Εργαστήριο Νεοελληνικών Διαλέκτων και η βάση δεδομένων GREED

ΑΓΓΕΛΙΚΗ ΡΑΛΛΗ ΔΗΜΗΤΡΙΟΣ ΠΑΠΑΖΑΧΑΡΙΟΥ ΑΘΑΝΑΣΙΟΣ ΚΑΡΑΣΙΜΟΣ
Πανεπιστήμιο Πατρών
ralli@upatras.gr papaz@upatras.gr akarasimos@upatras.gr

1. Εισαγωγή

Η Νέα Ελληνική είναι πλούσια σε διαλεκτικές ποικιλίες, οι οποίες χρησιμοποιούνται στον καθημερινό λόγο, ενώ υπάρχουν και κάποιες γλωσσικές ποικιλίες που περιορίζονται σε συγκεκριμένες ομάδων πρεσβύτερων/ γερόντων και αντιμετωπίζουν το φάσμα της εξαφάνισης και εξάλειψης (Trudgill 1998, Κοντοσόπουλος 2001).

Εντούτοις, οι διαλεκτικές ποικιλίες μελετήθηκαν ελάχιστα, αν και περιέχουν αξιοσημείωτη εμφάνιση φαινόμενων για τη γλωσσολογική ανάλυση. Αυτό το διαλεκτικό μωσαϊκό οφείλεται σε μεγάλο βαθμό σε συγκεκριμένες ιστορικές, πολιτικές και κοινωνικές συνθήκες και περιστάσεις που χαράχτηκαν στην Ιστορία του Νεότερου Ελληνικού Κράτος, που απελευθερώθηκε από την Οθωμανική Αυτοκρατορία στις αρχές του 19ου αιώνα και κατά την αρχική του σύσταση περιλάμβανε τις γεωγραφικές περιοχές της Πελοποννήσου, της Στερεάς Ελλάδας και κάποιων νησιών. Έως τότε, διάφορες ομάδες της σύγχρονης Ελλάδας έκαναν εσωτερική μετανάστευση στο νεοσύστατο κράτος (π.χ. από Κρήτη, Μακεδονία και Δωδεκάνησα), ενώ παράλληλα ένας σημαντικός αριθμός Ελλήνων διαλεκτόφωνων προσφύγων μετακινήθηκαν από την Τουρκία (Μικρά Ασία και Πόντος) στην Ελλάδα, με το πέρας της Μικρασιατικής καταστροφής το 1922 και την ανταλλαγή πληθυσμών.

Σήμερα, η Κοινή Νέα Ελληνική είναι κυρίως βασισμένη στην Πελοποννησιακή διάλεκτο, ενώ οι διάλεκτοι από τα υπόλοιπα γεωγραφικά διαμερίσματα εντός και εκτός Ελλάδος δημιουργούν ένα ιδιαίτερο, ξεχωριστό και ποικιλόχρωμο γλωσσικό μωσαϊκό, οι οποίες χρήζουν άμεσα να περιγραφούν, να αναλυθούν και να διατηρηθούν, προτού αυτές εξαλείψουν παντελώς.

Εντούτοις, προς τη συγκεκριμένη κατεύθυνση δεν έχουν γίνει σοβαρά και συστηματικά βήματα έρευνας. Στην Ελλάδα υπάρχει από το 1908 ένα εθνικό ερευνητικό κέντρο στην Ακαδημία Αθηνών, το οποίο ενδιαφέρεται για γραπτά και προφορικά διαλεκτικά δεδομένα, αλλά τα διαλεκτικά δεδομένα δεν είναι ψηφιοποιημένα, τα περισσότερα είναι αδημοσίευτα με αυξημένες δυσκολίες πρόσβασης για τους εξωτερικούς ερευνητές. Μη-ψηφιοποιημένα διαλεκτικά δεδομένα εντοπίζονται παράλληλα σε συγκεκριμένους συλλόγους και οργανισμούς από πρόσφυγες από κάθε γωνιά της Ελλάδος, όπως για παράδειγμα το Ιστορικό Αρχείο των Μικρασιατών Ελλήνων στη Θεσσαλονίκη, το κέντρο Μικρασιατικών σπουδών, η Ένωση Ποντίων στην Παναγία Σουμελά Ημαθίας, αλλά έχουν συλλεχθεί κυρίως με ιστορικά κριτήρια και στόχους και φυσικά δεν έχουν ταξινομηθεί και κατηγοριοποιηθεί συστηματικά.

Η πρώτη συστηματική προσπάθεια ψηφιοποίησης, καταλογογράφησης και κωδικοποίησης διαλεκτικών δεδομένων έγινε από το Εργαστήριο Νεοελληνικών Διαλέκτων του Πανεπιστημίου Πατρών με την υλοποίηση της ηλεκτρονικής βάσης GREED, η οποία περιέχει γλωσσολογικά και μετα-γλωσσολογικά corpora. Αυτά τα δεδομένα συλλέχθηκαν από έρευνες πεδίου, όπου καταγράφηκαν δεδομένα φυσικής και αυθόρμητης ομιλίας με στόχο το σχηματισμό μιας αντιπροσωπευτικής εικόνας της γλωσσολογικής κατάστασης συγκεκριμένων γεωγραφικών και κοινωνικών περιοχών της Ελλάδος. Παράλληλα, γίνεται προσπάθεια συλλογής και διαλεκτικών χειρογράφων και διαφόρων κειμένων, βιβλίων, έντυπων συλλογών, ώστε να δημιουργήσουμε ένα

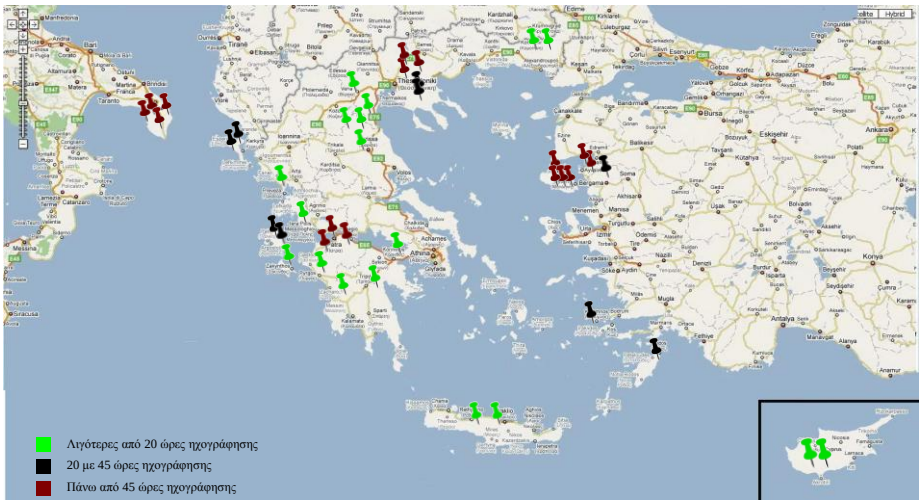
ψηφιοποιημένο σώμα κειμένων, ωστόσο ο τελευταίος στόχος αποτελεί μακροχρόνια προσπάθεια και έμμεση προτεραιότητα. Φιλοδοξία μας είναι η βάση GREED να αποτελεί πολύτιμο αρωγό για τη μελλοντική έρευνα της κατηγοριοποίησης και οργάνωσης των διαφόρων γλωσσολογικών φαινομένων – φωνολογικά, μορφολογικά, κοινωνιογλωσσολογικά κτλ. – που εντοπίζονται διαδιαλεκτικά. Επομένως, θα διευκολύνει αισθητά τις δημοσιεύσεις και εκδόσεις γλωσσάριων, λεξικών και γραμματικών των διαφόρων διαλέκτων της Νέας Ελληνικής.

2. GREED Corpus και συλλογή δεδομένων

Ο θεμέλιος λίθος για την ανάπτυξη της ηλεκτρονικής βάσης GREED αποτέλεσαν διάφορα ερευνητικά προγράμματα που αποσκοπούσαν στη διατήρηση συγκεκριμένων διαλέκτων:
"*Grico: Dialect spoken in the area of Salento, South Italy*" (Interreg II, Ευρωπαϊκή Ένωση, σύνολο 55 ωρών, συντονίστρια Αγγελική Ράλλη).
"Διαλεκτικές ποικιλίες της Ανατολικής Λέσβου. Σύγκριση με την μικρασιτική διάλεκτο των Κυδωνίων και Μοσχονησίων" (Υπουργείο Παιδείας, σύνολο 45 ωρών, συντονίστρια Αγγελική Ράλλη).
"Η μικρασιάτικη διάλεκτος των Κυδωνίων και Μοσχονησίων" (Υπουργείο Αιγαίου και Υπουργείο Παιδείας, σύνολο 112 ώρες, συντονίστρια Αγγελική Ράλλη).
"Cappadocian". Endangered Languages and Documentation Programme. University of London SOAS, σύνολο 40 ωρών, συντονιστές Mark Janse, Αγγελική Ράλλη και Δημήτρης Παπαζαχαρίου).
"*Διαλεκτική ποικιλία Πάτρας*" (Πανεπιστήμιο Πατρών, σύνολο 100 ωρών, συντονιστής Δημήτρης Παπαζαχαρίου).
"*Η διάλεκτος της Αγίας Παρασκευής Λέσβου*" (Δήμος Αγίας Παρασκευής, σύνολο 40 ωρών, συντονίστρια Αγγελική Ράλλη)
"*Τουρκοκρητικά Μικράς Ασίας*" (Υπουργείο Εξωτερικών, σύνολο 32 ωρών, συντονίστριες Αγγελική Ράλλη και ΧΧ)
"*Από το γλωσσικό ιδίωμα των Μεγάρων στο γλωσσικό ιδίωμα της Παλαιάς Αθήνας*" (Ίδρυμα Λεβέντη και Δήμος Μεγαρέων, σύνολο 44 ωρών, συντονίστριες Αγγελική Ράλλη και Αγγελική Σύρκου)
Παράλληλα η συλλογή υλικού γίνεται στα πλαίσια μαθημάτων, διπλωματικών εργασιών και διδακτορικών διατριβών που προσθέτουν στη βάση σημαντικό υλικό. Ο ακόλουθος πίνακας δίνει μια κατατοπιστική εικόνα του συνολικού υλικού της βάσης:

Διαλεκτική περιοχή	*Ώρες*	*Ποσοστό*	*Ομιλητές*	*Ποσοστό*
Καππαδοκικά	41 Ώρες	8%	82 Ομιλητές	12,77%
Μικρά Ασία	105 Ώρες	21,00%	78 Ομιλητές	12,14%
Κύπρος	2,5 Ώρες	0,50%	12 Ομιλητές	1,89%
Δωδεκάνησα	9,5 Ώρες	2%	13 Ομιλητές	2,02%
Ήπειρος	12 Ώρες	2,20%	17 Ομιλητές	2,60%
Επτάνησα	15 Ώρες	3,00%	33 Ομιλητές	5,10%
Μακεδονία	9 Ώρες	1,60%	16 Ομιλητές	2,50%
Λέσβος	128 Ώρες	25,30%	80 Ομιλητές	12,46%
Κάτω Ιταλία	55 Ώρες	11,00%	68 Ομιλητές	10,60%
Στερεά Ελλάδα	12 Ώρες	2,20%	21 Ομιλητές	3,27%
Θεσσαλία	8 Ώρες	2%	16 Ομιλητές	2,50%
Θράκη	8 Ώρες	2%	6 Ομιλητές	1%
Πελοπόννησος	100 Ώρες	20%	200 Ομιλητές	31,15%
Σύνολο	505 Ώρες	100,0 %	642 Ομιλητές	100,0%

Πίνακας 1: *Συνολική στατιστική της ηλεκτρονικής βάσης*



Εικόνα 1: *Οι γεωγραφικοί τόποι, όπου πραγματοποιήθηκαν διαλεκτικές ηχογραφήσεις από το Εργαστήριο Νεοελληνικών Διαλέκτων*

Ουσιαστικά υπάρχουν δύο διαφορετικοί τύποι του προφορικού υλικού: α) ηχογραφήσεις από αυθόρμητο προφορικό λόγο από συναντήσεις και β) στοχευμένες συνεντεύξεις για την εξαγωγή συγκεκριμένων γλωσσολογικών πληροφοριών από προφορικό υλικό. Η προσπάθεια αναμόχλευσης προσωπικών διηγήσεων και κυρίως παλαιότερων ιστοριών ήταν ηθελημένη επιλογή – στις πλειονότητα των ερευνητικών προγραμμάτων – για τη διαφύλαξη υλικού πολιτισμικής κληρονομιάς ταυτόχρονα με τη συλλογή των γλωσσικών δεδομένων. Πιο συγκεκριμένα, οι ηχογραφήσεις έγιναν από ερευνητές πεδίου που είχαν αποκτήσει κάποιες κοινωνικές σχέσεις και επαφές με την υπό διερεύνηση κοινότητα και τους πληροφορητές συγκεκριμένα, ή κυρίως με τη συνδρομή της φυσικής παρουσίας ενδιάμεσου, δηλαδή ενός μέλους της τοπικής κοινότητας ή άτομο που διατηρεί στενές επαφές με τους πληροφορητές (φίλος φίλου, συγγενής συγγενή, γείτονας). Για παράδειγμα, μιλώντας για τις προσωπικές εμπειρίες και δυσκολίες από δύσκολες περιόδους της ελληνικής ιστορίας ήταν πιο αποτελεσματικό για να τους κάνουμε να ανοιχτούν συναισθηματικά, να αισθανθούν άνετα και να μπορέσουν να εκφραστούν ελεύθερα μιλώντας διαλεκτικά και να καταφέρουν να αφαιρέσουν από το μυαλό την ιδέα της συνέντευξης και να αισθανθούν ότι βρίσκονται σε μια καθημερινή στιγμή. Σύμφωνα με τις αρχές της Μεθοδολογίας της Έρευνας, αυτή η μέθοδος παρέχει σημαντικές πληροφορίες για την προφορική ιστορία και τα γλωσσικά δεδομένα και αυξάνει σημαντικά τις πιθανότητες για συλλογή αυθόρμητου λόγου και περιορισμό του φαινόμενου της προσποίησης.

Στην GREED, οι διάλεκτοι είναι καταχωρημένες γεωγραφικά (καθότι αυτή πληροφορία θα βοηθήσει το διαλεκτικό χάρτη μελλοντικά) και οι πληροφορίες σχετικά με τα μεταδεδομένα είναι δομημένες σε επτά βασικές κατηγορίες: Ιδιότητες Αρχείων, Διάλεκτος, Ερευνητικό πρόγραμμα, Τεχνικές πληροφορίες, Επικοινωνιακή περίσταση, Πληροφορητές, Γλωσσολογικά δεδομένα. Αυτές οι βασικές κατηγορίες που χρησιμοποιήθηκαν και για τον χαρακτηρισμό όλου του προφορικού υλικού, έχουν πολλές υποκατηγορίες που παρέχουν πολλές επιλογές για τη δημιουργία μιας προχωρημένης μηχανής αναζήτησης. Στα σχήματα που ακολουθούν δίνονται δείγματα από δύο ομάδες μεταδεδομένων από τη συλλογή προφορικού υλικού από τις Νεοελληνικές διαλέκτους.

Αν και η δημιουργία της βάσης εξακολουθεί να είναι υπό δημιουργία, η GREED περιέχει πάνω από 460 ώρες προφορικού υλικού, συνοδευμένο από μεταδεδομένα και 40 ώρες του υλικού έχει ήδη απομαγνητοφωνηθεί συνοδευμένο από πρωτόκολλο χαρτογράφησης αρχείων.

Παράλληλα έχει ξεκινήσει η ψηφιοποίηση διαφόρων χειρογράφων και σπανίων βιβλίων, ώστε σύντομα η GREED να διαθέτει και την αντίστοιχη πλατφόρμα για την αξιοποίηση του γραπτού διαλεκτικού υλικού. Στο πλαίσιο του ερευνητικού προγράμματος έγινε μια οργανωμένη προσπάθεια ψηφιοποίησης χειρογράφων, κυρίως νομικής και συμβολαιογραφικής φύσεως που περιείχαν μεταξύ άλλων και διαλεκτικό υλικό. Τα 1500 και πλέον χειρόγραφα ψηφιοποιήθηκαν και βρίσκονται στη διαδικασία χαρακτηρισμού τους από επιλεγμένες πληροφορίες μεταδεδομένων. Είναι ενδεικτικό ότι ηλεκτρονικές βιβλιοθήκες χειρογράφων στο διαδίκτυο συνοδεύονται πάντα από πληροφορίες περιγραφής του χειρογράφου.[1]

3. Είδος αρχείων (αρχεία ήχου, μεταγραφές, πρωτόκολλα χαρτογράφησης και ψηφιακά χειρόγραφα)

Τα δεδομένα των ηχογραφήσεων των διαλέκτων συλλέχθηκαν με τη χρήση επαγγελματικών ψηφιακών κασετοφώνων Marantz. Η επιλογή των επαγγελματικών ψηφιακών συσκευών ελήφθει με βάση τις διεθνείς προδιαγραφές για ποιοτικές ηχογραφήσεις με τις ελάχιστες δυνατές απώλειας.

Οι πληροφορητές συνήθως ηχογραφήθηκαν κατά ζεύγη ή κατά μόνας με τη συνδρομή του ενδιάμεσου και ο μέσος χρόνος ηχογράφησης είναι περίπου στα εξήντα λεπτά. Όπως και στα πιο πρόσφατα ερευνητικά προγράμματα, οι ηχογραφήσεις πραγματοποιήθηκαν με ψηφιακές συσκευές εγγραφής (η επαγγελματική σειρά της Marantz), που εγγράφει τις συνομιλίες σε ασυμπίεστη μορφή αρχείου .wav και ελαχιστοποιεί την οποιαδήποτε διαδικασία ψηφιοποίησης των ηχητικών αρχείων. Παράλληλα, οι συγκεκριμένες συνομιλίες καταγράφονται στερεοφωνικά – σε αριστερό και δεξί κανάλι – με τη χρήση δύο μικροφώνων, ώστε να αντιστοιχείται ένα κανάλι ανά πληροφορητή, εφόσον είναι δυνατόν. Με αυτό τον τρόπο, καταφέραμε να μειώσουμε το περιβαλλοντικό θόρυβο (περίπου 40 db) για να επιτύχουμε την μέγιστη δυνατή ποιότητα εγγραφής και την ίδια στιγμή να μειώσουμε στο ελάχιστο το προβληματικό φαινόμενο της επικάλυψης, όταν δύο ομιλητές μιλάνε την ίδια χρονική στιγμή ή διακόπτει ο ένας τον άλλον.

Να σημειωθεί ότι τα ηχητικά αρχεία εισάγονται σε υπολογιστή συνδεμένο με βάση δεδομένων χωρίς καμία υποβάθμιση ποιότητας και αποθηκεύονται για λόγους ασφαλείας σε ένα σύστημα αποθήκευσης NAS για υψηλότερη ασφάλεια. Επίσης η εισαγωγή των ηχητικών αρχείων των διαλέκτων. Τυπικοί στόχοι επεξεργασίας συμπεριλαμβάνουν την ορθή ονοματοδοσία, διαχωρισμό καναλιών, αφαίρεση προσωπικών πληροφοριών, ενίσχυση των χαμηλής έντασης ηχογραφήσεων, μείωση του θορύβου και καθαρισμός του σήματος από έντονους μικροφωνισμούς.

Επομένως τα ηλεκτρονικά αρχεία των Νεοελληνικών διαλέκτων στην ηλεκτρονική βάση GREED είναι τα ακόλουθα:

(α.) Ψηφιακά Αρχεία ήχου: ηχογραφήσεις φυσικού διαλεκτικού λόγου σε μορφή στέρεο, καθώς και μονοκαναλικός διαχωρισμός.

(β.) Αρχεία περιγραφής των ηχογραφήσεων: (i.) μεταγραφές ομιλίας (εναλλαγές διαλόγου, απομαγνητοφώνηση ορθογραφική, φωνολογική (σπάνια) και μορφολογική σήμανση), (ii.) πρωτόκολλο χαρτογράφησης ηχητικού αρχείου (ανά δύο λεπτά χαρακτηρισμός αρχείου με συγκεκριμένα κριτήρια

[1] Ενδεικτικά η Schoenberg Database of Manuscripts (http://dla.library.upenn.edu/cocoon/dla/schoenberg/index.html), η National Mission for Manuscripts (http://www.namami.org/index.htm), η Leeds Verse Database (http://www.leeds.ac.uk/library/spcoll/bcmsv/intro.htm), η International Dunhuang Project: The Silk Road Online (http://idp.bl.uk/), η Medieval and Early Modern Manuscripts Collection: Database and Digital Images (http://research.hrc.utexas.edu/pubmnem/), η Old English Manuscript Database (http://www8.georgetown.edu/departments/medieval/labyrinth/subjects/mss/oe/oldeng.html) μεταξύ άλλων.

(γ.) Κείμενα και χειρόγραφα: κείμενα που έχουν γραφτεί πρωταρχικώς στη διάλεκτο.
Εκτός από τα ψηφιακά αρχεία ήχου, μια ικανοποιητική βάση δεδομένων πρέπει να εσωκλείει και μεταγραφές – απομαγνητοφωνήσεις των αρχείων. Υπάρχει μια μεγάλη συζήτηση από τους ερευνητές βάσεων δεδομένων για το ποιος είναι ο πλέον κατάλληλος τρόπος μεταγραφής των ηχητικών αρχείων (*φωνητικός*, *φωνολογικός* ή *ορθογραφικός*). Συμφωνώντας με τους Durand & Eriksson (2007) και τους Anderwald & Wagner (2007: 42-43) υποστηρίζουμε ότι τα μειονεκτήματα της φωνολογικής και φωνητικής απομαγνητοφώνησης είναι τέτοιας φύσεως για τα ελληνικά που προτιμήσαμε την ορθογραφική μεταγραφή των προφορικών συνομιλιών. Η επιλογή μας επηρεάστηκε σημαντικά από την προοπτική εκμετάλλευσης του διαλεκτικού υλικού για μορφολογικούς αναλυτές με τη χρήση του απομαγνητοφωνημένου υλικού για μορφολογικούς και λεξικογραφικούς σκοπούς. Παράλληλα κατά την απομαγνητοφώνηση βασιστήκαμε στις κωδικοποιήσεις της Ανάλυσης Λόγου αναφορικά με τις εναλλαγές διαλόγου, διακοπές, επικαλύψεις, παύσεις, επιμηκύνσεις, γρήγορος ή αργός ρυθμός ομιλίας, ένταση και χαμηλόφωνη ομιλία, είναι τα διάφορα μεταγλωσσικά φαινόμενα που μπορούν να επηρεάσουν φωνολογικά φαινόμενα και σημειώνονται κατά την απομαγνητοφώνηση και χαρτογράφηση του αρχείου.
Η ορθογραφική μεταγραφή δίνει τη δυνατότητα για πιο απρόσκοπτη διερεύνηση των μορφοσυντακτικών χαρακτηριστικών και κοινωνιογλωσσολογικών φαινόμενων, αλλά υπάρχουν εμφανή προβλήματα που αφορούν ζητήματα τεχνικής φύσεως, όπως για παράδειγμα, πως θα λειτουργήσει η φωνητική κωδικοποίησης σε λογισμικά όπως το Praat και το E-Lan.
Τέλος, μόνο η ορθογραφική μεταγραφή των δεδομένων θα καλύψει τις υπάρχουσες απαιτήσεις της βάσης: στόχος ενός ολοκληρωμένου corpus πρέπει να είναι η δυνατότητα να είναι μηχανικά-αναγνώσιμο (machine-readable), να επιτρέπει την εύκολη και γρήγορη διαχείριση αναζήτησης με διάφορα εργαλεία και το πλέον σημαντικό να συγκρίνεται με άλλα σώματα κειμένων όσον αφορά την απλότητα και την ευχρηστία. Επιπροσθέτως, η ορθογραφική μεταγραφή θα μας επιτρέψει να συγκρίνουμε τα δεδομένα με αντίστοιχα άλλων γραπτών και προφορικών συλλογών και μας επιτρέπουν να κάνουμε συγκρίσεις ανάμεσα σε διαφορετικούς ομιλητές, διαφορετικές διαλέκτους και διαλεκτικές περιοχές και διαφορετικά corpora.
Παρόλο που οι συνεντεύξεις είναι άμεσα προσβάσιμες λόγω της ηλεκτρονικής τους μορφής [ο κάθε ερευνητής μπορεί να έχει άμεση πρόσβαση στο αρχείο που επιθυμεί για ανάλυση, ακόμα και στην στερεοφωνική του μορφή], η απουσία φωνολογικής απομαγνητοφώνησης αποτρέπει την γρήγορη και ευρεία φωνολογική ανάλυση χωρίς τη χρήση των ηχητικών αρχείων. Όλα τα απομαγνητοφωνημένα αρχεία έχουν καταγραφεί και σε αρκετά σημεία φωνολογικά φαινόμενα έχουν χαρτογραφηθεί από την απομαγνητοφώνηση χωρίς την άμεση σύνδεση με τα ηχητικά αρχεία. Ελπίζουμε μελλοντικά πως η ηλεκτρονική βάση θα παρέχει την επιθυμητή ευθυγράμμιση ήχου και κειμένου, όπως στο Necte (βλ Allen *et al.* 2007) και στο ONZE[2] (βλ. Gordon *et al.* 2007)· προς το παρόν η ευθυγράμμιση επιτυγχάνεται μόνο μέσω του ELan και του Praat.
Για να καλυφθούν κάποια κενά της ορθογραφικής μεταγραφής, αλλά κυρίως για την δυνατότητα μιας γρήγορης χαρτογράφησης και «ακτινογραφίας» ενός ηχητικού αρχείου παρέχεται για αρκετές περιπτώσεις των διαλεκτικών δεδομένων το πρωτόκολλο χαρτογράφησης. Ανά δύο λεπτά χαρακτηρίζεται το αρχείο με βάση κάποια κριτήρια τεχνικά και περιγραφικά, όπως ποιότητα ηχογράφησης, ύπαρξη θορύβων, αριθμός ομιλητών, καθώς και με γλωσσολογικά κριτήρια, όπως καταγραφή ή σήμανση ενδιαφερόντων γλωσσικών φαινόμενων πάσης φύσεως (π.χ. σήμανση για αλλόμορφα, για ασυνήθιστο επιτονισμό, για συντακτικούς περιορισμούς κλπ.).

[2] http://www.lacl.canterbury.ac.nz/onze/news.html

4. Διαχείριση και ιστοσελίδα

Οι απαιτήσεις για τη βάση δεδομένων είναι για ένα σύστημα που να μπορεί να παρέχει πρόσβαση στα διαλεκτικά δεδομένα μέσω μιας κοινής διεπιφάνειας. Απαιτητικοί έλεγχοι πιστότητας πρωτοκόλλων και λοιποί κανόνες σχετικά με συνοχή και ασφάλεια των δεδομένων αποτελούν βασικές προϋποθέσεις. Κιας και πρωταρχικός στόχος είναι η υλοποίηση ενός εργαλείου βάσης δεδομένων που να είναι εύχρηστο, πολυχρηστικό και ανοιχτό για τη γλωσσολογική κοινότητα για αρκετό καιρό, δημιουργήθηκε μια διαδραστική ιστοσελίδα (έχοντας ως οδηγό τα ISCC χαρακτηριστικά, Dipper et al. 2007) με στόχο να μπορεί να αλληλεπιδρά με άλλα λογισμικά επεξεργασίας, όπως Praat. Το σύστημά μας υποστηρίζει ελληνικούς και λατινικούς χαρακτήρες. Το περιβάλλον εργασίας των χρηστών που παρέχεται στους ερευνητές είναι γρήγορο και εύκολο στη χρήση· επομένως ο χρόνος εκπαίδευσης είναι μειωμένος.

Η αρχιτεκτονική δομή της βάσης είναι χτισμένη πάνω σε τέσσερα αντικείμενα. Όλα τα αντικείμενά (*Metadata, Metadatadetails, mdListValues* [προ-εισαγμένες τιμές] και *FileAttribs* [πίνακας με όλα τα αρχεία]) είναι συνδεμένα αναμεταξύ τους με μια σχέση 'ένα προς πολλά', για παράδειγμα η τιμή 'dialect name' του Metadatadetails είναι συνδεμένη με τις τιμές 'Ποντιακά', 'Λεσβιακά', 'Κυπριακά' μεταξύ άλλων τιμών από το mdListValues. Το σύστημά είναι βασισμένο σε αρχιτεκτονική client-server (apache server), η οποία συσχετίζεται με μια συσχετιστική βάση δεδομένων τύπου MySQL. Όλες οι σελίδες είναι χτισμένες πάνω σε φόρμες template και επεξεργάζονται τα δεδομένα χρησιμοποιώντας μικρούς κώδικες σε PHP γλώσσα. Οι χρήστες έχουν πρόσβαση στα δεδομένα μέσω μιας PHP διεπιφάνειας με τη χρήση του HTML πρωτοκόλλου. Ένας σημαντικός λόγος επιλογής ενός client/ server δικτύου είναι επειδή επιτρέπει την πρόσβαση στη βάση δεδομένων την ίδια στιγμή και στα αρχεία που είναι αποθηκευμένα στον server.

Το βασισμένο στο διαδίκτυο σύστημα μας ακολουθεί τις αρχές ενός client/ server μοντέλου σχετικά με την προσκόμιση πληροφορίας των αρχείων. Βασισμένο σε ένα τέτοιο μοντέλο ο client υπολογιστής είναι συνδεμένος με τον server υπολογιστή, ο οποίος περιέχει τις πληροφορίες και φυσικά ο client υπολογιστής εξαρτάται άμεσα από τον server για την απόκτηση των απαραίτητων πληροφοριών. Βασισμένο στη δικτυακή τεχνολογία, είναι ανοιχτό για οποιοδήποτε λειτουργικό σύστημα που έχει φυλλομετρητή διαδικτύου (web browser). Για την ώρα, για τη διαφύλαξη της σταθερότητας του συστήματος, οι χρήστες μπορούν να ανεβάσουν αρχεία, αλλά οι τιμές των μεταδεδομένων πρέπει να εισαχθούν από τον διαχειριστή του συστήματος έπειτα από αίτηση του χρήστη. Στην παρούσα φάση της υλοποίησης, δουλεύουμε σε μια παραλλαγμένη TEI (Text Encoding Initiative) έκδοση για τα δεδομένα. Επιπλέον, το σύστημα παράγει αναφορές καταγραφής αλλαγών και προβλημάτων αυτόματα, ώστε να είναι δυνατή η γρήγορη εύρεση του προβλήματος, για παράδειγμα όταν ο διαμοιραστής αποτυγχάνει να αναβαθμίσει τις φόρμες των απαραίτητων μεταδεδομένων μέσα σε περιορισμένο χρονικό διάστημα (30 δευτερόλεπτα).

5. Εργαλεία ανάλυσης των Νεοελληνικών Διαλέκτων

Όπως αναφέραμε σε προηγούμενη ενότητα η βάση δεδομένων συνοδεύεται εκτός από τα ηχητικά αρχεία και από τα αντίστοιχα αρχεία μεταγραφής, για όσα αρχεία ήχου έχουν πραγματοποιηθεί. Η επιλογή συνοδευτικού λογισμικού δεν είναι εύκολη υπόθεση, αποτελεί αναπόσπαστο κομμάτι μιας καλής βάσης προφορικών δεδομένων και τα λογισμικά πρέπει να πληρούν βασικά κριτήρια[3]:

(1) Να είναι λογισμικά ανοιχτού κώδικα και ελεύθερα ως προς τη χρήση

(2) Να παρέχει μεγάλο εύρος σχεδιαστικών παραμέτρων

(3) Να υποστηρίζει αρχεία από διαφορετικά λογισμικά που χρησιμοποιούνται για τον σχολιασμό αρχείων σε διαφορετικά γλωσσολογικά επίπεδα

(4) Να επιτρέπει την χρήση πιθανών add-ons και plug-ins

[3] Για αυτό το λόγο επιλέχθηκαν τα λογισμικά Praat (μαζί με το Akustyk) και το ELAN.

(5) Να προσφέρεται συνεχής υποστήριξη από τους προγραμματιστές/ παραγωγούς του λογισμικού
(6) Να είναι πολυγλωσσικό ή τουλάχιστον σε αγγλική έκδοση και να επιτρέπει τη χρήση του Unicode πρωτοκόλλου

6. Μελλοντικά σχέδια

Η ηλεκτρονική διαλεκτική βάση GREED και η συλλογή υλικού από τις Νεοελληνικές διαλέκτους είναι έρευνα υπό εξέλιξη. Είναι στις επιθυμίες και στα σχέδια μας να παρέχουμε μια ολοκληρωμένη μορφή της βάσης, η οποία θα είναι ανοιχτή για όλη την ακαδημαϊκή – και όχι μόνο – κοινότητα. Σεβόμενοι τα μελλοντικά μας σχέδια για την ηλεκτρονική βάση διαλεκτικών δεδομένων, τα ακόλουθα σημεία θεωρούμε ότι οφείλουμε να τα υπογραμμίσουμε:
[Τεχνικά] Κατά την διάρκεια της έρευνας για τις Νεοελληνικές διαλέκτους, αναβαθμίσαμε σημαντικό την διεπιφάνεια επίδρασης του χρήστη με ένα εύκολο στη χρήση web περιβάλλον, όπου δεν απαιτείται η χρήση κανενός λογισμικού από τον χρήστη. Έχουμε τη δυνατότητα να παρέχουμε μια πληθώρα κατανοητών και κατατοπιστικών κοινωνιογλωσσολογικών μεταδεδομένων, όπως και συμπληρωματικές πληροφορίες για τα ηχητικά αρχεία. Εντούτοις, πρέπει να παρέχουμε κωδικοποιημένες πληροφορίες και μεταδεδομένα για τα ψηφιακά δεδομένα, τα οποία δεν έχουν καταχωρηθεί και καταλογογραφηθεί με ενιαίο τρόπο. Η δική μας έκδοση βρίσκεται σε στάδιο δοκιμής και αναβάθμισης, αλλά έχει αποδειχθεί μέχρι στιγμής αρκετά γρήγορη και φιλική προς τον χρήστη.
[Τεχνικά] Δημιουργούμε έναν πιο αναπτυγμένο σύστημα αναζήτησης με κριτήρια βασισμένα στα μεταδεδομένα. Στοχεύουμε να κάνουμε τη βάση πιο γρήγορη, χωρίς προβλήματα και με σταθερότητα κώδικα.
[Τεχνικά] Να ελέγξουμε τα υπάρχοντα αρχεία μεταγραφής και απομαγνητοφώνησης και να συνεχίσουμε την μεταγραφή των υπόλοιπων διαλεκτικών προφορικών αρχείων.
[Τεχνικά] Έναρξη ευρύτερων φωνολογικών/ φωνητικών μεταγραφών που να συνοδεύουν τις ορθογραφικές μεταγραφές και τη μορφολογικές αναλύσεις.
[Τεχνικά] Μια αξιολόγηση της βάσης από ερευνητές που έχουν ήδη δουλέψει με τη βάση, καθώς και από προσωπικό που έχει εμπειρία από άλλες ηλεκτρονικές βάσεις
[Γλωσσολογικά] Έναρξη διερεύνησης του σώματος όλων των διαλεκτικών δεομένων με τη χρήση του μορφολογικού αναλυτή, για παράδειγμα με το TOOLBOX, ώστε να δημιουργήσουμε ένα καλό λεξικό.
[Γλωσσολογικά] Εμπλουτισμός του διαλεκτικού υλικού, τόσο προφορικού, όσο και γραπτού, με την οργάνωση νέων αποστολών και συλλογών υλικού, καθώς και την ψηφιοποίηση του γραπτού υλικού που έχουμε στην κατοχή μας.
[Έρευνα] Σχεδιάζουμε την έκδοση λεξικών, λεξιλογίων και γραμματικών για τις διαλέκτους που έχουμε μεγάλο εύρος προφορικού υλικού.
[Έρευνα] Επιπλέον χορήγηση ερευνητικών προσπαθειών για οικονομική υποστήριξη με στόχο τη βελτίωση και εξέλιξη της ηλεκτρονικής βάσης GREED.
[Έρευνα] Χρήση της βάσης δεδομένων ως βοηθητικό εργαλείο για τη μελλοντική διαλεκτική έρευνα για διάφορα φωνολογικά και μορφολογικά φαινόμενα, τα οποία εντοπίζονται δια-διαλεκτικά και αποτελούν σημαντικότερο αρωγό για την παραγωγή άρθρων και μονογραφιών για τις διάφορες νεοελληνικές διαλέκτους.
[Έρευνα] Επικοινωνία και συνεργασία με τη διεθνή γλωσσολογική κοινότητα, ώστε να παρέχουμε τη δυνατότητα πρόσβασης σε ελληνικά διαλεκτικά δεδομένα και παράλληλα να διατηρήσουμε και να διασώσουμε μια εξαιρετικά σημαντικά πολιτιστική κληρονομιά.

Βιβλιογραφία

Allen, W., Beal, J., Corrigan, K., Maguire, Moisl, H. (2007). *A Linguistic 'Time Capsule'*: *The Newcastle Electronic Corpus of Tyneside English*. Creating and digitalizing Language Corpora Vol.2 (edited by Beal J. *et al.*). Palgrave McMillan Publication, pp. 16-48.

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

Anderwald, L., Wagner, S. (1997). *FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data.* Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

Barbiers, S., Cornips, L., Kunst, J.-P. (2007). *The Syntactic Atlas of the Dutch Dialects (SAND): A Corpus of Elicited Speech as an On-line Dynamic Atlas.* Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

Dipper, S., Goetze, M., Skopeteas, S. (2007). *Information Structure in Cross-linguistic corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. ISIS Working papers of the SFB 632.

Gordon, E., Maclagan, M., Hay, J. (2007). *The ONZE Corpus*. Creating and digitalizing Language Corpora Vol.2 (edited by Beal J. *et al.*). Palgrave McMillan Publication, pp. 82-104.

MacWhinney, B. (2007). *The Talkbank Project*. Creating and digitalizing Language Corpora Vol.1 (edited by Beal J. *et al.*). Palgrave McMillan Publication.

Ralli, A. (2006). Syntactic and Morphosyntactic Phenomena in Modern Greek Dialects: The State of the Art. 2007*. Journal of Greek Linguistics* 2006: 121-159. Ακαδημία Αθηνών (1933-). *Ιστορικόν Λεξικόν της Νέας Ελληνικής Γλώσσης, της τε Κοινώς Ομιλουμένως και των Ιδιωμάτων*. Αθήνα

Ιστορικό Αρχείο Ελλήνων Προσφύγων Καλαμαριάς Θεσσαλονίκης, (http://www.kalamaria.gr/index.php?option=com_content&task=view&id=85&Itemid=599)

Κοντοσόπουλος, Ν. (2006)[4]. *Διάλεκτοι και ιδιώματα της Νέας Ελληνικής*. Αθήνα: εκδ. Γρηγόρης.

Μηνάς, Κ. (2003), *Η γλώσσα των Δημοσιευμένων Μεσαιωνικών ελληνικών εγγράφων της Κάτω Ιταλίας και της Σικελίας*. (επανέκδοση από Ι.Λ.Ν.Ε.), Αθήνα.

Καραναστάσης, Α. (1986-1992). *Ιστορικόν Λεξικόν των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας*. Τόμοι Α-Ε. Αθήνα.

Καραναστάσης, Α. (1992). *Γραμματική των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας*. Αθήνα

Κωστάκης, Θ. (1986-1987). *Λεξικό της Τσακωνικής Διαλέκτου*. Τόμοι Α-Γ, Αθήνα.

Ράλλη, Α. (to appear). *Λεξικό των Ιδιωμάτων Κυδωνιών, Μοσχονησίων και Ανατολικής Λέσβου*. Πανεπιστήμιο Πατρών: Εργαστήριο Νεοελληνικών Διαλέκτων.

Ίδρυμα Μανώλη Τριανταφυλλίδη (http://ins.web.auth.gr/english.htm)