Προκλήσεις επισημείωσης ενός πολυ-διαλεκτικού, πολυ-επίπεδου σώματος γραπτών και προφορικών κειμένων των Νεοελληνικών Διαλέκτων

Αθανάσιος Καρασίμος[1,3], Ελένη Γαλιώτου[2], Νικήτας Καρανικόλας[2], Γιώργος Κορωνάκης[2], Κώστας Αθανασάκος[2], Δημήτρης Παπαζαχαρίου[1], Αγγελική Ράλλη[1]
Πανεπιστήμιο Πατρών[1], ΤΕΙ Αθηνών[2], Ακαδημία Αθηνών[3]
akarasimos@academyofathens.gr, egali@teiath.gr, nnk@teiath.gr, gkoronakis@gmail.com, k.athanasakos@gmail.com, papaz@upatras.gr, ralli@upatras.gr

0. Περίληψη

Στην παρούσα μελέτη που αποτελεί μέρος του προγράμματος «AMIGRE – Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιάτικης Ελληνικής Γλώσσας» παρουσιάζεται η επισημείωση ενός διαλεκτικού σώματος αρχείων που διαφέρει από τα υπόλοιπα σε δύο βασικά σημεία. Πρώτον, έχει συμπεριληφθεί ένα μεγάλο εύρος δειγμάτων από τις διαλεκτικές ποικιλίες του Πόντου, της Καππαδοκίας και του Αϊβαλιού και αποτελεί την πιο ευρεία κάλυψη των συγκεκριμένων διαλεκτικών περιοχών σε προφορικό και γραπτό υλικό. Επιπροσθέτως, παρέχονται τα αποτελέσματα από μια συστηματοποιημένη προσπάθεια επισημείωσης με κοινή στρατηγική σε γραπτά και προφορικά δεδομένα.

Το συγκεκριμένο διαλεκτικό σώμα κειμένων έχει μια μεγάλη ποικιλία χαρακτηριστικών που συνδυαστικά δημιουργούν ένα εξειδικευμένο εργαλείο για τη γλωσσολογική και διαλεκτολογική μελέτη. Αυτά τα χαρακτηριστικά είναι μεταξύ άλλων: γλωσσολογικό περιεχόμενο (διάλεκτοι από τρεις περιοχές που συσχετίζονται), multi-tiers επισημείωση (μεταγραφή και απεικόνιση προφορικού και γραπτού υλικού με βάση διεθνή στάνταρ, π.χ. SAMPA), πολυεπίπεδα μεταδεδομένα (TEI), αναβαθμισμένη μηχανή αναζήτησης (βασισμένη σε γλωσσολογική πληροφορία και μεταδεδομένα), ψηφιακές συλλογές χειρογράφων και ηχητικών αρχείων, χάρτες απεικόνισης των δεδομένων και συνοδευτικό multimedia τρι-διαλεκτικό λεξικό.

Σημαντικά ζητήματα για την επισημείωση σε φωνολογικό επίπεδο αντιμετωπίστηκαν κατά τη μελέτη καθότι έγινε μια συστηματική προσπάθεια να ενοποιηθούν όλες οι διαφορετικές μεταγραφές διαλεκτικού γραπτού υλικού που δεν υπήρχε κοινή στρατηγική απεικόνισης. Παράλληλα προτείνεται πολυεπίπεδη φωνολογική (παράλληλα με μορφολογική) επισημείωση του σώματος κειμένων καθιερώνοντας ένα βασικό πρότυπο επισημείωσης διαλεκτικού υλικού για τις Νεοελληνικές Διαλέκτους σε καθιερωμένα λογισμικά ανάλυσης ομιλίας.

Λέξεις-Κλειδιά: Ποντιακά, Καππαδοκικά, Αϊβαλιώτικά, επισημείωση, σώματα κειμένων, Υπολογιστική Διαλεκτολογία

1. Εισαγωγή

1.1. THALIS project AMiGrE

Στην παρούσα μελέτη που αποτελεί μέρος του προγράμματος «AMIGRE – Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιάτικης Ελληνικής Γλώσσας» παρουσιάζεται

η επισημείωση ενός διαλεκτικού σώματος αρχείων που διαφέρει από τα υπόλοιπα σε δύο βασικά σημεία. Από ένα μεγάλο εύρος δειγμάτων από τις διαλεκτικές ποικιλίες του Πόντου, της Καππαδοκίας και του Αϊβαλιού παρέχονται τα αποτελέσματα από μια συστηματοποιημένη προσπάθεια επισημείωσης με κοινή στρατηγική σε γραπτά και προφορικά δεδομένα.

Πιο συγκεκριμένα σκοπός του ερευνητικού προγράμματος είναι να μελετήσει συστηματικά τα Ποντιακά, τα Καππαδοκικά και τα Αϊβαλιώτικα, τρεις γλωσσικές ποικιλίες που απειλούνται με εξαφάνιση. Μεταξύ άλλων, επιδιώκεται η μελέτη των συγκεκριμένων διαλέκτων με σκοπό να αποκαλυφθούν οι ομοιότητες και οι διαφορές τους σε συγχρονικό επίπεδο, να επισημανθεί η εξέλιξή τους, να χαρτογραφηθεί η διαφοροποίησή τους, αλλά και να εντοπισθούν τα σημαντικότερα χαρακτηριστικά τους σε σχέση με τις υπόλοιπες Νεοελληνικές διαλέκτους. Επιπροσθέτως, γίνεται προσπάθεια για μία εμπεριστατωμένη ανάλυση συγκεκριμένων φωνητικών/ φωνολογικών, μορφολογικών και σημασιολογικών φαινομένων, καθώς και της επιρροής διαφορετικά τυπολογικών γλωσσικών συστημάτων, μιας και είναι εμφανής η επίδραση της Τουρκικής (συγκολλητική γλώσσας) στις συγκεκριμένες διαλέκτους της Νέας Ελληνικής (διαχυτικής γλώσσα). Για αυτό το λόγο έχει γίνει συστηματική αρχειοθέτηση και ψηφιοποίηση προφορικού και γραπτού υλικού μεγάλου εύρους και έχει οργανωθεί σε μία ψηφιακή βάση δεδομένων. Ένα σημαντικό μέρος του πρωτογενούς υλικού θα μεταγραφεί και θα σχολιαστεί με την χρήση του πιο σύγχρονου εξοπλισμού. Γραπτό υλικό θα ψηφιοποιηθεί, και ένα μέρος αυτού, που θα επιλεγεί σύμφωνα με αυστηρά ποιοτικά κριτήρια (χρονολόγηση, προέλευση, αξιοπιστία), θα μεταγραφεί.

1.2. Σώματα γραπτών κειμένων vs. Σώματα προφορικών κειμένων

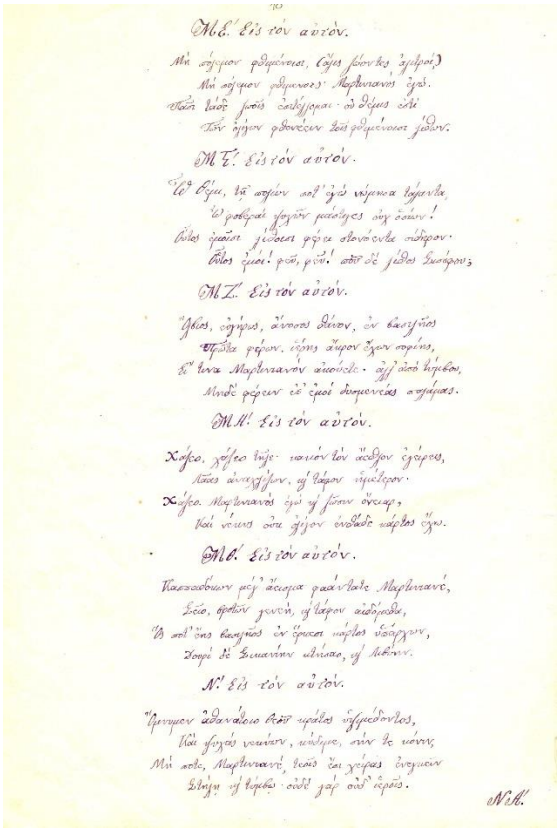
Η συνέπεια στην επισημείωση σωμάτων κειμένων είναι μια ουσιώδης ιδιότητα για τις πολλαπλές χρήσεις επισημειωμένων σωμάτων κειμένων στην υπολογιστική και θεωρητική γλωσσολογία. Παλαιότερες έρευνες εντόπισαν προβλήματα σε μορφολογική και POS επισημείωση (van Halteren 2000, Eskin 2000, Dickinson & Meurers 2003), ενώ πιο πρόσφατες εντόπισαν λάθη σε συντακτικό και δομικό επίπεδο (Ule & Simov 2004, Dickinson 2005).

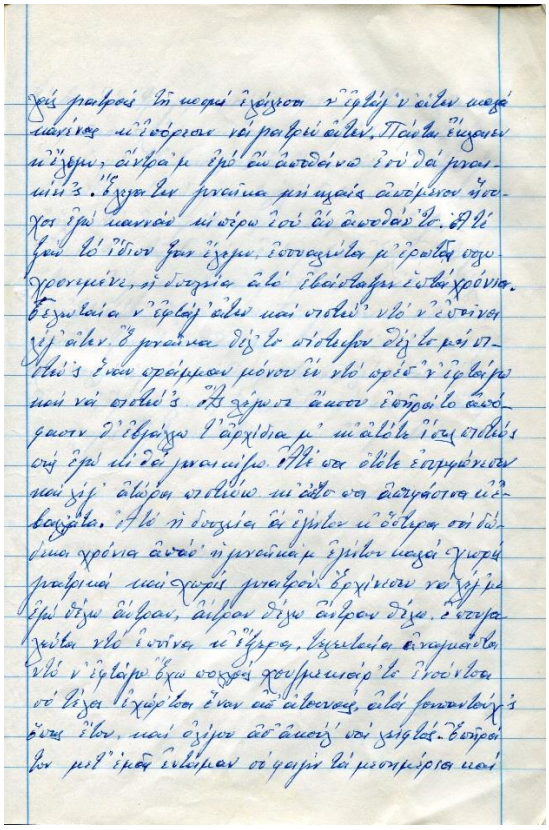
Τα *σώματα γραπτών κειμένων* είναι σαφώς περισσότερα ανά γλώσσα παγκοσμίως και σημαντικό κομμάτι της υπολογιστικής και διακειμενικής γλωσσολογίας έχει γίνει για την επισημείωση και τη αξιποίησή τους. Από την άλλη τα *σώματα προφορικών κειμένων* υστερούν σε όγκο και διαφέρουν σε πολλά σημεία από τα αντίστοιχα γραπτά, ωστόσο υπάρχει έλλειψη συντονισμένης επισημείωσης, ενώ το ζήτημα της ανίχνευσης σφαλμάτων στον σχολιασμό της ομιλούμενης γλώσσας σωμάτων δεν έχει ακόμη αντιμετωπιστεί συστηματικά. Αυτό είναι σημαντικό δεδομένου ότι τα σώματα προφορικών κειμένων αυξάνονται ιδιαίτερα, όπως φαίνεται στο Linguistic Data Consortium (www.ldc.upenn.edu). Το πρόβλημα εντείνεται όταν γίνεται προσπάθεια δημιουργίας κοινής στρατηγικής επισημείωσης σε σώματα προφορικών και γραπτών κειμένων και δη όταν το αντικείμενο είναι ιδιαίτερα εξειδικευμένο, όπως το προαναφερθέν διαλεκτικό σώμα.

2. State-of-the-Art σχεδιασμός συστήματος

2.1 Η φύση των δεδομένων

Το σώμα προφορικών κειμένων του έργου AMiGre αποτελείται από περίπου 180 ώρες (δηλαδή 60 ώρες ανά διάλεκτο), όπως αυτά συλλέχτηκαν για τη διαλεκτική βάση Gree.D. (Karasimos *et al*., 2008). Η συλλογή των ηχογραφήσεων έγινε με συσκευές ψηφιακής ηχογράφησης υψηλής ευκρίνειας, σε όσον το δυνατόν πιο ήσυχες συνθήκες και πάντα με συναίνεση των συνομιλητών. Η επιλογή των ομιλητών έγινε με μεγάλη προσοχή, όσο αυτό ήταν εφικτό· στόχος ήταν οι ομιλητές να έχουν καθαρή άρθρωση, να έχουν φυσική ροή ομιλίας, 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 την 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). Βασική προϋπόθεση για τον ενδιάμεσο ήταν η καλή σχέση και γνωριμία με τους ομιλητές καθώς και η άριστη γνώση και χρήση της διαλέκτου.





Εικόνα 1 & 2: *Δείγμα εικόνων από τα ψηφιοποιημένα χειρόγραφα (αριστερά Επιτάφια επιγράμματα του Λεβίδη· δεξιά Χειρόγραφα Καζαντζίδη)*

Αντιστοίχως, το σώμα γραπτών κειμένων αποτελείται από ψηφιοποιημένα χειρόγραφα έγγραφα συνόλου 2.000.000 λεξικών τύπων. Το σημαντικότερο ζήτημα για τη συλλογή γραπτών δεδομένων είναι η έλλειψη πρωτογενών πηγών και κυρίως χειρογράφων για τα Αϊβαλιώτικα· αναπόφευκτα η ισορροπία ανάμεσα στην αντιπροσωπευτικότητα του δείγματος κειμένων που ψηφιοποιήθηκαν δεν ήταν εφικτή. Πέραν αυτής της εγγενούς δυσκολίας, τα κείμενα επιλέχτηκαν με βάση συγκεκριμένα κριτήρια. Βασικό κριτήριο ήταν το ζήτημα πνευματικής ιδιοκτησίας για την ψηφιοποίηση και για αυτό το λόγο

επιλέχθηκαν κείμενα πριν το 1938. Επίσης επιλέχθηκαν κυρίως κείμενα πεζού λόγου με ελάχιστη επιλογή ποιημάτων και τραγουδιών. Εκτός από μια αντιπροσωπευτική αντιπροσώπευση ανάμεσα στα δημοσιευμένα κείμενα και τα χειρόγραφα, σημαντικό βάρος δόθηκε στην σπανιότητα μερικών εξ αυτών (αναλυτικά για τα κριτήρια στο Κολιοπούλου, Μαρκόπουλος & Παντελίδης (υπό έκδοση)). Τα δεδομένα των παραπάνω σωμάτων πέρασαν από επεξεργασία, επιλογή, επισημείωση και ανάλυση και επεξεργάζονται σύμφωνα με το μοντέλο 3A (annotation, abstraction, analysis) των Wallis & Nelson (2001) και τον προτεινόμενο μορφότυπο των Gries & Berez (υπό έκδοση). Για την περαιτέρω επεξεργασία, επισημείωση, ανάλυση και περιγραφή μεταδεδομένων έγιναν δύο υπο-σώματα κειμένων με 60 ώρες και 200.000 λέξεις αντίστοιχα. Η συγκεκριμένη επεξεργασία και ανάλυση έγινε εκτός από τη συνδρομή δημοφιλών γλωσσολογικών εργαλείων, με επτά νέες εφαρμογές που δημιουργήθηκαν στο πλαίσιο του προγράμματος (βλ. ενότητα 2.2).

2.2. Οι εφαρμογές του συστήματος

Το σύστημα διαθέτει επτά (7) βασικές εφαρμογές για την υποστήριξη της ανάλυσης των συγκεκριμένων διαλεκτικών σωμάτων, ενώ παράλληλα γίνεται η χρήση δύο εξαιρετικά δημοφιλών γλωσσολογικών εργαλείων, όπως είναι το Praat και το ELAN. Είναι οι ακόλουθες (αναλυτικότερα βλ. Karanikolas, Galiotou & Ralli 2014):

(α) **Phon Tagger** για την οριοθέτηση των λέξεων και χρησιμοποιείται τόσο στο προφορικό όσο και στο γραπτό σώμα κειμένων, ώστε να υπάρχει μια ενιαία αντιμετώπιση της πληροφορίας των μορφολογικών ορίων των λέξεων μεταξύ των δύο σωμάτων.

(β) **Morph Tagger** για τον μορφολογικό σχολιασμό των λέξεων, όπου πραγματοποιείται στο επίπεδο λέξης. Για κάθε μορφολογική λέξη παρέχονται πληροφορίες σχετικά με το μέρος του λόγου, γραμματικές ιδιοτήτες και μορφολογικά φαινομένα, όπως η παραγωγή και σύνθεση.

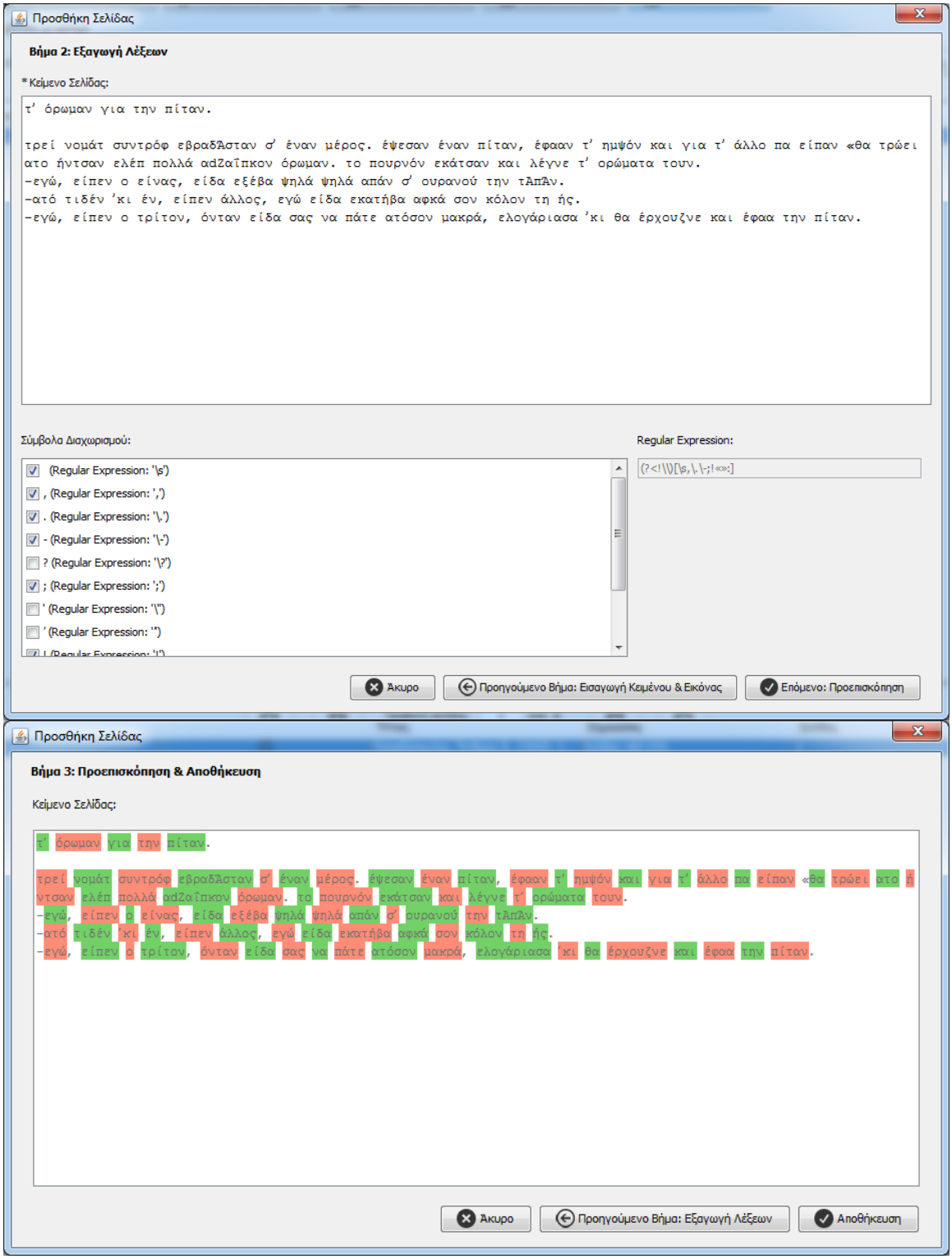
(γ) **Synt Tagger** για τη συντακτική ανάλυση και δομή φράσεων και προτάσεων· στην τρέχουσα κατάσταση του συστήματος, η επισημείωση γίνεται σε επίπεδο λέξης, όπου κάθε λέξη συνδέεται τουλάχιστον με μία συντακτική δομή. Η εφαρμογή παρέχει επίσης η δυνατότητα για επισημείωση σε μια φράση ή σε προτασιακό επίπεδο.

(δ) **Sem Tagger** για το σημασιολογικό σχολιασμό καταχωρώντας πληροφορίες όπως *δάνειο* (καθώς και την καταγωγή του), *ιδιωματική φράση,* κτλ.

(ε) **Text Imaging** για την προεπισκόπηση εικόνων από τα ψηφιοποιημένα κείμενα και χειρόγραφα,

(στ) **Text Transcription** για μεταγραφή των ψηφιοποιημένων κειμένων και των εικόνων και τέλος,

(ζ) **MOS** (Oral Metadata) για μια ολοκληρωμένη δομή μεταδεδομένων· αυτή η εφαρμογή παρέχει τη δυνατότητα διατήρησης και ενημέρωσης των μεταδεδομένων του σώματος προφορικών κειμένων και περιλαμβάνει πληροφορίες όπως *ηλικία*, *φύλο*, *πολιτισμικό υπόβαθρο* του ομιλητή μεταξύ άλλων (σημειώνεται ότι, οι πληροφορίες αυτές δεν είναι διαθέσιμες για τις γραπτές πηγές).

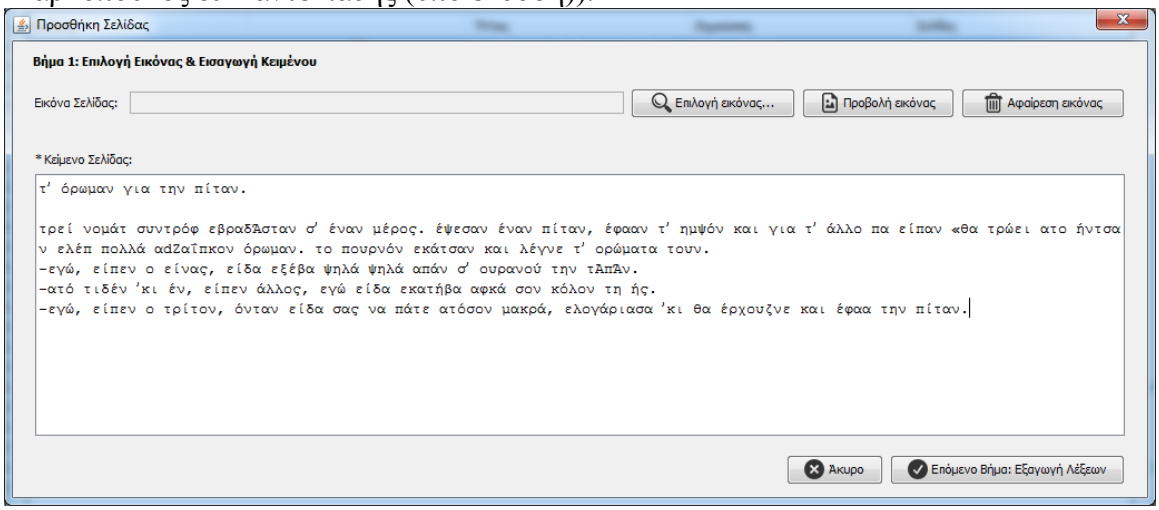


Εικόνα 3 & 4: *Δείγμα από την εφαρμογή - υποσύστημα οριοθέτησης μορφολογικών λέξεων.*

3. Προ-επεξεργασία σωμάτων γραπτών και προφορικών κειμένων

Η προ-επεξεργασία των δεδομένων μπορεί να συνοψιστεί ως εξής:

α) **Διαμόρφωση και Παραμετροποίηση**: Το κάθε πρωτογενές στερεοφωνικό ηχητικό αρχείο διαχωρίστηκε στα αντίστοιχα κανάλια του και έγινε επιλογή των κατάλληλων αρχείων με βάση συγκεκριμένων γλωσσολογικών και τεχνικών κριτηρίων (βλ. Karasimos *et al*. 2010). Επιπροσθέτως οι εικόνες πέρασαν από τεχνική επεξεργασία για απομόνωση των σελίδων, αποκοπή μαύρων πλαισίων και ρύθμιση της καθαρότητας τους.
β) **Επισημείωση**: Το σώμα γραπτών κειμένων πέρασε από μια συστηματική παραμετροποίηση φωνολογική και μορφολογική με βάση προεπιλεγμένες ετικέτες για ελεγχόμενες λίστες τιμών για την πλήρη κάλυψη των δύο επιπέδων.
Παράλληλα κωδικοποιηθήκε μια μικρή παραλλαγή του προτύπου SAMPA (Wells 1997) και ενοποιήθηκαν οι διαφορετικές ποικιλίες συμβόλων γραπτών κειμένων με βάση τη πρόταση των Μανωλέσσου, Μπέης & Μπασσέα (2012). Για τη επεξεργασία των προφορικών κειμένων έγινε μια αρχική προετοιμασία σύμφωνα με μια ανανεωμένη προσέγγιση παλαιότερης τακτικής επισημείωσης (Ράλλη, Παπαζαχαρίου & Καρασίμος, 2010). Συγκεκριμένα, από το σύνολο του ψηφιοποιημένου υλικού οι επιλεγμένες λέξεις μεταγράφηκαν «δια χειρός», χωρίς την βοήθεια αυτοματοποιημένου λογισμικού μεταγραφής, λόγω των δυσκολιών που ένα τέτοιο εγχείρημα ενδεχομένως να προκαλούσε, όπως είναι η δυσκολία αυτόματης αναγνώρισης πολυτονικού συστήματος, δυσκολία αυτόματης αναγνώρισης χαρακτήρων στο χειρόγραφο υλικό (Κολιοπούλου, Μαρκόπουλος & Παντελιάδης (υπό έκδοση)).



Εικόνα 5: *Δείγμα από την επισημείωση κειμένου ενός χειρόγραφου.*

γ) **Μεταδεδομένα**: Ακολουθήθηκε το πρωτόκολλο καταγραφής για τα προφορικά δεδομένα, όπου επιλέχθηκαν οι πληροφορίες που ταιριάζουν και για το σώματα γραπτών κειμένων με την παράλληλη εισαγωγή νέων ελεγχόμενων λιστών με τιμές για τα ψηφιοποιημένα κείμενα.

4. Επισημείωση

4.1. Επισημείωση σώματος γραπτών και προφορικών κειμένων

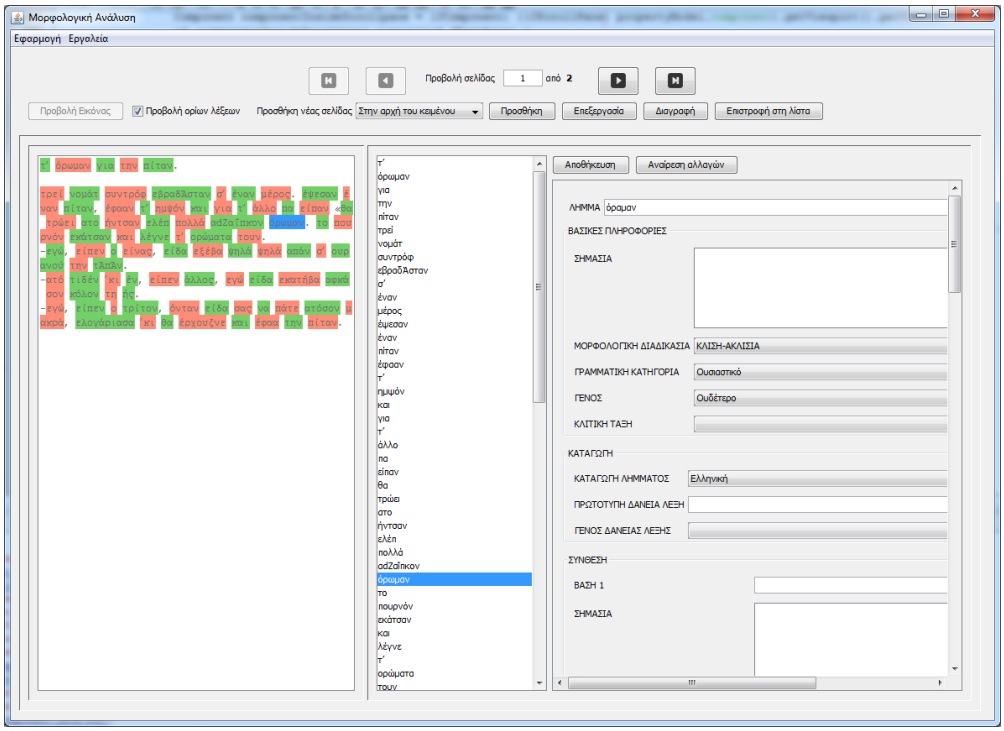
Για την επισημείωση των δύο σωμάτων ακολουθήθηκαν ίδιες στρατηγικές επισημείωσης, τουλάχιστον στα βασικά γλωσσικά επίπεδα. Η ουσιαστικότερη διαφοροποίηση, εντούτοις, εντοπίζεται στο φωνητικό—φωνολογικό επίπεδο, όπου είναι αναμενόμενα να υπάρχουν

διαφορετικά επίπεδα επισημείωσης που θα απουσιάζουν (αναλυτικότερα βλ. Κολιοπούλου, Μαρκόπουλος & Παντελίδης, υπό έκδοση).

4.1.1. Μορφολογικό επίπεδο

Και στα δύο σώματα οι κατηγορίες και υποκατηγορίες μορφολογικής ανάλυσης είναι ίδιες, όπου κυριαρχούν οι λίστες με τις προεπιλεγμένες τιμές στις περισσότερες περιπτώσεις. Οι κατηγορίες ανάλυσης περιέχουν πληροφορίες, όπως λήμμα, μορφολογική διαδικασία, γένος, κλιτική τάξη, γραμματική κατηγορία, καταγωγή, τύποι βάσεων/μορφημάτων/παραγωγικών προσφυμάτων/ κλιτικών προσφυμάτων (ανά γραμματική κατηγορία).

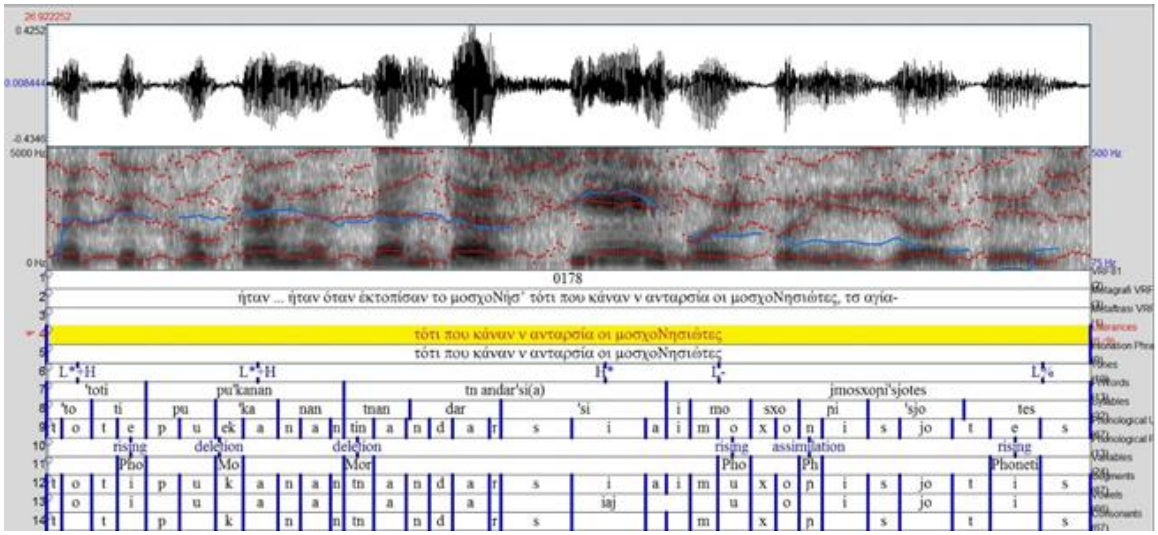
1. ΛΗΜΜΑ	2. ΜΟΡΦΟΛΟΓΙΚΗ ΔΙΑΔΙΚΑΣΙΑ	3. ΓΡΑΜΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΑ	4. ΓΕΝΟΣ	5. ΚΛΙΤΙΚΗ ΤΑΞΗ	6. ΚΑΤΑΓΩΓΗ ΛΗΜΜΑΤΟΣ
	Κλίση-Ακλισία	Επίθετο	Ουδέτερο	ΚΤ1-Ουσιαστικά	Τουρκική
	Παραγωγή-Κλίση	Ουσιαστικό	Αρσενικό-	ΚΤ2-Ουσιαστικά	Ελληνική
	Σύνθεση-Κλίση	Άρθρο	Αρσενικό	ΚΤ3-Ουσιαστικά	Ρομανική
	Σύνθεση-Παραγωγή-Κλίση	Ρήμα	Θηλυκό	ΚΤ4-Ουσιαστικά	Άλλη
	Παραγωγή-Σύνθεση-Κλίση	Επίρρημα	Χωρίς γένος	ΚΤ5-Ουσιαστικά	
		Αντωνυμία		ΚΤ6-Ουσιαστικά	
		Γερούνδιο		ΚΤ7-Ουσιαστικά	
		Απαρέμφατικός		ΚΤ8-Ουσιαστικά	
		Μετοχή		ΚΤ9-Ουσιαστικά	
		Επιφώνημα		ΚΤ10-Ουσιαστικά	
		Πρόθεση		ΚΤ1-Ρήματα	
		Σύνδεσμος		ΚΤ2Α-Ρήματα	
		Έκφραση		ΚΤ2Β-Ρήματα	
		Αριθμητικό		ΜΚΤ -Επίθετα	
				Άκλιτο	



Εικόνα 6 & 7: Δείγμα μορφολογικής ανάλυσης στο στάδιο προ-επεξεργασίας και στο στάδιο χρήσης του Morph Tagger

4.1.2. Φωνολογικό—φωνητικό επίπεδο

Η διαφοροποίηση μεταξύ των σωμάτων στο συγκεκριμένο επίπεδο είναι αναμενόμενη, Ενώ στο σώμα γραπτών κειμένων γίνεται εντοπισμός φαινομένων φωνηέντων και συμφώνων (ανάπτυξη, ανομοίωση, αποβολή, ανύψωση, αφομοίωση κτλ) με μονοεπίπεδο tier, στο σώμα προφορικών κειμένων οι πολυεπίπεδη χρήση tiers ανάλυσης συμπεριλαμβάνει ανάλυση έκφωνημάτων, φωνολογικών λέξεων, συλλαβών, φωνημάτων, επιτονισμού, συνεισφορών, κτλ. Γίνεται χρήση μιας τροποποιημένες έκδοσης του IPA για τη συνολική επισημείωση των ηχητικών αρχείων.



Εικόνα 8: *Πολυεπίπεδη φωνολογική επισημείωση διαλεκτικού υλικού στο Praat*

4.2. Προκλήσεις στην επισημείωση μεταξύ σωμάτων κειμένων

Η σημαντικότερη πρόκληση και τα σημαντικότερα ερευνητικά ζητήματα εντοπίζονται στην επισημείωση στα σώματα γραπτών κειμένων. Όπως επισημαίνουν οι Κολιοπούλου, Μαρκόπουλος & Παντελίδης (υπό έκδοση) δεν έγινε φωνητική/ φωνολογική μεταγραφή, γιατί:

α) τα ακριβή φωνολογικά χαρακτηριστικά των τριών διαλέκτων παραμένουν αμφίβολα, καθότι τα περισσότερα κείμενα είναι παλαιότερα των 75 ετών και η κωδικοποίηση των φαινομένων έγινε με τυχαίο, μη-επιστημονικό, αλλά συστηματικό τρόπο από τους συγγραφείς,

β) τα γραπτά κείμενα δεν ενδείκνυνται για φωνητική μεταγραφή, γιατί αρκετοί συμβολισμοί δεν μπορούν να αντιστοιχισθούν με σιγουριά στα αντίστοιχα σύμβολα του IPA και

(γ) η μη-επιστημικότητα των συντακτών, η αυθαιρεσία συμβόλων στη συγγραφή των κειμένων εμφανίζεται έντονα στο δείγμα· παράλληλα βαρύνουσας σημασίας είναι η απουσία συνοδευτικού ενδείκτη ή εισαγωγικού κειμένου που να εξηγούν τις όποιες αποφάσεις πήραν κατά τη συλλογή του υλικού ή την γραπτή απόδοση των προφορικών μαρτυριών.

Επομένως για να υπερκεραστούν τα προβλήματα (α) χρησιμοποιήθηκαν συμπεράσματα από την επισημείωση των προφορικών κειμένων για αμφίβολους χαρακτήρες, μιας και τα ηχητικά αρχεία ενδείκνυνται για τέτοια μεταγραφή, (β) έγινε επιβεβαίωση συμβόλων από άλλα κείμενα ίδιας περιόδου, όσο αυτό ήταν εφικτό και (γ) ακολουθήθηκε η χρήση

ελληνικού αλφαβήτου με την καθιερωμένη ιστορική ορθογραφία. Στο τελικό έλεγχο των επισημειώσεων τα δύο σώματα κειμένων θα λειτουργήσουν ως ελεγκτές ακρίβειας και συνέπειας για την επικαιροποίηση των προβληματικών επισημειώσεων. Ταυτόχρονα αποτελούν ένα αξιόπιστο δείγμα για επαλήθευση των πινάκων αντιστοιχίσης συμβόλων με το IPA.

5. Συμπεράσματα

Η συνέπεια στην επισημείωση σωμάτων κειμένων παραμένει σοβαρό ζήτημα για τη διακειμενική γλωσσολογία. Σημαντικά ζητήματα για την επισημείωση σε φωνολογικό επίπεδο αντιμετωπίστηκαν κατά τη μελέτη καθότι έγινε μια συστηματική προσπάθεια να ενοποιηθούν όλες οι διαφορετικές μεταγραφές διαλεκτικού γραπτού υλικού μιας και δεν υπήρχε προηγουμένως κοινή στρατηγική απεικόνισης. Παράλληλα προτείνεται πολυεπίπεδη φωνολογική (παράλληλα με μορφολογική) επισημείωση του σώματος κειμένων καθιερώνοντας ένα βασικό πρότυπο επισημείωσης διαλεκτικού υλικού για τις Νεοελληνικές Διαλέκτους σε καθιερωμένα λογισμικά ανάλυσης ομιλίας, ενώ γίνεται η χρήση των επισημειώσεων για δια-σωματική επικαιροποίηση της συνέπειας και της ακρίβειας της συνολικής επισημείωσης.

Βιβλιογραφία

Dickinson, M & W. Detmar-Meurers (2003) Detecting errors in part-of-speech annotation. In *Proceedings of EACL-03*, pp. 107–114, Budapest, Hungary.

Dickinson, M. (2005) *Error detection and correction in annotated corpora*. Ph.D. thesis, The Ohio State University.

Eskin, E. (2000) Automatic corpus correction with anomaly detection. In *Proceedings of NAACL-00*, pp. 148–153, Seattle, Washington.

Gries, S. Th. & A. L. Berez (to appear). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*. Berlin & New York: Springer.

Karanikolas, N. Galiotou, E. & A. Ralli (2014). Towards a Unified Exploitation of Electronic Dialectal Corpora: Problems and Perspectives. In P. Sojka *et al*. (eds.) *TSD 2014*, LNAI 8655, pp. 257–266. Switzerland: Springer.

Karasimos A., Melissaropoulou D., Ralli A., Papazachariou D. & D. Asimakopoulos (2008) GREED: Cataloguing and Encoding Modern Greek Dialectal Spoken Corpora. Presented in *CatCod 2008*, 4-5 December, Orleans, France.

Ralli A., Papazachariou D. & A. Karasimos (2009) Laboratory of Modern Greek Dialects and the GREED project. In A. Ralli *et al.* (eds.) *Proceeding of 4th International Conference of Modern Greek Dialects and Linguistic Theory*.

Ule, T. & K. Simov (2004) Unexpected Productions May Well be Errors. In *Proceeding of LREC 2004*, pp. 1795-1798.

van Halteren H. (2000). The detection of inconsistency in manually tagged text. In A. Abeill´e, T. Brants & H. Uszkoreit (eds), *Proceedings of LINC-00*, Luxembourg.

Wallis, S.A. & G. Nelson (2001) Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, 15, pp. 307-340.

Κολιοπούλου, Μ., Μαρκόπουλος, Θ. & Ν. Παντελίδης (υπό έκδοση) Πόντος, Καππαδοκία, Αιβαλί: προκλήσεις ενός ψηφιακού σώματος γραπτού υλικού. Στα *Proceedings of ICGL11* (Ρόδος, 26-29/09/2013).
Μανωλέσσου Ι., Μπένης Σ. & Χ. Μπασσέα (2012). Η φωνητική μεταγραφή των Νεοελληνικών Διαλέκτων. *Λεξικογραφικόν Δελτίον 26,* σσ. 161-222.