Κεφάλαιο 6

Προσδιορισμός Συγγραφέα

6.1 Εισαγωγή

Όπως αναφέρθηκε και στο κεφάλαιο 4, οι μέχρι σήμερα προσεγγίσεις στον προσδιορισμό συγγραφέα μέσω στατιστικών μεθόδων είναι βοηθούμενες από υπολογιστή παρά βασισμένες σε υπολογιστή. Με άλλα λόγια, δεν έχει αναπτυχθεί έως σήμερα ένα αυτοματοποιημένο σύστημα ικανό να επεξεργαστεί οποιοδήποτε κείμενο και να εκπαιδευτεί σε οποιονδήποτε συγγραφέα. Επίσης, οι υφολογικοί δείκτες που χρησιμοποιούνται αφορούν σχεδόν αποκλειστικά λεξιλογική πληροφορία [44], με αποτέλεσμα είτε να εξαρτώνται άμεσα από το μέγεθος του προς ανάλυση κειμένου (συναρτήσεις πλούτου λεξιλογίου, βλ. § 4.2.3), είτε να μην μπορούν να εφαρμοστούν σε διαφορετικές ομάδες συγγραφέων (συχνότητες εμφάνισης λειτουργικών λέξεων, βλ. § 4.2.4).

Οι απόπειρες προσδιορισμού συγγραφέα που έχουν γίνει μέχρι σήμερα ασχολούνται, ως επί το πλείστον, με την αναγνώριση της πατρότητας ανώνυμων ή αμφισβητούμενων λογοτεχνικών κειμένων. Έτσι, έχουν προταθεί τεχνικές που αποδεικνύουν ότι το *The Revenger's Tragedy* γράφτηκε από τον Middleton και όχι

από τον Tourneur [88], ότι η πρώτη και η δεύτερη πράξη του *Pericles* είναι πιο πιθανό να γράφτηκε από τον Wilkins παρά από τον Shakespeare [87], ότι ο συγγραφέας των 12 αμφισβητούμενων *Federalist Papers* (μία συλλογή δοκιμίων του 18ου αιώνα) ήταν ο Madison και όχι ο Hamilton [44, 66, 94] κ.ά. Ωστόσο, η χρησιμοποίηση τέτοιων περιπτώσεων για τον έλεγχο υφολογικών θεωριών έχει κάποια πολύ σοβαρά μειονεκτήματα:

- Ο αριθμός των υποψήφιων συγγραφέων είναι συνήθως περιορισμένος (δύο ή τρεις). Η τεχνική που ελέγχεται, λοιπόν, είναι πιθανό να αποδειχτεί λιγότερο ακριβής εφόσον εφαρμοστεί αργότερα σε περιπτώσεις με περισσότερους υποψήφιους συγγραφείς (περισσότερους από πέντε).
- Τα λογοτεχνικά κείμενα είναι συνήθως μεγάλα σε μέγεθος (αρκετές χιλιάδες λέξεις). Επομένως, εφόσον αναπτυχθεί μία μέθοδος που θα απαιτεί κείμενα με μεγάλο μήκος για να δώσει ικανοποιητικά αποτελέσματα, δεν θα μπορεί να εφαρμοστεί σε περιπτώσεις με κείμενα περιορισμένου μήκους.
- Τα λογοτεχνικά κείμενα συνήθως δεν είναι ομοιογενή καθώς μπορεί να περιέχουν διάλογους, αφηγήσεις κτλ. Επομένως, μία ολοκληρωμένη προσέγγιση θα απαιτούσε την ανάπτυξη εργαλείων αυτόματης επιλογής των πιο χαρακτηριστικών τμημάτων του κειμένου στα οποία το ύφος του συγγραφέα αντικατοπτρίζεται καλύτερα.

Σε αυτό το κεφάλαιο παρουσιάζουμε ένα πλήρως αυτόματο σύστημα ικανό να εκτελεί αναγνώριση συγγραφέα. Ειδικότερα, περιγράφουμε πειράματα αυτόματης ταξινόμησης ενός σώματος κειμένων από εφημερίδες με γνώμονα τον συγγραφέα τους. Η μέθοδός μας βασίζεται στο σύνολο των υφολογικών δεικτών που προτάθηκε στο κεφάλαιο 4 σε συνδυασμό με τις μεθόδους κατηγοριοποίησης που εφαρμόστηκαν στην αναγνώριση είδους κειμένου (βλ. § 5.2). Επιπλέον, η σύγκριση της μεθόδου μας με την πιο σύγχρονη λεξιλογική προσέγγιση δείχνει ότι η πρότασή μας επιτυγχάνει καλύτερα αποτελέσματα. Παρ' όλα αυτά, ο συνδυασμός των δύο μεθόδων είναι η πιο αξιόπιστη λύση. Επίσης, παρουσιάζεται μία προσέγγιση αυτόματης επιβεβαίωσης συγγραφέα που δεν απαιτεί περαιτέρω εκπαίδευση του συστήματος.

Στο επόμενο τμήμα περιγράφεται το σώμα κειμένων που χρησιμοποιήθηκε σε αυτήν τη μελέτη. Το τμήμα 6.3 περιλαμβάνει τα πειράματα που πραγματοποιήθηκαν

σχετικά με την αναγνώριση συγγραφέα. Επίσης, γίνεται συγκριτική μελέτη των αποτελεσμάτων με βάση λεξιλογικές μεθόδους και διερευνάται η σχέση μεγέθους σώματος εκπαίδευσης και ακρίβειας ταξινόμησης καθώς και η σημαντικότητα των προτεινόμενων υφολογικών δεικτών. Στο τμήμα 6.4 περιγράφονται τα πειράματα επιβεβαίωσης συγγραφέα και, τέλος, το τμήμα 6.5 περιλαμβάνει περιληπτικά τα συμπεράσματα που αποκομίστηκαν.

6.2 Σώμα Κειμένων ανά Συγγραφέα

Το σώμα κειμένων που χρησιμοποιήθηκε σε αυτήν την εργασία αποτελείται από κείμενα που βρέθηκαν στην ηλεκτρονική έκδοση της εβδομαδιαίας εφημερίδας *Το Βήμα*[1]. Η επιλογή αυτής της συγκεκριμένης εφημερίδας έγινε λόγω του ότι η ηλεκτρονική της έκδοση περιέχει μεγάλη ποικιλία κειμένων. Συνήθως, είναι διαθέσιμα ολόκληρα τα κείμενα και όχι μόνο κάποια αποσπάσματα από αυτά. Ακόμη, υπάρχει ένα μεγάλο σύνολο δημοσιογράφων, επιστημόνων, λογοτεχνών, κτλ. που δημοσιεύουν κείμενά τους σε εβδομαδιαία βάση σε αυτήν την εφημερίδα. Επομένως, ήταν δυνατή η συλλογή ενός ικανοποιητικού αριθμού κειμένων από κάθε συγγραφέα. *Το Βήμα* αποτελείται από εννιά τμήματα, όπως φαίνεται στον πίνακα 6.1.

Κωδικός τμήματος	**Τίτλος**	**Περιγραφή**
Α	*Το Βήμα*	Άρθρα, ρεπορτάζ, πολιτική, διπλωματία, αθλητικά
Β	*Νέες Εποχές*	Ένθετο περί πολιτισμού
Γ	*Το Άλλο Βήμα*	Περιοδικό ποικίλης ύλης
Δ	*Ανάπτυξη*	Οικονομικά άρθρα, νέα επιχειρήσεων
Ε	*Η Δραχμή σας*	Προσωπική οικονομία
Ι	*Ειδική Έκδοση*	Ειδικό εβδομαδιαίο ένθετο
Σ	*Βιβλία*	Ένθετο περί νέων βιβλίων
Ζ	*Τέχνες και Καλλιτέχνες*	Ένθετο περί τέχνης
Τ	*Ταξίδια*	Ένθετο με ταξιδιωτικές περιηγήσεις, πληροφορίες

Πίνακας 6.1. Η δομή της εφημερίδας *Το Βήμα*.

Για τον καλύτερο έλεγχο της προτεινόμενης μεθοδολογίας σχηματίσαμε δύο ομάδες συγγραφέων, όπως φαίνεται και στον πίνακα 6.2:

[1] http://tovima.dolnet.gr

	Κωδικός	Όνομα συγγραφέα	Κείμενα	Λέξεις (μέσος όρος)	Θεματική περιοχή
Ομάδα Α	Α01	Ν. Νικολάου	20	797	Οικονομία
	Α02	Ν. Μαράκης	20	871	Διπλωματία
	Α03	Δ. Ψυχογιός	20	535	Πολιτική
	Α04	Γ. Μπήτρος	20	689	Πολιτική, κοινωνία
	Α05	Δ. Νικολακόπουλος	20	1.162	Πολιτική, κοινωνία
	Α06	Θ. Λιανός	20	696	Κοινωνία
	Α07	Κ. Χαλβατζάκης	20	1.061	Τεχνολογία
	Α08	Γ. Λακόπουλος	20	1.248	Πολιτική
	Α09	Ρ. Σωμερίτης	20	721	Πολιτική, κοινωνία
	Α10	Δ. Μητρόπουλος	20	888	Διπλωματία
Ομάδα Β	Β01	Δ. Μαρωνίτης	30	572	Πολιτισμός, κοινωνία
	Β02	Μ. Πλωρίτης	30	1.166	Πολιτισμός, κοινωνία
	Β03	Κ. Τσουκαλάς	30	1.380	Διπλωματία
	Β04	Χ. Κιοσσέ	30	1.689	Αρχαιολογία
	Β05	Σ. Αλαχιώτης	30	1.005	Βιολογία
	Β06	Γ. Μπαμπινιώτης	30	1.158	Γλωσσολογία
	Β07	Θ. Τάσιος	30	1.020	Τεχνολογία, κοινωνία
	Β08	Γ. Δερτιλής	30	894	Ιστορία, κοινωνία
	Β09	Α. Λιάκος	30	1.256	Ιστορία, κοινωνία
	Β10	Γ. Βώκος	30	985	Φιλοσοφία

Πίνακας 6.2. Το σώμα κειμένων ανά συγγραφέα.

- **Ομάδα Α:** Αποτελείται από δέκα συγγραφείς των οποίων τα κείμενα εμφανίζονται ως επί το πλείστον στο τμήμα Α. Αυτό το τμήμα περιέχει κείμενα δημοσιογράφων πάνω σε θέματα της επικαιρότητας. Αξίζει να σημειωθεί ότι ένας συγγραφέας μπορεί να υπογράφει κείμενα από διαφορετικά είδη κειμένων (π.χ. ρεπορτάζ και άρθρα). Η επιλογή των συγκεκριμένων συγγραφέων ήταν τυχαία.
- **Ομάδα Β:** Αποτελείται από δέκα συγγραφείς των οποίων τα κείμενα εμφανίζονται ως επί το πλείστον στο τμήμα Β. Αυτό το τμήμα περιέχει δοκίμια τα οποία έχουν γραφτεί από επιστήμονες, λογοτέχνες, κτλ. και σπάνια από δημοσιογράφους. Συνήθως, σε τέτοια κείμενα το προσωπικό ύφος του κάθε συγγραφέα δεν επισκιάζεται από τα χαρακτηριστικά του είδους κειμένου. Η επιλογή των συγκεκριμένων συγγραφέων ήταν τυχαία.

Τα κείμενα που συγκεντρώθηκαν δεν υπέστησαν καμία χειρονακτική προεπεξεργασία (επιλογή συγκεκριμένων τμημάτων ή δειγματοληψία). Για να ελαχιστοποιηθεί η πιθανότητα αλλαγής του προσωπικού ύφους ενός συγκεκριμένου συγγραφέα κατά το πέρασμα του χρόνου, επιλέξαμε κείμενα που δημοσιεύτηκαν σε φύλλα της εφημερίδας από το 1997 μέχρι τους πρώτους μήνες του 1999. Η τελευταία στήλη του

πίνακα 6.2 αναφέρεται στην θεματική περιοχή της πλειοψηφίας των κειμένων του αντίστοιχου συγγραφέα. Να σημειωθεί ότι αυτή η πληροφορία δεν λήφθηκε υπ' όψιν κατά την κατασκευή του παρουσιαζόμενου σώματος.

Παρατηρούμε ότι το μέσο μήκος του κειμένου παρουσιάζει αρκετά μεγάλη διαφορά από συγγραφέα σε συγγραφέα. Σε γενικές γραμμές, οι συγγραφείς της ομάδας Α έχουν χαμηλότερο μέσο μήκος κειμένου από αυτούς της ομάδας Β.

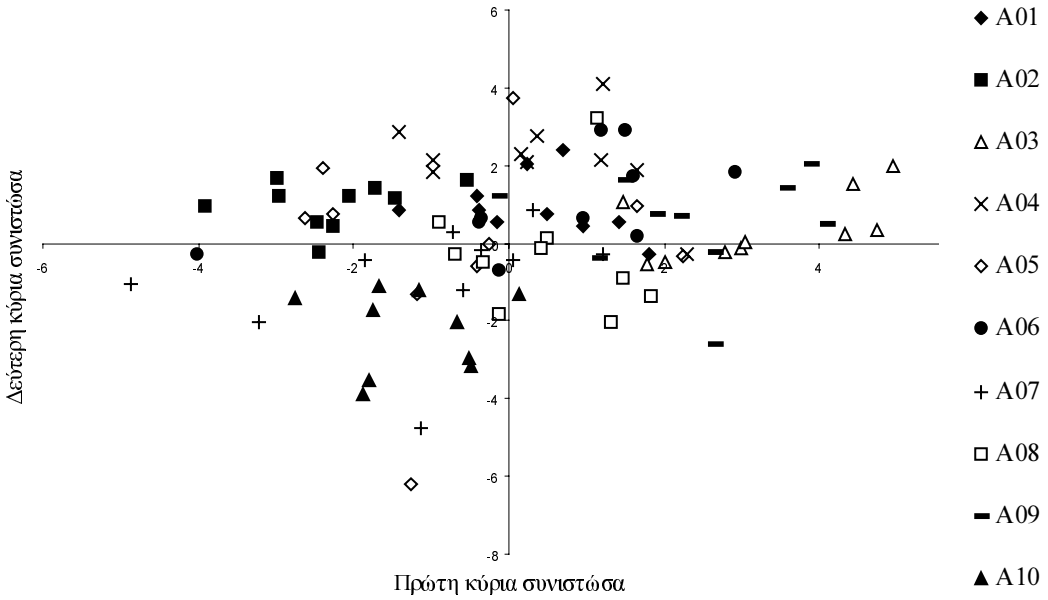
Από κάθε συγγραφέα 10 κείμενα χρησιμοποιήθηκαν ως σώμα εκπαίδευσης και άλλα 10 ως σώμα ελέγχου. Τα υπόλοιπα 10 κείμενα των συγγραφέων της ομάδας Β χρησιμοποιήθηκαν μόνο στα πειράματα των τμημάτων 6.3.1, 6.3.2 και 6.3.4.

6.3 Αναγνώριση Συγγραφέα

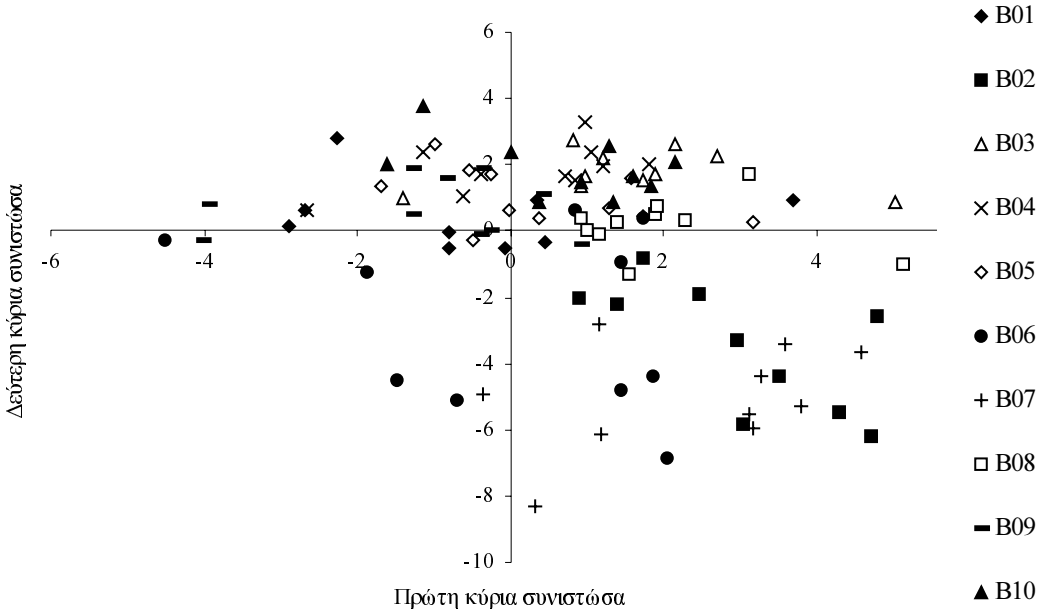
Η μεθοδολογία που παρουσιάστηκε στην αναγνώριση είδους κειμένου εφαρμόστηκε και στην αναγνώριση συγγραφέα. Όλα τα κείμενα αναλύθηκαν από τον ανιχνευτή ορίων περιόδων και φράσεων και έτσι έγινε διαθέσιμο ένα διάνυσμα 22 υφολογικών παραμέτρων για το κάθε κείμενο.

Πριν περάσουμε στην διαδικασία αυτόματης ταξινόμησης, θεωρούμε χρήσιμο να δοθεί μία πρώτη εντύπωση για τις βασικές ομοιότητες και διαφορές μεταξύ των συγγραφέων των δύο ομάδων. Προς αυτήν την κατεύθυνση, εφαρμόστηκε η στατιστική τεχνική της ανάλυσης κυρίων συνιστωσών στο σώμα ελέγχου της καθεμίας ομάδας συγγραφέων. Η αναπαράσταση λοιπόν, του σώματος ελέγχου των ομάδων Α και Β στο χώρο που ορίζουν οι δύο πρώτες κύριες συνιστώσες φαίνεται στα σχήματα 6.1 και 6.2 αντίστοιχα. Είναι φανερό ότι τα κείμενα του ίδιου συγγραφέα βρίσκονται στην ίδια περίπου περιοχή. Ωστόσο, οι περιοχές αυτές δεν μπορούν να διακριθούν σαφώς μεταξύ τους. Ακόμη, η σύγκριση με την αντίστοιχη αναπαράσταση του σώματος κειμένων ανά είδος (βλ. σχήμα 5.2) δείχνει ότι το σώμα κειμένων ανά συγγραφέα (και οι δύο ομάδες) είναι πιο συμπαγές, δηλ. αποτελείται από κείμενα πιο ομοιογενή υφολογικά, καθώς υπάρχει σημαντική διαφορά στην τάξη μεγέθους του διαστήματος τιμών των δύο πρώτων κυρίων συνιστωσών. Για παράδειγμα, η πρώτη κύρια συνιστώσα παίρνει τιμές στο διάστημα [–5, +5] στην περίπτωση του σώματος κειμένων ανά συγγραφέα (και στις δύο ομάδες) ενώ στην περίπτωση του σώματος κειμένων ανά είδος παίρνει τιμές στο διάστημα [–10, +8].

Επομένως θα πρέπει να αναμένουμε χειρότερα αποτελέσματα ταξινόμησης σε σχέση με αυτά της αναγνώρισης κειμένου.

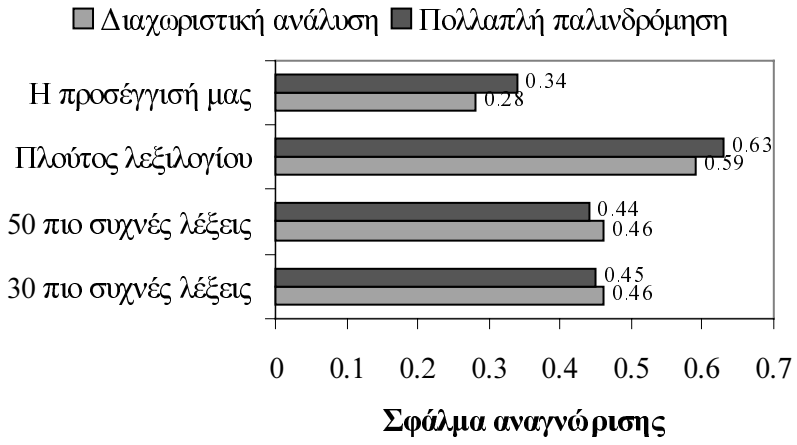


Σχήμα 6.1. Αναπαράσταση του σώματος ελέγχου της ομάδας Α στο χώρο των δύο πρώτων κυρίων συνιστωσών.

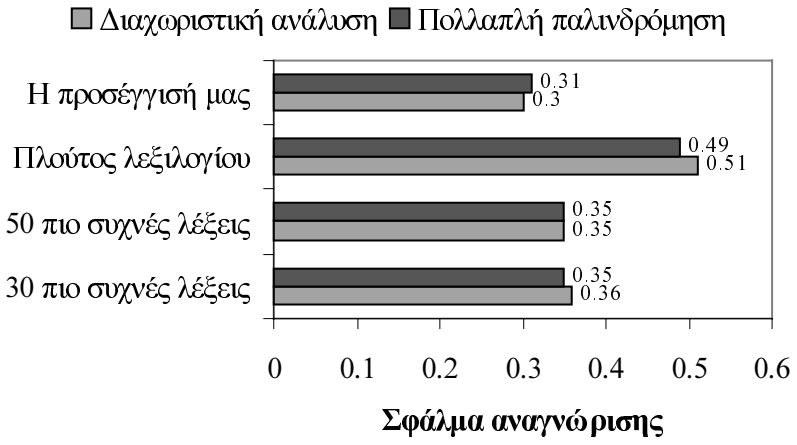


Σχήμα 6.2. Αναπαράσταση του σώματος ελέγχου της ομάδας Β στο χώρο των δύο πρώτων κυρίων συνιστωσών.

Στην συνέχεια, με βάση το σώμα εκπαίδευσης, εκπαιδεύσαμε το σύστημα αναγνώρισης συγγραφέα για την κάθε μία ομάδα συγγραφέων ξεχωριστά, τόσο με βάση τη μέθοδό μας όσο και με βάση καθεμία από τις λεξιλογικές μεθόδους που χρησιμοποιήθηκαν και στην αναγνώριση είδους κειμένου (βλ. § 5.3). Συγκριτικά αποτελέσματα της απόδοσης των μοντέλων κατηγοριοποίησης που προέκυψαν στο αντίστοιχο σώμα ελέγχου (αξιολόγηση κλειστού συνόλου) φαίνονται στα σχήματα 6.3 και 6.4, για την ομάδα Α και Β αντίστοιχα.



Σχήμα 6.3. Συγκριτικά αποτελέσματα για την αναγνώριση συγγραφέα στην ομάδα Α.



Σχήμα 6.4. Συγκριτικά αποτελέσματα για την αναγνώριση συγγραφέα στην ομάδα Β.

Όπως και στην περίπτωση της αναγνώρισης είδους κειμένου, η απόδοση των συναρτήσεων πλούτου του λεξιλογίου είναι πολύ φτωχή συγκριτικά με τις υπόλοιπες μεθόδους. Απ' την άλλη, οι 30 και 50 πιο συχνές λέξεις επιτυγχάνουν σημαντικά

καλύτερα αποτελέσματα για την ομάδα Β σε σχέση με την ομάδα Α. Σε κάθε περίπτωση και για τις δύο ομάδες συγγραφέων η μέθοδός μας επιτυγχάνει τα καλύτερα αποτελέσματα.

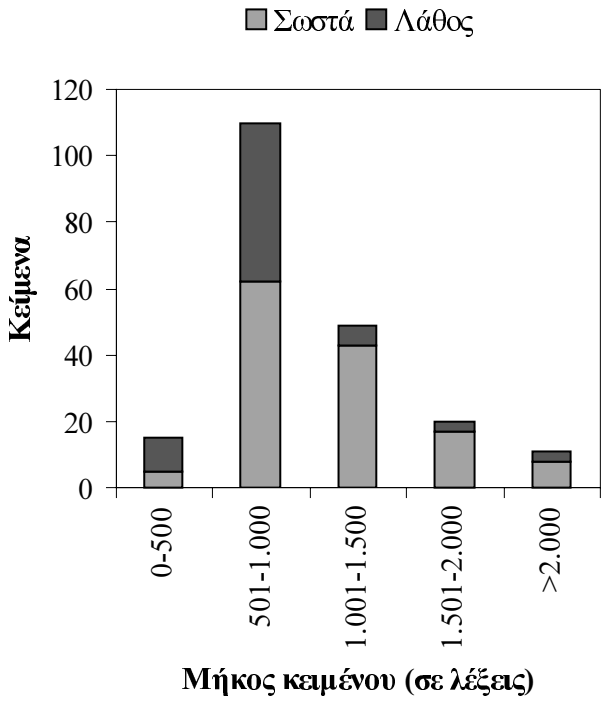
Αναλυτικά αποτελέσματα της αυτόματης ταξινόμησης που επιτεύχθηκαν στο σώμα ελέγχου με την μέθοδό μας φαίνονται στον πίνακα 6.3. Βλέπουμε ότι επιβεβαιώνεται η αρχική μας υπόθεση αφού η ακρίβεια ταξινόμησης είναι σαφώς πιο χαμηλή σε σχέση με την αναγνώριση είδους κειμένου. Όσον αφορά την ομάδα Α, υπάρχει μία σημαντική διαφορά στην απόδοση των δύο μεθόδων κατηγοριοποίησης. Επιπλέον, τρεις συγγραφείς είναι υπεύθυνοι για το 50% του σφάλματος αναγνώρισης και με τις δύο τεχνικές: Α01, Α03 και Α06. Να σημειωθεί ότι το μέσο μήκος των κειμένων αυτών των συγκεκριμένων συγγραφέων (βλ. πίνακα 6.2) είναι σχετικά μικρό (μικρότερο από 800 λέξεις).

Ομάδα Α				***Ομάδα Β***			
Κωδικός	**Σφάλμα αναγνώρισης**			**Κωδικός**	**Σφάλμα αναγνώρισης**		
Α01	*Πολλαπλή παλινδρόμηση*	0,5	*Διαχωριστική ανάλυση* 0,4	Β01	*Πολλαπλή παλινδρόμηση*	0,7	*Διαχωριστική ανάλυση* 0,6
Α02		0,3	0,2	Β02		0,0	0,0
Α03		0,6	0,5	Β03		0,2	0,4
Α04		0,2	0,1	Β04		0,1	0,1
Α05		0,3	0,3	Β05		0,7	0,4
Α06		0,7	0,5	Β06		0,3	0,3
Α07		0,3	0,3	Β07		0,0	0,1
Α08		0,1	0,1	Β08		0,6	0,6
Α09		0,2	0,3	Β09		0,1	0,1
Α10		0,2	0,1	Β10		0,4	0,4
Μέσος όρος		**0,34**	**0,28**	**Μέσος όρος**		**0,31**	**0,30**

Πίνακας 6.3. Αποτελέσματα αυτόματης αναγνώρισης συγγραφέα.

Απ' την άλλη, οι δύο τεχνικές κατηγοριοποίησης δίνουν παρόμοια αποτελέσματα για την ομάδα Β όσον αφορά τον μέσο όρο σφάλματος αναγνώρισης. Γενικά, όπως και στην αναγνώριση είδους κειμένου, το σφάλμα αναγνώρισης είναι πιο ομαλά κατανεμημένο με βάση την διαχωριστική ανάλυση. Οι συγγραφείς Β01, Β05 και Β08 είναι υπεύθυνοι για το 65% και το 55% του σφάλματος με χρήση της πολλαπλής παλινδρόμησης και της διαχωριστικής ανάλυσης αντίστοιχα. Όπως και στην περίπτωση της ομάδας Α, οι τρεις αυτοί συγγραφείς διακρίνονται για το σχετικά μικρό μέσο μήκος κειμένου (μικρότερο από 1.000 λέξεις).

Φαίνεται λοιπόν ότι στην αναγνώριση συγγραφέα το μήκος κειμένου παίζει πολύ σημαντικό ρόλο. Θυμίζουμε ότι στην αναγνώριση είδους κειμένου το μήκος κειμένου δεν επηρέασε σημαντικά τα αποτελέσματα αυτόματης ταξινόμησης (βλ. § 5.5). Στο σχήμα 6.5 φαίνεται η ακρίβεια ταξινόμησης σε σχέση με το μήκος των κειμένων και των δύο ομάδων, με χρήση της πολλαπλής παλινδρόμησης. Περίπου το 80% (53 στα 65) των συνολικών κειμένων που ταξινομήθηκαν λάθος είχαν λιγότερες από 1.000 λέξεις. Φαίνεται λοιπόν, ότι κείμενα με μήκος μεγαλύτερο από 1.000 λέξεις κατορθώνουν να αναπαραστήσουν πιο αξιόπιστα τα υφολογικά χαρακτηριστικά του συγγραφέα τους.

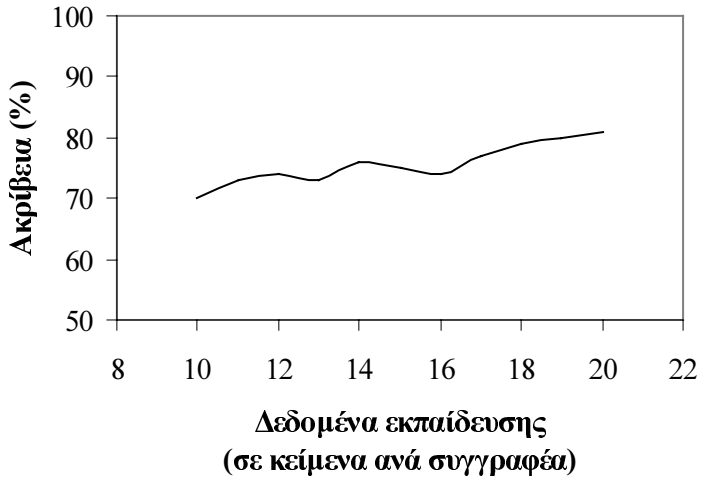


Σχήμα 6.5. Κατανομή του σώματος ελέγχου και των δύο ομάδων συναρτήσει του μήκους τους και της ακρίβειας ταξινόμησης.

6.3.1 Μέγεθος σώματος εκπαίδευσης

Για να διαπιστωθεί κατά πόσο εξαρτάται η ακρίβεια ταξινόμησης από το μέγεθος του σώματος εκπαίδευσης πραγματοποιήσαμε το ακόλουθο πείραμα: το σύστημα εκπαιδεύτηκε με βάση διαφορετικά σώματα εκπαίδευσης, που κυμαινόταν από 10 έως 20 κείμενα από κάθε συγγραφέα. Το πείραμα εφαρμόστηκε στην ομάδα Β και ως μέθοδος ταξινόμησης χρησιμοποιήθηκε η διαχωριστική ανάλυση. Το σώμα ελέγχου

ήταν σε όλες τις περιπτώσεις το ίδιο (δέκα κείμενα από κάθε συγγραφέα). Τα αποτελέσματα φαίνονται στο σχήμα 6.6.



Σχήμα 6.6. Η ακρίβεια ταξινόμησης συναρτήσει του μεγέθους σώματος εκπαίδευσης της ομάδας Β.

Κωδικός	Κατηγοριοποίηση										Σφάλμα
	B01	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B09**	**B10**	
B01	**7**	0	0	0	1	0	0	0	1	1	**0,3**
B02	0	**10**	0	0	0	0	0	0	0	0	**0,0**
B03	0	0	**8**	0	0	0	0	2	0	0	**0,2**
B04	0	0	0	**10**	0	0	0	0	0	0	**0,0**
B05	0	0	0	3	**6**	0	0	0	0	1	**0,4**
B06	0	0	0	0	0	**10**	0	0	0	0	**0,0**
B07	0	0	0	0	1	0	**9**	0	0	0	**0,1**
B08	0	0	1	0	2	2	0	**4**	1	0	**0,6**
B09	0	0	0	0	0	0	0	0	**10**	0	**0,0**
B10	0	0	2	0	1	0	0	0	0	**7**	**0,3**
									Μέσος όρος:		**0,19**

Πίνακας 6.4. Πίνακας σύγχυσης της ομάδας Β με βάση 20 κείμενα από κάθε συγγραφέα ως σώμα εκπαίδευσης.

Γενικά, η ακρίβεια ταξινόμησης αυξάνεται με την αύξηση του μεγέθους του σώματος εκπαίδευσης αν και αυτή η αύξηση δεν είναι γραμμική. Σε μερικές περιπτώσεις μάλιστα, η απόδοση του συστήματος χειροτερεύει (π.χ. με χρήση 16 κειμένων από κάθε συγγραφέα). Χρησιμοποιώντας 20 κείμενα από κάθε συγγραφέα ως σώμα εκπαίδευσης επιτυγχάνεται η καλύτερη ακρίβεια ταξινόμησης (81% ή αλλιώς σφάλμα αναγνώρισης ίσο με 19%). Ο πίνακας σύγχυσης (confusion matrix) σε αυτήν την περίπτωση δίνεται στον πίνακα 6.4.

Υπενθυμίζεται ότι η κάθε γραμμή αυτού του πίνακα αντιστοιχεί στα δέκα κείμενα του σώματος ελέγχου για ένα συγκεκριμένο συγγραφέα και η κάθε στήλη στα αποτελέσματα ταξινόμησης αυτών των κειμένων. Επομένως, η διαγώνιος περιέχει τα σωστά ταξινομημένα κείμενα. Οι συγγραφείς με χαμηλό μέσο μήκος κειμένου (Β01, Β05 και Β08) είναι υπεύθυνοι για το 65% του συνολικού σφάλματος αναγνώρισης.

6.3.2 Συνδυασμός με λεξιλογική προσέγγιση

Για την καλύτερη εκτίμηση των αποτελεσμάτων που αποκομίστηκαν από την μέθοδό μας, αποφασίσαμε να διερευνήσουμε τις ιδιότητες του συνδυασμού της προσέγγισής μας με μία λεξιλογική προσέγγιση. Για το σκοπό αυτό χρησιμοποιήσαμε τις 50 πιο συχνές λέξεις, μία προσέγγιση που στις περισσότερες περιπτώσεις έδωσε αρκετά καλά αποτελέσματα και απαιτεί ελάχιστο υπολογιστικό κόστος για την υλοποίησή της.

ακόμη	έχει	μια	που	τη
αλλά	η	μόνο	πρέπει	την
αν	ή	μπορεί	σε	της
από	ήταν	να	στα	τις
αυτή	θα	ο	στη	το
αυτό	και	οι	στην	τον
για	κατά	όμως	στις	του
δεν	κι	οποία	στο	τους
είναι	μας	όπως	στον	των
ένα	με	ότι	τα	ως

Πίνακας 6.5. Οι 50 πιο συχνά εμφανιζόμενες λέξεις του σώματος εκπαίδευσης της ομάδας Β.

Ως πεδίο σύγκρισης επιλέξαμε την ομάδα Β και την χρήση 20 κειμένων από κάθε συγγραφέα ως σώμα εκπαίδευσης. Τα σώματα εκπαίδευσης και ελέγχου που χρησιμοποιήθηκαν είναι τα ίδια με αυτά του τμήματος 6.3.1. Αρχικά, βρέθηκε η λίστα των 50 πιο συχνά εμφανιζόμενων λέξεων στο σώμα εκπαίδευσης, η οποία φαίνεται στον πίνακα 6.5 με αλφαβητική σειρά. Σε αυτήν την λίστα δεν υπάρχουν κύρια ονόματα. Στην συνέχεια, μετρήθηκαν οι συχνότητες εμφάνισης των λέξεων αυτών σε κάθε κείμενο του σώματος εκπαίδευσης και στα διανύσματα των 50 υφολογικών δεικτών που προέκυψαν εφαρμόσαμε διαχωριστική ανάλυση. Να σημειωθεί ότι για να κανονικοποιηθούν οι υφολογικοί δείκτες, οι συχνότητες

εμφάνισης των λέξεων αυτών σε ένα κείμενο διαιρέθηκαν προς το συνολικό μήκος του κειμένου. Τέλος, εφαρμόσαμε τη διαδικασία αυτόματης ταξινόμησης στο σώμα ελέγχου (πάντα με βάση τις συχνότητες εμφάνισης των 50 πιο συχνά εμφανιζόμενων λέξεων σε κάθε κείμενο). Τα αποτελέσματα φαίνονται στον πίνακα 6.6.

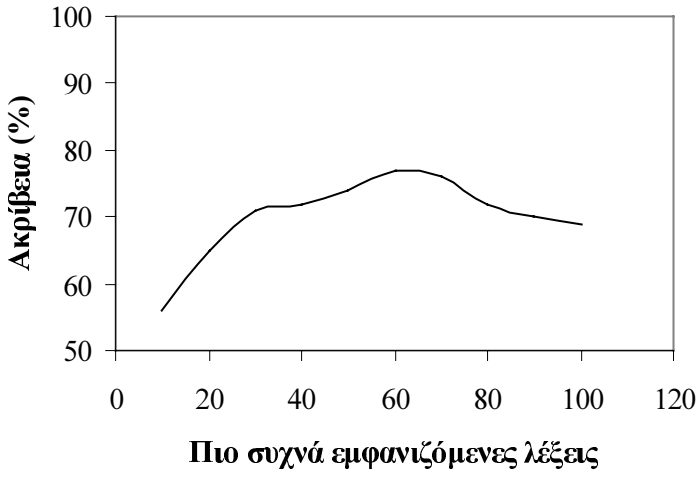
Κωδικός	**Κατηγοριοποίηση**										**Σφάλμα**
	B01	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B09**	**B10**	
B01	**5**	1	0	1	2	0	0	1	0	0	**0,5**
B02	0	**10**	0	0	0	0	0	0	0	0	**0,0**
B03	0	0	**7**	2	1	0	0	0	0	0	**0,3**
B04	0	0	0	**10**	0	0	0	0	0	0	**0,0**
B05	0	0	0	1	**5**	0	1	1	0	2	**0,5**
B06	0	0	0	0	0	**9**	1	0	0	0	**0,1**
B07	0	0	0	1	1	0	**8**	0	0	0	**0,2**
B08	1	0	2	2	0	0	1	**3**	0	1	**0,7**
B09	0	0	0	0	0	0	0	0	**9**	1	**0,1**
B10	0	1	1	0	0	0	0	0	0	**8**	**0,2**
									Μέσος όρος:		**0,26**

Πίνακας 6.6. Πίνακας σύγχυσης της ομάδας Β βάσει της λεξιλογικής προσέγγισης.

Παρατηρούμε ότι η λεξιλογική προσέγγιση επιτυγχάνει ακρίβεια ταξινόμησης της τάξης του 74% (ή αλλιώς 0,26 σφάλμα αναγνώρισης) πράγμα που σημαίνει ότι υστερεί κατά 7% σε σχέση με τα αντίστοιχα αποτελεσμάτων της δικής μας μεθόδου (βλ. πίνακα 6.4). Επίσης, αξίζει να σημειωθεί ότι το 65% του σφάλματος αναγνώρισης οφείλεται στους συγγραφείς B01, B05 και B08 που διακρίνονται για το χαμηλό μέσο μήκος κειμένου τους. Φαίνεται λοιπόν, ότι και η λεξιλογική προσέγγιση εξαρτάται από το μήκος του κειμένου.

Οι 50 πιο συχνά εμφανιζόμενες λέξεις στο σώμα εκπαίδευσης αντιστοιχούν περίπου στο 40% των συνολικών λέξεων στο σώμα αυτό. Ακόμη, οι 100 πιο συχνά εμφανιζόμενες λέξεις αντιστοιχούν περίπου στο 45% των συνολικών λέξεων του σώματος εκπαίδευσης. Για να διαπιστωθεί κατά πόσο εξαρτάται η ακρίβεια ταξινόμησης της λεξιλογικής προσέγγισης από τον αριθμό των πιο συχνά εμφανιζόμενων λέξεων στο σώμα εκπαίδευσης, πραγματοποιήσαμε το ίδιο πείραμα για σύνολα λέξεων που κυμαινόταν από τις 10 έως τις 100 πιο συχνά εμφανιζόμενες λέξεις. Τα αποτελέσματα αυτού του πειράματος φαίνονται στο σχήμα 6.7. Παρατηρούμε ότι η υψηλότερη ακρίβεια επιτυγχάνεται χρησιμοποιώντας τις 60 πιο συχνά εμφανιζόμενες λέξεις (77%), η οποία όμως συνεχίζει να είναι χαμηλότερη σε σχέση με τη δική μας μέθοδο (81%). Επίσης, ενώ στο διάστημα από 10 έως 60 λέξεις

η ακρίβεια βελτιώνεται αισθητά, στην συνέχεια, στο διάστημα από 60 έως 100 λέξεις η ακρίβεια φαίνεται να πέφτει. Επομένως, η χρησιμοποίηση περισσότερων λέξεων (άρα και περισσότερων υφολογικών δεικτών) πέρα από κάποιο όριο δεν συνεπάγεται ότι βελτιώνει την ακρίβεια ταξινόμησης. Το αντίθετο μάλιστα. Κατά την γνώμη μας αυτό συμβαίνει επειδή οι λέξεις που είναι πιο σημαντικές για την διάκριση μεταξύ υφολογικών κατηγοριών είναι οι πιο συχνά εμφανιζόμενες αλλά και οι πιο σπάνια εμφανιζόμενες. Επομένως, η χρησιμοποίηση των συχνοτήτων εμφάνισης όλο και περισσότερων συχνά εμφανιζόμενων λέξεων, πέρα από κάποιο όριο, δεν βελτιώνει την απόδοση ενός συστήματος αυτόματης ταξινόμησης. Απ' την άλλη, η χρησιμοποίηση των πιο σπάνια εμφανιζόμενων λέξεων δεν συνιστάται αφού τέτοιες λέξεις είναι στενά συνδεδεμένες με το θεματικό περιεχόμενο του κειμένου.



Σχήμα 6.7. Ακρίβεια ταξινόμησης της λεξιλογικής προσέγγισης συναρτήσει του αριθμού των πιο συχνά εμφανιζόμενων λέξεων.

Ωστόσο, οι δύο προσεγγίσεις μπορούν εύκολα να συνδυαστούν. Σε αυτήν την περίπτωση χρησιμοποιούμε 72 συνολικά υφολογικούς δείκτες: 22 από την δική μας προσέγγιση και 50 από την λεξιλογική προσέγγιση. Εφαρμόσαμε την διαχωριστική ανάλυση και σε αυτήν την περίπτωση, με βάση σώμα εκπαίδευσης αποτελούμενο από 20 κείμενα από κάθε συγγραφέα της ομάδας Β. Η διαδικασία αυτόματης ταξινόμησης εφαρμόστηκε στο σώμα ελέγχου και τα αποτελέσματα δίνονται στον πίνακα 6.7. Παρατηρούμε ότι ο συνδυασμός των δύο μεθόδων επιτυγχάνει μεγάλη ακρίβεια ταξινόμησης (της τάξης του 87%). Το ποσοστό του σφάλματος αναγνώρισης που οφείλεται στους συγγραφείς που διακρίνονται για το χαμηλό μέσο μήκος κειμένου

τους (B01, B05 και B08) είναι 85%, δηλ. σημαντικά αυξημένο σε σχέση με το αντίστοιχο ποσοστό της κάθε μεθόδου.

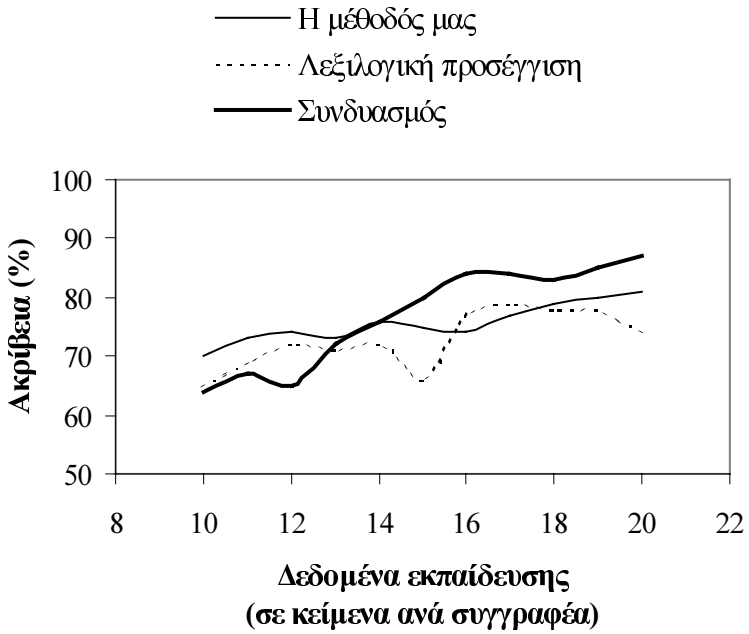
Κωδικός	**Κατηγοριοποίηση**										**Σφάλμα**
	B01	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B09**	**B10**	
B01	**7**	0	2	1	0	0	0	0	0	0	**0,3**
B02	0	**10**	0	0	0	0	0	0	0	0	**0,0**
B03	0	0	**10**	0	0	0	0	0	0	0	**0,0**
B04	0	0	0	**10**	0	0	0	0	0	0	**0,0**
B05	0	0	0	3	**6**	0	0	1	0	0	**0,4**
B06	0	1	0	0	0	**9**	0	0	0	0	**0,1**
B07	0	0	0	0	0	0	**10**	0	0	0	**0,0**
B08	1	0	0	1	0	0	1	**6**	1	0	**0,4**
B09	0	0	0	0	0	0	0	0	**10**	0	**0,0**
B10	0	0	0	0	0	0	0	1	0	**9**	**0,1**
										Μέσος όρος:	**0,13**

Πίνακας 6.7. Πίνακας σύγχυσης της ομάδας Β βάσει του συνδυασμού της δικής μας μεθόδου και της λεξιλογικής προσέγγισης.

Στο τμήμα 6.3.1 εξετάστηκε η σχέση ακρίβειας ταξινόμησης – μεγέθους σώματος εκπαίδευσης. Πώς επηρεάζει όμως το μέγεθος του σώματος εκπαίδευσης την λεξιλογική προσέγγιση αλλά και τον συνδυασμό των δύο προσεγγίσεων; Για να απαντήσουμε σε αυτήν την ερώτηση εκπαιδεύσαμε τόσο την λεξιλογική προσέγγιση όσο και τον συνδυασμό των δύο μεθόδων με βάση διαφορετικά σώματα εκπαίδευσης, αποτελούμενα από 100 έως 200 συνολικά κείμενα (ή αλλιώς από 10 έως 20 κείμενα από κάθε συγγραφέα). Τα αποτελέσματα δίνονται στο σχήμα 6.8. Στο σχήμα αυτό φαίνεται και η αντίστοιχη καμπύλη της δικής μας μεθόδου (βλ. σχήμα 6.6) για συγκριτικούς λόγους. Τόσο τα σώματα εκπαίδευσης όσο και το σώμα ελέγχου ήταν σε όλες τις περιπτώσεις τα ίδια.

Παρατηρούμε ότι σε όλες τις μεθόδους η ακρίβεια γενικά βελτιώνεται με την αύξηση του σώματος εκπαίδευσης. Παρ' όλα αυτά η βελτίωση αυτή δεν είναι καθόλου ομαλή με βάση την λεξιλογική προσέγγιση. Αντίθετα, η δική μας μέθοδος παρουσιάζει τις μικρότερες διακυμάνσεις. Επίσης, η λεξιλογική προσέγγιση συμπεριφέρεται χειρότερα από την δική μας σε όλο το φάσμα του σώματος εκπαίδευσης, εκτός από την περίπτωση χρήσης 16 και 17 κειμένων από κάθε συγγραφέα. Ο συνδυασμός των δύο προσεγγίσεων επιτυγχάνει αρκετά χαμηλή ακρίβεια για σώμα εκπαίδευσης με λιγότερα από 14 κείμενα από κάθε συγγραφέα. Ωστόσο, όταν αυξάνεται αρκετά το μέγεθος του σώματος εκπαίδευσης τα αποτελέσματα του συνδυασμού των δύο

προσεγγίσεων βελτιώνονται θεαματικά. Αξίζει να σημειωθεί ότι για σώμα εκπαίδευσης με 10 κείμενα από κάθε συγγραφέα, η δική μας προσέγγιση είναι η πιο αξιόπιστη. Αυτό το γεγονός αποκτά μεγάλη σημασία αν αναλογιστεί κανείς ότι σε πολλές περιπτώσεις ο αριθμός των διαθέσιμων προς εκπαίδευση κειμένων είναι πολύ περιορισμένος.



Σχήμα 6.8. Συγκριτικά αποτελέσματα ακρίβειας ταξινόμησης συναρτήσει του μεγέθους σώματος εκπαίδευσης.

6.3.3 Σημαντικότητα υφολογικών δεικτών

Για την διερεύνηση της σημαντικότητας του κάθε υφολογικού δείκτη χρησιμοποιήθηκαν, όπως και στην περίπτωση της αναγνώρισης είδους κειμένου (βλ. § 5.5.2), οι απόλυτες τιμές του στατιστικού-*t* των συντελεστών παλινδρόμησης. Ο πίνακας 6.8 περιέχει τις μέσες τιμές του στατιστικού-*t* κάθε υφολογικού δείκτη για το σύνολο των ομάδων Α και Β. Η σειρά σπουδαιότητας των υφολογικών επιπέδων παραμένει η ίδια με αυτήν του πειράματος αναγνώρισης είδους κειμένου. Ωστόσο, παρατηρείται μία αύξηση της σπουδαιότητας των υφολογικών δεικτών του επιπέδου δείγματος και του επιπέδου φράσης και μείωση της αντίστοιχης σπουδαιότητας του επιπέδου ανάλυσης (λαμβάνοντας υπ' όψιν τους μέσους όρους του κάθε επιπέδου).

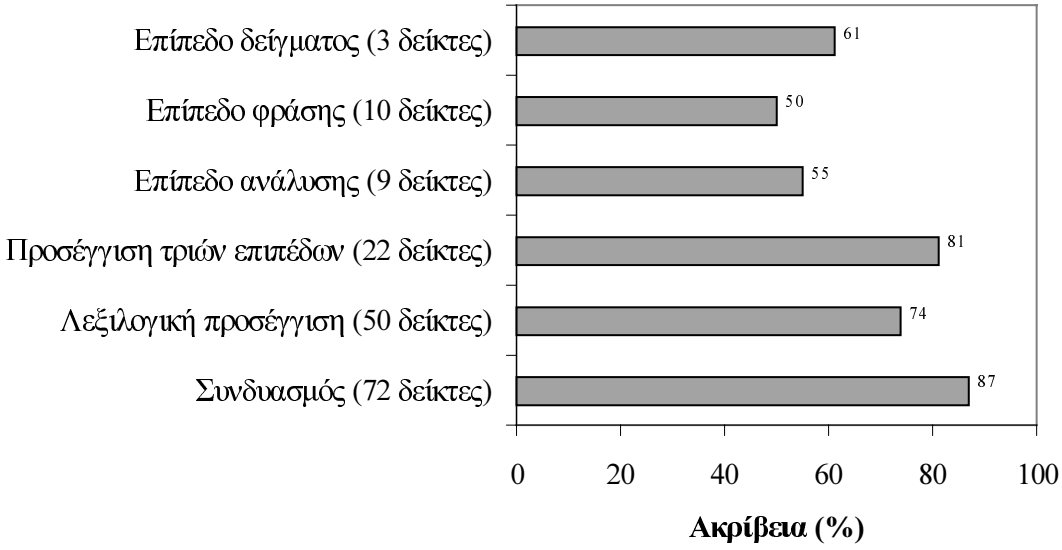
Ο πιο σημαντικός δείκτης είναι ο Δ03 (περίοδοι προς πιθανά όρια περιόδων). Εκτός από τους άλλους δύο υφολογικούς δείκτες επιπέδου δείγματος, πολύ σημαντικοί είναι οι Δ14 (λέξεις-κλειδιά προς λέξεις) και Δ17 (μορφολογικές περιγραφές φράσεων προς συνολικές φράσεις). Επίσης, αξίζει να σημειωθεί ότι, όπως και στην περίπτωση της αναγνώρισης είδους κειμένου, οι υφολογικοί δείκτες που σχετίζονται με τον αριθμό των λέξεων που περικλείονται σε κάθε είδος φράσης (Δ09-Δ13) είναι πιο σημαντικοί από τις συχνότητες εμφάνισης των ειδών των φράσεων (Δ04-Δ08).

Επίπεδο	**Δείκτης ύφους**	**Απόλυτη τιμή *t***	**Μέσος όρος**
Δείγματος	Δ01	1,80	1,88
	Δ02	1,85	
	Δ03	1,98	
Φράσεων	Δ04	0,76	0,86
	Δ05	0,77	
	Δ06	0,77	
	Δ07	0,75	
	Δ08	0,76	
	Δ09	0,98	
	Δ10	0,85	
	Δ11	0,90	
	Δ12	1,07	
	Δ13	0,97	
Ανάλυσης	Δ14	1,30	1,01
	Δ15	1,05	
	Δ16	0,79	
	Δ17	1,42	
	Δ18	1,06	
	Δ19	0,84	
	Δ20	0,86	
	Δ21	0,90	
	Δ22	0,84	

Πίνακας 6.8. Μέσες τιμές *t* των συντελεστών παλινδρόμησης συνολικά για τις ομάδες Α και Β.

Με βάση αυτά τα αποτελέσματα μπορούμε να υποθέσουμε ότι ένα σύστημα που θα βασιζόταν αποκλειστικά σε ένα υφομετρικό επίπεδο θα είχε χειρότερα αποτελέσματα από τον συνδυασμό των τριών επιπέδων και πως τα καλύτερα αποτελέσματα θα αποκομιζόταν από το επίπεδο δείγματος. Για να επιβεβαιώσουμε αυτήν την υπόθεση εφαρμόσαμε τρεις φορές διαχωριστική ανάλυση στο σώμα εκπαίδευσης της ομάδας Β (20 κείμενα από κάθε συγγραφέα) χρησιμοποιώντας κάθε φορά τις παραμέτρους ενός μόνο υφολογικού επιπέδου. Στην συνέχεια, εφαρμόσαμε την διαδικασία ταξινόμησης στο σώμα ελέγχου (που ήταν πάντα το ίδιο) και τα αποτελέσματα φαίνονται στο σχήμα 6.9. Στο ίδιο σχήμα δίνονται, για συγκριτικούς λόγους, τα αντίστοιχα

αποτελέσματα της προσέγγισης τριών επιπέδων, της λεξιλογικής προσέγγισης καθώς και του συνδυασμού των δύο μεθόδων. Βλέπουμε ότι η αρχική μας υπόθεση επιβεβαιώνεται και στην πράξη. Η σειρά σημαντικότητας των τριών επιπέδων αντικατοπτρίζεται στην απόδοσή τους στο σώμα ελέγχου.



Σχήμα 6.9. Ακρίβεια ταξινόμησης για κάθε επίπεδο ξεχωριστά.

6.4 Επιβεβαίωση Συγγραφέα

Στα πειράματα αναγνώρισης συγγραφέα, έχοντας ως δεδομένο ένα σύνολο συγγραφέων (ομάδα Α ή Β), προσπαθήσαμε να απαντήσουμε στο ερώτημα ποιος απ' αυτούς είναι πιο πιθανό να έχει γράψει το κείμενο. Ωστόσο, πολλές εφαρμογές απαιτούν την επιβεβαίωση της υπόθεσης ότι ένα συγκεκριμένο άτομο είναι (ή δεν είναι) ο συγγραφέας ενός κειμένου. Σε μια τέτοια περίπτωση η διαδικασία ταξινόμησης είναι πιο απλή μιας και οι πιθανές απαντήσεις είναι μόνο δύο: *ναι*, το άτομο αυτό είναι ο συγγραφέας του κειμένου ή *όχι*, το κείμενο δεν γράφτηκε από αυτόν τον συγγραφέα.

Η υλοποίηση ενός υπολογιστικού συστήματος αυτόματης επιβεβαίωσης συγγραφέα έχει δύο βασικές απαιτήσεις:

- Την εύρεση μιας **συνάρτησης απόκρισης** για ένα συγκεκριμένο συγγραφέα. Αυτή η συνάρτηση θα πρέπει, με βάση την τιμή των υφολογικών δεικτών, να επιστρέφει μία τιμή για κάθε κείμενο.
- Τον ορισμό ενός **κατωφλιού**. Κάθε κείμενο που έχει τιμή της συνάρτησης απόκρισης μεγαλύτερη από το κατώφλι γίνεται αποδεκτό ως κείμενο του συγγραφέα. Σε αντίθετη περίπτωση απορρίπτεται.

Επιπλέον, για την αντικειμενική μέτρηση της ακρίβειας μιας μεθόδου επιβεβαίωσης συγγραφέα ορίζουμε το *σφάλμα απόρριψης* (false rejection) και το *σφάλμα αποδοχής* (false acceptance) ως ακολούθως:

Σφάλμα απόρριψης = *κείμενα του συγγραφέα που απορρίφθηκαν προς συνολικά κείμενα του συγγραφέα*

Σφάλμα αποδοχής = *κείμενα άλλων συγγραφέων που έγιναν αποδεκτά προς συνολικά κείμενα των άλλων συγγραφέων*

Αυτές οι παράμετροι χρησιμοποιούνται ευρέως στην επεξεργασία ομιλίας και ειδικότερα κατά την επιβεβαίωση ομιλητή (speaker verification) [35], μία εφαρμογή που έχει πολλές κοινές ιδιότητες με την επιβεβαίωση συγγραφέα.

Στην παρούσα εργασία θεωρήσαμε ως συνάρτηση απόκρισης, για τον κάθε συγγραφέα των ομάδων Α και Β, τη συνάρτηση πολλαπλής παλινδρόμησης που έγινε διαθέσιμη μετά την εκπαίδευση του συστήματος αναγνώρισης συγγραφέα για τις δύο αυτές ομάδες. Όσον αφορά το κατώφλι, θεωρήθηκε ότι η καλύτερη λύση είναι να εκφραστεί ως συνάρτηση του συντελεστή πολλαπλής συσχέτισης R (βλ. § 5.2.1). Θυμίζουμε ότι ο συντελεστής πολλαπλής συσχέτισης ισούται με το 1 όταν η συνάρτηση πολλαπλής παλινδρόμησης περνάει από όλα τα σημεία του σώματος εκπαίδευσης (με άλλα λόγια, όταν υπάρχει μηδενικό σφάλμα) [48]. Απ' την άλλη, ισούται με μηδέν όταν οι ανεξάρτητες μεταβλητές (δηλ. οι υφολογικοί δείκτες) δεν έχουν καμία επίδραση στην τιμή απόκρισης. Επομένως η τιμή αυτού του συντελεστή είναι μία πολύ καλή ένδειξη σχετικά με την αξιοπιστία της συνάρτησης απόκρισης του κάθε συγγραφέα. Ως κατώφλι λοιπόν, ορίσαμε το $R/2$.

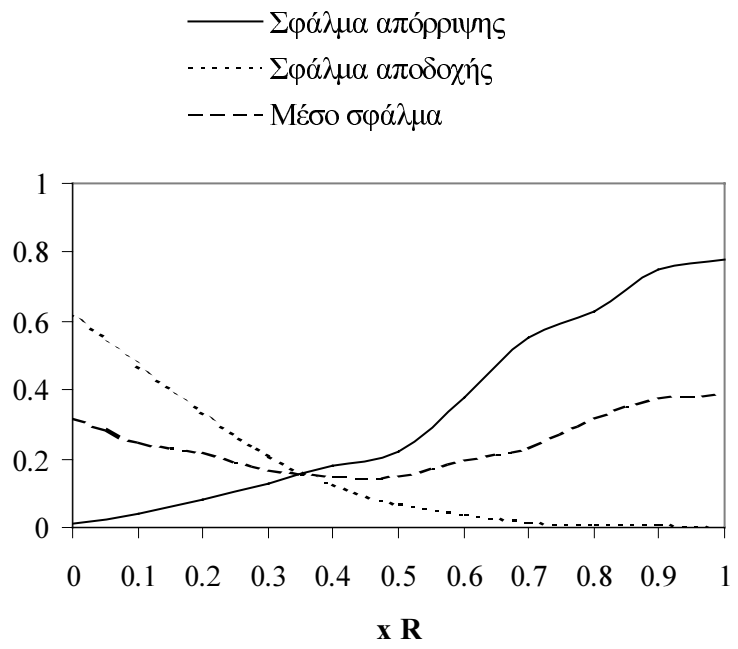
Η προσέγγιση αυτή εφαρμόστηκε στους συγγραφείς των ομάδων Α και Β. Οι τιμές των σφαλμάτων απόρριψης και αποδοχής που αποκομίστηκαν δίνονται στον πίνακα

6.9. Να σημειωθεί ότι ο υπολογισμός του σφάλματος αποδοχής έγινε λαμβάνοντας υπ' όψιν τους συγγραφείς της ίδιας ομάδας μόνο (αξιολόγηση κλειστού συνόλου). Παρατηρούμε ότι υπάρχει σημαντική διαφορά στην τάξη μεγέθους του σφάλματος απόρριψης σε σχέση με το σφάλμα αποδοχής, πράγμα που σημαίνει ότι γενικά, κατά την διαδικασία επιβεβαίωσης ενός συγγραφέα, πολύ λίγα κείμενα άλλων συγγραφέων γίνονται εσφαλμένα αποδεκτά.

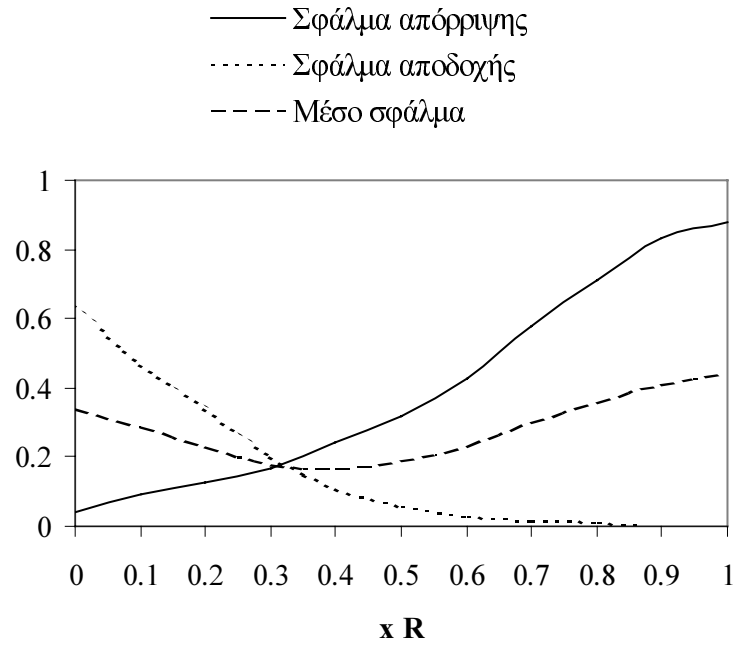
	Κωδικός	**R/2**	**Σφάλμα απόρριψης**	**Σφάλμα αποδοχής**
Ομάδα Α	A01	0,33	0,500	0,033
	A02	0,33	0,300	0,011
	A03	0,36	0,600	0,044
	A04	0,36	0,200	0,111
	A05	0,35	0,300	0,067
	A06	0,35	0,700	0,044
	A07	0,34	0,200	0,044
	A08	0,31	0,100	0,111
	A09	0,35	0,200	0,055
	A10	0,35	0,100	0,089
	Μέσος όρος	**0,35**	**0,320**	**0,061**
Ομάδα Β	B01	0,32	0,300	0,022
	B02	0,42	0,000	0,044
	B03	0,33	0,000	0,155
	B04	0,33	0,100	0,089
	B05	0,28	0,600	0,144
	B06	0,36	0,200	0,011
	B07	0,38	0,000	0,022
	B08	0,30	0,600	0,100
	B09	0,36	0,000	0,055
	B10	0,40	0,400	0,033
	Μέσος όρος	**0,35**	**0,220**	**0,068**

Πίνακας 6.9. Αποτελέσματα επιβεβαίωσης συγγραφέα για τις ομάδες Α και Β (κατώφλι=*R/2*).

Επίσης, μεταξύ των συγγραφέων της ομάδας Α δεν υπάρχουν σημαντικές διαφορές στην τιμή του *R/2*. Αντίθετα, στην ομάδα Β το *R/2* κυμαίνεται μεταξύ 0,28 και 0,42. Το μεγαλύτερο ποσοστό του μέσου σφάλματος απόρριψης οφείλεται στους συγγραφείς που διακρίνονται για το χαμηλό μήκος κειμένου τους, δηλ. από την ομάδα Α οι A01, A03 και A06 ενώ από την ομάδα Β οι B01, B05 και B08. Απ' την άλλη, το σφάλμα αποδοχής φαίνεται να εξαρτάται από την τιμή του κατωφλιού. Όσο μικρότερο το κατώφλι, τόσο μεγαλύτερο το σφάλμα αποδοχής.



Σχήμα 6.10. Μέσος όρος σφάλματος απόρριψης, σφάλματος αποδοχής και μέσου σφάλματος συναρτήσει του κατωφλιού για την ομάδα Α.



Σχήμα 6.11. Μέσος όρος σφάλματος απόρριψης, σφάλματος αποδοχής και μέσου σφάλματος συναρτήσει του κατωφλιού για την ομάδα Β.

Εδώ πρέπει να τονιστεί ότι ο καθορισμός του κατωφλιού εξαρτάται άμεσα από την εφαρμογή. Υπάρχουν εφαρμογές που απαιτούν ελάχιστο σφάλμα απόρριψης αδιαφορώντας για το σφάλμα αποδοχής, ενώ άλλες απαιτούν ελάχιστο σφάλμα αποδοχής αδιαφορώντας για το σφάλμα απόρριψης. Σε γενικές γραμμές, η ελαχιστοποίηση του *μέσου σφάλματος* (ο μέσος όρος των σφαλμάτων απόρριψης και αποδοχής) ικανοποιεί την πλειοψηφία των εφαρμογών. Στα σχήματα 6.10 και 6.11 φαίνεται η διακύμανση του μέσου όρου των τιμών του σφάλματος απόρριψης, του σφάλματος αποδοχής και του μέσου σφάλματος συναρτήσει των τιμών του κατωφλιού, για τις ομάδες Α και Β αντίστοιχα. Παρατηρούμε χαμηλές τιμές του κατωφλιού αντιστοιχούν σε ελάχιστο σφάλμα απόρριψης και μέγιστο σφάλμα αποδοχής ενώ υψηλές τιμές του κατωφλιού αντιστοιχούν σε μέγιστο σφάλμα απόρριψης και ελάχιστο σφάλμα αποδοχής. Και στις δύο ομάδες το ελάχιστο μέσο σφάλμα επιτυγχάνεται για κατώφλι περίπου ίσο με *R/2*.

6.5 Περίληψη - Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε ένα αυτοματοποιημένο σύστημα προσδιορισμού συγγραφέα. Σε αντίθεση με τις σύγχρονες μεθόδους, οι υφολογικές παράμετροι δεν βασίζονται σε λεξιλογική πληροφορία αλλά στο σύνολο των υφολογικών δεικτών που παρουσιάστηκε στο κεφάλαιο 4. Τα αποτελέσματα που επιτεύχθηκαν στα πειράματα αναγνώρισης συγγραφέα κρίνονται πολύ ικανοποιητικά αφού ξεπερνούν την ακρίβεια που επιτυγχάνει η πιο σύγχρονη και αξιόπιστη λεξιλογική προσέγγιση. Ωστόσο, η μεγαλύτερη ακρίβεια επιτυγχάνεται με το συνδυασμό των δύο μεθόδων.

Ως πεδίο ελέγχου της προτεινόμενης μεθόδου επιλέξαμε δύο ομάδες συγγραφέων μιας εβδομαδιαίας εφημερίδας. Η ομάδα Α αποτελείται από συγγραφείς που μπορεί να υπογράφουν κείμενα από διαφορετικά είδη κειμένων (π.χ. άρθρα και ρεπορτάζ), ενώ τα κείμενα της ομάδας Β είναι πιο ομοιογενή. Τα αποτελέσματα ταξινόμησης είναι και στις δύο ομάδες συγκρίσιμα. Φαίνεται λοιπόν, ότι το σύνολο των υφολογικών δεικτών καταφέρνει να διακρίνει τα προσωπικά χαρακτηριστικά ενός συγγραφέα ακόμα και όταν τα εν λόγω κείμενα ανήκουν σε διαφορετικά είδη. Να σημειωθεί ότι παρατηρήθηκε διαφορά στην απόδοση των 30 και 50 πιο συχνών λέξεων για την ομάδα Α και Β που πιθανώς οφείλεται σ' αυτό το γεγονός (βλ. σχήματα 6.3 και 6.4).

Ένας πολύ σημαντικός παράγοντας που παίζει ρόλο στην αξιοπιστία του συστήματος είναι το μήκος του κειμένου. Πιο συγκεκριμένα, θέτοντας ως κατώτατο όριο τις 1.000 λέξεις ανά κείμενο εξασφαλίζουμε πολύ υψηλά ποσοστά ακρίβειας ταξινόμησης. Όμως, η θέσπιση ενός τέτοιου ορίου δεν συμβαδίζει με την προδιαγραφή περί δυνατότητας επεξεργασίας κειμένου χωρίς περιορισμούς (βλ. § 1.4). Πρέπει πάντως να σημειωθεί ότι και η λεξιλογική προσέγγιση παρουσίασε άσχημα αποτελέσματα στην αναγνώριση συγγραφέων με χαμηλό μέσο μήκος κειμένου. Επίσης, είναι δεδομένη η αδυναμία άλλων λεξιλογικών μεθόδων, όπως οι συναρτήσεις πλούτου του λεξιλογίου, να αντιμετωπίσουν αυτό το πρόβλημα.

Η ακρίβεια ταξινόμησης φαίνεται να βελτιώνεται σημαντικά με την αύξηση του σώματος εκπαίδευσης. Η βελτίωση αυτή είναι ιδιαίτερα αισθητή στο σύστημα που συνδυάζει την δική μας μέθοδο με την λεξιλογική προσέγγιση. Συγκριτικά με την λεξιλογική προσέγγιση αλλά και με τον συνδυασμό των δύο μεθόδων, η πρότασή μας παρουσιάζει την πιο ομαλή βελτίωση. Επίσης, η μέθοδός μας υπερτερεί των άλλων όταν χρησιμοποιούνται σχετικά λίγα κείμενα από κάθε συγγραφέα για εκπαίδευση (γύρω στα 10), πράγμα που ισχύει στις περισσότερες περιπτώσεις.

Ο έλεγχος της σημαντικότητας των υφολογικών δεικτών έδειξε ότι η σειρά σπουδαιότητας των υφομετρικών επιπέδων είναι ίδια με αυτήν της αναγνώρισης είδους κειμένου. Ωστόσο, η διαφορά του επιπέδου δείγματος από τα άλλα δύο αυξήθηκε. Όσον αφορά τους επιμέρους δείκτες, εκτός από τις παραμέτρους του επιπέδου δείγματος, πολύ σημαντικό ρόλο στην αναγνώριση συγγραφέα παίζει η συχνότητα των λέξεων κλειδιών και η συχνότητα των μορφολογικών περιγραφών των φράσεων που έχουν ανιχνευθεί.

Ακόμη, παρουσιάστηκε μία προσέγγιση αυτόματης επιβεβαίωσης συγγραφέα που βασίζεται στην εκπαίδευση του συστήματος αναγνώρισης συγγραφέα. Τα πειράματα στις ομάδες συγγραφέων Α και Β έδειξαν ότι για κατώφλι ίσο με *R/2,* που αντιστοιχεί στο ελάχιστο μέσο σφάλμα, το σφάλμα απόρριψης είναι κατά πολύ μεγαλύτερο του σφάλματος αποδοχής. Με άλλα λόγια, κατά την επιβεβαίωση ενός συγγραφέα πολύ λίγα κείμενα άλλων συγγραφέων γίνονται αποδεκτά. Ασφαλώς, το πιο σημαντικό θέμα στην επιβεβαίωση συγγραφέα είναι η επιλογή του κατωφλιού. Επιλέξαμε να εκφράσουμε το κατώφλι ως συνάρτηση του συντελεστή πολλαπλής συσχέτισης, ενός δείκτη αξιοπιστίας των συναρτήσεων πολλαπλής παλινδρόμησης. Ωστόσο, η επιλογή

ενός συγκεκριμένου κατωφλιού εξαρτάται άμεσα από τις ανάγκες της εκάστοτε εφαρμογής.