοιτητών της ειδίκευσης σε διαλεκτικά φαινόμενα από την Κύπρο μέχρι την Θράκη, οι οποίες προϋπέθεταν ηχογραφήσεις διαλεκτικών συνομιλιών
- Σύνολο: **505** ώρες μαγνητοφωνημένου διαλεκτικού λόγου

Διαλεκτικός χάρτης



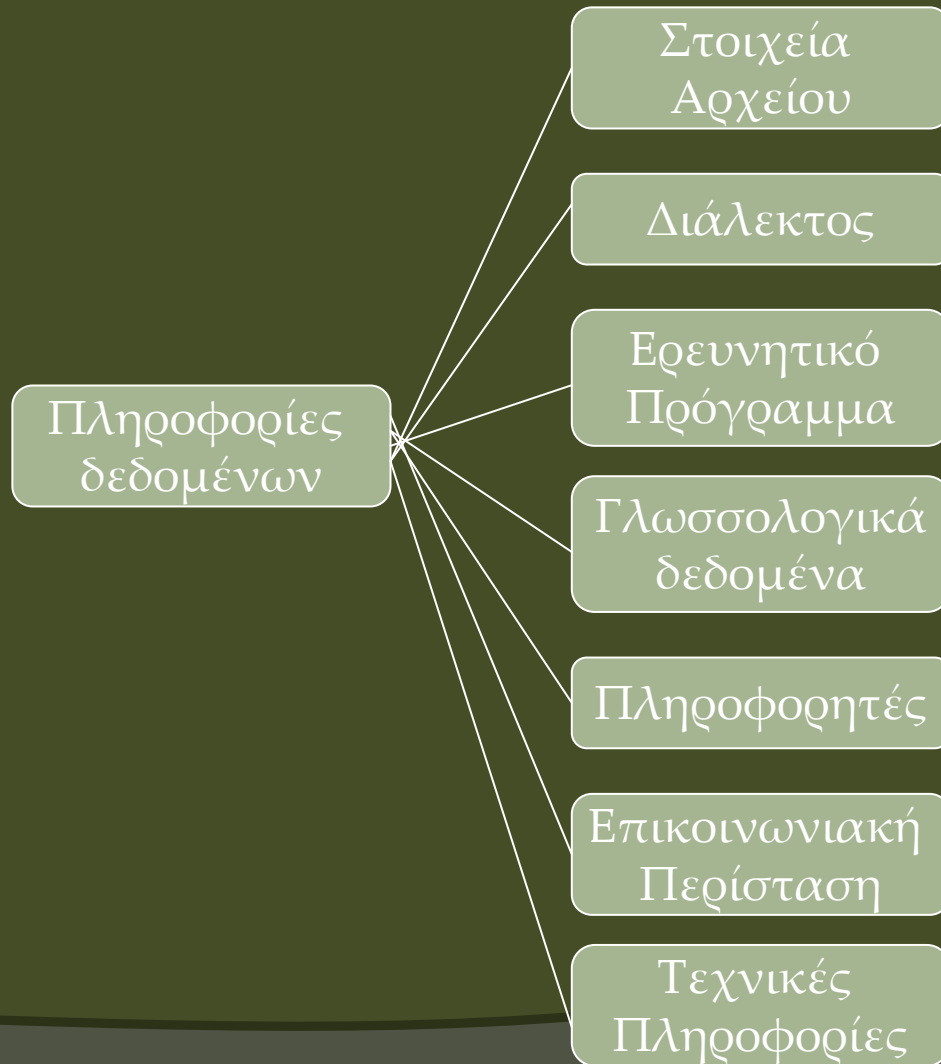
A) Δεδομένα (επεξεργασμένα)

- ◎ Απομαγνητοφωνήσεις των ηχητικών αρχείων
(ορθογραφική μεταγραφή)

B) Γλωσσικά μεταδεδομένα

- ◎ Κωδικοποίηση (annotation)
(Φωνητική & Προσωδιακή, Φωνολογική,
Μορφολογική)

B) Μη Γλωσσικά Μεταδεδομένα



B) Μη γλωσσικά μεταδεδομένα: υποκατηγοριοποίηση



Λογισμικό – Κριτήρια Επιλογής

- ⊙ Λογισμικά αιχμής και ευρείας αποδοχής
- ⊙ Λογισμικά ανοιχτού κώδικα και ελεύθερα στην χρήση
- ⊙ Πληθώρα επιλογών για παραμετροποίηση
- ⊙ Δυνατότητα συνδυασμού των λογισμικών μεταξύ τους
- ⊙ Συνεχή υποστήριξη από τους δημιουργούς/προγραμματιστές
- ⊙ Υποστήριξη ελληνικής γλώσσας και unicode γραμματοσειρών

Λογισμικό για μεταγραφή (ELAN)

The screenshot displays the ELAN software interface. The main window shows an audio file named "Elan - KAL_MK02_MKF87_GKM66_MKM55_NKF60.eaf". The interface includes a menu bar (File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help) and a toolbar with various playback and editing controls. Below the toolbar, there are volume and rate sliders, both set to 100. The audio waveform is visible, with a selection range from 00:00:00.000 to 00:00:00.000. The transcription window shows several tiers with text segments and time markers. The tiers are labeled as follows:

- default [0]
- Maria [290]
- Artemis [712]
- MKF87 [239]
- GKM66 [525]
- MKM55 [253]
- NKF60 [68]

The transcription text includes:

0063 ναι α ναι το παιι μου | 0065 ήταν τ'

0064 ελχετε πάει, α | 0066 και το 'λεγα στο Μαράκι που | 0069 παα

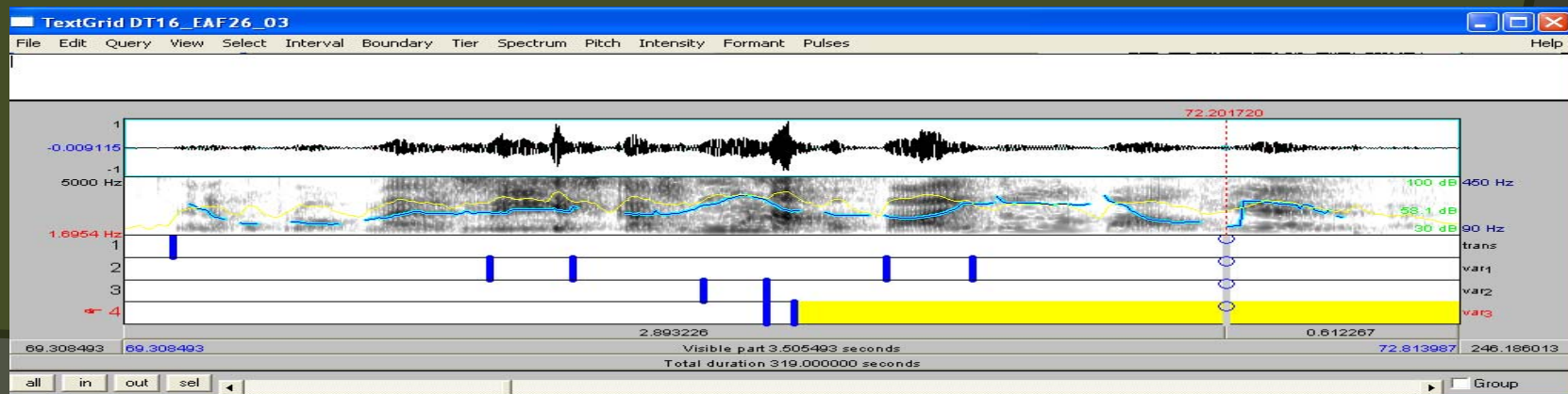
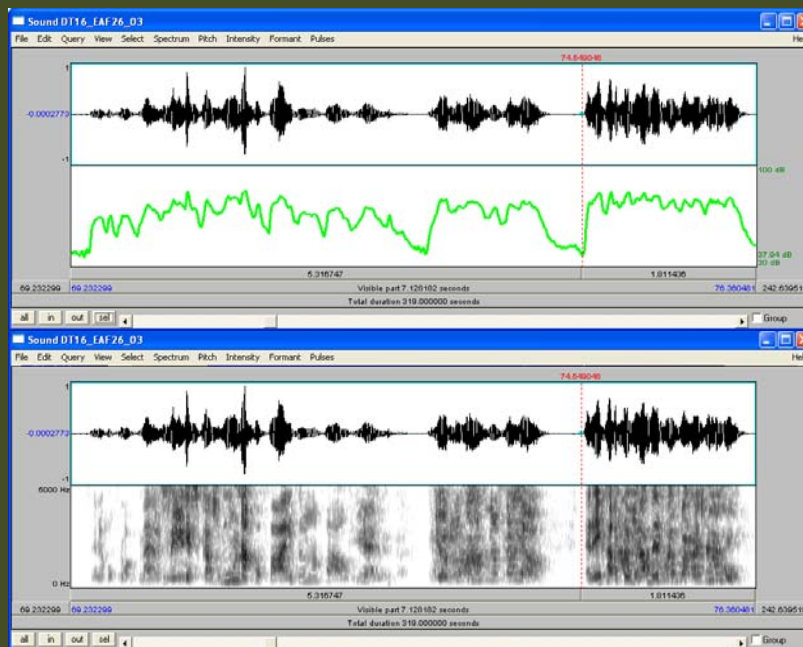
0067 ελχε μπουζούκια ελχε σαμπούνα ελχε φαγοπότι

0068 κάτσαμε και παίζαμε μέχρι της τρεη

The "About ELAN" dialog box is open, displaying the following information:

- About ELAN...** (selected tab)
- Acknowledgments** (tab)
- ELAN - ELAN Linguistic Annotator**
- Version: 3.5.0**
- Copyright © 2001 - 2008**
- Max-Planck-Institute for Psycholinguistics**
- Nijmegen, The Netherlands**
- Source code for this version available under GPL (<http://www.gnu.org>)
- OK** button

Λογισμικό για Φωνητική / Φωνολογική Ανάλυση (Praat)



Λογισμικό για Μορφολογική Ανάλυση (Toolbox)

The screenshot displays the 'Toolbox - Dictionary' application window. The main window shows a morphological analysis of the Greek phrase 'Τα παιδιά της γειτονιάς παίζουν και'. The analysis is presented in a table format with columns for the original text, morphological tags, and English glosses.

Original Text	Morphological Tag	English Gloss
Τα παιδιά της γειτονιάς παίζουν και	child nom.pl.neu gen.sing.fem neighbour n.der.suffix gen.sing.fem play and	child nom.pl.neu gen.sing.fem neighbour n.der.suffix gen.sing.fem play and
γελοῦσ	inf.suf art n der.suf inf.suf v comp	inf.suf art n der.suf inf.suf v comp
γελάσ	past.inf.suf 3.sing.Pr full acc.sing.neu joy and happy n.der.suffix	past.inf.suf 3.sing.Pr full acc.sing.neu joy and happy n.der.suffix
χαρά και ευτυχία	adj inf.suf n comp adj der.suf	adj inf.suf n comp adj der.suf

Below the main window, there are two smaller windows: 'Dictionary.txt:2' and 'Dictionary.txt:1'. 'Dictionary.txt:2' shows a list of lexemes, with 'ιά' selected. 'Dictionary.txt:1' shows the detailed entry for 'χαρά', including its part of speech (n), gloss (joy), and date of last edit (02/Jul/2008).

The Windows taskbar at the bottom shows the system tray with the time 1:36 μμ and the date 5/23/2008. The active window is 'Toolbox - Dictionary.txt'.

Γ) Περιβάλλον Βάσης Δεδομένων

Βρέθηκαν συνολικά **66** εγγραφές.

metadata Παρακαλώ μην επιλέγετε <Μεταβλητή Συστήματος>

attribute EL

attribute EN

Αναζήτηση (1=ΝΑΙ, 0=ΟΧΙ)

Εμφάνιση (1=ΝΑΙ, 0=ΟΧΙ)

Υποβολή

Metadata EL	attribute EL	Metadata EN	attribute EN	Πεδίο αναζήτησης	Πεδίο Εμφάνισης		
Στοιχεία Αρχείου	Αύξων αριθμός αρχείου	File Properties	File Number	ΟΧΙ	ΝΑΙ		
Στοιχεία Αρχείου	Ελεύθερο/Κλειδωμένο	File Properties	Free/ Under processing	ΟΧΙ	ΟΧΙ		
Στοιχεία Αρχείου	Όνομα Αρχείου	File Properties	File Name	ΟΧΙ	ΟΧΙ		
Στοιχεία Αρχείου	Φάκελος	File Properties	Folder	ΟΧΙ	ΝΑΙ		
Διάλεκτος	Γεωγραφικός προσδιορισμός διαλέκτου	dialect	Geographical specification of Dialect	ΝΑΙ	ΝΑΙ		
Διάλεκτος	Μέρος ηχογράφησης	dialect	Meros hoxografisis	ΝΑΙ	ΝΑΙ		
Διάλεκτος	Όνομα διαλέκτου	dialect	Dialect Name	ΝΑΙ	ΝΑΙ		
Διάλεκτος	Συνοχή Τόπου Ηχογράφησης	dialect	Sinoxhi topou ixigrafisis	ΟΧΙ	ΟΧΙ		
Διάλεκτος	Τόπος ηχογράφησης (Διάλεκτος)	dialect	Place of Recording	ΝΑΙ	ΝΑΙ		
Διάλεκτος	Χρόνος ηχογράφησης	dialect	Date of Recording	ΟΧΙ	ΝΑΙ		
Ερευνητικό Πρόγραμμα	Απομαγνητοφωνητής	Research program	Transcriber	ΟΧΙ	ΟΧΙ		
Ερευνητικό Πρόγραμμα	Ενδιάμεσος	Research	Local Contact	ΝΑΙ	ΝΑΙ		

Find: όνομα δ Match case

Done

Ο Διαχειριστής της βάσης δεδομένων μπορεί να εισάγει όσες υποκατηγορίες metadata (attributes) επιθυμεί.

Γ) Αναζήτηση δεδομένων











Όνομα διαλέκτου Γεωγραφικός προσδιορισμός διαλέκτου Τόπος ηχογράφησης (Διάλεκτος) Όνομα Ερευνητικού Προγράμματος Επιστημονικός υπεύθυνος

Ερευνητική ομάδα Υπεύθυνος συλλογής υλικού

Ερευνητής πεδίου Ενδιάμεσος Είδος πρωτότυπης ηχογράφησης

Είδος λόγου Φύλο Τόπος Ηχογράφησης (Τεχν. Πλ) Μέρας ηχογράφησης

Βρέθηκαν συνολικά 471 εγγραφές

	Αύξων αριθμός αρχείου	Όνομα διαλέκτου	Γεωγραφικός προσδιορισμός διαλέκτου	Τόπος ηχογράφησης (Διάλεκτος)	Χρόνος ηχογράφησης	Όνομα Ερευνητικού Προγράμματος	Επιστημονικός υπεύθυνος	Ερευνητική ομάδα	Υπεύθυνος συλλογής υλικού
 	EP_GG01_ASM??_MC??	Eptanisian	-	Zakinthos	1/7/2001	-	-	-	-
 	ND_VK02_--F18	Northern Dialects	-	Veroia	1/11/2002	UC	Dimitris Papazachariou	-	-
 	ND_VK02_E-F45	Northern Dialects	-	Veroia	1/11/2002	UC	Dimitris Papazachariou	-	-
 	ND_VK01_--F66_--F65	Northern Dialects	-	Veroia	1/11/2002	UC	Dimitris Papazachariou	-	-
 	ND_VK01_--F44	Northern Dialects	-	Veroia	1/11/2002	UC	Dimitris Papazachariou	-	-

Find: Match case

Τα πεδία αναζήτησης εμφανίζονται ως combo boxes και παράγονται στη σελίδα δυναμικά. Αυτό σημαίνει ότι ανάλογα με τη συμπλήρωση των προηγούμενων σελίδων, από τον διαχειριστή του συστήματος, παράγονται δυναμικά τα πεδία αναζήτησης καθώς και τα πεδία εμφάνισης.

Μελλοντικοί Στόχοι

- ◎ Παλαιά κείμενα και ιστορικά έγγραφα (αρχεία εικόνας)
- ◎ Μηχανή αναζήτησης που να συνδυάζει και γλωσσικές και μη γλωσσικές πληροφορίες
- ◎ Διαβαθμισμένα επίπεδα πρόσβασης για τους χρήστες
- ◎ Περισσότερες μεταγραφές / περισσότερη κωδικοποίηση