Η ΔΙΑΛΕΚΤΟΣ ΤΩΝ ΜΕΓΑΡΩΝ ΣΤΗΝ ΗΛΕΚΤΡΟΝΙΚΗ ΒΑΣΗ *GREED*

ΑΘΑΝΑΣΙΟΣ Ν. ΚΑΡΑΣΙΜΟΣ

Abstract

Greece contains a rich variety of dialects (cf. Kontosopoulos 2006), some of which are used systematically in every-day speech, while others are restricted to specific groups of elders and are in danger of extinction. No attempt has been made to digitize, catalogue and encode dialectal data until very recently, when the *Laboratory of Modern Greek Dialects* started developing the GREED Database, with the aim to facilitate access to dialectal data, both linguistic and metalinguistic, in order to share dialectal information with the international linguistic community and ultimately preserve a significant linguistic heritage. More particularly, we present the dialect of Megara. The base has been built on established standards, such as the TEI Header, OLAC and IMDI, and its main goal is the enhancement of the material, the exchange of data and the support of academic research activities. It is not dependent on specific operational systems (OS) and commercial software, and a principal characteristic is the three-dimensional structuring concerning (i) the *dialectal data*, (ii) the *metalinguistic data* and (iii) the *combined browsing* of both. Although the work is still in progress, GREED contains 20 hours of natural dialectal speech from Megara, which is accompanied by metalinguistic information (overall: 400 hours of dialectal speech and 40 hours of material are already annotated and transcribed). Among the future benefits of the database will be to help future dialectal research, in categorizing and organizing various phonological and morphological phenomena, which are found cross-dialectally, and make easier the publication of dictionaries and grammars of the various Modern Greek dialects.

Λέξεις-Κλειδιά: διάλεκτος των Μεγάρων, GREED, βάσεις δεδομένων, ηλεκτρονικές βιβλιοθήκες

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 φαινομένων 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 εκλείψουν 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:

1. "*Grico: Dialect spoken in the area of Salento, South Italy*" (Interreg II, Ευρωπαϊκή Ένωση, σύνολο 55 ωρών, συντονίστρια Αγγελική Ράλλη).
2. *"Διαλεκτικές ποικιλίες της Ανατολικής Λέσβου. Σύγκριση με την μικρασιτική διάλεκτο των Κυδωνίων και Μοσχονησίων"* (Υπουργείο Παιδείας, σύνολο 45 ωρών, συντονίστρια Αγγελική Ράλλη).
3. *"Η μικρασιάτικη διάλεκτος των Κυδωνίων και Μοσχονησίων"* (Υπουργείο Αιγαίου και Υπουργείο Παιδείας, σύνολο 112 ώρες, συντονίστρια Αγγελική Ράλλη).
4. "*Cappadocian"*. Endangered Languages and Documentation Programme. University of London SOAS, σύνολο 40 ωρών, συντονιστές Mark Janse, Αγγελική Ράλλη και Δημήτρης Παπαζαχαρίου).

5. "*Διαλεκτική ποικιλία Πάτρας*" (Πανεπιστήμιο Πατρών, σύνολο 100 ωρών, συντονιστής Δημήτρης Παπαζαχαρίου).
6. "*Η διάλεκτος της Αγίας Παρασκευής Λέσβου*" (Δήμος Αγίας Παρασκευής, σύνολο 40 ωρών, συντονίστρια Αγγελική Ράλλη)
7. "*Τουρκοκρητικά Μικράς Ασίας*" (Υπουργείο Εξωτερικών, σύνολο 32 ωρών, συντονίστριες Αγγελική Ράλλη)
8. "*Από το γλωσσικό ιδίωμα των Μεγάρων στο γλωσσικό ιδίωμα της Παλαιάς Αθήνας*" (Ίδρυμα Λεβέντη και Δήμος Μεγαρέων, σύνολο 20 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*

Ομιλητές		*Άτομα*	*Ποσοστό*
α. Ηλικία	έως 20	16	2%
	20-30	78	10%
	30-40	37	4,7%
	40-50	18	2,3%
	50-60	184	23,5%
	60-70	162	20,7%
	70-80	92	11,7%
	80-90	108	13,8%
	90 -	60	7,7%
	άγνωστο	28	3,6%
β. Φύλο	Γυναίκες	474	59,3%
	Άνδρες	325	40,7%
Σύνολο		799	100,0%

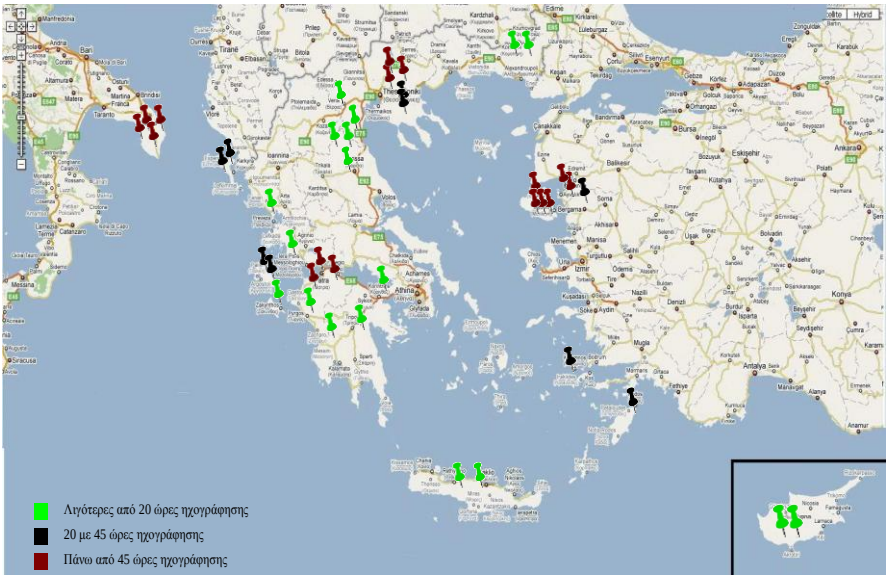
Πίνακας 2: *Πληροφορίες για τους ομιλητές*

Χρόνος Εγγραφής	*Αριθμός*	*Ποσοστό*
2000	54	7,8%
2001	18	2,6%
2002	156	22,6%
2003	169	24,5%
2004	80	11,6%
2005	12	1,7%
2006	12	1,7%
2007	58	8,4%
2008	40	5,8%
2009-2010	92	13,3%
Σύνολο	691	100,0%

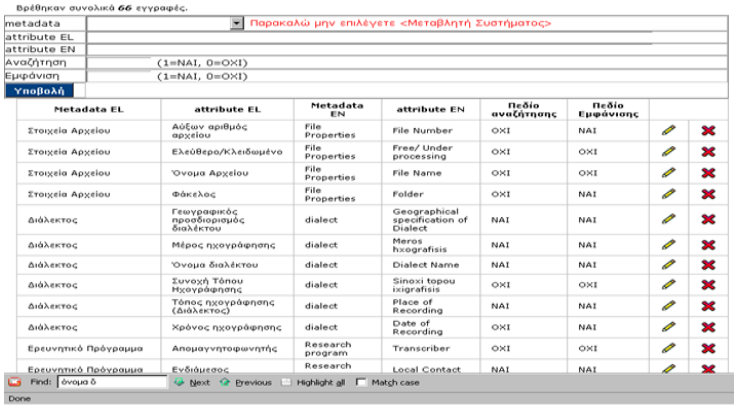
Πίνακας 3: *Πληροφορίες για τις ηχογραφήσεις*

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. Στην περίπτωση της διαλέκτου των Μεγάρων η συλλογή προφορικού υλικού αποτελείται από επικοινωνιακές περιστάσεις αυθόρμητου προφορικού λόγου με πλήθος διηγήσεων και σε ελάχιστα σημεία με ερωτηματολογικού τύπου συνεντεύξεις. Παράλληλα πραγματοποιήθηκε συλλογή γραπτού υλικού από συμβολαιογραφικά χειρόγραφα του 18ου–19ου αιώνα και άλλα διάφορα προσωπικά έγγραφα γραμμένα στην τοπική διάλεκτο. Είναι σημαντικό να υπογραμμιστεί ότι τα μεγαρικά δεδομένα περιέχουν ένα μεγάλο αριθμό από προφορικές ιστορίες και διηγήσεις παλαιότερων καταστάσεων και γεγονότων που διαδραματίστηκαν στην τριγύρω περιοχή. Η προσπάθεια αναμόχλευσης προσωπικών διηγήσεων και κυρίως παλαιότερων ιστοριών ήταν ηθελημένη επιλογή – στην πλειονότητα των ερευνητικών προγραμμάτων – για τη διαφύλαξη υλικού πολιτισμικής κληρονομιάς ταυτόχρονα με τη συλλογή των

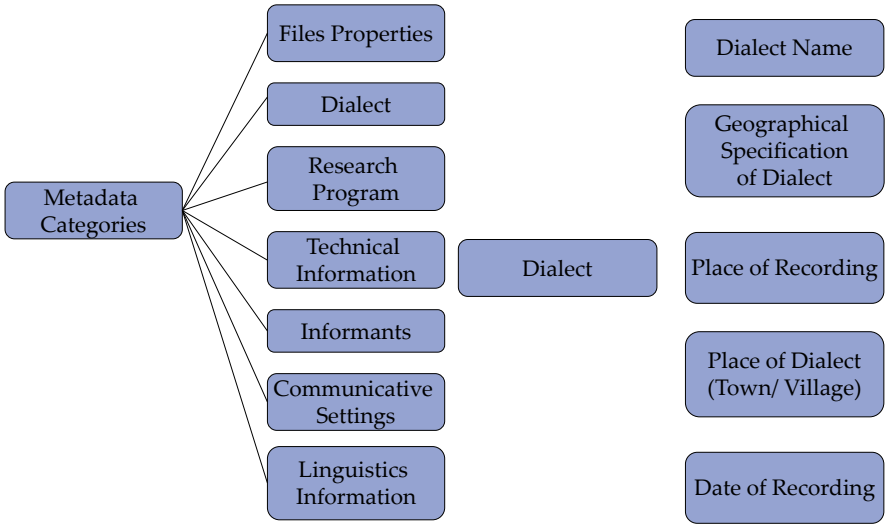
γλωσσικών δεδομένων. Πιο συγκεκριμένα, οι ηχογραφήσεις έγιναν από ερευνητές πεδίου που είχαν αποκτήσει κάποιες κοινωνικές σχέσεις και επαφές με την υπό διερεύνηση κοινότητα και τους πληροφορητές συγκεκριμένα, ή κυρίως με τη συνδρομή της φυσικής παρουσίας ενδιάμεσου, δηλαδή ενός μέλους της τοπικής κοινότητας 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.



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



Εικόνα 2: *Η διεπιφάνεια χρήσης του διαχειριστή στη βάση δεδομένων*



Εικόνα 3: [αριστερά] *Δείγμα των επτά βασικών κατηγοριών μεταδεδομένων και* [δεξιά] *δείγμα των υποκατηγοριών της κατηγορίας ΔΙΑΛΕΚΤΟΣ (Dialect)*

Στην GREED, οι διάλεκτοι είναι καταχωρημένες γεωγραφικά (καθότι αυτή η 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 τη διάλεκτο των Μεγάρων. 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.

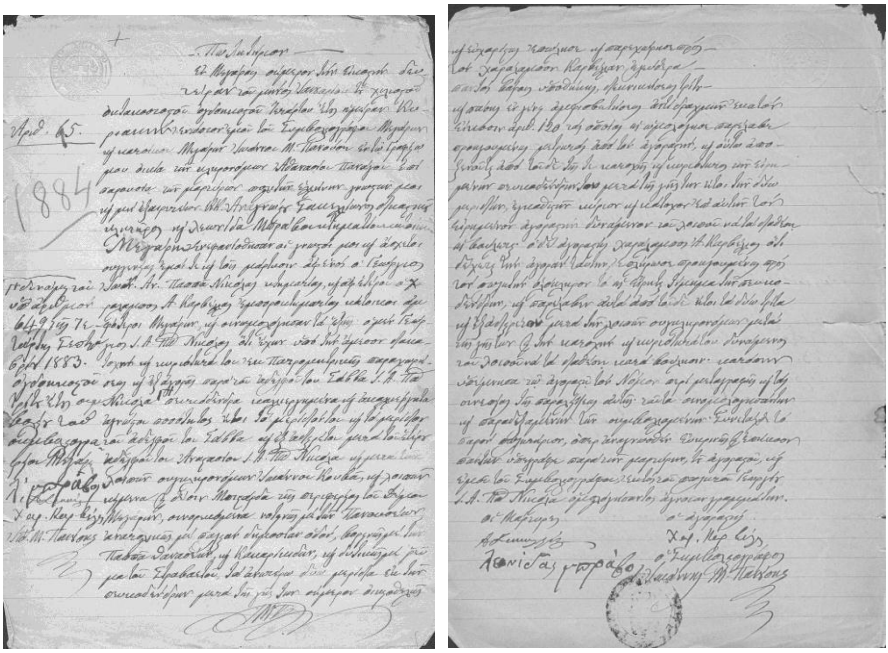
Παράλληλα έχει ξεκινήσει η ψηφιοποίηση διαφόρων χειρογράφων και σπανίων βιβλίων, ώστε σύντομα η GREED να διαθέτει και την αντίστοιχη πλατφόρμα για την αξιοποίηση του γραπτού διαλεκτικού υλικού. Στο πλαίσιο του ερευνητικού προγράμματος έγινε μια οργανωμένη προσπάθεια ψηφιοποίησης χειρογράφων, κυρίως νομικής 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 της διαλέκτου των Μεγάρων συλλέχθηκαν με τη χρήση επαγγελματικών ψηφιακών κασετοφώνων Marantz. Η επιλογή των επαγγελματικών ψηφιακών συσκευών ελήφθη με βάση τις διεθνείς προδιαγραφές για ποιοτικές ηχογραφήσεις με τις ελάχιστες δυνατές απώλειες.

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 τον 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.

[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), 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.



Εικόνα 4: *Χειρόγραφα από την περιοχή των Μεγάρων του 1884 και 1885*

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 της διαλέκτου των Μεγάρων έγινε και σε εφεδρικά συστήματα αποθήκευσης. Τυπικοί στόχοι επεξεργασίας συμπεριλαμβάνουν την ορθή ονοματοδοσία, διαχωρισμό καναλιών, αφαίρεση προσωπικών πληροφοριών, ενίσχυση των χαμηλής έντασης ηχογραφήσεων, μείωση του θορύβου και καθαρισμό του σήματος από έντονους μικροφωνισμούς.

Επομένως τα ηλεκτρονικά αρχεία της διαλέκτου των Μεγάρων 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

(γ) Κείμενα και χειρόγραφα: κείμενα που έχουν γραφτεί πρωταρχικώς στη διάλεκτο.

ΧΡΟΝΟΣ	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10	10 - 12
ΣΥΜΜΕΤΟΧΗ Α ΟΜΙΛΗΤΗ	!	!	!	!	!	!
ΣΥΜΜΕΤΟΧΗ Β ΟΜΙΛΗΤΗ	vv	v	v	v	v	v
ΥΠΑΡΞΗ ΕΝΑΛΛΑΓΩΝ	>7	4 - 5	>7	>7	>7	4 - 5
ΥΠΑΡΞΗ ΔΙΑΚΟΠΩΝ	3 - 4	0	3 - 4	3 - 4	3 - 4	1 - 2
ΥΠΑΡΞΗ ΑΛΛΗΛΕΠΙΚΑΛΥΨΕΩΝ	3 - 4	0	3 - 4	3 - 4	3 - 4	1 - 2
ΓΕΝΙΚΗ ΠΟΙΟΤΗΤΑ ΗΧΟΓΡΑΦΗΣΗΣ	Κ	Κ	Κ	Κ	Κ	Κ
ΠΟΙΟΤΗΤΑ ΗΧΟΓΡΑΦΗΣΗΣ Α' ΟΜΙΛΗΤΗ	Κ	Κ	Κ	Κ	Κ	Κ
ΠΟΙΟΤΗΤΑ ΗΧΟΓΡΑΦΗΣΗΣ Β' ΟΜΙΛΗΤΗ	Κ	Κ	Κ	Κ	Κ	Κ
ΕΠΙΚΟΙΝΩΝΙΑΚΗ ΠΕΡΙΣΤΑΣΗ	Φ	Φ	Φ	Φ	Φ	Φ
ΣΥΜΜΕΤΟΧΗ ΕΡΕΥΝΗΤΗ	Ως Β	Ως Β	Ως Β	Ως Β	Ως Β	Ως Β
ΕΝΔΙΑΦΕΡΟΝΤΑ ΠΡΟΣΩΔΙΑΚΑ ΦΑΙΝΟΜΕΝΑ	Ο	Ο	Ο	Ο	Ο	Ο
ΕΝΔΙΑΦΕΡΟΝΤΑ ΦΩΝΟΛΟΓΙΚΑ ΦΑΙΝΟΜΕΝΑ	Ν	Ν	Ν	Ν	Ν	Ν
ΕΝΔΙΑΦΕΡΟΝΤΑ ΜΟΡΦΟΛΟΓΙΚΑ ΦΑΙΝΟΜΕΝΑ	Ν	Ν	Ο	Ν	Ν	Ν
ΕΝΔΙΑΦΕΡΟΝΤΑ ΣΥΝΤΑΚΤΙΚΑ ΦΑΙΝΟΜΕΝΑ	Ο	Ο	Ο	Ο	Ο	Ο
ΕΝΔΙΑΦΕΡΟΝΤΑ ΛΕΞΙΚΑ ΦΑΙΝΟΜΕΝΑ	Ο	Ο	Ο	Ο	Ν	Ν
ΕΝΔΙΑΦΕΡΟΝΤΑ ΠΡΑΓΜΑΤΟΛΟΓΙΚΑ ΦΑΙΝΟΜΕΝΑ	Ο	Ο	Ο	Ο	Ο	Ο

Εικόνα 5: *Πρωτόκολλο χαρτογράφησης μεγαρικού διαλεκτικού ηχητικού αρχείου*

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, πώς θα λειτουργήσει η φωνητική κωδικοποίησης στα λογισμικά PRAAT και ELAN.

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 βάσεων γραπτών και προφορικών δεδομένων και μας επιτρέπει 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 σε όλες τις περιπτώσεις των διαλεκτικών δεδομένων των Μεγάρων το πρωτόκολλο χαρτογράφησης. Ανά δύο λεπτά χαρακτηρίζεται το αρχείο, δηλαδή «χαρτογραφείται» 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.). Να σημειωθεί ότι η χαρτογράφηση των ηχητικών αρχείων εφαρμόστηκε για πρώτη φορά εξ ολοκλήρου στα διαλεκτικά δεδομένα των Μεγάρων και στοχεύουμε να το εφαρμόσουμε στο εγγύς μέλλον και στα υπόλοιπα διαλεκτικά δεδομένα της βάσης του GREED.

Η ηλεκτρονική βάση GREED έχει σχεδιαστεί με στόχο την δημιουργία μιας ηλεκτρονικής βάσης προφορικού υλικού· στην παρούσα στιγμή βρίσκεται στο στάδιο υλοποίησης και αναβάθμισης και δεν έχουν ολοκληρωθεί όλα τα συστατικά της. Δυστυχώς στην τρέχουσα έκδοσή της δεν παρέχεται η δυνατότητα για εισαγωγή και καταλογογράφηση των ψηφιακών χειρογράφων (για περισσότερα βλ. Σύρκου στον παρόντα τόμο). Μετά την απαραίτητη επεξεργασία των εικόνων, έγινε η περιγραφή των χειρογράφων με κάποια μεταδεδομένα έχοντας υπόψιν τις συνοδευτικές πληροφορίες που δίνονται από τις προαναφερθείσες διαδικτυακές συλλογές χειρογράφων (βλ. υποσημ. 1).

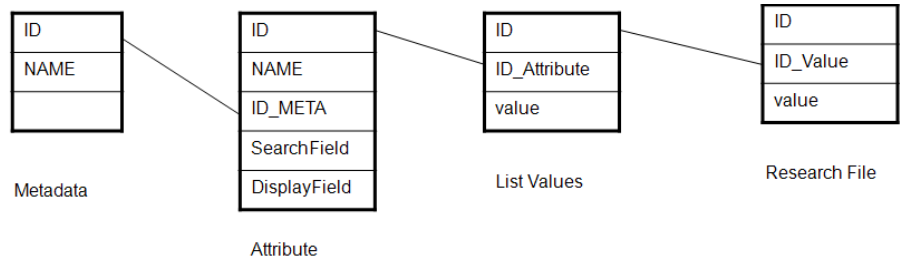
4. Διαχείριση και ιστοσελίδα

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. Αν 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.

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

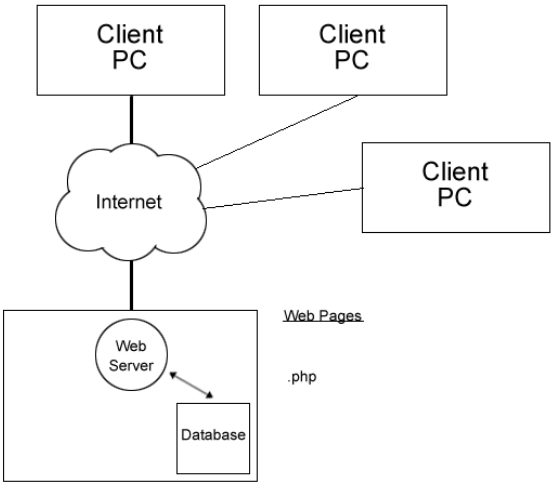
Όλα τα αντικείμενά (*Metadata, Metadatadetails, mdListValues* [προ-εισαγμένες τιμές] και *FileAttribs* [πίνακας με όλα τα αρχεία]) είναι συνδεμένα αναμεταξύ τους με μια σχέση 'ένα προς πολλά', για παράδειγμα η τιμή 'dialect name' του Metadatadetails είναι συνδεμένη με τις τιμές 'Ποντιακά', 'Λεσβιακά', '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.



Εικόνα 6: *Δείγμα της αρχιτεκτονικής «ένα προς πολλά» του συστήματος*[3]

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.

[3] **Metadata***: Κατηγορίες των μεταδεδομένων όπως «Διάλεκτος», «Ερευνητικό Πρόγραμμα». **Attribute***: Υποκατηγορίες των μεταδεδομένων όπως «Όνομα διαλέκτου», «Μέρος ηχογράφησης». **List Values***: Δημιουργία κατηγοριών για κάθε attribute. Για παράδειγμα για το attribute «Όνομα διαλέκτου» υπάρχουν οι τιμές «Μεγαρικά», «Ποντιακά», «Λεσβιακά» κλπ. **Research File**: Είναι το αρχείο του ερευνητή με τις τιμές που συμπλήρωσε.



Εικόνα 7: *Η αρχιτεκτονική της βάσης δεδομένων GREED.*

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 της διαλέκτου των Μεγάρων

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:
(1) Να είναι λογισμικά ανοιχτού κώδικα και ελεύθερα ως προς τη χρήση
(2) Να παρέχουν μεγάλο εύρος σχεδιαστικών παραμέτρων
(3) Να υποστηρίζουν αρχεία από διαφορετικά λογισμικά που χρησιμοποιούνται για τον σχολιασμό αρχείων σε διαφορετικά γλωσσολογικά επίπεδα
(4) Να επιτρέπουν την χρήση πιθανών add-ons και plug-ins
(5) Να προσφέρεται συνεχής υποστήριξη από τους προγραμματιστές/ παραγωγούς του λογισμικού
(6) Να είναι πολυγλωσσικά ή τουλάχιστον σε αγγλική έκδοση και να επιτρέπουν τη χρήση του Unicode πρωτοκόλλου

5.1. Για ποιο λόγο επιλέχθηκαν ELAN και PRAAT?

Το ELAN είναι ένα επιτυχημένο λογισμικό με δεκαετή παρουσία στον τομέα των εργαλείων ομιλίας και είναι εξοπλισμένο με συνεχή αναβάθμιση και υποστήριξη από το

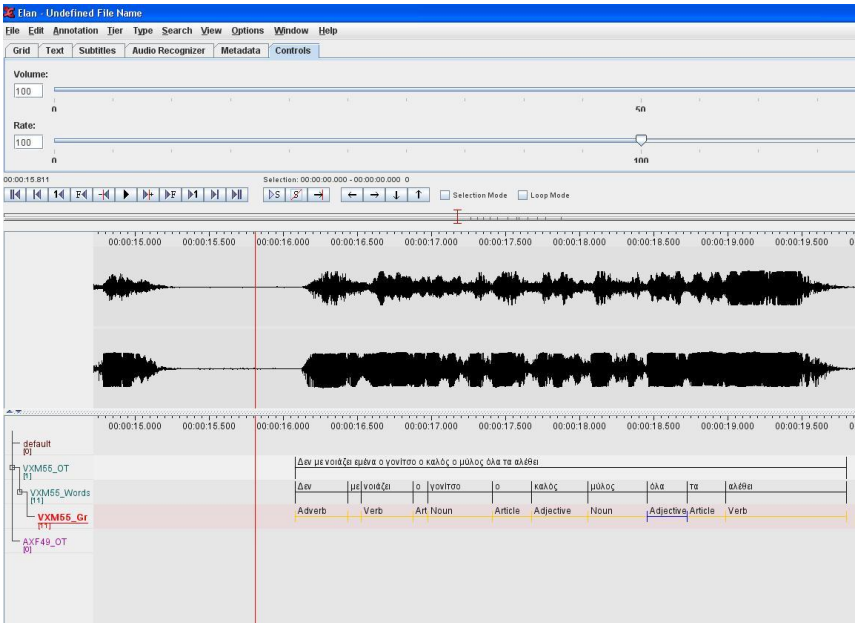
ερευνητικό κέντρο Max Planck Institute[4]· με το ELAN ο χρήστης μπορεί να προσθέσει έναν απεριόριστο αριθμό σχολίων σε αρχεία ήχου, αλλά και εικόνας. Οι προγραμματιστές παρέχουν εξαιρετικά λεπτομερείς οδηγίες χρήσης για σχολιασμό (το εγχειρίδιο χρήσης του ELAN) και ένα σημαντικό σύνολο από συνοδευτικό λογισμικό, όπως μετατροπέας των δημοφιλών προγραμμάτων (ECONV), έναν δημιουργό περιγραφής μεταδεδομένων σε λογική βάσης (IMDI), ένα εργαλείο αναζήτησης σχολιασμών (ANNEX) μεταξύ άλλων. Αυτό το λογισμικό παρέχεται δωρεάν από την ιστοσελίδα του ιδρύματος. Ένας μεγάλος αριθμός ερευνητών που συσχετίζονται με τη ψυχογλωσσολογία, την εργαστηριακή φωνητική/ φωνολογία, τη διαλεκτολογία και την ανάλυση λόγου έχει χρησιμοποιήσει ήδη το ELAN.

Τα πλεονεκτήματα του ELAN είναι αρκετά δυνατά. Ο σχολιασμός ή μεταγραφή του αρχείου μπορεί να είναι μια πρόταση, μία λέξη, μια γραμματική κατηγορία, ένα φώνημα, ένα σχόλιο, μια μετάφραση ή μια περιγραφή οποιαδήποτε χαρακτηριστικού εντοπίζεται στο αρχείο. Τα σχόλια μπορούν να ενταχθούν σε πολλαπλές σειρές σχολιασμού (tiers), το οποίο μας επιτρέπει να εισάγουμε ή να εξάγουμε tiers προς και από το PRAAT. Το κείμενο των σχολιασμών μπορεί να είναι και σε Unicode κωδικοποίηση και η μεταγραφή αποθηκεύεται σε αρχείο μορφής XML· είναι κοινός τόπος η χρήση του XML για την περιγραφή μεταδεδομένων στο διαδίκτυο με τη συνοδεία πανίσχυρων εργαλείων για την ανάλυση δεδομένων και είναι σχεδιασμένο πρωτίστως για να διαβάζεται από υπολογιστή παρά από άνθρωπο. Το ELAN παρέχει διαφορετικές οπτικές γωνίες παρακολούθησης των σχολιασμών· κάθε οπτική είναι συνδεμένη και συγχρονισμένη με μια μπάρα αναπαραγωγής. Το ELAN είναι γραμμένο σε γλώσσα JAVA και ο πηγαίος κώδικας είναι διαθέσιμος για τον καθένα για μη-εμπορική χρήση. Επίσης είναι σημαντικό ότι τρέχει σε όλα τα δημοφιλή λειτουργικά συστήματα, όπως Windows, Mac OS X και Linux OS.

Τα βασικά χαρακτηριστικά του προγράμματος επιτρέπουν στον χρήστη να καλύψει μεγάλο εύρος των αναγκών του. Το πρόγραμμα έχει τα ακόλουθα:

(1) πλοήγηση στο αρχείο ήχου/ βίντεο με διαφορετικά βήματα και τρόπους
(2) εύκολη και γρήγορη πλοήγηση ανάμεσα στους σχολιασμούς
(3) ύπαρξη κυματομορφής για τα .wav αρχεία
(4) υποστήριξη των εγγράφων οδηγών (templates) για σχολιασμό
(5) διάφοροι μέθοδοι εισαγόμενου από διαφορετικού τύπου λογισμικά
(6) δυνατότητα αναζήτησης κανονικών εκφράσεων στα πολλαπλά tiers εντός ενός αρχείου μεταγραφής ή μιας ομάδας αρχείων μεταγραφής
(7) υποστήριξη από τον χρήστη φτιαγμένων ελεγχόμενων λεξιλογίων
(8) εισαγωγή και εξαγωγή αρχείων από Shoebox/Toolbox, CHAT, Transcriber (εισαγωγή μόνο), Praat και csv/tab-delimited αρχείων σχολιασμού
(9) εξαγωγή του σχολιασμού σε μορφή διασχολιασμένου κειμένου, html, smil και υποτιτλισμένου κειμένου
(10) εκτύπωση των σχολιασμών αυτόνομα
(11) πολλαπλή χρήση αναίρεσης και επανάληψης ενέργειας

[4] http://www.mpg.de/english/portal/index.html

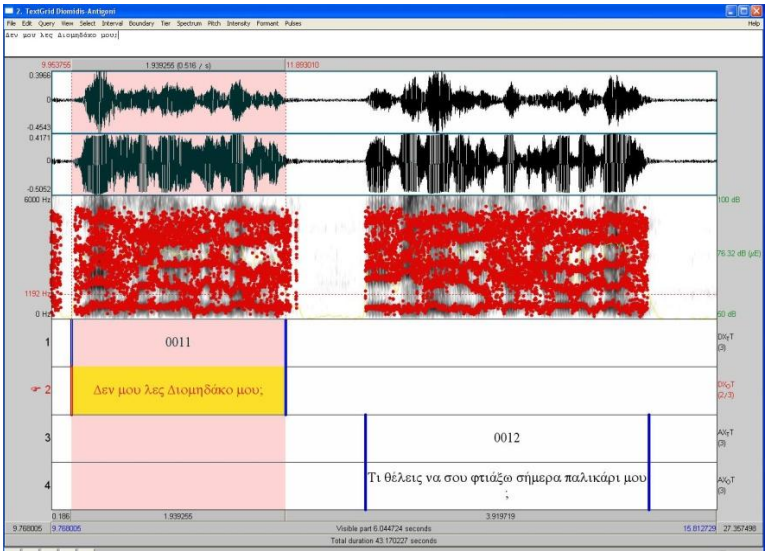


Εικόνα 8: *Εικόνα από απομαγνητοφώνηση ηχητικού αρχείου με το* ELAN

Από την άλλη πλευρά υπάρχει το χαρακτηριζόμενο ως αντίπαλο δέος[5], το PRAAT[6], που αναπτύχθηκε από τους φωνητικούς Paul Boersma και David Weenink (University of Amsterdam) και είναι ένα πανίσχυρο εργαλείο – με σχεδόν εικοσαετή ιστορία ύπαρξης – στην κορυφή της τεχνολογίας επεξεργασίας ομιλίας· είναι ένα εξαιρετικά (υπολογιστικά) ελαφρύ λογισμικό μέσω του οποίου δίνεται η δυνατότητα ανάλυσης, σύνθεσης, επεξεργασίας και χειραγώγησης της ομιλίας, καθώς και δημιουργίας υψηλής ευκρίνειας εικόνων (οπτικοποίηση του ηχητικού σήματος). Αν και το PRAAT σχεδιάστηκε αρχικά ως εργαλείο για τη φωνητική έρευνα, μπορεί να χρησιμοποιηθεί επίσης και για φωνολογικούς λόγους, καθώς και για μεταγραφή/ απομαγνητοφώνηση. Ωστόσο, εξαιτίας κάποιων ιδιαίτερων προβλημάτων που προέκυψαν από τη χρήση Unicode γραμματοσειρών για την απεικόνιση της ελληνικής, χρησιμοποιήσαμε το PRAAT για την εναλλαγή διαλόγου (turn-taking) και τις όποιες φωνητικές μετρήσεις για τη διάλεκτο των Μεγάρων, ενώ για την πολυεπίπεδη ορθογραφική (και όχι μόνο) μεταγραφή το ELAN. Και τα δύο λογισμικά επιτρέπουν στον ερευνητή να ευθυγραμμίσει την ομιλία με το κείμενο απευθείας, με τέτοιο τρόπο που να παρέχεται γρήγορος και επιτυχής τρόπος για τον εντοπισμό των υπό αναζήτηση ηχητικών κομματιών. Επιπροσθέτως, σημαντικό πλεονέκτημα του PRAAT είναι ότι επιτρέπει στο χρήστη να χωρίσει τη μεταγραφή σε διάφορα tiers, τα οποία συνήθως αντιστοιχούν σε διαφορετικούς ομιλητές του αρχείου ή σε διαφορετικό γλωσσικό επίπεδο έρευνας ανά ομιλητή, όπως π.χ. ο επιτονισμός του ομιλητή1.

[5] Να υπογραμμίσουμε ότι διαφωνούμε με τη λογική ότι το ELAN και το PRAAT είναι εργαλεία που χρησιμοποιούνται για την επιτέλεση της ίδιας δουλειάς, καθότι ανταποκρίνονται σε διαφορετικές ανάγκες ερευνητών με κάποιες κοινές ενέργειες.

[6] http://www.fon.hum.uva.nl/praat/



Εικόνα 9: *Εικόνα ανάλυσης του ηχητικού σημείου με το PRAAT*

Συγκεντρώνοντας τα βασικά χαρακτηριστικά του λογισμικού αναφέρουμε:
(1) Ανάλυση ομιλίας, σύνθεση ομιλίας και πειράματα ακουστικής φωνητικής
(2) Διαχωρισμός και ονοματοδοσία
(3) Επεξεργασία και χειραγώγηση ηχητικού σήματος
(4) Αλγόριθμοι εκμάθησης και στατιστική
(5) Δυνατότητα προγραμματισμού, φορητότητας και προσωπικής διαμόρφωσης

5.2. Για ποιο λόγο επιλέχθηκε το Toolbox?

Τέλος, αναφέρουμε ότι για μια πρωταρχική μορφολογική και τυπολογική της διαλέκτου των Μεγάρων επιλέχθηκε το λογισμικό Field Linguist's Toolbox[7], το οποίο είναι πρόγραμμα διαχείρισης δεδομένων και εργαλείο ανάλυσης για τους ερευνητές πεδίου και τους παραδοσιακούς μορφολόγους. Είναι εξαιρετικά χρήσιμο για τη διατήρηση και δημιουργία λεξικών δεδομένων, για τη μορφολογική ανάλυση και για το διαγραμματισμένο κείμενο (interlinearizing text)[8], αλλά μπορεί να χρησιμοποιηθεί για οποιαδήποτε εικονική διαχείριση δεδομένων.

Το Toolbox είναι ένας είδος βάσης δεδομένων προσανατολισμένο στα κείμενα ή στις μεταγραφές με συνεπικουρικό σύστημα διαχείρισης των δεδομένων με πρόσθετες σχεδιαστικές προσθήκες για να καλυφθούν οι ανάγκες ενός ερευνητή πεδίου. Για την ευκολία της χρήσης του, το πακέτο εγκατάστασης του Toolbox περιλαμβάνει ήδη προεγγραφές κατηγοριών λεξικού και ορισμούς βάσεων δεδομένων και ένα τυπικό λεξιλόγιο και ένα σώμα κειμένων.

Η βάση του Toolbox με το σύστημα διαχείρισης προσφέρει πολύ ισχυρή χρηστικότητα όπως παραμετροποιήσιμη ταξινόμηση, πολλαπλές οπτικές γωνίες των

[7] http://www.sil.org/computing/toolbox/
[8] Αφορά το κείμενο που περιέχει το αρχικό κείμενο, την ανάλυση μορφημάτων, τις γραμματικές κατηγορίες, τη μετάφραση και ό,τι επιπλέον προσθέσει ο ερευνητής. Βλ. την εικόνα του Toolbox.

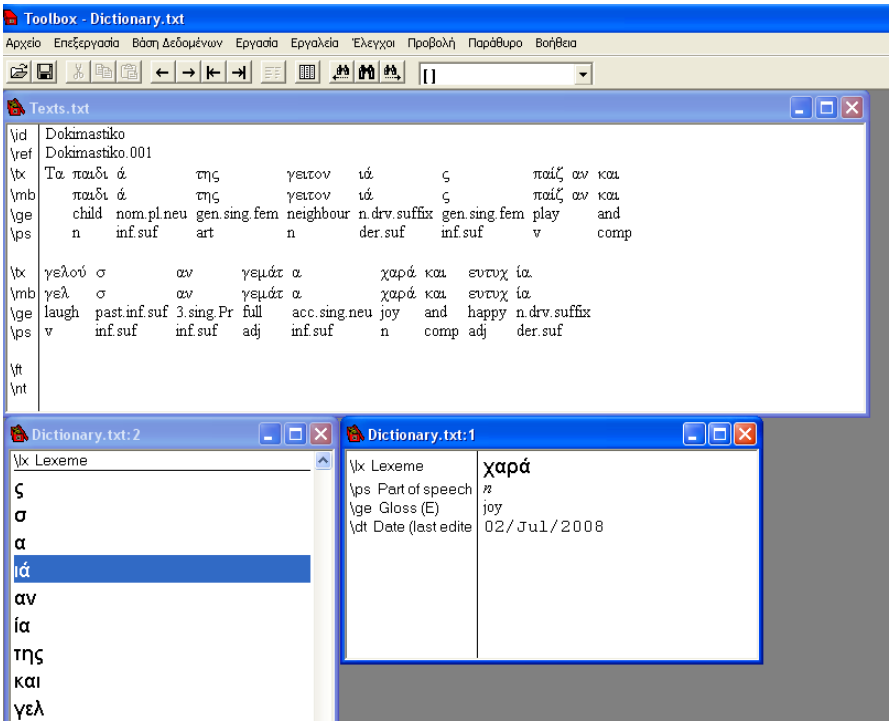
δεδομένων, οπτική εξερεύνησης για την παρατήρηση των δεδομένων σε διαφορετικές μορφές και πολλαπλά φίλτρα για την παρουσίαση των υποσυνόλων των δεδομένων. Επίσης παρέχει τη δυνατότητα για μικρά προγράμματα κώδικα, ώστε να αυτοματοποιηθεί κάποια διαδικασία με αρκετές επιλογές παραμετροποίησης. Παρέχει την επιλογή χρήσης Unicode γραμματοσειράς και συμβόλων. Να σημειωθεί ότι έχουμε ήδη συνεργαστεί με τη σχεδιαστική ομάδα του Toolbox και πλέον υφίσταται ελληνική έκδοση του προγράμματος μαζί με τον οδηγό εκμάθησης και το εγχειρίδιο χρήσης.

Μία από τις αρετές του Toolbox είναι η δυναμική γλωσσολογική λειτουργικότητα. Περιλαμβάνει ένα μορφολογικό αναλυτή που μπορεί να επεξεργαστεί σχεδόν όλα τα μορφολογικά φαινόμενα. Χρησιμοποιεί μια τεχνική παραγωγής και ανάλυσης διαγραμμισμένου κειμένου ορισμένη από τον χρήστη, η οποία χρησιμοποιεί τον μορφολογικό αναλυτή, το λεξικό και τις λεξικές φόρμουλες για να παράγει μορφολογικά σχολιασμένο κείμενο. Το διαγραμματισμένο κείμενο μπορεί να εξαχθεί σε μορφή κατάλληλη και για κείμενα και άρθρα γλωσσολογικά. Το Toolbox εσωκλείει και άλλες δυνατότητες εξαγωγής όχι μόνο για άλλα λογισμικά, αλλά για την παραγωγή ενός δημοσιεύσιμου λεξιλογίου από το λεξιλόγιο των δεδομένων.

Αν και το Toolbox είναι πολύ ισχυρό εργαλείο και ελαφρύ υπολογιστικά, είναι σχεδιασμένο να είναι εύκολο στη χρήση. Ο χρήστης μπορεί να χρησιμοποιήσει τις απλές αρχικές ρυθμίσεις και βαθμιαία να προσθέσει μερικά επιπρόσθετα δυνατά χαρακτηριστικά, εφόσον το επιθυμεί. Το πακέτο του Toolbox παρέχει και ένα εικονογραφημένο πακέτο εκπαίδευσης του χρήστη, αλλά ταυτόχρονα οδηγό για τη χρήση του σε διδασκαλία, το οποίο αφορά μια «φτιαχτή» γλώσσα και είναι μεταφρασμένο στα ελληνικά. Το Toolbox διανέμεται δωρεάν στους χρήστες και ευτυχώς τυχόν σοβαρά προβλήματα επιλύονται το συντομότερο δυνατό. Οι συνεχείς αναβαθμίσεις του προγράμματος και η συνεχής υποστήριξη επηρέασαν θετικά την επιλογή μας για το Toolbox.

6. Διατήρηση, αδειοδότηση και back-up

Όλα τα προφορικά και γραπτά δεδομένα της διαλέκτου των Μεγάρων βρίσκονται αποθηκευμένα μαζί με τα υπόλοιπα δεδομένα της βάσης σε ψηφιακή μορφή σε σκληρούς δίσκους πολλαπλής σύνδεσης και ασφάλειας. Φυσικά και δεν μπορούμε να κρατήσουμε γνήσια αντίγραφα των ηχογραφήσεων, όπως και των σπανίων χειρογράφων, που μετά την ψηφιοποίηση τους παραδόθηκαν ξανά στους κατόχους τους· εντούτοις ό,τι διαλεκτικό υλικό έχει εκδοθεί ή δωρισθεί στο εργαστήριο παραμένει στο εργαστήριο. Για την χρήση των αντιγράφων από τα μεγαρικά δεδομένα οι ερευνητές αποκτούν πρόσβαση στα δεδομένα μετά από παροχή άδειας από τον διευθυντή του προγράμματος. Παράλληλα καταχωρούμε και τις αιτήσεις των φορμών συγκατάθεσης των ομιλητών.



Εικόνα 10: *Εικόνα μορφολογικής ανάλυσης με το Toolbox*

Σύμφωνα με τη φόρμα δίνεται η συγκατάθεση να επεξεργαστούν οι προσωπικές πληροφορίες των ομιλητών για τους στόχους της έρευνας· γνωστοποιείται ότι αυτές οι πληροφορίες είναι ανώνυμες σε όλα τα επίπεδα χρήσης της έρευνας και ότι αντιμετωπίζονται ως άκρως εμπιστευτικά στοιχεία και διαχειρίζονται σε πλήρη συμφωνία με τις διατάξεις του Νόμου για την προστασία του πολίτη από την επεξεργασία δεδομένων προσωπικού χαρακτήρα (Ν. 2472/97).

Όλα τα υπολογιστικά αρχεία που σχετίζονται με τα μεγαρικά δεδομένα, αλλά και το σύνολο των δεδομένων της βάσης βρίσκονται αποθηκευμένα σε δύο σκληρούς δίσκους διαδικτύου υπό την επίβλεψη του τεχνικού του εργαστηρίου· εντούτοις χρησιμοποιούμε μια εβδομαδιαία χειροκίνητη διατήρηση αρχείων στα δεδομένα αποθηκεύοντάς τα στο υψηλής ασφάλειας σύστημα αποθήκευσης NAS Storage. Όταν γίνονται σημαντικές και μεγάλες αλλαγές στη βάση μας, σημαντικές αναβαθμίσεις του server ή συντήρηση του συστήματος, δημιουργούμε δύο εικόνες-αντίγραφα του συστήματος της βάσης· εντούτοις, στην περίπτωση αντιμετώπισης σοβαρού προβλήματος με τη νεότερη έκδοση, υφίσταται πάντα η προηγούμενη έκδοση, για να επιστρέψουμε σε αυτή με ασφάλεια.

7. Μελλοντικά σχέδια

Η ηλεκτρονική διαλεκτική βάση GREED και η συλλογή υλικού από τη διάλεκτο των Μεγάρων 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 τη διάλεκτο των Μεγάρων, αναβαθμίσαμε σημαντικά 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 τις 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] Μετά την έκδοση του παρόντος τόμου για τη διάλεκτο των Μεγάρων, σχεδιάζουμε την έκδοση λεξικών, λεξιλογίων και γραμματικών για τις υπόλοιπες διαλέκτους που έχουμε σε μεγάλο εύρος.

[Έρευνα] Εύρεση επιδοτήσεων ερευνητικών προσπαθειών για οικονομική υποστήριξη με στόχο τη βελτίωση και εξέλιξη της ηλεκτρονικής βάσης GREE.D.

[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. In Beal et al. (Eds.), *Creating and digitalizing Language Corpora, Vol. 2 (*pp. 16-48). Palgrave McMillan Publication.

Anderson, J., Beavan, D. & Kay, C. (2007). SCOTS: Scottish Corpus of Texts and Speech. In

Beal J. et al. (Eds.), *Creating and digitalizing Language Corpora, Vol. 1* (pp. 17-34). Palgrave McMillan Publication.

Anderwald, L. & Wagner, S. (1997). FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data. In Beal et al. (Eds.), *Creating and digitalizing Language Corpora, Vol.1*. 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. In Beal et al. (Eds.), *Creating and digitalizing Language Corpora, Vol.1*. 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. In Beal et al. (Eds.), *Creating and digitalizing Language Corpora, Vol.2* (pp. 82-104). Palgrave McMillan Publication.

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

Ralli, A. (2006). Syntactic and Morphosyntactic Phenomena in Modern Greek Dialects: The State of the Art. *Journal of Greek Linguistics, 7*, 121-159.

Ακαδημία Αθηνών (1933-). *Ιστορικόν Λεξικόν της Νέας Ελληνικής Γλώσσης, της τε Κοινώς Ομιλουμένης και των Ιδιωμάτων*. Αθήνα

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

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

Κοντοσόπουλος, Ν. (2006). *Διάλεκτοι και ιδιώματα της Νέας Ελληνικής* (4η έκδοση). Αθήνα: εκδόσεις 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. (1984-1992). *Ιστορικόν Λεξικόν των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας,* τόμ. Α-Ε. Αθήνα.

Καραναστάσης, Α. (1997). *Γραμματική των Ελληνικών Ιδιωμάτων της Κάτω Ιταλίας*. Αθήνα.

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

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

ΑΘΑΝΑΣΙΟΣ Ν. ΚΑΡΑΣΙΜΟΣ
ΕΡΓΑΣΤΗΡΙΟ ΝΕΟΕΛΛΗΝΙΚΩΝ ΔΙΑΛΕΚΤΩΝ
ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
akarasimos@upatras.gr