ΑΡΧΕΙΟΘΕΤΗΣΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΨΗΦΙΟΠΟΙΗΜΕΝΩΝ ΠΗΓΩΝ ΕΛΛΗΝΙΚΩΝ ΔΙΑΛΕΚΤΩΝ

Ε. Γαλιώτου a, *, N. Καρανικόλας a, Α. Ράλλη b

a Τμήμα Μηχανικών Πληροφορικής Τ.Ε., Τ.Ε.Ι. Αθήνας, Αγ. Σπυρίδωνα, 122 10 Αιγάλεω, Αθήνα - (egali, nnk)@teiath.gr
b Τμήμα Φιλολογίας, Πανεπιστήμιο Πατρών, Πανεπιστημιούπολη, 265 04 Ρίο, Πάτρα- ralli@upatras.gr

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ελληνικές διάλεκτοι, Σώματα κειμένων, Αρχειοθέτηση, Ψηφιοποίηση, Πολυμεσικές βάσεις, Υπολογιστική διαλεκτολογία

ΠΕΡΙΛΗΨΗ:

Οι ελληνικές διάλεκτοι της Μικράς Ασίας, όπως τα Ποντιακά, τα Καππαδοκικά και τα Αϊβαλιώτικα, θεωρούνται ιδανικές περιπτώσεις μελέτης για τη διαλεύκανση της εξέλιξης της Ελληνικής, καθώς και για διάφορα φαινόμενα γλωσσικής επαφής, λόγω της μακρόχρονης επαφής με την Τουρκική και της σχετικής απομόνωσής τους από τις υπόλοιπες ελληνικές διαλέκτους. Οι συγκεκριμένες διάλεκτοι εκφράζουν μία πλούσια πολιτισμική και γλωσσική κληρονομιά, αλλά αντιμετωπίζουν σοβαρό κίνδυνο εξαφάνισης, ιδιαίτερα τα Καππαδοκικά. Επομένως, η περιγραφή και διατήρηση αυτής της πολύτιμης κληρονομιάς προβάλλει ως επιτακτική ανάγκη. Η χρήση των νέων τεχνολογιών στην επεξεργασία των διαλέκτων συμβάλλει με καθοριστικό τρόπο στη διάσωση και την ανάδειξη αυτής της σημαντικής πολιτιστικής κληρονομιάς. Στην εργασία, περιγράφεται ένα καινοτόμο σύστημα αρχειοθέτησης και επεξεργασίας ψηφιοποιημένων σωμάτων γραπτών και προφορικών τεκμηρίων των τριών ελληνικών διαλέκτων της Μικράς Ασίας (Ποντιακά. Καππαδοκικά, Αϊβαλιώτικα), τα οποία έχουν συλλεγεί στο πλαίσιο του ερευνητικού προγράμματος AMiGre (ΘΑΛΗΣ). Το σύστημα έχει ως πυρήνα μία πολυτροπική βάση δεδομένων η οποία επιτρέπει την παράλληλη εμφάνιση πρωτογενών και επεξεργασμένων δεδομένων καθώς και την κωδικοποίηση μεγάλου αριθμού πληροφοριών μεταδεδομένων. Το υποσύστημα αναζήτησης επιτρέπει: (α) συνδυασμένη αναζήτηση σε διαφορετικά επίπεδα γλωσσικής αναπαράστασης (φωνολογικό, μορφολογικό), (β) πρόσβαση σε μεταδεδομένα και (γ) συνδυασμένη αναζήτηση στα τεκμήρια τόσο των γραπτών όσο και των προφορικών πηγών.

1. ΕΙΣΑΓΩΓΗ

Μετά την ανταλλαγή πληθυσμών μεταξύ Ελλάδας και Τουρκίας, που επιβλήθηκε με τη Συνθήκη της Λωζάνης (1923), ομιλητές των ελληνικών διαλέκτων της Μικράς Ασίας εγκαταστάθηκαν σε διάφορα μέρη της Ελλάδας, αφήνοντας πίσω τους γη και περιουσία, αλλά παίρνοντας μαζί τους την ιστορία, τα ήθη και τα έθιμα, τις παραδόσεις και τη γλώσσα τους. Οι διάλεκτοι αυτές, όπως τα Ποντιακά, τα Καππαδοκικά και τα Αϊβαλιώτικα, θεωρούνται ιδανικές περιπτώσεις μελέτης για τη διαλεύκανση της εξέλιξης της Ελληνικής, καθώς και για διάφορα φαινόμενα γλωσσικής επαφής, λόγω της μακρόχρονης επαφής με την Τουρκική και της σχετικής απομόνωσής τους από τις υπόλοιπες ελληνικές διαλέκτους. Τα Καππαδοκικά (Dawkins 1916) προέρχονται από μία ελληνική διάλεκτο της ύστερης αρχαιότητας, αλλά δέχτηκαν ισχυρή επίδραση από την Τουρκική ύστερα από την κατάκτηση της Καππαδοκίας, αρχικά από τους Σελτζούκους Τούρκους τον 11ο αι. και στη συνέχεια από τους Οθωμανούς τον 14ο αι. Σήμερα, αποτελούν ένα διαλεκτικό συνεχές που ομιλείται από μερικές δεκάδες πρόσφυγες (κυρίως δεύτερης και τρίτης γενιάς), κατοίκους προσφυγικών χωριών της Κεντρικής και Βόρειας Ελλάδας. Τα Ποντιακά καταγράφονται συνεχώς από την ύστερη αρχαιότητα και εφεξής. Διατηρούν ακόμη ορισμένα αρχαϊκά χαρακτηριστικά, και έχουν επίσης επηρεαστεί από την Τουρκική (Παπαδόπουλος 1955, Drettas 1997). Σήμερα, ομιλητές της Ποντιακής απαντούν σε όλη την Ελλάδα, αλλά κυρίως σε θύλακες στην Ήπειρο, τη Μακεδονία και τη Δυτική Θράκη. Είναι αξιοσημείωτο ότι η διάλεκτος ομιλείται ακόμη στον Πόντο από έναν αριθμό μουσουλμάνων κατοίκων (Mackridge 1990), καθώς και σε ορισμένες περιοχές της Γεωργίας και του Βορείου Καυκάσου. Τέλος, οι πρώτες μαρτυρίες για τα Αϊβαλιώτικα απαντούν τον 16ο αιώνα (βλ. Σάκκαρης 1940). Η διάλεκτος αυτή εμφανίστηκε μετά την εγκατάσταση κυρίως Λεσβίων αποίκων στις Κυδωνίες (Αϊβαλί) και τα Μοσχονήσια. Ανήκει στην ομάδα των βορείων ελληνικών διαλέκτων και έχει δεχτεί ισχυρή επίδραση από την Τουρκική, κυρίως στο μορφολογικό επίπεδο. Σήμερα, μερικές εκατοντάδες ομιλητές των Αϊβαλιώτικων βρίσκονται στη Λέσβο, όπου οι πρόγονοί τους εκτοπίστηκαν μετά την ανταλλαγή των πληθυσμών. Οι τρεις διάλεκτοι εκφράζουν μία πλούσια πολιτισμική και γλωσσική κληρονομιά, αλλά αντιμετωπίζουν σοβαρό κίνδυνο εξαφάνισης: ο αριθμός των προσφύγων πρώτης γενιάς είναι σχεδόν ανύπαρκτος, ενώ οι επόμενες γενιές απορροφώνται σταδιακά από το αντίστρωμα (adstratum) της Κοινής Νέας Ελληνικής (ΚΝΕ), τόσο πολιτισμικά όσο και γλωσσικά. Επομένως, η περιγραφή και διατήρηση αυτής της πολύτιμης κληρονομιάς προβάλλει ως επιτακτική ανάγκη.

Το ερευνητικό πρόγραμμα ΘΑΛΗΣ : «Πόντος, Καππαδοκία, Αϊβαλί: Στα χνάρια της Μικρασιατικής Ελληνικής» (AMiGre) είχε ως στόχο τη συστηματική μελέτη των τριών συγκεκριμένων διαλέκτων της Μικράς Ασίας τόσο αυτοτελώς όσο και συγκριτικά μεταξύ τους. Το έργο αυτό αποτελεί την πρώτη απόπειρα για μια συνολική συγκριτική γλωσσολογική μελέτη των μικρασιατικών διαλέκτων της Ελληνικής. Αποτελεί επίσης την πρώτη απόπειρα στην Ελλάδα να συνδυαστούν η Θεωρητική Γλωσσολογία, η Πληροφορική και η Τεχνολογία της Πληροφορίας με στόχο την επιστημονική παρουσίαση ελληνικών διαλεκτικών δεδομένων στην ακαδημαϊκή

*Corresponding author.

κοινότητα, με τη μορφή ενός πολυμεσικού τριδιαλεκτικού λεξικού και μίας πολυτροπικής βάσης δεδομένων (Galiotou *et al*., 2014). Στην παρούσα εργασία παρουσιάζουμε το σύστημα αρχειοθέτησης και διαχείρισης γραπτών και προφορικών διαλεκτικών δεδομένων που αναπτύχθηκε στα πλαίσια του ερευνητικού προγράμματος. Πυρήνας του συστήματος είναι η πολυτροπική βάση δεδομένων η οποία η οποία επιτρέπει την παράλληλη εμφάνιση πρωτογενών και επεξεργασμένων δεδομένων καθώς και την κωδικοποίηση μεγάλου αριθμού πληροφοριών μεταδεδομένων. Πρωτογενή ονομάζονται δεδομένα όπως οι βιντεοσκοπήσεις και ηχογραφήσεις φυσικών ομιλητών καθώς και οι φωτογραφίες παλαιών βιβλίων και χειρογράφων. Επεξεργασμένα δεδομένα θεωρούνται οι κωδικοποιήσεις των πρωτογενών δεδομένων, όπως οι μεταγραφές και οι επισημειώσεις. Τέλος, τα μεταδεδομένα περιλαμβάνουν πληροφορίες σχετικές με πεδία και δομές δεδομένων.

2. ΨΗΦΙΑΚΟ ΣΩΜΑ ΓΡΑΠΤΩΝ ΚΑΙ ΠΡΟΦΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

2.1 Γραπτά δεδομένα

Η δημιουργία του ψηφιακού σώματος γραπτών δεδομένων έγινε σε 4 στάδια (Κολιοπούλου *et al.* 2015):

1. Εντοπισμός, συλλογή και καταλογογράφηση κειμένων πρωτογενών πηγών (χειρογράφων και δημοσιευμένων κειμένων) καθώς και δευτερογενών πηγών (κειμένων με μεταγλωσσικές πληροφορίες)
2. Ψηφιοποίηση μεγάλου μέρους του υλικού
3. Μεταγραφή μέρους του ψηφιοποιημένου σώματος κειμένων σε επεξεργάσιμη μορφή
4. Επισημείωση μέρους του μεταγεγραμμένου υλικού με κριτήρια επιλογής τη χρονολόγηση και την προέλευση του κειμένου και την αξιοπιστία των γλωσσικών και μεταγλωσσικών πληροφοριών

Συλλογή και καταλογογράφηση: Αρχικά δημιουργήθηκε ένας βιβλιογραφικός κατάλογος πρωτογενών και δευτερογενών πηγών. Οι πρωτογενείς πηγές περιλαμβάνουν αφηγήσεις, μύθους, θρύλους, τραγούδια, λαογραφικές περιγραφές, αφηγήσεις ή μαρτυρίες γεγονότων. Οι δευτερογενείς πηγές περιλαμβάνουν γλωσσολογικές μελέτες, γραμματικές, λεξικά ή γλωσσάρια. Οι πηγές αυτές εντοπίστηκαν κυρίως στο *Κέντρο Μικρασιατικών Σπουδών*, στο *Σπουδαστήριο Λαογραφίας του Ε.Κ.Π.Α*. αλλά και διάσπαρτες σε άλλες βιβλιοθήκες και σπουδαστήρια ελληνικών και ξένων εκπαιδευτικών ιδρυμάτων και ερευνητικών κέντρων, καθώς και στις καταγραφές ποντιακών συλλόγων, όπως της *Επιτροπής Ποντιακών Μελετών*. Επίσης, χρησιμοποιήθηκαν περιοδικά όπως : *Αθηνά, Ελληνική Διαλεκτολογία, Λαογραφία, Λεξικογραφικόν Δελτίον, Μελέτες για την Ελληνική Γλώσσα, Μικρασιατικά Χρονικά και Νεοελληνική Διαλεκτολογία*. Δημιουργήθηκε μία βιβλιογραφική βάση που περιλαμβάνει πάνω από 1250 βιβλιογραφικές αναφορές. Το μεγαλύτερο ποσοστό των πηγών αφορούν τα Ποντιακά ενώ η εκπροσώπηση των Αϊβαλιώτικων είναι εξαιρετικά περιορισμένη. Προφανώς η ανισοκατανομή αυτή εμφανίζεται και στο ψηφιακό σώμα των διαλέκτων.

Ψηφιοποίηση: Στη συνέχεια, έγινε η επιλογή των κειμένων προς ψηφιοποίηση με στόχο τη δημιουργία ψηφιακού σώματος κειμένων έκτασης 2.000.000 λέξεων, με βάση συγκεκριμένα κριτήρια: Ο πρώτος περιορισμός που τέθηκε ήταν η νομιμότητα της ψηφιοποίησης δεδομένης της προστασίας των πνευματικών δικαιωμάτων. Συνεπώς, τα κείμενα που επελέγησαν είχαν εκδοθεί πριν από το 1940. Ωστόσο, η αξία του ψηφιακού σώματος κειμένων δεν μειώθηκε διότι, η σημασία των πηγών πριν από το 1922, δηλαδή πριν τη μετεγκατάσταση των πληθυσμών στον Ελλαδικό χώρο, είναι πρωταρχική για τη μελέτη των μικρασιατικών διαλέκτων. Επιπλέον, στο δείγμα συμπεριελήφθησαν και νεότερα κείμενα, όπως του Κέντρου Μικρασιατικών Σπουδών το οποίο παρεχώρησε άδεια ψηφιοποίησης του υλικού αυτού στα πλαίσια της συμμετοχής του Κέντρου στο ερευνητικό πρόγραμμα. Όσον αφορά τα είδη των κειμένων, ψηφιοποιήθηκαν κείμενα πεζού λόγου. Το δείγμα δεν περιέχει τραγούδια ή ποιήματα διότι, αφενός, εμφανίζουν ιδιαίτερα γλωσσικά φαινόμενα που σχετίζονται άμεσα με τον έμμετρο λόγο και αφετέρου, η καταγραφή τους φέρει συνήθως ανεπαρκή στοιχεία προέλευσης. Δημιουργήθηκε επίσης αντιπροσωπευτικό δείγμα από χειρόγραφα με βασικά κριτήρια τη σπανιότητα, την παλαιότητα και το είδος του κειμένου που διασώζουν. Επελέγησαν κυρίως κείμενα από σπάνιες ή προσωπικές εκδόσεις, αλλά και γνωστά και σημαντικά εγχειρίδια για τη μελέτη των συγκεκριμένων διαλέκτων. Η διαδικασία της ψηφιοποίησης έγινε με φορητό σαρωτή έτσι ώστε να διασφαλιστεί η ποσότητα και η ποιότητα του αποτελέσματος και ταυτόχρονα να αποφευχθεί η αλλοίωση της ποιότητας των κειμένων. Τέλος, οι ψηφιακές εικόνες που δημιουργήθηκαν έτυχαν περαιτέρω επεξεργασίας έτσι ώστε να έχουν ομοιόμορφη εμφάνιση στο τελικό σώμα κειμένων.

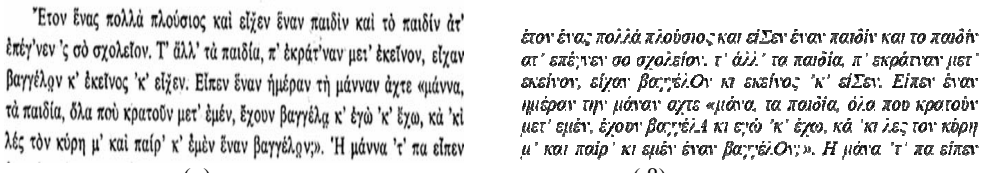
Μεταγραφή: Από το σύνολο του ψηφιοποιημένου σώματος κειμένων, επελέγησαν κείμενα έκτασης 200.000 λέξεων τα οποία μεταγράφηκαν χωρίς να γίνει χρήση λογισμικού οπτικής αναγνώρισης χαρακτήρων (Optical Character Recognition - OCR) ή αναγνώρισης χειρογράφων κειμένων (Hand-written Text Recognition - HCR). Αφενός, ένα λογισμικό OCR θα αντιμετώπιζε δυσκολίες αναγνώρισης του πολυτονικού συστήματος και αναγνώρισης χειρόγραφου κειμένου. Αφετέρου, η εκπαίδευση ενός λογισμικού HTR θα ήταν χρονικά απαγορευτική λόγω της μεγάλης ανομοιογένειας του υπό ψηφιοποίηση υλικού (πολλοί διαφορετικοί γραφικοί χαρακτήρες, πολλοί διαφορετικοί ειδικοί συμβολισμοί, πολλές διαφορετικές ποιότητες χαρτιού και μελανιού). Τα κείμενα μεταγράφηκαν με το ελληνικό αλφάβητο και την καθιερωμένη ιστορική ορθογραφία, ενώ οι διαλεκτικές ιδιαιτερότητες στο φωνητικό/φωνολογικό επίπεδο αποδόθηκαν με χρήση κεφαλαίων ελληνικών χαρακτήρων και ελάχιστων λατινικών μόνο όταν σημειώνονταν φωνολογικές πληροφορίες από τους εκδότες ή τους συγγραφείς. Το συγκεκριμένο σύστημα μεταγραφής είναι μία πρωτότυπη δημιουργία της ερευνητικής ομάδας του AMiGre και είναι προσαρμοσμένο στα δεδομένα και τους σκοπούς του ερευνητικού προγράμματος. Με την προσέγγιση αυτή, δημιουργήθηκε ένα σώμα κειμένων προερχόμενων και από τις τρεις διαλέκτους, με ενιαίο σύστημα μεταγραφής ανεξαρτήτως της πηγής που διευκολύνει τη τις αναζητήσεις κάθε είδους γλωσσικών φαινομένων. Η Ποντιακή και η Καππαδοκική διάλεκτος εκπροσωπούνται από περίπου 95.000 λέξεις η κάθε μία ενώ η Αϊβαλιώτικη από περίπου 10.000 λέξεις. Η διαφορά αυτή στην εκπροσώπηση των διαλέκτων οφείλεται στο γεγονός ότι οι γραπτές πηγές των Αϊβαλιώτικων είναι πολύ περιορισμένες σε αριθμό και έκταση. Για τον λόγο αυτό, ψηφιοποιήθηκε και ένας αριθμός προφορικών πηγών από τα Αϊβαλιώτικα, έτσι ώστε το δείγμα να αποκτήσει μία σχετική αντιπροσωπευτικότητα από όλες τις

διαλέκτους. Τόσο η φύση των κειμένων όσο και η έως τώρα έλλειψη συστηματικής μελέτης των διαλέκτων προκάλεσαν μία σειρά προβλημάτων στη διαδικασία της μεταγραφής. Για παράδειγμα, πολλά χειρόγραφα ήταν ιδιαιτέρως δυσανάγνωστα οπότε προτιμήθηκε η μεταγραφή περισσότερων δημοσιευμένων πηγών έτσι ώστε να ελαχιστοποιηθεί η πιθανότητα επισφαλών μεταγραφών. Το σημαντικότερο όμως πρόβλημα είναι η μη τυποποιημένη γραφή διαλέκτων η οποία οφείλεται στην τάση των ερευνητών να ακολουθούν ο καθένας προσωπικό σύστημα καταγραφής κυρίως σε σχέση με τις φωνολογικές ιδιαιτερότητες που δεν είναι δυνατόν να καταγραφούν με το καθιερωμένο σύστημα καταγραφής της ΚΝΕ. Για τον λόγο αυτό, δεν μεταγράφηκαν κείμενα τα οποία περιείχαν πλήθος αδιευκρίνιστων στοιχείων. Επίσης, η ομάδα του ερευνητικού μας προγράμματος δημιούργησε έναν πίνακα αντιστοίχισης των ιδιαίτερων συμβόλων που χρησιμοποιήθηκαν με τις φωνολογικές τους αξίες (βλ. Πίνακα 1). Ο πίνακας αυτός παρέχεται στους ερευνητές – χρήστες της πολυτροπικής βάσης δεδομένων έτσι ώστε να είναι απολύτως σαφείς οι φωνολογικές αντιστοιχίες των συμβόλων.

Διεθνές Φωνητικό Αλφάβητο	Συνήθη σύμβολα πηγών	AMiGre
æ	Ä	A
œ, ø	Ö	O
ɯ	I	I
ə	ə	E
y	Ü	Y
b, d, g	b, d, g	b, d, g ή μπ, ντ, γκ, γγ
mb, nd, ng	μb, νd, νg	μb, νd, νg
q	Q	q
ʎ	λ˙, λ̂	Λ
ɲ	ν˙	N
ʃ	σ˙, σ̂, χ˙, χ̂	Σ
ʒ	ζ˙, ζ̂	Z
tʃ	τσ˙, τσ̂	τΣ
dʒ	dζ˙, ντζ̂	dZ
c	κ˙, κ̂	K
ç	χ˙, χ̂	X
ɟ	γ˙, γ̂	Γ
ɟ	γκ˙, γκ̂	G
p^h t^h k^h	‘, p t k	πh, τh, κh

Πίνακας 1. Σύστημα μεταγραφής

Η τελική μορφή των μεταγεγραμμένων κειμένων (βλ. Εικόνα 1) μπορεί να φανεί χρήσιμη τόσο σε όσους ασχολούνται ερασιτεχνικά με τις συγκεκριμένες διαλέκτους όσο και τους ερευνητές. Οι πρώτοι διευκολύνονται στην ανάγνωση λόγω της χρήσης του απλού αλφαβήτου με προσθήκη λίγων συμβόλων ενώ οι δεύτεροι επωφελούνται από ένα σχετικά συντηρητικό σύστημα μεταγραφής το οποίο ωστόσο παρέχει όλες τις πληροφορίες που βρίσκονται στις κειμενικές πηγές.


Ἔτον ἕνας πολλὰ πλούσιος καὶ εἶχεν ἕναν παιδίν καὶ τὸ παιδίν ἀτ' ἐπέγ'νεν 'ς σὸ σχολεῖον. Τ' ἄλλ' τὰ παιδία, π' ἐκράτ'ναν μετ' ἐκεῖνον, εἶχαν βαγγέλ̧ν κ' ἐκεῖνος 'κ' εἶχεν. Εἶπεν ἕναν ἡμέραν τὴ μάνναν ἀχτε «μάννα, τὰ παιδία, ὅλα ποὺ κρατοῦν μετ' ἐμέν, ἔχουν βαγγέλ̧ κ' ἐγὼ 'κ' ἔχω, κὰ 'κὶ λές τὸν κύρη μ' καὶ παίρ' κ' ἐμὲν ἕναν βαγγέλ̧ν;». Ἡ μάννα 't' πα εἶπεν

(α)

ἔτον ἕνας πολλὰ πλοῦσιος καὶ εἴΣεν ἕναν παιδίν καὶ το παιδὶν ατ' επέ;νεν σο σχολείον. τ' ἀλλ' τα παιδία, π' εκράταν μετ' εκείνον, είχαν βα;;έλ.Ον κι εκείνος 'κ' είΣεν. Είπεν έναν ημέραν την μάναν αχτε «μάνα, τα παιδία, όλο που κρατούν μετ' εμέν, έχουν βα;;έλ.Α κι εγώ 'κ' έχω, κά 'κι λες τον κύρη μ' και παίρ' κι εμέν έναν βα;;έλ.Ον;». Η μάνα 't' πα είπεν

(β)

Εικόνα 1. (α) Δημοσιευμένη μορφή και (β) μεταγραφή αποσπάσματος ποντιακού κειμένου

Επισημείωση: Το τελευταίο στάδιο της δημιουργίας του ψηφιακού σώματος γραπτών κειμένων αφορούσε την επισημείωση μέρους των μεταγεγραμμένων κειμένων έκτασης 50.000 λέξεων. Για την επιλογή του υλικού προς επισημείωση ακολουθήθηκε η αρχή της αντιπροσωπευτικότητας όπως και στα προηγούμενα στάδια. Δεν επισημειώθηκαν κείμενα από την Αϊβαλιώτικη διάλεκτο λόγω της έλλειψης κειμένων στη διάλεκτο αυτή. Οπότε, η κατανομή των λέξεων στο επισημειωμένο σώμα κειμένων είναι 25.000 λέξεις για τα Ποντιακά και 25.000 λέξεις για τα Καππαδοκικά. Η επισημείωση έγινε σε αρχεία .xls από τα οποία τα δεδομένα εισήχθησαν στην πολυτροπική βάση. Η γλωσσολογική επισημείωση έγινε σε φωνολογικό και μορφολογικό επίπεδο ενώ προστίθενται πληροφορίες που αφορούν δάνειες λέξεις και αρχαϊσμούς. Όσον αφορά τη φωνολογία, η επισημείωση πραγματοποιήθηκε σε επίπεδο μονάδας (φωνήεντος-συμφώνου) ή συλλαβής εφόσον αφορούσε δύο τεμάχια. Στα πλαίσια αυτά δηλώνεται μία πλειάδα φαινομένων που χαρακτηρίζουν τις διαλέκτους και ανήκουν στις ακόλουθες ευρείες κατηγορίες: ανάπτυξη, αποβολή, αφομοίωση, ανομοίωση,ανύψωση, τσιτακισμός, εξασθένωση, ενίσχυση, διφθογγοποίηση, συνάρθρωση, μετάθεση, τροπή, φωνητικός αρχαϊσμός, δάνειο φώνημα. Για την επισημείωση του φαινομένου γίνεται ιστορική αναγωγή στον προγενέστερο γνωστό πρόγονο της διαλέκτου εφόσον είναι σχετικά γνωστή η ιστορική προέλευση του τύπου. Στις περισσότερες φορές η αναγωγή φτάνει μέχρι τη

Μεσαιωνική Ελληνική, ενώ σε κάποιες περιπτώσεις ως προγενέστερος γνωστός πρόγονος θεωρείται η Ελληνιστική Κοινή. Όσον αφορά τη μορφολογία, η επισημείωση γίνεται στο επίπεδο λέξης. Παρέχει πλήρη γραμματική αναγνώριση και δίνει έμφαση σε πληροφορίες που αφορούν την κλίση και την παραγωγή. Σ΄ αυτό το επίπεδο της μορφολογικής επισημείωσης παρέχονται συγχρονικές πληροφορίες για την πληρέστερη περιγραφή των διαλέκτων οπότε δεν υπάρχει η ιστορική διάσταση που επιβάλλει η φωνολογική επισημείωση. Αξίζει να σημειωθεί ότι η πλήρης φωνολογική και μορφολογική επισημείωση συνάντησε σημαντικές δυσκολίες λόγω της μη συστηματικής προγενέστερης μελέτης των διαλέκτων - κυρίως των Καππαδοκικών – και της έλλειψης βασικών βοηθημάτων (λεξικά, γραμματικές).

2.2 Προφορικά δεδομένα

Όσον αφορά τα δεδομένα προφορικών πηγών, η βάση περιλαμβάνει, ψηφιοποιημένες ηχογραφήσεις φυσικών ομιλητών, ορθογραφική μεταγραφή του 1/3 των ηχογραφημένων συνομιλιών και μετάφρασή τους, μορφολογική επισημείωση της παραγωγής, της σύνθεσης και της κλίσης, καθώς και φωνητική/φωνολογική επισημείωση επιτονικών φράσεων, λέξεων, συλλαβών και φθόγγων (Παπαζαχαρίου & Καρασίμος 2015). Έγιναν ηχογραφήσεις διάρκειας περίπου 180 ωρών (δηλαδή 60 ώρες ανά διάλεκτο) με συσκευές ψηφιακής ηχογράφησης υψηλής ευκρίνειας, σε όσον το δυνατόν πιο ήσυχες συνθήκες, συνήθως στα σπίτια των πληροφορητών και πάντα με την προηγούμενη συναίνεσή τους. Οι ερευνητές πεδίου ήταν φυσικοί ομιλητές της διαλέκτου υπό μελέτη, ικανοί να εφαρμόσουν εθνογραφικές μεθόδους συλλογής δεδομένων και να ηχογραφούν φυσικές και αβίαστες καθημερινές ομιλίες. Στα πλαίσια του εφικτού, οι ομιλητές επελέγησαν ώστε να έχουν καθαρή άρθρωση, φυσική ροή ομιλίας και να κάνουν συστηματική χρήση της διαλέκτου στην καθημερινότητά τους. Επίσης, στις περισσότερες περιπτώσεις, ήταν απαραίτητη η ύπαρξη του ενδιάμεσου στις ηχογραφήσεις, ώστε οι ομιλητές να αισθάνονται πιο οικεία κατά τη διάρκεια της ηχογράφησης και να ελαχιστοποιηθούν τα σημεία διαλόγου όπου θα γινόταν αλλαγή γλωσσικού συστήματος επικοινωνίας (εγκατάλειψη της διαλέκτου και χρήση της Κοινής Νέας Ελληνικής).

Τα πρωτογενή δεδομένα συνοδεύονται από την ορθογραφική μεταγραφή ενός μέρους τους (όσον αφορά τα Ποντιακά και τα Καππαδοκικά) καθώς και από τη μετάφρασή τους. Η μεταγραφή και η μετάφραση των συνομιλιών για τα Καππαδοκικά και τα Ποντιακά έγινε από τους ερευνητές πεδίου, ενώ για τα Αϊβαλιώτικα έγινε από φυσικούς ομιλητές της λεσβιακής διαλέκτου η οποία είναι πολύ κοντινή με τα Αϊβαλιώτικα. Η ορθογραφική μεταγραφή η οποία οριοθετείται από τις συνεισφορές του κάθε ομιλητή και προσδιορίζεται από τα όρια των επιτονικών φράσεων που απαρτίζουν κάθε συνεισφορά έγινε με τη βοήθεια του λογισμικού Praat (Boersma & Weenink 2017), το οποίο επιτρέπει την παράλληλη εμφάνιση ηχητικού αρχείου και λωρίδων κειμένου για μεταγραφή και επισημείωση. Όσον αφορά την επισημείωση των προφορικών πηγών, στις ηχογραφήσεις που μεταγράφηκαν και μεταφράστηκαν προσδιορίστηκαν: οι συνεισφορές κάθε ομιλητή, οι επιτονικές φράσεις, οι λέξεις, οι συλλαβές και, τέλος, οι φθόγγοι.

Τα μεταδεδομένα αφορούν τη διάλεκτο, τους πληροφορητές και την επικοινωνιακή περίσταση. Πιο συγκεκριμένα, ως προς τη διάλεκτο, παρέχονται πληροφορίες σχετικά με το όνομα της διαλέκτου, τον τόπο προέλευσης του πληροφορητή, τον τόπο ηχογράφησης και την πληθυσμιακή οργάνωση του τόπου ηχογράφησης (π.χ., μεικτή ή ομοιογενής). Ως προς τους πληροφορητές, παρέχονται πληροφορίες σχετικά το φύλο, την ηλικία, το μορφωτικό επίπεδο, την ομάδα καταγωγής, το καθεστώς ομάδας καταγωγής στον τόπο ηχογράφησης, το είδος γειτονιάς και καθημερινές σχέσεις ανάμεσα στους πληροφορητές και στους συγγενείς τους, τους γείτονες και τους συναδέλφους τους. Τέλος, ως προς την επικοινωνιακή περίσταση, παρέχονται πληροφορίες σχετικά με τον αριθμό των συμμετεχόντων στην ηχογράφηση, την κοινωνική τους σχέση, καθώς και το είδος της ηχογράφησης (όπως φιλική συνομιλία μεταξύ γνωστών, τυπική συνομιλία μεταξύ αγνώστων, συνέντευξη, προφορικό ερωτηματολόγιο, κ.λπ.)

3. ΣΥΣΤΗΜΑ ΑΡΧΕΙΟΘΕΤΗΣΗΣ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗΣ ΔΙΑΛΕΚΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

3.1 Διεθνής και Ελληνική εμπειρία

Για την ανάδειξη και επεξεργασία των διαλεκτικών δεδομένα που αποθησαυρίσθηκαν από το Εργαστήριο Νεοελληνικών Διαλέκτων ήταν απαραίτητη η αποθήκευσή τους σε ηλεκτρονικό μέσον και η χρήση κατάλληλου λογισμικού για την αναζήτηση στο αποθηκευμένο υλικό. Ενδεικτικά αναφέρουμε αντίστοιχες περιπτώσεις από τη διεθνή εμπειρία: 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 & Kleiweg 2003). Σε μια άλλη εργασία (Ubul et al. 2015), περιγράφεται η προσπάθεια δημιουργίας μιας πολυμεσικής βάσης με σκοπό τη μελέτη των υπό εξαφάνιση διαλέκτων των περιοχών της Ιαπωνίας. Σε ευρωπαϊκό επίπεδο, ενδεικτικά αναφέρουμε το DynaSAND (Dynamic Syntactic Atlas of Dutch Dialects), ένα online εργαλείο για την επεξεργασία των ολλανδικών συντακτικών ποικιλιών (Barbiers et al. 2006). Το τελευταίο διαθέτει μια διαδικτυακή υπηρεσία (web service) για την αξιοποίηση του σώματος από άλλες εφαρμογές (Kunst & Wesseling 2010). Επίσης, δεδομένα καταλανικών διαλέκτων αποθησαυρίζονται στο ηλεκτρονικό σώμα COD (Corpus Oral Dialectal) (Clua & Lloret 2006). Τέλος, ένα έργο που έχει σε μεγάλο βαθμό παρόμοιους στόχους με το έργο μας είναι το SCOTS (Scottish Corpus of Text and Speech) καθώς αποτελεί ένα μεγάλης έκτασης ηλεκτρονικό σώμα γραπτών και προφορικών τεκμηρίων των γλωσσών της Σκοτίας (Anderson et al. 2007). Σχετικά με τις ελληνικές διαλέκτους, η μόνη προγενέστερη της δικής μας απόπειρα, αφορά τη δημιουργία της βάσης Gree.D (Ράλλη *et al*., 2010) η οποία εμπλουτίζεται συνεχώς με προφορικά δεδομένα που συλλέγονται από το Εργαστήριο Νεοελληνικών Διαλέκτων του Πανεπιστημίου Πατρών. Υπάρχουν επίσης αρκετά εργαλεία λογισμικού, που διευκολύνουν την επισημείωση σε διαφορετικά επίπεδα. Στη συνέχεια αναφέρουμε ορισμένα από αυτά: Το Praat (Boersma & Weenink 2017) είναι ένα δωρεάν διαθέσιμο λογισμικό ανάλυσης και επεξεργασίας ακουστικών σημάτων και ήχων το οποίο βασίζεται σε ψηφιακές ηχητικές καταγραφές. Το ELAN - EUDICO Linguistic Annotator- (Sloetjes & Wittenburg 2008) είναι λογισμικό που επιτρέπει την εισαγωγή επισημειώσεων και σχολιασμών σε αρχεία ψηφιακού ήχου και βίντεο. Οι επισημειώσεις μπορούν να γίνονται σε πολλά επίπεδα. Το LaBB-CAT (Fromont & Hay 2008) είναι ένα πρόγραμμα περιήγησης που βασίζεται σε γλωσσολογικά εργαλεία και αποθηκεύει ηχογραφήσεις ή βιντεοσκοπήσεις, μεταγραφές κειμένου και άλλες επισημειώσεις. Το (Field Linguist's) Toolbox (Buseman & Buseman 2007) είναι ένα εργαλείο διαχείρισης και ανάλυσης δεδομένων στο γλωσσολογικό τομέα. Είναι ιδιαίτερα χρήσιμο για τη διατήρηση λεξικολογικών δεδομένων, για τη μορφολογική ανάλυση κείμενων, αλλά μπορεί να χρησιμοποιηθεί και για τη διαχείριση

διαφόρων ειδών δεδομένων. Παρόλες τις δυνατότητες που έχουν, τα ανωτέρω λογισμικά δεν είναι σε θέση να ανταποκριθούν στις απαιτήσεις του προγράμματός μας οι οποίες συνοψίζονται στα εξής:

(i) Επισημειώσεις σε πολλά διαφορετικά γλωσσολογικά επίπεδα.

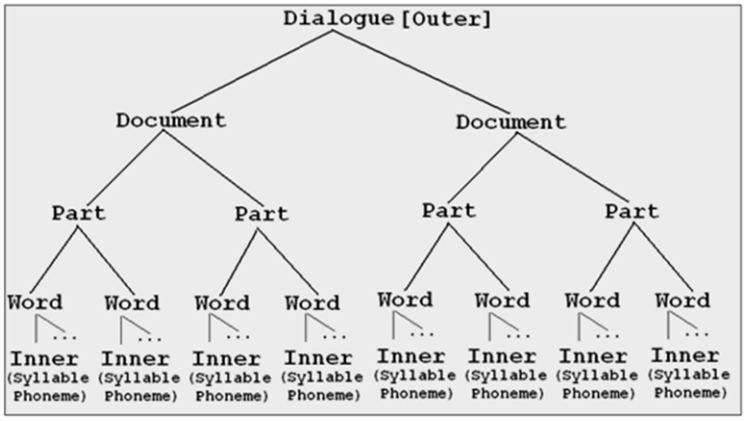
(ii) Συνδυασμένη αναζήτηση σε διαφορετικά επίπεδα αναπαράστασης

(iii) Συνδυασμένη αναζήτηση σε τεκμήρια τόσο γραπτών όσο και προφορικών πηγών.

3.2 Σχεδίαση και ανάπτυξη του συστήματος αρχειοθέτησης και διαχείρισης γραπτών και προφορικών πηγών

3.2.1 Αναπαράσταση δεδομένων

Η απαίτηση για συνδυασμένη αναζήτηση στα γραπτά και προφορικά δεδομένα, δημιούργησε την πρόκληση για ενοποιημένη αναπαράσταση δεδομένων διαφορετικών τύπων και δομών (Καρανικόλας *et al.*, 2015). Τα τεκμήρια των γραπτών μας πηγών περιλαμβάνουν βιβλία, άρθρα, δακτυλογραφημένα και χειρόγραφα κείμενα τα οποία απαρτίζονται από σελίδες που περιλαμβάνουν μορφολογικές λέξεις. Οι επισημειώσεις γίνονται σε 3 επίπεδα (κείμενο, σελίδα, λέξη). Τα τεκμήρια των προφορικών πηγών περιέχουν ηχογραφήσεις ενός ομιλητή ή μιας μικρής ομάδας συνομιλητών. Για κάθε ομιλητή διατίθεται η επισημείωση της ομιλίας του, υποδιαιρούμενη σε εναλλαγές (turn takings). Μία εναλλαγή αντιστοιχεί σε ένα εκφώνημα του ομιλητή (επιτονική πρόταση) και υποδιαιρείται σε μορφολογικές λέξεις. Με τη σειρά τους, οι μορφολογικές λέξεις υποδιαιρούνται σε συλλαβές και οι συλλαβές σε φωνήματα (φωνήεντα ή σύμφωνα). Συνεπώς τα τεκμήρια προφορικών πηγών επισημειώνονται σε περισσότερα από 3 επίπεδα. Οι δύο δομές των γραπτών και προφορικών τεκμηρίων ενοποιήθηκαν σε μία ενιαία γενική δομή η οποία αντιστοιχεί τόσο στο γραπτό όσο και το προφορικό τεκμήριο. Στα πλαίσια αυτά έγινα οι εξής αντιστοιχίσεις: Το επίπεδο **Dialogue (Outer)** αντιπροσωπεύει το **διάλογο** μεταξύ ομιλητών στα προφορικά τεκμήρια. Το επίπεδο **Document** αντιστοιχεί στον **ομιλητή** και στο **γραπτό τεκμήριο**. Το επίπεδο **Part** αντιστοιχεί στο **εκφώνημα του ομιλητή** και στη **σελίδα γραπτού τεκμήριου**. Το επίπεδο **Word** αντιστοιχεί στις **μορφολογικές λέξεις** και στις δύο ομάδες τεκμηρίων. Τέλος, το κατώτερο επίπεδο **Inner** αντιστοιχεί στα τμήματα στα οποία επιμερίζεται μία λέξη όπως **συλλαβές, φωνήματα,** κλπ στα προφορικά τεκμήρια. Στην εικόνα 2 απεικονίζεται η ιεραρχική δομή που αντιστοιχεί στον γενικό τύπο τεκμηρίου. Εκτός από το γλωσσολογικό ενδιαφέρον, τα διαλεκτικά δεδομένα παρουσιάζουν και ιδιαίτερο ενδιαφέρον ως προς την αναπαράσταση και επεξεργασία τους σε ηλεκτρονικά μέσα λόγω της ποικιλίας των μορφών και τύπων αρχείων προς επεξεργασία από το σύστημα αρχειοθέτησης και επεξεργασίας. Όσον αφορά τα προφορικά τεκμήρια, το σύστημα διατηρεί: ψηφιακές ηχογραφήσεις συνήθως σε αρχεία WAV, αρχικές επισημειώσεις συνήθως σε αρχεία TextGrid (αρχεία εξόδου του Praat) και υπολογιστικά επεξεργάσιμες επισημειώσεις για όλα τα επίπεδα που υποστηρίζονται. Πιο συγκεκριμένα: οι υπολογιστικά επεξεργάσιμες επισημειώσεις ανά επίπεδο αναπαράστασης είναι: μεταδεδομένα ομιλητή (Document), επισημειώσεις εκφωνημάτων όπως ορθογραφική μεταγραφή, μετάφραση στην ΚΝΕ (Part), επισημειώσεις μορφολογικών λέξεων όπως ορθογραφική μεταγραφή, μεταγραφή και φωνολογικά φαινόμενα της λέξης (Word) και επισημειώσεις σε επιμέρους στοιχεία μίας λέξης ,όπως συλλαβές, θέση και τόνος φωνηέντων, σύμφωνα (Inner).



Εικόνα 2: Αφηρημένη ιεραρχική δομή για την κάλυψη όλων των τεκμηρίων

Όσον αφορά τα γραπτά τεκμήρια, το σύστημα διατηρεί: ψηφιοποιημένες σελίδες από τα πρωτότυπα, συνήθως αρχεία JPG, μεταγραφές που ομοιογενοποιούν τα σύμβολα από τα πρωτότυπα τεκμήρια (αρχεία κειμένου) και υπολογιστικά επεξεργάσιμες επισημειώσεις για όλα τα επίπεδα που υποστηρίζονται. Πιο συγκεκριμένα, οι υπολογιστικά επεξεργάσιμες επισημειώσεις ανά επίπεδο αναπαράστασης είναι: μεταδεδομένα τεκμηρίου (Document), επισημειώσεις σελίδων (Part), μορφολογικές επισημειώσεις (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). Το σύστημα παρέχει επίσης τη δυνατότητα προσθήκης συντακτικών και σημασιολογικών επισημειώσεων εάν υπάρξει ανάγκη στο μέλλον. Με βάση τα παραπάνω δημιουργήθηκαν τέσσερις συλλογές αρχείων δεδομένων (Karanikolas *et al*., 2014): ψηφιακές ηχογραφήσεις προφορικών δεδομένων (αρχεία WAV), αρχικές επισημειώσεις προφορικών (αρχεία TextGrid), ψηφιοποιημένες σελίδες από τα πρωτότυπα γραπτά (αρχεία Image), μεταγραφές σελίδων των γραπτών. Τα υπολογιστικά επεξεργάσιμα στοιχεία αποθηκεύονται σε τρεις βάσεις δεδομένων: (α) Βάση δεδομένων "Struct" : σύνολο πινάκων που υλοποιούν την αφηρημένη ιεραρχική δομή που καλύπτει γραπτά και προφορικά τεκμήρια. (β) Βάση δεδομένων EAV: καλύπτει τις επισημειώσεις. Λόγω το ότι δεν ήταν δυνατόν να προσδιοριστούν εξαρχής όλες οι οντότητες και οι ιδιότητές τους, έγινε εισαγωγή ενός σχήματος EAV (Entity-Attribute-Value) (Anhøj 2003). Σ΄αυτό έγιναν επεκτάσεις ώστε να υποστηρίζει ελεύθερα και

καθορισμένα σύνολα τιμών (λεξιλόγια), πολλαπλές τιμές ιδιοτήτων και εξαρτήσεις εμφάνισης ιδιοτήτων. (γ) Βάση δεδομένων Inner: ακολουθεί επίσης το σχήμα EAV και αποθηκεύει επιμέρους τμήματα λέξεων και τις αντίστοιχες επισημειώσεις.

3.2.2 Επισκόπηση των εφαρμογών

Το σύστημα διαχείρισης δεδομένων γραπτών και προφορικών πηγών έχει υλοποιηθεί γύρω από δύο βασικά υποπρογράμματα , ένα για τα γραπτά (G.Written) και ένα για τα προφορικά (G.Oral) τεκμήρια. (Το πρόθημα G. αντιστοιχεί στο ακρωνύμιο GUI -Graphical User Interface). Τα δύο υποσυστήματα σχεδιάσθηκαν και υλοποιήθηκαν ως ένα τρίπτυχο (Part – Words - Επισημειώσεις). Παρέχονται χειριστήρια για την πλοήγηση μεταξύ των Parts, με ενέργειες οι οποίες εξαρτώνται από τη μορφή του τεκμηρίου (Καρανικόλας *et al*., 2015). Οι επιμέρους ενότητες λογισμικού που λειτουργούν σε ένα τεκμήριο είναι **:**

- Ο φωνολογικός επισημειωτής (Ph. Tagger) εφαρμόζεται σε γραπτά και προφορικά τεκμήρια.
- Ο μορφολογικός επισημειωτής (Morph. Tagger**)** εφαρμόζεται σε γραπτά και προφορικά τεκμήρια.
- Ο συντακτικός επισημειωτής (Syn. Tagger**)** αποδίδει συντακτικές κατηγορίες σε μία ακολουθία λέξεων.
- Ο σημασιολογικός επισημειωτής (Sem. Tagger) αποδίδει σημασιολογικές πληροφορίες σε μία ακολουθία λέξεων

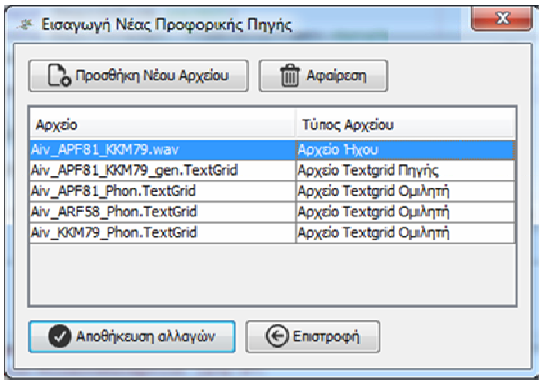
Η εισαγωγή των γραπτών τεκμηρίων είναι μία βηματική (σελίδα-σελίδα) διαδικασία Αυτή η βασική ενότητα λογισμικού χρησιμοποιεί άλλες δύο ενότητες για να ενσωματώσει την ψηφιοποιημένη σελίδα και να εισάγει την ομοιογενοποιημένη μεταγραφή της σελίδας. Όσον αφορά τα προφορικά τεκμήρια γίνεται μαζική εισαγωγή: Ο χρήστης προσδιορίζει μεσω της διεπαφής όλα τα αρχεία TextGrid (αποτελέσματα επεξεργασίας με το Praat) των ομιλητών και τις ψηφιακές ηχογραφήσεις και μέσω της κατάλληλης διαδικασίας ελέγχου γίνεται η ένταξη του προφορικού τεκμηρίου στο σύστημα. Για τη διαχείριση των μεταδεδομένων γραπτών πηγών (τεκμήριο) και προφορικών πηγών (ομιλητής – ηλικία, φύλο, καταγωγή κλπ) απαιτούνται δύο ενότητες λογισμικού αντίστοιχα. Τέλος, υπάρχουν ξεχωριστές ενότητες λογισμικού για περιήγηση στα γραπτά τεκμήρια, περιήγηση στα προφορικά καθώς και για την αναζήτηση και εμφάνιση των τεκμηρίων. Αναπτύχθηκαν επίσης, δύο ακόμη ενότητες λογισμικού που χρησιμοποιήθηκαν κατά την αρχική λειτουργία του συστήματος ως υπηρεσίες για την μαζική εισαγωγή γραπτών και προφορικών πηγών

4. ΥΠΟΣΥΣΤΗΜΑ ΕΙΣΑΓΩΓΗΣ & ΕΠΕΞΕΡΓΑΣΙΑΣ ΔΕΔΟΜΕΝΩΝ

4.1 Εισαγωγή δεδομένων

Στη συνέχεια, παρουσιάζουμε με συντομία τη διαδικασία εισαγωγής γραπτών και προφορικών τεκμηρίων στο σύστημα με χρήση διεπαφής χρήστη. Όσον αφορά τα γραπτά τεκμήρια, πρόκειται για μία βηματική διαδικασία η οποία συνοψίζεται στα εξής: (α) μιουργία κενού γραπτού τεκμηρίου. (β) Προσθήκη σελίδων γραπτού τεκμηρίου μία-μία. (γ) για κάθε σελίδα εισαγωγή τόσο της ψηφιοποιημένης της μορφής (Text Imaging) όσο και της μεταγραφής της (Text Transcription). Οι επισημειώσεις (μορφολογικές, φωνολογικές, κλπ) γίνονται από το χρήστη σε επόμενο στάδιο επεξεργασίας των δεδομένων που περιγράφουμε στην επόμενη παράγραφο. Στην εικόνα 3 φαίνεται ένα παράδειγμα εισαγωγής μεταγεγραμμένου κειμένου και καθορισμού των συμβόλων διαχωρισμού του σε λέξεις. Σε αντίθεση με τις γραπτές πηγές, η εισαγωγή των προφορικών πηγών δεν γίνεται βήμα-βήμα. Ο χρήστης επιλέγει όλα τα αρχεία μεταγραφών (TextGrid) και τα ηχητικά αρχεία (Wav) που συνθέτουν το προφορικό τεκμήριο προς εισαγωγή. Στην εικόνα 4 φαίνεται ένα τέτοιο παράθυρο προσδιορισμού των μεταγραφών (TextGrid) και των ηχητικών αρχείων που συνθέτουν ένα προφορικό τεκμήριο. Πρόκειται για ένα διάλογο τριών ομιλητών στα Αϊβαλιώτικα (και ισάριθμες μεταγραφές, μία ανά ομιλητή), μία ηχητική ψηφιακή καταγραφή της συνομιλίας και μία γενική μεταγραφή ανεξάρτητη ομιλητή. Αξίζει να σημειωθεί ότι οι μεταγραφές και οι επισημειώσεις βρίσκονται στα TextGrid αρχεία και αφού περάσουν από ένα βήμα ελέγχου ενσωματώνονται στις υπολογιστικά οργανωμένες και διαχειριζόμενες πληροφορίες και είναι διαθέσιμες για αναζήτηση από το χρήστη. Οι συμπληρωματικές επισημειώσεις (μορφολογικές, φωνολογικές, κλπ) γίνονται από το χρήστη σε επόμενο στάδιο επεξεργασίας των δεδομένων που περιγράφουμε στη συνέχεια. Εκτός από τη διαδικασία εισαγωγής τεκμηρίων σε αλληλεπίδραση με τον χρήστη, το σύστημα παρέχει τη δυνατότητα μαζικής εισαγωγής του συνόλου των αρχείων που περιέχουν μεταγραφές και επισημειώσεις.

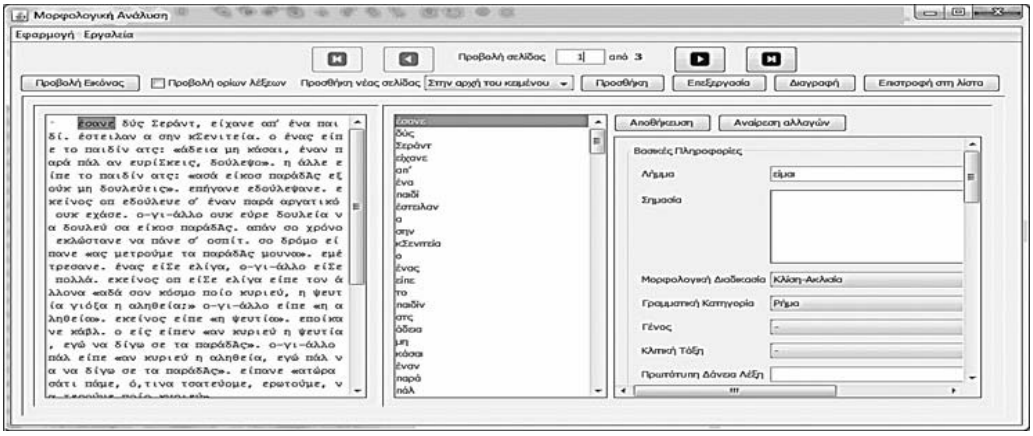
Εικόνα 3: Διαδικασία καθορισμού συμβόλων διαχωρισμού σε λέξεις ενός μεταγεγραμμένου κειμένου.



Εικόνα 4: Εισαγωγή προφορικού τεκμηρίου

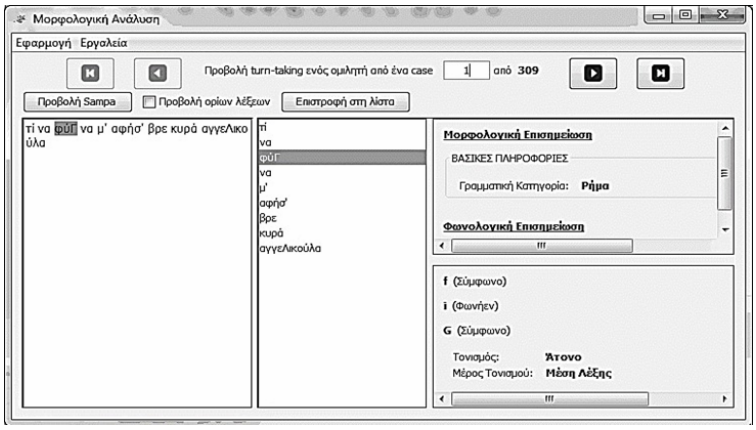
4.2 Επεξεργασία δεδομένων

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 επιλεγμένης λέξης. Το τρίπτυχο παρέχει χειριστήρια για τη μετακίνηση μεταξύ των σελίδων του γραπτού τεκμηρίου. Το G.Written παρέχει, επίσης, χειριστήριο για την προσθήκη σελίδας γραπτού τεκμηρίου. Στην Εικόνα 5 βλέπουμε ένα στιγμιότυπο του τριπτύχου γραπτών τεκμηρίων όπου εμφανίζονται μόνο οι ιδιότητες (επισημειώσεις) της λέξης που έχουν τιμές. Το σύνολο των ιδιοτήτων εμφανίζεται μόνο όταν κληθεί η φόρμα ενημέρωσης.



Εικόνα 5: Τρίπτυχο γραπτών τεκμηρίων

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 και δεξιά εμφανίζονται οι επισημειώσεις μίας επιλεγμένης λέξης. Το τρίπτυχο παρέχει χειριστήρια για τη μετακίνηση μεταξύ των εκφωνημάτων του ομιλητή του προφορικού τεκμηρίου. Στην εικόνα 6 βλέπουμε το τρίπτυχο προφορικών τεκμηρίων:



Εικόνα 6: Τρίπτυχο προφορικών τεκμηρίων

5. ΑΝΑΖΗΤΗΣΗ ΔΕΔΟΜΕΝΩΝ

Οι απαιτήσεις που τέθηκαν για το υποσύστημα αναζήτησης πληροφορίας προσδιορίζονται επακριβώς στην επόμενη λίστα:

- Διαισθητική χρήση.
- Υποστήριξη ερωτημάτων που λαμβάνουν υπ' όψη την ύπαρξη πολλαπλών τιμών για τα πεδία επισημειώσεων. (Karanikolas & Skourlas 2014).
- Δύο τύποι περιορισμών για κάθε κριτήριο: περιορισμοί τιμών, συνθήκες απόστασης
- Πολλαπλά κριτήρια και σύζευξη κριτηρίων
- Κάθε κριτήριο να εστιάζει σε κάποιο από τα επίπεδα (Document, Part, Word, Inner).
- Έκφραση απαιτήσεων ανάκτησης για: πραγματικά δεδομένα, σύνοψη δεδομένων (data aggregations), τεχνητά δεδομένα (artifacts-on the fly created data).

Όπως αναφέρθηκε, στο ίδιο ερώτημα αναζήτησης μπορούν να συνδυαστούν περισσότερα από ένα κριτήρια. Στον *Πίνακα 2* απεικονίζεται η δομή ενός ερωτήματος.

Word/Token/Phenomenon			**Location**				
<Value>	{Between, And, Or, Exact}	<Value>	<EAV subschema> (επισημειώσεις / μεταπληροφορίες)	<Attribute>	<Part distances>	<Word distances>	<Interval_no distances>

Πίνακας 2: Δομή ερωτήματος

Εκτός των κριτηρίων αναζήτησης, η διεπαφή προσδιορίζει και ποιες πληροφορίες εμφανίζονται ως απάντηση σε ένα ερώτημα. Στον *Πίνακα 3* βλέπουμε τη δομή του επιθυμητού αποτελέσματος. 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.

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

Πίνακας 3: Δομή επιθυμητού αποτελέσματος

Στον *Πίνακα 4*, παρουσιάζουμε ένα στιγμιότυπο διεπαφής της ενότητας λογισμικού *Αναζήτησης*. Η απεικονιζόμενη διεπαφή, προσδιορίζει ότι αναζητούνται Documents (ομιλητές που συμμετείχαν σε διαλόγους στην περίπτωση προφορικών πηγών). Για την εύρεσή τους προσδιορίζονται μεταδεδομένα των ομιλητών (φύλο, ηλικία και καταγωγή), καθώς και μεταδεδομένα (ονοματεπώνυμο) του εποπτεύοντα μελετητή ή επισημειωτή του διαλόγου.

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

Output	Document	-

Πίνακας 4: Διεπαφή Αναζήτησης

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή παρουσιάσαμε τη σχεδίαση και ανάπτυξη ενός συστήματος δημιουργίας και διαχείρισης σώματος τεκμηρίων από γραπτές και προφορικές πηγές τριών Ελληνικών διαλέκτων της Μικράς Ασίας. Το σύστημα που αναπτύχθηκε με πυρήνα μία πολυτροπική βάση δεδομένων, επιτρέπει την παράλληλη εμφάνιση πρωτογενών και επεξεργασμένων δεδομένων καθώς και την κωδικοποίηση ενός μεγάλου αριθμού μεταδεδομένων. Κατά τη σχεδίασή του επιδιώχθηκε μια γενίκευση των προδιαγραφών, έτσι ώστε να μην είναι αυστηρά εξαρτημένο από συγκεκριμένες διαλέκτους και συγκεκριμένα γλωσσικά φαινόμενα. Πιστεύουμε ότι το σύστημα στο οποίο καταλήξαμε μπορεί να χρησιμοποιηθεί και σε άλλα έργα δημιουργίας και διαχείρισης σώματος τεκμηρίων διαφορετικών γλωσσών και διαλέκτων. Δηλαδή, πρόκειται για μία δυναμική πλατφόρμα που επιτρέπει την προσαρμογή της σε νέες ανάγκες που θα προκύψουν σε επόμενες έρευνες.

Βιβλιογραφία

Anderson, J., Beavan, D., Kay, C., 2007. SCOTS: Scottish Corpus of Texts and Speech. *J. Beal, K. Corrigan, H. Moisl (επιμ.), Creating and digitizing language corpora, 1,* Palgrave McMillan Publication, 17-34.

Anhøj, J., 2003. Generic design of web-based clinical databases. *Journal medical internet research 4*. [Διαθέσιμο στο: https://www.jmir.org/2003/4/e27/]

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

Boersma, P., Weenink, D., 2017. Praat: doing phonetics by computer. [Computer program], Version 6.0.35. [Ανακτήθηκε από: http://www.praat.org/, στις 29 Οκτωβρίου 2017].
Buseman, A., K. Buseman, K., 2007. Toolbox self-training: how to use the field linguist's Toolbox. [Ανακτήθηκε από: http://www.ling.helsinki.fi/kit/2009k/clt234/docs/Toolbox_Self- Training.pdf, στις 17 Νοεμβρίου 2015]

Clua, E., M-R. Lloret, M-R., 2006 New tendencies in geographical dialectology: the Catalan Corpus Oral Dialectal (COD). *J.-P. Y. Montreuil (επιμ.), New perspectives on romance linguistics, 2 (phonetics, phonology, and dialectology)*. Amsterdam/Philadelphia: John Benjamins. [Ανακτήθηκε από: http://pages.uv.es/foncat/cat/Treballs/10.Clua-Lloret.pdf, στις 15 Νοεμβρίου 2015].

Dawkins, R., 1916. *Modern Greek in Asia Minor: a study of the dialects of Sílli, Cappadocia and Phárasa with grammar, texts, translations and glossary*. Cambridge: Cambridge University Press.

Drettas, G., 1997. *Aspects Pontiques*. Paris: ARP.

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

Galiotou, E., Karanikolas, N., Manolessou, I., Pantelidis, N., Papazachariou, D., Ralli, A., Xydopoulos, G., 2014. Asia Minor Greek: Towards a computational processing, *Procedia – Social and Behavioral Sciences*, 147, Special issue: Proc. IC-ININFO 2013, Elsevier, pp. 458-466

Καρανικόλας Ν., Γαλιώτου Ε., Αθανασάκος, Κ., Κορωνάκης, Γ., 2015. Ένα πολυτροπικό σύστημα αρχειοθέτησης και διαχείρισης γραπτών και προφορικών πηγών μελέτης της γλώσσας και των γλωσσικών ιδιωμάτων, *Α. Ράλλη (εκδ.) Πρόγραμμα ΘΑΛΗΣ: «Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιατικής Ελληνικής»,* Παν. Πατρών, 69-98

Karanikolas, N., Galiotou, E., Papazachariou, D., Athanasakos, K., Koronakis, G., Ralli, A., 2015. Towards a computational processing of oral dialectal data, *Proceedings of the 19th PCI 2015* (Athens, Oct. 1-3, 2015), ACM Press, pp. 337-341

Karanikolas, N., Galiotou, E., Ralli, A., 2014. Towards a unified exploitation of electronic dialectal corpora: Problems and perspectives, *Text, Speech and Dialogue: Proceedings of the 17th Int. Conference TSD 2014* (Brno, Czech. Rep., Sept. 8-12 2014), LNAI 8655, Springer, pp. 257-266

Karanikolas N., Skourlas, C. 2014. Personal digital libraries: a self-archiving approach. *Library review 63* (6/7): 436-451

Κολλιοπούλου, Μ., Μανωλέσσου, Ι., Μαρκόπουλος, Θ., Παντελίδης, Ν. 2015. Ένα ψηφιακό σώμα κειμένων για τρεις μικρασιατικές διαλέκτους, *Α. Ράλλη (εκδ.) Πρόγραμμα ΘΑΛΗΣ: «Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιατικής Ελληνικής»*, Παν. Πατρών, 43-54

Kunst, J. P., Wesseling, F., 2010. Dialect corpora taken further: The DynaSAND corpus and its application in newer tools. *R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto & Y. Harada (επιμ.), Proceedings of the 24th Pacific Asia conference on language, information and computation*, Tohoku University, Nov. 4-7, 2010. Waseda University, 759-767.

Mackridge, P., 1990. *Some pamphlets on dead Greek dialects*. The annual of the British school at Athens 85: 201-212.

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

Παπαδόπουλος, Ά., 1955. *Ιστορική γραμματική της ποντικής διαλέκτου*. Αρχείον Πόντου. Παράρτημα 1. Αθήνα: Επιτροπή Ποντιακών Μελετών.

Παπαζαχαρίου, Δ., Καρασίμος, Α., 2015. Οργάνωση και κωδικοποίηση προφορικών πηγών σε πολυτροπική βάση δεδομένων αιχμής: η περίπτωση του AMiGre corpus, *Α. Ράλλη (εκδ.) Πρόγραμμα ΘΑΛΗΣ: «Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιατικής Ελληνικής»*, Παν. Πατρών, 55-68

Ράλλη, Α., Παπαζαχαρίου, Δ., Καρασίμος, Α., 2010. Εργαστήριο Νεοελληνικών Διαλέκτων και η βάση δεδομένων Gree.D. *M. Janse, B. Joseph, A. Ralli & A. Karasimos (επιμ.), Proceedings of the 4th International Conference on Modern Greek Dialects and Linguistic Theory*. Patras: University of Patras Press, 7-15

Σάκκαρης, Γ., 1940. *Περί της διαλέκτου των Κυδωνιέων εν συγκρίσει προς τας λεσβιακάς*. Μικρασιατικά Χρονικά 3: 74-141.

Sloetjes, H., P. Wittenburg, P., 2008. Annotation by category-ELAN and ISO DCR. *K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (επιμ.), Proceedings of the 6th international conference on Language Resources and Evaluation (LREC'08),* Marrakech, Morocco. European Language Resources Association, 816-820

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. [Ανακτήθηκε από: https://www.ijcat.org/articles/2-2/Research-on-Oral-Map-in-Regional-Dialect-Using-Google-Map.html, στις 15 Νοεμβρίου 2015]

Wells, J. C., 1997. SAMPA computer readable phonetic alphabet. *D. Gibbon, R. Moore & R. Winski (επιμ.), Handbook of standards and resources for spoken language systems.* Berlin & New York: Mouton de Gruyter.
[Διαθέσιμο στο: https://www.phon.ucl.ac.uk/home/sampa/]

Ευχαριστίες
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. Ευχαριστούμε θερμά τους συνεργάτες μας: Νίκο Κουτσούκο, Μαρία Κολιοπούλου, Θόδωρο Μαρκόπουλο, Δημήτρη Παπαζαχαρίου, Αθανάσιο Καρασίμο (Πανεπιστήμιο Πατρών), Νίκο Παντελίδη (ΕΚΠΑ), Ιώ Μανωλέσσου (Ακαδημία Αθηνών), Κώστα Αθανασάκο, Γιώργο Κορωνάκη (ΤΕΙ Αθήνας).