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

THALIS AMiGre Project

Αθανάσιος Καρασίμος (*Ακαδημία Αθηνών & Πανεπιστήμιο Πατρών*), Ελένη Γαλιώτου (*ΤΕΙ Αθηνών*), Νικήτας Καρανικόλας (*ΤΕΙ Αθηνών*), Γιώργος Κορωνάκης (*ΤΕΙ Αθηνών*), Κώστας Αθανασάκος (*ΤΕΙ Αθηνών*), Δημήτρης Παπαζαχαρίου (*Πανεπιστήμιο Πατρών*), Αγγελική Ράλλη (Πανεπιστήμιο Πατρών)

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.

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 (Dickinson και Meurers 2005, 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 αξιοποίησή 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

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). Αντιστοίχως, το σώμα γραπτών κειμένων αποτελείται από ψηφιοποιημένα χειρόγραφα έγγραφα συνόλου 2.000.000 λεξικών τύπων. 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 45 ώρες και 200.000 λέξεις αντίστοιχα.

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. Οι εφαρμογές του συστήματος είναι οι ακόλουθες:

(α) **Phon Tagger** για την οριοθέτηση των λέξεων στα σώματα, (β) **Morph Tagger** για τον μορφολογικό σχολιασμό των λέξεων (POS, διαδικασία σχηματισμού, μορφολογικά φαινόμενα κτλ), (γ) **Synt Tagger** για τη συντακτική ανάλυση και δομή φράσεων και προτάσεων, (δ) **Sem Tagger** για το σημασιολογικό σχολιασμό, (ε) **Text Imaging** για την προεπισκόπηση εικόνων, (στ) **Text Transcription** για μεταγραφή των κειμένων των εικόνων και (ζ) **MOS** (Oral Metadata) για μια ολοκληρωμένη δομή μεταδεδομένων.

3. Προ-επεξεργασία σωμάτων γραπτών και προφορικών κειμένων

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

> **Διαμόρφωση και Παραμετροποίηση**: Τα ηχητικά αρχεία διαχωρίστηκαν στα αντίστοιχα κανάλια του και έγινε επιλογή των κατάλληλων αρχείων με βάση συγκεκριμένων γλωσσολογικών και τεχνικών κριτηρίων (βλ. Karasimos et al. 2001). Επιπροσθέτως οι εικόνες πέρασαν από τεχνική επεξεργασία για απομόνωση των σελίδων, αποκοπή μαύρων πλαισίων και ρύθμιση της καθαρότητας τους.

> **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).

> **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.

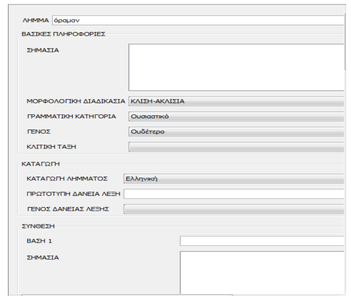
4. Επισημείωση

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 (2014).

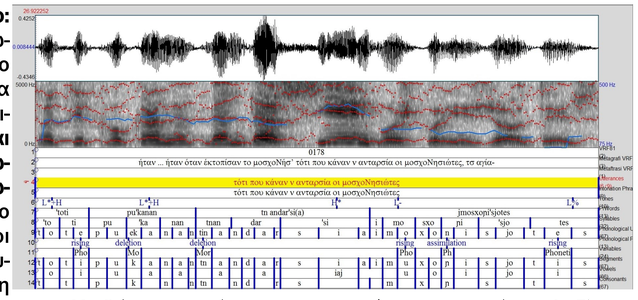
Μορφολογικό επίπεδο: Και στα δύο σώματα οι κατηγορίες και υποκατηγορίες μορφολογικής ανάλυσης είναι ίδιες, όπου κυριαρχούν οι λίστες με τις προεπιλεγμένες τιμές στις περισσότερες περιπτώσεις. Οι κατηγορίες ανάλυσης περιέχουν πληροφορίες, όπως **λήμμα, μορφολογική διαδικασία, γένος, κλιτική τάξη, γραμματική 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	



Δείγμα μορφολογικής ανάλυσης στο στάδιο προ-επεξεργασίας και στο στάδιο χρήσης του Morph Tagger

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.



4.2. Προκλήσεις στην επισημείωση μεταξύ σωμάτων κειμένων

Η σημαντικότερη πρόκληση στην επισημείωση εντοπίζεται στα σώματα γραπτών κειμένων. Όπως επισημαίνουν οι Κολιοπούλου, Μαρκόπουλος & Παντελίδης (2014) δεν έγινε φωνητική/ φωνολογική μεταγραφή, γιατί: α) τα ακριβή φωνολογικά χαρακτηριστικά των τριών διαλέκτων παραμένουν αμφίβολα, β) τα γραπτά κείμενα δεν ενδείκνυνται για φωνητική μεταγραφή και (γ) η μη-επιστημικότητα των συντακτών και η αυθαιρεσία συμβόλων στη συγγραφή των κειμένων εμφανίζεται έντονα στο δείγμα. Επομένως για να υπερκεραστούν τα προβλήματα (α) χρησιμοποιήθηκαν συμπεράσματα από την επισημείωση των προφορικών κειμένων για αμφίβολους χαρακτήρες (καθότι ενδείκνυνται για φωνητική μεταγραφή), (β) έγινε επιβεβαίωση συμβόλων από άλλα κείμενα ίδιας περιόδου 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.

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.