Κεφάλαιο 5

Αναγνώριση Είδους Κειμένου

5.1 Εισαγωγή

Στην βιβλιογραφία δεν υπάρχει ένας κοινά αποδεκτός ορισμός της έννοιας του είδους κειμένου. Πολλοί ερευνητές, μάλιστα, χρησιμοποιούν διαφορετικούς όρους για να το εκφράσουν. Έτσι, έχουν προταθεί κατά καιρούς οι όροι: είδος κειμένου (text genre) [61], τύπος κειμένου (text type) [8], ιδίωμα (register) [36], υπογλώσσα (sublanguage) [57], λειτουργικό ύφος (functional style) [76, 97] κ.ά.

Σε αυτήν την εργασία το είδος ενός κειμένου καθορίζεται από το λειτουργικό του ρόλο. Κάθε είδος κειμένου κωδικοποιεί μία επικοινωνιακή περίσταση καθώς περιλαμβάνει κείμενα που έχουν γραφτεί κάτω από τις ίδιες επικοινωνιακές συνθήκες, απευθύνονται στο ίδιο κοινό, και πληρούν την ίδια κοινωνική αποστολή. Παράγοντες όπως ο συγγραφέας ή το θεματικό περιεχόμενο του κειμένου δεν καθορίζουν το είδος του. Τυπικά παραδείγματα ειδών κειμένου αποτελούν τα δημόσια έγγραφα, τα επιστημονικά άρθρα, οι συνεντεύξεις κ.ά.

Η αυτόματη αναγνώριση του είδους κειμένου έχει αποκτήσει μεγάλη σημασία τα τελευταία χρόνια. Η ραγδαία ανάπτυξη βάσεων δεδομένων με κείμενα σε ηλεκτρονική μορφή, και κυρίως η ανάπτυξη του World-Wide Web, δημιούργησε για πρώτη φορά το πρόβλημα ετερογενών πηγών, δηλ. πηγών που περιέχουν περισσότερα από ένα είδη κειμένων. Έτσι, η αναζήτηση ενός κειμένου που να ταιριάζει με τις απαιτήσεις του χρήστη (εφαρμογές ανάκτησης πληροφορίας), σε ένα ετερογενές περιβάλλον, συνδέεται άμεσα με την αναγνώριση του είδους κειμένου που επιθυμεί ο χρήστης. Για να γίνει αυτό πιο αντιληπτό φανταστείτε ότι όταν κάποιος πηγαίνει σε μία βιβλιοθήκη δεν ψάχνει απλά κάτι που να έχει σχέση με ένα θεματικό αντικείμενο. Αντίθετα, έχει συγκεκριμένες απαιτήσεις, όπως επιστημονικά άρθρα για τον ανθρώπινο εγκέφαλο, ρεπορτάζ εφημερίδων για την εξωτερική πολιτική, συνταγές κινέζικης κουζίνας κτλ.

Η αναγνώριση του είδους κειμένου μπορεί να οδηγήσει και στη βελτίωση της απόδοσης διαφόρων συστημάτων επεξεργασίας φυσικής γλώσσας. Πολλές έννοιες λέξεων συνδέονται μόνο με συγκεκριμένα είδη κειμένων (π.χ. η λέξη *μπρίκι* όταν συναντάται σε συνταγές συνδέεται πάντα με την έννοια του μαγειρικού σκεύους και όχι του πλοίου). Μία λέξη που αντιστοιχεί σε περισσότερα από ένα μέρη-του-λόγου, είναι πιθανό να εμφανίζεται συχνότερα με τη μορφή ενός συγκεκριμένου μέρους-του-λόγου σε κάποιο δεδομένο είδος κειμένου (π.χ. οι λέξεις *ήπια, ήπιες* είναι πιο πιθανό να εμφανίζονται με τη μορφή επιθέτου, παρά ρήματος, σε επιστημονικά κείμενα). Επίσης, διάφορες συντακτικές δομές είναι στενά συνδεδεμένες με συγκεκριμένα είδη κειμένων και συναντιούνται σπάνια σε άλλα είδη (π.χ. στα επιστημονικά κείμενα χρησιμοποιείται συνήθως παθητική σύνταξη ενώ στις συνταγές όχι).

Παρ' όλα αυτά, ο αριθμός των μελετών αυτόματης υπολογιστικής αναγνώρισης του είδους κειμένου δεν είναι ανάλογος άλλων περιοχών της επεξεργασίας κειμένου. Η πιο πλήρης έρευνα πάνω στο θέμα διεξήχθη από τον Biber, ο οποίος μελέτησε τις διαφορές μεταξύ γραπτού και προφορικού λόγου [7] και τις διαφορές μεταξύ των ειδών κειμένων για διάφορες γλώσσες [11]. Πιο συγκεκριμένα, για την Αγγλική γλώσσα ο Biber προτείνει ένα σύνολο από 67 δείκτες ύφους, το οποίο περιλαμβάνει λεξιλογικές και συντακτικές παραμέτρους (μερικές αρκετά περίπλοκες) και με βάση αυτές αναλύει ένα μεγάλο σώμα αποτελούμενο από 23 είδη κειμένων. Στην συνέχεια, κατατάσσει τα είδη κειμένων ως προς εφτά διαστάσεις, που προκύπτουν από την

εφαρμογή της στατιστικής μεθόδου της παραγοντικής ανάλυσης (factor analysis), και προσπαθεί να ερμηνεύσει αυτές τις διαστάσεις χρησιμοποιώντας όρους όπως «πληροφοριακό περιεχόμενο ↔ πλοκή», «αφήγηση ↔ μη-αφήγηση» κ.ά. Στην ουσία, η εργασία αυτή δεν στοχεύει στην δημιουργία ενός αυτοματοποιημένου συστήματος αλλά προσπαθεί να ερμηνεύσει γλωσσολογικά τις ομοιότητες και τις διαφορές μεταξύ των ειδών κειμένων καθώς και να εξηγήσει τη λειτουργία των γλωσσολογικών χαρακτηριστικών ενός κειμένου.

Οι Karlgren και Cutting [49] προτείνουν ένα πολύ πιο απλό σύνολο υφολογικών δεικτών, αποτελούμενο από 20 παραμέτρους και υποσύνολο του Biber, με γνώμονα την εύκολη μέτρηση αυτών των παραμέτρων σε οποιοδήποτε κείμενο. Στην συνέχεια εφαρμόζουν διαχωριστική ανάλυση (discriminant analysis) στα διανύσματα που προκύπτουν από την ανάλυση των κειμένων του σώματος *Brown,* επιτυγχάνοντας 65% ακρίβεια στην αναγνώριση 15 ειδών κειμένου. Να σημειωθεί ότι τα αποτελέσματα αυτά αναφέρονται στο σώμα εκπαίδευσης. Μία κάπως διαφορετική προσέγγιση προτείνεται από τον Kessler [54], ο οποίος ταξινομεί τα κείμενα με βάση ένα σύνολο γενικών ιδιοτήτων (generic facets), τις οποίες ονομάζει: *Brow* (έχει να κάνει με το πνευματικό επίπεδο των αναγνωστών του κειμένου), *Narrative* (η οποία υποδεικνύει το βαθμό αφήγησης του κειμένου) και *Genre* (που ταιριάζει με το είδος κειμένου όπως το έχουμε ορίσει). Επίσης, χρησιμοποιεί ακόμα πιο απλουστευμένο σύνολο υφολογικών δεικτών για να αποφύγει το συντακτικό σχολιασμό του κειμένου. Το σύστημα αυτό επιτυγχάνει στην καλύτερη περίπτωση ακρίβεια ταξινόμησης 79% ως προς την ιδιότητα *Genre* για ένα υποσύνολο του σώματος *Brown* (συνολικά 6 είδη κειμένων).

Επίσης, μία εμπειρική περιγραφή τριών επιπέδων για το λειτουργικό ύφος της Νεοελληνικής γλώσσας προτείνεται από το Μίχο [63]. Το κατώτερο επίπεδο περιλαμβάνει συνολικά 16 γλωσσολογικά χαρακτηριστικά που ανιχνεύονται στο κείμενο (λεξιλογικά και συντακτικά), τα οποία ομαδοποιούνται στο πιο πάνω επίπεδο, που αποτελείται από τα γνωρίσματα του ύφους (επισημότητα, γλαφυρότητα κ.ά.). Τα γνωρίσματα του ύφους, τέλος, ομαδοποιούνται στο κορυφαίο επίπεδο που αποτελείται από γενικές κατηγορίες του λειτουργικού ύφους, όπως επιστημονικό ύφος, δημοσιογραφικό ύφος, λογοτεχνικό ύφος, κ.ά.

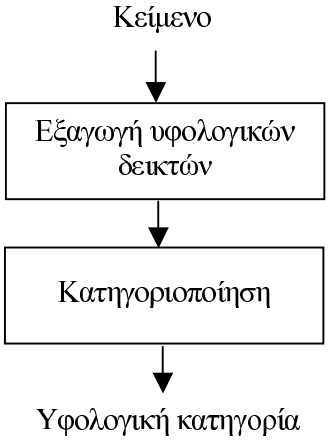
Σε αυτό το κεφάλαιο παρουσιάζουμε την εφαρμογή των υφολογικών δεικτών που παρουσιάστηκαν στο προηγούμενο κεφάλαιο για την αυτόματη αναγνώριση είδους κειμένου. Στο επόμενο τμήμα περιγράφονται οι μέθοδοι κατηγοριοποίησης που επιλέξαμε. Στο τμήμα 5.3 περιγράφονται οι λεξιλογικές προσεγγίσεις με τις οποίες συγκρίναμε την απόδοση της μεθόδου μας και στο τμήμα 5.4 περιγράφεται το σώμα κειμένων που χρησιμοποιήθηκε στα πειράματα που παρουσιάζονται στο τμήμα 5.5. Τέλος, στο τμήμα 5.6 δίνεται μια περίληψη του κεφαλαίου και συνοψίζονται τα βασικότερα συμπεράσματα που προέκυψαν από αυτά τα πειράματα.

5.2 Μέθοδοι Κατηγοριοποίησης

Μετά την εξαγωγή των υφολογικών δεικτών από το κείμενο, είναι απαραίτητο να εφαρμόσουμε μία διαδικασία κατηγοριοποίησης που θα κατατάξει το κείμενο σε μία υφολογική κατηγορία (στην περίπτωσή αυτού του κεφαλαίου, ένα είδος κειμένου), όπως φαίνεται στο σχήμα 5.1. Για την ταξινόμηση ενός διανύσματος παραμέτρων σε ένα μέλος ενός συνόλου προκαθορισμένων κατηγοριών έχουν προταθεί διάφορες μέθοδοι. Ενδεικτικά μπορούμε να αναφέρουμε τεχνικές της πολυπαραγοντικής στατιστικής, όπως την πολλαπλή παλινδρόμηση (multiple regression) και τη διαχωριστική ανάλυση (discriminant analysis), τα νευρωνικά δίκτυα (neural networks) και τεχνικές της μηχανικής εκμάθησης, όπως τα δέντρα αποφάσεων (decision trees) και τους ταξινομητές Bayes (Bayesian classifiers). Όλες αυτές οι τεχνικές, αφού εκπαιδευτούν, με βάση ένα σύνολο δεδομένων εκπαίδευσης από κάθε κατηγορία, μπορούν να ταξινομήσουν ένα νέο διάνυσμα σε μία από αυτές τις κατηγορίες.

Πρόσφατα, ο Yang [101] μελέτησε την απόδοση αρκετών τεχνικών ταξινόμησης κατά την εφαρμογή τους σε κατηγοριοποίηση κειμένων και κατέληξε στο συμπέρασμα ότι στατιστικές τεχνικές πολυπαραγοντικής ανάλυσης, όπως η πολλαπλή παλινδρόμηση και η μέθοδος των k-πιο κοντινών γειτόνων (k-nearest neighbours), καθώς και τα νευρωνικά δίκτυα παρουσιάζουν τα καλύτερα αποτελέσματα. Από την άλλη, με βάση την απόδοση αυτών των τεχνικών συναρτήσει του αριθμού των δεδομένων που χρησιμοποιούνται για εκπαίδευση, φάνηκε ότι η ακρίβεια κατηγοριοποίησης συγκλίνει, όταν χρησιμοποιούνται πάνω από 300 κείμενα από κάθε κατηγορία για εκπαίδευση. Όμως, σε περιπτώσεις όπου χρησιμοποιούνται λίγα

κείμενα για εκπαίδευση (γύρω στα 10 από κάθε κατηγορία), οι στατιστικές τεχνικές πολυπαραγοντικής ανάλυσης υπερτερούν αισθητά έναντι των νευρωνικών δικτύων και των ταξινομητών Bayes [102].



Σχήμα 5.1. Διαδικασία κατηγοριοποίησης.

Λαμβάνοντας υπ' όψιν αυτά τα συμπεράσματα καταλήξαμε στην χρησιμοποίηση δύο τεχνικών της πολυπαραγοντικής στατιστικής: της πολλαπλής παλινδρόμησης και της διαχωριστικής ανάλυσης που θα περιγραφούν στα επόμενα τμήματα. Εκτός των πλεονεκτημάτων που αναφέρθηκαν πιο πριν, αυτές οι τεχνικές απαιτούν ελάχιστο χρονικό κόστος τόσο για την εκπαίδευσή τους όσο και για την απόκρισή τους. Επομένως, είναι δυνατή η ενσωμάτωσή τους σε ένα αυτόματο σύστημα αναγνώρισης είδους κειμένου που θα απαιτούσε απόκριση σε πραγματικό χρόνο (real-time response system). Επιπλέον, η χρήση δύο τεχνικών δίνει τη δυνατότητα συγκριτικών αποτελεσμάτων όσον αφορά την ακρίβεια ταξινόμησης αλλά και της εξαγωγής πιο γενικών συμπερασμάτων.

5.2.1 Πολλαπλή παλινδρόμηση

Η ανάλυση παλινδρόμησης είναι μία στατιστική μεθοδολογία που χρησιμοποιείται για την πρόβλεψη των τιμών μίας ή περισσότερων εξαρτημένων μεταβλητών από τις τιμές ενός συνόλου ανεξάρτητων μεταβλητών [32, 48]. Στην περίπτωση αναγνώρισης είδους κειμένου, οι εξαρτημένες μεταβλητές είναι οι υφολογικές κατηγορίες, δηλ. τα είδη κειμένων, και οι ανεξάρτητες μεταβλητές είναι οι υφολογικοί δείκτες. Η ανεξάρτητη μεταβλητή εκφράζεται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών όπως φαίνεται πιο κάτω:

$$y_i=b_{0i} + z_1b_{1i} + z_2b_{2i} +...+ z_rb_{ri} + e_i$$

όπου το y_i είναι η τιμή απόκρισης για την i εξαρτημένη μεταβλητή, $z_1, z_2, ..., z_r$ είναι οι ανεξάρτητες μεταβλητές, $b_{0i}, b_{1i}, b_{2i}, ..., b_{ri}$ είναι οι συντελεστές παλινδρόμησης και e_i το σφάλμα για την i ανεξάρτητη μεταβλητή. Κατά τη διάρκεια της εκπαίδευσης, υπολογίζονται οι συντελεστές παλινδρόμησης για κάθε κατηγορία με βάση την *εκτίμηση ελαχίστων τετραγώνων* (least-squares estimation). Πιο συγκεκριμένα, αυτή η μέθοδος εκτίμησης επιλέγει τους συντελεστές παλινδρόμησης έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγωνικών αποστάσεων της πραγματικής τιμής από την εκτιμούμενη τιμή. Τελικά, η συνάρτηση υπολογισμού των συντελεστών αυτών δίνεται από τον πιο κάτω τύπο:

$$\hat{\beta}_i = (Z'Z)^{-1}Z'Y_i$$

όπου $\hat{\beta}_i$ είναι ο πίνακας-στήλη των συντελεστών παλινδρόμησης για την i εξαρτημένη μεταβλητή, Z είναι ο πίνακας $(n \times (r+1))$ των r ανεξάρτητων μεταβλητών για n δεδομένα εκπαίδευσης και Y_i είναι ο πίνακας-στήλη των πραγματικών τιμών της i εξαρτημένης μεταβλητής.

Η ποιότητα του εξαγόμενου μοντέλου, το κατά πόσο δηλαδή καταφέρνει να ταιριάξει με τα δεδομένα εκπαίδευσης, μπορεί να μετρηθεί μέσω του συντελεστή προσδιορισμού (coefficient of determination) R^2 που ορίζεται πιο κάτω:

$$R^2 = \frac{\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

όπου n είναι το πλήθος των δεδομένων εκπαίδευσης, $\bar{y}$ είναι η μέση τιμή απόκρισης, ενώ $\hat{y}_j$ είναι η εκτιμούμενη τιμή και y_j είναι η πραγματική τιμή της απόκρισης. Ο συντελεστής προσδιορισμού δείχνει το ποσοστό της μεταβολής της y_j που «εξηγήθηκε» (ή προσδιορίστηκε) από τις ανεξάρτητες μεταβλητές $z_1, z_2, ..., z_r$. Το R^2 (και ο *συντελεστής πολλαπλής συσχέτισης* (multiple correlation coefficient) $R = +\sqrt{R^2}$) ισούται με 1 εφόσον η γραμμική εξίσωση περνάει από όλα τα δεδομένα εκπαίδευσης.

Στην άλλη ακραία περίπτωση, όταν οι ανεξάρτητες μεταβλητές δεν έχουν καμία επίδραση στην τιμή απόκρισης, ισούται με μηδέν.

Η πολλαπλή παλινδρόμηση μπορεί επίσης να χρησιμοποιηθεί και για την εκτίμηση της επίδρασης των ανεξάρτητων μεταβλητών στην τιμή της απόκρισης. Με άλλα λόγια, είναι δυνατόν να μετρηθεί η σημαντικότητα της κάθε ανεξάρτητης μεταβλητής. Πιο συγκεκριμένα, το ποσοστό κατά το οποίο μειώνεται ο συντελεστής προσδιορισμού όταν αφαιρεθεί μία συγκεκριμένη ανεξάρτητη μεταβλητή από την εξίσωση της παλινδρόμησης μπορεί να μετρηθεί μέσω του τετραγώνου της *ημι-μερικής συσχέτισης* (semipartial correlation) που ορίζεται ως ακολούθως [92]:

$$sr_i^2 = \frac{t_i^2}{df_{res}}(1 - R^2)$$

όπου sr_i είναι η ημιμερική συσχέτιση και t_i η τιμή του στατιστικού-*t* για την *i* ανεξάρτητη μεταβλητή ενώ df_{res} είναι οι βαθμοί ελευθερίας των διαφορών των εκτιμούμενων τιμών από τις πραγματικές τιμές (residuals degrees of freedom). Εφόσον τα υπόλοιπα συστατικά της εξίσωσης είναι σταθερά, μπορούμε να πούμε ότι η σημαντικότητα μιας ανεξάρτητης μεταβλητής είναι συνάρτηση της απόλυτης τιμής του στατιστικού-*t*. Όσο μεγαλύτερη η απόλυτη τιμή του στατιστικού-*t* για μία ανεξάρτητη μεταβλητή τόσο πιο σημαντική η συμβολή της στην τιμή απόκρισης. Η τιμή του στατιστικού-*t* του *j* συντελεστή παλινδρόμησης b_j καθορίζεται ως εξής:

$$t_{b_j} = \frac{b_j}{S_{b_j}}$$

όπου S_{b_j} είναι το τυπικό σφάλμα (standard error) του *j* συντελεστή παλινδρόμησης.

Η ταξινόμηση ενός νέου διανύσματος σε μία κατηγορία γίνεται ως εξής: Αρχικά, υπολογίζεται η τιμή απόκρισης της κάθε ανεξάρτητης μεταβλητής. Στην συνέχεια επιλέγεται η ανεξάρτητη μεταβλητή (δηλ. η υφολογική κατηγορία) με την μεγαλύτερη τιμή απόκρισης.

Για τον εφαρμογή της ανάλυσης παλινδρόμησης στα πειράματα που περιγράφονται στην συνέχεια χρησιμοποιήθηκε το υπολογιστικό εργαλείο *MACANOVA*[1].

5.2.2 Διαχωριστική ανάλυση

Ο στόχος της διαχωριστικής ανάλυσης είναι η ταξινόμηση ενός διανύσματος παραμέτρων σε μία κατηγορία από ένα σύνολο προκαθορισμένων κατηγοριών [33]. Από μαθηματική άποψη, η διαχωριστική ανάλυση προσπαθεί μέσω ενός γραμμικού συνδυασμού των παραμέτρων να επιτύχει ελάχιστη διασπορά μεταξύ των δεδομένων της ίδιας κατηγορίας και μέγιστη διασπορά μεταξύ των διαφόρων κατηγοριών.

Η μέθοδος αυτή εξάγει ένα σύνολο γραμμικών διαχωριστικών συναρτήσεων (discriminant functions) από τα δεδομένα εκπαίδευσης. Ο αριθμός των συναρτήσεων που απαιτούνται για το διαχωρισμό των κατηγοριών είναι ανάλογος του αριθμού των κατηγοριών. Γενικά για n κατηγορίες απαιτούνται n-1 συναρτήσεις: η πρώτη συνάρτηση διαχωρίζει μία κατηγορία από τις υπόλοιπες n-1, η δεύτερη συνάρτηση διαχωρίζει μία άλλη κατηγορία από τις υπόλοιπες n-2, κ.ο.κ.

Για την ταξινόμηση ενός διανύσματος, εκτός των δεδομένων εκπαίδευσης, σε μία από τις κατηγορίες έχουν προταθεί διάφορες μέθοδοι. Ο πιο απλός τρόπος είναι η χρήση των συναρτήσεων ταξινόμησης (classification functions). Η κάθε συνάρτηση ταξινόμησης αντιστοιχεί σε μία κατηγορία και αποτελείται από ένα γραμμικό συνδυασμό των παραμέτρων, όπως και στην περίπτωση της πολλαπλής παλινδρόμησης:

$$y_i = c_{0i} + z_1 c_{1i} + z_2 c_{2i} + \ldots + z_r c_{ri}$$

όπου y_i είναι η τιμή ταξινόμησης για την i κατηγορία, $z_1, z_2, \ldots, z_r$ είναι οι τιμές των παραμέτρων διαχωρισμού (δηλ. των υφολογικών δεικτών), c_{0i} είναι μία σταθερά και $c_{1i}, c_{2i}, \ldots, c_{ri}$ είναι οι συντελεστές ταξινόμησης για την i κατηγορία. Οι συντελεστές ταξινόμησης υπολογίζονται ως εξής:

$$C_i = W^{-1} * M_i$$

[1] http://www.stat.umn.edu/~gary/macanova/macanova.home.html

όπου C_i είναι ο πίνακας των συντελεστών ταξινόμησης της *i* κατηγορίας, W είναι πίνακας διασποράς μεταξύ των κατηγοριών και M_i είναι ο πίνακας των μέσων τιμών των παραμέτρων *z* για την *i* κατηγορία. Τέλος, η σταθερά c_{0i} υπολογίζεται ως εξής:

$$c_{0i} = (-1/2)\, C_i * M_i$$

Ως πιο πιθανή κατηγορία για ένα διάνυσμα παραμέτρων υποδεικνύεται η κατηγορία της οποίας η συνάρτηση ταξινόμησης έδωσε την πιο υψηλή τιμή απόκρισης.

Ωστόσο, για την επίτευξη όσο το δυνατόν καλύτερων αποτελεσμάτων αποφασίσαμε να χρησιμοποιήσουμε μία ελαφρώς πιο περίπλοκη μέθοδο ταξινόμησης που βασίζεται στην απόσταση *Mahalanobis*. Η απόσταση Mahalanobis είναι μία μέτρηση της απόστασης μεταξύ δύο σημείων σε ένα χώρο που ορίζεται από πολλαπλές συσχετιζόμενες μεταβλητές. Το βασικότερο σημείο στο οποίο υπερτερεί της Ευκλείδειας απόστασης είναι ότι λαμβάνει υπ' όψιν της τις συσχετίσεις μεταξύ των μεταβλητών. Έτσι, για κάθε κατηγορία αρχικά εντοπίζεται ο *κεντροειδής* της, δηλ. το σημείο που αναπαριστά το μέσο όρο όλων των μεταβλητών στον πολυδιάστατο χώρο που καθορίζεται από τις ανεξάρτητες μεταβλητές. Στην συνέχεια για κάθε διάνυσμα υπολογίζεται η απόσταση Mahalanobis από τους κεντροειδείς των κατηγοριών και το διάνυσμα ταξινομείται στην κατηγορία με την μικρότερη απόσταση. Η απόσταση Mahalanobis *r* ενός διανύσματος $\boldsymbol{x}$ από ένα μέσο διάνυσμα $\boldsymbol{m}_x$ υπολογίζεται ως εξής:

$$r^2 = (\boldsymbol{x} - \boldsymbol{m}_x)' \, C_x^{-1} \, (\boldsymbol{x} - \boldsymbol{m}_x)$$

όπου C_x είναι ο πίνακας συνδιασποράς (covariance matrix) του $\boldsymbol{x}$. Η χρήση αυτής της μεθόδου ταξινόμησης προσφέρει επίσης την δυνατότητα υπολογισμού της πιθανότητας να ανήκει ένα διάνυσμα σε μία συγκεκριμένη κατηγορία. Προσεγγιστικά, η πιθανότητα αυτή, που καλείται μεταγενέστερη πιθανότητα (posterior probabilities), είναι ανάλογη της απόστασης Mahalanobis από τον κεντροειδή της συγκεκριμένης κατηγορίας.

Για την εφαρμογή της διαχωριστικής ανάλυσης στα πειράματα που περιγράφονται στην συνέχεια χρησιμοποιήθηκε το υπολογιστικό εργαλείο *XLSTAT*[1].

[1] http://www.xlstat.com

5.3 Λεξιλογικές Υφομετρικές Προσεγγίσεις

Για την καλύτερη αξιολόγηση της συνεισφοράς της προτεινόμενης μεθόδου αποφασίσαμε να υλοποιήσουμε τις δύο πιο σύγχρονες και αξιόπιστες υφομετρικές προσεγγίσεις και να ελέγξουμε την απόδοσή τους στο ίδιο πεδίο ελέγχου. Ειδικότερα, υλοποιήσαμε (i) ένα πολυπαραγοντικό μοντέλο συναρτήσεων πλούτου του λεξιλογίου [43] και (ii) ένα μοντέλο που βασίζεται στις συχνότητες εμφάνισης των πιο συχνών λέξεων του σώματος εκπαίδευσης [19].

Για τη μέτρηση του πλούτου του λεξιλογίου χρησιμοποιήσαμε ένα σύνολο πέντε συναρτήσεων: το K που έχει προταθεί από τον Yule [103], το R που έχει προταθεί από τον Honore [46], το W που έχει προταθεί από τον Brunet [17], το S που έχει προταθεί από τον Sichel [82] και το D που έχει προταθεί από τον Simpson [84] και ορίζονται ως ακολούθως:

$$K = \frac{10^4 (\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2}$$

$$R = \frac{(100 \log N)}{(1 - (\frac{V_1}{V}))}$$

$$W = N^{V^{-\alpha}}$$

$$S = \frac{V_2}{V}$$

$$D = \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)}$$

όπου το V_i είναι ο αριθμός των λέξεων που εμφανίζονται ακριβώς i φορές στο κείμενο (τα V και N ορίζονται στην παράγραφο 4.2.3) και το α είναι μία παράμετρος που συνήθως τίθεται ίση με 0,17. Το ίδιο σύνολο συναρτήσεων χρησιμοποιήθηκε από τον Baayen και τους συναδέλφους του για παρόμοιους σκοπούς [6]. Για κάθε κείμενο υπολογίζονται αυτές οι συναρτήσεις και παράγεται ένα διάνυσμα πέντε παραμέτρων. Αυτά τα διανύσματα μπορούν στην συνέχεια να κατηγοριοποιηθούν στο πιο πιθανό

είδος κειμένου εφαρμόζοντας μία από τις τεχνικές ταξινόμησης που περιγράψαμε στο προηγούμενο τμήμα.

Η δεύτερη λεξιλογική μέθοδος χρησιμοποιεί ως δείκτες ύφους τις συχνότητες εμφάνισης των πιο συχνών λέξεων του σώματος εκπαίδευσης. Συνήθως χρησιμοποιούνται 30 ή 50 πιο συχνές λέξεις. Για λόγους σύγκρισης αποφασίσαμε να υλοποιήσουμε και τις δύο αυτές προτάσεις. Έτσι, για κάθε κείμενο υπολογίζεται ένα διάνυσμα 30 (ή 50) παραμέτρων που υποδεικνύει τις συχνότητες εμφάνισης των πιο συχνών λέξεων του σώματος εκπαίδευσης (κανονικοποιημένες ως προς το μήκος του κειμένου). Όπως πριν, αυτά τα διανύσματα μπορούν να κατηγοριοποιηθούν στο πιο πιθανό είδος κειμένου.

5.4 Σώμα Κειμένων ανά Είδος

Εφόσον για τα Νέα Ελληνικά δεν υπήρχε διαθέσιμο κάποιο σώμα κειμένων που να είναι ταξινομημένο ανά είδη κειμένων, αποφασίσαμε να κατασκευάσουμε ένα από την αρχή. Το σώμα που χρησιμοποιήθηκε στα πειράματα του Μίχου [63] περιλαμβάνει ένα περιορισμένο αριθμό προσεκτικά επιλεγμένων κειμένων τα οποία είχαν μετατραπεί χειρονακτικά σε ηλεκτρονική μορφή και ταξινομήθηκαν σε γενικές κατηγορίες (π.χ. δημοσιογραφικά, λογοτεχνικά, επιστημονικά, κ.ά.). Επίσης, όπως έχει επισημανθεί από πολλούς ερευνητές, η χρήση ήδη υπαρχόντων σωμάτων κειμένων που δεν έχουν ταξινομηθεί με βάση το είδος τους δεν είναι το κατάλληλο πεδίο αξιολόγησης ενός συστήματος αναγνώρισης είδους κειμένου [54]. Και αυτό γιατί μια γενική κατηγορία (π.χ. δημοσιογραφικά κείμενα) μπορεί να μην είναι ομογενής υφολογικά.

Το σώμα που χρησιμοποιήθηκε σε αυτήν τη μελέτη περιλαμβάνει κείμενα που πληρούν τις πιο κάτω συνθήκες:

- **Κείμενο χωρίς περιορισμούς:** τα κείμενα πρέπει να είναι ήδη σε ηλεκτρονική μορφή με πιθανή συνέπεια να περιέχουν διάφορα είδη λαθών (π.χ. τυπογραφικά, ορθογραφικά, κ.ά.).
- **Ακατέργαστο κείμενο:** τα κείμενα δεν απαιτείται να έχουν υποστεί κάποιο σχολιασμό ούτε να έχουν υποστεί χειρονακτική προεπεξεργασία.

- **Ολόκληρο κείμενο:** κανένας περιορισμός δεν λαμβάνεται υπ' όψιν όσον αφορά το ελάχιστο ή το μέγιστο μέγεθος του κειμένου. Τα κείμενα πρέπει να είναι διαθέσιμα όπως ακριβώς εμφανίζονται στην πηγή τους.

Έτσι, κατασκευάσαμε ένα σώμα κειμένων ταξινομημένο ανά είδος συλλέγοντας Νεοελληνικά κείμενα από διάφορες ιστοσελίδες του Διαδικτύου. Να σημειωθεί ότι μέχρι τώρα δεν έχει οριστεί το πλήρες σύνολο ειδών κειμένων για τα Νέα Ελληνικά. Επίσης, το σύνολο των ειδών κειμένων μπορεί να διαφέρει από γλώσσα σε γλώσσα, αν και για τις ινδοευρωπαϊκές γλώσσες υπάρχουν πολύ μεγάλες ομοιότητες. Ως πρότυπο, λοιπόν, επιλέξαμε το σύνολο ειδών κειμένων που προτείνει ο Biber [11] για την Αγγλική γλώσσα, και το ακολουθήσαμε όπου αυτό ήταν δυνατό. Το σώμα κειμένων ταξινομημένο ανά είδος που κατασκευάστηκε φαίνεται στον πίνακα 5.1. Ασφαλώς, το σύνολο των ειδών κειμένου που παρουσιάζουμε εδώ δεν είναι πλήρες αλλά καλύπτει ένα αρκετά ευρύ φάσμα κειμένων της Νεοελληνικής γλώσσας και προσφέρεται για την αξιολόγηση της προτεινόμενης μεθόδου κατηγοριοποίησης.

Κωδικός	**Είδος κειμένου**	**Κείμενα**	**Λέξεις (μέσος όρος)**	**Πηγή**
E01	Άρθρα εφημερίδας	25	729	Εφημερίδα *Το Βήμα*[1]
E02	Ρεπορτάζ εφημερίδας	25	902	Εφημερίδα *Το Βήμα*
E03	Επιστημονικά άρθρα	25	2.120	Περιοδικό *Αρχεία Παθολογικής Ανατομικής*[2]
E04	Επίσημα έγγραφα	25	1.059	Αποφάσεις ανωτάτου δικαστηρίου[3], Υπουργικές αποφάσεις[4]
E05	Λογοτεχνία	25	1.508	Διάφορες σελίδες
E06	Μαγειρικές συνταγές	25	109	Περιοδικό *NetLife*[5]
E07	Βιογραφικά σημειώματα	25	333	Διάφορες σελίδες
E08	Συνεντεύξεις	25	2.625	Εφημερίδα *Το Βήμα*
E09	Προγραμματισμένες ομιλίες	25	2.569	Υπουργείο Εθνικής Αμύνης[6]
E10	Ραδιοφωνικές ειδήσεις	25	137	Ραδιοσταθμός *Flash 9.61*[7]

Πίνακας 5.1. Το σώμα κειμένων ανά είδος.

[1] http://tovima.dolnet.gr
[2] http://www.mednet.gr/hsap/apagr.htm
[3] http://senanet.com/cgi-bin/dsearch/form
[4] http://www.labor-ministry.gr/intr1_gr.htm
[5] http://www.netlife.gr/kitchen/recipes/index.htm
[6] http://www.mod.gr/
[7] http://www.flash.gr/

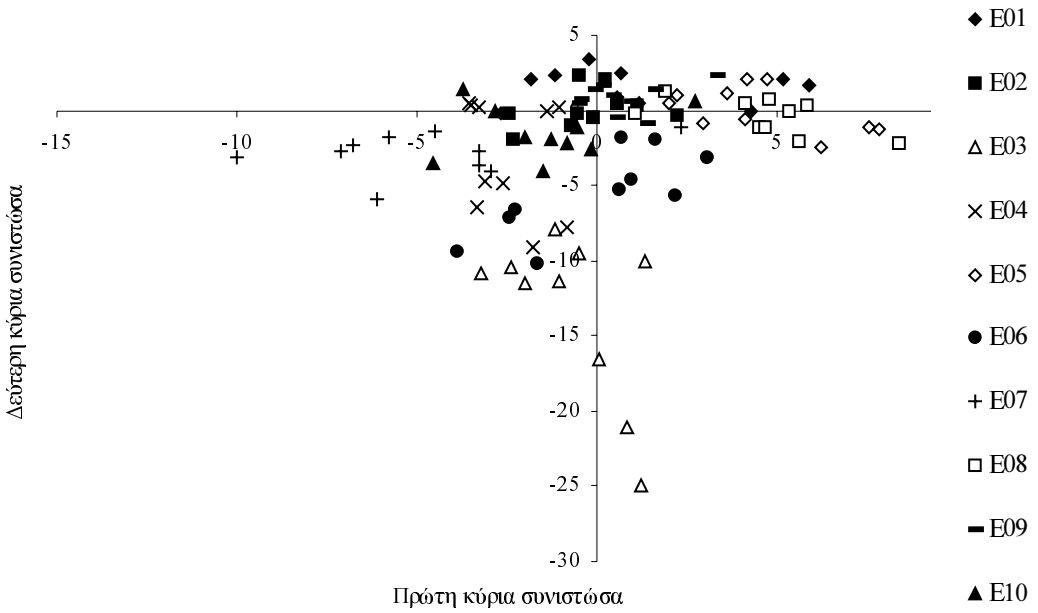
Είναι φανερό πως υπάρχει μεγάλη διαφορά στην τάξη μεγέθους του κειμένου μεταξύ των ειδών. Επίσης, αξίζει να σημειωθεί ότι τα τελευταία τρία είδη (Ε08, Ε09 και Ε10) αναφέρονται στον προφορικό λόγο. Οι προγραμματισμένες ομιλίες και οι ραδιοφωνικές ειδήσεις γράφτηκαν πριν την στιγμή που εκφωνήθηκαν ενώ οι συνεντεύξεις μεταφέρθηκαν στο χαρτί μετά την πραγματοποίησή τους. Τα υπόλοιπα είδη (Ε01 ως Ε07) αναφέρονται στον κατεξοχήν γραπτό λόγο.

5.5 Αξιολόγηση

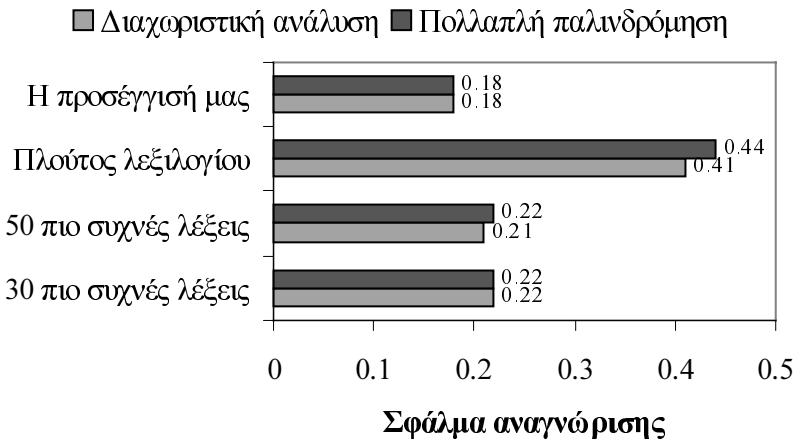
Το σώμα που περιγράφηκε στο προηγούμενο τμήμα αναλύθηκε από τον ανιχνευτή ορίων περιόδων και φράσεων, ο οποίος για κάθε κείμενο εξήγαγε ένα διάνυσμα 22 υφολογικών παραμέτρων, όπως περιγράφηκε στο κεφάλαιο 4. Δέκα κείμενα από κάθε είδος χρησιμοποιήθηκαν ως σώμα εκπαίδευσης και δέκα ως σώμα ελέγχου. Τα υπόλοιπα πέντε κείμενα από κάθε είδος χρησιμοποιήθηκαν στο πείραμα του τμήματος 5.5.1.

Πριν προχωρήσουμε στην διαδικασία κατηγοριοποίησης των κειμένων κρίνουμε σκόπιμο να δώσουμε μια πρώτη ένδειξη των βασικών ομοιοτήτων και διαφορών μεταξύ των κατηγοριών. Γι' αυτό το λόγο, εφαρμόσαμε την στατιστική τεχνική ανάλυσης κυρίων συνιστωσών (principal components analysis) στο σώμα ελέγχου. Η αναπαράσταση των 100 κειμένων του σώματος ελέγχου στον χώρο που ορίζεται από την πρώτη και τη δεύτερη κύρια συνιστώσα (που είναι υπεύθυνες του 43% της συνολικής διασποράς) φαίνεται στο σχήμα 5.2. Είναι ξεκάθαρο ότι κείμενα του ίδιου είδους κειμένου βρίσκονται περίπου στην ίδια περιοχή του διανυσματικού αυτού χώρου. Όμως, οι περιοχές αυτές δεν είναι σαφώς διακριτές μεταξύ τους.

Τα κείμενα του σώματος εκπαίδευσης χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων κατηγοριοποίησης τόσο για την πολλαπλή παλινδρόμηση όσο και για την διαχωριστική ανάλυση. Εκτός από την μέθοδό μας εκπαιδεύσαμε το μοντέλο με βάση τις συναρτήσεις πλούτου του λεξιλογίου καθώς και με βάση τις συχνότητες εμφάνισης των 30 και 50 πιο συχνών λέξεων. Η απόδοση των μοντέλων κατηγοριοποίησης που προέκυψαν ελέγχθηκε στο αντίστοιχο σώμα ελέγχου. Τα συγκριτικά αποτελέσματα ως προς το σφάλμα αναγνώρισης (λάθος αναγνωρισμένα κείμενα προς συνολικά κείμενα) φαίνονται στο σχήμα 5.3.



Σχήμα 5.2. Το σώμα ελέγχου στον χώρο των δύο πρώτων κυρίων συνιστωσών.



Σχήμα 5.3. Συγκριτικά αποτελέσματα για την αναγνώριση είδους κειμένου.

Γενικά, η διαχωριστική ανάλυση φαίνεται να διακρίνει καλύτερα τα κείμενα του σώματος ελέγχου. Η απόδοση των συναρτήσεων πλούτου λεξιλογίου είναι πολύ φτωχή. Αυτό οφείλεται στο πολύ μικρό μήκος των περισσότερων κειμένων του σώματος ανά είδος [95]. Η μέθοδός μας υπερτερεί σε κάθε περίπτωση των 30 και 50 πιο συχνών λέξεων.

Αναλυτικά αποτελέσματα του σφάλματος αναγνώρισης της μεθόδου μας δίνονται στον πίνακα 5.2. Αν και ο μέσος όρος σφάλματος αναγνώρισης είναι ο ίδιος και για τις δύο στατιστικές τεχνικές, υπάρχουν αρκετά σημαντικές διαφορές στα επιμέρους

σφάλματα αναγνώρισης μερικών ειδών (π.χ. E01 και E05). Γενικά, η κατανομή του σφάλματος είναι πιο ομαλή με χρήση της διαχωριστικής ανάλυσης. Περίπου το 60% του σφάλματος αναγνώρισης με χρήση της πολλαπλής παλινδρόμησης οφείλεται στα είδη E01 και E07, ενώ το 65% του σφάλματος αναγνώρισης με χρήση της διαχωριστικής ανάλυσης οφείλεται στα είδη E01, E05 και E07. Αντίθετα, τα είδη E04, E06, E08 και E10 έχουν πολύ καλά αποτελέσματα και στις δύο περιπτώσεις. Πρέπει να σημειωθεί ότι τα είδη κειμένων του προφορικού λόγου (E08-E10) αναγνωρίστηκαν με μεγαλύτερη ακρίβεια (μέσος όρος σφάλματος=0.10) από αυτά του γραπτού λόγου (μέσος όρος σφάλματος=0.21).

Κωδικός		**Σφάλμα αναγνώρισης**		
E01		0,7		0,4
E02		0,2		0,1
E03		0,0		0,0
E04		0,1		0,2
E05		0,1		0,4
E06		0,0		0,0
E07		0,4		0,4
E08		0,1		0,0
E09		0,2		0,2
E10		0,0		0,1
Μέσος όρος		**0,18**		**0,18**

Πολλαπλή παλινδρόμηση *Διαχωριστική ανάλυση*

Πίνακας 5.2. Τα αποτελέσματα αναγνώρισης είδους κειμένου.

Κωδικός	**Κατηγοριοποίηση**										**Σφάλμα**
	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**	**E08**	**E09**	**E10**	
E01	**6**	3	0	0	0	0	0	0	1	0	**0,4**
E02	0	**9**	0	0	0	0	0	0	0	1	**0,1**
E03	0	0	**10**	0	0	0	0	0	0	0	**0,0**
E04	0	1	0	**8**	0	0	0	0	0	1	**0,2**
E05	0	0	0	0	**6**	0	0	2	2	0	**0,4**
E06	0	0	0	0	0	**10**	0	0	0	0	**0,0**
E07	0	0	0	3	0	0	**6**	0	0	1	**0,4**
E08	0	0	0	0	0	0	0	**10**	0	0	**0,0**
E09	1	0	0	0	0	0	0	1	**8**	0	**0,2**
E10	0	0	0	1	0	0	0	0	0	**9**	**0,1**
										Μέσος όρος:	**0,18**

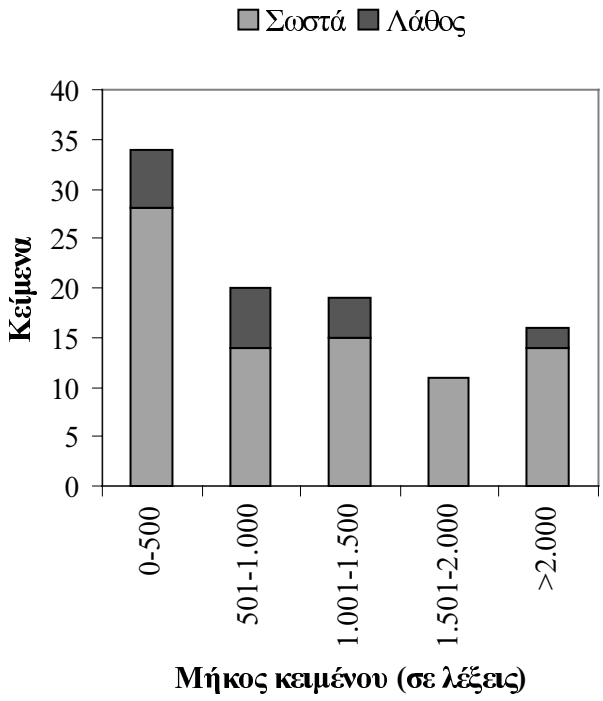
Πίνακας 5.3. Κατηγοριοποίηση του σώματος ελέγχου με βάση τη διαχωριστική ανάλυση.

Τα πλήρη αποτελέσματα ταξινόμησης σύμφωνα με τη μέθοδο διαχωριστικής ανάλυσης φαίνονται στον πίνακα 5.3. Κάθε γραμμή αυτού του πίνακα αντιστοιχεί στα 10 κείμενα ελέγχου ενός είδους κειμένου ενώ κάθε στήλη αναφέρεται στο αποτέλεσμα της κατηγοριοποίησης αυτών των κειμένων. Έτσι, η διαγώνιος του

πίνακα περιλαμβάνει τα κείμενα που έχουν κατηγοριοποιηθεί σωστά. Ο πίνακας αυτός είναι κατάλληλος για την εύρεση των πιο συχνών περιπτώσεων λάθους. Πιο συγκεκριμένα, τα συχνότερα λάθη γίνονται μεταξύ των πιο κάτω συνδυασμών:

- *Άρθρα εφημερίδας* → *ρεπορτάζ εφημερίδας*. Πρόκειται για κείμενα που ανήκουν στην ίδια εφημερίδα. Να σημειωθεί ότι *Το Βήμα* την περίοδο δειγματοληψίας κειμένων εκδιδόταν σε εβδομαδιαία βάση με αποτέλεσμα τα ρεπορτάζ να περιέχουν στοιχεία ανασκόπησης των γεγονότων μιας εβδομάδας.
- *Βιογραφικά σημειώματα* → *επίσημα έγγραφα*. Και τα δύο αυτά είδη χαρακτηρίζονται από υψηλό βαθμό αφαιρετικού ύφους.
- *Λογοτεχνία* → *συνεντεύξεις* και *προγραμματισμένες ομιλίες*. Αυτά τα τρία είδη χαρακτηρίζονται συνήθως από αφηγηματικό ύφος.

Βλέπουμε, λοιπόν, ότι τα πιο συχνά λάθη κατηγοριοποίησης είναι δυνατόν να εξηγηθούν με βάση τους παραδοσιακούς υφολογικούς όρους.



Σχήμα 5.4. Κατανομή του σώματος ελέγχου σύμφωνα με την ακρίβεια κατηγοριοποίησης και το μήκος κειμένου.

Στο σχήμα 5.4 φαίνεται η κατανομή των κειμένων του σώματος ελέγχου συναρτήσει του μήκους τους (σε λέξεις) και των αποτελεσμάτων της κατηγοριοποίησης σύμφωνα με τη μέθοδο πολλαπλής παλινδρόμησης. Όπως φαίνεται, η ακρίβεια των κειμένων με μέγεθος μικρότερο των 500 λέξεων είναι σχετικά υψηλή. Αυτό οφείλεται στον υψηλό βαθμό υφολογικής ομοιογένειας των ειδών *μαγειρικές συνταγές* και *ραδιοφωνικές ειδήσεις*. Γενικά πάντως, πιο αξιόπιστα αποτελέσματα επιτυγχάνουμε για κείμενα μεγαλύτερα των 1.500 λέξεων. Να σημειωθεί ότι ο Biber [9, 10] υποστηρίζει ότι ένα κειμένου της τάξης μεγέθους των 1.000 λέξεων είναι αρκετό για να αναπαρασταθούν επαρκώς τα γλωσσολογικά γνωρίσματα μιας υφολογικής κατηγορίας.

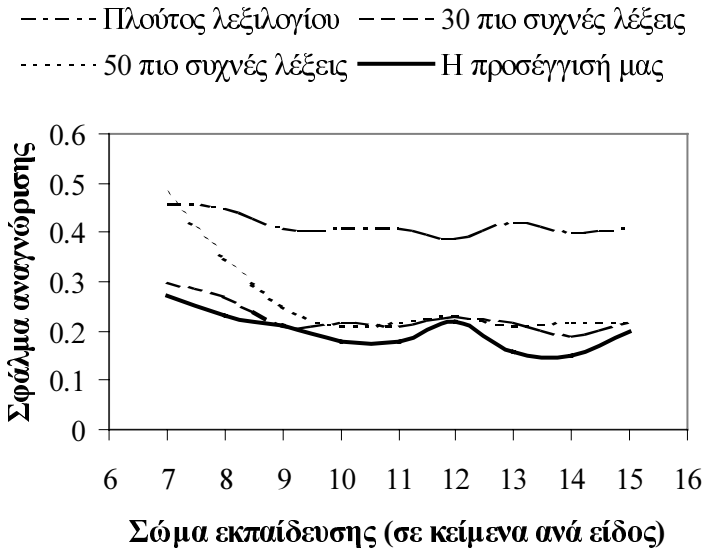
5.5.1 Μέγεθος σώματος εκπαίδευσης

Ένα άλλο συμπέρασμα στο οποίο κατέληξε ο Biber [9, 10] είναι ότι σχετικά λίγα κείμενα (γύρω στα δέκα) αρκούν για να αναπαραστήσουν επαρκώς τα γλωσσολογικά γνωρίσματα μιας υφολογικής κατηγορίας. Κατά το χωρισμό του σώματος κειμένων σε σώμα εκπαίδευσης και ελέγχου, ακολουθήσαμε αυτή τη θεώρηση, η οποία έρχεται σε συμφωνία με τις απαιτήσεις ενός πλήρως αυτοματοποιημένου συστήματος που πρέπει να μπορεί να εκπαιδεύεται εύκολα για μία νέα υφολογική κατηγορία. Να σημειωθεί ότι σε πολλές περιπτώσεις υπάρχουν ελάχιστα δεδομένα για εκπαίδευση. Παρ' όλα αυτά, κρίνουμε σκόπιμο να εξερευνήσουμε τη σχέση μεγέθους του σώματος εκπαίδευσης και ακρίβειας κατηγοριοποίησης.

Έτσι, χρησιμοποιήσαμε τη μέθοδο πολλαπλής παλινδρόμησης για να εκπαιδεύσουμε το σύστημα αναγνώρισης ειδών κειμένου βασιζόμενοι σε μέγεθος σώματος εκπαίδευσης που κυμαινόταν από 7 έως 15 κείμενα από κάθε είδος. Το σώμα ελέγχου ήταν σε όλες τις περιπτώσεις το ίδιο (αποτελούμενο από 10 κείμενα για κάθε είδος). Η ίδια διαδικασία ακολουθήθηκε και για τις λεξιλογικές προσεγγίσεις (δηλ. πλούτος λεξιλογίου, 30 πιο συχνές λέξεις, 50 πιο συχνές λέξεις). Στο σχήμα 5.5 φαίνονται συγκριτικά αποτελέσματα της σχέσης του μέσου όρου του σφάλματος αναγνώρισης και του σώματος εκπαίδευσης.

Η απόδοση των συναρτήσεων πλούτου του λεξιλογίου δεν επηρεάζεται σημαντικά από την αύξηση του σώματος εκπαίδευσης. Απ' την άλλη, το σφάλμα αναγνώρισης των 30 πιο συχνών λέξεων, των 50 πιο συχνών λέξεων και της μεθόδου μας γενικά μειώνεται όταν χρησιμοποιούνται περισσότερα κείμενα για εκπαίδευση. Η απόδοση

των 30 πιο συχνών λέξεων είναι πιο σταθερή σε σχέση με τις 50 πιο συχνές λέξεις αλλά σε κάθε περίπτωση η μέθοδός μας επιτυγχάνει καλύτερα αποτελέσματα.



Σχήμα 5.5. Το σφάλμα αναγνώρισης συναρτήσει του μεγέθους του σώματος εκπαίδευσης.

Το μικρότερο σφάλμα αναγνώρισης (0,15) επιτυγχάνεται από την μέθοδό μας χρησιμοποιώντας 14 κείμενα από κάθε είδος ως σώμα εκπαίδευσης. Ωστόσο, στο διάστημα από 11 ως 15 κείμενα, το σφάλμα αναγνώρισης δεν μειώνεται γραμμικά. Έτσι, το σφάλμα αναγνώρισης με χρήση 12 και 15 κειμένων ανά είδος ως σώμα εκπαίδευσης είναι υψηλότερο από αυτό των 10 κειμένων ανά είδος. Επομένως, χρησιμοποιώντας 10 κείμενα ανά είδος ως σώμα εκπαίδευσης επιτυγχάνουμε σχετικά καλά αποτελέσματα.

5.5.2 Σημαντικότητα των υφολογικών δεικτών

Όπως τονίστηκε στο κεφάλαιο 4, οι υφολογικοί δείκτες που χρησιμοποιούμε διακρίνονται σε τρία επίπεδα: δείγματος, φράσης και ανάλυσης. Θα ήταν, λοιπόν, πολύ χρήσιμο να διερευνηθεί η συμβολή του κάθε επιπέδου δεικτών στην διαδικασία κατηγοριοποίησης. Για την μέτρηση της σημαντικότητας της κάθε υφολογικής παραμέτρου χρησιμοποιήθηκαν οι απόλυτες τιμές του στατιστικού-*t* των συντελεστών των συναρτήσεων γραμμικής παλινδρόμησης. Υπενθυμίζουμε ότι όσο μεγαλύτερη είναι η απόλυτη τιμή του *t* για κάποιον συντελεστή παλινδρόμησης τόσο

πιο σημαντική είναι η ανεξάρτητη μεταβλητή (δηλ. ο υφολογικός δείκτης) με την οποία σχετίζεται, όσον αφορά τη συνεισφορά της στην τιμή απόκρισης (βλ. § 5.2.1).

Επίπεδο	**Δείκτης ύφους**	**Απόλυτη τιμή *t***	**Μέσος όρος**
Δείγματος	Δ01	1,06	1,67
	Δ02	2,52	
	Δ03	1,43	
Φράσεων	Δ04	0,57	0,75
	Δ05	0,58	
	Δ06	0,56	
	Δ07	0,57	
	Δ08	0,57	
	Δ09	0,77	
	Δ10	0,93	
	Δ11	0,59	
	Δ12	1,72	
	Δ13	0,67	
Ανάλυσης	Δ14	1,03	1,11
	Δ15	2,11	
	Δ16	1,45	
	Δ17	1,08	
	Δ18	0,72	
	Δ19	1,14	
	Δ20	1,00	
	Δ21	0,81	
	Δ22	0,65	

Πίνακας 5.4. Μέσες τιμές *t* των συντελεστών παλινδρόμησης.

Στον πίνακα 5.4 δίνονται οι μέσοι όροι των απόλυτων τιμών του *t* για καθένα από τους δείκτες ύφους που χρησιμοποιήθηκαν στο πείραμα αναγνώρισης είδους κειμένου, με χρήση της πολλαπλής παλινδρόμησης και με βάση 10 κείμενα από κάθε είδος ως σώμα εκπαίδευσης. Το πιο σημαντικό επίπεδο δεικτών είναι αυτό του δείγματος. Το επίπεδο ανάλυσης, απ' την άλλη, αποδεικνύεται πιο σημαντικό από αυτό της φράσης. Πιο σημαντικές παράμετροι για το διαχωρισμό των κατηγοριών είναι τα σημεία στίξης ανά λέξη και οι ειδικές λέξεις ανά λέξη (Δ02 και Δ15 αντίστοιχα). Επίσης, όσον αφορά το επίπεδο φράσης, πρέπει να τονιστεί ότι οι δείκτες που σχετίζονται με τον αριθμό των λέξεων που εσωκλείονται σε κάθε είδος φράσης (Δ09-Δ13) είναι πιο σημαντικοί από τις συχνότητες εμφάνισης των ειδών των φράσεων (Δ04-Δ08).

5.5.3 Ελαττωματική ανάλυση

Καθώς ο ανιχνευτής ορίων περιόδων και φράσεων που χρησιμοποιείται για την εξαγωγή των υφολογικών δεικτών είναι ένα αυτοματοποιημένο εργαλείο επεξεργασίας κειμένου είναι πολύ χρήσιμο να διερευνηθεί σε ποιο βαθμό η ακρίβεια της ανάλυσης που παρέχει επηρεάζει την ακρίβεια κατηγοριοποίησης αλλά και τη σημαντικότητα των δεικτών ύφους. Έτσι, πραγματοποιήσαμε το ακόλουθο πείραμα. Δημιουργήσαμε τεχνητή ελαττωματική ανάλυση κειμένου επεμβαίνοντας σε ορισμένα σημεία της διαδικασίας ανίχνευσης ορίων περιόδων και φράσεων. Πιο συγκεκριμένα:

- Όσον αφορά τη διαδικασία ανίχνευσης ορίων περιόδων, μόνο οι τελείες θεωρήθηκαν ως πιθανά όρια περιόδου.
- Όσον αφορά τη διαδικασία ανίχνευσης ορίων φράσεων, το πέμπτο πέρασμα ανάλυσης δεν λήφθηκε υπ' όψιν.

Αυτές οι επεμβάσεις επηρέασαν σημαντικά την απόδοση του ανιχνευτή ορίων περιόδων και φράσεων. Επαναλάβαμε το πείραμα αναγνώρισης ειδών κειμένων με χρήση της πολλαπλής παλινδρόμησης, με βάση αυτή τη φορά τα ελαττωματικά αποτελέσματα ανάλυσης. Το σφάλμα αναγνώρισης αυξήθηκε κατά 25% περίπου (0,23). Επομένως, η ακρίβεια κατηγοριοποίησης είναι άμεσα εξαρτημένη από την ακρίβεια ανάλυσης του κειμένου.

Υφομετρικό Επίπεδο	**Μέσος όρος απόλυτης τιμής *t***	
	Κανονική ανάλυση	**Ελαττωματική ανάλυση**
Δείγματος	1,67	1,55
Φράσεων	0,75	0,97
Ανάλυσης	1,11	1,29

Πίνακας 5.5. Συγκριτικές τιμές του μέσου όρου του απόλυτου *t* για κανονική και ελαττωματική ανάλυση.

Εκτός από την ακρίβεια κατηγοριοποίησης επηρεάστηκε και η σημαντικότητα των υφολογικών επιπέδων. Συγκριτικά αποτελέσματα των μέσων όρων των τιμών *t* των τριών επιπέδων για την κανονική και την ελαττωματική ανάλυση δίνονται στον πίνακα 5.5. Η διάταξη των τριών επιπέδων ως προς τη σημαντικότητά τους παραμένει

η ίδια. Ωστόσο, παρατηρείται αύξηση της σημαντικότητας των επιπέδων ανάλυσης και φράσεων και μείωση της σημαντικότητας του επιπέδου δείγματος.

5.6 Περίληψη - Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε την εφαρμογή του προτεινόμενου συνόλου των υφολογικών δεικτών (βλ. κεφάλαιο 4) στο πρόβλημα της αναγνώρισης είδους κειμένου. Στα πειράματα που πραγματοποιήθηκαν χρησιμοποιήθηκαν δείγματα από 10 είδη κειμένων σε ηλεκτρονική μορφή, τα οποία βρέθηκαν στο Διαδίκτυο. Ασφαλώς, τα είδη αυτά δεν είναι το πλήρες σύνολο των ειδών κειμένων της Νέας Ελληνικής γλώσσας. Παρ' όλα αυτά καλύπτουν ένα αρκετά ευρύ φάσμα κειμένων και προσφέρουν την δυνατότητα εξαγωγής αξιόπιστων συμπερασμάτων. Τα αποτελέσματα κρίνονται πολύ ικανοποιητικά αφού η ακρίβεια ταξινόμησης είναι της τάξης του 82-85% που είναι πολύ καλύτερη από τις επιδόσεις παρόμοιων συστημάτων για την Αγγλική γλώσσα [49, 54]. Στον πίνακα 5.6 δίνονται συγκριτικά αποτελέσματα για την απόδοση των δύο πιο γνωστών συστημάτων για την Αγγλική γλώσσα. Να σημειωθεί ότι η απόδοση του συστήματος των Karlgren και Cutting αναφέρεται στο σώμα εκπαίδευσης.

Σύστημα	**Είδη κειμένων**	**Ακρίβεια**
Karlgren & Cutting [48]	15	65%
Kessler et al. [53]	6	61-79%
Το δικό μας	10	82-85%

Πίνακας 5.6. Συγκριτικά αποτελέσματα συστημάτων αναγνώρισης είδους κειμένου.

Παρατηρήθηκε ότι η πλειοψηφία των λαθών προκλήθηκε από είδη κειμένων που εξ ορισμού δεν χαρακτηρίζονται από απόλυτη υφολογική ομοιογένεια (άρθρα εφημερίδων, λογοτεχνία και βιογραφικά σημειώματα). Με άλλα λόγια είναι πολύ πιθανό αυτά τα συγκεκριμένα είδη να μπορούν να διακριθούν σε υφολογικά ομοιογενείς υποκατηγορίες (π.χ. η λογοτεχνία μπορεί να διακριθεί σε νουβέλες, διηγήματα, μυθιστορήματα, κ.ά.). Επίσης, τα πιο συχνά λάθη ταξινόμησης εξηγούνται με όρους της παραδοσιακής υφολογίας. Είναι φανερό λοιπόν, ότι το προτεινόμενο σύνολο υφολογικών δεικτών είναι ικανό να ανιχνεύσει την υφολογική ομοιογένεια.

Η πολλαπλή παλινδρόμηση και η διαχωριστική ανάλυση, που χρησιμοποιήθηκαν για την αυτόματη ταξινόμηση των κειμένων ανά είδος, ανήκουν στις παραδοσιακές τεχνικές της πολυπαραγοντικής στατιστικής. Καθώς απαιτούν ελάχιστο χρονικό κόστος εκπαίδευσης και απόκρισης συμβαδίζουν απόλυτα με τις προδιαγραφές μιας πρακτικής εφαρμογής. Ακόμη, η ικανότητά τους να επιτυγχάνουν πολύ ικανοποιητικά αποτελέσματα βασιζόμενες σε σχετικά λίγα δεδομένα εκπαίδευσης τούς δίνει ένα επιπλέον πλεονέκτημα. Γενικά, οι δύο αυτές τεχνικές δίνουν παρόμοια αποτελέσματα αν και είναι φανερό πως η διαχωριστική ανάλυση επιτυγχάνει πιο ομαλή κατανομή του σφάλματος αναγνώρισης.

Συγκριτικά πειράματα με χρήση διαφορετικών μεγεθών του σώματος εκπαίδευσης έδειξε ότι η ακρίβεια ταξινόμησης μπορεί να βελτιωθεί με την αύξηση του σώματος εκπαίδευσης αν και αυτή η βελτίωση δεν είναι πάντα σημαντική. Η χρήση 10 κειμένων από κάθε είδος κειμένου ως σώμα εκπαίδευσης, δίνει πολύ ικανοποιητικά αποτελέσματα και αυτό αποκτά ιδιαίτερη σημασία αν αναλογιστούμε ότι στην πλειοψηφία των περιπτώσεων δεν είναι διαθέσιμα αρκετά κείμενα για εκπαίδευση.

Η σημαντικότητα των υφολογικών δεικτών μετρήθηκε με αντικειμενικά κριτήρια (συνεισφορά της κάθε παραμέτρου στην συνάρτηση απόκρισης της πολλαπλής παλινδρόμησης). Τα αποτελέσματα έδειξαν ότι το πιο σημαντικό υφομετρικό επίπεδο είναι αυτό του δείγματος. Πολύ ενδιαφέρον είναι το γεγονός ότι το επίπεδο ανάλυσης, που είναι ένας εναλλακτικός τρόπος σύλληψης της υφολογικής πληροφορίας, είναι πιο σημαντικό από το επίπεδο φράσης. Όσον αφορά τους επιμέρους δείκτες, πολύ σημαντικό ρόλο στην αναγνώριση είδους κειμένου φαίνεται να παίζουν η χρήση των σημείων στίξης και η συχνότητα των λέξεων που δεν ταιριάζουν με καμία κοινή κατάληξη της Νέας Ελληνικής γλώσσας (ειδικές λέξεις).

Το ολοκληρωμένο σύστημα αναγνώρισης είδους κειμένου μπορεί να χρησιμοποιηθεί σε οποιαδήποτε εφαρμογή απαιτεί ταξινόμηση κειμένων σύμφωνα με το είδος τους. Ενδεικτικά, αναφέρουμε τις εφαρμογές ανάκτησης και εξαγωγής πληροφορίας. Οι απαιτήσεις σε υπολογιστικό κόστος είναι ελάχιστες αφού ο ανιχνευτής ορίων περιόδων και φράσεων, στον οποίο βασίζεται η εξαγωγή των υφολογικών δεικτών, βασίζεται σε ελάχιστους πόρους και η διαδικασία κατηγοριοποίησης βασίζεται στον υπολογισμό απλών γραμμικών συναρτήσεων. Η εκπαίδευση του συστήματος για την αναγνώριση ενός συγκεκριμένου συνόλου ειδών κειμένου είναι απλή και με βάση

περίπου 10 κείμενα από κάθε είδος ως σώμα εκπαίδευσης επιτυγχάνονται πολύ ικανοποιητικά αποτελέσματα. Επίσης, το σύστημα αυτό θα μπορούσε να χρησιμοποιηθεί για την αυτόματη ανίχνευση ομοιογένειας, όσον αφορά το υφολογικό επίπεδο, σε μεγάλα σώματα κειμένων [55].