Ένα πολυτροπικό σύστημα αρχειοθέτησης και διαχείρισης γραπτών και προφορικών πηγών μελέτης της γλώσσας και των γλωσσικών ιδιωμάτων[1]

Νικήτας Ν. Καρανικόλας[2], Ελένη Γαλιώτου, Κωνσταντίνος Αθανασάκος & Γεώργιος Κορωνάκης
Τμήμα Μηχανικών Πληροφορικής
Τεχνολογικό Εκπαιδευτικό Ίδρυμα Αθηνών
nnk@teiath.gr, egali@teiath.gr, k.athanasakos@gmail.com, gkoronakis@gmail.com

Περίληψη

Η παρούσα εργασία παρουσιάζει το σχεδιασμό και την υλοποίηση ενός συστήματος δημιουργίας και διαχείρισης σώματος τεκμηρίων από γραπτές και προφορικές πηγές. Η δομή του συστήματος επιτρέπει και το περιεχόμενο του σώματος αξιοποιεί μια ιδιαίτερα μεγάλη ποικιλία μεταγραφών και επισημειώσεων. Περαιτέρω, οι δομές δεδομένων του συστήματος είναι τέτοιες που επιτρέπουν τη δυναμική εξέλιξη του συστήματος σε δεδομένα (database schema evolution) και σε λειτουργικότητα (πολλά υποσυστήματα επεξεργασίας δημιουργούνται δυναμικά από το ίδιο προσαρμόσιμο υποσύστημα). Επιπροσθέτως το σύστημα έχει σχεδιαστεί και παρέχει ένα ιδιαίτερα έξυπνο υποσύστημα δημιουργίας ερωτημάτων ανάκτησης (query builder) που μπορεί να συνδυάζει συνθήκες από διαφορετικά επίπεδα επισημειώσεων (φωνολογικά, μορφολογικά, συντακτικά, μεταδεδομένων, κλπ). Η εφαρμογή του συστήματος έχει γίνει σε τεκμήρια (γραπτά και προφορικά) τριών διαλέκτων της Μικράς Ασίας (Ποντιακά, Καππαδοκικά, Αϊβαλιώτικα). Εκτιμούμε ότι μπορεί να χρησιμοποιηθεί για την κατασκευή και διαχείριση σωμάτων τεκμηρίων άλλων γλωσσών ή διαλέκτων.

Λέξεις - Κλειδιά: σώμα τεκμηρίων, γλωσσικά εργαλεία, υπολογιστική γλωσσολογία, Ελληνικές διάλεκτοι

1. Εισαγωγή

1.1 Σκοπός

Ο τίτλος του έργου στο οποίο συμμετείχαμε και κατέληξε σε παραδοτέα μεταξύ των οποίων και το παρόν άρθρο ήταν «Pontus, Cappadocia, Aivali: In search of Asia Minor

[1] Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: ΘΑΛΗΣ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

[2] Corresponding author

Greek». Μια απόδοση στα Ελληνικά είναι «Πόντος, Καππαδοκία, Αϊβαλή: στα ίχνη του Ελληνισμού της Μικράς Ασίας». Από τον τίτλο και μόνο προκύπτει ο σκοπός καταγραφής στοιχείων που χαρακτηρίζουν τον Ελληνισμό της Μικράς Ασίας. Σημαντικός τρόπος έκφρασης του κάθε λαού / έθνους είναι η γλώσσα του όπως αυτή πραγματώνεται και χρησιμοποιείται σε κάθε μορφή γραπτής και προφορικής επικοινωνίας. Έτσι σε αυτό το έργο επιχειρήθηκε μια συστηματική οργάνωση τεκμηρίων, γραπτών και προφορικών πηγών, της Ελληνικής γλώσσας και ορισμένων εκ των διαλέκτων αυτής στη Μικρά Ασία, προκειμένου να συμβάλουμε στην διατήρηση αυτών και στην καλύτερη αξιοποίηση τους από επόμενες γενιές. Το σημαντικότερο ίσως γεγονός, που ενισχύει την αναγκαιότητα για το σκοπό αυτού του έργου, αποτέλεσε η συνθήκη της Λωζάνης (1923). Η τότε ανταλλαγή πληθυσμών και η χαλάρωση της μνήμης που αναγκαστικά επέρχεται σε κάθε επόμενη γενιά οδηγεί σε σταδιακή απομάκρυνση από τον πολιτισμό, τις παραδώσεις και τη γλώσσα των απογόνων των Ελλήνων της Μικράς Ασίας. Είναι λοιπόν σημαντικό να διατηρήσουμε τις πηγές αλλά επιπλέον και να τις οργανώσουμε σε προσβάσιμα συστήματα προκειμένου να επιτρέπουμε την ευκολότερη επίσκεψη αυτών καθώς και τη διατήρηση των καταγεγραμμένων φαινόμενων και παρατηρήσεων που έχουν κατά καιρούς κάνει οι μελετητές της γλώσσας και των διαλέκτων της Μικράς Ασίας.

Τα κυριότερα συστήματα που αναπτύχθηκαν σε αυτό το έργο είναι δύο (2). Ένα τριδιαλεκτικό λεξικό και μία πολυτροπική βάση αρχειοθέτησης γραπτών και προφορικών πηγών καθώς και φαινόμενων και παρατηρήσεων που έχουν κατά καιρούς κάνει οι μελετητές των πηγών. Στο άρθρο αυτό θα εστιάσουμε στο δεύτερο σύστημα (πολυτροπική βάση). Πρόκειται για ένα σώμα τεκμηρίων (corpora) και ένα σύνολο λογισμικών εργαλείων για την οργάνωση και διαχείριση του σώματος. Πρώιμες σκέψεις για την πολυτροπική βάση έχουν τεκμηριωθεί σε άλλες πρόσφατες εργασίες μας (Galiotou et al 2014˙ Karanikolas et al 2014˙ Karasimos et al 2014). Πρώιμες σκέψεις για το εργαλείο (module) αναζήτησης δεδομένων έχουν τεκμηριωθεί πολύ πρόσφατα (Karanikolas et al 2015).

1.2 Άλλες προσπάθειες

Υπάρχει σημαντικός αριθμός προσπαθειών για τη δημιουργία σωμάτων κειμένων και/ή εργαλείων για τη διαχείριση σωμάτων κειμένων. Το LAMSAS project (Linguistic Atlas of Middle and South Atlantic States) αφορά την αρχειοθέτηση και επεξεργασία δεδομένων από παράκτιες προς τον Ατλαντικό περιοχές των ΗΠΑ (Nerbonne et al 2003). Σε μια άλλη εργασία (Ubul et al 2015) περιγράφεται η προσπάθεια δημιουργίας μιας πολυμεσικής βάσης με σκοπό τη μελέτη (την αντιστοίχηση – voice language map) των υπό εξαφάνιση διαλέκτων των διαμερισμάτων (regions) της Ιαπωνίας. Σχετικά με την Ευρώπη μπορούμε να αναφέρουμε το DynaSAND (Dynamic Syntactic Atlas of Dutch Dialects) on-line εργαλείο για την επεξεργασία των Ολλανδικών συντακτικών ποικιλιών (Barbiers et al 2006). Το τελευταίο διαθέτει μια web service διεπαφή για την αξιοποίηση του σώματος από άλλες εφαρμογές (Kunst et al 2010). Το SCOTS (Scottish Corpus of Text and Speech) είναι ίσως το εγγύτερο από άποψη στόχων με τη δική μας προσπάθεια καθώς στοχεύει σε ένα μεγάλης έκτασης ηλεκτρονικό σώμα γραπτών και προφορικών τεκμηρίων των γλωσσών της Σκωτίας (Anderson et al 2007). Για τις Καταλανικές διαλέκτους μπορούμε να αναφέρουμε το ηλεκτρονικό σώμα COD (Corpus Oral Dialectal) (Clua et al 2006).

2. Πηγές και αναπαράσταση τεκμηρίων στο σύστημα

Οι πηγές τις οποίες οργανώνει και αρχειοθετεί το σύστημα γραπτών και προφορικών πηγών, φαινόμενων και παρατηρήσεων χωρίζονται σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει βιβλία, περιοδικά, ανθολόγια, ποιήματα, παραμύθια, χειρόγραφα και γενικότερα γραπτά τεκμήρια με κείμενα τριών διαλέκτων της Μικράς Ασίας. Στις περισσότερες περιπτώσεις τα τεκμήρια αυτά είναι συνταγμένα από επιστημονικά καταρτισμένους μελετητές αλλά και μη επιστημονικά καταρτισμένους μελετητές που καταγράφουν (υποδεικνύουν) τα παρατηρούμενα φαινόμενα της γλώσσας/διαλέκτου με τη χρήση συμβόλων (συνήθως χαρακτήρων) που μπορεί να μην εντάσσονται στο αλφάβητο της μελετώμενης γλώσσας. Μάλιστα μπορεί διαφορετικοί μελετητές να συμβολίζουν το ίδιο φαινόμενο με διαφορετικά σύμβολα.

Το σύστημα, για αυτή (την πρώτη) κατηγορία πηγών διατηρεί την αρχική επισημείωση του αρχικού μελετητή αλλά και μία μεταγραφή που χρησιμοποιεί ομογενοποιημένα σύμβολα που είναι τα ίδια σε όλες τις μεταγραφές και είναι ανεξάρτητα του αρχικού μελετητή. Πέρα από αυτές τις πληροφορίες (αρχική μορφή τεκμηρίου ως έγινε από τον αρχικό μελετητή και μεταγραφή με ομογενοποιημένα σύμβολα) το σύστημα διατηρεί μορφολογικές και φωνολογικές επισημειώσεις με σημείο αναφοράς τη λέξη. Έχει επίσης προβλεφθεί ώστε το σύστημα να μπορεί να διατηρεί συντακτικές και σημασιολογικές επισημειώσεις σε επίπεδο ακολουθίας από διαδοχικές λέξεις (φράσεις, προτάσεις, κλπ).

Η δεύτερη κατηγορία πηγών περιλαμβάνει ηχογραφήσεις προφορικού λόγου. Πρόκειται για διαλόγους ή συνεντεύξεις από εναπομείναντες σήμερα ομιλητές κάποιας εκ των μελετώμενων διαλέκτων. Στις ηχογραφήσεις αυτές συνήθως επιβλέπει ένας σύγχρονος μελετητής ή μπορεί ακόμα και να συμμετέχει σε αυτές. Η κάθε ηχογράφηση συνοδεύεται από επισημειώσεις του ήχου με σύγχρονα εργαλεία που υπήρχαν πριν από το έργο AMiGre. Οι επισημειώσεις αυτές γίνονται σε δίαφορα επίπεδα μικρότερης ή μεγαλύτερης διάρκειας (ενδεικτικά αναφέρουμε επιτονικές προτάσεις, επιτονικές λέξεις, μορφολογικές λέξεις και φωνήματα) και μπορεί να ακολουθούν πρότυπα (π.χ. IPA (IPAa˙ IPAb), SAMPA (Wells 1997˙ SAMPA)) ή να είναι περισσότερο ελεύθερες (χωρίς συγκεκριμένο πρότυπο ή με ιδιωτικό πρότυπο). Επίσης στην δεύτερη κατηγορία πηγών μπορεί να υπάρχουν επισημειώσεις που αφορούν στιγμιαία φαινόμενα (π.χ. τονισμός).

Το σύστημα, για τη δεύτερη κατηγορία πηγών (προφορικές πηγές) διατηρεί ψηφιακή έκδοση του ηχητικού τεκμηρίου και όλες τις μεταγραφές που υπάρχουν για όλα τα επίπεδα επιμερισμού (ευρύτερης ή στενότερης) χρονικής διάρκειας (όπως για παράδειγμα επιτονικές προτάσεις, επιτονικές λέξεις, μορφολογικές λέξεις και φωνήματα) και για όλες τις στιγμιαίες επισημειώσεις του τεκμήριου. Στην περίπτωση των προφορικών πηγών το σύνολο των επισημειώσεων που διατηρεί το σύστημα για το ίδιο τεκμήριο είναι από τη γέννηση τους (που ενδεικτικά αναφέρουμε ότι μπορεί να γινόταν από το εργαλείο praat ή το élan) ένα ενιαίο αντικείμενο (ψηφιακό αρχείο).

3. Αναπαράσταση τεκμηρίων πριν τη δημιουργία του συστήματος

Στις προηγούμενες παραγράφους αναφερθήκαμε στις πηγές και στην αναπαράσταση των τεκμηρίων των πηγών στο σύστημα γραπτών και προφορικών πηγών που εδώ

παρουσιάζουμε. Στη συνέχεια θα αναφερθούμε στην ψηφιακή αναπαράσταση των τεκμηρίων όταν δεν υπήρχε (πριν δημιουργηθεί) το σύστημα γραπτών και προφορικών πηγών. Τότε υπήρξε μια σημαντική κατάσταση που πρέπει να αναδειχθεί. Στην προ συστήματος γραπτών και προφορικών πηγών εποχή, για κάθε τεκμήριο οι πληροφορίες αυτού (ψηφιοποιημένη μορφή του πρωτότυπου τεκμηρίου, μεταγραφή και φωνολογικές και μορφολογικές επισημειώσεις σε επίπεδο λέξης) δεν ευρίσκονταν σε ένα ενιαίο ψηφιακό αρχείο (εγγραφή), ούτε και ήταν επεξεργάσιμες από ένα και μόνο λογισμικό. Για παράδειγμα ένα τεκμήριο γραπτής πηγής μπορεί να είχε ένα αρχείο tiff ή jpeg για το ψηφιοποιημένο πρωτότυπο, ένα αρχείο επεξεργαστή κειμένου ή απλού κειμένου (απλό text) για την μεταγραφή και ένα ή περισσότερα αρχεία λογιστικών φύλλων (spreadsheets) για τις μορφολογικές και φωνολογικές επισημειώσεις των λέξεων του τεκμηρίου. Έτσι, καθώς αυτά τα αρχεία απαιτούν διαφορετικά λογισμικά επεξεργασίας, δεν μπορούσαν να εκληφθούν ως ένα ψηφιακό αρχείο (μία εγγραφή).

Σε ανάλογη κατάσταση ήταν και οι πληροφορίες των τεκμηρίων προφορικών πηγών καθώς δεν μπορούσαν να ευρίσκονται σε ένα ενιαίο (ή δυνάμενο να εκληφθεί ως ενιαίο) ψηφιακό αρχείο (εγγραφή) και δεν ήταν επεξεργάσιμες (στο σύνολο των πληροφοριών του τεκμηρίου προφορικής πηγής) από ένα λογισμικό. Ένα τεκμήριο προφορικής πηγής μπορεί να το αποτελούσαν ένα αρχείο ψηφιακού ήχου, ένα αρχείο επισημειώσεων (έστω ένα αρχείο που παράγεται από το λογισμικό Praat, βλέπε παρακάτω) και ένα λογιστικό φύλλο με μεταπληροφορίες κυρίως για τους ομιλητές (ηλικία, φύλο, επάγγελμα, καταγωγή, γραμματικές γνώσεις, κλπ). Τα αρχεία που συνέθεταν το ψηφιακό τεκμήριο προφορικής πηγής δεν ήταν στο σύνολο τους επεξεργάσιμα από το ίδιο λογισμικό. Ως εκ τούτου κάθε τεκμήριο προφορικής πηγής δεν μπορούσε να εκληφθεί ως ένα ψηφιακό αρχείο (μία εγγραφή).

Η κατακερματισμένη μορφή των τεκμηρίων σε περισσότερα από ένα ψηφιακά αρχεία ήταν ένας από τους λόγους που οδήγησαν στη δημιουργία του συστήματος γραπτών και προφορικών πηγών.

4. Διαθέσιμα εργαλεία – δυνατότητες και περιορισμοί

4.1 Praat

Το Praat (Boersma 2012˙ Boersma et al 2013) είναι ένα δωρεάν διαθέσιμο λογισμικό ανάλυσης και επεξεργασίας ακουστικών σημάτων και ήχων. Βασίζεται σε ψηφιακές ηχητικές καταγραφές. Τέτοιες (μονοφωνικές ή στερεοφωνικές ηχογραφήσεις) μπορούν είτε να δημιουργηθούν μέσα από το praat ή να ανοιχθούν από ένα αρχείο που έχει δημιουργηθεί με άλλο τρόπο (πρόγραμμα). Μεταξύ των δυνατοτήτων που παρέχει, επιτρέπει να φιλτράριστεί το σήμα (να εξαιρεθεί κάποιο εύρος συχνοτήτων) και να ενισχυθούν κάποιες άλλες συχνότητες. Σπουδαιότερη, κατά την άποψη μας, είναι η δυνατότητα να τεμαχισθεί μια κυματομορφή ομιλίας σε διάφορα τμήματα (διαστήματα) της και να επισυναφθούν επισημειώσεις (ετικέτες, χαρακτηρισμοί, προσδιορισμοί ιδιοτήτων και φαινόμενων) για κάθε τμήμα στο οποίο έχει τεμαχίσει η κυματομορφή. Ο τεμαχισμός και η επισημείωση μπορεί να γίνει σε περισσότερα του ενός επίπεδα με διαφορετικού μήκους τμήματα κυματομορφής. Ενδεικτικά: μια ομιλία (μια κυματομορφή) μπορεί να χωρισθεί σε επιτονικές προτάσεις, επιτονικές φράσεις, μορφολογικές λέξεις, φωνήματα, κλπ. Η επισημείωση με το Praat μπορεί να αφορά και σημεία. Έτσι, για παράδειγμα, μπορεί να επισημειωθεί ο τονισμός των λέξεων και ο

επιτονισμός των φράσεων. Ολες οι επισημειώσεις του Praat φυλάγονται σε text αρχείο (με κατάληξη .TextGrid).

4.2 ELAN

Το ELAN (the Eudico Linguistic Annotator, (ELAN˙ Sloetjes et al 2008)) είναι λογισμικό που επιτρέπει την εισαγωγή επισημειώσεων και σχολιασμών σε αρχεία ψηφιακού ήχου και σε αρχεία ψηφιακού βίντεο. Η επισημειώσεις μπορούν να είναι σε πολλά επίπεδα. Για παράδειγμα μπορεί να είναι ένας απλός τίτλος που αφορά όλο το αρχείο (ήχου ή βίντεο), να είναι μια ελεύθερη μετάφραση των διαλόγων, να είναι ένας διαχωρισμός σε λέξεις ή ακόμα και σε φωνήματα. Μπορούν επίσης να μπουν και άλλες επισημειώσεις όπως θεματικός προσδιορισμός των διαφορετικών ενοτήτων του ψηφιακού αρχείου. Θα μπορούσε κανείς να εξάγει το συμπέρασμα ότι το ELAN είναι περίπου το ίδιο με το Praat και επιπλέον υποστηρίζει και ψηφιακό βίντεο. Υπάρχουν όμως βασικές διαφορές καθώς τα επίπεδα στο Praat δεν είναι συσχετισμένα μεταξύ τους ενώ στο ELAN είναι συσχετισμένα ιεραρχικά. Έτσι στο ELAN ελέγχεται αν μια ακολουθία από διαδοχικά τμήματα ενός επιπέδου (tier) έχει ακριβώς τα ίδια όρια (συμπίπτει) με ένα τμήμα του πατρικού επιπέδου (parent tier). Επιπλέον στο ELAN οι τιμές σχολιασμού (επισημειώσεων) μπορούν (προαιρετικά) να είναι προκαθορισμένες και να ελέγχονται (ελεγχόμενο λεξιλόγιο). Τα ελεγχόμενα λεξιλόγια δίνουν στο ELAN καλύτερες δυνατότητες για μορφολογικές, συντακτικές και άλλες επισημειώσεις. Για παράδειγμα μπορούμε να ορίσουμε ένα ελεγχόμενο λεξιλόγιο που περιέχει τα μέρη του λόγου της Ελληνικής γλώσσας και να μην έχει δυνατότητα ο χρήστης να επισημειώσει (στο tier της μορφολογικής επισημείωσης) κάτι άλλο που δεν είναι μέρος του λόγου. Οι επισημειώσεις του ELAN φυλάγονται σε text αρχείο (με κατάληξη .eaf και δομή XML).

4.3 LaBB-CAT (ONZE Miner)

Το LaBB-CAT (Fromont et al 2008˙ LaBB-CAT) είναι ένα πρόγραμμα περιήγησης, βασίζεται σε γλωσσολογικά εργαλεία και αποθηκεύει ηχογραφήσεις ή βιντεοσκοπήσεις, μεταγραφές κειμένου, και άλλες επισημειώσεις. Οι επισημειώσεις των διαφόρων τύπων μπορούν να παραχθούν αυτόματα ή χειροκίνητα ή να προστεθούν από άλλα εργαλεία. Οι μεταγραφές και οι επισημειώσεις μπορούν να αναζητηθούν σε συγκεκριμένο κείμενο ή με κανονικές (κανονιστικές) εκφράσεις (regular expressions). Τα αποτελέσματα της αναζήτησης ή ολόκληρες μεταγραφές μπορεί να ελεχθούν ή να αποθηκευτούν σε μια ποικιλία μορφών και τα σχετικά τμήματα των καταγραφών μπορούν να παίχθούν ή να ανοιχτούν από κάποιο λογισμικό ακουστικής ανάλυσης. Το LaBB-CAT είναι ουσιαστικά ένα είδος αποθετηρίου για τα χρονικά ευθυγραμμισμένα απομαγνητοφωνημένα αποσπάσματα ήχου/βίντεο (αποθετήριο για ευθυγραμμισμένα κείμενα επισημείωσης με τις αντίστοιχες θέσεις στα αρχεία/σήματα βίντεο/ήχου). Οι ευθυγραμμισμένες μεταγραφές μπορεί να έχουν παραχθεί με τη χρήση άλλων εργαλείων (όπως TranscriberAG, Praat). Μέσα από το LaBB-CAT επιτρέπονται πρόσθετες δυνατότητες για την περαιτέρω επεξεργασία και αποθήκευση. Συνοπτικά: το LaBB-CAT επιτρέπει στο χρήστη να αρχειοθετήσει ψηφιακές ηχογραφήσεις και ψηφιακό βίντεο μαζί με μεταγραφές και άλλες επισημειώσεις.

4.4 Toolbox

Το «the Field Linguist's Toolbox» (Buseman) (στο εξής Toolbox) είναι ένα εργαλείο διαχείρισης και ανάλυσης δεδομένων στο γλωσσολογικό τομέα. Είναι ιδιαίτερα χρήσιμο για τη διατήρηση λεξικολογικών δεδομένων, για την μορφολογική ανάλυση κείμενων, αλλά μπορεί να χρησιμοποιηθεί για τη διαχείριση διαφόρων ειδών δεδομένων. Αποτελεί την εξέλιξη και προέκταση μιας βελτιωμένης έκδοσης του Shoebox. Το Toolbox είναι ένα κείμενο-προσανατολισμένο σύστημα διαχείρισης βάσεων δεδομένων με προστιθέμενη λειτουργικότητα μιας και σχεδιάστηκε να καλύψει τις ανάγκες ενός γλωσσολόγου (κυρίως τυπολόγου και μορφολόγου). Το υποκείμενο σύστημα διαχείρισης βάσεων δεδομένων (DBMS) προσφέρει πλήρη ευελιξία στον χρήστη στο σχεδιασμό οποιουδήποτε τύπου βάσης δεδομένων. Προκειμένου να διευκολύνει στην κατανόηση των δυνατοτήτων του και για την διευκόλυνση των χρηστών του περιλαμβάνει προκαθορισμένους ορισμούς δεδομένων για ένα τυπικό λεξικό και σώμα κειμένων. Το σύστημα διαχείρισης βάσεων δεδομένων προσφέρει ισχυρή λειτουργικότητα, όπως προσαρμοσμένη διαλογή, πολλαπλές όψεις της ίδιας βάσης δεδομένων, προεπισκόπηση (preview) των δεδομένων σε μορφή πίνακα, καθώς και φιλτράρισμα για να δείξει υποσύνολα της βάσης δεδομένων. Μπορεί να χειριστεί οποιοδήποτε αριθμό σεναρίων (συνδυαστικού τύπου αναζήτηση) στην ίδια βάση δεδομένων. Κάθε σενάριο έχει τη δική του μορφοποίηση σε unicode κωδικοποίηση. Το Toolbox έχει επίσης σημαντικές γλωσσικές λειτουργίες. Περιλαμβάνει ένα μορφολογικό αναλυτή που μπορεί να χειριστεί σχεδόν όλα τα είδη των μορφοφωνολογικών διεργασιών. Έχει πρότυπο δείγμα λεξημάτων που επιτρέπει στον γλωσσολόγο να περιγράψει όλα τα πιθανά μοτίβα προσφυμάτων που μπορούν να εμφανιστούν. Παρέχει ένα σύστημα διαστίχοισης / διαγραμμάτισης / ενδογράμμισης (Interlinearization) πληροφοριών για (τη βηματική) απόδοση λέξεων (π.χ. βηματική ερμηνεία αρχαίων λέξεων, βηματική ενσωμάτωση δάνειων λέξεων από άλλες γλώσσες). Το Toolbox μπορεί να εξάγει το ενδογραμμισμένο κειμένο σε μορφή κατάλληλη για χρήση στις δημοσιεύσεις, ενώ παράλληλα εξάγει τα δεδομένα σε xml. Συνοπτικά: είναι εργαλείο για λεξικογραφία και μορφοφωνολογική επεξεργασία.

5. Motivation – Data alignment

Η αρχική σκέψη ήταν ότι προηγμένα λογισμικά, όπως για παράδειγμα το Labb-CAT, τα οποία επιτρέπουν στο χρήστη να αρχειοθετήσει ψηφιακές ηχογραφήσεις και ψηφιακό βίντεο μαζί με μεταγραφές και άλλες επισημειώσεις θα μπορούσαν να είναι ικανοποιητικά για την αποθήκευση και επεξεργασία των τεκμηρίων (γραπτών ή προφορικών) του AMiGre. Τελικά αποδείχθηκε ότι δεν μπορούσαν να επιτρέψουν όλες τις βασικές προδιαγραφές απαιτήσεων που είχαν οριστεί στο έργο:

(α) Επισημειώσεις σε πολλά διαφορετικά γλωσσολογικά επίπεδα,
(β) Συνδυασμένη αναζήτηση σε διαφορετικά επίπεδα αναπαράστασης (φωνολογικά, μορφολογικά, μεταδεδομένα και δυνητικά σε συντακτικά και σημασιολογικά),
(γ) Συνδυασμένη αναζήτηση σε αμφότερα τα τεκμήρια γραπτών και προφορικών πηγών.

Αναγκαστικά λοιπόν θα έπρεπε να δημιουργήσουμε (να σχεδιάσουμε και να υλοποιήσουμε) ένα λογισμικό που θα ήταν κομμένο και ραμμένο στις ανάγκες του AMiGre και θα μπορούσε να δεχτεί σαν είσοδο και να συνθέσει σε μία ενότητα όλες τις

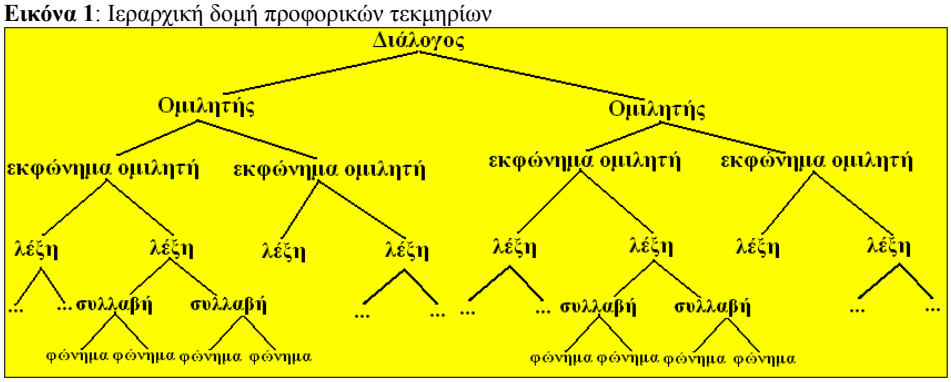
επιμέρους πληροφορίες αυτού (που προέρχονταν από το ELAN ή το Praat ή ήταν αρχεία ψηφιακού ήχου, ή ήταν ψηφιοποιημένα έγγραφα ή ήταν πληροφορίες που βρίσκονταν σε αρχεία λογιστικών φύλλων, ή ήταν μεταγραφές σε αρχεία κειμένου, κλπ).

Με δεδομένο ότι η συλλογή τεκμηρίων του AMiGre βασίζεται σε δύο διαφορετικές ομάδες τεκμηρίων (γραπτών και προφορικών), δημιουργήθηκε η πρόκληση ενοποίησης των δομών τους σε μια κοινή δομή που θα επέτρεπε την κοινή αντιμετώπιση τους. Τα τεκμήρια των γραπτών μας πηγών περιλαμβάνουν βιβλία, άρθρα, δακτυλογραφημένα και χειρόγραφα κείμενα τα οποία απαρτίζονται από σελίδες και οι σελίδες περιέχουν (αποτελούνται από) μορφολογικές λέξεις. Οι επισημειώσεις γίνονται σε όλα τα 3 προαναφερθέντα επίπεδα. Στην εικόνα 1 βλέπουμε τα 3 επίπεδα γραπτών πηγών.

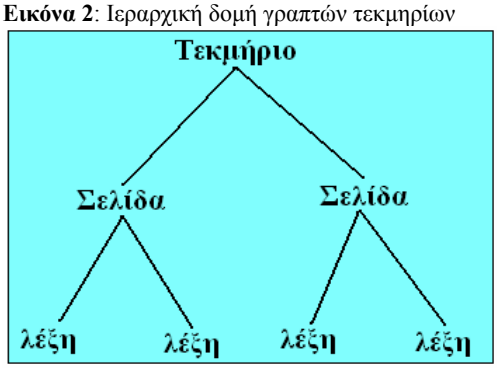
Τα τεκμήρια των προφορικών μας πηγών περιέχουν ηχογραφήσεις ενός ομιλητή ή μιας μικρής ομάδας συνομιλητών. Για κάθε ομιλητή διατίθεται η επισημείωση του λόγου/ομιλίας αυτού, υποδιαιρούμενη στις εναλλαγές (turn takings) που τον αφορούν και αυτές υποδιαιρούμενες σε λέξεις. Μεταξύ των δυνατοτήτων για επιτονικές ή μορφολογικές λέξεις επιλέχθηκαν οι δεύτερες. Αρχικά η κάθε εναλλαγή εκτεινόταν σε όλο το μήκος ενός χωρίς διακοπές τμήματος της ομιλίας του ομιλητή. Τελικά επιλέχθηκε να θεωρήσουμε ως εναλλαγή (turn taking) το κάθε εκφώνημα του ομιλητή (συνήθως την κάθε επιτονική πρόταση). Επιπροσθέτως οι μορφολογικές λέξεις υποδιαιρούνται σε συλλαβές και οι συλλαβές σε φωνήματα (phonemes). Τα φωνήματα μπορεί να πραγματώνονται ως φωνήεντα (vowels) ή σύμφωνα (consonants). Συνεπώς τα τεκμήρια προφορικών πηγών επισημειώνονται σε περισσότερα από τα 3 (των γραπτών πηγών) επίπεδα. Στην εικόνα 2 βλέπουμε τα επίπεδα των προφορικών πηγών.

Για την ενοποίηση των δομών των δύο ομάδων τεκμηρίων προβήκαμε σε αντιστοιχίσεις. Έτσι το αφηρημένο επίπεδο του Document ανταποκρίνεται στον ομιλητή και στο γραπτό τεκμήριο. Το αφηρημένο επίπεδο Part ανταποκρίνεται στο εκφώνημα του ομιλητή και στη σελίδα γραπτού τεκμήριου. Το αφηρημένο επίπεδο Word ανταποκρίνεται στις μορφολογικές λέξεις και στις δύο ομάδες τεκμηρίων. Το ανώτερο αφηρημένο επίπεδο Outer συναντάται μόνο στα προφορικά τεκμήρια και αντιπροσωπεύει το διάλογο (περισσοτέρων του ενός) ομιλητών. Το κατώτερο αφηρημένο επίπεδο Inner το συναντάμε μόνο στα προφορικά τεκμήρια και αντιπροσωπεύει κάθε τι στο οποίο μπορεί να επιμερίζεται μία λέξη (συλλαβές, φωνήματα, κλπ). Το αφηρημένο ιεραρχικό διάγραμμα που τελικά (κατόπιν των αντιστοιχήσεων) υιοθετήσαμε εμφανίζεται στην εικόνα 3.

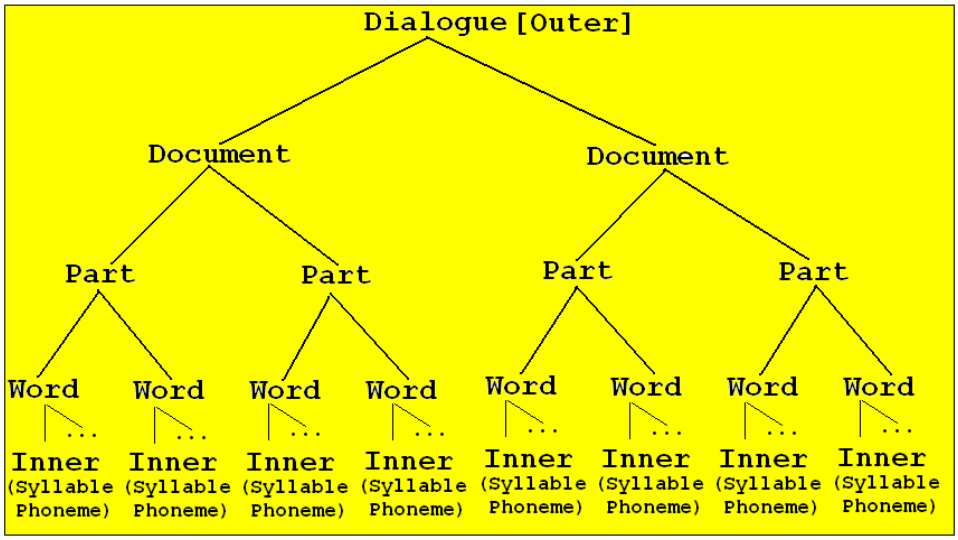
Εικόνα 1: Ιεραρχική δομή προφορικών τεκμηρίων



Εικόνα 2: Ιεραρχική δομή γραπτών τεκμηρίων



Εικόνα 3: Αφηρημένη ιεραρχική δομή για την κάλυψη όλων των τεκμηρίων



6. Overview εφαρμογών και δεδομένων

6.1 Δεδομένα

Τα δεδομένα που διατηρεί το σύστημα για τα προφορικά τεκμήρια είναι ψηφιακές ηχογραφήσεις (συνήθως αρχεία WAV), αρχικές επισημειώσεις (συνήθως αρχεία TextGrid) και υπολογιστικά επεξεργάσιμες επισημειώσεις για όλα τα επίπεδα που υποστηρίζονται. Οι υπολογιστικά επεξεργάσιμες επισημειώσεις είναι ευρύτερες (περισσότερες) από αυτές που υπάρχουν στις αρχικές επισημειώσεις (TextGrid) καθώς μπορούν να συμπληρωθούν στο σύστημα μετά την ένταξη (import) της αρχικής επισημείωσης (TextGrid). Αναλυτικά οι υπολογιστικά επεξεργάσιμες επισημειώσεις είναι: μεταδεδομένα ομιλητή – επίπεδο Document, επισημειώσεις εκφωνημάτων (ορθογραφική μεταγραφή, μετάφραση στην KNE) – επίπεδο Part, επισημειώσεις μορφολογικών λέξεων (ορθογραφική μεταγραφή, Greek Sampa (Greek Sampa) και φωνολογικά φαινόμενα της λέξης) – επίπεδο Word και επισημειώσεις για κάθε τι στο

οποίο μπορεί να επιμερίζεται μία λέξη (συλλαβές σε Greek Samba, θέση και τόνος φωνηέντων, σύμφωνα) – επίπεδο Inner. Δυνητικά (καθώς υποστηρίζονται από τις δομές αποθήκευσης και τα modules του ανεπτυγμένου λογισμικού) μπορούν να υπάρχουν και μορφολογικές επισημειώσεις σε επίπεδο λέξης και συντακτικές και σημασιολογικές επισημειώσεις σε επίπεδο ακολουθίας λέξεων (φράσης, πρότασης, κλπ).

Τα δεδομένα που διατηρεί το σύστημα για τα γραπτά τεκμήρια είναι ψηφιοποιημένες σελίδες από τα πρωτότυπα (συνήθως αρχεία JPG), μεταγραφές (transcriptions) που ομογενοποιούν τα σύμβολα από τα πρωτότυπα τεκμήρια (αρχεία κειμένου) και υπολογιστικά επεξεργάσιμες επισημειώσεις για όλα τα επίπεδα που υποστηρίζονται. Στην περίπτωση των γραπτών κειμένων υπάρχει μεγάλος όγκος μορφολογικών επισημειώσεων. Οι υπολογιστικά επεξεργάσιμες επισημειώσεις είναι ευρύτερες (περισσότερες) από αυτές που υπάρχουν στις αρχικές επισημειώσεις (συνήθως excel αρχεία) καθώς μπορούν να συμπληρωθούν στο σύστημα μετά την ένταξη (import) της αρχικής επισημείωσης (από τα excel αρχεία). Αναλυτικά οι υπολογιστικά επεξεργάσιμες επισημειώσεις είναι: μεταδεδομένα τεκμηρίου – επίπεδο Document, επισημειώσεις σελίδων (ελάχιστες ή καθόλου) – επίπεδο Part και μορφολογικές επισημειώσεις λέξεων – επίπεδο Word. Ενδεικτικά αναφέρουμε: γραμματική κατηγορία, αν η λέξη είναι δάνεια και ποια η προέλευση της, αν είναι αρχαϊσμός, αν παρατηρείται διαφοροποίηση γένους, αν είναι απλή ή πολύπλοκη και ποιες είναι οι μορφολογικές διαδικασίες δημιουργίας της (παραγωγή, σύνθεση, συμφυρμός). Δυνητικά μπορούν να υπάρχουν και συντακτικές και σημασιολογικές επισημειώσεις σε επίπεδο ακολουθίας λέξεων (φράσης, πρότασης, κλπ).

Με βάση τα παραπάνω είναι σχεδόν αυτονόητη η ανάγκη για τέσσερις συλλογές αρχείων. Αυτές είναι ψηφιακές ηχογραφήσεις προφορικών (WAV files), αρχικές επισημειώσεις προφορικών (TextGrid files), ψηφιοποιημένες σελίδες από τα πρωτότυπα γραπτά (Image files), μεταγραφές σελίδων των γραπτών (Transcripted Written Sources). Τα δεδομένα αυτά απεικονίζονται (μεταξύ άλλων) στην εικόνα 4. Δεν κρίθηκε σκόπιμο να διατηρηθούν τα αρχικά excel αρχεία μορφολογικής επισημείωσης (σε συλλογή) καθώς αφομοιώνονται πλήρως στην EAV database (που θα εξετάσουμε παρακάτω).

Τα δεδομένα που διατηρεί το σύστημα συμπληρώνονται από βάσεις δεδομένων για τα υπολογιστικά επεξεργάσιμα στοιχεία. Τρεις είναι οι βάσεις δεδομένων (database subschemas) που χρησιμοποιεί το σύστημα. Η πρώτη είναι η Struct Database. Πρόκειται για ένα σύνολο πινάκων που υλοποιούν την αφηρημένη ιεραρχική δομή για την κάλυψη όλων των τεκμηρίων. Οι πληροφορίες που φυλάγονται σε αυτή προσδιορίζουν στοιχεία όπως ποιο είναι το document (ομιλητής / γραπτό τεκμήριο), από ποια parts (εκφωνήματα / σελίδες) αποτελείται, ποιες words (μορφολογικές λέξεις) συνθέτουν το part. Σε αυτή τη βάση δεδομένων (Struct Database) ομαδοποιούνται και τα documents (ομιλητές) που συνθέτουν ένα προφορικό τεκμήριο (ηχογραφημένο διάλογο).

Η δεύτερη βάση που απαιτούνταν για το σύστημά μας είναι η βάση που θα χρησιμοποιούσε για κάθε λογής επισημειώσεις. Μια κλασική σχεσιακή υλοποίηση θα απαιτούσε μεγάλη σχεδιαστική προσπάθεια (για να προβλεφθούν όλες οι οντότητες και όλες οι ιδιότητές τους) και τελικά το σύστημα θα υπέφερε από διαρκείς βελτιώσεις έτσι ώστε να ενσωματώνονται καινούργιες οντότητες (entities) και καινούργιες ιδιότητες (attributes). Το παραπάνω συμπέρασμα προέκυψε από τις αρχικές συζητήσεις μεταξύ των ομάδων εργασίες καθώς παρατηρήθηκε αδυναμία να προσδιοριστούν εξ αρχής οι οντότητες και οι ιδιότητές τους. Πρόκειται για το γνωστό πρόβλημα Schema evolution.

Επιπλέον μια σχεσιακή υλοποίηση θα χαρακτηρίζονταν από αραιά δεδομένα, καθώς είναι συχνό σε αυτό τον τομέα (γλωσσολογικό) να μην διαθέτουν πάντα οι ιδιότητες τιμές. Ένα άλλο θέμα που προέκυψε σύντομα ήταν ότι οι τιμές των ιδιοτήτων μπορούσαν να λαμβάνουν τιμές από καθορισμένα σύνολα τιμών (λεξιλόγια) και σε άλλες περιπτώσεις να λαμβάνουν τιμές ελεύθερα (χωρίς λεξιλόγια). Ένα ακόμα χαρακτηριστικό ήταν ότι μπορούσαν να υπάρχουν ιδιότητες που δέχονται (λαμβάνουν) πολλαπλές τιμές. Τέλος παρατηρήθηκαν εξαρτήσεις εμφάνισης (ύπαρξης) ιδιοτήτων από τις τιμές άλλων ιδιοτήτων. Όλες αυτές οι παρατηρήσεις έκαναν αναγκαία την εισαγωγή ενός Entity-Attribute-Value (EAV) σχήματος (Anhøj 2003) που αντιμετωπίζει τα προβλήματα του Schema evolution και των αραιών δεδομένων. Σε αυτό έγιναν παρεμβάσεις (επεκτάσεις) ώστε να υποστηρίζει ελεύθερα και καθορισμένα σύνολα τιμών (λεξιλόγια), πολλαπλές τιμές ιδιοτήτων και εξαρτήσεις εμφάνισης ιδιοτήτων (dependencies). Τη βάση αυτή (των επισημειώσεων) την ονομάζουμε EAV database.

Η τρίτη βάση του συστήματός μας απαιτούταν για τις επισημειώσεις που γίνονται σε υποδιαιρέσεις των λέξεων (συλλαβές, φωνήματα, κλπ). Η υλοποίηση της βάσης για αυτές τις επισημειώσεις μπορεί να γίνει με δύο τρόπους: να ενταχθεί στην παραπάνω EAV database και να υλοποιηθεί ανεξάρτητα. Ανεξάρτητα από το ποια υλοποίηση χρησιμοποιήθηκε, εννοιολογικά θα την εξετάζουμε σαν ανεξάρτητη database για να της δώσουμε ξεχωριστή υπόσταση και για να είναι συμβατή (η εννοιολογική αναπαράσταση του συστήματος) με την εμπειρία που έχει ο χρήστης στο σύστημα.

6.2 Εφαρμογές

Για τη διαχείριση των τεκμηρίων απαιτείται μια κεντρική σελίδα για κάθε τεκμήριο. Για το σκοπό αυτό αποφασίσθηκε να δημιουργηθούν δύο modules (φόρμες / υποπρογράμματα), ένα για γραπτά τεκμήρια και ένα για προφορικά τεκμήρια. Αυτά τα modules ονομάσθηκαν **G.Written** και **G.Oral**, αντίστοιχα. Το πρόθεμα G. ανταποκρίνεται στα ακρωνύμιο GUI (για το Graphical User Interface) αλλά μπορεί και να εκληφθεί ως Glue (κόλλα, καθώς λειτουργεί σαν κόλλα που συνενώνει τα επιμέρους modules και λειτουργίες που μπορούν να γίνουν σε ένα τεκμήριο).

Αμφότερα σχεδιάσθηκαν και υλοποιήθηκαν ως ένα τρίπτυχο που αριστερά παρουσιάζει το Part (εκφώνημα προφορικού τεκμηρίου / σελίδα γραπτού τεκμηρίου), στη μέση παρουσιάζει τα Words (μορφολογικές λέξεις του Part του προφορικού ή γραπτού τεκμηρίου) και δεξιά επιτρέπει να εμφανισθούν (και να ενημερωθούν) όλες οι επισημειώσεις μίας (κάθε φορά) λέξης (ή ακολουθίας λέξεων ή συστατικού λέξης). Το τρίπτυχο σχεδιάστηκε και υλοποιήθηκε έτσι ώστε να παρέχει χειριστήρια για τη μετακίνηση (navigation) μεταξύ των Parts (εκφωνημάτων προφορικού τεκμηρίου / σελίδων γραπτού τεκμηρίου). Προφανώς, παρέχουν εξειδικευμένες ενέργειες που εξαρτώνται από τη μορφή του τεκμηρίου (για παράδειγμα το G.Oral παρέχει χειριστήριο για εναλλαγή μεταξύ ορθογραφικής μεταγραφής και Greek Samba μεταγραφής των λέξεων στο μεσαίο πάνελ και το G.Written παρέχει χειριστήριο για προσθήκη σελίδας γραπτού τεκμηρίου).

Τα επιμέρους modules και λειτουργίες που μπορούν να γίνουν σε ένα τεκμήριο παρουσιάζονται ακολούθως:

Ο φωνολογικός επισημειωτής (**Ph. Tagging**) χρησιμοποιείται για να αποδώσει (συσχετίσει) φαινόμενα (π.χ. τσιτακισμός) που παρατηρούνται στη λέξη. Έχει εφαρμογή σε αμφότερα τα γραπτά και τα προφορικά τεκμήρια.

Ο μορφολογικός επισημειωτής (**Morph. Tag**) χρησιμοποιείται για να αποδώσει (συσχετίσει) φαινόμενα και ιδιότητες (π.χ. γραμματική κατηγορία ή αν η λέξη είναι δάνεια και ποια η προέλευση της) που παρατηρούνται στη λέξη. Έχει εφαρμογή σε αμφότερα τα γραπτά και τα προφορικά τεκμήρια.

Ο συντακτικός επισημειωτής (**Syn. Tag**) χρησιμοποιείται για να αποδώσει συντακτικές κατηγορίες (ρηματική φράση – VP, ονοματική φράση – NP, προθεματική φράση – PP, πρόταση – S, κλπ) σε μία ακολουθία από λέξεις.

Ο σημασιολογικός επισημειωτής (**Sem. Tagger**) χρησιμοποιείται για να αποδώσει σημασιολογικές πληροφορίες σε μία ακολουθία από λέξεις.

Για τη διαχείριση των μεταδεδομένων που στα γραπτά αφορούν το τεκμήριο και στα προφορικά αφορούν τον ομιλητή (ηλικία, φύλο, καταγωγή, κλπ) απαιτούνται δύο modules. Αυτά τα modules ονομάσθηκαν **Meta Written** και **Meta Oral**, αντίστοιχα.

Για την εισαγωγή νέων τεκμηρίων απαιτούνται αντίστοιχα modules. Έτσι σχεδιάστηκαν και αναπτύχθηκαν τα modules **P.I. Written** και **I. Oral**. Για τα γραπτά η εισαγωγή γίνεται βηματικά (σελίδα – σελίδα) και για αυτό το λόγο, το module έλαβε όνομα Partial (ή Page) Import Written (εν συντομία P.I. Written). Για τα προφορικά η εισαγωγή γίνεται συνολικά και για αυτό το λόγο, το module έλαβε όνομα Import Oral (εν συντομία I. Oral). Στην τελευταία περίπτωση ο χρήστης προσδιορίζει (μέσω διεπαφής) όλα τα TextGrids των ομιλητών και την (ή τις) ψηφιακές ηχογραφήσεις και συντελείται το κατάλληλο parsing που εντάσσει το προφορικό τεκμήριο στο σύστημα. Στα γραπτά κείμενα το βασικό module (P.I. Written) αξιοποιεί τις δυνατότητες δύο άλλων modules (**T. Imaging** – Text Imaging και **T. Transcription** – Text Transcription) για να ενσωματώσει την ψηφιοποιημένη σελίδα και να εισάγει την ομογενοποιημένη μεταγραφή της σελίδας.

Για την περιήγηση και την εύρεση των τεκμηρίων που απαιτούνται κάθε φορά υπάρχουν άλλα τρία modules. Αυτά είναι **B. Oral** (Browse Oral) για περιήγηση στα προφορικά τεκμήρια, **B. Written** (Browse Written) για περιήγηση στα γραπτά τεκμήρια και **Search & Retrieve** για εύρεση και εμφάνιση τεκμηρίων. Το τελευταίο θα αναπτυχθεί ιδιαιτέρως σε επόμενη ενότητα του άρθρου.

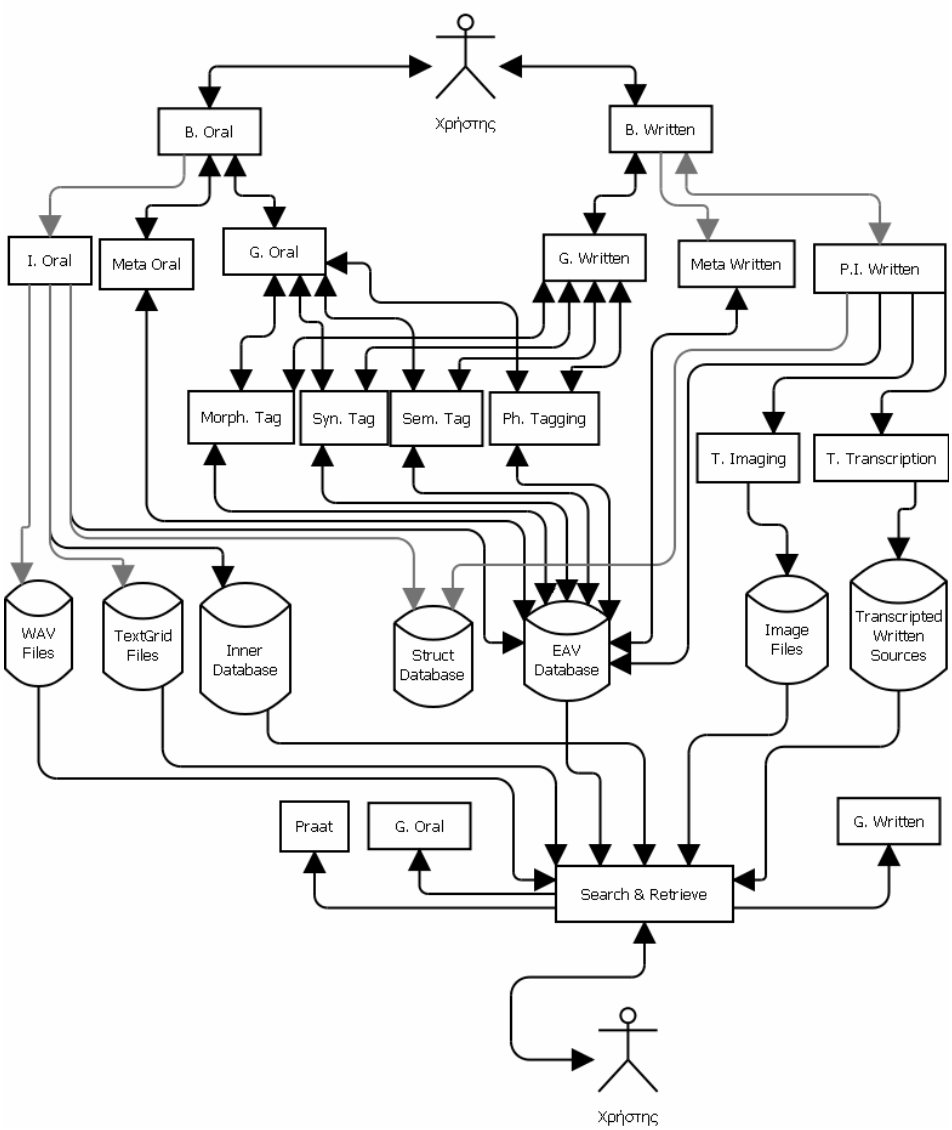
Υπάρχουν άλλα δύο modules για την μαζική εισαγωγή γραπτών και προφορικών πηγών. Αυτά ονομάζονται **M.I. Written** (Massive Import Written) και **M.I. Oral** (Massive Import Oral). Δεν αναπτύσσονται σε βάθος, καθώς χρησιμοποιήθηκαν ως υπηρεσία και δεν έχουν την ωριμότητα που απαιτείται για να δοθούν σε τελικούς χρήστες.

Ο πίνακας 1 περιέχει μια συνοπτική παρουσίαση των ονομάτων των modules και των δεδομένων που εξετάσαμε παραπάνω. Ο πίνακας βοηθάει στην κατανόηση του αρχιτεκτονικού διαγράμματος του συστήματος (εικόνα 4).

7. Θεμελιώδη στοιχεία για τη ΒΔ

Τα δεδομένα που διατηρεί το σύστημα συμπληρώνονται από βάσεις δεδομένων για τα υπολογιστικά επεξεργάσιμα στοιχεία. Στη συνέχεια εξετάζουμε τις τρεις αυτές βάσεις δεδομένων (database subschemas) που χρησιμοποιεί το σύστημά μας.

Εικόνα 4: Εφαρμογές και δεδομένα



7.1 Struct

Η Struct Database αποτελείται από ένα σύνολο πινάκων που υλοποιούν την αφηρημένη ιεραρχική δομή που παρουσιάστηκε παραπάνω (εικόνα 3) για την κάλυψη όλων των τεκμηρίων. Οι πληροφορίες που φυλάγονται σε αυτή προσδιορίζουν στοιχεία για το document, τις υποδιαιρέσεις του document (τα parts) και τις υποδιαιρέσεις των υποδιαιρέσεων των documents (words). Μπορούμε να πούμε ότι η Struct Database οργανώνει σε διαδοχικά επίπεδα εκλέπτυνσης τα συνθετικά των τεκμηρίων

προκειμένου αυτά να αποκτήσουν ταυτότητα (identifier – ID) και να επισημειωθούν με τη βοήθεια της EAV Database (που εξετάζεται αργότερα).

Η υλοποίηση της αφηρημένης μορφής της Struct Database (εικόνα 3) έγινε με δύο σχεσιακά σχήματα τα οποία μοιάζουν πολύ μεταξύ τους. Η διαφοροποίηση τους είναι ότι η υλοποίηση για τα Oral – προφορικά – τεκμήρια αποτελείται από όλα (τα 5) επίπεδα της αφηρημένης μορφής, ενώ η υλοποίηση για τα Written – γραπτά τεκμήρια – περιέχει μόνο τα 3 ενδιάμεσα επίπεδα της αφηρημένης μορφής (καθώς τα άλλα 2 περιφερειακά επίπεδα (Dialogue και Inner) δεν έχουν εφαρμογή στα γραπτά τεκμήρια).

Οι εικόνα 5 παρουσιάζει τη δομή υλοποίησης της Struct Database για τα προφορικά τεκμήρια (εξ ου και το πρόθεμα oral_ στους database tables) και η εικόνα 6 παρουσιάζει τη δομή υλοποίησης της Struct Database για τα γραπτά τεκμήρια (εξ ου και το πρόθεμα wrt_ στους database tables). Για την καλύτερη κατανόηση των δύο δομών υλοποίησης πρέπει κανείς να θυμηθεί τις αντιστοιχήσεις (data alignment) που αναφέρθηκαν σε προηγούμενη ενότητα. Εδώ τις υπενθυμίζουμε με τον πίνακα 2 στον οποίο υπάρχει και το αφηρημένο ισοδύναμο (πρώτη στήλη του πίνακα 2). Στα διαγράμματα (εικόνες 5 και 6) δίπλα από τους database tables υπάρχει το αφηρημένο όνομα που παραπέμπει στην πρώτη στήλη του πίνακα 2.

Πίνακας 1: Ορολογία για την αρχιτεκτονική του συστήματος

Όνομα	**Ερμηνεία**
B. Oral	Browse Oral
B. Written	Browse Written
I. Oral	Import Oral
Meta Oral	Metadata for Oral
G. Oral	τρίπτυχο προφορικών
G. Written	τρίπτυχο γραπτών
Meta Written	Metadata for Written
P.I. Written	Page (ή Part) Import Written
Morph. Tag	Morphological Tagging
Syn. Tag	Syntactic Tagging
Sem. Tag	Semantic Tagging
Ph. Tagger	Phonological Tagging
T. Imaging	Text Imaging
T. Transcription	Text Transcription
M.I. Oral	Massive Import Oral
M.I. Written	Massive Import Written
Struct Database	Dialogue, Document, Part, Word
Inner Database	συστατικά των λέξεων (συλλαβές, φωνήματα, κλπ) για τα προφορικά τεκμήρια (έχει υλοποιηθεί με EAV)
άλλο	τα υπόλοιπα λεκτικά στα διαγράμματα δεν απαιτούν ερμηνεία

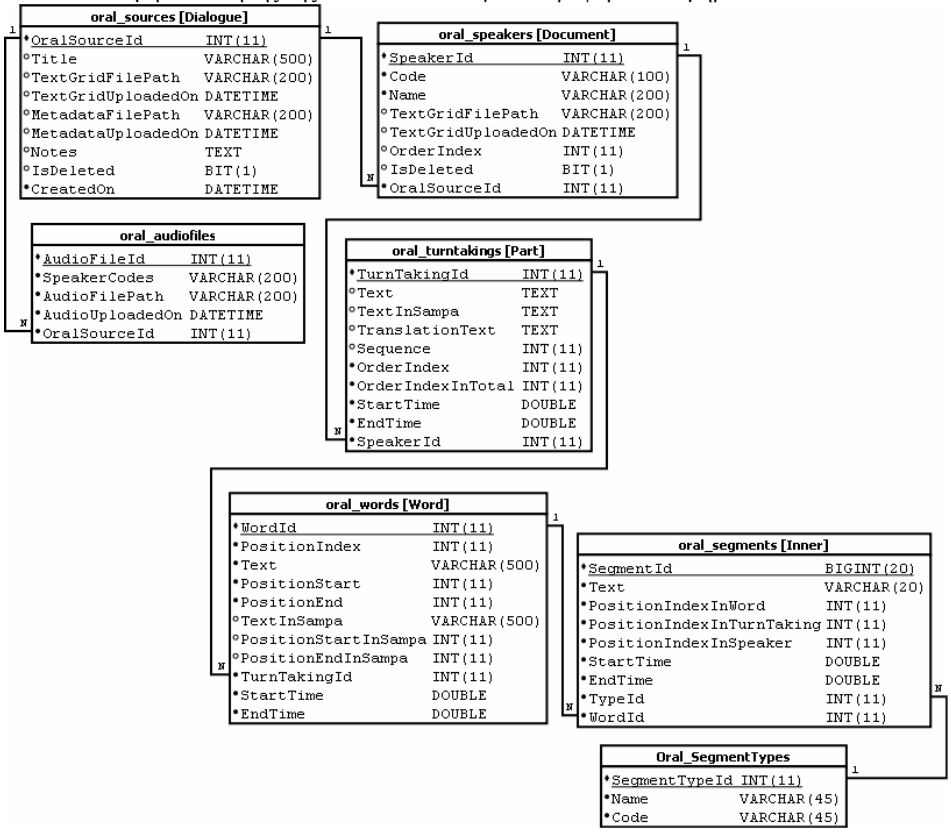
Σε αμφότερα τα διαγράμματα βλέπουμε συσχετίσεις ένα προς πολλά για κάθε ζεύγος ενός επιπέδου με τα συστατικά του αμέσως επόμενου επιπέδου. Για παράδειγμα ένας Διάλογος έχει πολλούς Ομιλητές (στην εικόνα 5, η σχέση του oral_sources με τον oral_speakers είναι 1:n) και ένα γραπτό τεκμήριο έχει πολλές σελίδες (στην εικόνα 6, η σχέση του wrt_documents με τον wrt_pages είναι 1:n). Στα διαγράμματα υπάρχουν και ορισμένοι βοηθητικοί database tables. Στο διάγραμμα της εικόνας 5 εμφανίζονται και οι πίνακες oral_audiofiles και oral_SegmentTypes. Ο πίνακας (database table) oral_audiofiles είναι για την αποθήκευση του ενός (αλλά μπορεί να είναι και

περισσότερα του ενός) ψηφιακού ηχητικού αρχείου που αντιστοιχεί σε ένα προφορικό τεκμήριο. Ο πίνακας (database table) oral_SegmentTypes περιέχει τα τμήματα στα οποία μπορεί να υποδιαιρείται μία μορφολογική λέξη (μέχρι στιγμής αυτά είναι συλλαβές, φωνήεντα, σύμφωνα αλλά μπορούν να συμπληρωθούν και άλλα). Επίσης στο διάγραμμα της εικόνας 6 υπάρχει ο πίνακας wrt_wordextractconfigurations. Σε αυτόν φυλάγονται πληροφορίες που καθορίζουν πως έγινε (με ποιες ρυθμίσεις) η διάσπαση της μεταγραφής μιας σελίδας ενός γραπτού τεκμηρίου σε λέξεις (tokenization configuration). Δηλαδή στον πίνακα αυτό φυλάγονται οι διαχωριστικοί χαρακτήρες που χρησιμοποιήθηκαν, η κανονική έκφραση (regular expression) και οι θέσεις της μεταγραφής στις οποίες δεν ελήφθη υπόψη (εξαιρέθηκε) η σημασία των διαχωριστικών χαρακτήρων.

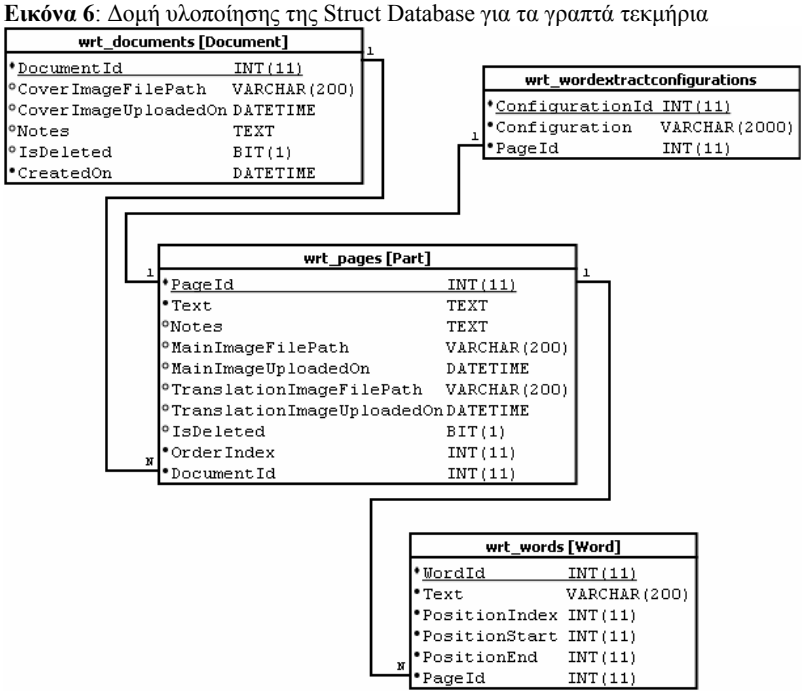
Πίνακας 2: αντιστοιχήσεις (data alignment) γραπτών και προφορικών τεκμηρίων

Αφηρημένο	**Προφορικά**	**Γραπτά**
Dialogue	Διάλογος – συνολικό προφορικό τεκμήριο	--
Document	Ένας ομιλητής / συνομιλητής	συνολικό γραπτό τεκμήριο
Part	εκφώνημα του ομιλητή	σελίδα γραπτού τεκμήριου
Wort	μορφολογικές λέξεις	μορφολογικές λέξεις
Inner	Συλλαβές, φωνήεντα, σύμφωνα, κλπ	--

Εικόνα 5: Δομή υλοποίησης της Struct Database για τα προφορικά τεκμήρια



Εικόνα 6: Δομή υλοποίησης της Struct Database για τα γραπτά τεκμήρια



Περιορισμοί στην έκταση του άρθρου δεν μας επέτρεψαν να παρέχουμε ένα λεξικό δεδομένων που θα διευκόλυνε στην καλύτερη κατανόηση των πεδίων της struct database (π.χ. των πεδίων Text, TextInSampa, TranslationText του oral_turntakings).

7.2 EAV

Η δεύτερη βάση που χρησιμοποιεί το σύστημα μας είναι η βάση που καταγράφονται κάθε λογής επισημειώσεις. Όπως προαναφέρθηκε αυτή βασίζεται στην Entity-Attribute-Value αναπαράσταση για να μας απαλλάξει από το πρόβλημα του διαρκούς Schema Evolution και να επιλύσει το πρόβλημα της εκτεταμένη χρήσης null values. Οι επισημειώσεις που καταχωρούνται σε αυτό αφορούν οντότητες (πλειάδες) των database tables oral_speakers, oral_turntakings, oral_words, oral_segments, wrt_documents, wrt_pages και wrt_words. Με άλλα λόγια το entity (Ε του EAV) λαμβάνει τιμές από τα πρωτεύοντα κλειδιά (primary keys) των επτά (7) πινάκων που αναφέρθηκαν.

Σε προηγούμενη ενότητα αναφέραμε πως έγιναν παρεμβάσεις (επεκτάσεις) προκειμένου το υποσχήμα (database subschema) να υποστηρίζει ελεύθερα και καθορισμένα σύνολα τιμών (λεξιλόγια), πολλαπλές τιμές ιδιοτήτων και εξαρτήσεις εμφάνισης ιδιοτήτων (dependencies). Μια άλλη σημαντική παρατήρηση (που δεν αναφέρθηκε προηγούμενα) είναι ότι όλες οι επισημειώσεις παρουσιάζουν/απαιτούν παρόμοια λειτουργικότητα. Έτσι τα εννέα modules επισημείωσης (μορφολογικής, φωνολογικής, συντακτικής και σημασιολογικής, καθένα για δύο τύπους πηγών και επιπλέον επισημείωση συστατικών λέξης για τα προφορικά) θα μπορούσαν να γίνουν ένα module που θα ενημέρωνε κάθε φορά διαφορετικό σύνολο ιδιοτήτων. Επιπλέον οι μεταπληροφορίες θα μπορούσαν να διαχειριστούν με το ίδιο module καθώς απαιτούν

την ίδια λειτουργικότητα αλλά ενημερώνουν επίσης διαφορετικά σύνολα ιδιοτήτων. Έτσι τα 11 modules (9 επισημειώσεων και 2 μεταπληροφορίας) θα μπορούσαν να γίνουν ένα module που προσαρμόζεται με βάση το σύνολο ιδιοτήτων που διαχειρίζεται, αρκεί να μπορούσαν να ορισθούν αυτά τα 11 σύνολα ιδιοτήτων.

Όλες οι παραπάνω απαιτήσεις οδήγησαν στην σχεδίαση ενός επεκτεταμένου EAV σχήματος που υποστηρίζει όλες τις απαιτήσεις. Το σχήμα εμφανίζεται στην εικόνα 7. Διακρίνουμε 9 πίνακες τους οποίους θα εξηγήσουμε στη συνέχεια.

Ο πίνακας app_sourcetypes καθορίζει (περιέχει) του δύο τύπους τεκμηρίων (γραπτά και προφορικά). Ο πίνακας app_modules καθορίζει (περιέχει) τα 11 modules που χρειάζονται (5 για γραπτά και 6 για προφορικά). Ο πίνακας propertygroups χρησιμοποιείται για δύο λόγους: (α) για τον ορισμό θεματικών υποσυνόλων ιδιοτήτων (τέτοια έχουμε στη μορφολογική επισημείωση γραπτών πηγών και στα μεταδεδομένα προφορικών πηγών) και (β) για τον ορισμό προκαθορισμένων συνόλων τιμών των ιδιοτήτων (lookups στην ορολογία των πληροφορικών, λεξιλόγια στην ορολογία των γλωσσολόγων). Ο πίνακας propertygrouptypes είναι βοηθητικός και καθορίζει (περιέχει) του δύο σκοπούς για τους οποίους χρησιμοποιείται ο propertygroups. Ο πίνακας properties χρησιμοποιείται για να καθορίζει (περιέχει) όλες τις ιδιότητες που αξιοποιούνται από τα 11 modules. Ο πίνακας propertytypes είναι βοηθητικός και περιέχει του τέσσερις τύπους που μπορεί να έχει μια ιδιότητα (αλφαριθμητικό, αλφαριθμητικό με πολλαπλές τιμές, προκαθορισμένης τιμής από lookup/λεξιλόγιο και πολλαπλής προκαθορισμένης τιμής από lookup/λεξιλόγιο). Είδαμε ότι η δεύτερη χρήση του πίνακα propertygroups είναι για τον ορισμό προκαθορισμένων συνόλων τιμών ιδιοτήτων, δηλαδή να ορίζει ονομασίες ταυτοποίησης των lookups/λεξιλογίων. Όμως ένα lookup/λεξιλόγιο πρέπει, εκτός από ονομασία, να έχει και το σύνολο (μια λίστα) με τις αποδεκτές τιμές. Αυτή τη λίστα τιμών του κάθε lookup/λεξιλογίου ορίζουμε στον πίνακα propertyvalues.

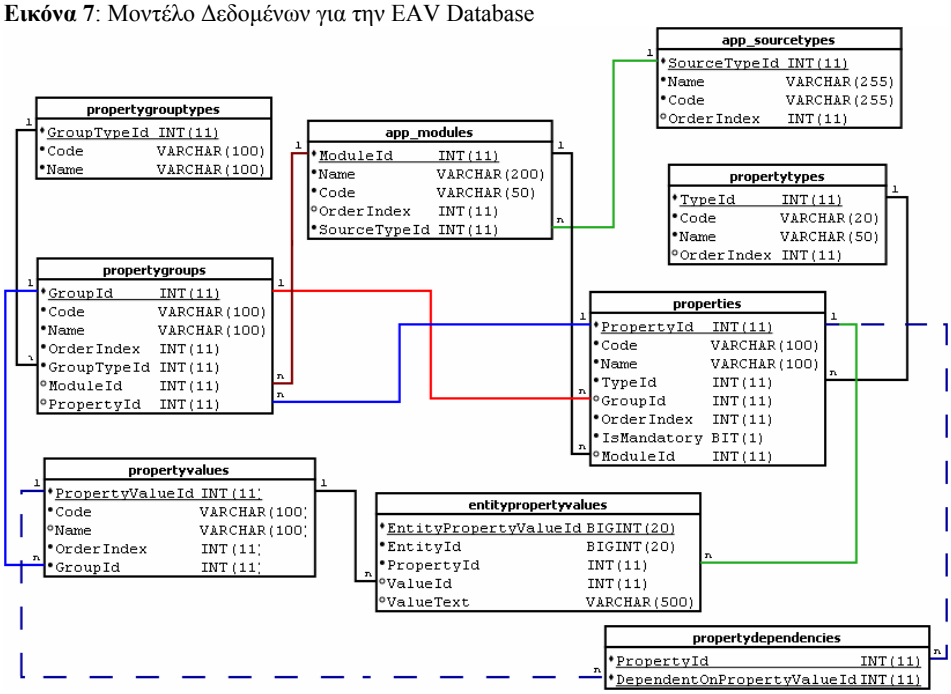
Ο πίνακας entitypropertyvalues είναι ο κεντρικός της EAV database. Εδώ καταχωρούνται τα Entities (στο πεδίο EntityId), τα Attributes (στο πεδίο PropertyId) και τα Values (στο πεδίο ValueId ή στο πεδίο ValueText). Όπως προαναφέραμε το πεδίο τιμών του EntityId είναι τα πρωτεύοντα κλειδιά (primaty keys) επτά (7) βασικών πινάκων του Struct Database (oral_speakers, oral_turntakings, oral_words, oral_segments, wrt_documents, wrt_pages και wrt_words). Το πεδίο τιμών του PropertyId είναι το πρωτεύον κλειδί του πίνακα properties. Το πεδίο τιμών του ValueId είναι το πρωτεύον κλειδί του πίνακα propertyvalues. Στην περίπτωση που το Attribute (αυτό που προσδιορίζει το πεδίο PropertyId) έχει τύπο αλφαριθμητικού ή αλφαριθμητικού με πολλαπλές τιμές, τότε συμπληρώνεται το ValueText και όχι το ValueId.

Στον πίνακα propertydependencies προσδιορίζονται οι ιδιότητες που εμφανίζονται υπό τον περιορισμό ότι μία άλλη ιδιότητα έχει λάβει κάποια συγκεκριμένη τιμή. Αν μία ιδιότητα δεν περιέχεται στον πίνακα propertydependencies τότε είναι ελεύθερη από εξαρτήσεις και θα εμφανίζεται πάντα στο module (ή στο θεματικό υποσύνολο ιδιοτήτων του module) που εντάσσεται.

Στο διάγραμμα της εικόνας 7 εμφανίζονται και ορισμένοι περιορισμοί (constraints) που χαρακτηρίζουν το μοντέλο. Αυτοί οι περιορισμοί είναι κυρίως Foreign Key constraints. Τα πρωτεύοντα κλειδιά εμφανίζονται υπογραμμισμένα και επιπλέον (αριστερά της ονομασίας τους) έχουν ένα ρόμβο. Τα πεδία που είναι not null έχουν το σύμβολο του σφαιριδίου (solid bullet). Τα πεδία που αποδέχονται null τιμές έχουν το σύμβολο του κουλουριού (circle). Οι περιορισμοί αυτοί δεν επαρκούν για την πλήρη

λειτουργικότητα του μοντέλου. Μετά το διάγραμμα, δίδονται (με φραστικό τρόπο) επιπλέον περιορισμοί που μπορεί να εκφράζονται εντός της database (συνήθως με triggers) ή εκτός αυτής (συνήθως στο business logic του λογισμικού).

Εικόνα 7: Μοντέλο Δεδομένων για την EAV Database



Οι επόμενες παράγραφοι της παρούσας υπό-ενότητας (7.2) αποτελούν επιπλέον λογικούς περιορισμούς και μπορούν να παραλειφθούν σε πρώτη ανάγνωση.

Το properties.GroupId μπορεί να είναι null ή να δείχνει κάποιο GroupId του propertygroups. Δηλαδή τα properties που δεν εντάσσονται σε ομάδες και εξαρτώνται απευθείας από modules δεν αντιστοιχούν σε groups (γραμμές) του propertygroups.

Στον propertygroups έχουν οριστεί (ως τώρα) ομάδες μόνο για τη μορφολογική επισημείωση γραπτών και τα metadata προφορικών.

Το propertygroups.ModuleId έχει νόημα μόνο για ιδιότητες. Το propertygroups.PropertyId έχει νόημα μόνο για σύνολα τιμών ιδιότητας. Δηλαδή, μόνο το ένα ή το άλλο (από τα ModuleId και PropertyId) θα έχει τιμή σε μία γραμμή του propertygroups.

Τα πεδία properties.ModuleId και properties.GroupId είναι αμοιβαία αποκλειόμενα. Δηλαδή μια ιδιότητα (μια γραμμή του πίνακα properties) δείχνει είτε στο module που ανήκει είτε στην ομάδα ιδιοτήτων στην οποία ανήκει και ποτέ και στα δύο.

Στον propertydependencies κάθε γραμμή προσδιορίζει ότι η ιδιότητα (στην οποία δείχνει το πεδίο) PropertyId θα ενεργοποιείται μόνο αν η DependentOnPropertyValueId αποτελεί τιμή μια άλλης (κατάλληλης για την τιμή DependentOnPropertyValueId) ιδιότητας. Αν μία ιδιότητα (η ταυτότητα της ιδιότητας) δεν περιέχεται στον πίνακα propertydependencies τότε είναι ελεύθερη από εξαρτήσεις και θα εμφανίζεται πάντα στο module (ή στην ομάδα ιδιοτήτων) που εντάσσεται.

Το πεδίο propertygroups.OrderIndex έχει τιμές για τη διάταξη των ομάδων ιδιοτήτων που ανήκουν στο ίδιο module (έχουν την ίδια τιμή propertygroups.ModuleId και αυτό δεν είναι null).

Το πεδίο propertyvalues.OrderIndex έχει τιμές για τη διάταξη των αποδεκτών τιμών της φιλόξενης ιδιότητας (propertyvalues.PropertyId).

Το πεδίο properties.OrderIndex (α) έχει τιμές για τη διάταξη των ιδιοτήτων που ανήκουν στην ίδια ομάδα (η ομάδα προσδιορίζεται στο properties.GroupId και δεν είναι null) ή (β) έχει τιμές για τη διάταξη των ιδιοτήτων που ανήκουν στο ίδιο module (το module προσδιορίζεται στο properties.ModuleId και το properties.GroupId είναι null).

Το entitypropertyvalues.EntityId πρέπει να ελέγχετε ότι κάνει αναφορά στο PK ενός εκ των 7 βασικών πινάκων του struct subschema. Αυτό είναι στο business logic της εφαρμογής (εναλλακτικά θα μπορούσε να ελεγχθεί με triggers).

Το entitypropertyvalues.ValueId και το entitypropertyvalues.ValueText είναι αμοιβαία αποκλειόμενα στην ίδια εγγραφή. Μόνο ένα από τα δύο έχει κάθε φορά τιμή (το άλλο είναι null).

Το entitypropertyvalues.PropertyId θα μπορούσε να είναι περιττό καθώς προκύπτει από το ValueId. Όμως επειδή κάποιες εγγραφές έχουν entitypropertyvalues.ValueText (και δεν έχουν entitypropertyvalues.ValueId) το πεδίο PropertyId είναι αναγκαίο στον πίνακα entitypropertyvalues.

Ο πίνακας entitypropertyvalues λειτουργεί ως bag για τις πολλαπλές τιμές που μπορεί να έχει ο ίδιος συνδυασμός τιμών των EntityId και PropertyId. Αν θέλαμε διάταξη των πολλαπλών τιμών θα χρειαζόταν ένα πεδίο OrderIndex.

Όταν με μια εγγραφή του propertygroups ορίζεται μια ομάδα τιμών ιδιότητας (δηλαδή το propertygroups.PropertyId δεν είναι null) τότε το propertygroups.PropertyId δείχνει στο properties.PropertyId (είναι FK σε κάποια ιδιότητα που ορίζεται στον πίνακα properties. Αναρωτιέται κανείς γιατί αυτός ο περιορισμός (FK constraint) δεν είναι 1:1. Η απάντηση είναι ότι μπορεί να έχουμε πολλά υποσύνολα τιμών για την ίδια ιδιότητα (το lookup χωρίζεται σε ενότητες τιμών).

Για να μην υπάρξει παρανόηση τονίζουμε ότι μια λίστα προκαθορισμένων τιμών (lookup ή λεξιλόγιο) μιας ιδιότητας ορίζεται αποκλειστικά από τον propertygroups και τον propertyvalues. Δηλαδή ο ορισμός γίνεται ακολουθώντας την FK διαδρομή του properties στον propertygroups (1:n) και του propertygroups στον propertyvalues (1:n).

7.3 Inner

Ο σκοπός της Inner Database είναι να διαχειριστούν όλες οι υποδιαιρέσεις που μπορεί να οριστούν σε μία λέξη και να μπορούν να καταχωρηθούν επισημειώσεις για κάθε ορισμένη υποδιαίρεση. Η απαίτηση που είχαμε από τους γλωσσολόγους (συγκεκριμένα από φωνολόγους που ασχολούνταν με τις προφορικές πηγές) ήταν πως θα ήθελαν ανάλυση των μορφολογικών λέξεων σε συλλαβές, φωνήματα, σύμφωνα και φωνήεντα. Για τις συλλαβές, τα φωνήματα, και τα σύμφωνα τους αρκούσε μόνο η μεταγραφή τους (σε φωνητικό αλφάβητο sampa). Για τα φωνήεντα ήθελαν να γνωρίζουν αν είναι τονισμένα, το είδος του τονισμού και αν η θέση του τονισμένου φωνήεντος είναι στην αρχή, στη μέση, στο τέλος της μορφολογικής λέξης ή ακόμα και αν είναι στο τέλος της επιτονικής φράσης. Μάλιστα είχαν ορίσει και μία δική τους κωδικοποίηση που συγκέντρωνε όλες τις (δύο) ιδιότητες (τύπος τονισμού και θέση τονισμένου φωνήεντος) σε ένα interval (μία θέση – ένα place holder) του tier (επιπέδου ανάλυσης) των

φωνηέντων στο εργαλείο που χρησιμοποιούσαν (Praat). Μια πρώτη ανάλυση οδήγησε στο σχεδιασμό ενός πίνακα (που θα ελάμβανε τη θέση του πίνακα oral_segments στο σχήμα της εικόνας 5). Αυτός ο database table (πίνακας) παρουσιάζεται μέσα από ένα παράδειγμα, στον Πίνακα 3.

Πίνακας 3: Η αρχική σκέψη (δημιουργίας database table) για τη μεταγραφή και την επισημείωση υποδιαιρέσεων λέξης

Phenomenon	**WID**	**start**	**stop**	**Level**	**interval_no**
u_s_e	30852	4.1234	4.2345	Vowel	22
t	30852	3.9876	4.1233	Consonant	27
u	30852	4.1234	4.2345	Segment	28
tu	30852	3.9876	4.2345	Syllable	12

Πρόκειται για μία λύση που επιτρέπει τη διπλή αξιοποίηση των δεδομένων. Μπορούμε να ψάξουμε για το φαινόμενο u_s_e (που σύμφωνα με την κωδικοποίηση των φωνολόγων αφορά το φωνήεν [u], Τονισμένο (stressed) και στο τέλος (end) της λέξης). Από την άλλη μπορούμε να ψάξουμε με την ταυτότητα μιας λέξης (word identifier – WID) και να ανακαλέσουμε όλα τα συστατικά και υποσυστατικά της λέξης. Μάλιστα μπορούμε να διατάξουμε όλα τα ανακτηθέντα intervals (θέσεις) του ιδίου επιπέδου (tier) υποδιαίρεσης (συλλαβή, φώνημα, φωνήεν και σύμφωνο) της λέξης. Η διάταξη ανά επίπεδο γίνεται με βάση την ιδιότητα interval_no (που προέρχεται από τα αρχεία TextGrid και είναι η αρχική διάταξη των στοιχείων του ιδίου επιπέδου). Έτσι μπορούμε να εκφράσουμε ερωτήματα απόστασης μεταξύ των συστατικών κάποιου επιπέδου (αυτό θα γίνει καλύτερα κατανοητό σε επόμενη ενότητα). Οι τιμές των ιδιοτήτων *phenomenon*, *start* και *stop* προέρχονται επίσης από το αρχικό αρχείο επισημείωσης (TextGrid).

Δηλαδή αυτή η δομή επιτρέπει την ιεραρχική αναπαραγωγή των δεδομένων από τα κατώτερα επίπεδα (συλλαβή, φώνημα, σύμφωνο και φωνήεν) της εικόνας 1 και ταυτοχρόνως τη σειριακή προσπέλαση των συστατικών στο ίδιο κατώτερο επίπεδο. Από την άλλη αυτή η δομή έχει περιορισμούς για την επέκταση των πληροφοριών επισημείωσης καθώς τόσο η μεταγραφή όσο και οι δύο ιδιότητες μπόρεσαν να συνδυαστούν και να συναποθηκευτούν στο ίδιο πεδίο (Phenomenon). Αν όμως προστεθούν και άλλες ιδιότητες (schema evolution) σε κάποιο ή κάποια από τα επίπεδα των συστατικών της λέξης, τότε πόσες ιδιότητες και πόσο σωστά μπορούν να συνωστιστούν στο ίδιο πεδίο; Για να διατηρήσει το σύστημα μας την επεκτασιμότητα του (να μην υποφέρει από schema evolution στα συστατικά της λέξης) πρέπει και οι ιδιότητες τονισμού και θέσης τονισμού των φωνηέντων να αποθηκευτούν όχι στον σχετικό πίνακα της database struct (όχι στον πίνακα oral_segments ή κάποιον άλλο στη θέση του) αλλά στην EAV database. Αυτό τελικά και κάναμε (υλοποιήσαμε). Βάλαμε έναν ελάχιστα διαφορετικό database table (από αυτόν του παραδείγματος στον πίνακα 3) στην struct database (βάλαμε δηλαδή τον πίνακα oral_segments που βλέπουμε στην εικόνα 5) και τον υποστηρίξαμε με το βοηθητικό πίνακα oral_SegmentTypes. Τώρα στον oral_segments αποθηκεύουμε μόνο την επισημείωση και τις ιδιότητες του συστατικού (όταν υπάρχουν) τις αποθηκεύουμε στην EAV database. Προφανώς και πρέπει να οριστούν οι δύο νέες ιδιότητες (τονισμού και θέσης τονισμού) στην EAV Database και πρέπει επίσης να προσδιοριστούν οι τιμές (lookup/λεξιλόγιου) της κάθε ιδιότητας. Έτσι η υλοποίηση μας βασίζεται στους πίνακες oral_segments και oral_SegmentTypes στην Struct Database και στη συμπλήρωση τιμών σε δύο εκ των

πινάκων (Properties και PropertyValues) της EAV Database. Όλα αυτά εμφανίζονται μέσα από ένα παράδειγμα στους πίνακες 4, 5, 6, 7 και 8. Το παράδειγμα είναι ισοδύναμο με αυτό που είδαμε σαν μια πρώτη προσέγγιση/ανάλυση (στον πίνακα 3).

Πίνακας 4: Δεδομένα στον πίνακα oral_SegmentTypes

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

Πίνακας 5: Απόσπασμα δεδομένων από τον πίνακα Oral_Segments

Segment Id	Text	Word Id	Position Index In Word	Position Index In Turn Taking	Position Index In Speaker	Start Time	End Time	TypeId
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

Πίνακας 6: Απόσπασμα δεδομένων από τον πίνακα Properties

Property Id	Name	TypeId	Is Mandatory	ModuleId
148	Τονισμός	3	0	11
149	Μέρος Τονισμού	3	0	11

Πίνακας 7: Απόσπασμα δεδομένων από τον πίνακα 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

Πίνακας 8: Απόσπασμα δεδομένων από τον πίνακα EntityPropertyValues

Entity Property Value Id	EntityId	Property Id	Value Id	ValueText
11240	8501	148	1446	NULL
11241	8501	149	1450	NULL

8. Τρόπος εισαγωγής δεδομένων

Τα δεδομένα των τεκμηρίων διακρίνονται από τρεις διαφορετικές καταστάσεις. Αυτές είναι:

- αρχική/πρωτότυπη μορφή,
- ψηφιακή αποτύπωση χωρίς υπολογιστικά οργανωμένη αναζήτηση και διαχείριση,
- ψηφιακή μορφή με υπολογιστικά οργανωμένη αναζήτηση και διαχείριση.

Στην πρώτη κατάσταση έχουμε τα πρωτότυπα τεκμήρια (έγγραφα, βιβλία, δακτυλογραφημένα και χειρόγραφα κείμενα για τα γραπτά καθώς και ηχογραφήσεις διαλόγων). Στη δεύτερη κατάσταση έχουμε ψηφιακές εκδόσεις των προηγούμενων (όπως ψηφιοποιημένες σελίδες των γραπτών σε αρχεία ηλεκτρονικών υπολογιστών και

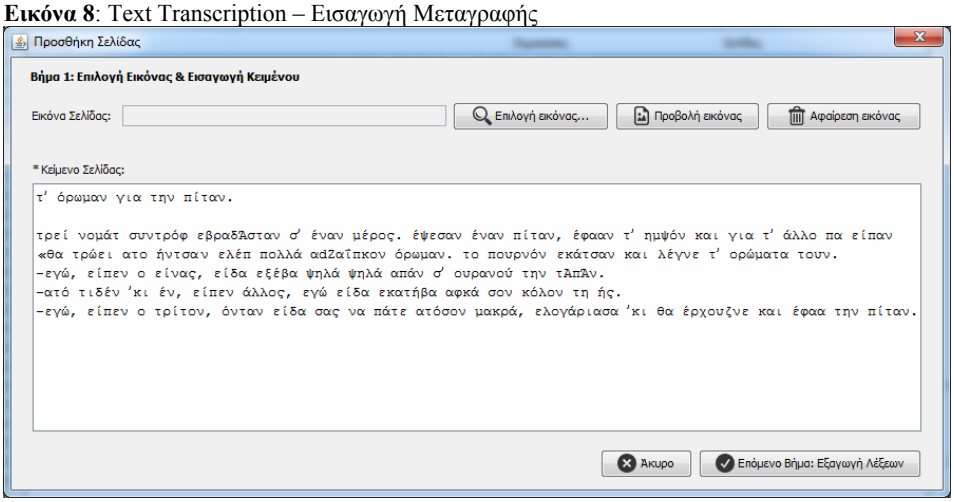
ψηφιοποιημένες εκδόσεις των ηχογραφήσεων και πάλι σε αρχεία ηλεκτρονικών υπολογιστών). Στη δεύτερη μορφή το υλικό συμπληρώνεται από μεταγραφές και επισημείωσεις (σε διάφορα δομικά επίπεδα των τεκμηρίων). Αμφότερες μεταγραφές και επισημειώσεις αποθηκεύονται σε επιπλέον αρχεία ηλεκτρονικών υπολογιστών. Στην τρίτη κατάσταση έχουμε συλλογές αρχείων (όπως είδαμε στην εικόνα 4 μπορούμε να έχουμε συλλογές από TextGrid, από Images, από Wav, από Μεταγραφές – text, κλπ) και βάσεις δεδομένων. Οι βάσεις δεδομένων έχουν διπλό σκοπό: (α) οργανώνουν τα δεδομένα των συλλογών αρχείων, (β) υποκαθιστούν πλήρως ορισμένα δεδομένα της δεύτερης κατάστασης. Για παράδειγμα τα αρχεία επισημειώσεων που (στη δεύτερη κατάσταση) βρίσκονται σε υπολογιστικά φύλλα μετασχηματίζονται πλήρως σε δεδομένα (της EAV) βάσης δεδομένων (στην τρίτη κατάσταση). Το ίδιο συμβαίνει και για τα υπολογιστικά φύλλα των μεταδεδομένων (κυρίως βιβλιογραφικά στοιχεία για τα γραπτά τεκμήρια και δημογραφικά και πολιτιστικά στοιχεία για τους ομιλητές των προφορικών τεκμηρίων) που επίσης μετασχηματίζονται πλήρως σε δεδομένα (της EAV) βάσης δεδομένων (στην τρίτη κατάσταση).

Σκοπός της παρούσας ενότητας είναι να παρουσιάσουμε τη διαδικασία για τη μετάβαση από τη δεύτερη στην τρίτη κατάσταση. Αυτό (η μετάβαση) μπορεί να γίνεται σε αλληλεπίδραση με το χρήστη (online) αλλά και μαζικά (για την αναδρομική εισαγωγή – import – του υλικού που είχε συγκεντρωθεί πριν την ανάπτυξη των πληροφοριακών συστημάτων). Πρώτα εξετάζουμε την μετάβαση/εισαγωγή – import – σε αλληλεπίδραση με το χρήστη. Μετά θα κάνουμε σύντομη αναφορά στη μαζική (αναδρομική) εισαγωγή.

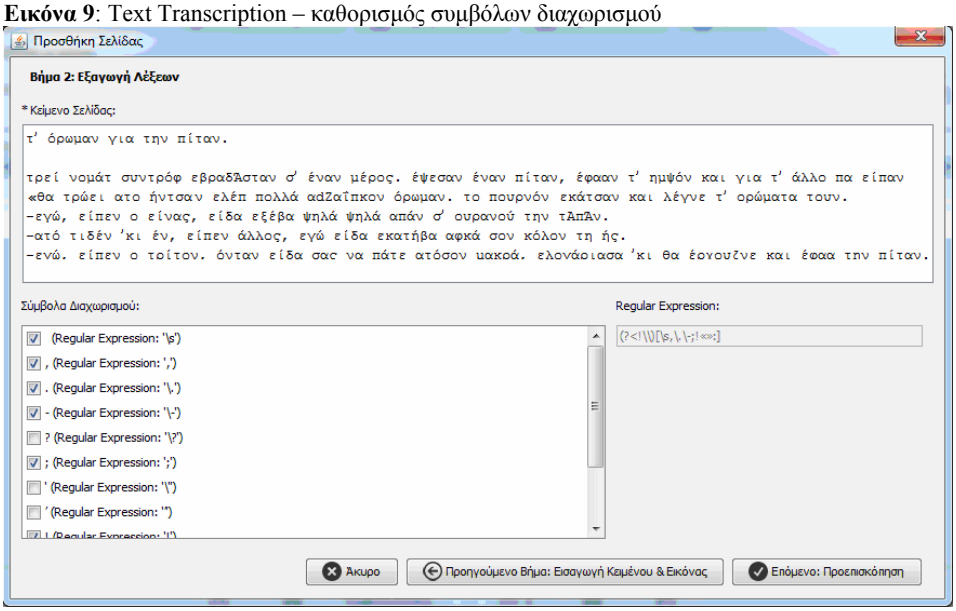
8.1 Εισαγωγή γραπτών τεκμηρίων σε αλληλεπίδραση με το χρήστη

Η εργασία αυτή γίνεται βηματικά. Πρώτα θα πρέπει να δημιουργηθεί ένα κενό γραπτό τεκμήριο. Αυτό (το κενό γραπτό τεκμήριο) δημιουργείται τόσο από browse (με τη χρήση του χειριστηρίου που παρέχει το module Browse Written) όσο και από το αρχικό menu. Στη συνέχεια ο χρήστης καλεί το P.I. Written, επαναληπτικά για κάθε σελίδα του γραπτού τεκμηρίου. Η προσθήκη της σελίδας γραπτού τεκμηρίου μπορεί να ξεκινήσει τόσο από browse (σε επιλεγμένο έγγραφο) όσο και από το τρίπτυχο γραπτών (G. Written) ενός επεξεργαζόμενου γραπτού τεκμήριου. Η προσθήκη σελίδας (P.I. Written) καλεί το Text Imaging για να κάνει την εισαγωγή εικόνας (ψηφιοποιημένης σελίδας) και το Text Transcription για να κάνει τη μεταγραφή αυτής. Για τη μεταγραφή, ο χρήστης, εισαγάγει το κείμενο μεταγραφής, καθορίζει τα σύμβολα διαχωρισμού (tokenization – διάσπασης σε λέξεις) και δοκιμάζει (βλέπει) αν ο διαχωρισμός τον ικανοποιεί. Στις εικόνες 8 και 9 βλέπουμε ένα παράδειγμα εισαγωγής μεταγραφής και καθορισμού των συμβόλων διαχωρισμού της σε λέξεις.

Εικόνα 8: Text Transcription – Εισαγωγή Μεταγραφής



Εικόνα 9: Text Transcription – καθορισμός συμβόλων διαχωρισμού



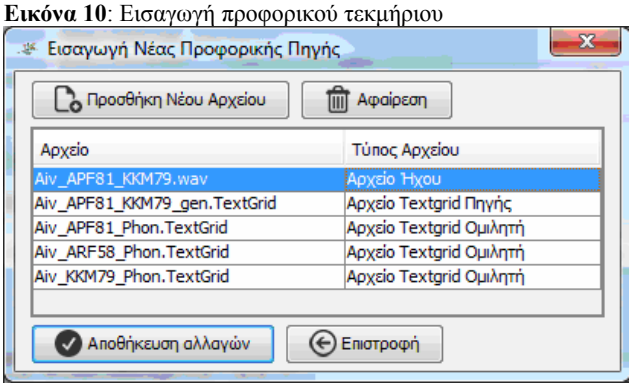
Αξίζει να σημειωθεί ότι όλες οι επισημειώσεις (μορφολογικές, φωνολογικές, κλπ) γίνονται (αργότερα) από το χρήστη (με τη βοήθεια του τρίπτυχου γραπτών και των επιμέρους modules επισημείωσης που το τρίπτυχο επικαλείται).

8.2 Εισαγωγή προφορικών τεκμηρίων σε αλληλεπίδραση με το χρήστη

Η εργασία αυτή (σε αντίθεση με την προηγούμενη) δεν γίνεται βηματικά. Ο χρήστης ζητά να δημιουργήσει (εισαγάγει) ένα προφορικό τεκμήριο χρησιμοποιώντας τις

κατάλληλες επιλογές (χειριστήρια) που του δίνονται από το browse (Browse Oral) και από το αρχικό menu. Το σύστημα τότε εμφανίζει στο χρήστη ένα παράθυρο (το παράθυρο του module I. Oral) και ο χρήστης καλείται και συμπληρώνει όλα τα αρχεία μεταγραφών (TextGrid) και τα ηχητικά αρχεία (Wav) που συνθέτουν το προφορικό τεκμήριο (στη δεύτερη κατάσταση). Στην εικόνα 10 βλέπουμε ένα τέτοιο παράθυρο προσδιορισμού των μεταγραφών (TextGrid) και των ηχητικών αρχείων που συνθέτουν ένα προφορικό τεκμήριο. Οι συντελεστές του έργου μπορούν να διακρίνουν ότι σε αυτό το παράδειγμα έχουμε ένα διάλογο τριών ομιλητών (και ισάριθμες μεταγραφές, μία ανά ομιλητή), μία ηχητική ψηφιακή καταγραφή της συνομιλίας και μία γενική μεταγραφή ανεξάρτητη ομιλητή. Πρόκειται επίσης για Αϊβαλιώτικα όπου ο εποπτεύοντας μελετητής είναι φυσικός ομιλητής της διαλέκτου και συμμετέχει και ο ίδιος στο διάλογο.

Εικόνα 10: Εισαγωγή προφορικού τεκμήριου



Αξίζει να σημειωθεί ότι οι μεταγραφές και οι επισημειώσεις (μορφολογικές, φωνολογικές, και επιπλέον σε επίπεδο συστατικών λέξης) βρίσκονται στα TextGrid αρχεία και αφού περάσουν από ένα βήμα ελέγχου (parsing) ενσωματώνονται στις υπολογιστικά οργανωμένες και διαχειριζόμενες πληροφορίες και είναι διαθέσιμες για αναζήτηση από το χρήστη. Αυτό δεν αποκλείει το χρήστη από το να προβεί (αργότερα και κυρίως) σε συμπληρωματικές μορφολογικές και φωνολογικές επισημειώσεις, (με τη βοήθεια του τρίπτυχου προφορικών και των επιμέρους modules επισημείωσης που το τρίπτυχο επικαλείται).

8.3 Μαζική εισαγωγή γραπτών τεκμηρίων

Πρόκειται για λογισμικό που εισαγάγει (import) το σύνολο των αρχείων ηλεκτρονικών υπολογιστών που περιέχουν μεταγραφές και επισημειώσεις. Δηλαδή, το σύνολο των αρχείων που συνθέτουν ένα τεκμήριο στη δεύτερη κατάσταση, ελέγχεται (περνάει από parsing) και ενσωματώνεται στις υπολογιστικά οργανωμένες και διαχειριζόμενες πληροφορίες (databases και συλλογές αρχείων). Για καλύτερη κατανόηση αναφέρουμε ότι για κάθε τεκμήριο, το σύνολο των ψηφιοποιημένων σελίδων (αρχεία εικόνας), το σύνολο των μεταγραφών (αρχεία κειμένου, όσα και οι εικόνες/σελίδες), το σύνολο των μορφολογικών επισημειώσεων (αρχεία λογιστικών φύλλων) και τα μεταδεδομένα του τεκμηρίου (ένα ακόμα αρχείο λογιστικού φύλλου) συγκεντρωνονται σε ένα κατάλογο για να αποτελέσουν ένα τεκμήριο στη δεύτερη κατάσταση. Πολλοί τέτοιοι κατάλογοι

(οργανωμένοι σε ένα ευρύτερο/προγονικό κατάλογο) τροφοδοτούν το λογισμικό (καλύτερα υπηρεσία) μαζικής εισαγωγής γραπτών (Massive Import Written – M.I. Written) και αυτό επαναλαμβάνει το έργο του (parsing και ενσωμάτωση) για κάθε τεκμήριο. Είναι προφανές ότι το σύνολο των μεταγραφών και επισημειώσεων του κάθε τεκμηρίου οργανώνεται και γίνεται διαχειρίσιμο υπολογιστικά.

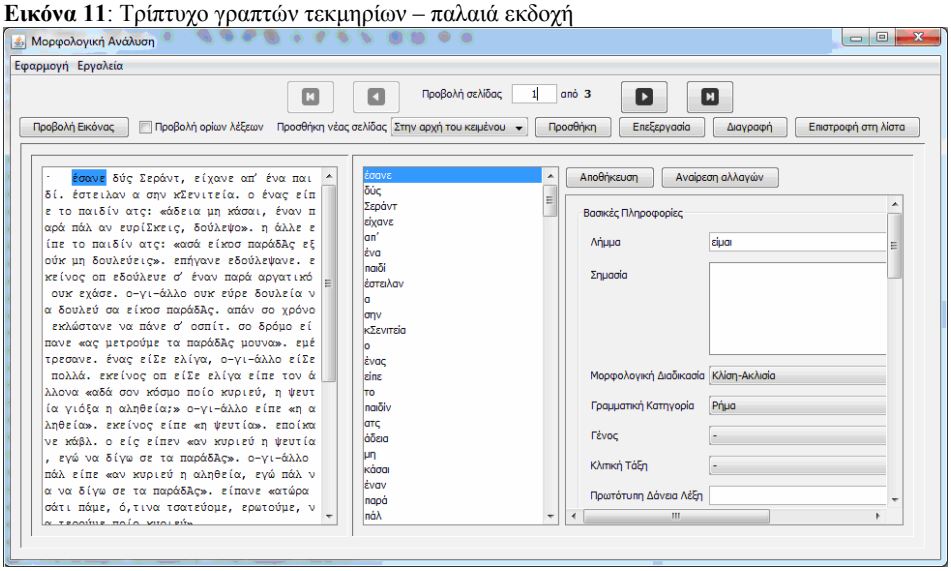
8.4 Μαζική εισαγωγή προφορικών τεκμηρίων

Πρόκειται για λογισμικό (Massive Import Oral – M.I. Oral) που εισαγάγει πολλά προφορικά τεκμήρια. Για κάθε τεκμήριο εισαγάγει (import) το σύνολο των αρχείων ηλεκτρονικών υπολογιστών που περιέχουν μεταγραφές και επισημειώσεις και συνθέτουν ένα προφορικό τεκμήριο στη δεύτερη κατάσταση. Στην περίπτωση αυτή δεν υπάρχει ουσιαστική διαφορά μεταξύ του I. Oral και του M.I. Oral. Η μοναδική διαφορά είναι ότι το τελευταίο επενεργεί σε πολλούς καταλόγους (πολλά προφορικά τεκμήρια) και για κάθε ένα κάνει την ίδια δουλειά που θα έκανε το I. Oral.

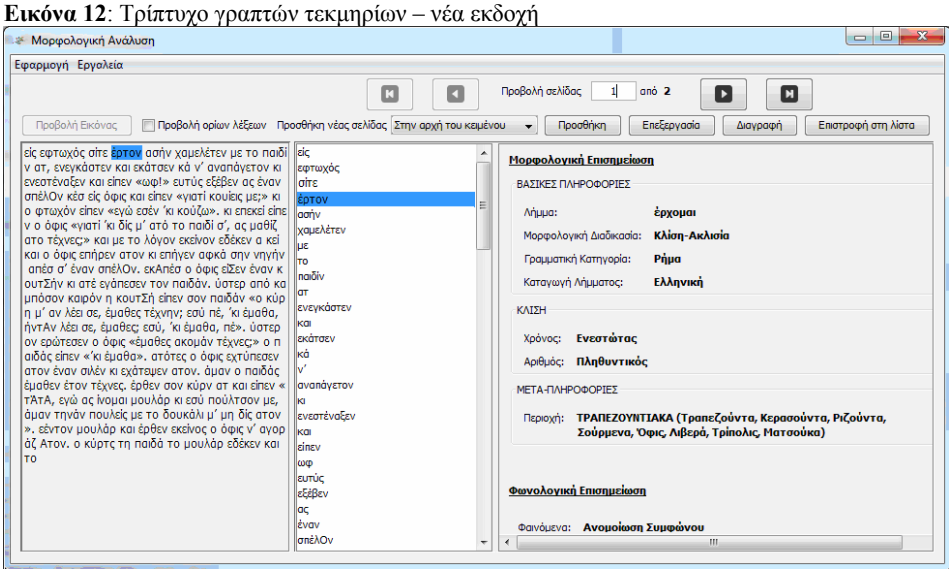
9. Επεξεργασία δεδομένων

Για την εποπτεία ενός γραπτού τεκμηρίου παρέχεται ένα τρίπτυχο (module G.Written). Στο τρίπτυχο γραπτών τεκμηρίων, αριστερά εμφανίζεται η μεταγραφή σελίδας, στη μέση εμφανίζονται οι μορφολογικές λέξεις της σελίδας και δεξιά εμφανίζονται και μπορούν να ενημερωθούν οι επισημειώσεις μίας (κάθε φορά) λέξης. Το τρίπτυχο παρέχει χειριστήρια για τη μετακίνηση μεταξύ των σελίδων του γραπτού τεκμηρίου. Το G.Written παρέχει επίσης χειριστήριο (κουμπί) για την προσθήκη σελίδας γραπτού τεκμηρίου (που καλεί το P.I. Written). Στην εικόνα 11 βλέπουμε τη μορφή που είχε το τρίπτυχο γραπτών τεκμηρίων (G.Written module). Σε μια νεότερη έκδοση (εικόνα 12) εμφανίζονται μόνο οι ιδιότητες (επισημειώσεις) της λέξης που έχουν τιμές. Σε αυτή την έκδοση, το σύνολο των ιδιοτήτων εμφανίζεται μόνο όταν κληθεί η φόρμα ενημέρωσης.

Εικόνα 11: Τρίπτυχο γραπτών τεκμηρίων – παλαιά εκδοχή

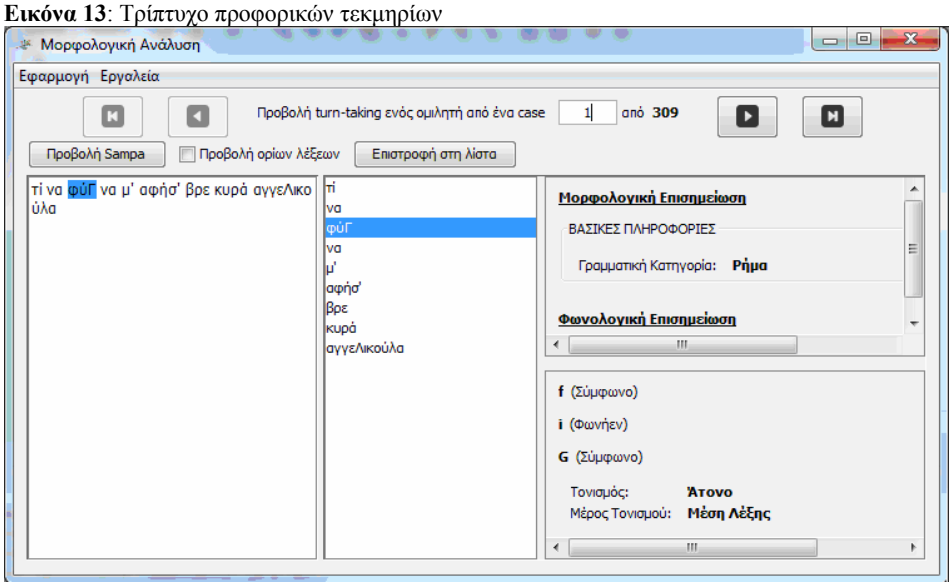


Εικόνα 12: Τρίπτυχο γραπτών τεκμηρίων – νέα εκδοχή



Για την εποπτεία ενός προφορικού τεκμηρίου παρέχεται ένα τρίπτυχο (module G.Oral). Στο τρίπτυχο προφορικών τεκμηρίων, αριστερά εμφανίζεται η μεταγραφή ενός εκφωνήματος του ομιλητή, στη μέση εμφανίζονται οι μορφολογικές λέξεις του εκφωνήματος σε ορθογραφική μεταγραφή ή σε αλφάβητο sampa (toggle) και δεξιά εμφανίζονται οι επισημειώσεις μίας (κάθε φορά) λέξης. Το τρίπτυχο παρέχει χειριστήρια για τη μετακίνηση μεταξύ των εκφωνημάτων του ομιλητή του προφορικού τεκμηρίου. Στην εικόνα 13 βλέπουμε το τρίπτυχο προφορικών τεκμηρίων.

Εικόνα 13: Τρίπτυχο προφορικών τεκμηρίων



10. Αναζήτηση δεδομένων

Οι απαιτήσεις που τέθηκαν για το υποσύστημα αναζήτησης πληροφορίας (search module) προσδιορίζονται επακριβώς στην επόμενη λίστα:

- Διαισθητική χρήση,
- Υποστήριξη ερωτημάτων που λαμβάνουν υπόψη την ύπαρξη πολλαπλών τιμών για τα πεδία επισημειώσεων. Με άλλα λόγια, θα πρέπει να επιτρέπει στο χρήστη να απαιτήσει να βρεθούν ταυτοχρόνως δύο (ή δυνητικά/μελλοντικά και περισσότερες από δύο) τιμές σε ένα πεδίο μιας εγγραφής (αντικείμενο στο ίδιο επίπεδο της struct database). Δηλαδή, το υποσύστημα αναζήτησης, πέρα από τις απαιτήσεις για ακριβές (Exact) ταίριασμα, ταίριασμα εύρους (Range) και διάζευξης (Disjunction) τιμών, πρέπει να υποστηρίζει και την λογική σύζευξη τιμών και να παρέχει αντίστοιχο τελεστή (*And*) μεταξύ τιμών για το ίδιο κριτήριο αναζήτησης (Karanikolas et al 2014b).
- Κάθε κριτήριο να μπορεί να επιβάλει 2 τύπων περιορισμούς (*τιμών* – *βασικούς* – και *απόστασης* – βλέπε και παρακάτω),
- Πολλαπλά κριτήρια,
- Σύζευξη κριτηρίων (υπονοούμενη χρήση του τελεστή *And* μεταξύ διαφορετικών κριτηρίων) (Karanikolas et al 2011),
- Έκφραση συνθηκών απόστασης (προαναφέρθηκαν ως *περιορισμοί απόστασης* – *distance conditions*).
- Κάθε κριτήριο να εστιάζει σε κάποιο από τα επίπεδα (*Document, Part, Word, Inner*).
- Έκφραση απαιτήσεων ανάκτησης για: πραγματικά δεδομένα, σύννοψη δεδομένων (data aggregations), τεχνητά δεδομένα (artifacts - on the fly created data),

10.1 Η διεπαφή

Όπως αναφέρθηκε μπορούν (στο ίδιο ερώτημα αναζήτησης) να συνδυαστούν (με υπονοούμενο And) περισσότερα από ένα κριτήρια. Μια πρόσφορη μορφή για την έκφραση του κάθε κριτηρίου ή (με βάση τον τρόπο που υλοποιείται) η μορφή της κάθε γραμμής ερωτήματος σχηματοποιείται σύμφωνα με το template του πίνακα 9.

Πίνακας 9: Δομή γραμμής ερωτήματος

Word/ token /phenomenon			**Location**				
<Value>	{ Between, And, Or, Exact}	<Value>	<EAV subschema>*	<Attribute>	<Part distances>	<Word distances>	<Interval no distances >

* αναφορά/υπόδειξη ενός εκ των 11 modules (9 επισημειώσεων και 2 μεταπληροφορίας)

Πέρα από το ένα ή περισσότερα κριτήρια που προσδιορίζουν τι θα αναζητηθεί (Search), η διεπαφή πρέπει να προσδιορίζει και τι θα ανακληθεί (Retrieve) για να εμφανιστεί. Αυτό (τι θα εμφανιστεί) απαιτείται να προσδιορίζεται μία φορά για το σύνολο των κριτηρίων. Στον πίνακα 10 βλέπουμε ένα template που επιλέχθηκε για να

χρησιμοποιηθεί. Οι τιμές για τη θέση <Result Type> είναι: *Document, Part, Word, Inner*. Οι τιμές για τη θέση <Aggregate> (συνάθροιση) είναι: *count, count_documents, count_parts, count_words, count_inners* and *null*. Οι τιμές για τη θέση <Artifact> είναι: *Mini praat* (ένα TextGrid που αφορά, συνήθως, μια λέξη) και *null*.

Πίνακας 10: Δομή για τον προσδιορισμό του επιθυμητού αποτελέσματος

Output	<Result Type>	<Aggregate> or <Artifact>

Στους πίνακες 11, 12 και 13, παρουσιάζουμε παραδείγματα από συμπληρωμένες διεπαφές του search (ή ακριβέστερα Search and Retrieve) module. Η απεικονιζόμενη στον πίνακα 11 διεπαφή, προσδιορίζει ότι αναζητούνται parts (σελίδες στην περίπτωση των γραπτών πηγών) που περιλαμβάνουν το φαινόμενο του *φωνηεντικού αρχαϊσμού* (*Vowel Archaism*), ακολουθούμενου από ένα *επίθετο* (*adjective*) το οποίο είναι μια δάνεια λέξη (loan word) και αυτή η δάνεια λέξη είναι μέρος του λόγου Ουσιαστικό (*Noun*) και έχει γένος *Αρσενικό* (*Masculin*). Η απεικονιζόμενη στον πίνακα 12 διεπαφή, προσδιορίζει ότι αναζητούνται parts (επιτονικές προτάσεις στην περίπτωση προφορικών πηγών) που περιέχουν άτονο (unstressed) φωνήεν στο τέλος φράσης. Μάλιστα στην περίπτωση αυτή θα μετρηθούν τα parts (πλήθος επιτονικών προτάσεων) που περιέχουν το ζητούμενο, για κάθε document (ομιλητή) στον οποίο εμφανίζεται το φαινόμενο. Η απεικονιζόμενη στον πίνακα 13 διεπαφή, προσδιορίζει ότι αναζητούνται documents (ομιλητές στην περίπτωση προφορικών πηγών που συμμετείχαν σε διαλόγους). Για την εύρεση τους προσδιορίζονται μεταδεδομένα των ομιλητών (φύλο, ηλικία και καταγωγή) καθώς και μεταδεδομένα (ονοματεπώνυμο) του εποπτεύοντα μελετητή (ή επισημειωτή) του διαλόγου (στον οποίο συμμετείχαν οι ομιλητές).

Πίνακας 11: Παράδειγμα αναζήτησης

Word/ token /phenomenon			**Location**				
vowel archaism	--		Morphological Written	-	-	X	-
Adjective	--		Morphological Written	PART OF SPEECH	-	Y in (X+1, X+10)	-
Noun	--		Morphological Written	PART OF SPEECH OF	-	Y	-
Masculin	--		Morphological Written	GENDER OF LOAN WORD	-	Y	-

Output	Part	-

Πίνακας 12: Παράδειγμα αναζήτησης

Word/ token /phenomenon			**Location**				
?_u_f	--		Inner Oral	Vowel	-	-	-

Output	Document	count_part

Πίνακας 13: Παράδειγμα αναζήτησης

Word/ token /phenomenon			**Location**				
Ifigenia Zisi	Or	Mary Karra	Metadata Oral	Annotator	-	-	-
Male	--		Metadata Oral	Inf. Sex	-	-	-
75	Between	100	Metadata Oral	Inf. Age	-	-	-
cappadocians	--		Metadata Oral	Inf. Origin	-	-	-

Output	Document	-

10.2 BNF για τα <Part distances>, <Word distances> και <Interval no distances>

Η απόσταση των στοιχείων (στις θέσεις <Part distances>, <Word distances> και <Interval_no distances>) προσδιορίζεται από μία απλή γραμματική. Η σύνταξη της δίδεται σε συμβολισμό BNF:

<Location> ::= <Location single> | <Location range>
<Location range> ::= <μεταβλητή> in (<Location single> , <Location single>)
<Location single> ::= <μεταβλητή> [<operator> <αριθμός>]
<μεταβλητή> ::= <latin char>
<latin char> ::= {a|b|c|d|e|f|g|h}
<operator> ::= {+|-}
<αριθμός> ::= <ψηφίο> [<ψηφίο>]
<ψηφίο> ::= {0|1|2|3|4|5|6|7|8|9}

Αρχικό σύμβολο είναι το <Location>. Οι χαρακτήρες που επιλέχθηκαν για τη μεταβλητή (unbound variable) είναι τέτοιοι που όταν είναι σε πεζά ξεχωρίζουν και δεν υπάρχει κίνδυνος να μπερδευτούν με Ελληνικούς χαρακτήρες. Έτσι αν ο χρήστης γράψει κάποια μεταβλητή με Ελληνικούς χαρακτήρες, να μπορεί το σύστημα (με το parsing) να εντοπίσει το λάθος και να δείξει το συνολικό statement με πεζούς χαρακτήρες, υποδεικνύοντας τη θέση του λάθους (χωρίς να υπάρχει πιθανότητα να μη γίνει κατανοητή η φύση του λάθους).

11. Συμπεράσματα

Στην εργασία αυτή παρουσιάσαμε τη σχεδίαση και ανάπτυξη ενός συστήματος δημιουργίας και διαχείρισης σώματος τεκμηρίων από γραπτές και προφορικές πηγές. Το σύστημα έχει σχεδιαστεί κατά τρόπον ώστε να μπορεί να χρησιμοποιηθεί για την κατασκευή και διαχείριση σωμάτων τεκμηρίων άλλων γλωσσών ή διαλέκτων.

Το σύστημα που αναπτύχθηκε (ως πολυτροπική βάση δεδομένων) επιτρέπει την παράλληλη εμφάνισης πρωτογενών και επεξεργασμένων δεδομένων καθώς και την κωδικοποίηση ενός μεγάλου αριθμού μεταδεδομένων. Κατά τη σχεδίαση του επιδιώχθηκε μια γενίκευση των προδιαγραφών έτσι ώστε να μην είναι το σύστημα αυστηρά εξαρτημένο από συγκεκριμένα γλωσσολογικά φαινόμενα και τρόπους αντίληψης / χειρισμού αυτών. Πιστεύουμε ότι το σύστημα στο οποίο καταλήξαμε

μπορεί να χρησιμοποιηθεί και σε άλλα έργα δημιουργίας και διαχείρισης σώματος τεκμηρίων ακόμα και όταν αυτά δεν έχουν τις ίδιες ανάγκες, στόχους και προδιαγραφές. Δηλαδή, πρόκειται για μία δυναμική πλατφόρμα που επιτρέπει την προσαρμογή της σε νέες ανάγκες που θα προκύψουν σε επόμενες έρευνες.

References

Anderson, J., Beavan, D., Kay, C. (2007) SCOTS: Scottish Corpus of Texts and Speech. In Beal J. (ed.), *Creating and digitalizing Language Corpora Vol.1*. Palgrave McMillan Publication, 17-34.

Anhøj, J. (2003) Generic Design of Web-Based Clinical Databases. *Journal Medical Internet Research, 4*.
doi: 10.2196/jmir.5.4.e27

Barbiers, S. et al (2006) *Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND)*. Amsterdam, Meertens Institute.
http://www.meertens.knaw.nl/sand/

Boersma, P. (2012) The use of Praat in corpus research. In Durand, J., Gut, U. & Kristofferson, G. (eds.), *Handbook of corpus phonology*. Oxford: Oxford University Press.

Boersma, P. & Weenink, D. (2013). *Praat: Doing phonetics by computer*.
http://www.praat.org

Buseman, A. & Buseman, K. *Toolbox Self-Training: How to use the Field Linguist's Toolbox*.
http://www.ling.helsinki.fi/kit/2009k/clt234/docs/Toolbox_Self-Training.pdf

Clua, E. & Lloret, M-R. (2006) New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD). In Jean-Pierre Y. Montreuil (ed.), *New Perspectives on Romance Linguistics, vol. 2 (Phonetics, Phonology, and Dialectology)*. Amsterdam/Philadelphia: John Benjamins.
http://pages.uv.es/foncat/cat/Treballs/10.Clua-Lloret.pdf

ELAN, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.
http://tla.mpi.nl/tools/tla-tools/elan/

Fromont, R. & Hay, J. (2008) ONZE Miner: the development of a browser-based research tool, *Corpora, 3(2)*: 173-193.

Galiotou, E., Karanikolas, N.N., Manolessou, I., Pantelidis, N., Papazachariou, D., Ralli, A. & Xydopoulos, G. (2014) Asia Minor Greek: Towards a Computational Processing. *Procedia, Social and Behavioral Sciences, vol. 147*: 458-466.
doi: 10.1016/j.sbspro.2014.07.138

Greek Sampa.
https://www.phon.ucl.ac.uk/home/sampa/greek.htm

The International Phonetic Alphabet and the IPA Chart.
https://www.internationalphoneticassociation.org/content/ipa-chart

IPA Chart With Sounds.
http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/

Karanikolas, N.N. (2011) Search Culture. In *Proceedings of the 15th Panhellenic Conference on Informatics, PCI'2011*, Kastoria, Greece, Sept. 30– Oct. 2, 2011. IEEE CPS. doi: 10.1109/PCI.2011.23

Karanikolas, N.N., Galiotou, E. & Ralli, A. (2014) Towards a Unified Exploitation of Electronic Dialectal Corpora: Problems and Perspectives. *TSD'2014: Text, Speech and Dialogue, 17th International Conference*. Brno, Czech Republic, September 8–12, 2014. Springer, LNAI 8655: 257-266. doi: 10.1007/978-3-319-10816-2_32

Karanikolas N.N. & Skourlas, C. (2014) Personal Digital Libraries: a self-archiving approach. *Library Review, 63 (6/7)*: 436-451. doi: 10.1108/LR-06-2014-0073

Karanikolas, N.N., Galiotou, E., Papazachariou, D., Athanasakos, K., Koronakis, G. & Ralli, A. (2015) Towards a computational processing of oral dialectal data. *Proceedings of the 19th Panhellenic Conference on Informatics, PCI 2015*, Athens, Greece, October 01 - 03, 2015. ACM 978-1-4503-3551-5. doi: 10.1145/2801948.2801966

Karasimos A., Galiotou E., Karanikolas, N.N., Koronakis, G., Athanasakos, K. Papazachariou, D. & Ralli, A. (2014) Challenges of Annotating a Multi-Dialect, Multi-Level Corpus of Spoken and Written Modern Greek Dialects. *MGDLT6: 6th International Conference on Modern Greek Dialects & Linguistic Theory*, Patras, Greece, 25-28 Sept. 2014.

Kunst, J.P. & Wesseling, F. (2010) Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Tohoku University, Nov. 4-7, 2010, 759-767. http://www.aclweb.org/anthology/Y10-1088

LaBB-CAT (formerly known as ONZE Miner). http://onzeminer.sourceforge.net/

Nerbonne, J. & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities 37 (3)*: 339-357.

SAMPA - computer readable phonetic alphabet. https://www.phon.ucl.ac.uk/home/sampa/

Sloetjes, H. & Wittenburg, P. (2008) Annotation by category – ELAN and ISO DCR. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Ubul, A., Kake, H., Sakoguchi, Y. & Kishie, S. (2015) Research on Oral Map in Regional Dialect Using Google Map. *Int. Jour. Comp. Tech. 2 (2):* 31-35. http://ijcat.org/IJCAT-2015/2-2/Research-on-Oral-Map-in-Regional-Dialect-Using-Google-Map.pdf

Wells, J.C. (1997) SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R. & Winski, R. (eds.), *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter.