ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΑΧΩΡΙΣΜΟΣ ΕΙΔΩΝ ΛΟΓΟΥ ΚΑΙ ΟΜΙΛΗΤΩΝ ΑΠΟ ΓΡΑΠΤΑ ΚΕΙΜΕΝΑ

Χαιρετάκης Νικόλαος

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ :

Γ. Καραγιάννης

ΑΘΗΝΑ 2000

ΔΙΑΧΩΡΙΣΜΟΣ ΕΙΔΩΝ ΛΟΓΟΥ ΚΑΙ ΟΜΙΛΗΤΩΝ ΑΠΟ ΓΡΑΠΤΑ ΚΕΙΜΕΝΑ

Διπλωματική Εργασία

Οκτώβριος 2000

Χαιρετάκης Νικόλαος

Εθνικό Μετσόβιο Πολυτεχνείο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ινστιτούτο Επεξεργασίας του Λόγου

ΔΙΑΧΩΡΙΣΜΟΣ ΕΙΔΩΝ ΛΟΓΟΥ ΚΑΙ ΟΜΙΛΗΤΩΝ ΑΠΟ ΓΡΑΠΤΑ ΚΕΙΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ

Πρόλογος ..... ***vi***

ΚΕΦΑΛΑΙΟ 1 : ΓΕΝΙΚΑ ***1***

1.1 Εισαγωγή 1
1.2 Αναζήτηση πληροφοριών (*Information Retrieval*) 1
1.3 Ποιο είναι το πρόβλημα 3
1.4 Χαρακτηριστικά Ελληνικής Γλώσσας 5
1.4.1 Πολυτυπία 5
1.4.2 Αρνήσεις 7
1.5 Ανάλογες προσπάθειες και αναφορές σε επιστημονικές εργασίες. 7
1.6 Χαρακτηριστικά Corpus 10

ΚΕΦΑΛΑΙΟ 2 : ΕΡΓΑΛΕΙΑ ***13***

2.1 Εισαγωγή 13
2.2 AMP (*Automated Morphological Processor*) 13
2.3 Άλλα μετροπρογράμματα 15
2.4 Tagger – Lemmatizer (*Λημματοποιητής*) 16
2.5 Στατιστικά Προγράμματα 16
2.6 SOM Toolbox 18

ΚΕΦΑΛΑΙΟ 3 : ΔΙΑΧΩΡΙΣΜΟΣ ΕΙΔΩΝ ΛΟΓΟΥ ***23***

3.1 Εισαγωγή 23
3.2 Κείμενα 23
3.3 Ρηματική Πολυτυπία 23
3.4 Μετρήσεις 25
3.4.1 Μετρήσεις Μορφολογικών Χαρακτηριστικών 25
3.4.2 Μετρήσεις Δομικών Χαρακτηριστικών 26
3.4.3 Μέρη του Λόγου 27
3.5 Αποτελέσματα 28
3.6 Επεξεργασία 32
3.6.1 Στατιστική Επεξεργασία 32
3.6.2 Νευρωνικά Δίκτυα 36
3.7 Συμπεράσματα 37

ΚΕΦΑΛΑΙΟ 4 : ΔΙΑΧΩΡΙΣΜΟΣ ΟΜΙΛΗΤΩΝ..... 42

4.1 Εισαγωγή 42
4.2 Κείμενα 42
4.3 Μετρήσεις 43
4.3.1 Μέτρηση Μορφολογικών Χαρακτηριστικών 43
4.3.2 Μέτρηση Δομικών Χαρακτηριστικών 45
4.3.3 Μέρη του Λόγου 46
4.3.4 Μέτρηση Λεκτικών Χαρακτηριστικών 46
4.3.5 Άλλες Μετρήσεις 47
4.4 Αποτελέσματα 47
4.5 Επεξεργασία 56
4.5.1 Στατιστική Επεξεργασία 56
4.5.2 Νευρωνικά Δίκτυα 62
4.6 Συμπεράσματα 66

ΚΕΦΑΛΑΙΟ 5 : ΣΥΜΠΕΡΑΣΜΑΤΑ..... 68

5.1 Εισαγωγή 68
5.2 Συμπεράσματα 68
5.3 Προτάσεις για βελτιώσεις 69

Παράρτημα I..... 72

Παράρτημα II 80

Παράρτημα III..... 89

Παράρτημα IV 93

Παράρτημα V..... 94

Παράρτημα VI 95

Ευρετήριο 96

Βιβλιογραφία 97

Πρόλογος

Η Ελληνική γλώσσα είναι μια γλώσσα με ιστορία 40 και πλέον αιώνων. Όπως γίνεται αντιληπτό μια τέτοια πορεία στο χρόνο επηρέασε και διαμόρφωσε τη γλώσσα την οποία ομιλούμε και σήμερα, τη Νέα Ελληνική Γλώσσα. Έτσι, μπορούμε να ισχυριστούμε ότι η Νέα Ελληνική Γλώσσα είναι ένα κράμα διαφόρων γλωσσικών συστατικών από όλες τις περιόδους της ιστορίας της.

Εκμεταλλευόμενοι λοιπόν την αναπόφευκτη ποικιλομορφία η οποία διέπει τη Νέα Ελληνική Γλώσσα προσπαθήσαμε να κινηθούμε σε δύο επίπεδα : Το πρώτο ήταν η ανάπτυξη ενός συστήματος το οποίο να μπορεί με μεγάλη ακρίβεια να καθορίσει και να ταξινομήσει τα είδη του λόγου τα οποία χρησιμοποιούνται σε ένα σώμα κειμένων. Το δεύτερο ήταν μέσα σε ένα συγκεκριμένο είδος λόγου να μπορέσουμε επίσης με μεγάλη ακρίβεια να διαχωρίσουμε τους συγγραφείς (ή στην περίπτωση μας, τους ομιλητές). Είναι αυτονόητο ότι το δεύτερο μέρος της εργασίας είναι σαφώς πιο απαιτητικό και δύσκολο να ολοκληρωθεί από το πρώτο και γι' αυτό πιστεύω ότι η έρευνα πάνω σε αυτόν τον τομέα θα συνεχιστεί.

Όσον αφορά την δομή αυτής της εργασίας, το ***Κεφάλαιο 1*** αναφέρεται στα γενικά χαρακτηριστικά που διέπουν την όλη εργασία καθώς και τις χρήσεις και εφαρμογές που μπορεί να έχει, το ***Κεφάλαιο 2*** αναφέρεται στα προγράμματα και τα λοιπά εργαλεία που χρησιμοποιήθηκαν για τις μετρήσεις και την επεξεργασία τους, το ***Κεφάλαιο 3*** μελετάει τις προσπάθειες μας για τον διαχωρισμό των ειδών του λόγου και το ***Κεφάλαιο 4*** επικεντρώνεται στην προσπάθεια διαχωρισμού ομιλητών από γραπτά κείμενα. Τέλος, το ***Κεφάλαιο 5*** συγκεντρώνει όλα τα συμπεράσματα και δίνει κάποιες κατευθύνσεις πάνω στις οποίες θα μπορούσε να κινηθεί στο μέλλον μια επέκταση της εργασίας αυτής.

Ολοκληρώνοντας, θα ήθελα να ευχαριστήσω όσους συνέβαλλαν στην εκπόνηση αυτής της διπλωματικής εργασίας. Τον καθηγητή κ. Γιώργο Καραγιάννη (*gkara@ilsp.gr*) ο οποίος μου έδωσε την ευκαιρία να εργαστώ πάνω σε αυτό το τόσο ενδιαφέρον θέμα, τον Δρ. Γεώργιο Ταμπουρατζή (*giorg_t@ilsp.gr*) για την αμέριστη και καθημερινή συμπαράστασή του, την Δρ. Στέλλα Μαρκαντωνάτου (*marks@ilsp.gr*) για τη βοήθεια σε γλωσσολογικά θέματα και την Δίδα Μαρίνα Βασιλείου (*mvas@ilsp.gr*) για όλη τη βοήθεια και την άψογη συνεργασία.

Αθήνα, Οκτώβριος 2000

Χαιρετάκης Νικόλαος

(*nhaire@softlab.ntua.gr*)

ΚΕΦΑΛΑΙΟ 1
ΓΕΝΙΚΑ

1.1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μια εκτενής αναφορά στους επιστημονικούς τομείς και τις εφαρμογές στις οποίες είναι απαραίτητος ο διαχωρισμός ειδών του λόγου και ο διαχωρισμός συγγραφέων ή ομιλητών. Αναφέρουμε αναλυτικά ποια ακριβώς χαρακτηριστικά της Ελληνικής Γλώσσας εκμεταλλευτήκαμε για να επιτύχουμε την ταξινόμηση των κειμένων και ποιες ανάλογες προσπάθειες έχουν γίνει για άλλες γλώσσες. Τέλος, γίνεται αναφορά στα σώματα κειμένων που χρησιμοποιήσαμε για τις μετρήσεις μας.

1.2 Αναζήτηση πληροφοριών (Information Retrieval)

Η ταξινόμηση μιας συλλογής κειμένων έτσι ώστε η εύρεση τους να γίνεται εύκολα και γρήγορα είναι μια αρκετά δύσκολη εργασία, ιδίως στην περίπτωση που περισσότεροι του ενός χρήστη πρόκειται να χρησιμοποιήσουν τη συλλογή αυτή.

Η παραδοσιακή μέθοδος οργάνωσης κειμένων και βιβλίων είναι η ταξινόμηση τους σε κατηγορίες οι οποίες έχουν προεπιλεγεί. Τα βασικά μειονεκτήματα της μεθόδου αυτής είναι ότι ο χρήστης θα πρέπει να εξοικειωθεί με το σύστημα ταξινόμησης – συχνά είναι δύσκολο να ξεχωρίσει σε ποια κατηγορία ανήκει ένα κείμενο – και ότι ένα κείμενο μπορεί λογικά και διαισθητικά να ανήκει σε περισσότερες από μια κατηγορίες. Μια λύση στα προβλήματα αυτά είναι η δημιουργία ενός ευρετηρίου (*index*) για τη συλλογή κειμένων. Τα ευρετήρια συνήθως δημιουργούνται από υπαλλήλους οι οποίοι διαβάζουν τα κείμενα και τα αρχειοθετούν κάτω από τις αντίστοιχες κατηγορίες. Η χειρωνακτική δημιουργία ενός ευρετηρίου είναι μια αρκετά δύσκολη, ίσως βαρετή εργασία, η οποία αντιμετωπίζει και το μειονέκτημα ότι κανείς δεν μπορεί να προβλέψει ποια ακριβώς μπορεί να είναι η

μελλοντική χρήση ενός κειμένου έτσι ώστε να το κατατάξει κάτω από όλες τις ανάλογες κατηγορίες. Συνεπώς, οδηγούμαστε στο συμπέρασμα ότι αυτή είναι μια εργασία η οποία πρέπει να αυτοματοποιηθεί. Όμως ακόμη και η αυτόματη δημιουργία ενός ευρετηρίου δεν είναι κάτι εύκολο να γίνει.

Οι αλγόριθμοι αναζήτησης πληροφοριών βασίζονται στις έννοιες της ακρίβειας (*precision*) και της ανάκλησης πληροφορίας (*recall*) [Yan99]. Η *ακρίβεια* στα κείμενα τα οποία ανευρέθησαν μετράται ως ο λόγος του αριθμού των σχετικών με το θέμα αναζήτησης κειμένων προς τον συνολικό αριθμό κειμένων. Η *ανάκληση πληροφορίας* μετράται ως ο λόγος του αριθμού των σχετικών ανευρεθέντων κειμένων προς τον αριθμό των συνολικών σχετικών κειμένων. Εάν ένας αλγόριθμος βρίσκει κάθε φορά όλα τα κείμενα της βάσης τότε έχει ποσοστό ανάκλησης 100%, αλλά μικρή ακρίβεια. Οι έννοιες της ακρίβειας και της ανάκλησης πληροφορίας είναι συνεπώς αντιστρόφως ανάλογες.

Σε ορισμένες εργασίες η ικανότητα άντλησης πληροφορίας έχει διαφορετικό αντίκτυπο στις έννοιες της ακρίβειας και της ανάκλησης πληροφορίας. Για παράδειγμα, εάν κάποιος εργάζεται πάνω σε μια ποινική υπόθεση ή μια ευρεσιτεχνία ίσως χρειάζεται να έχει στην κατοχή του όλα τα έγγραφα μιας βάσης κειμένων. Άρα το ποσοστό της ανάκλησης πληροφορίας σε αυτή την περίπτωση παίζει σημαντικό ρόλο. Ένας άλλος χρήστης όμως ίσως να αναζητεί απλώς μια απάντηση σε κάποιο ερώτημά του. Σε αυτή την περίπτωση το πρώτο κείμενο που θα βρεθεί ίσως του δώσει την απάντηση και θα αγνοήσει όλα τα υπόλοιπα.

Επιπλέον, υπάρχουν χαρακτηριστικά των κειμένων τα οποία ίσως μας οδηγήσουν σε λανθασμένη αντίληψη για το πόσο σχετικά είναι με το θέμα που αναζητάμε. Για παράδειγμα, νομικές ή ιατρικές συμβουλές οι οποίες κυκλοφορούν ελεύθερα στο internet, χωρίς να έχουν ελεγχθεί από αρμόδιες υπηρεσίες, μπορούν να περιέχουν ανακρίβειες ή παραπλανητικά στοιχεία. Επίσης, η παλαιότερη έκδοση ενός βιβλίου ή ενός εγχειριδίου μπορεί να αναιρείται πλήρως από τις νεότερες εκδόσεις και συνεπώς να μην μας ενδιαφέρει. Τέτοιου είδους προβλήματα είναι πολύ δύσκολο να προβλεφθούν και ίσως αδύνατον να αντιμετωπιστούν. Επειδή λοιπόν δεν μπορούμε να βρούμε και να εφαρμόσουμε ένα μέτρο για την σχετικότητα των

κειμένων, καταλήγουμε στο συμπέρασμα ότι κάποια κείμενα μπορεί να είναι μόνο *μερικώς* σχετικά.

Συνοψίζοντας, πρέπει να αναφέρουμε ότι ενώ η ακρίβεια και το ποσοστό ανάκλησης πληροφορίας είναι πολύ χρήσιμες έννοιες για την δοκιμή και την αξιολόγηση των αλγορίθμων αναζήτησης πληροφοριών, η χρησιμότητα τους σε πραγματικά συστήματα τείνει να είναι δυσδιάκριτη. Οι αλγόριθμοι είναι μόνο ένα μέρος του συστήματος αναζήτησης πληροφοριών. Ο ακριβής προσδιορισμός της εργασίας που θέλουμε να φέρουμε εις πέρας, οι προτιμήσεις του χρήστη, τα χαρακτηριστικά κάθε κειμένου, οι περιορισμοί σε χρόνο και κόστος και άλλοι παράγοντες παίζουν σημαντικό ρόλο στην αξιολόγηση ενός συστήματος.

1.3 Ποιο είναι το πρόβλημα

Στα συστήματα αναζήτησης πληροφοριών (*information retrieval systems*) η σχέση μεταξύ του ερωτήματος που θέτουμε στο σύστημα και του κειμένου που αναζητάμε προσδιορίζεται κυρίως από τον αριθμό και την συχνότητα των από κοινού τους όρων [HG96]. Τα εργαλεία που χρησιμοποιούμε σήμερα για αναζήτηση πληροφοριών εντός μιας βάσης κειμένων βασίζονται στην εμφάνιση συγκεκριμένων όρων σε ένα κείμενο. Ο ερευνών εισάγει τους όρους που αναζητεί και λαμβάνει ως απάντηση μια λίστα κειμένων τα οποία περιέχουν τους συγκεκριμένους όρους ή στενά σχετιζόμενους με αυτούς όρους [KBD+98]. Η μέθοδος αυτή λειτουργεί καλά μέχρι ενός σημείου. Αυτό είναι διαισθητικά κατανοητό, μιας και για ικανούς χρήστες και για καλά ορισμένες βάσεις κειμένων, τα αποτελέσματα θα είναι από μέτρια έως καλά. Όμως η μέθοδος αυτή παρουσιάζει και προφανή μειονεκτήματα. Η χρήση συχνοτήτων όρων μας δίνει μια γενικά φτωχή απεικόνιση του περιεχομένου του κειμένου για δύο κυρίως λόγους. Πρώτον, γιατί το κάθε κείμενο έχει περισσότερα χαρακτηριστικά από όσα μπορεί να του αποδώσει μια λίστα όρων (ιδίως εάν οι όροι αυτοί δεν είναι αντιπροσωπευτικοί) και δεύτερον γιατί είναι συνήθως δύσκολο για τους χρήστες να σημειώσουν με ακρίβεια μία ή περισσότερες λέξεις-κλειδιά οι οποίες θα τους δώσουν τα βέλτιστα αποτελέσματα.

Τα κείμενα δεν διαφοροποιούνται μόνο ως προς το θέμα τους. Οι διαφορές σε ύφος (*style*) μεταξύ κειμένων του ιδίου θέματος είναι συχνά το ίδιο καταφανείς όσο και οι διαφορές θέματος μεταξύ κειμένων διαφορετικών θεμάτων, αλλά του ίδιου ύφους ή είδους [Kar96]. Οι διαφορές ύφους μπορούν να οφείλονται σε ένα από τους ακόλουθους λόγους : επιλογή λέξεων, επιλογή συντακτικών δομών κ.ο.κ. Οι διαφορές αυτές οφείλονται είτε σε προσωπικές προτιμήσεις, είτε στο ότι τα κείμενα απευθύνονται σε κάποιο συγκεκριμένο κοινό, είτε στην ανάγκη ύπαρξης συνέχειας μεταξύ παρόμοιων κειμένων. Ως ύφος ορίζουμε την συνεπή και διακριτή τάση να χρησιμοποιούμε κάποιες από τις ανωτέρω γλωσσικές επιλογές [Kar99]. Ένας γενικότερος ορισμός του ύφους είναι η διαφορά μεταξύ δύο τρόπων περιγραφής του ίδιου πράγματος. Συνεπώς οι διαφορές μεταξύ των κειμένων – που δεν είναι κατ' αρχήν θεματικές – είναι διαφορές ύφους. Φυσικά, είναι αδύνατο να διαχωρίσουμε πλήρως τις θεματικές διαφορές από τις υφολογικές διαφορές. Ορισμένα κείμενα πρέπει ή τείνουν να εκφέρονται σε συγκεκριμένο ύφος : νομικά θέματα, δημοσιογραφικά κείμενα, τεχνική ορολογία κ.τ.λ. Για παράδειγμα, με διαφορετικό είδους λόγου θα περιγραφεί ένα έγκλημα από μια σκανδαλοθηρική εφημερίδα και με τελείως διαφορετικό είδος λόγου θα συνταχθεί η ποινική δικογραφία η οποία θα αφορά το ίδιο έγκλημα.

Όσον αφορά την αναζήτηση πληροφοριών, είναι γενικά πιο ενδιαφέρον να μελετήσουμε την περίπτωση των διαφορών σε ύφος σε σχέση με κείμενα παρόμοιων θεματικών ενοτήτων. Οι διαφορές οι οποίες θα συμπληρώσουν καλύτερα την αναζήτηση πληροφοριών σε κείμενα παρόμοιων θεματικών ενοτήτων δεν είναι οι διαφορές μεταξύ συγγραφέων (ή ομιλητών), ούτε οι διαφορές μεταξύ αυτόνομων κειμένων, αλλά οι συνεπείς, προβλέψιμες και ευδιάκριτες διαφορές μεταξύ ομάδων κειμένων. Συνεπώς τα πειράματα τα οποία θα πραγματοποιηθούν θα πρέπει να εστιαστούν σε χαρακτηριστικά των κειμένων τα οποία μπορούν να μετρηθούν και τα οποία οδηγούν σε μεταβλητές που μπορούν να διαχωρίσουν ή να ταξινομήσουν κείμενα. Ο σκοπός είναι να βρούμε μεταβλητές μέσω των οποίων θα μπορέσουμε να καταλάβουμε τι ύφους είναι το νέο ή άγνωστο κείμενο και συνεπώς να προβλέψουμε εάν θα ενδιαφέρει τον χρήστη ή τον αναγνώστη.

Ένα δεύτερο πρόβλημα είναι το πρόβλημα της ταξινόμησης κειμένων. Η ταξινόμηση κειμένων συνίσταται στην απόδοση σε μια κατηγορία, εντός μιας

δεδομένης ιεραρχίας κατηγοριών, των νεοεισερχομένων κειμένων που έρχονται στην κατοχή μας [DKM+98] και μπορεί να χρησιμοποιηθεί για να υποστηρίξει την αναζήτηση ή την εξαγωγή πληροφοριών καθώς και το φιλτράρισμα κειμένων. [Mou96].

Η χειρωνακτική ταξινόμηση κειμένων είναι μια ιδιαίτερα δαπανηρή και χρονοβόρα διαδικασία. Ιδιαίτερα σήμερα, οι ανάγκες της ταξινόμησης κειμένων έχουν αυξηθεί δραματικά, λόγω του πλούτου των πηγών κειμένων, όπως το Internet. Επιπλέον αντιμετωπίζουμε κι άλλα προβλήματα, όπως για παράδειγμα ότι ένα κείμενο δεν μπορεί να αποδοθεί ως ένα διάνυσμα χαρακτηριστικών -ακόμη κι εάν γίνει αυτό τα χαρακτηριστικά του μπορούν να είναι πάρα πολλά- ή ότι η πληροφορία που θα μας οδηγήσει στη σωστή ταξινόμηση του ίσως δεν είναι ευδιάκριτη ή τέλος ότι είναι απόλυτα φυσιολογικό για ένα κείμενο να περιέχει λάθη (ορθογραφικά, συντακτικά κ.τ.λ.) [LT96]. Όλα τα παραπάνω κάνουν το πρόβλημα που αντιμετωπίζουμε ιδιαίτερα ενδιαφέρον.

1.4 Χαρακτηριστικά Ελληνικής Γλώσσας

Η Νέα Ελληνική Γλώσσα, λόγω της αδιάλειπτης παρουσίας της τους τελευταίους 40 αιώνες έχει προσλάβει στοιχεία τα οποία διατηρούνται ακόμα και σήμερα. Κάποια από τα στοιχεία τα οποία μπορούν να καθορίσουν σε μεγάλο βαθμό το ύφος της γλώσσας που χρησιμοποιείται παρουσιάζονται στις επόμενες παραγράφους.

1.4.1 Πολυτυπία

Ένα τέτοιο στοιχείο, το οποίο εκμεταλλευτήκαμε στην συγκεκριμένη εργασία [ΤΜΗ+00a] είναι η *πολυτυπία*. Με τον όρο *πολυτυπία* εννοούμε την ύπαρξη πολλών διαφορετικών τύπων οι οποίοι χρησιμοποιούνται ευρέως και έχουν την ίδια σημασία ή αποδίδουν το ίδιο νόημα. Ένα παράδειγμα εμφάνισης *πολυτυπίας* δίνεται στην παρακάτω περίπτωση :

έβαλαν – βάλαν - βάλανε (1.1)

όπου και οι τρεις όροι χρησιμοποιούνται ευρέως στη Νέα Ελληνική Γλώσσα. Επίσης πολύ συχνά συναντάμε το φαινόμενο ότι παραπάνω από μία λέξεις χρησιμοποιούνται για την ίδια έννοια, όπως π.χ. :

ύδωρ – νερό, τριαντάφυλλο – ρόδο (1.2)

Η *πολυτυπία* στην Ελληνική γλώσσα ευνοήθηκε από την ύπαρξη της *Καθαρεύουσας* η οποία ήταν η επίσημη γλώσσα του Ελληνικού κράτους έως το 1979, οπότε και αντικαταστάθηκε από την *Δημοτική*. Η Νέα Ελληνική Γλώσσα, παρότι βρίσκεται πιο κοντά στην *Δημοτική*, εν τούτοις περιέχει ένα μεγάλο αριθμό χαρακτηριστικών της *Καθαρεύουσας*. Τέτοια χαρακτηριστικά απαντώνται σε διαφορετικές αναλογίες στα διάφορα είδη του λόγου. Γενικά, τα είδη του λόγου τα οποία εμφανίζουν ένα πιο επίσημο ύφος (όπως π.χ. ο ακαδημαϊκός λόγος ή η νομοθεσία) χρησιμοποιούν περισσότερα χαρακτηριστικά της *Καθαρεύουσας* από ότι η λογοτεχνία ή ο προφορικός λόγος. Πρέπει εδώ να τονιστεί ότι αρκετές από τις γραμματικά ισοδύναμες λεκτικές μονάδες απαντώνται αρκετά συχνά στη Νέα Ελληνική Γλώσσα. Ωστόσο, επειδή μερικές από αυτές είναι χαρακτηριστικές ενός επίσημου ύφους ή ενός ανεπίσημου ύφους, η χρήση τους μας φανερώνει την προτίμηση σε ένα συγκεκριμένο γλωσσικό είδος.

Η συγκεκριμένη εργασία χρησιμοποιεί τόσο μορφολογικά όσο και δομικά χαρακτηριστικά για την αυτόματη κατηγοριοποίηση κειμένων. Στην περίπτωση της Νέας Ελληνικής Γλώσσας, το σύστημα μας εκμεταλλεύεται το φαινόμενο της *πολυτυπίας* όπως επίσης και την έντονα κλιτική φύση της γλώσσας[1]. Έτσι, συλλέγουμε πληροφορίες για τις ρηματικές καταλήξεις, καθώς επίσης και για δομικά χαρακτηριστικά, όπως το μέγεθος των λέξεων και των προτάσεων. Όλα τα μορφολογικά και δομικά χαρακτηριστικά συγκεντρώνονται σε ένα αρχείο δεδομένων,

[1] Η έντονα κλιτική φύση της Ελληνικής Γλώσσας είναι ευδιάκριτη κυρίως αν τη συγκρίνουμε με άλλες γλώσσες, όπως η Αγγλική. Για παράδειγμα και αναφερόμενοι στα ρήματα, η Ελληνική Γλώσσα έχει ξεχωριστές καταλήξεις για κάθε πρόσωπο και αριθμό (κάν -ω, -εις, -ει, -ουμε, -ετε, -ουν) σε αντίθεση με τα ρήματα της Αγγλικής τα οποία –και σε συγκεκριμένους μόνο χρόνους- διαφοροποιούν τις καταλήξεις τους μόνο στο τρίτο πρόσωπο ενικό (I/you do, he/she/it does, we/you/they do). Επίσης, το ίδιο συμβαίνει και στα ουσιαστικά, όπου σε αντίθεση με την Ελληνική Γλώσσα (άλογ –ο, -ου, -ο, -α, -ων, -α) τα ουσιαστικά στην Αγγλική διαφοροποιούν μόνο την κατάληξη του πληθυντικού (horse – horses). Ακόμη πιο έντονη είναι η διαφορά στα επίθετα, όπου εκτός των όσων αναφέρθηκαν προηγουμένως, τα αγγλικά επίθετα διατηρούν την ίδια μορφή και στα τρία γένη (καλός, καλή, καλό – good).

τον οποίο στην συνέχεια επεξεργαζόμαστε με στατιστικές μεθόδους καθώς και με την χρήση νευρωνικών δικτύων.

1.4.2 Αρνήσεις

Η χρήση ορισμένων αρνητικών λέξεων (*ουδείς, ουδέποτε, ουδαμού, άνευ*) δηλώνει καθαρά μια προτίμηση για την Καθαρεύουσα, ενώ η χρήση άλλων (*δίχως, μήτε, χωρίς*) είναι ενδεικτική προτίμησης της Δημοτικής. Εν τούτοις, οι πιο συχνά χρησιμοποιούμενες αρνητικές λέξεις (*όχι, μην, δεν*) δεν είναι χαρακτηριστικές του ενός ή του άλλου είδους [MT00].

1.5 Ανάλογες προσπάθειες και αναφορές σε επιστημονικές εργασίες.

Προσπάθειες ανάλογες με τη δική μας έχουν γίνει, κυρίως για την Αγγλική Γλώσσα. Για παράδειγμα, ο *Adam Kilgariff* επιλέγει την μέθοδο της μέτρησης των συχνοτήτων των λέξεων. Στο άρθρο του "*Which words are characteristic of a text ? A survey of statistical approaches*" [Kil96] αναφέρει ότι οι λέξεις-κλειδιά είναι ιδιαίτερα χρήσιμες στην περίπτωση της αναζήτησης πληροφοριών. Οι λέξεις τις οποίες αναζητούμε έχουν γενικά μεγαλύτερη αξία σε όσο λιγότερα κείμενα κάνουν την εμφάνιση τους. Ένας επιπλέον λόγος ο οποίος αυξάνει την σημαντικότητα μιας λέξης σε ένα κείμενο είναι ο αριθμός εμφανίσεων της στο κείμενο αυτό. Επίσης, μελετώντας ένα κείμενο, είναι αναμενόμενο ότι εάν μια λέξη έχει ήδη εμφανιστεί μια φορά σε αυτό, τότε είναι πολύ πιθανότερο να την ξανασυναντήσουμε στο ίδιο κείμενο, παρά εάν μέχρι τώρα δεν έχει εμφανιστεί καθόλου.

Σε ένα άλλο άρθρο του [KR98] ο ίδιος προσπαθεί να αναπτύξει μια μέθοδο για την μέτρηση της ομοιότητας μεταξύ σωμάτων κειμένων. Επισημαίνει βέβαια ότι υπάρχουν εγγενή προβλήματα στην προσπάθεια αυτή όπως π.χ. ότι τα αποτελέσματα έχουν μεγάλη εξάρτηση από την αντίληψη του καθενός. Αυτό συμβαίνει για δύο κυρίους λόγους : πρώτον, γιατί κανείς δεν μπορεί εύκολα να διαβάσει ένα μεγάλο

σώμα κειμένων και να το συγκρίνει με κάποιο άλλο και δεύτερον γιατί η σύγκριση πολύπλοκων και πολυδιάστατων αντικειμένων δεν μπορεί να δώσει μια απλή απάντηση στο ερώτημα : "πόσο όμοια είναι τα δύο αντικείμενα ;"

Η *Ellen Riloff* ακολουθεί μια διαφορετική προσέγγιση στην αντιμετώπιση του προβλήματος. Συγκεκριμένα, ερευνά το κατά πόσο οι πολύ συχνά (και συνήθως μικρές σε μέγεθος) χρησιμοποιούμενες λέξεις μπορούν να είναι τελικά χρήσιμες και σημαντικές στην ταξινόμηση κειμένων [Ril95]. Το αποτέλεσμα στο οποίο καταλήγει είναι ότι -αντιθέτως με το τι πιστεύουν οι περισσότεροι- οι λέξεις αυτές όντως μπορούν να παίξουν καθοριστικό ρόλο στο συγκεκριμένο ερευνητικό τομέα. Έτσι, λέξεις όπως οι προθέσεις και οι αρνήσεις, αλλά ακόμη και η ύπαρξη ενός ουσιαστικού στον ενικό ή τον πληθυντικό αριθμό, ή ενός ρήματος σε διαφορετικούς χρόνους μπορούν να μας βοηθήσουν δραματικά στην κατηγοριοποίηση κειμένων.

Ο *Mehran Sahami* εφαρμόζει τον αλγόριθμο *του Multiple Cause Mixture Model* για να δημιουργήσει μια δομή πολλαπλών κατηγοριών ταξινόμησης σε ένα μη χαρακτηρισμένο σώμα κειμένων [SHS96]. Ο αλγόριθμος αυτός μέσα από ένα στάδιο εκμάθησης άνευ επίβλεψης μπορεί να εκτιμήσει σε ποια από τις κατηγορίες θα καταλήξει ένα νεοεισερχόμενο κείμενο. Το βασικό μειονέκτημα της μεθόδου αυτής είναι η πολύ μεγάλη υπολογιστική πολυπλοκότητα του αλγορίθμου που καθιστά απαγορευτική τη χρήση της σε συστήματα πραγματικού χρόνου (*real time*).

Η *Yiming Yang* συγκρίνει τρεις τεχνολογίες αυτόματου διαχωρισμού κειμένων [Yan96].

i. "Το ταίριασμα με βάση λέξεις" (*word-matching*), το οποίο συνταιριάζει κείμενα σε κατηγορίες με βάση τις κοινές λέξεις ανάμεσα στα κείμενα και τις κατηγορίες. Βασικό μειονέκτημα της μεθόδου αυτής είναι ότι εάν τα κείμενα και οι κατηγορίες δεν έχουν κάποιες κοινές λέξεις, τότε δεν συνταιριάζονται ακόμη και αν διαισθητικά θα έπρεπε.

ii. "Το ταίριασμα με βάση λεξικό συνωνύμων" (*thesaurus-based matching*), το οποίο χρησιμοποιεί λεκτικούς συνδέσμους (*lexical links*) –κατασκευασμένους χειρωνακτικά ή αυτόματα εκ των προτέρων- για να συσχετίσει ένα κείμενο με

τα ονόματα ή τις περιγραφικές φράσεις μιας κατηγορίας. Βασικό μειονέκτημα είναι ότι οι λεκτικοί σύνδεσμοι είναι συνήθως στατικοί και δεν μπορούν να αποδώσουν με ακρίβεια το νόημα μιας λέξης όταν αυτό εξαρτάται από τα συμφραζόμενα. Επίσης, αν το λεξικό συνωνύμων είναι κατασκευασμένο χειρωνακτικά τότε έχουμε να αντιμετωπίσουμε μεγάλο κόστος και χαμηλή προσαρμοστικότητα σε διαφορετικά είδη του λόγου.

iii. "Την εμπειρική εκμάθηση σχέσης λέξης-κατηγορίας" *(Empirical learning of term-category associations*) η οποία επιτυγχάνεται από την εκμάθηση ενός συνόλου κειμένων και τις κατηγορίες οι οποίες τους έχουν αποδοθεί. Η μέθοδος αυτή στηρίζεται στην ανθρώπινη αξιολόγηση της σχετικότητας των κειμένων και δεν έχει χρησιμοποιηθεί τόσο πολύ μέχρι σήμερα γιατί αντικειμενικά υπάρχουν χιλιάδες κατηγορίες στις οποίες μπορούν να ταξινομηθούν κείμενα.

Τέλος, ο *Jussi Karlgren* στην προσπάθεια του να διαχωρίσει γλωσσικά είδη χρησιμοποιεί κυρίως δομικές πληροφορίες καθώς και πληροφορίες για τα μέρη του λόγου [Kar99]. Ο *Karlgren* χρησιμοποιεί ως στατιστική μέθοδο την διακριτική ανάλυση [KC94] και τα ποσοστά επιτυχίας των πειραμάτων του κυμαίνονται από 52% (έχοντας επιλέξει 15 κατηγορίες διαχωρισμού) έως 96% (για 2 κατηγορίες διαχωρισμού). Βλέπουμε ότι το ποσοστό επιτυχίας μειώνεται δραματικά όσο αυξάνονται οι κατηγορίες διαχωρισμού, κάτι το οποίο είναι αναμενόμενο. Ο ίδιος εξηγεί ότι το φαινόμενο αυτό έχει να κάνει κυρίως με την επιλογή των κατηγοριών και με το τι έχει οριστεί να περιλαμβάνει η κάθε μία. Επίσης, αναφέρει ότι μια άλλη στατιστική τεχνική για να ανακαλύψουμε κατηγορίες είναι η παραγοντική ανάλυση (*factor analysis*). Το πρόβλημα αυτής της τεχνικής είναι ότι ενώ μεν οι κατηγορίες που προκύπτουν μπορεί να έχουν λογική έννοια, ίσως θα είναι δύσκολο να εξηγηθούν σε κάποιο μη ειδικό ο οποίος προσπαθεί να αναζητήσει πληροφορίες. Κλείνοντας, επισημαίνει ότι κάποιοι παράμετροι μπορεί να έχουν μεγάλη διακύμανση ή ακόμη και ασύμμετρη κατανομή σε ορισμένα κείμενα. Αυτό δεν είναι δύσκολο να καθοριστεί μέσω της τυπικής απόκλισης, αλλά πάντως είναι κάτι το οποίο πρέπει να διερευνηθεί.

1.6 Χαρακτηριστικά Corpus

Τα κείμενα που χρησιμοποιήσαμε για τις μετρήσεις μας χωρίζονται σε δύο μεγάλα σώματα κειμένων.

1. Το πρώτο σώμα κειμένων (*πίνακας 1.1*), το οποίο χρησιμοποιήθηκε για τον διαχωρισμό των ειδών του λόγου και πιο συγκεκριμένα τον διαχωρισμό μεταξύ πολιτικού, ιστορικού και λογοτεχνικού λόγου, αποτελείται από :

    i. 12 συνεδριάσεις της Ελληνικής Βουλής από το πρώτο μισό του έτους 1999, συνόλου 509.917 λέξεων (*πολιτικός λόγος*).

    ii. 32 κείμενα, 24 διαφορετικών συγγραφέων, γύρω από την ιστορία της Μακεδονίας, συνόλου 360.933 λέξεων (*ιστορικός λόγος*).

    iii. 24 λογοτεχνικά κείμενα, 6 διαφορετικών συγγραφέων, συνόλου 363.873 λέξεων (*λογοτεχνία*).

Σώμα Κειμένων Ι	*Κείμενα*	*Λέξεις*
Πολιτικός Λόγος	12	509.917
Ιστορικός Λόγος	32	360.933
Λογοτεχνία	24	363.873
Σύνολο	**68**	**1.234.723**

Πίνακας 1.1 – Το σώμα κειμένων Ι

2. Το δεύτερο σώμα κειμένων (*πίνακας 1.2*), το οποίο χρησιμοποιήθηκε για τον διαχωρισμό των ομιλητών και πιο συγκεκριμένα τον διαχωρισμό μεταξύ πέντε (από εδώ και στο εξής ομιλητές Α, Β, Γ, Δ, Ε) προσωπικοτήτων της Ελληνικής Βουλής (έναν από κάθε κόμμα το οποίο εκπροσωπούνταν στη Βουλή την περίοδο 1997 - 2000), αποτελούνταν από :

    i. 418 ομιλίες του ομιλητή Α, συνόλου 463.680 λέξεων.
    ii. 85 ομιλίες του ομιλητή Β, συνόλου 177.853 λέξεων.
    iii. 245 ομιλίες του ομιλητή Γ, συνόλου 241.882 λέξεων.
    iv. 154 ομιλίες του ομιλητή Δ, συνόλου 217.305 λέξεων.

v. 103 ομιλίες του ομιλητή Ε, συνόλου 190.601 λέξεων.

Όλες αυτές οι ομιλίες έλαβαν χώρα κατά τη διάρκεια της περιόδου Ιανουάριος 1997 - Μάρτιος 2000, ενώ όπως θα αναφέρουμε και στο αντίστοιχο κεφάλαιο εκτός από το σύνολο των ομιλιών χρησιμοποιήσαμε και ένα υποσύνολο τους για τη λήψη αποτελεσμάτων.

Σώμα Κειμένων ΙΙ	*1997*	*1998*	*1999*	*2000*	*Ομιλίες*	*Λέξεις*
Ομιλητής Α	161	107	125	25	418	463.680
Ομιλητής Β	36	28	18	3	85	177.853
Ομιλητής Γ	81	71	67	26	245	241.882
Ομιλητής Δ	72	49	30	3	154	217.305
Ομιλητής Ε	16	42	38	7	103	190.601
Σύνολο	366	297	278	64	**1005**	**1.291.321**

Πίνακας 1.2 – Το σώμα κειμένων ΙΙ

Πρέπει εδώ να επισημανθεί ότι όλα τα κείμενα τα οποία χρησιμοποιήσαμε είχαν διορθωθεί για τυχόν ορθογραφικά ή άλλα λάθη. Επιπλέον, τα πρακτικά της Βουλής έχουν υποστεί και μια προεπεξεργασία με κύριο σκοπό την απάλειψη των διαλόγων μικρής έκτασης ή των παρεμβάσεων (ελλειπτικές προτάσεις), φαινόμενα τα οποία είναι τυπικά για συνεδριάσεις της Βουλής, αλλά δεν αποτελούν δείγμα ολοκληρωμένου λόγου και δεν αναμένεται να χαρακτηρίζουν κάποιο ομιλητή.

ΚΕΦΑΛΑΙΟ 2
ΕΡΓΑΛΕΙΑ

2.1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται αναφορά σε όλα τα εργαλεία τα οποία χρησιμοποιήσαμε κατά την διάρκεια των πειραμάτων, τόσο για την λήψη των απαραίτητων μετρήσεων όσο και για την επεξεργασία των αποτελεσμάτων.

2.2 AMP (Automated Morphological Processor)

Το πρόβλημα της μορφολογικής επεξεργασίας λεκτικών μονάδων έχει αντιμετωπιστεί με δύο κυρίως προσεγγίσεις. Η πρώτη έγκειται στη χρήση γλωσσολογικών κανόνων [PW93] για την μοντελοποίηση των μορφολογικών φαινομένων, την συμπίεση της πληροφορίας στα μορφολογικά λεξικά, την λημματοποίηση και τέλος την παραγωγή των μορφολογικών τύπων. Η δεύτερη προσέγγιση χρησιμοποιεί υπολογιστικές αρχιτεκτονικές οι οποίες προσομοιώνουν τη δομή βιολογικών νευρωνικών δικτύων, ώστε να επιτευχθεί η μορφολογική επεξεργασία των λέξεων [GAS94]. Το πρόγραμμα που χρησιμοποιήσαμε για τις μετρήσεις μας (AMP) χρησιμοποιεί και αξιοποιεί γλωσσολογικούς κανόνες, αλλά ταυτόχρονα ενσωματώνει κάποια στατιστικά στοιχεία για να επιτύχει τον σκοπό του.

Οι μετρήσεις των μορφολογικών χαρακτηριστικών γίνονται μέσω του AMP (*Automated Morphological Processor).* Το AMP έχει σχεδιασθεί έτσι ώστε να εκμεταλλεύεται πλήρως την άκρως κλιτική φύση της Ελληνικής Γλώσσας, επιτρέποντας και διευκολύνοντας την παραγωγή μορφολογικών λεξικών με αυτόματο τρόπο [TK99]. Η χειρωνακτική παραγωγή μορφολογικών λεξικών είναι μια επίπονη και χρονοβόρα εργασία και το AMP επιτρέπει την αυτόματη δημιουργία τους με μεγάλη ακρίβεια. Πιο συγκεκριμένα, η ακρίβεια που επιτυγχάνει το AMP

συγκρίνοντας τα αποτελέσματά του με τα περιεχόμενα του μορφολογικού λεξικού, το οποίο έχει κατασκευαστεί χειρωνακτικά στο ΙΕΛ, αγγίζει το 96%. [AM95].

Το AMP βασίζεται στην τεχνική της ταύτισης και απόκρυψης *(matching and masking technique*), η οποία χρησιμοποιείται ευρέως σε εφαρμογές αναγνώρισης προτύπων (*pattern recognition*). Σύμφωνα με την τεχνική αυτή, το κάθε πρότυπο διαιρείται σε τμήματα και επιχειρείται η ταύτιση ομοειδών τμημάτων των προτύπων ενώ ταυτόχρονα αποκρύπτονται προσωρινά τα υπόλοιπα τμήματα των προτύπων. Στην συγκεκριμένη περίπτωση, όπου τα πρότυπα είναι λεκτικές μονάδες, η κάλυψη και απόκρυψη αναφέρεται στα θέματα και τις αντίστοιχες καταλήξεις των λεκτικών μονάδων. Έτσι, αν υποτεθεί ότι υπάρχουν δύο συγκεκριμένες λέξεις οι οποίες παριστάνονται ως $\theta_1\kappa_1$ και $\theta_2\kappa_2$, όπου θ_i το εκάστοτε θέμα και κ_i η αντίστοιχη κατάληξη, τότε η επιτυχής ταύτιση συνεπάγεται ότι είτε $\theta_1=\theta_2$ (ταύτιση των θεμάτων των δύο λέξεων), είτε $\kappa_1=\kappa_2$ (ταύτιση των καταλήξεων των δύο λέξεων).

Η διαδικασία εντοπισμού πιθανών θεμάτων και καταλήξεων υλοποιείται με την επαναληπτική εφαρμογή της τεχνικής ταύτισης και απόκρυψης. Σε κάθε επανάληψη, μελετώνται όλες οι λεκτικές μονάδες που υπάρχουν στο σώμα κειμένων ώστε από τις ήδη προσδιορισθείσες καταλήξεις και θέματα να προσδιορισθούν νέες πιθανές καταλήξεις και θέματα. Έτσι, σταδιακά δημιουργείται ένα σύνολο από πιθανές καταλήξεις και θέματα για κάθε λεκτική μονάδα. Φυσικά, το σύστημα συμπληρώνεται από ένα σύνολο γραμματικών περιορισμών ως προς τις αποδεκτές και μη μορφές θεμάτων και καταλήξεων της Ελληνικής Γλώσσας.

Με την παραπάνω διαδικασία, για κάθε λέξη του σώματος παράγεται ένας αριθμός από δυνατές λύσεις διαχωρισμού σε θέμα και κατάληξη. Αυτές οι λύσεις κατατάσσονται ανάλογα με την πιθανότητά τους να αποτελούν τον βέλτιστο διαχωρισμό της δεδομένης λεκτικής μονάδας σε θέμα και κατάληξη χρησιμοποιώντας ένα κριτήριο κατάταξης, και πιο συγκεκριμένα το ελάχιστο μήκος κατάληξης.

Για τους σκοπούς του διαχωρισμού ειδών λόγου και ομιλητών είναι απαραίτητο να μετρήσουμε τις συχνότητες εμφάνισης συγκεκριμένων καταλήξεων, όπως αναφέρουμε αναλυτικά στα Κεφ. 3 και 4. Συνεπώς, στη περίπτωση μας το AMP περιορίζεται να χρησιμοποιήσει δεδομένες καταλήξεις, οι οποίες είναι

χαρακτηριστικές για το κάθε είδος, ως a priori γνώση. Για κάθε μία από τις καταλήξεις αυτές το AMP μετράει τη συχνότητα εμφάνισης της, καταγράφει τα αντίστοιχα θέματα και μετράει τη συχνότητα εμφάνισης των.

Επειδή το AMP δεν είχε αρχικά σχεδιασθεί για την εκτέλεση αυτών των εξειδικευμένων μετρήσεων, χρειάστηκε να γίνουν κάποιες μετατροπές και βελτιώσεις στο αρχικό πρόγραμμα. Επίσης, λόγω του ότι η λίστα των καταλήξεων που χρησιμοποιήσαμε ήταν αρκετά εκτενής, το πρόγραμμα μπορούσε να εκτελεστεί μόνο σε περιβάλλον Linux και όχι σε SunOS που ήταν αρχικά η πλατφόρμα υλοποίησης του.

2.3 Άλλα μετροπρογράμματα

Οι μετρήσεις των δομικών χαρακτηριστικών γίνονται με την βοήθεια άλλων μετροπρογραμμάτων τα οποία έχουν γραφεί σε γλώσσα προγραμματισμού C και τρέχουν σε περιβάλλον Linux, καθώς και σε shell scripts τα οποία εκτελούνται επίσης σε περιβάλλον Linux. Ειδικότερα, χρησιμοποιήσαμε κυρίως δύο προγράμματα :

i. Το πρώτο δέχεται ως είσοδο ένα κείμενο και μετράει τον αριθμό των λέξεων, προτάσεων, κομμάτων, παρενθέσεων, παυλών, ερωτηματικών, καθώς και τις κατανομές λέξεων μήκους x σε γράμματα και τις κατανομές προτάσεων μήκους x σε λέξεις.

ii. Το δεύτερο πρόγραμμα (*shell script*) το οποίο στηρίζεται στη χρήση της εντολής grep του unix, δέχεται ως είσοδο τα λημματοποιημένα αρχεία από τον tagger (§ 2.4) και μετράει τη συχνότητα των μερών του λόγου, τα ουσιαστικά και επίθετα τα οποία βρίσκονται σε γενική πτώση, τα πρόσωπα ανά αριθμό των ρημάτων, τις αρνήσεις και τα όποια λήμματα χρησιμοποιήσαμε.

iii. Τέλος, χρησιμοποιήθηκαν και άλλα ήσσονος σημασίας προγράμματα για άλλες εργασίες, όπως η ταξινόμηση (*sorting*) λημμάτων, καθώς και ένα macro του Microsoft Word 97 το οποίο "καθάριζε" τα κείμενα από τα σημεία στίξης, αφήνοντας μόνο τις λέξεις.

2.4 Tagger – Lemmatizer (Λημματοποιητής)

Για την κατηγοριοποίηση των λέξεων των κειμένων σε μέρη του λόγου (*PoS*) χρησιμοποιήθηκε ο *tagger*, ο οποίος έχει αναπτυχθεί στο ΙΕΛ. Ο *tagger* αυτός έχει ακρίβεια πάνω από 96% [PPG^{+}00] στην ταξινόμηση των μερών του λόγου, ποσοστό το οποίο είναι αρκετά ικανοποιητικό για τις εφαρμογές στις οποίες τον χρησιμοποιήσαμε.

2.5 Στατιστικά Προγράμματα

Η ανάλυση των μετρήσεων έγινε κατ' αρχήν με την βοήθεια στατιστικών μεθόδων. Για το σκοπό αυτό χρησιμοποιήσαμε τα πακέτα Statgraphics 3.1 και Statistica 5.0. Οι δύο στατιστικές μέθοδοι που χρησιμοποιήσαμε είναι η ανάλυση σε ομάδες (*Cluster Analysis*) και η διακριτική ανάλυση (*Discriminant Analysis*).

Ο όρος *cluster analysis* (ανάλυση σε ομάδες) στην πραγματικότητα περικλείει ένα μεγάλο αριθμό από διαφορετικούς αλγόριθμους ταξινόμησης. Ένα βασικό ερώτημα το οποίο αντιμετωπίζουν πολλοί ερευνητές σε θέματα αναζήτησης πληροφορίας είναι το πως μπορούν να οργανωθούν οι παρατηρήσεις σε δομές οι οποίες να έχουν νόημα. Για παράδειγμα, ο βιολόγος ο οποίος προσπαθεί να οργανώσει τα διάφορα είδη ζώων χωρίς να λάβει κατ' αρχήν υπόψη του τις διαφορές μεταξύ τους, κατατάσσει τον άνθρωπο στα πρωτεύοντα θηλαστικά, τα θηλαστικά, τα σπονδυλωτά και τα ζώα. Στην κατάταξη αυτή παρατηρούμε ότι όσα περισσότερα μέλη έχει μια ομάδα, τόσο λιγότερο όμοια είναι αυτά τα μέλη μεταξύ τους. Εν προκειμένω ο άνθρωπος έχει περισσότερα κοινά με τα άλλα πρωτεύοντα θηλαστικά (π.χ. πίθηκος) παρά με τα άλλα θηλαστικά (π.χ. σκύλος). Αναλυτικότερα, δεδομένου ενός αριθμού αντικειμένων ή παρατηρήσεων που έχουν γίνει σε n χαρακτηριστικά (n μεταβλητές), η ανάλυση σε ομάδες (*cluster analysis*) ασχολείται με την εύρεση των "αποστάσεων" (ανομοιοτήτων) της καθεμιάς παρατήρησης από όλες τις υπόλοιπες και στη συνέχεια χωρίζει το σύνολο των παρατηρήσεων σε ένα αριθμό Κ ομάδων κατά τέτοιο τρόπο ώστε δύο παρατηρήσεις που ομαδοποιούνται στην ίδια ομάδα να είναι όσο το δυνατόν πιο όμοιες μεταξύ τους και παρατηρήσεις που βρίσκονται σε

διαφορετικές ομάδες να είναι το δυνατόν πιο "ανόμοιες". Προφανώς, στη μέθοδο αυτή ανάλυσης σημαντικό ρόλο διαδραματίζει η μέθοδος μέτρησης της απόστασης.

Η *discriminant analysis* (διακριτική ανάλυση) χρησιμοποιείται για να καθορίσουμε ποιες μεταβλητές διαφοροποιούν δύο ή περισσότερες ομάδες. Μέσω της *discriminant analysis*, η οποία δέχεται ως είσοδο μια ομάδα από προ-ταξινομημένα στοιχεία και όλες τις ανεξάρτητες μεταβλητές τους, λαμβάνουμε ως αποτέλεσμα ένα σετ από συναρτήσεις διαχωρισμού οι οποίες ταξινομούν τις αντίστοιχες ομάδες (για n ομάδες έχουμε n-1 συναρτήσεις διαχωρισμού) [Man86]. Οι συναρτήσεις αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν για να προβλέψουν την κατηγορία στην οποία θα ταξινομηθούν τα καινούρια στοιχεία, κάτι το οποίο έχει εφαρμόσει και ο *Karlgren* [Kar00]. Αναλυτικότερα, δεδομένου ενός αριθμού αντικειμένων ή παρατηρήσεων που έχουν γίνει σε n χαρακτηριστικά (n μεταβλητές), η διακριτική ανάλυση (*discriminant analysis*) ασχολείται με τον προσδιορισμό K διακριτικών συναρτήσεων (δηλ. K γραμμικών συνδυασμών των n μεταβλητών) που χωρίζουν τον χώρο των n διαστάσεων σε K+1 υπόχωρους στους οποίους κατατάσσονται όλες οι παρατηρήσεις κατά τέτοιο τρόπο ώστε το ποσοστό των παρατηρήσεων που δεν έχουν τοποθετηθεί στο σωστό τους υπόχωρο να είναι το ελάχιστο δυνατόν.

Εκτός από το πλήρες μοντέλο *discriminant analysis* υπάρχουν δύο ακόμα, τα οποία χρησιμοποιήσαμε και εξηγούμε παρακάτω :

i. Εμπρός Βηματική Ανάλυση (*Forward Stepwise Analysis*). Με τη μέθοδο αυτή χτίζουμε ένα μοντέλο διακριτικής ανάλυσης βήμα-βήμα. Το μοντέλο αυτό αρχικά δεν περιλαμβάνει καμία μεταβλητή. Η διαδικασία βήμα προς βήμα εξαρτάται από τις δύο τιμές F1 και F2, που έχουν καθορισθεί για την προσθήκη και για την αφαίρεση μεταβλητών στο μοντέλο. Συγκεκριμένα, σε κάθε βήμα (i) εξετάζονται όλες οι μεταβλητές που ευρίσκονται εκτός του μοντέλου, προσδιορίζεται αυτή που συνεισφέρει περισσότερο στο διαχωρισμό σε ομάδες και εφόσον η συνεισφορά αυτή υπερβαίνει την τιμή F1, η μεταβλητή εισάγεται στο μοντέλο και (ii) εξετάζονται όλες οι μεταβλητές που ευρίσκονται εντός του μοντέλου, προσδιορίζεται αυτή που συνεισφέρει το λιγότερο στον ορθό διαχωρισμό σε ομάδες και - εφόσον η συνεισφορά αυτή είναι μικρότερη από την

τιμή F2 - η μεταβλητή αφαιρείται από το μοντέλο. Με τον ίδιο τρόπο συνεχίζουμε μέχρι να χτίσουμε το τελικό μοντέλο.

ii. Πίσω Βηματική Ανάλυση (*Backward Stepwise Analysis*). Το μοντέλο στην αρχή περιλαμβάνει όλες τις μεταβλητές και στην συνέχεια, αποβάλλονται αλλά και επανεισάγονται μεταβλητές με διαδικασία ανάλογη με αυτή της Εμπρός Βηματικής Ανάλυσης. Συνεπώς, στην περίπτωση της επιτυχούς ανάλυσης, θα κρατήσουμε μόνο τις σημαντικές μεταβλητές στο μοντέλο.

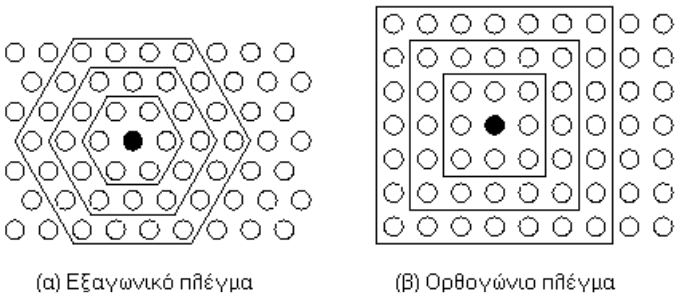
Τόσο η *Forward Stepwise Analysis* όσο και η *Backward Stepwise Analysis* χρησιμοποιούνται στις περιπτώσεις που έχουμε μεγάλο αριθμό παραμέτρων ώστε να εξετασθεί εάν ένας μικρότερος αριθμός χαρακτηριστικών μπορεί να διαχωρίσει με την ίδια τάξη ακρίβειας τις δεδομένες παρατηρήσεις.

Η διαδικασία βήμα προς βήμα εξαρτάται από τις τιμές F1 για την προσθήκη και F2 για την αποβολή μεταβλητών. Η τιμή F μιας μεταβλητής υποδεικνύει την στατιστική σημαντικότητά της στο διαχωρισμό σε ομάδες, δηλαδή το μέτρο του κατά πόσο η συγκεκριμένη μεταβλητή έχει μοναδική συμβολή στην πρόβλεψη των ομάδων. Γενικότερα το μοντέλο θα συνεχίσει να προσθέτει μεταβλητές όσο οι αντίστοιχες τιμές F κάποιων μεταβλητών είναι μεγαλύτερες από την τιμή της F1 που καθόρισε ο χρήστης, ενώ θα αφαιρεί μεταβλητές από το μοντέλο, εάν η σημαντικότητα τους υπολείπεται της F2 που καθόρισε ο χρήστης.

2.6 SOM Toolbox

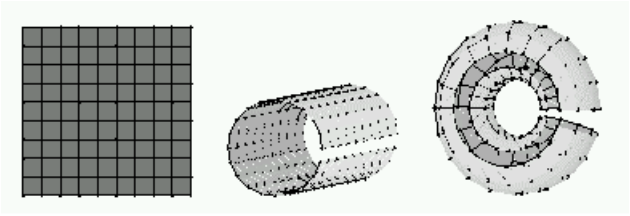
Στη συνέχεια η ανάλυση των μετρήσεων έγινε με τη χρήση νευρωνικού δικτύου και πιο συγκεκριμένα με το πακέτο *SOM Toolbox* για το περιβάλλον της *Matlab*. Το SOM (*Self Organizing Map*) είναι ένα μοντέλο νευρωνικού δικτύου και αλγορίθμου ο οποίος υλοποιεί μια χαρακτηριστική μη γραμμική απεικόνιση από ένα χώρο μεγάλων διαστάσεων σε ένα πίνακα νευρώνων μικρών διαστάσεων [KOS$^+$96]. Αυτό βέβαια γίνεται μεθοδικά και ακολουθώντας κάποιους κανόνες. Ο χάρτης που προκύπτει τείνει να διατηρήσει τις τοπολογικές σχέσεις των δεδομένων. Πιο αναλυτικά, το SOM αποτελείται από νευρώνες οργανωμένους σε ένα κανονικό μονοδιάστατο ή

διδιάστατο πλέγμα. Περισσοτέρων διαστάσεων πλέγματα μπορούν επίσης να χρησιμοποιηθούν, αλλά το συγκεκριμένο *Toolbox* δεν υποστηρίζει την οπτικοποίησή τους. Στην περίπτωση του διδιάστατου πλέγματος οι νευρώνες παίρνουν τη μορφή είτε ορθογώνιου, είτε εξαγωνικού πλέγματος (*σχήμα 2.1*).



Σχήμα 2.1

Εάν οι πλευρές του χάρτη συνδέονται μεταξύ τους, το σχήμα του χάρτη γίνεται κυλινδρικό ή με μορφή τόρου (*σχήμα 2.2*). Ο αριθμός των νευρώνων μπορεί να ποικίλλει από μερικές δεκάδες έως μερικές χιλιάδες. Κάθε νευρώνας αντιπροσωπεύεται από ένα κ-διάστατο πίνακα βαρών $x=[x_{1,\ldots,}x_{\kappa}]$, όπου κ είναι η διάσταση του πίνακα εισόδου. Οι νευρώνες συνδέονται με τους γειτονικούς τους μέσω μιας σχέσης γειτνίασης, η οποία διέπει την τοπολογία ή τη δομή του χάρτη. Στο *SOM Toolbox*, η τοπολογία διανέμεται σε δύο παράγοντες: την τοπική δομή του δικτυώματος και το σχήμα του χάρτη [VHA^{+}00].



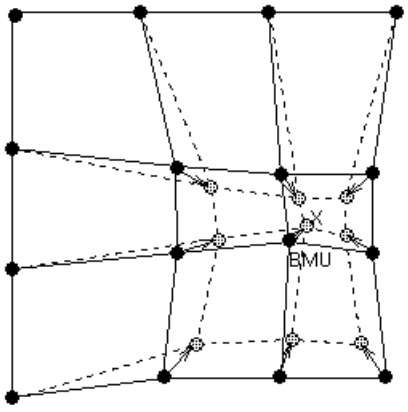
Σχήμα 2.2

Ο αλγόριθμος εκμάθησης του SOM ομοιάζει με αλγόριθμους κβαντοποίησης διανυσμάτων (*vector quantization*), όπως ο *k-means*, ή ο *LVQ (Learning Vector Quantization)*. Η σημαντική διαφορά είναι ότι εκτός του διανύσματος βάρους ενημερώνονται και οι γείτονες στο χάρτη : η περιοχή γύρω από τον καλύτερα ταιριαστό πίνακα προσαρμόζεται στο παρόν δείγμα εκμάθησης. Το τελικό αποτέλεσμα είναι ότι οι νευρώνες του δικτυώματος είναι διατεταγμένοι και οι γειτονικοί νευρώνες μετά το τέλος της διαδικασίας εκμάθησης (άνευ επίβλεψης)

έχουν παρόμοια διανύσματα βάρους. Τα βάρη του επικρατούντος νευρώνα ρυθμίζονται με τη χρήση του κανόνα εκμάθησης Kohonen. Υποθέτοντας ότι επικρατεί ο i^{ος} νευρώνας, η i^{η} γραμμή του πίνακα βαρών (*IW*) προσαρμόζεται ως εξής :

$$_iIW^{1,1}(q) = {}_iIW^{1,1}(q-1) + a(p(q) - {}_iIW^{1,1}(q-1))$$

Συνεπώς, ο νευρώνας του οποίου ο πίνακας βαρών είναι πλησιέστερος στον πίνακα εισόδου ενημερώνεται έτσι ώστε να βρεθεί ακόμα πιο κοντά (*σχήμα 2.3*). Το αποτέλεσμα είναι ότι ο επικρατών νευρώνας (*Best Matching Unit - BMU*) έχει περισσότερες πιθανότητες να επικρατήσει ξανά εάν εμφανιστεί παρόμοιος πίνακας εισόδου ενώ αντίθετα έχει λιγότερες πιθανότητες επικράτησης εάν ο πίνακας εισόδου είναι πολύ διαφορετικός. Όσο περισσότερες είσοδοι παρουσιάζονται, τόσο κάθε νευρώνας προσαρμόζεται καλύτερα. Τελικά, εάν υπάρχουν αρκετοί νευρώνες, κάθε ομάδα παρόμοιων εισόδων θα έχει ένα νευρώνα ο οποίος θα δίνει έξοδο 1 (ή κοντά στο 1) όταν παρουσιάζεται μια είσοδος από την παραπάνω ομάδα και έξοδο 0 (ή κοντά στο 0) για κάθε άλλη είσοδο. Άρα, το δίκτυο μαθαίνει να κατηγοριοποιεί τις νέες εισόδους.



Σχήμα 2.3 – Ενημέρωση του σημείου BMU και των γειτόνων του ως προς τη νέα είσοδο x. Οι συνεχόμενες και διακεκομμένες γραμμές αντιστοιχούν στην κατάσταση πριν και μετά την ενημέρωση, αντίστοιχα.

Το SOM μπορεί επίσης προβλέψει το επόμενο δείγμα, λαμβάνοντας υπόψη του τις πληροφορίες που περιέχουν τα προηγούμενα δείγματα. Με τη χρήση της γραμμικής παλινδρόμησης (*linear regression*) ένα νευρωνικό δίκτυο μπορεί να

“μάθει” την ιστορία του συστήματος και στην συνέχεια να την εφαρμόσει στην πρόβλεψη των μελλοντικών αποτελεσμάτων [KK96].

Αφού τα διανύσματα βάρους του SOM έχουν καλά ορισμένες συντεταγμένες στον χάρτη, το SOM μπορεί να θεωρηθεί και ως ένας αλγόριθμος διανυσματικής προβολής. Οι πρωτότυποι πίνακες και οι προβολές τους ορίζουν ένα ολίγων-διαστάσεων χάρτη του αρχικού (πολλών διαστάσεων) τοπολογικού χάρτη.

Οι πιο χρήσιμες εφαρμογές του SOM είναι η οπτικοποίηση συστημάτων μεγάλων διαστάσεων και η επεξεργασία και εύρεση κατηγοριών από μη επεξεργασμένα δεδομένα. Έτσι, το SOM βρίσκει πεδίο εφαρμογής στην αναγνώριση προτύπων (computer vision, speech recognition), στη ρομποτική και στις τηλεπικοινωνίες [Koh90].

Όσον αφορά το SOM και την αναζήτηση πληροφοριών ή την ταξινόμηση κειμένων, πρέπει να αναφέρουμε ότι η συγκεκριμένη μέθοδος έχει τύχει ευρύτατης χρήσης από πολλούς ερευνητές (για παράδειγμα [Mer99] και [Hyo96]). Οι χάρτες οι οποίοι προκύπτουν απεικονίζουν την ομοιότητα που έχουν τα κείμενα τα οποία μελετάμε. Έτσι, όσα κείμενα είναι παρόμοια τοποθετούνται σε γειτονικές περιοχές του χάρτη. Βέβαια, λόγω της απεικόνισης ενός χώρου προτύπων πολλών διαστάσεων σε ένα μόνο χάρτη δύο διαστάσεων προκύπτουν κάποιοι περιορισμοί όσον αφορά την ακρίβεια απεικόνισης. Μία ακόμη έλλειψη που παρουσιάζει η μέθοδος αυτή είναι η αδυναμία της να ορίσει επακριβώς τα όρια των ομάδων που προκύπτουν [Mer97].

Το SOM Toolbox είναι μια υλοποίηση και απεικόνιση του SOM στο περιβάλλον της Matlab. Το Toolbox αυτό μπορεί να χρησιμοποιηθεί για την επεξεργασία δεδομένων, την αρχικοποίηση και εκπαίδευση SOM χρησιμοποιώντας μια πλειάδα από τοπολογίες, την απεικόνιση του SOM με διάφορους τρόπους και την ανάλυση των ιδιοτήτων του SOM και των δεδομένων (ποιότητα SOM, ομάδες στον χάρτη και συσχετίσεις μεταξύ των μεταβλητών).

ΚΕΦΑΛΑΙΟ 3
ΔΙΑΧΩΡΙΣΜΟΣ ΕΙΔΩΝ ΛΟΓΟΥ

3.1 Εισαγωγή

Το τρίτο κεφάλαιο εστιάζεται στη προσπάθεια που κάναμε να διαχωρίσουμε τα τρία προαναφερθέντα είδη λόγων : πολιτικό, ιστορικό και λογοτεχνικό. Στις παραγράφους που ακολουθούν, επισυνάπτουμε όλες τις σχετικές πληροφορίες για τα κείμενα που μελετήσαμε, τις μετρήσεις που κάναμε, την επεξεργασία τους και τέλος κάποια συμπεράσματα για την ακρίβεια της μεθόδου που ακολουθήσαμε.

3.2 Κείμενα

Τα κείμενα που χρησιμοποιήσαμε στη συγκεκριμένη ενότητα είναι τα κείμενα που αποτελούν το πρώτο σώμα κειμένων (*πίνακας 1.1*).

3.3 Ρηματική Πολυτυπία

Η *πολυτυπία*, όπως αναφέρθηκε και προηγουμένως, είναι ένα αρκετά διαδεδομένο φαινόμενο στη Νέα Ελληνική Γλώσσα, ενώ παρατηρείται κυρίως στους ρηματικούς τύπους (*1.1*). Σε μικρότερο βαθμό, η *πολυτυπία* παρατηρείται στα ουσιαστικά και τα επιρρήματα (*1.2*) [ΜΚ00]. Στα ουσιαστικά η πολυτυπία είναι περιορισμένη σε ένα πολύ μικρό αριθμό όρων, ενώ τα επιρρήματα είναι σαφώς πιο σπάνια από τα ρήματα. Επιπλέον, στην περίπτωση των επιρρημάτων, υπάρχει αλληλεπίδραση μορφολογικών και σημασιολογικών θεμάτων με άμεση συνέπεια η αυτόματη επεξεργασία να καθίσταται αδύνατη [ΤΜΗ^{+}00a]. Έχοντας υπόψη μας τα παραπάνω γεγονότα αποφασίσαμε να περιορίσουμε τις μετρήσεις μας στην περίπτωση της *ρηματικής πολυτυπίας*.

Η *ρηματική πολυτυπία* της Ελληνικής Γλώσσας οφείλεται κατά κύριο λόγο στις καταλήξεις των ρημάτων και σε μικρότερο βαθμό στα προθέματα. Μερικές πάντως από τις διαφορετικές καταλήξεις δεν συνιστούν απαραίτητα κάποια διαφορά ανάμεσα σε Καθαρεύουσα και Δημοτική Γλώσσα. Στην εργασία αυτή εστιάζουμε σε αυτές ακριβώς τις καταλήξεις οι οποίες μπορούν να μας δώσουν μια καθαρή εικόνα για το είδος του λόγου που χρησιμοποιείται σε κάθε περίπτωση.

Οι διαφορετικές καταλήξεις οφείλονται στις εξελικτικές τάσεις της Ελληνικής Γλώσσας [ΚΜ99]. Μία τέτοια τάση της Δημοτικής είναι να έχει λέξεις που καταλήγουν σε 'ανοικτές' συλλαβές. Π.χ. οι καταλήξεις σε ***–ν*** του 3^{ου} προσώπου πληθυντικού γίνονται ***–νε*** (*3.1*)

έλεγαν (Καθ.) – λέγανε (Δημ.) (3.1)

Μια δεύτερη τάση της Δημοτικής είναι να μετατρέπει τις ομάδες συμφώνων οι οποίες αποτελούνται από δύο τυρβώδη ή εκρηκτικά σύμφωνα σε ομάδες που αποτελούνται από ένα τυρβώδες και ένα εκρηκτικό σύμφωνο. Σε ομάδες συμφώνων οι οποίες περιέχουν το ***σ***, το μη οξύ σύμφωνο που το συνοδεύει μετατρέπεται σε εκρηκτικό. Σε ομάδες συμφώνων με σιωπηρά τυρβώδη σύμφωνα τα οποία δεν περιέχουν το ***σ***, το πρώτο εκ των συμφώνων είναι τυρβώδες και το δεύτερο εκρηκτικό, όπως φαίνεται και στα επόμενα παραδείγματα :

πεισθώ – πειστώ (3.2)
καλυφθώ – καλυφτώ (3.3)
απαλλάχθηκα – απαλλάχτηκα (3.4)

Τρίτον, στη Νέα Ελληνική Γλώσσα υπάρχουν κλάσεις ρημάτων τα οποία είτε ακολουθούν το κλιτικό παράδειγμα των αρχαίων ρημάτων τα οποία περιέχουν ένα ***α*** (*3.5*), ***ο*** (*3.6*) ή ***ε*** (*3.7*) ως θεματικό φωνήεν, είτε επιλέγουν την κατάληξη τους από μια από τις καταλήξεις της Δημοτικής Γλώσσας.

εξαρτάται (Καθ.) – εξαρτιέται (Δημ.) (3.5)
αξιούμε (Καθ.) – αξιώνουμε (Δημ.) (3.6)
θεωρείτο (Καθ.) – θεωρούνταν (Δημ.) (3.7)

Μερικές φορές η Δημοτική χρησιμοποιεί ρηματικές καταλήξεις οι οποίες είναι παρόμοιες αλλά όχι ταυτόσημες με τις καταλήξεις της Καθαρεύουσας (*3.8*) :

δεικνύω (Καθ.) – δείχνω (Δημ.) (3.8)

Τέλος, υπάρχει μια πλειάδα ρημάτων τα οποία είναι φανερό ότι προέρχονται από την Καθαρεύουσα, όμως είτε δεν έχουν αντικατασταθεί στη Δημοτική (*3.9*), είτε το αντίστοιχο ρήμα της Δημοτικής χαρακτηρίζεται ως καθομιλούμενο (*3.10*) :

προΐσταμαι (3.9)
διάκειμαι (Καθ.) – νιώθω (Δημ.) (3.10)

Όλες αυτές οι μορφολογικές αντιθέσεις επεκτείνονται μέσω της κλιτικής φύσης των ελληνικών ρημάτων και επηρεάζουν εκατοντάδες μορφές λέξεων. Μελετώντας την κατανομή των μορφών αυτών, θα δείξουμε ότι τα μορφολογικά χαρακτηριστικά παίζουν ένα σημαντικότατο ρόλο στην αναγνώριση του γλωσσολογικού είδους. Στην συγκεκριμένη εργασία χρησιμοποιήσαμε 230 χαρακτηριστικές καταλήξεις ρημάτων (*Παράρτημα ΙΙΙ*) για να διακρίνουμε τις μορφολογικές πληροφορίες που μας δίνει κάθε κείμενο.

3.4 Μετρήσεις

Η επιλογή των χαρακτηριστικών που θα εξετάσουμε έγινε λαμβάνοντας υπόψη μας παλαιότερες μελέτες, αλλά και με βάση την υπολογιστική πολυπλοκότητα που απαιτούσε η κάθε μέτρηση.

3.4.1 Μετρήσεις Μορφολογικών Χαρακτηριστικών

Τα μορφολογικά χαρακτηριστικά συνίστανται στη μέτρηση των συχνοτήτων των ρηματικών καταλήξεων (*Παράρτημα ΙΙΙ*) οι οποίες εκφράζουν τις διαφορές ανάμεσα στη Καθαρεύουσα και τη Δημοτική. Οι 230 αυτές συχνότητες εμφάνισης μας δίνουν τελικά 14 μεταβλητές, εκ των οποίων οι 12 είναι γραμμικά ανεξάρτητες (*πίνακας*

3.1). Το άθροισμα των λόγιων και δημοτικών καταλήξεων μετρήθηκε για να μας δώσει μια γενικότερη εικόνα, αλλά επειδή είναι γραμμικός συνδυασμός των επιμέρους μετρήσεων δεν χρησιμοποιείται στην στατιστική ανάλυση.

1ο Ενικό Λόγιο	1ο Ενικό Δημοτική
2ο Ενικό Λόγιο	2ο Ενικό Δημοτική
3ο Ενικό Λόγιο	3ο Ενικό Δημοτική
1ο Πληθυντικό Λόγιο	1ο Πληθυντικό Δημοτική
2ο Πληθυντικό Λόγιο	2ο Πληθυντικό Δημοτική
3ο Πληθυντικό Λόγιο	3ο Πληθυντικό Δημοτική
Σύνολο Λόγιων	*Σύνολο Δημοτικής*

Πίνακας 3.1 – Μορφολογικά Χαρακτηριστικά

3.4.2 Μετρήσεις Δομικών Χαρακτηριστικών

Η μέτρηση των δομικών χαρακτηριστικών (μήκος λέξης, μήκος πρότασης κ.τ.λ.) είναι η πιο κλασσική μεθοδολογία στην προσπάθεια διαχωρισμού γλωσσικών ειδών. Στην συγκεκριμένη έρευνα μετρήσαμε τα ακόλουθα :

i. Αριθμός λέξεων και προτάσεων ανά κείμενο.
ii. Αριθμός κομμάτων και λοιπών σημείων στίξης (παρενθέσεις, παύλες).
iii. Μέσο μήκος λέξεων και προτάσεων.
iv. Συχνότητα λέξεων μήκους x [$1 \leq x \leq 30$ γράμματα].
v. Συχνότητα προτάσεων μήκους x [$1 \leq x \leq 150$ λέξεις].

Ο σκοπός μέτρησης των (i,iii,iv,v) είναι προφανής. Ο λόγος που μας οδήγησε να μετρήσουμε και το (ii) ήταν ότι η έντονη χρήση σημείων στίξης συχνά φανερώνει την ύπαρξη υποπροτάσεων μέσα σε μια πρόταση. Συνεπώς είναι κι αυτή μια ένδειξη διαφορετικού ύφους σε ένα κείμενο.

Παρόμοια δομικά χαρακτηριστικά μετρήθηκαν και από τον *Karlgren* [Kar00], με την προσθήκη των μεγάλων λέξεων (οι οποίες είναι χαρακτηριστικές κάποιων ειδών του λόγου στην Αγγλική Γλώσσα) και των αριθμών. Στην εργασία μας τα αριθμητικά συμπεριλήφθηκαν στην μέτρηση των λέξεων.

Ο *πίνακας 3.2* μας δίνει μια εικόνα για τις μέσες τιμές και το εύρος τιμών κάποιων από τις παραπάνω μεταβλητές στα κείμενα που μελετήσαμε :

	Πολιτικός Λόγος	*Ιστορικός Λόγος*	*Λογοτεχνία*
Λέξεις	509.917	360.933	363.873
Προτάσεις	29.269	17.653	24.831
Μήκος Λέξεων	5,3 (5,2 – 5,5)	5,7 (5,4 – 6,0)	4,7 (4,5 – 5,0)
Μήκος Προτάσεων	17,4 (15,5 – 20,1)	20,4 (13,2 – 36,6)	14,7 (7,5 – 29,3)

Πίνακας 3.2 – Απόλυτες τιμές για τον αριθμό λέξεων και προτάσεων και μέσες τιμές (σε παρένθεση τα εύρη τιμών) για τα μήκη λέξεων και προτάσεων ανά κείμενο.

3.4.3 Μέρη του Λόγου

Το τελευταίο μέρος των μετρήσεων αφορούσε τις μετρήσεις για τα μέρη του λόγου. Πιο συγκεκριμένα, μέσω των αποτελεσμάτων του *tagger* μπορέσαμε να διαχωρίσουμε τις λέξεις των κειμένων στα 11 μέρη του λόγου (*πίνακας 3.3*), προσθέτοντας έτσι άλλες 11 μεταβλητές στο σύστημα.

1-Επίθετα	7-Αντωνυμίες
2-Προθέσεις	8-Αριθμητικά
3-Επιρρήματα	9-Μόρια
4-Άρθρα	10-Ρήματα
5-Σύνδεσμοι	11-Λοιπά
6-Ουσιαστικά	

Πίνακας 3.3 – Μέρη του Λόγου

Οι μετρήσεις αυτές εκφράστηκαν σαν ποσοστά επί τοις εκατό, κατά συνέπεια έχουμε 10 γραμμικά ανεξάρτητες μεταβλητές.

3.5 Αποτελέσματα

Προτού προχωρήσουμε στην επεξεργασία των δεδομένων και επειδή τα κείμενα ήταν διαφόρων μεγεθών, κανονικοποιήσαμε τις μεταβλητές έτσι ώστε όλες οι τιμές να αντιστοιχούν σε κείμενα μεγέθους 100.000 λέξεων. Όπου ήταν δυνατό (π.χ. μέρη του λόγου), εκφράσαμε τις μεταβλητές σε ποσοστά επί τοις εκατό.

Τα αποτελέσματα των μετρήσεων που κάναμε δίνονται συνοπτικά στους *πίνακες 3.4 – 3.7* με την μορφή μέσων όρων και τυπικών αποκλίσεων για κάθε ένα από τα μετρηθέντα χαρακτηριστικά. Ο *πίνακας 3.4* περιέχει τις συχνότητες εμφάνισης των ρηματικών καταλήξεων της Καθαρεύουσας, ενώ ο *πίνακας 3.5* τις αντίστοιχες συχνότητες εμφάνισης των ρηματικών καταλήξεων της Δημοτικής. Μιλάμε φυσικά για τις 230 καταλήξεις του *Παραρτήματος Ι*. Επιπλέον, επειδή η χρήση ενός πίνακα με 230 μεταβλητές θα ήταν υπολογιστικά αδύνατη, αλλά και επειδή τα μερικά αποτελέσματα θα ήταν στατιστικά ασήμαντα, οι καταλήξεις εκφράζονται αθροιστικά για κάθε αριθμό και πρόσωπο.

Παρατηρώντας τους πίνακες αυτούς, μπορούμε να κάνουμε κάποιες αξιοπρόσεκτες παρατηρήσεις [ΤΜΗ^{+}00a] σε σχέση με τα τρία είδη λόγου που μελετάμε :

- Στον πολιτικό λόγο (*πίνακες 3.4* και *3.5*) οι ρηματικές καταλήξεις οι οποίες αντιστοιχούν στην Καθαρεύουσα είναι γενικότερα πιο συχνές από αυτές της Δημοτικής, σε σύγκριση με τα άλλα δύο είδη λόγου.

- Παραδόξως όμως, στην περίπτωση του δευτέρου προσώπου τόσο του ενικού όσο και του πληθυντικού στον πολιτικό λόγο η συχνότητα των καταλήξεων της Δημοτικής είναι συγκριτικά υψηλότερη από αυτή των καταλήξεων Καθαρεύουσας. Αυτό είναι περισσότερο εμφανές στη περίπτωση του δευτέρου προσώπου ενικού, όπου η συχνότητα εμφάνισης των καταλήξεων Δημοτικής είναι πάνω από 50 φορές υψηλότερη από αυτή της Καθαρεύουσας. Αντιθέτως, παρατηρούμε το αντίστροφο στο πρώτο και τρίτο πρόσωπο Ενικού και Πληθυντικού, όπου οι συχνότητες εμφάνισης καταλήξεων της Καθαρεύουσας είναι μεγαλύτερες.

- Στον ιστορικό λόγο η πλειοψηφία των ρηματικών καταλήξεων αντιστοιχούν στο τρίτο πρόσωπο (ενικού ή πληθυντικού), γεγονός το οποίο υποδηλώνει την επικράτηση ενός αφηγηματικού ύφους.

- Το τρίτο πρόσωπο είναι επίσης πιο συχνά χρησιμοποιούμενο και στη λογοτεχνία, ωστόσο σε μικρότερο βαθμό σε σχέση με τον ιστορικό λόγο.

	1ο Ενικό	*2ο Ενικό*	*3ο Ενικό*	*1ο Πληθυντικό*	*2ο Πληθυντικό*	*3ο Πληθυντικό*
Πολιτικός Λόγος	166,0 (51,9)	1,0 (1,6)	247,3 (62,1)	308,0 (38,0)	51,8 (20,6)	472,7 (100,8)
Ιστορικός Λόγος	0,7 (2,9)	0,2 (1,1)	208,3 (129,8)	43,8 (51,0)	0,4 (1,9)	343,9 (140,3)
Λογοτεχνία	157,2 (108,5)	33,5 (41,3)	55,7 (56,6)	103,8 (182,0)	2,0 (9,1)	123,3 (67,3)

Πίνακας 3.4 – Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) των ρηματικών καταλήξεων της Καθαρεύουσας, κανονικοποιημένες σε 100.000 λέξεις, για καθένα από τα τρία είδη λόγου.

	1ο Ενικό	*2ο Ενικό*	*3ο Ενικό*	*1ο Πληθυντικό*	*2ο Πληθυντικό*	*3ο Πληθυντικό*
Πολιτικός Λόγος	50,4 (32,9)	67,0 (16,0)	180,2 (32,4)	82,2 (28,9)	48,6 (26,9)	77,9 (16,6)
Ιστορικός Λόγος	0,0 (0,0)	37,7 (53,3)	228,3 (102,0)	8,9 (16,1)	0,0 (0,0)	153,0 (72,6)
Λογοτεχνία	69,0 (57,2)	85,9 (65,8)	330,8 (154,2)	49,3 (74,3)	37,3 (64,2)	67,2 (45,6)

Πίνακας 3.5 – Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) των ρηματικών καταλήξεων της Δημοτικής, κανονικοποιημένες σε 100.000 λέξεις, για καθένα από τα τρία είδη λόγου.

	Προτάσεις	*Κόμματα*	*Παρενθέσεις*	*Παύλες*	*Καταλήξεις Καθαρεύουσας*	*Καταλήξεις Δημοτικής*
Πολιτικός Λόγος	5772,4 (463,5)	6726,0 (442,0)	56,2 (49,5)	570,1 (69,8)	1252,5 (137,7)	506,3 (49,5)
Ιστορικός Λόγος	4678,2 (1164,1)	6834,9 (1422,4)	849,1 (463,1)	495,2 (254,5)	597,4 (221,5)	427,9 (161,1)
Λογοτεχνία	8649,0 (2975,0)	7728,7 (1435,8)	40,7 (95,4)	1793,8 (1386,2)	475,5 (294,4)	639,5 (223,5)

Πίνακας 3.6 – Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) για τα δομικά χαρακτηριστικά κανονικοποιημένα σε 100.000 λέξεις, για καθένα από τα τρία είδη λόγου. Οι δύο τελευταίες στήλες καταγράφουν τα συγκεντρωτικά αποτελέσματα των πινάκων 3.4 και 3.5. Επειδή δεν είναι γραμμικά ανεξάρτητες μεταβλητές δεν λαμβάνουν μέρος στην στατιστική επεξεργασία και απλώς τα αναφέρουμε για λόγους πληρότητας.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Πολιτικός Λόγος	7,91	6,52	5,10	12,85	6,73	19,70	5,18	7,31	14,78	13,03	0,50
Ιστορικός Λόγος	11,15	7,78	5,15	16,22	5,18	24,02	1,72	4,02	15,55	8,46	0,76
Λογοτεχνία	6,40	5,94	6,92	11,80	6,31	16,35	4,84	10,01	14,39	16,61	0,43

Πίνακας 3.7 – Επί τοις εκατό (%) αναλογία για τα 11 μέρη του λόγου, για καθένα από τα τρία είδη λόγου. Οι αντιστοιχίες (1=επίθετα κ.τ.λ.) δίνονται στον πίνακα 3.3.

- Η λογοτεχνία παρουσιάζει μεγαλύτερο αριθμό προτάσεων και υπό-προτάσεων (αν λάβουμε υπόψη μας συλλογικά τα κόμματα, τις παρενθέσεις και τις παύλες) σε σχέση με τα άλλα δύο είδη λόγου, τα οποία έχουν υψηλό βαθμό ομοιότητας σε αυτά τα χαρακτηριστικά.

- Εξετάζοντας τις τυπικές αποκλίσεις, παρατηρούμε ότι η λογοτεχνία έχει τη μεγαλύτερη διακύμανση σε όλα σχεδόν τα χαρακτηριστικά. Ακολουθεί σε διακύμανση ο ιστορικός λόγος και τελευταίος είναι ο πολιτικός λόγος ο οποίος έχει τη μικρότερη διακύμανση. Άρα κατ' αρχάς μπορούμε να βγάλουμε το συμπέρασμα ότι ο πολιτικός λόγος είναι πιο συμπαγής και ίσως να έχουμε να κάνουμε και με μια ειδική μορφή λόγου που χρησιμοποιείται αποκλειστικά από τους πολιτικούς.

Στις επόμενες παραγράφους θα κάνουμε μια πιο εκτενή ανάλυση των δεδομένων ακολουθώντας στατιστικές και άλλες μεθόδους. Πρέπει να σημειώσουμε ότι στην ανάλυση αυτή χρησιμοποιήσαμε ένα υποσύνολο των προαναφερθέντων χαρακτηριστικών, με απώτερο σκοπό να πετύχουμε ένα ικανοποιητικό ποσοστό αναγνώρισης με όσο το δυνατόν λιγότερες μεταβλητές.

3.6 Επεξεργασία

Ακολουθεί η επεξεργασία των παραπάνω μετρήσεων με τη χρήση στατιστικής μεθόδου αλλά και με τη βοήθεια νευρωνικών δικτύων.

3.6.1 Στατιστική Επεξεργασία

Προκειμένου να συγκρίνουμε τα τρία είδη λόγου επιλέξαμε τη μέθοδο του διαχωρισμού σε ομάδες (*cluster analysis*). Πριν από οποιαδήποτε προσπάθεια διαχωρισμού κανονικοποιήσαμε όλες τις παραμέτρους, έτσι ώστε όλες να κυμαίνονται στο ίδιο εύρος τιμών, να απαλλαγούμε από την επίδραση των μονάδων μετρήσεων και κάθε μεταβλητή να συνεισφέρει ομοιόμορφα στη διαμόρφωση του τελικού αποτελέσματος. Επιπλέον, χρησιμοποιήσαμε όλα τα δυνατά κριτήρια

(*nearest, furthest, median, centroid* και *average*) για την ανάλυση χωρίς αρχικοποίηση (*non-seeded analysis*). Όπως φάνηκε όλες οι παραπάνω μέθοδοι ομαδοποίησης παρήγαγαν παρόμοια αποτελέσματα και κατά συνέπεια δεν θα αναφερθούμε ξεχωριστά στη κάθε μία.

Στην αρχή πειραματιστήκαμε με την ομαδοποίηση χωρίς αρχικοποίηση (*non-seeded*) προσπαθώντας να βρούμε ποιες ομάδες υπάρχουν στο σώμα κειμένων. Όταν χρησιμοποιήσαμε έξι ομάδες, οι πέντε από αυτές περιείχαν ένα ή το πολύ δύο κείμενα από τη Λογοτεχνία ενώ η έκτη περιείχε τα υπόλοιπα Λογοτεχνικά κείμενα μαζί με όλα τα Πολιτικά και τα Ιστορικά κείμενα.

Στη συνέχεια επιλέξαμε τη χρησιμοποίηση 10 ομάδων για την ταξινόμηση των κειμένων. Το πείραμα αυτό κατέδειξε ότι η Λογοτεχνία είναι το είδος λόγου με τη μεγαλύτερη μεταβλητότητα, μιας και τα κείμενα της κατέλαβαν εννέα από τις δέκα διαθέσιμες ομάδες. Από την άλλη πλευρά, όλα τα κείμενα του Πολιτικού Λόγου ξεχώρισαν καταλαμβάνοντας μια ομάδα, η οποία δεν περιείχε άλλα κείμενα. Όσον αφορά τον Ιστορικό Λόγο, όλα τα κείμενα του κατέλαβαν μια ομάδα, η οποία όμως περιείχε και κείμενα της Λογοτεχνίας.

Από τα πρώτα αυτά αποτελέσματα έχουμε μια καθαρή ένδειξη ότι τα κείμενα της Βουλής αποτελούν μια συμπαγή ομάδα στο χώρο προτύπων που ορίζεται από τον πίνακα χαρακτηριστικών, κάτι το οποίο επιβεβαιώνει και την παρατήρηση της *παραγράφου 3.5*. Ο Ιστορικός Λόγος ήταν επίσης σχετικά συμπαγής, ενώ η Λογοτεχνία περιείχε αρκετά κείμενα απομακρυσμένα από τον βασικό πυρήνα συγκέντρωσης των λογοτεχνικών κειμένων. Για να αποτιμήσουμε καλύτερα αυτές τις παρατηρήσεις προχωρήσαμε επιπλέον στην ανάλυση σε ομάδες (*cluster analysis*) των κειμένων κάθε είδους λόγου. Η ανάλυση αυτή έδειξε ότι ο Πολιτικός και Ιστορικός Λόγος σχηματίζουν αρκετά συμπαγείς ομάδες, σε αντίθεση με τη Λογοτεχνία, πέντε μέλη της οποίας ήταν αρκετά διαφορετικά σε σχέση με τα υπόλοιπα.

Κατόπιν, προχωρήσαμε στην ανάλυση με αρχικοποίηση (*seeded analysis*). Αρχικά, χρησιμοποιήσαμε έξι ομάδες, ήτοι δύο ομάδες ανά είδος λόγου. Ακολούθως, ο αριθμός των ομάδων μειώθηκε σε τέσσερις, με τον Πολιτικό και τον Ιστορικό Λόγο να καταλαμβάνουν από μία και τη Λογοτεχνία τις άλλες δύο. Τέλος, οι ομάδες έγιναν

τρεις, με κάθε είδος λόγου να καταλαμβάνει από μία. Τα πειράματα αυτά έγιναν με τη χρήση των εξής πινάκων δεδομένων :

i. Πίνακας 16 μεταβλητών αποτελούμενος από τις μετρήσεις για τις ρηματικές καταλήξεις όλων των προσώπων και αριθμών Δημοτικής και Καθαρεύουσας καθώς και των προτάσεων, κομμάτων, παρενθέσεων και παυλών.

ii. Πίνακας 14 μεταβλητών αποτελούμενος από τις μετρήσεις για τις ρηματικές καταλήξεις όλων των προσώπων και αριθμών Δημοτικής και Καθαρεύουσας καθώς και των προτάσεων και κομμάτων.

iii. Πίνακας 12 μεταβλητών αποτελούμενος από τις μετρήσεις για τις ρηματικές καταλήξεις όλων των προσώπων και αριθμών Δημοτικής και Καθαρεύουσας.

Οι αρχικές τιμές (*seeds*) για τον Πολιτικό και τον Ιστορικό Λόγο επιλέχθηκαν τυχαία. Η μία από τις αρχικές τιμές για τη Λογοτεχνία επιλέχτηκε έτσι ώστε να μην είναι ένα από τα πέντε τα οποία ξεφεύγουν από τα υπόλοιπα. Τα αποτελέσματα για τους διάφορους πίνακες σε συνάρτηση με τον αριθμό των ομάδων δίνονται στον *πίνακα 3.8* που ακολουθεί :

	16 μεταβλητές	*14 μεταβλητές*	*12 μεταβλητές*
6 ομάδες	98,5 %	92,7 %	95,6 %
4 ομάδες	98,5 %	94,1 %	97,1 %
3 ομάδες	97,1 %	92,7 %	95,6 %

Πίνακας 3.8 – Ακρίβεια ομαδοποίησης με αρχικοποίηση (seeded) σαν συνάρτηση του αριθμού των ομάδων και των μεταβλητών.

Τα αποτελέσματα που καταγράφονται στον *πίνακα 3.8* δείχνουν ότι η βέλτιστη ομαδοποίηση επιτυγχάνεται όταν χρησιμοποιούμε και τις 16 μεταβλητές, δηλαδή και τα δομικά αλλά και τα μορφολογικά χαρακτηριστικά. Παρατηρούμε ότι εάν αφαιρέσουμε μέρος των δομικών χαρακτηριστικών (καταλήγοντας σε πίνακα 14 μεταβλητών) το ποσοστό αναγνώρισης πέφτει σε 93-94%. Στη συγκεκριμένη περίπτωση αποδεικνύεται ότι είναι καλύτερα να αφαιρέσουμε όλες τις σχετιζόμενες με δομικά χαρακτηριστικά μεταβλητές, αφού το ποσοστό αναγνώρισης του πίνακα 12 μεταβλητών είναι μεγαλύτερο αυτού των 14 μεταβλητών.

Σύμφωνα με τον *πίνακα 3.8*, ακόμη και εάν χρησιμοποιήσουμε τρεις ομάδες το αποτέλεσμα είναι παραπάνω από ικανοποιητικό, με μόνο ελάχιστες λανθασμένες ταξινομήσεις. Μάλιστα, τα καλύτερα αποτελέσματα λαμβάνονται με τη χρήση τεσσάρων ομάδων, κάτι το οποίο δείχνει ότι αφαιρώντας τις περιττές ομάδες από τα καλά ορισμένα είδη λόγου (Πολιτικό και Ιστορικό) αυξάνεται το ποσοστό αναγνώρισης.

Στη συνέχεια, επαυξάνουμε τον πίνακα δεδομένων με τις πληροφορίες για τα μέρη του λόγου, έχοντας έτσι :

i. Πίνακας 27 μεταβλητών αποτελούμενος από τις μετρήσεις για τις ρηματικές καταλήξεις όλων των προσώπων και αριθμών Δημοτικής και Καθαρεύουσας των προτάσεων, κομμάτων, παρενθέσεων παυλών και των μερών του λόγου.

ii. Πίνακας 15 μεταβλητών αποτελούμενος από τις μετρήσεις των προτάσεων, κομμάτων, παρενθέσεων, παυλών καθώς και των μερών του λόγου.

Πραγματοποιήσαμε μια καινούρια σειρά πειραμάτων χρησιμοποιώντας τη μέθοδο του Ward με τη τετραγωνισμένη Ευκλείδεια απόσταση [TMH^{+}00b] για να ομαδοποιήσουμε τα δεδομένα με τη μέθοδο χωρίς αρχικοποίηση (*non-seeded*). Τα αποτελέσματα για τους διάφορους πίνακες σε συνάρτηση με τον αριθμό των ομάδων δίνονται στον *πίνακα 3.9* που ακολουθεί :

	12 μεταβλητές	*16 μεταβλητές*	*27 μεταβλητές*	*15 μεταβλητές*
6 ομάδες	95,5 %	100,0 %	100,0 %	100,0 %
5 ομάδες	95,5 %	100,0 %	100,0 %	89,6 %
4 ομάδες	94,1 %	98,5 %	100,0 %	83,4 %
3 ομάδες	94,1 %	98,5 %	98,5 %	83,4 %

Πίνακας 3.9 – Ακρίβεια ομαδοποίησης χωρίς αρχικοποίηση (*non-seeded*) σαν συνάρτηση του αριθμού των ομάδων και των μεταβλητών.

Είναι φανερό ότι η προσθήκη των μετρήσεων για τα μέρη του λόγου βελτιώνει την απόδοση του συστήματος, η οποία φτάνει ακόμη και το 100%. Ακόμη και αν χρησιμοποιούσαμε μόνο τα μέρη του λόγου θα είχαμε αρκετά καλά

αποτελέσματα αν και βέβαια όχι τόσο καλά όσο χρησιμοποιώντας τα μορφολογικά και δομικά χαρακτηριστικά.

3.6.2 Νευρωνικά Δίκτυα

Η επεξεργασία των αποτελεσμάτων μέσω νευρωνικών δικτύων και πιο συγκεκριμένα του SOM έγινε στο περιβάλλον της Matlab. Πιο συγκεκριμένα, η επεξεργασία ακολούθησε τα εξής βήματα: Ανάγνωση και κανονικοποίηση των δεδομένων μέσω ενός γραμμικού μετασχηματισμού, αρχικοποίηση (κανονικά και τυχαία) και εκπαίδευση του χάρτη, και τέλος αυτόματη απόδοση ετικετών στα στοιχεία του χάρτη και απεικόνιση του. Το σετ εντολών το οποίο χρησιμοποιήσαμε προκειμένου να λάβουμε τους χάρτες δίνεται στο *παράρτημα IV*.

Χρησιμοποιήσαμε χάρτες μεγέθους 8x4, 6x3 και 3x2. Για κάθε έναν από τους χάρτες αυτούς το SOM διαχωρίζει πλήρως και επιτυχώς τα τρία είδη λόγου. Στον *πίνακα 3.10* που ακολουθεί αναφέρουμε σε πόσους νευρώνες διασπάται το κάθε είδος λόγου για καθέναν από τους ανωτέρω χάρτες.

	Πολιτικός Λόγος	*Ιστορικός Λόγος*	*Λογοτεχνία*	*Κενά*
8x4	2	11	7	12
6x3	1	5	5	7
3x2	1	2	2	1

Πίνακας 3.10 – Κατανομή των ειδών λόγου σε νευρώνες ανά χάρτη.

Παρατηρούμε πάλι την τάση του Πολιτικού Λόγου να συγκεντρωθεί σε όσο το δυνατόν λιγότερα κελιά (τελικά σε ένα), κάτι το οποίο αποδεικνύει το πόσο συμπαγής είναι. Ο Ιστορικός Λόγος και η Λογοτεχνία από την άλλη, εξαπλώνονται αρκετά, αλλά πάντα σε γειτονικά κελιά. Τα ίδια αποτελέσματα πήραμε και στην περίπτωση που η αρχικοποίηση του χάρτη έγινε με τυχαίο τρόπο χρησιμοποιώντας τη συνάρτηση som_randinit.

Στο *σχήμα 3.1*, το οποίο ακολουθεί, παραθέτουμε ενδεικτικά τον χάρτη μεγέθους 6x3 για την παραπάνω ανάλυση. Ο U-matrix (*Unified distance matrix*)

οπτικοποιεί σε δύο διαστάσεις τις αποστάσεις μεταξύ των γειτονικών κελιών, και μας βοηθά να διακρίνουμε την δομή του χάρτη. Υψηλές τιμές του U-matrix καταδεικνύουν το όριο της περιοχής μιας ομάδας, ενώ χαμηλές ομοιόμορφες τιμές υποδηλώνουν την ύπαρξη μιας ή περισσοτέρων ομάδων. Μπορούμε να δούμε τον U-matrix και τα επίπεδα τιμών των 29 συνιστωσών (μεταβλητών) οι οποίες συντελούν στον τελικό διαχωρισμό των ειδών του λόγου. Αναλυτικά οι μεταβλητές αυτές δίνονται στο *Παράρτημα V*. Όπως μπορούμε να διαπιστώσουμε, ο χάρτης χρησιμοποιεί θερμοκρασιακή κλίμακα, όπου τα κελιά που απεικονίζονται με βαθύ κόκκινο χρώμα υποδεικνύουν υψηλές τιμές μεταβλητών ενώ τα κελιά που απεικονίζονται με ελαφρύ θαλασσί χρώμα αντιστοιχούν σε χαμηλές τιμές των μεταβλητών. Βέβαια, τα κελιά του *σχήματος 3.1* αντιστοιχούν πλήρως στα κελιά του χάρτη του *σχήματος 3.2* (χάρτης Labels).

Ο U-matrix φαίνεται ακόμη καλύτερα στο *σχήμα 3.2* μαζί με την τελική κατανομή των κειμένων τα οποία αποτελούν το σώμα κειμένων Ι (με Μ συμβολίζουμε τα Ιστορικά κείμενα, με Λ τα λογοτεχνικά και με Β τα πολιτικά κείμενα).

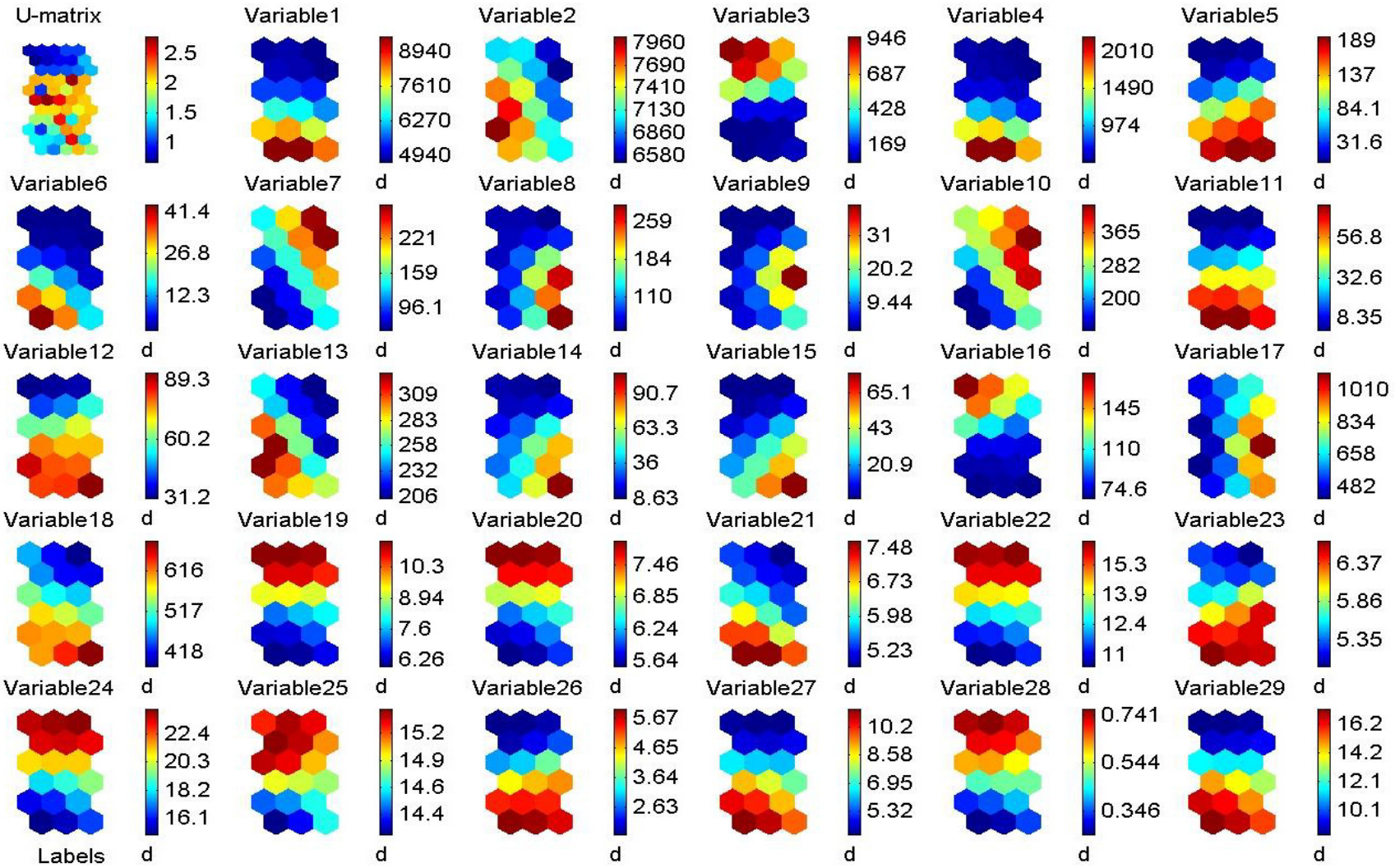
3.7 Συμπεράσματα

Μελετώντας τα συγκεκριμένα κείμενα καταλήγουμε στο συμπέρασμα ότι τα τρία είδη λόγου έχουν διαφορετικά χαρακτηριστικά. Τα κείμενα του Πολιτικού Λόγου σχηματίζουν μια καλά ορισμένη κλάση στο χώρο προτύπων, όπως και ο Ιστορικός Λόγος, αν και σε μικρότερο βαθμό. Αντίθετα η Λογοτεχνία σχηματίζει μια λιγότερο συμπαγή κλάση και για την επιτυχή αναγνώρισή των μελών της απαιτείται η χρήση μεγαλύτερου αριθμού αρχικών τιμών και ομάδων για να καλυφθεί ικανοποιητικά ο χώρος προτύπων.

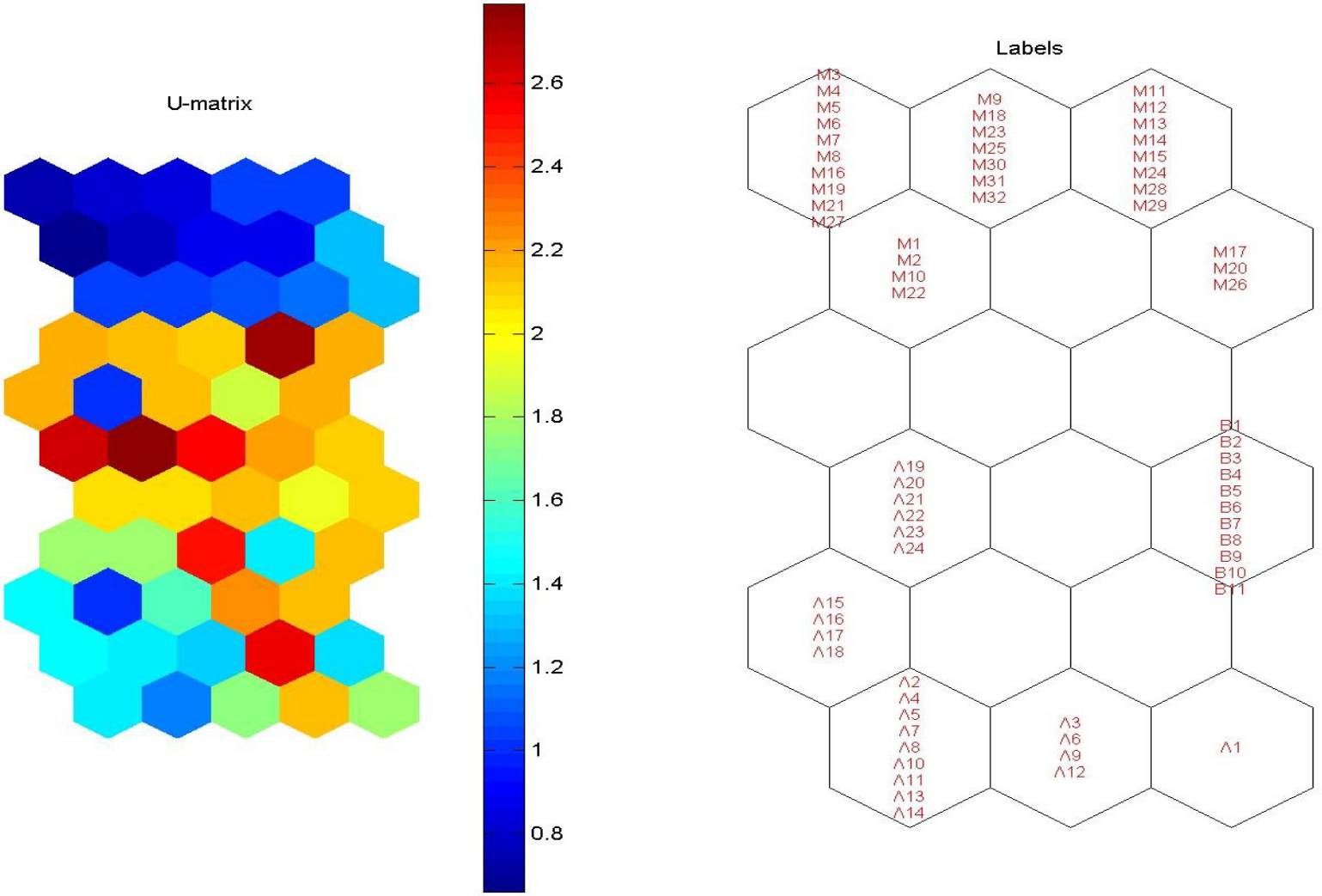
Εν τούτοις, όλες οι στατιστικές μέθοδοι ομαδοποίησης που χρησιμοποιήσαμε είχαν απόδοση μεγαλύτερη του 98% και συνεπώς η συγκεκριμένη μεθοδολογία μπορεί με μεγάλη ακρίβεια να αναγνωρίσει το είδος λόγου ενός τυχαίου κειμένου που ανήκει σε ένα από τα χρησιμοποιούμενα είδη λόγου. Βέβαια, τα αποτελέσματα αυτά

ελήφθησαν με ένα περιορισμένο αριθμό ειδών του λόγου και κειμένων, όμως τα αποτελέσματα είναι τόσο ακριβή ώστε να μπορούμε να ισχυριστούμε ότι αυτή η μέθοδος θα οδηγεί σε καλά αποτελέσματα τόσο για μεγαλύτερα σώματα κειμένων όσο και για περισσότερα είδη λόγου.

Επιπλέον, έχουμε το πλεονέκτημα ότι τα αποτελέσματα αυτά επιβεβαιώθηκαν στην πράξη και με τη μέθοδο των νευρωνικών δικτύων.


U-matrix
2.5
2
1.5
1
Variable1
8940
7610
6270
4940
Variable2
7960
7690
7410
7130
6860
6580
Variable3
946
687
428
169
Variable4
2010
1490
974
Variable5
189
137
84.1
31.6
Variable6
41.4
26.8
12.3
Variable7
221
159
96.1
Variable8
259
184
110
Variable9
31
20.2
9.44
Variable10
365
282
200
Variable11
56.8
32.6
8.35
Variable12
89.3
60.2
31.2
Variable13
309
283
258
232
206
Variable14
90.7
63.3
36
8.63
Variable15
65.1
43
20.9
Variable16
145
110
74.6
Variable17
1010
834
658
482
Variable18
616
517
418
Variable19
10.3
8.94
7.6
6.26
Variable20
7.46
6.85
6.24
5.64
Variable21
7.48
6.73
5.98
5.23
Variable22
15.3
13.9
12.4
11
Variable23
6.37
5.86
5.35
Variable24
22.4
20.3
18.2
16.1
Variable25
15.2
14.9
14.6
14.4
Variable26
5.67
4.65
3.64
2.63
Variable27
10.2
8.58
6.95
5.32
Variable28
0.741
0.544
0.346
Variable29
16.2
14.2
12.1
10.1
Labels
d

Σχήμα 3.1


U-matrix
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
Labels
M3
M4
M5
M6
M7
M8
M16
M19
M21
M27
M9
M18
M23
M25
M30
M31
M32
M11
M12
M13
M14
M15
M24
M28
M29
M1
M2
M10
M22
M17
M20
M26
Λ19
Λ20
Λ21
Λ22
Λ23
Λ24
B1
B2
B3
B4
B5
B6
B7
B8
B9
B10
B11
Λ15
Λ16
Λ17
Λ18
Λ2
Λ4
Λ5
Λ7
Λ8
Λ10
Λ11
Λ13
Λ14
Λ3
Λ6
Λ9
Λ12
Λ1

Σχήμα 3.2

ΚΕΦΑΛΑΙΟ 4
ΔΙΑΧΩΡΙΣΜΟΣ ΟΜΙΛΗΤΩΝ

4.1 Εισαγωγή

Στο τέταρτο κεφάλαιο επικεντρώνουμε την προσοχή μας στον δεύτερο στόχο αυτής της εργασίας, ο οποίος είναι ο διαχωρισμός των κειμένων των πέντε ομιλητών της Βουλής. Ακολουθούν οι πληροφορίες για τα κείμενα που μελετήσαμε, οι μετρήσεις που κάναμε, η επεξεργασία τους και τέλος κάποια συμπεράσματα για την ακρίβεια της μεθόδου που ακολουθήσαμε.

4.2 Κείμενα

Επειδή, όπως είναι φανερό (*πίνακας 1.2*), το αρχικό μας δείγμα δεν είναι σταθμισμένο (κυρίως λόγω της ύπαρξης μεγάλου αριθμού κειμένων του ομιλητή Α), αναγκαστήκαμε κατά την διαδικασία της επεξεργασίας των μετρήσεων να επιλέξουμε κάποια κείμενα ώστε η παρουσία των ομιλητών να είναι ισοβαρής μέσα στο όλο δείγμα. Γι' αυτό, εκτός από το αρχικό σώμα κειμένων ΙΙ, χρησιμοποιήσαμε για τα πειράματά μας και ένα υποσύνολο του, το σώμα κειμένων ΙΙΙ. Το σώμα κειμένων ΙΙΙ αποτελείται από 127 πρωτολογίες που εκφωνήθηκαν τα έτη 1999-2000 και πιο συγκεκριμένα :

Σώμα Κειμένων ΙΙΙ	*1999-2000*	*Λέξεις*
Ομιλητής Α	36	90090
Ομιλητής Β	16	73988
Ομιλητής Γ	30	64905
Ομιλητής Δ	21	60071
Ομιλητής Ε	24	75550
Σύνολο	**127**	**364604**

Πίνακας 4.1 – Το σώμα κειμένων ΙΙΙ

Μάλιστα, οι πρωτολογίες αυτές επιλέχθηκαν να προέρχονται από συνεδριάσεις στις οποίες τουλάχιστον δύο από τους πέντε ομιλητές έλαβαν το λόγο. Όπως θα φανεί αργότερα το σώμα κειμένων ΙΙΙ είναι σαφώς πιο ισορροπημένο (*balanced*) και με μικρότερες διακυμάνσεις.

4.3 Μετρήσεις

Για το δεύτερο αυτό σκέλος της εργασίας εμπλουτίσαμε τις μετρήσεις μας με κάποιες παραπάνω μεταβλητές.

4.3.1 Μέτρηση Μορφολογικών Χαρακτηριστικών

Εκτός από τα μορφολογικά χαρακτηριστικά τα οποία αναφέρονται στον *πίνακα 3.1* προσθέσαμε ακόμη τα εξής :

i. Την χρήση εσωτερικών αυξήσεων (κατηγορίες *α* και *β*) στους παρελθοντικούς χρόνους των ρημάτων[2] (δύο επιπλέον μεταβλητές). Η μέτρηση αυτή είναι και η μόνη η οποία έγινε χειρωνακτικά από την γλωσσολογική ομάδα, αλλά επιλέξαμε να προχωρήσουμε στην πραγματοποίηση της για να δούμε αν όντως είναι σημαντική στο διαχωρισμό των ομιλητών.

   Οι ρηματικοί τύποι που απαρτίζουν την κατηγορία *α* είναι τύποι που έχουν αύξηση (συλλαβική, χρονική). Αυτή η αύξηση είναι :

   - εσωτερική, π.χ. απ***έ***φυγε, παρ***ή***λασαν
   - εξωτερική, π.χ. ***έ***δωσαν, ***ή***θελαν

[2] Για τις ανάγκες των πειραμάτων μας, μετρήσαμε τα ρήματα τα οποία βρίσκονται σε παρελθοντικούς χρόνους. Πήραμε δηλαδή τη λίστα των λημμάτων που ο tagger είχε σημειώσει ως ρήματα (tag : Vb) και από αυτήν εξάγαμε όσα βρίσκονται σε παρελθοντικό χρόνο (tag : Pa). Στη συνέχεια έγινε χειρωνακτική επεξεργασία των ρημάτων αυτών για την εξαγωγή κάποιων συμπερασμάτων. Μέσω αυτής της επεξεργασίας που έγινε από την γλωσσολογική ομάδα, μετρήσαμε και το ποσοστό ακρίβειας του tagger στη συγκεκριμένη εργασία, το οποίο ήταν 98,15% (98.142 σωστά παρελθοντικές ρηματικές μορφές σε σύνολο 99.989 που μας έδωσε ο tagger).

Οι ρηματικοί τύποι της κατηγορίας *β* είναι αναύξητοι, δεν έχουν δηλαδή αύξηση, π.χ. κλάψανε, απάντησε.

Στόχος του εντοπισμού των εν λόγω ρηματικών τύπων είναι να δούμε τη χρήση της αύξησης στην παρούσα φάση της Ελληνικής, προκειμένου να βγουν και συμπεράσματα για την ποικιλία Καθαρεύουσας και Δημοτικής.

Η χρονική αύξηση (παρ***ή***λασαν, απ***ή***ντησε) είναι μάλλον σπάνια, ενώ η συλλαβική αύξηση, το –***ε***-, είναι η κατά κόρον χρησιμοποιούμενη. Υποστηρίζεται [Μπα72] ότι η αύξηση αυτή εμφανίζεται ως φορέας του τόνου. Για παράδειγμα, έχουμε τους παρελθοντικούς τύπους "γράψαμε", "γράψατε", χωρίς αύξηση, αλλά παράλληλα υπάρχουν οι τύποι "έγραψα", "έγραψες", κ.ο.κ., όπου η αύξηση εμφανίζεται, για να φέρει τον τόνο.

Με τις κατηγορίες α και β προσπαθήσαμε να δούμε πότε η αύξηση είναι απαραίτητη, πότε είναι δυνατός ο εναλλακτικός τύπος και πότε η αύξηση είναι εντελώς περιττή. Για παράδειγμα :

- ενδι***έ***φερε – ενδι***ά***φερε : Εδώ η αύξηση είναι απαραίτητη και η απουσία της έχει ως αποτέλεσμα αντιγραμματικό τύπο.
- έδωσαν – δώσανε : Εδώ η αύξηση, παρόλο που φέρει τον τόνο, δεν είναι άκρως απαραίτητη, αφού είναι δυνατός και ο αναύξητος τύπος.
- δι***ε***τέθη – διατέθηκε : Εδώ η αύξηση δεν είναι απαραίτητη, καθόσον μάλιστα δεν φέρει τον τόνο. Η χρήση της έχει ως αποτέλεσμα ένα ρηματικό τύπο της Καθαρεύουσας.

ii. Το πρόσωπο και τον αριθμό των ρηματικών μορφών (*πίνακας 4.2*), προσθέτοντας έτσι άλλες 6 μεταβλητές στο σύστημα μας.

1ο Πρόσωπο Ενικό	1ο Πρόσωπο Πληθυντικό
2ο Πρόσωπο Ενικό	2ο Πρόσωπο Πληθυντικό
3ο Πρόσωπο Ενικό	3ο Πρόσωπο Πληθυντικό

Πίνακας 4.2 – Επιπλέον Μορφολογικά Χαρακτηριστικά

4.3.2 Μέτρηση Δομικών Χαρακτηριστικών

Εκτός των όσων σημείων στίξης αναφέρονται στην *παράγραφο 3.4.2* μετρήσαμε επιπλέον και τα ερωτηματικά, ενώ όπως θα αναφέρουμε και στην επεξεργασία των μετρήσεων συμπτύξαμε τα κόμματα, τις παρενθέσεις και τις παύλες σε μια μεταβλητή κάτω από το όνομα "*σημεία στίξης*". Στον πίνακα 4.3 παραθέτουμε τις μέσες τιμές και το εύρος τιμών για κάποιες από τις μεταβλητές αυτές :

	Ομιλητής Α	*Ομιλητής Β*	*Ομιλητής Γ*	*Ομιλητής Δ*	*Ομιλητής Ε*
Λέξεις	463.680	177.853	241.882	217.305	190.601
Προτάσεις	23.564	10.944	12.832	11.520	10.312
Μήκος Λέξεων	5,3 (4,6 – 5,9)	5,2 (4,7 – 5,6)	5,4 (4,8 – 6,2)	5,6 (4,9 – 6,3)	5,5 (4,8 – 5,9)
Μήκος Προτάσεων	19,7 (8,1 – 40,6)	16,3 (6,8 – 24,5)	18,8 (9,7 – 38,2)	18,9 (8,6 – 34,6)	18,5 (11,8 – 31,8)

Πίνακας 4.3 – Απόλυτες τιμές για τον αριθμό λέξεων και προτάσεων και μέσες τιμές (σε παρένθεση τα εύρη τιμών) για τα μήκη λέξεων και προτάσεων ανά κείμενο.

Επιπλέον αξιοποιήσαμε σε μεγαλύτερο βαθμό τις μετρήσεις για τα μήκη λέξεων και προτάσεων εισάγοντας τις εξής 17 μεταβλητές (*πίνακας 4.4*) :

1-3 Γράμματα ανά Λέξη	1 Λέξη ανά Πρόταση
4-8 Γράμματα ανά Λέξη	2-5 Λέξεις ανά Πρόταση
9-15 Γράμματα ανά Λέξη	6-10 Λέξεις ανά Πρόταση
16-20 Γράμματα ανά Λέξη	11-15 Λέξεις ανά Πρόταση
21-30 Γράμματα ανά Λέξη	16-20 Λέξεις ανά Πρόταση
	21-25 Λέξεις ανά Πρόταση
	26-30 Λέξεις ανά Πρόταση
	31-40 Λέξεις ανά Πρόταση
	41-50 Λέξεις ανά Πρόταση
	51-75 Λέξεις ανά Πρόταση
	76-100 Λέξεις ανά Πρόταση
	101-150 Λέξεις ανά Πρόταση

Πίνακας 4.4 – Επιπλέον Δομικά Χαρακτηριστικά

Επιλέξαμε να ομαδοποιήσουμε τις λέξεις που αποτελούνται από ένα έως τρία γράμματα ακριβώς επειδή εκεί περιλαμβάνονται όλα τα άρθρα καθώς και άλλες

λέξεις οι οποίες είναι αρκετά συχνές (όπως οι *και, να*) και μη χαρακτηριστικές. Επίσης χρησιμοποιήσαμε μια μεταβλητή για τις προτάσεις μήκους μίας λέξης.

4.3.3 Μέρη του Λόγου

Όπως και προηγουμένως μετρήθηκαν πάλι τα 11 μέρη του λόγου (*πίνακας 3.3*) και επιπλέον τα ποσοστά ουσιαστικών και επιθέτων τα οποία βρίσκονται σε γενική πτώση. Η μέτρηση αυτή επιλέχθηκε να πραγματοποιηθεί γιατί η γενική πτώση μας δίνει ένα μέτρο της χρήσης ουσιαστικών και επιθέτων στη θέση των ρημάτων που συνήθως χρησιμοποιούνται στον προφορικό λόγο.

4.3.4 Μέτρηση Λεκτικών Χαρακτηριστικών

Στις μετρήσεις μας, επιλέξαμε να συμπτύξουμε τις λέξεις *ουδείς*, *ουδέποτε* και *ουδαμού* κάτω από μια μεταβλητή, λόγω της εξαιρετικά μικρής συχνότητας εμφάνισης τους. Κλείνοντας αυτή την αναφορά στις αρνητικές λέξεις, πρέπει να πούμε ότι έγιναν επίσης μετρήσεις για τα εξής διγράμματα / τριγράμματα (*πίνακας 4.5*). Οι μετρήσεις αυτές δεν αξιοποιήθηκαν στην συγκεκριμένη έρευνα, είναι όμως σίγουρο ότι θα χρησιμοποιηθούν στο μέλλον.

μη + ουσιαστικό	όχι + ουσιαστικό
μη + επίθετο	όχι + επίθετο
μη + μετοχή	όχι + μετοχή
μη + επίρρημα + επίθετο	όχι + επίρρημα + επίθετο
μη + επίρρημα + μετοχή	όχι + επίρρημα + μετοχή

Πίνακας 4.5 – Διγράμματα / Τριγράμματα των *μη* και *όχι*

Ένα ζευγάρι λέξεων το οποίο επίσης καταδεικνύει τη διαφορά ανάμεσα σε Καθαρεύουσα και Δημοτική είναι οι *οποίος* (*Καθ.*) και *που* (*Δημ.*). Οι λέξεις αυτές μπαίνουν στο σύστημα κάτω από μια μεταβλητή, τον λόγο εμφάνισης *οποίος / που*.

Κατά την διάρκεια των πειραμάτων κάναμε δειγματοληπτικά μετρήσεις για τα πιο συχνά χρησιμοποιούμενα λήμματα σε κάποια κείμενα και τη συχνότητα εμφάνισης τους. Περιληπτικά αναφέρουμε ότι τα πιο συχνά λήμματα στις ομιλίες των πέντε πολιτικών φαίνεται να είναι τα : *ο, και, να, στου, είμαι*. Από αυτές τις λίστες λημμάτων παρατηρήσαμε έντονες διαφορές στη συχνότητα εμφάνισης ορισμένων λημμάτων και αποφασίσαμε να τα μελετήσουμε πιο επισταμένα. Κάποια από αυτά τα λήμματα περιέχονται στον *πίνακα 4.6* ενώ συνολικά χρησιμοποιήσαμε 17 λήμματα τα οποία προσθέτουν στο σύστημα άλλες 17 γραμμικά ανεξάρτητες μεταβλητές.

Εγώ	Λαός
Μπορώ	Θέλω
Κάνω	Υπάρχω

Πίνακας 4.6 – Λήμματα

4.3.5 Άλλες Μετρήσεις

Στα δεδομένα μας προσθέσαμε την πληροφορία για την χρονιά στην οποία έλαβε χώρα η κάθε ομιλία (όπως θα αναφέρουμε αργότερα, φαίνεται να υπάρχει μια μεταβολή των χαρακτηριστικών κάποιων ομιλητών σε βάθος χρόνου) καθώς και την τάξη της κατά την ημερήσια διάταξη της συνεδρίασης (πρωτολογία, δευτερολογία κ.τ.λ.).

4.4 Αποτελέσματα

Προτού προχωρήσουμε στη λήψη οποιουδήποτε αποτελέσματος και επειδή τα κείμενα ήταν διαφόρων μεγεθών, κανονικοποιήσαμε τις μεταβλητές έτσι ώστε όλες οι τιμές να αντιστοιχούν σε κείμενα μεγέθους 100.000 λέξεων. Όπου ήταν δυνατό (π.χ. μέρη του λόγου ή διάφορες κατανομές γραμμάτων και λέξεων), εκφράσαμε τις μεταβλητές σε ποσοστά επί τοις εκατό.

	1ο Ενικό	*2ο Ενικό*	*3ο Ενικό*	*1ο Πληθυντικό*	*2ο Πληθυντικό*	*3ο Πληθυντικό*
Ομιλητής Α	114,30 (1,98)	0,43 (0,07)	208,76 (2,92)	199,28 (3,30)	58,66 (1,40)	429,39 (6,26)
Ομιλητής Β	66,91 (1,89)	1,12 (0,15)	191,73 (4,83)	423,95 (10,19)	50,04 (2,01)	424,51 (10,51)
Ομιλητής Γ	74,00 (1,48)	0,41 (0,06)	247,23 (3,43)	329,91 (6,09)	10,75 (0,46)	577,14 (8,91)
Ομιλητής Δ	63,97 (1,17)	0,92 (0,16)	174,87 (3,30)	185,45 (4,37)	84,21 (1,91)	473,07 (9,00)
Ομιλητής Ε	110,18 (1,98)	0,00 (0,00)	277,02 (5,95)	290,66 (5,95)	123,82 (3,90)	381,43 (8,44)

Πίνακας 4.7 - Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) των ρηματικών καταλήξεων της Καθαρεύουσας, κανονικοποιημένες σε 100.000 λέξεις, για καθέναν από τους πέντε ομιλητές.

	1ο Ενικό	*2ο Ενικό*	*3ο Ενικό*	*1ο Πληθυντικό*	*2ο Πληθυντικό*	*3ο Πληθυντικό*
Ομιλητής Α	25,88 (0,60)	70,52 (1,27)	180,94 (2,66)	68,80 (1,44)	51,11 (1,42)	72,46 (1,33)
Ομιλητής Β	16,31 (0,66)	82,65 (2,46)	183,86 (4,68)	125,95 (3,73)	57,35 (1,75)	89,40 (2,60)
Ομιλητής Γ	15,30 (0,39)	67,39 (1,37)	186,87 (2,62)	109,97 (1,93)	13,23 (0,64)	96,74 (1,67)
Ομιλητής Δ	13,35 (0,50)	115,97 (2,39)	239,76 (4,20)	69,95 (1,83)	71,79 (1,66)	124,25 (2,44)
Ομιλητής Ε	24,66 (0,70)	76,60 (1,90)	202,52 (4,10)	129,59 (3,40)	79,22 (2,26)	81,85 (2,02)

Πίνακας 4.8 - Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) των ρηματικών καταλήξεων της Δημοτικής, κανονικοποιημένες σε 100.000 λέξεις, για καθέναν από τους πέντε ομιλητές.

	Προτάσεις	*Κόμματα*	*Παρενθέσεις*	*Παύλες*	*Ερωτηματικά*	*Καταλήξεις Καθαρεύουσας*	*Καταλήξεις Δημοτικής*
Ομιλητής Α	5081,95 (61,75)	8222,48 (106,15)	136,73 (2,37)	829,45 (12,19)	386,69 (7,64)	1010,83 (12,72)	469,72 (6,22)
Ομιλητής Β	6153,40 (127,80)	7207,64 (153,22)	83,21 (2,33)	852,95 (19,91)	691,02 (16,75)	1156,01 (26,60)	555,52 (12,45)
Ομιλητής Γ	5305,07 (73,39)	7146,05 (96,09)	133,54 (1,60)	348,10 (5,82)	191,42 (3,73)	1239,45 (17,86)	489,49 (6,60)
Ομιλητής Δ	5301,30 (86,91)	7442,08 (119,25)	103,08 (1,49)	235,61 (5,70)	557,74 (10,20)	982,49 (16,64)	635,05 (10,21)
Ομιλητής Ε	5410,25 (105,21)	8403,94 (156,00)	164,22 (4,64)	643,23 (12,84)	447,53 (10,66)	1183,10 (22,83)	594,44 (10,89)

Πίνακας 4.9 - Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) για τα δομικά χαρακτηριστικά κανονικοποιημένα σε 100.000 λέξεις, για καθέναν από τους πέντε ομιλητές. Οι δύο τελευταίες στήλες καταγράφουν τα συγκεντρωτικά αποτελέσματα των πινάκων 3.4 και 3.5. Επειδή δεν είναι γραμμικά ανεξάρτητες μεταβλητές δεν λαμβάνουν μέρος στην στατιστική επεξεργασία και απλώς τα αναφέρουμε για λόγους πληρότητας.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Ομιλητής Α	8,48	7,14	5,59	12,76	7,64	20,30	4,43	7,27	14,18	11,91	0,30
Ομιλητής Β	7,80	5,94	5,67	12,07	7,72	18,09	5,73	7,42	15,21	13,80	0,56
Ομιλητής Γ	8,23	6,38	4,00	14,46	5,71	20,83	4,86	6,50	15,98	12,78	0,26
Ομιλητής Δ	9,54	6,51	4,23	14,02	6,51	21,91	4,33	6,24	14,91	11,47	0,32
Ομιλητής Ε	9,34	6,60	5,77	12,05	7,13	20,11	4,68	7,17	14,38	12,54	0,23

Πίνακας 4.10 - Επί τοις εκατό (%) αναλογία για τα 11 μέρη του λόγου, για καθέναν από τους πέντε ομιλητές. Οι αντιστοιχίες (1=επίθετα κ.τ.λ.) δίνονται στον πίνακα 3.3.

	Γενικές Επιθέτων	*Γενικές Ουσιαστικών*	*1ο Ενικό Ρημάτων*	*2ο Ενικό Ρημάτων*	*3ο Ενικό Ρημάτων*	*1ο Πληθυντικό Ρημάτων*	*2ο Πληθυντικό Ρημάτων*	*3ο Πληθυντικό Ρημάτων*
Ομιλητής Α	21,57	23,62	8,64	0,52	45,89	8,54	11,58	19,24
Ομιλητής Β	16,35	20,72	5,26	0,70	39,35	11,86	10,45	17,28
Ομιλητής Γ	20,50	23,42	7,08	0,18	50,73	16,63	1,91	19,71
Ομιλητής Δ	25,13	29,54	5,35	0,14	45,42	7,24	11,36	18,26
Ομιλητής Ε	20,49	24,49	7,92	0,09	39,66	9,44	9,61	11,58

Πίνακας 4.11 – Επί τοις εκατό (%) αναλογία γενικών σε επίθετα και ουσιαστικά. Επί τοις εκατό (%) αναλογία προσώπου-αριθμού για όλα τα ρήματα (όχι μόνο όσα έχουν τις 230 προαναφερθείσες καταλήξεις).

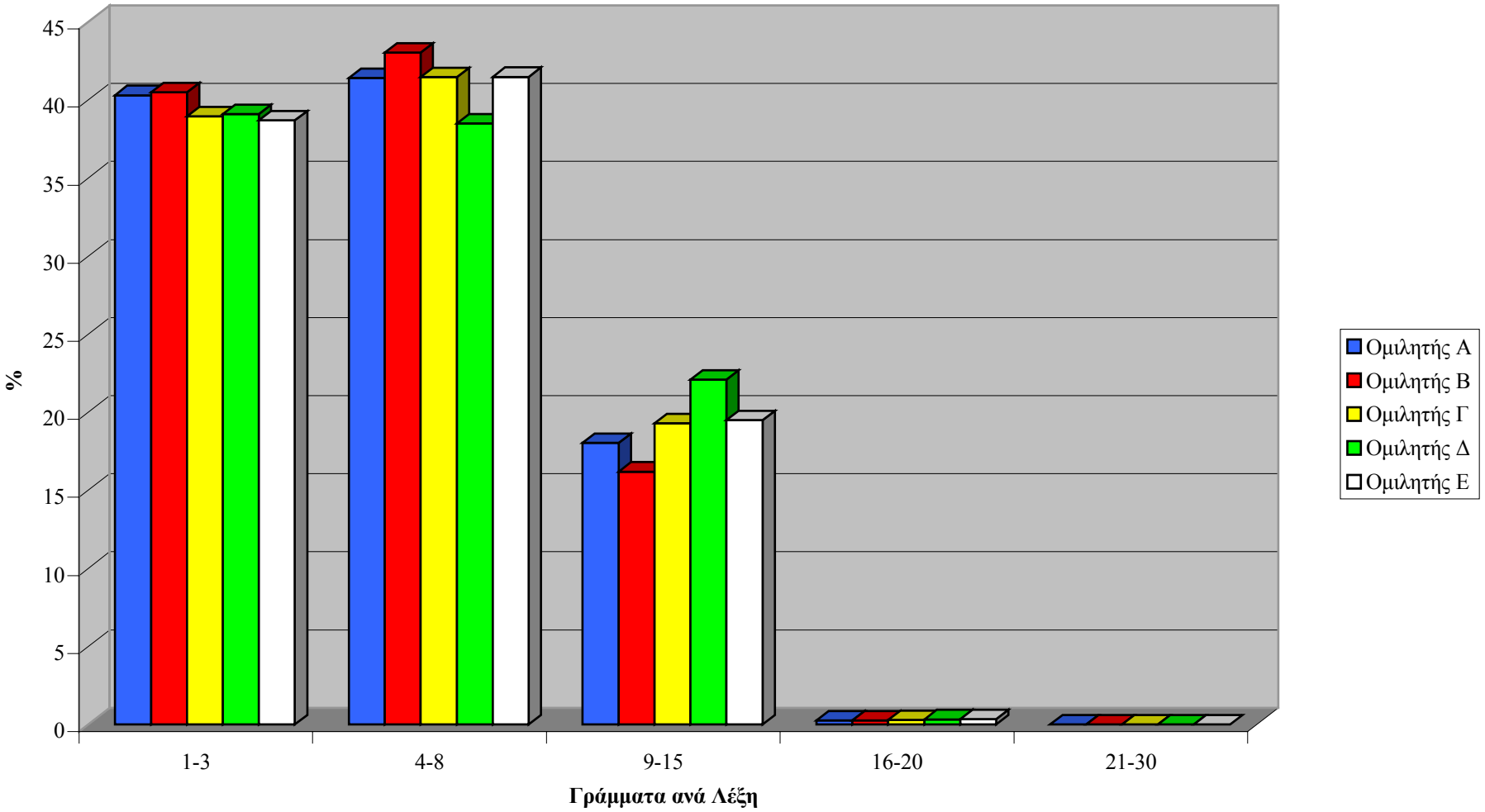
	όχι	*ουδείς*	*ουδέποτε*	*ουδαμού*	*άνευ*	*δε(ν)*	*μη(ν)*
Ομιλητής Α	219,98 (3,69)	0,65 (0,08)	2,59 (0,22)	0,00 (0,00)	6,47 (0,31)	1735,90 (22,93)	421,41 (6,69)
Ομιλητής Β	299,69 (7,98)	2,25 (0,21)	3,37 (0,26)	0,00 (0,00)	3,37 (0,30)	2237,24 (45,78)	468,93 (11,36)
Ομιλητής Γ	101,29 (1,77)	0,00 (0,00)	2,07 (0,17)	0,00 (0,00)	1,24 (0,14)	1904,23 (28,20)	384,49 (7,09)
Ομιλητής Δ	145,42 (3,69)	0,00 (0,00)	0,46 (0,08)	0,00 (0,00)	2,76 (0,19)	1878,00 (27,94)	500,22 (8,28)
Ομιλητής Ε	200,94 (4,25)	4,20 (0,30)	6,30 (0,45)	0,00 (0,00)	7,87 (0,51)	1955,39 (36,90)	459,60 (10,57)

Πίνακας 4.12 – Μέσος όρος συχνοτήτων και τυπική απόκλιση (σε παρένθεση) για τις λέξεις οι οποίες εκφράζουν άρνηση, κανονικοποιημένες σε 100.000 λέξεις.

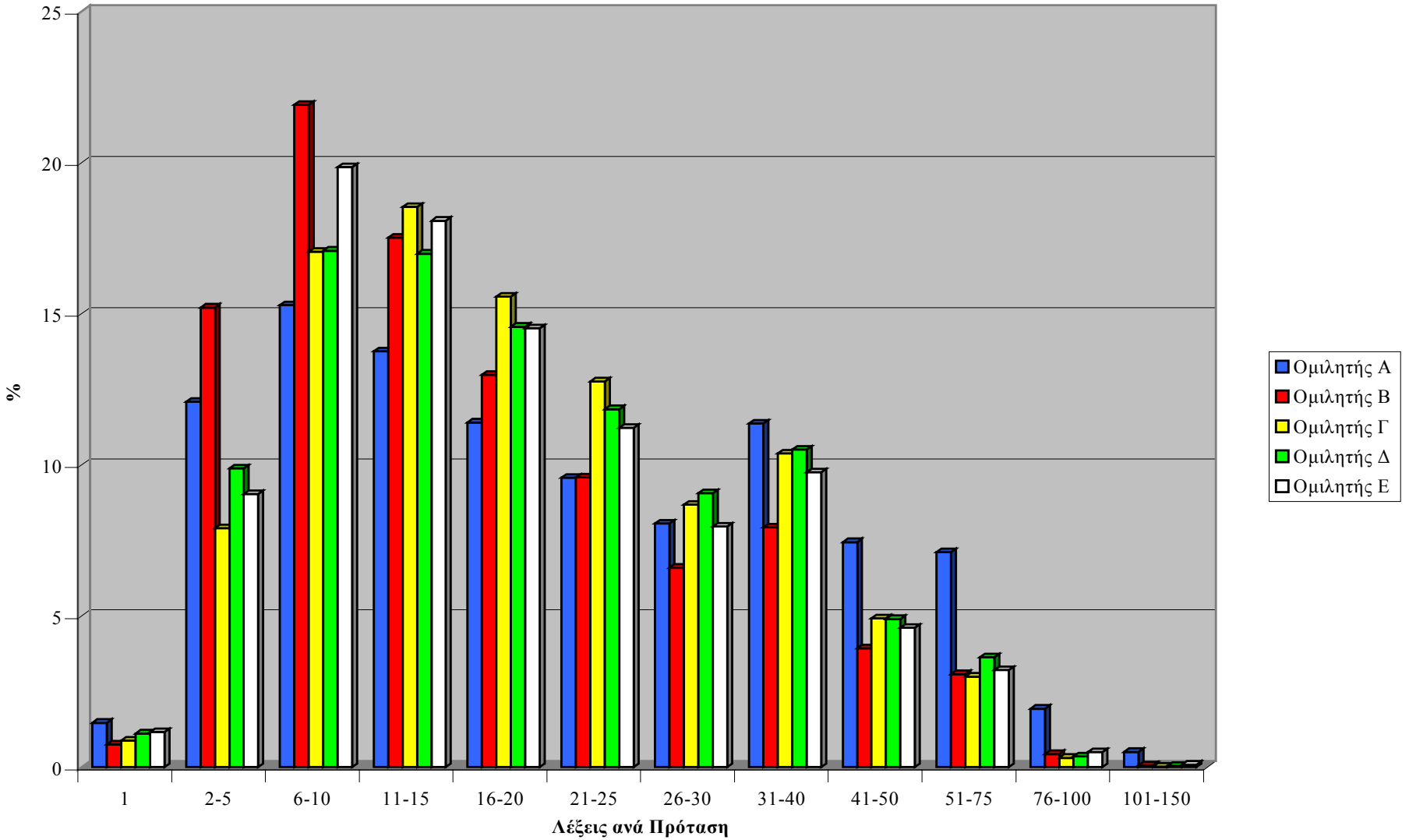
	α	*β*	*οποίος / που*
Ομιλητής Α	11,01 (4,00)	1,82 (1,00)	0,115 (0,16)
Ομιλητής Β	9,31 (4,97)	1,24 (0,95)	0,117 (0,20)
Ομιλητής Γ	12,92 (4,68)	1,58 (0,93)	1,041 (23,67)
Ομιλητής Δ	10,56 (4,28)	1,14 (0,79)	0,243 (0,36)
Ομιλητής Ε	11,19 (5,42)	1,06 (1,11)	0,144 (0,22)

Πίνακας 4.13 – Επί τοις εκατό (%) αναλογία και τυπική απόκλιση (σε παρένθεση) των εσωτερικών αυξήσεων τύπου α και β στα ρήματα παρελθοντικών χρόνων. Επίσης, η αναλογία εμφάνισης των λέξεων *οποίος / που*.

Σχήμα 4.1


45
40
35
30
25
20
15
10
5
0
%
1-3
4-8
9-15
16-20
21-30
Γράμματα ανά Λέξη
Ομιλητής Α
Ομιλητής Β
Ομιλητής Γ
Ομιλητής Δ
Ομιλητής Ε

Σχήμα 4.2


25
20
15
10
5
0
%
1
2-5
6-10
11-15
16-20
21-25
26-30
31-40
41-50
51-75
76-100
101-150
Λέξεις ανά Πρόταση
Ομιλητής Α
Ομιλητής Β
Ομιλητής Γ
Ομιλητής Δ
Ομιλητής Ε

Τα αποτελέσματα των μετρήσεων που κάναμε δίνονται συνοπτικά στους *πίνακες 4.7 – 4.13* και στα *σχήματα 4.1* και *4.2*. Πιο αναλυτικά, ο *πίνακας 4.7* περιέχει τις συχνότητες εμφάνισης των ρηματικών καταλήξεων της Καθαρεύουσας, ενώ ο *πίνακας 4.8* τις αντίστοιχες συχνότητες εμφάνισης των ρηματικών καταλήξεων της Δημοτικής. Επειδή η χρήση ενός πίνακα με 230 μεταβλητές θα ήταν υπολογιστικά αδύνατη αλλά και τα αποτελέσματα που πιθανώς θα έδινε δεν θα βοηθούσαν στην εξαγωγή συμπερασμάτων, οι καταλήξεις εκφράζονται αθροιστικά για κάθε αριθμό και πρόσωπο στον *πίνακα 4.9*, ο οποίος επίσης περιέχει τα δομικά χαρακτηριστικά. Ο *πίνακας 4.10* περιέχει τα στοιχεία για τα μέρη του λόγου, ενώ ο *πίνακας 4.11* τις γενικές ουσιαστικών και επιθέτων καθώς και τα στατιστικά για τα πρόσωπα και αριθμούς των ρημάτων. Ο *πίνακας 4.12* έχει τα στοιχεία για τις αρνήσεις και ο *πίνακας 4.13* στοιχεία για τις εσωτερικές αυξήσεις. Τέλος, απεικονίσαμε σχηματικά τις επί τοις εκατό αναλογίες των γραμμάτων ανά λέξη και λέξεων ανά πρόταση στα *σχήματα 4.1* και *4.2*.

Παρατηρώντας τους πίνακες αυτούς, μπορούμε να κάνουμε κάποιες αξιοπρόσεκτες παρατηρήσεις σε σχέση με τους πέντε ομιλητές που μελετήσαμε :

- Όπως ήταν αναμενόμενο από την ανάλυση των ειδών του λόγου και για τους πέντε ομιλητές οι ρηματικές καταλήξεις οι οποίες αντιστοιχούν στη Καθαρεύουσα είναι γενικότερα πιο συχνές από αυτές της Δημοτικής, εκτός από την περίπτωση του δευτέρου προσώπου ενικού και πληθυντικού. Παρόλα αυτά δεν εμφανίζονται έντονες διακυμάνσεις μεταξύ των πέντε ομιλητών, το αντίθετο μάλιστα, θα μπορούσαμε να ισχυριστούμε ότι όσον αφορά τη ρηματική πολυτυπία υπάρχει μια ομοιομορφία μεταξύ τους.

- Όσον αφορά τα δομικά χαρακτηριστικά παρατηρούμε μια τάση του Ομιλητή Α να χρησιμοποιεί λιγότερες προτάσεις, κάτι το οποίο διαφαίνεται και από την έντονη χρήση σημείων στίξης.

- Ο Ομιλητής Δ χρησιμοποιεί περισσότερο τη γενική από τους υπόλοιπους τέσσερις ομιλητές τόσο στα ουσιαστικά όσο και στα επίθετα, ενώ ο Ομιλητής Γ εκφράζεται περισσότερο από τους συναδέλφους του στο 1ο Πληθυντικό ρημάτων σε αντίθεση

με το 2ο Πληθυντικό, το οποίο χρησιμοποιεί ελάχιστα. Επίσης, ο Ομιλητής Β κάνει τη μεγαλύτερη χρήση αρνητικών μορίων και λέξεων από όλους.

- Μελετώντας τις λεκτικές μετρήσεις, παρατηρούμε ότι ο Ομιλητής Γ χρησιμοποιεί τη λέξη *οποίος* στην ίδια περίπου αναλογία με τη λέξη *που***,** σε αντίθεση με τους υπόλοιπους τέσσερις για τους οποίους η αναλογία αυτή είναι περίπου 1 προς 10.

- Τελειώνοντας με τις κατανομές γραμμάτων ανά λέξη και λέξεων ανά πρόταση παρατηρούμε ότι στη κατηγορία γράμματα ανά λέξη και οι πέντε ομιλητές ακολουθούν παρόμοιες κατανομές, ενώ στις λέξεις ανά πρόταση είναι εμφανής η προτίμηση του Ομιλητή Β στις προτάσεις έως 10 λέξεων, του Ομιλητή Γ στις προτάσεις από 11 έως 25 λέξεις, και τέλος –όπως αναμέναμε- του Ομιλητή Α στις μεγάλες προτάσεις (από 31 έως 150 λέξεις).

Στις επόμενες παραγράφους θα ακολουθήσουμε στατιστικές και άλλες μεθόδους για να κάνουμε μια πιο αντικειμενική ανάλυση των δεδομένων. Πρέπει να επισημάνουμε ότι στην ανάλυση αυτή θα χρησιμοποιήσουμε ένα υποσύνολο των προαναφερθέντων χαρακτηριστικών, με απώτερο στόχο να πετύχουμε ένα ικανοποιητικό ποσοστό αναγνώρισης με όσο το δυνατόν λιγότερες μεταβλητές.

4.5 Επεξεργασία

Ακολουθεί η επεξεργασία των παραπάνω μετρήσεων με τη χρήση στατιστικής μεθόδου αλλά και με τη βοήθεια νευρωνικών δικτύων.

4.5.1 Στατιστική Επεξεργασία

Στην περίπτωση του διαχωρισμού ομιλητών, η μέθοδος της ανάλυσης σε ομάδες (*cluster analysis*) δεν δίνει πολύ καλά αποτελέσματα, λόγω του ότι οι διαφορές σε προσωπικά στυλ είναι σαφώς λιγότερο έντονες από αυτές που εμφανίζονται ανάμεσα στα είδη λόγου. Ακόμη και όταν χρησιμοποιούσαμε δύο μόνο ομιλητές, οι ομάδες

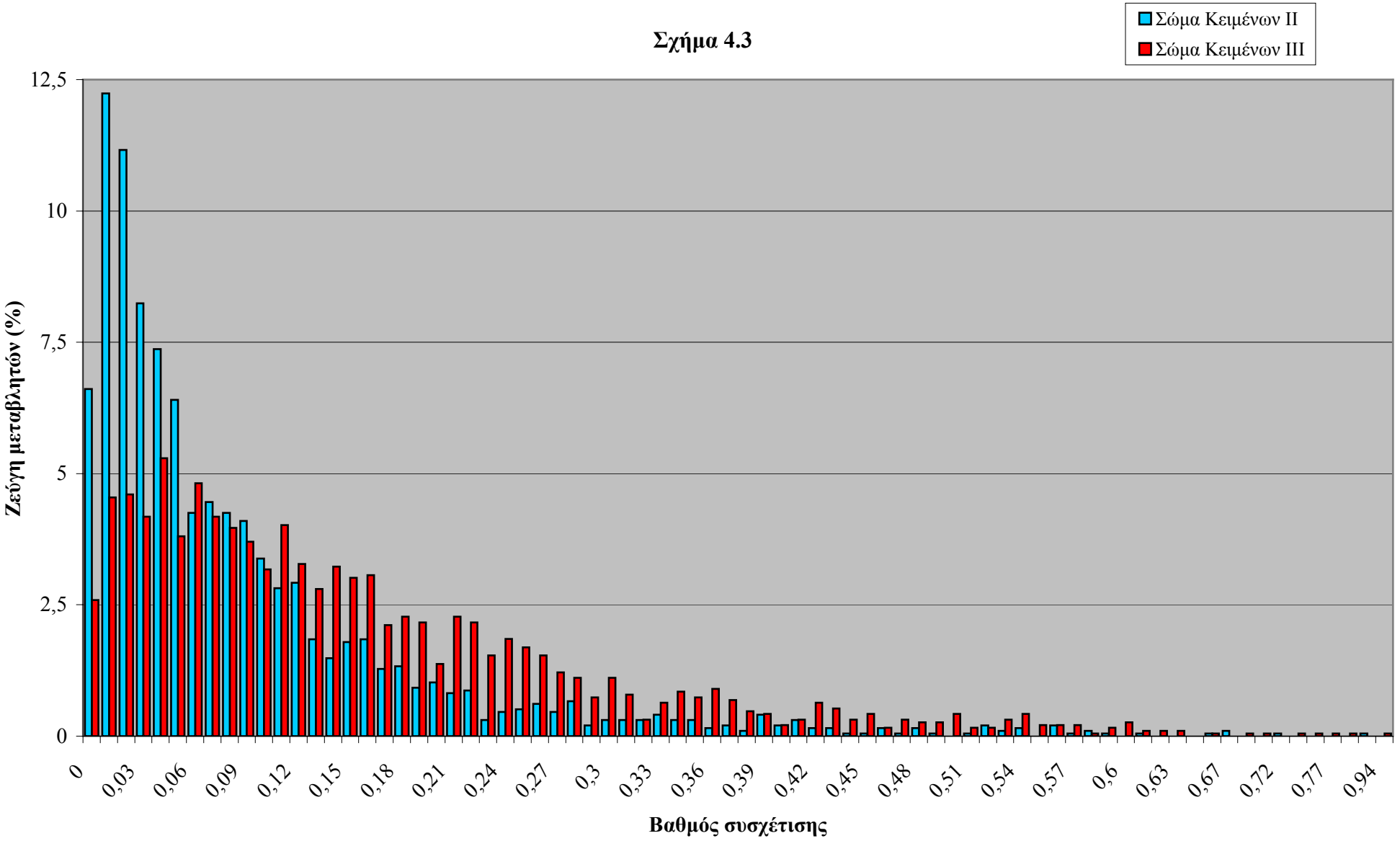
που σχηματίστηκαν περιελάμβαναν κείμενα και από τους δύο [TMH^{+}00b]. Γι' αυτό καταλήξαμε στη χρήση της διακριτικής ανάλυσης (*discriminant analysis*).

Ένα σημαντικό στοιχείο που έπρεπε να μελετήσουμε είναι το εάν οι μεταβλητές που επιλέξαμε να μετρήσουμε είχαν ισχυρή συσχέτιση. Αν όντως υπήρχαν ισχυρές συσχετίσεις, τότε αυτές θα μπορούσαν να χρησιμοποιηθούν ώστε να μειωθούν οι διαστάσεις του χώρου προτύπων. Για τις συγκεκριμένες μετρήσεις χρησιμοποιήσαμε 63 γραμμικά ανεξάρτητες μεταβλητές γιατί όπως θα φανεί αργότερα ο αριθμός των μεταβλητών αυτών είναι υπεραρκετός για την λήψη ικανοποιητικών αποτελεσμάτων.

Στο *σχήμα 4.3* φαίνεται ο αριθμός των συσχετίσεων σε ποσοστό επί τοις εκατό (%) ανά ζεύγος μεταβλητών τόσο για το σώμα κειμένων ΙΙ όσο και για το σώμα κειμένων ΙΙΙ. Όπως βλέπουμε, στο σώμα κειμένων ΙΙ, βαθμό συσχέτισης πάνω από 0,5 έχουμε μόνο στο 1,23% των ζευγών μεταβλητών, κάτι το οποίο υποδεικνύει τη χαμηλή συσχέτιση των παραμέτρων. Επιπλέον από τα 1953 ζεύγη μεταβλητών, μόνο το ένα έχει συσχέτιση μεγαλύτερη του 0,8 (συγκεκριμένα 0,88). Παρόμοια είναι και τα αποτελέσματα για το σώμα κειμένων ΙΙΙ, στο οποίο μόνο το 3,33% των ζευγών μεταβλητών έχουν συσχέτιση πάνω από 0,5 και μόνο ένα έχει συσχέτιση μεγαλύτερη του 0,8 (0,94). Άρα, καταλήγουμε στο συμπέρασμα ότι η πλειονότητα των μεταβλητών είναι μεταξύ τους ασυσχέτιστη .

Στη συνέχεια μελετήσαμε παράλληλα τα σώματα κειμένων ΙΙ και ΙΙΙ με τη μέθοδο της διακριτικής ανάλυσης (*discriminant analysis*). Οι 63 μεταβλητές χρησιμοποιήθηκαν για τη δημιουργία διαχωριστικών συναρτήσεων οι οποίες με ακρίβεια αναγνωρίζουν την ταυτότητα των ομιλητών. Επιπλέον χρησιμοποιήσαμε και τα τρία μοντέλα που προαναφέραμε προηγουμένως (πλήρες μοντέλο, εμπρός και πίσω βηματικό μοντέλο) με τιμή της παραμέτρου F1 = F2 ίση με 4 και μέγιστο αριθμό βημάτων τα 65.

Σχήμα 4.3


Σώμα Κειμένων II
Σώμα Κειμένων III
12,5
10
7,5
5
2,5
0
Ζεύγη μεταβλητών (%)
0
0,03
0,06
0,09
0,12
0,15
0,18
0,21
0,24
0,27
0,3
0,33
0,36
0,39
0,42
0,45
0,48
0,51
0,54
0,57
0,6
0,63
0,67
0,72
0,77
0,94
Βαθμός συσχέτισης

Τα αποτελέσματα της διακριτικής ανάλυσης δίνονται στον πίνακα 4.16 ο οποίος ακολουθεί :

	Πλήρες	*Εμπρός Βηματικό*	*Πίσω Βηματικό*
Σώμα Κειμένων ΙΙ	92,31 (63)	90,63 (33)	91,81 (41)
Σώμα Κειμένων ΙΙΙ	100,00 (62)	97,64 (16)	97,64 (19)

Πίνακας 4.16 - Αποτελέσματα της διακριτικής ανάλυσης για τα σώματα κειμένων ΙΙ και ΙΙΙ και σε παρένθεση ο αριθμός των μεταβλητών που χρησιμοποιεί το μοντέλο.

Αναλυτικότερα για κάθε ομιλητή και σώμα κειμένων τα παραπάνω αποτελέσματα διαμορφώνονται ως εξής :

Σώμα Κειμένων ΙΙ	*Πλήρες*	*Εμπρός Βηματικό*	*Πίσω Βηματικό*
Ομιλητής Α	95,90	95,20	95,42
Ομιλητής Β	84,71	77,65	84,71
Ομιλητής Γ	91,39	91,39	92,21
Ομιλητής Δ	90,26	88,31	87,66
Ομιλητής Ε	89,32	84,47	88,35

Πίνακας 4.17 - Ποσοστά αναγνώρισης ομιλητών του σώματος κειμένων ΙΙ για τα μοντέλα που αναφέρονται στον πίνακα 4.16.

Σώμα Κειμένων ΙΙΙ	*Πλήρες*	*Εμπρός Βηματικό*	*Πίσω Βηματικό*
Ομιλητής Α	100,00	100,00	97,22
Ομιλητής Β	100,00	100,00	100,00
Ομιλητής Γ	100,00	96,67	96,67
Ομιλητής Δ	100,00	95,24	95,24
Ομιλητής Ε	100,00	95,83	100,00

Πίνακας 4.18 - Ποσοστά αναγνώρισης ομιλητών του σώματος κειμένων ΙΙΙ για τα μοντέλα που αναφέρονται στον πίνακα 4.16.

Παρατηρούμε ότι η απόδοση του συστήματος βελτιώνεται στις μετρήσεις που αφορούν το σώμα κειμένων ΙΙΙ. Αυτό οφείλεται στο ότι τα συγκεκριμένα κείμενα είναι πρωτολογίες, δηλαδή κείμενα μεγαλύτερα σε μέγεθος και καλύτερα προετοιμασμένα από κάθε ομιλητή, με άμεση συνέπεια να είναι πιο αντιπροσωπευτικά για κάθε ομιλητή (μικρότερες διακυμάνσεις των παραμέτρων τους).

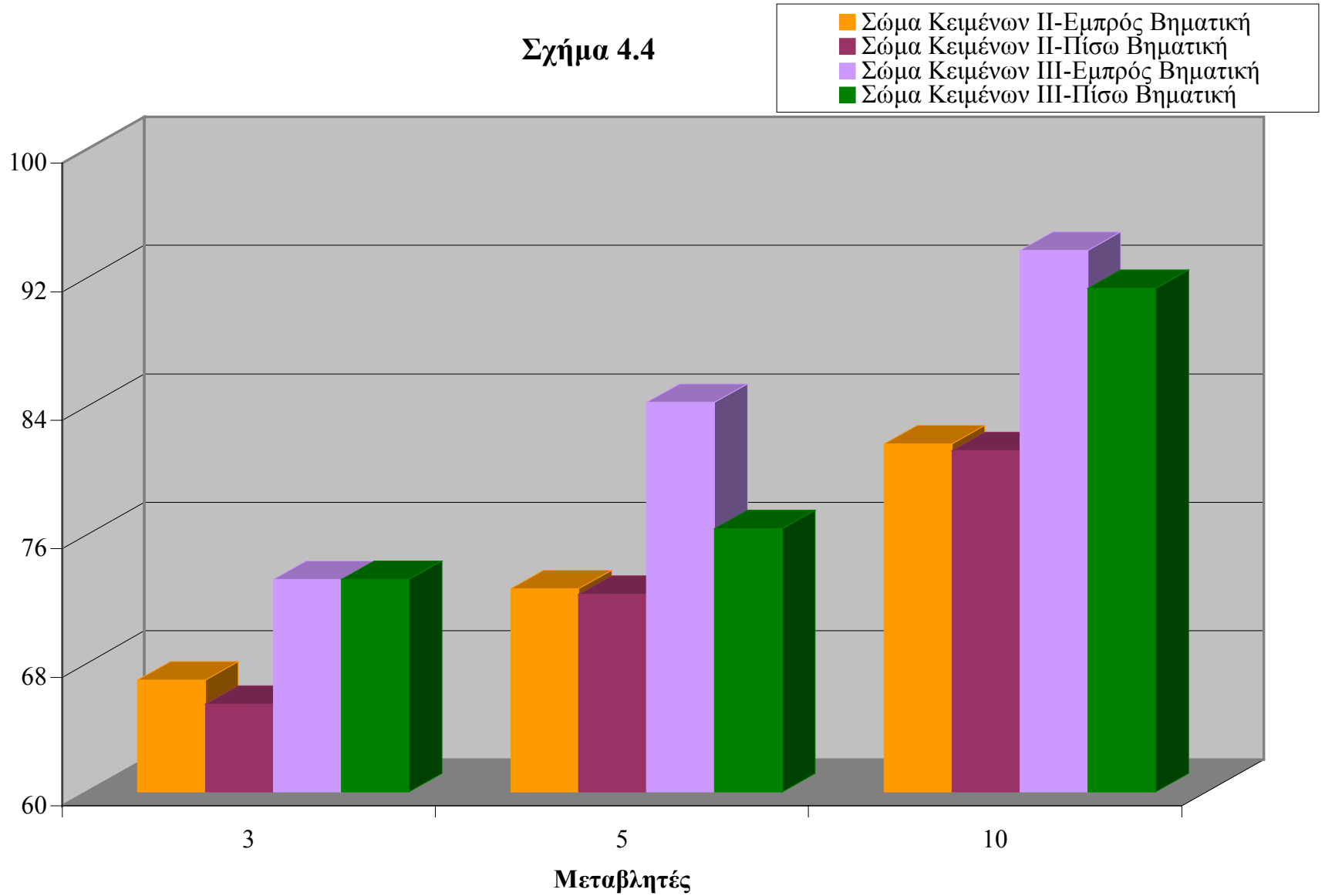
Όπως αναφέραμε και προηγουμένως επιθυμούμε το σύστημα μας να χρησιμοποιεί όσο το δυνατόν λιγότερες μεταβλητές. Για το σκοπό αυτό προχωρήσαμε και σε μια ακόμη ανάλυση των σωμάτων κειμένων ΙΙ και ΙΙΙ, τα αποτελέσματα της οποίας δίνονται παρακάτω.

	Εμπρός Βηματικό	*Πίσω Βηματικό*
Σώμα Κειμένων ΙΙ	66,97 (3)	65,47 (3)
Σώμα Κειμένων ΙΙ	72,64 (5)	72,31 (5)
Σώμα Κειμένων ΙΙ	81,66 (10)	81,26 (10)
Σώμα Κειμένων ΙΙΙ	73,23 (3)	73,23 (3)
Σώμα Κειμένων ΙΙΙ	84,25 (5)	76,38 (5)
Σώμα Κειμένων ΙΙΙ	93,70 (10)	91,34 (10)

Πίνακας 4.19 – Αποτελέσματα discriminant analysis για μοντέλα με λίγες μεταβλητές (σε παρένθεση ο αριθμός των μεταβλητών).

Παρατηρούμε ότι μπορούμε να έχουμε ικανοποιητική αναγνώριση ομιλητών ακόμα και με τρεις μεταβλητές, πόσο μάλλον με πέντε ή δέκα, γεγονός το οποίο είναι αρκετά ενθαρρυντικό για τις περαιτέρω έρευνες.

Σχήμα 4.4


Σώμα Κειμένων ΙΙ-Εμπρός Βηματική
Σώμα Κειμένων ΙΙ-Πίσω Βηματική
Σώμα Κειμένων ΙΙΙ-Εμπρός Βηματική
Σώμα Κειμένων ΙΙΙ-Πίσω Βηματική
100
92
84
76
68
60
3
5
10
Μεταβλητές

4.5.2 Νευρωνικά Δίκτυα

Η επεξεργασία των αποτελεσμάτων έγινε και πάλι μέσω του SOM στο περιβάλλον της Matlab, αλλά μόνο για το σώμα κειμένων III. Χρησιμοποιήσαμε το ίδιο σετ εντολών που αναφέρεται στο παράρτημα IV, χωρίς πάντως τη χρήση της τυχαίας αρχικοποίησης.

Χρησιμοποιήσαμε χάρτες μεγέθους 10x5, 8x4, 6x3 και 3x2. Επίσης χρησιμοποιήσαμε τόσο το σύνολο των μεταβλητών (85) αλλά και υποσύνολα του για να δούμε ποιες μεταβλητές συνεισφέρουν περισσότερο στο σωστό διαχωρισμό των ομιλητών.

	Δημ.–Καθ. (14)	*Ιστογράμματα (17)*	*Γενικές-Κλίσεις (8)*	*PoS (11)*	*Λήμματα (17)*	*Συνολικά (85)*
10x5	57,48	69,29	73,23	72,44	78,74	85,83
8x4	52,75	69,29	58,27	63,78	73,23	81,89
6x3	44,09	62,99	53,54	59,84	64,57	77,16
3x2	40,16	53,54	53,54	48,82	59,06	65,35

Πίνακας 4.20 – Αποτελέσματα μέσω νευρωνικού δικτύου για διάφορα μεγέθη χάρτη.

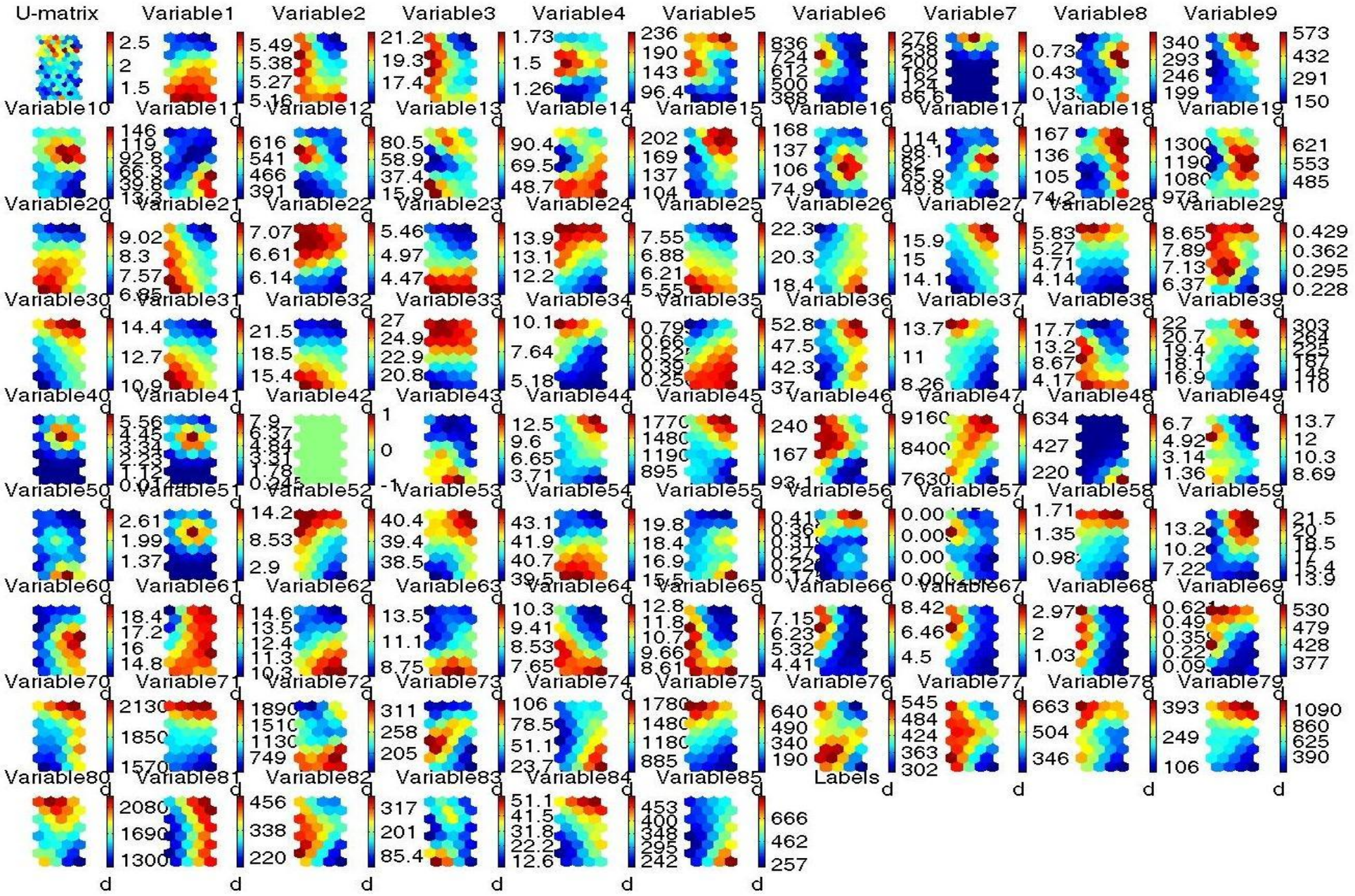
	Ομιλητής Α	*Ομιλητής Β*	*Ομιλητής Γ*	*Ομιλητής Δ*	*Ομιλητής Ε*
10x5	91,67	93,75	93,33	76,19	70,83
8x4	75,00	93,75	83,33	80,95	83,33
6x3	72,22	87,50	93,33	80,95	54,17
3x2	63,89	81,25	80,00	66,67	37,50

Πίνακας 4.21 – Αποτελέσματα (επί των 85 μεταβλητών) για κάθε ομιλητή ξεχωριστά.

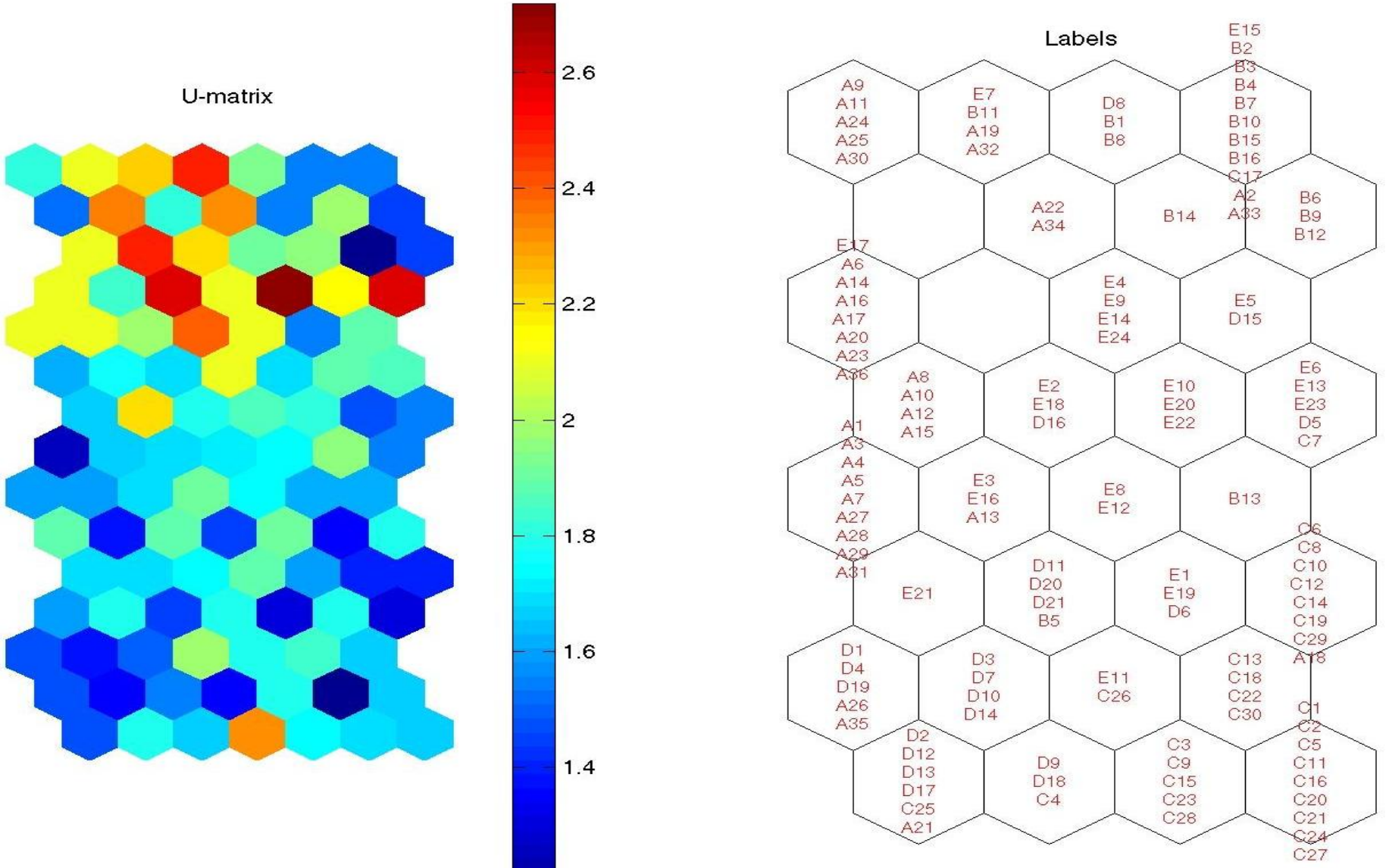
Παρατηρούμε ότι τα ποσοστά διαχωρισμού εμφανίζονται μικρότερα όσο μικραίνει το μέγεθος του χάρτη, κάτι το οποίο είναι αναμενόμενο γιατί σε ένα μικρό χάρτη τα κείμενα μπορούν να κατανεμηθούν σε λιγότερα κελιά. Αυτό έχει ως αποτέλεσμα κάποια κείμενα να ανατίθενται αναγκαστικά σε κελιά τα οποία ήδη περιέχουν κείμενα άλλων ομιλητών.

Επίσης παρατηρούμε ότι κάποιοι ομιλητές –και συγκεκριμένα οι Β και Γ- αναγνωρίζονται συστηματικά καλύτερα από τους άλλους, ενώ αντίθετα ο ομιλητής Ε αναγνωρίζεται πάντα με περισσότερη δυσκολία.

Στα *σχήματα 4.5* και *4.6* που ακολουθούν παραθέτουμε τους χάρτες 8x4 που προκύπτουν για κάθε μία από τις 85 μεταβλητές (αναλυτικά στο *Παράρτημα VI*) καθώς και τον αντίστοιχο U-matrix. Ο Ομιλητής Α συμβολίζεται με Α, ο Β με Β ο Γ με C, ο Δ με D και ο Ε με Ε. Είναι φανερό ότι ο διαχωρισμός σε αυτή την περίπτωση δεν επιτυγχάνεται με τόσο μεγάλη ακρίβεια -σε σχέση πάντα με την διακριτική ανάλυση, η οποία πάντως είναι μέθοδος με επίβλεψη (supervised) και όχι χωρίς επίβλεψη (unsupervised), όπως το νευρωνικό δίκτυο- μιας και αρκετά κείμενα ομιλητών ανατίθενται σε περιοχές όπου επικρατεί κάποιος άλλος ομιλητής.


U-matrix
Variable1
Variable2
Variable3
Variable4
Variable5
Variable6
Variable7
Variable8
Variable9
Variable10
Variable11
Variable12
Variable13
Variable14
Variable15
Variable16
Variable17
Variable18
Variable19
Variable20
Variable21
Variable22
Variable23
Variable24
Variable25
Variable26
Variable27
Variable28
Variable29
Variable30
Variable31
Variable32
Variable33
Variable34
Variable35
Variable36
Variable37
Variable38
Variable39
Variable40
Variable41
Variable42
Variable43
Variable44
Variable45
Variable46
Variable47
Variable48
Variable49
Variable50
Variable51
Variable52
Variable53
Variable54
Variable55
Variable56
Variable57
Variable58
Variable59
Variable60
Variable61
Variable62
Variable63
Variable64
Variable65
Variable66
Variable67
Variable68
Variable69
Variable70
Variable71
Variable72
Variable73
Variable74
Variable75
Variable76
Variable77
Variable78
Variable79
Variable80
Variable81
Variable82
Variable83
Variable84
Variable85
Labels

Σχήμα 4.5


U-matrix
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
Labels
A9 A11 A24 A25 A30
E7 B11 A19 A32
D8 B1 B8
E15 B2 B3 B4 B7 B10 B15 B16 C17
A22 A34
B14
A2 A33
B6 B9 B12
E17 A6 A14 A16 A17 A20 A23 A36
E4 E9 E14 E24
E5 D15
A8 A10 A12 A15
E2 E18 D16
E10 E20 E22
E6 E13 E23 D5 C7
A1 A3 A4 A5 A7 A27 A28 A29 A31
E3 E16 A13
E8 E12
B13
E21
D11 D20 D21 B5
E1 E19 D6
C6 C8 C10 C12 C14 C19 C29 A18
D1 D4 D19 A26 A35
D3 D7 D10 D14
E11 C26
C13 C18 C22 C30
C1 C2 C5 C11 C16 C20 C21 C24 C27
D2 D12 D13 D17 C25 A21
D9 D18 C4
C3 C9 C15 C23 C28

Σχήμα 4.6

4.6 Συμπεράσματα

Τα αποτελέσματα καταδεικνύουν ότι τα κείμενα των πέντε ομιλητών και γενικότερα του Πολιτικού Λόγου δεν μπορούν να διαχωριστούν εύκολα με βάση μόνο τα μορφολογικά χαρακτηριστικά και πιο συγκεκριμένα τις διαφορές Καθαρεύουσας – Δημοτικής. Αυτό εξηγείται, όπως έχει ήδη αναφερθεί, από το γεγονός ότι οι ομιλίες της Βουλής φαίνονται να σχηματίζουν μια καλά ορισμένη μορφή λόγου η οποία χρησιμοποιείται αποκλειστικά από τους πολιτικούς. Σε αντίθεση με τη διαδικασία διαχωρισμού των ειδών λόγου, οι λεκτικές παράμετροι παίζουν σημαντικό ρόλο στο διαχωρισμό ομιλητών. Αυτό φαίνεται και από το γεγονός ότι οι αρνητικές λέξεις και μόρια (όπως το δεν) έχουν κάποια σημασία στο διαχωρισμό ομιλητών, ενώ αντίθετα δεν μας απασχόλησαν καθόλου στο διαχωρισμό ειδών λόγου, ίσως πάντως και επειδή είχαμε ήδη επιτύχει μεγάλο ποσοστό διαχωρισμού και χωρίς την χρήση τους.

Τα δομικά χαρακτηριστικά είναι επίσης σημαντικά : Το μέσο μήκος λέξης, η χρήση σημείων στίξης και ερωτηματικών, όπως και κάποια μέρη του λόγου (άρθρα, επιρρήματα και ειδικά τα ρήματα).

Τέλος, πρέπει να επισημάνουμε ότι για ορισμένους ομιλητές προκύπτουν και κάποια αξιοσημείωτα συμπεράσματα. Για παράδειγμα, ο Ομιλητής Γ γενικά αναγνωρίζεται συστηματικά με μεγάλη ακρίβεια, σε αντίθεση με τον Ομιλητή Ε (αυτό γίνεται ιδιαίτερα αντιληπτό στην αναγνώριση με το νευρωνικό δίκτυο). Επιπλέον, ο Ομιλητής Α ομοιάζει αρκετά με τον Ομιλητή Β, όπως και ο Ομιλητής Δ με τον Ομιλητή Ε. Στις περιπτώσεις αυτές ίσως μια άλλη επιλογή μεταβλητών θα βοηθούσε ακόμη περισσότερο στο διαχωρισμό τους.

ΚΕΦΑΛΑΙΟ 5
ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Εισαγωγή

Στο πέμπτο κεφάλαιο συνοψίζουμε τα συμπεράσματα τα οποία προκύπτουν έως τώρα, ενώ ασχολούμαστε και με τις προοπτικές και κατευθύνσεις τις οποίες θα μπορούσε να ακολουθήσει στο μέλλον η εργασία αυτή με στόχο τη βελτιστοποίηση των αποτελεσμάτων και την εξαγωγή νέων συμπερασμάτων.

5.2 Συμπεράσματα

Όπως αναφέραμε και αναλυτικότερα στα αντίστοιχα κεφάλαια, η ακρίβεια που μπορέσαμε να επιτύχουμε με το σύστημα μας είναι κάτι παραπάνω από ικανοποιητική. Πιο συγκεκριμένα, λάβαμε πάρα πολύ καλά αποτελέσματα (με ποσοστό επιτυχίας έως και 100%) στην περίπτωση του διαχωρισμού των ειδών του λόγου με τη χρήση στατιστικών μεθόδων, ενώ ανάλογα καλά αποτελέσματα είχαμε και με τη χρήση των αυτό-οργανωμένων χαρτών (*SOM*). Επίσης πολύ καλά –αν και όχι στον ίδιο βαθμό- αποτελέσματα είχαμε και στον διαχωρισμό των ομιλητών, για τον οποίο πάντως αναγκαστήκαμε να μετρήσουμε μια πλειάδα μεταβλητών μιας και στην περίπτωση αυτή τα μορφολογικά χαρακτηριστικά δεν ήταν αρκετά. Το πρόβλημα που αντιμετωπίσαμε κατά την διάρκεια όλης της εργασίας ήταν ότι κάτι ανάλογο δεν είχε επιχειρηθεί μέχρι τώρα για την Ελληνική Γλώσσα και έτσι αναγκαστήκαμε να ακολουθήσουμε προτάσεις ερευνητών οι οποίοι είχαν ασχοληθεί με παρόμοια ζητήματα για την Αγγλική κυρίως Γλώσσα.

Συνεπώς, καταλήγουμε στο συμπέρασμα ότι ενώ μεν τα αποτελέσματα μας είναι αρκετά καλά, υπάρχουν ακόμη σημαντικά περιθώρια βελτιώσεων, κυρίως σε κάποιους τομείς τους οποίους δεν εξερευνήσαμε ενδελεχώς και τους οποίους αναφέρουμε στην επόμενη παράγραφο.

5.3 Προτάσεις για βελτιώσεις

Αν και η εργασία αυτή μας έδωσε άκρως ικανοποιητικά και ενθαρρυντικά αποτελέσματα, εν τούτοις δεν θα μπορούσαμε να ισχυριστούμε ότι δεν υπάρχουν περαιτέρω περιθώρια για βελτιώσεις και προσθήκες.

Έτσι, η έρευνα μας θα μπορούσε στο μέλλον να εμπλουτιστεί με την προσθήκη επιπλέον μετρήσεων, όπως για παράδειγμα τη μέτρηση παραθετικών δομών, ή τη δημιουργία λιστών με την κατάταξη των χρησιμοποιούμενων λημμάτων από κάθε ομιλητή ή συγγραφέα. Επίσης θα μπορούσαμε να εισάγουμε το θέμα κάθε κειμένου ως μια παράμετρο της εργασίας μας. Έχοντας υπόψη μας ότι η Λογοτεχνία (από την πλευρά του διαχωρισμού ειδών λόγου) και ο Ομιλητής Ε (για τον διαχωρισμό ομιλητών) αναγνωρίζονται συστηματικά με μικρότερη ακρίβεια θα πρέπει να στοχεύσουμε στην εύρεση και εισαγωγή νέων παραμέτρων μέσω των οποίων θα αυξηθεί η ακρίβεια του συστήματος.

Επιπλέον, θα μπορούσαμε να υιοθετήσουμε κάποιες μεθόδους και μετρήσεις οι οποίες ακολουθούνται από άλλους ερευνητές, όπως για παράδειγμα η *Inverse Document Frequency* [CG95], καθώς επίσης και να ερευνήσουμε αν ισχύουν κάποια αποτελέσματα τα οποία έχουν εξαχθεί για την Ελληνική Γλώσσα και έχουν δημοσιευθεί στην ξένη βιβλιογραφία (για παράδειγμα [Hud94]).

Βέβαια, οι μετρήσεις και τα αποτελέσματά μας θα πρέπει να διευρυνθούν (π.χ. με την προσθήκη περισσότερων ειδών λόγου, όπως ο θρησκευτικός ή λόγος που περιέχει τεχνική ορολογία) και να επιβεβαιωθούν με την πραγματοποίηση πειραμάτων μεγαλύτερης κλίμακας, καθώς και με τη χρήση περισσότερων στατιστικών ή και άλλων μεθόδων. Ο τομέας της αναγνώρισης ομιλητών, ο οποίος είναι πιο απαιτητικός και σαφώς πιο ενδιαφέρων, πρέπει επίσης να ερευνηθεί σε μεγαλύτερο βάθος και να εμπλουτιστεί με νέα δεδομένα.

Τέλος, ενδιαφέρον παρουσιάζουν και διάφορα άλλα στοιχεία που προέκυψαν αμυδρά κατά τη διάρκεια των πειραμάτων, όπως π.χ. το ότι τα χαρακτηριστικά κάποιων ομιλητών μεταβάλλονται σε βάθος χρόνου και ίσως αξίζει να ερευνήσουμε εάν μπορεί να γίνει πρόβλεψη για την μεταβολή αυτή. Επίσης θα είχε ενδιαφέρον να

μελετήσουμε το πως μεταβάλλεται ο τρόπος ομιλίας κάτω από διαφορετικές περιστάσεις και συνθήκες.

Παράρτημα I

Automatic Style Categorisation of Corpora in the Greek Language.

George Tambouratzis, Stella Markantonatou, Nikolaos Hairetakis, George Carayannis

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
{giorg_t, marks, nhaire, gcara}@ilsp.gr

Abstract

In this article, a system is proposed for the automatic style categorisation of text corpora in the Greek language. This categorisation is based to a large extent on the type of language used in the text, for example whether the language used is representative of formal Greek or not. To arrive to this categorisation, the highly inflectional nature of the Greek language is exploited. For each text, a vector of both structural and morphological characteristics is assembled. Categorisation is achieved by comparing this vector to given archetypes using a statistical-based method. Experimental results reported in this article indicate an accuracy exceeding 98% in the categorisation of a corpus of texts spanning different registers.

1. Introduction

The identification of the language style characterising the constituent parts of a corpus is very important to several applications. For example, in information retrieval applications, where large corpora of texts need to be searched efficiently, it is useful to have information about the language style used in each text, in order to improve the accuracy of the search (Karlgren, 1999). Language style can thus serve as a factor increasing the accuracy of the search itself. In fact, the criteria regarding language style may differ for each search and therefore – due to the large number of texts – there is a requirement to perform style categorisation in an automated manner.

Modern Greek presents itself as an interesting case study in style issues. As a whole, the Greek language has evolved continuously during the past 3000 years. One of the results of this evolution is that the correspondence between a given set of grammatical features, say, *Verb+Past+Imperfect+3rd+Plural* and the surface morphological markers is one-to-many, e.g. for the set of grammatical features given above and the verb root: *ντυν-* [din-] (to dress someone) the following forms are in widespread use (in the following phonetic transcriptions, a stress following a vowel identifies the intonation point in the word.:

(1) *έντυναν* [e'dinan], *ντύναν* [di'nan], *ντύνανε* [di'nane] (=dressed).

Politipia is the Greek term reserved for this phenomenon. In addition, it is often the case that more than one words are available for the same sense (e.g. *ύδωρ* [i'*δ*or] / *νερό* [nero'] (=water)). Syntax also presents cases of politipia.

Politipia in Modern Greek was endorsed by the so-called *Katharevousa* which was used as the official language of the Greek state (and the secondary and tertiary eduction) until 1979 – when it was replaced by the so-called *Demotiki*. *Katharevousa* was a somehow haphazard mix of colloquial and ancient Greek (Holton, Mackridge & Philippaki-Warburton, 1997). Modern Greek, while being closer to *Demotiki*, contains a considerable number of features which are easily identified as belonging to *Katharevousa*. Such features occur in varying proportions in the different registers (Biber, 1993; Biber, Conrad & Reppen, 1998). Typically, registers employing a more "formal" linguistic style, such as academic prose and the language of the law, contain more *Katharevousa* features than, say, fiction and spoken language.

It is important to stress here that several of the grammatically equivalent word forms may be encountered relatively frequently in Modern

Greek. However, as some of them normally characterise formal versus informal speech, it is the case that the choice of a particular word form signifies a particular linguistic style.

In this paper, a system is proposed that allows for the automatic categorisation of the texts contained in a corpus on the basis of the style employed by the respective author. The proposed system draws on morphological and structural features of the texts. In the case of Modern Greek, the system takes advantage of the phenomenon of politipia and draws heavily on the inflectional nature of this language. Information regarding the type of endings used is collected using automated tools. The system also employs information regarding structural features such as the size and structure of words and sentences. The morphological and structural features are assembled to form a data vector representing that text. This vector is then compared with the use of statistical methods to archetype vectors of the different styles taught to the system and is classified as belonging to the class with the highest likelihood.

In the next section, the problem of text-style categorisation is presented in more detail and related work is reviewed. In section 3, the morphological characteristics this study has relied on are presented. In section 4, the structural characteristics that were searched for are detailed. In section 5, the methods used to extract the structural and morphological information are presented. The corpora used in our experiments and the experimental results are described in section 6. Finally, section 7 concludes this paper with a review and discussion of the system structure and performance.

2. Categorisation of text style.

The recent dramatic developments in information technology assist in the integrated storage and processing of linguistic resources. For example, textual data is generated in an electronic format and therefore can be readily transferred and replicated by electronic means. Additionally, as a result of the improvement in the processing power and storage capacity of computer systems, the storage of very large corpora of texts has become possible. By virtue of the increased quality and capabilities of telecommunications it is possible to search through a number of distributed databases in order to retrieve the required information. As the cost of such computer systems falls, their use becomes more widespread, giving rise to the need for advanced information retrieval systems.

It has been argued (Karlgren, 1999) that in information retrieval applications, the language style used by the authors of the different texts is important to the accuracy of the search, this being determined objectively as the number of desired texts over the number of actually retrieved texts. Evidently, in such systems, this accuracy depends strongly on the selection of the appropriate feature set. This feature set, in turn, is dependent on the language used in the texts, though the system principles should hold over different languages, by selecting the appropriate characteristics. In the following two sections, the characteristics selected for the style classification of texts in the Greek language shall be presented.

3. Morphological Characteristics.

As noted earlier, politipia is a prevalent phenomenon in Modern Greek. Politipia is the phenomenon whereby a set of grammatical features is mapped onto more than one surface forms. Politipia is observed mainly in the verb system of Modern Greek as noted earlier in example (1). To a lesser degree, politipia can be observed in the noun and adverb systems of Modern Greek (see also Mikros & Carayannis, 2000). Adverbs, however, are not as frequent as verbs. In addition, in the case of adverbs, morphological issues interfere with semantic issues rendering automated processing intractable. On the other hand, politipia in the noun system is restricted to a very small number of forms. In the light of these facts, the research described in this article is confined to the study of verbal politipia.

Verbal politipia in Greek is mainly due to suffixes and, to a lesser degree, to prefixes (1). That is, verbal politipia in Greek is mainly due to the multitude of endings available for the same cluster of grammatical features. Some of the grammatically equivalent endings do not reflect a difference between formal (*Katharevousa*) and informal (*Demotiki*) speech but others do. This work focuses on the distinction between formal and informal speech.

Alternative endings are due to active evolutional tendencies of the language. Such is the tendency of *Demotiki* to have words ending with an 'open' syllable. So, *3rd Plural* endings in *–n* are augmented to *–ne* (2).

(2) *έλεγαν* [e'leγan] *(Kath)* / *λέγανε* [le'γane] *(Dem)* (=they said)

A second active tendency of *Demotiki* is to convert *Katharevousa's* consonant clusters consisting of two fricatives or two plosives into clusters consisting of one fricative and one plosive. In clusters containing an /s/, the non-strident partner converts to a plosive. In clusters with voiceless fricatives not containing an /s/, the first of the consonants is a fricative and the second one is a plosive (Holton, Mackridge & Philippaki-Warburton, 1997, pp. 14) as shown in examples (3)-(5).

(3) *πεισθώ* [pisθo'] */πειστώ* [pisto'] (=to be convinced)
(4) *καλυφθώ* [kalifθo'] */καλυφτώ* [kalifto'] (=to be covered / protected)
(5) *απαλλάχθηκα* [apala'xθika] / *απαλλάχτηκα* [apala'xtika] (=got rid of)

Third, in Modern Greek there exist classes of verbs which may either follow the inflectional paradigm of ancient verbs having an /a/ (6) or an /o/ (7) or an /e/ (8) as a thematic vowel or choose the set of endings offered by *Demotiki* (Clairis and Babiniotis, 1999).

(6) *εξαρτάται* [eksarta'te] *(Kath)* / *εξαρτιέται* [eksartiete] *(Dem)* (=depends)
(7) *αξιούμε* [aksiu'me] *(Kath)* / *αξιώνουμε* [aksio'nume] *(Dem)* (=demand)
(8) *θεωρείτο* [θeori'to] *(Kath)* / *θεωρούνταν* [θeoru'dan] *(Dem)* (= was considered)

It is sometimes the case that *Demotiki* uses a verbal root which is similar but not identical to the *Katharevousa* one (9).

(9) *δεικνύω* [δikni'o] *(Kath)* / *δείχνω* [δi'xno] *(Dem)* (=to show, to point)

Finally, there is a bulk of verbs (and their compounds) which are clearly inherited from *Katharevousa* but either they have not been replaced by some word of *Demotiki* (10) or, the existing *Demotiki* equivalent, is marked as colloquial (11) (Clairis and Babiniotis, 1999).

(10) *προϊσταμαι* [proi'stame] (=supervise)
(11) *διάκειμαι* [δia'kime] *(Kath)* / *νιώθω* [nio'θo] *(Dem)* (=I am disposed)

These morphological contrasts run through the inflectional paradigms of Greek verbs and affect hundreds of verbal wordforms. By studying the quantitative distribution of these forms, it shall be shown that morphological features play an important role in the identification of linguistic style (Biber, Konrad & Reppen, 1998). A total of approximately 230 endings which are characteristic of *Katharevousa* or *Demotiki* have been selected to determine the morphological information within a given text.

4. Structural Characteristics.

In the experiments, several different structural characteristics were measured in each text contained in the corpus:

a) Word count and sentence count.
b) Count of commas and other punctuation signs (parentheses, dashes).
c) Average word and sentence length.
d) Frequency of words with length *x* [where *x* varies from 1 to 30 letters].
e) Frequency of sentences composed by *x* words [where *x* varies from 1 to 150 words].

Apart from the obvious need to measure the characteristics mentioned in (a,c,d,e), characteristics (b) were also introduced since punctuation signs such as brackets, dashes and commas frequently indicate the existence of sub-sentences inside a sentence. Thus, the frequency-of-occurrence of these punctuation marks provides an indication of the stylistic variation in a text. It should be noted that in our measurements digits were counted as words.

Broadly similar factors were used in the experiments of Karlgren (1999), who counted characteristics (a, c and e) with the addition of long words and digits. Long words have not been used in this piece of research as they are not expected to be indicative of a given style in the Greek language, in contrast to other languages.

5. Extraction of Characteristics.

As noted earlier, the classification of text styles relies on two main categories of characteristics, morphological and structural. To automate the extraction of these characteristics, specialised programs are employed.

5.1. Extraction of Structural Characteristics

The structural characteristics mentioned above were measured via a custom-built program running under Linux. This program calculated all structural metrics for each text in a single pass and the results were processed with the help of a spreadsheet software package.

5.2. Extraction of Morphological Characteristics

The extraction of morphological characteristics is performed via the Automated Morphological Processor (AMP). The AMP has been designed to take advantage of the highly inflectional nature of the Greek language and allow the generation of morphological lexica in an automated manner (Tambouratzis & Carayannis, 1999). The generation of morphological lexica is a labour-intensive task, which requires several man-years to be completed. To encompass the language evolution throughout the ages, one would need to generate manually a large number of different morphological lexica, each one covering a specific **Language Evolution Sample** (hereafter denoted as LES) corresponding to a particular time-period in the

history of the Greek language and/or a geographical area. The use of the AMP allows the automated generation of morphological lexica with a large degree of accuracy. This is of particular importance in text retrieval and information extraction purposes from texts in the Greek language, where in order to represent fully the inflectional evolution of the language through time it is necessary to use two or three different morphological dictionaries.

The AMP is based on a rule-based iterative masking-and-matching principle, which relies on matching parts of different patterns while ignoring the remainders of these patterns. For example, if two patterns (here words) x_1x_2 and y_1y_2 are to be compared, the technique might focus on the possible similarity between x_1 and y_1 (attempting to match these parts) while ignoring x_2 and y_2 (temporarily masking off these parts).

This principle may be augmented by a modest amount of information regarding the specific LES being used, in order to optimise the accuracy of the morphological processing. Indeed, different approaches (involving various levels of a priori knowledge) have been followed to investigate the system behaviour (Tambouratzis & Carayannis, 1999). These indicate that the inclusion of a modest amount of LES-specific information (for example, a handful of a priori endings) can improve substantially the segmentation accuracy. The optimal morphological segmentation accuracy achieved by the AMP is equal to 96%, measured against the contents of the morphological lexicon, which has been constructed manually at the ILSP.

As noted, the AMP is able to operate on texts belonging to several different LESs. This is ensured by providing the AMP with a number of interchangeable modules, which may be selected according to the current LES requirements, while the core of the system remains the same. The interchangeable modules are of a declarative nature and may incorporate a priori knowledge.

For the purposes of text-style characterisation, it is necessary to calculate the frequencies-of-occurrence of specific endings, as detailed in section 3. Therefore the AMP is constrained to use the selected endings which are representative of style as a priori knowledge. For each of these endings, the frequency-of-occurrence as well as the corresponding root and its frequency-of-occurrence is recorded by the AMP. The results obtained for the given corpus are detailed in the following section.

6. Experiments.

6.1. Corpus Composition and Pre-processing

To study the performance of the proposed style-characterising system, a corpus has been created. This consists of three different types of text:

(i) excerpts from novels, which represent the Fiction register. These novels have been composed during the period 1988-1997 and are contained in the ILSP corpus of texts (Gavrilidou et al., 1998).

(ii) a collection of essays concerning the history of the Greek province of Macedonia, which represent the academic register. This collection of essays was created in 1992 and forms part of the ILSP corpus of texts. This collection is hereafter denoted as 'History'.

(iii) official transcripts from randomly-chosen sessions of the Greek Parliament, held during the period January 1999 - May 1999. This set of transcripts is hereafter denoted collectively as 'Parliament'.

To perform the experiments detailed in this article, all texts were spell-checked. A more extensive pre-processing was performed in the case of the Parliament register. The aim of this piece of research has been to process relatively large portions of text which provide a sufficiently clear view of the style of speech used at the Greek Parliament. Therefore, the texts contained in the parliament sub-corpus were processed and very short pieces of dialogue (which typically were related to the Parliament regulations) were removed from the transcripts. Such pre-processing need not be performed for the other two registers. The size and characteristics of the different registers of texts in the corpus are shown in table 1.

register	sample texts	size (in words)
Fiction	24	363,783
History	32	360,993
Parliament	12	509,917
Total	**68**	**1,234,693**

Table 1: Corpus composition in terms of register.

The tools detailed in section 5 were used to process each text contained in the corpus, generating a vector of characteristics. Since the texts are of varying sizes, it was decided to normalise them in order to remove possible

sources of variability prior to performing the statistical analysis. Thus, all characteristics were normalised so that the values reported corresponded to occurrences per 100,000 words of text. The values of these vectors are summarised in tables 2, 3 and 4, by providing for each register a set of average values and of the corresponding standard deviations for all the characteristics studied. In table 2, the frequencies of characteristic *Katharevousa* verb endings are detailed, while table 3 contains a similar table for *Demotiki* verb endings. It should be noted that these frequencies do not represent all verb endings but are restricted to the 230 endings that are grammatically representative of *Katharevousa* and *Demotiki*, respectively, as determined in section 4. Furthermore, as the use of a vector with 230 characteristics would be computationally intractable, in both tables 2 and 3, the endings are displayed collectively for each combination of the grammatical categories 'person' and 'number'. Tables 2, 3 and 4 provide a macroscopic view of the vectors of characteristics used to describe each of the texts contained in the corpus.

According to these collective results for the three registers (while being restricted to the set of 230 style-characteristic endings), certain patterns seem to emerge:

- In the Parliament register (tables 2 and 3), the verb endings corresponding to *Katharevousa* are as a whole more frequent than these of *Demotiki*, in comparison to the other registers.
- Interestingly, though, in the case of the 2nd person (both plural and singular) in the Parliament register the frequency of *Demotiki* endings is comparatively higher than that of *Katharevousa*. This is more marked in the 2nd person singular, where the frequency-of-occurrence of *Demotiki*-type endings is 50 times higher than that of the *Katharevousa*-type endings. On the contrary, the situation is reversed for the 1st and 3rd persons (both singular and plural), where *Katharevousa* endings are as a rule more frequent.
- In History, the majority of verb endings correspond to the third person (singular or plural), indicating a predominantly narrative style.
- The third person is also most frequently used in the Fiction register, though to a lesser extent than in the History register.
- Texts from the Fiction register have a considerably larger number of sentences and sub-sentences (if commas, brackets and dashes are collectively examined) as compared to the other two registers, which have a relatively high degree of similarity (see table 4).
- Collectively, the Fiction register shows a higher degree of variance over all characteristics than the other registers. The History register has a considerably lower variance of its characteristics while the Parliament register has the lowest variance, indicating possibly more tightly clustered representative vectors for its members. This, in conjunction with the previous observations, indicates the possible existence of a sub-language specific to politicians.

In the following paragraph, a more formal analysis of the data shall be provided using statistical methods. It should be noted that a selected subset of characteristics were used in this analysis, the aim of the vector being to study whether the three registers may be separated from each other with a sufficient degree of confidence. This is the subject of the following section.

	1st singular	2nd singular	3rd singular	1st plural	2nd plural	3rd plural
Fiction	157.2	33.5	55.7	103.8	2.0	123.3
	(108.5)	*(41.3)*	*(56.6)*	*(182.0)*	*(9.1)*	*(67.3)*
History	0.7	0.2	208.3	43.8	0.4	343.9
	(2.9)	*(1.1)*	*(129.8)*	*(51.0)*	*(1.9)*	*(140.2)*
Parliament	166.0	1.0	247.3	308.0	51.8	472.7
	(51.9)	*(1.6)*	*(62.1)*	*(38.0)*	*(20.6)*	*(100.8)*

Table 2: Average Frequency (in normal typescript) and standard deviation (in italics) of *Katharevousa* verb endings over 100,000 words, for each of the three text registers.

	1st singular	2nd singular	3rd singular	1st plural	2nd plural	3rd plural
Fiction	69.0	85.9	330.8	49.3	37.3	67.2
	(57.2)	*(65.8)*	*(154.2)*	*(74.3)*	*(64.2)*	*(45.6)*
History	0.0	37.7	228.3	8.9	0.0	153.0
	(0.0)	*(53.3)*	*(102.0)*	*(16.1)*	*(0.0)*	*(72.6)*
Parliament	50.4	67.0	180.2	82.2	48.6	77.9
	(32.9)	*(16.0)*	*(32.4)*	*(28.9)*	*(26.9)*	*(16.6)*

Table 3: Average Frequency (in normal typescript) and standard deviation (in italics) of *Demotiki* verb endings over 100,000 words, for each of the three text registers.

	sentences	commas	brackets	dashes	Katharevousa verb endings	Demotiki verb endings
Fiction	8649.0	7728.7	40.7	1793.8	475.5	639.5
	(2975.0)	*(1435.8)*	*(95.4)*	*(1386.2)*	*(294.43)*	*(223.5)*
History	4678.2	6834.9	849.1	495.2	597.37	427.93
	(1164.1)	*(1422.4)*	*(463.1)*	*(254.5)*	*(221.5)*	*(161.1)*
Parliament	5722.4	6726.0	56.2	570.1	1252.5	506.3
	(463.5)	*(442.0)*	*(49.5)*	*(69.8)*	*(137.7)*	*(49.5)*

Table 4: Average frequency (in normal typescript) and standard deviation (in italics) of macroscopic structural and morphological characteristics over 100,000 words, for each of the three text registers. The final two columns represent cumulative results for the measurements summarised in tables 2 and 3 respectively and are included here for reasons of clarity, though they do not represent independent measurements and thus are not used as vector elements.

6.2. Experimental Results.

To compare and contrast the three registers, a cluster analysis approach was selected. During the experiments, normalisation was performed over all vector parameters prior to the clustering operation, so that they occupied the same ranges of values. Additionally, different criteria (nearest, furthest, median, centroid and average) were studied when using non-seeded clustering experiments. It was found that all these clustering methods generated similar results and thus they shall not be distinguished henceforth, the results reported being those obtained by the majority of criteria.

Initially a non-seeded clustering approach was studied, to indicate the natural clusters existing in the corpus. When 6 clusters were used, five out of these each contained one or at most two texts from the Fiction register, while the sixth cluster contained the remaining texts from the Fiction register together with all texts from the History and Parliament registers.

Following that, 10 clusters were used to cluster the text vectors. This experiment showed that the register with the highest variability was the Fiction register, whose members occupied 9 out of the 10 available clusters. Out of the three registers, only the Parliament register was fully separated from the other registers, its members occupying a single cluster, which contained no texts from the other registers. Finally, all members of the History register occupied a single cluster, though this also contained texts belonging to the Fiction register.

These preliminary results indicated that the Parliament register was the one whose members were most closely spaced on in the pattern space defined by
the characteristics vector, confirming the conclusions of sub-section 6.1.. The History register was also relatively tightly-spaced, while the Fiction register contained several "outlier" members which were at a large distance from the majority of Fiction members. To evaluate these observations, a cluster analysis of the elements of each register was performed. This confirmed that the Parliament and History registers were the most tightly coupled. On the contrary, the Fiction register contained 5 elements, which were significantly different to the others.

Following this analysis, a seeded clustering approach was chosen. Initially, a cluster number of 6 was used, the clusters being equally distributed between the 3 registers, resulting in two clusters per register. Following that, the number of clusters was reduced to 4 (and then 3), in both cases one cluster being devoted to Parliament and History and two (and then one) being devoted to the Fiction register. These experiments were carried out using vectors consisting of:

- the endings frequencies for all persons in *Demotiki* and *Katharevousa* as well as the number of sentences, commas, parentheses and dashes (giving a 16-element vector);
- the endings frequencies for all persons in *Demotiki* and *Katharevousa* as well as the number of sentences and commas (giving a 14-element vector);
- the endings frequencies for all persons in *Demotiki* and *Katharevousa* (giving a 12-element vector consisting solely of morphological characteristics).

Seeds for the Parliament and History registers were chosen randomly. The seeds for the Fiction register were chosen so that at least one of them would not be an "outlier" of the Fiction register. Representative results are shown in table 5 for the different vectors and numbers of clusters. In each case, the classification rate quoted corresponds to the number of text elements correctly classified (according to the register of the respective seed).

	16-elem.	14-elem.	12-elem.
6 clust.	98.5%	92.7%	95.6%
4 clust.	98.5%	94.1%	97.1%
3 clust.	97.1%	92.7%	95.6%

Table 5: Clustering accuracy as a function of the number of clusters used and the size of the characteristic vector used.

This table indicates that the optimal clustering performance is obtained when the number of characteristics in the vector is equal to 16, that is when both structural and morphological information is considered. On the contrary, if part of the morphological information is suppressed (using a 14-element vector), the recognition rate is reduced (though it still remains around 93-94%). Indeed, the experimental results indicate that it is then probably beneficial to remove all structural information, recognition rate being higher in the case of 12-element vectors as opposed to 14-element vectors.

According to table 5, even when using 3 clusters the clustering result accurately represents the 3 registers, with very few misclassifications being performed. Indeed, the optimal results for each vector configuration are obtained using a 4 cluster classifier. This indicates that removing redundant clusters from the well-defined classes (that is, the Parliament and History registers) increases the recognition rate.

7. Discussion and Conclusions.

In this article, a system has been proposed for the automatic style categorisation of corpora in the Greek language. This categorisation is based on the type of language used in the text. To arrive to this categorisation, the highly inflectional nature of the Greek language is used. Characteristics reflecting this inflectional nature are combined with characteristics based on the structure of sentences to provide automatic style categorisation of texts.

The study of the corpus indicates that the three registers used here have different characteristics. The texts belonging to the Parliament register form a well-defined class in the pattern space. Similarly, historical texts also form a well-defined class, though to a slightly smaller extent than the Parliament register. On the contrary, texts belonging to the Fiction register form a less tight class. Thus, to successfully recognise this register, a comparatively large number of seeds is required to sufficiently cover the register space. However, as a whole, the seeded clustering approach using both 4 and 6 clusters achieves a recognition accuracy exceeding 98%. Thus, the presented method can be used to accurately define the register of a given text. Of course, these results have been obtained with a relatively constrained set of registers, but the results are of such a high accuracy as to support the conclusion that this line of work may lead to an accurate style-characterising system.

Future work on this area is planned to focus on confirming these results with larger-scale experiments. Additionally, there exist a number of parameters that may also be introduced in an effort to more accurately define the register of texts (such as negation parameters measured in (Markantonatou & Tambouratzis, 2000)). In this respect, the results obtained for the Fiction register are indicative of the existing scope for improvement. Indeed, techniques ranging from alternative statistical techniques to other pattern recognition methods (such as neural networks) are being considered for the recogniser part of this system.

8. Acknowledgements

The authors wish to acknowledge the assistance of the Secretariat of the Hellenic Parliament in obtaining the session transcripts studied in this piece of research.

9. References

Biber, D., 1993. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics,* Vol. 19, No. 2, pp. 219-241.

Biber, D., S. Conrad & R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press

Clairis, C. & G. Babiniotis. 1999. Grammar of Modern Greek – II Verbs. *Ellinika Grammata*, Athens (in Greek).

Gavrilidou, M., Labropoulou P., Papakostopoulou N., Spiliotopoulou S., Nassos N. 1998. Greek Corpus Documentation, Parole LE2-4017/10369, WP2.9-WP-ATH-1.

Holton D., P. Mackridge and I. Philippaki-Warburton. 1997. *Greek: A Comprehensive Grammar of the Modern Language.* Routledge, London and New York

Karlgren, J., 1999. Stylistic Experiments in Information Retrieval. In T. Strzalkowski (ed.), *Natural Language Information Retrieval,* pp. 147-166. Dordrecht: Kluwer.

Markantonatou, S. & Tambouratzis, G. 2000. Some quantitative observations regarding the use of grammatical negation in Modern Greek. To appear in *Proceedings of the 21st Annual Meeting of the Department of Linguistics*, Faculty of Philosophy of the Aristoteleian University of Thessaloniki, May 2000 (in print/in Greek)

Mikros, G. & Carayannis, G. (2000) Modern Greek Corpus Taxonomy. *Proceedings of the 2nd Conference on Language Resources and Evaluation*, Athens, Greece, 31 May-2 June (in print).

Tambouratzis, G. & Carayannis, G., 1999. Automated Construction of Morphological Lexica Possessing Terminology Wealth on the Basis of Term-Intensive Documents. *Proceedings of the Second Conference of Hellenic Language and Terminology*, Athens, 21-23 October 1999, pp. 149-155. ISBN 960-86069-1-8 (in Greek).

Παράρτημα II

Discriminating the registers and styles in the Modern Greek language

George Tambouratzis*, Stella Markantonatou*, Nikolaos Hairetakis*, Marina Vassiliou*, Dimitrios Tambouratzis^, George Carayannis*

* Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 151 25 Maroussi, Greece
{giorg_t, marks, nhaire, mvas, gkara} @ilsp.gr

^ Agricultural University of Athens,
Iera Odos 75, 118 55 , Athens, Greece.
{dtamb@aua.gr}

Abstract

This article investigates (a) whether register discrimination can successfully exploit linguistic information reflecting the evolution of a language (such as the diglossia phenomenon of the Modern Greek language) and (b) what kind of linguistic information and which statistical techniques may be employed to distinguish among individual styles within one register. Using clustering techniques and features reflecting the diglossia phenomenon, we have successfully discriminated registers in Modern Greek. However, diglossia information has not been shown sufficient to distinguish among individual styles within one register. Instead, a large number of linguistic features need to be studied with methods such as discriminant analysis in order to obtain a high degree of discrimination accuracy.

1. Introduction

The identification of the language style characterising the constituent parts of a corpus is very important to several applications. For example, in information retrieval applications, where large corpora of texts need to be searched efficiently, it is useful to have information about the language style used in each text, to improve the accuracy of the search (Karlgren, 1999). In fact, the criteria regarding language style may differ for each search and therefore – due to the large number of texts – there is a requirement to perform style categorisation in an automated manner. Such systems normally use statistical methods to evaluate the properties of given texts. The complexity of the studied properties varies. Kilgarriff (1996) employs mainly the frequency-of-occurrence of words while Karlgren (1999) applies statistical methods primarily on structural and part-of-speech information.

Baayen et al. (1996), who study the topic of author identification, apply statistical measures and methods on syntactic rewrite rules resulting by processing a given set of texts. They report that the accuracy thus obtained is higher than when applying the same statistical measures to the original text. On the other hand, Biber (1995) uses Multidimensional Analysis coupled with a large number of linguistic features to distinguish among registers. The underlying idea is that, rather than being distinguished on the basis of a set of linguistic features, registers are distinguished on the basis of combinations of weighted linguistic features, the so-called “dimensions”.

This article reports on the discrimination of texts in written Modern Greek. The ongoing research described here has followed two distinct directions. First, we have tried to distinguish among registers of written Modern Greek. In a second phase, our research has focused on distinguishing among individual styles within one register and, more specifically, among speakers of the Greek Parliament. To achieve that, structural, morphological and part-of-speech information

is employed. Initially (in section 2) emphasis is placed on distinguishing among the different registers used. In section 3, the task of author identification is tested with selected statistical methods. In both sections, we describe the set of linguistic features measured, we argue for the statistical method employed and we comment on the results. Section 4 contains a description of future plans for extending this line of research while in section 5 the conclusions of this article are provided.

2. Distinguishing Registers

To distinguish among registers, we successfully exploited a particular feature of Modern Greek, namely the contrast between Katharevousa and Demotiki. These are variations of Modern Greek which correspond (if only roughly) to formal and informal speaking. Katharevousa was the official language of the Greek State until 1979 when it was replaced by Demotiki. By that time, Demotiki was the established language of literature while, in times, it had been the language of elementary education. Compared to Demotiki, Katharevousa bears an important resemblance to Ancient Greek manifested explicitly on the morphological level and the use of the lexicon. At a second step, we dropped the Katharevousa-Demotiki approach and relied on part-of-speech information, which is often exploited in text categorisation experiments (for instance, see Biber et al. 1998). Again, we obtained satisfactory results.

2.1 Method of work

The variables used to distinguish among registers may be grouped into the following categories:

1. *Morphological variables*: These were verbal endings quantifying the contrast Katharevousa / Demotiki. Although the morphological differences between these two variations of Greek are not limited to the verb paradigm, we focused on the latter since it better highlights the contrast under consideration (Tambouratzis et al., 2000). A total of 230 verbal endings were selected, split into 145 Demotiki and 85 Katharevousa endings (see also the Appendix). These 230 frequencies-of-occurrence were grouped into 12 variables for use in the statistical analysis.
2. *Lexical variables*: Certain negation particles (*ουδείς, ουδέποτε, ουδαμού, άνευ*) clearly signify a preference for Katharevousa while others (*δίχως, μήτε, χωρίς*) are clear indicators of Demotiki. However, the most frequently used negation particles (*όχι, μην, δεν*) are not characteristic of either of the two variations.
3. *Structural macro-features*: average sentence length, number of commas, dashes and brackets (total of 4 variables).
4. After the completion of the experiments with variables of type 1-3 (Tambouratzis et al., 2000), Part-of-Speech (PoS) counts were introduced. The PoS categories were adjectives, adjunctions, adverbs, articles, conjunctions, nouns, pronouns, numerals, particles, verbs and a hold-all category (for non-classifiable entries), resulting in 11 variables expressed as percentages.

These variables are more similar to the characteristics used by Karlgren (1999), and differ considerably from those used by Kilgarriff (1996) and Baayen et al. (1996). For the metrics of the first and third categories, a custom-built program was used running under Linux. This program calculated all structural and morphological metrics for each text in a single pass and the results were processed with the help of a spreadsheet package. The metrics of the second category were calculated using a custom-built program in the C programming language. PoS counts were obtained using the ILSP tagger (Papageorgiou et al., 2000) coupled with a number of custom-built programs to determine the actual frequencies-of-occurrence from the tagged texts. Finally, the STATGRAPHICS package was used for the statistical analysis.

The dataset selected consisted of examples from three registers:

(i) fiction (364 Kwords - 24 texts),
(ii) texts of academic prose referring to historical issues, also referred to as the history register (361 Kwords – 32 texts) and
(iii) political speeches obtained from the proceedings of the Greek parliament sessions, also referred to as the parliament register (509 Kwords – 12 texts).

The texts of registers (I) and (II) were retrieved from the ILSP corpus (Gavrilidou et al., 1998), all of them dating from the period 1991-1999. The texts of register (III) were transcripts of the Greek Parliament sessions held during the first half of 1999.

This dataset was processed using both seeded and unseeded clustering techniques with between 3 and 6 clusters. The unseeded

approach confirmed the existence of distinct natural classes, which correspond to the three registers. The seeded approach confirmed the ability to accurately separate these three registers and to cluster their elements together. Initially, a 'short' data vector containing only the 12 morphological variables quantifying the Demotiki/Katharevousa contrast was used (Tambouratzis et al. 2000), as well as a 16-element vector combining structural and morphological characteristics. The seeds for the Parliament and History registers were chosen randomly. The seeds for the Fiction register were chosen so that at least one of them would not be an "outlier" of the Fiction register. Representative results are shown in Table 1 for the different vectors and numbers of clusters. In each case, the classification rate quoted corresponds to the number of text elements correctly classified (according to the register of the respective seed).

	12-elem.	16-elem.
6 clust.	95.6%	98.5%
4 clust.	97.1%	98.5%
3 clust.	95.6%	97.1%

Table 1 - Seeded clustering accuracy as a function of the cluster number and vector size.

The vector size was augmented with PoS information, resulting in a 27-element data vector. A new set of clustering experiments were performed using Ward's method with the squared Euclidean distance measure to cluster the data in an unseeded manner. Finally, a 15-element data vector was used with PoS and structural information but without any morphological information. The results obtained (Table 2) show that PoS information improves the clustering performance.

2.2 Comments on the Results

Our results strongly suggest that registers of written Modern Greek can be discriminated accurately on the basis of the contrast Katharevousa / Demotiki manifested with morphological variation. Languages with a different history may not be suited to such a categorisation method. This is evident in Biber's work (1995) for the English language, where a variety of grammatical and macro-structural linguistic features but no morphological variation features were employed. It seems then that corpora of languages which are characterised by the phenomenon of diglossia, may be successfully categorisable on the basis of morphological information (or other reflexes of diglossia). Such a discrimination method may give results as satisfactory as approaches which are closer to the Biber (1995) spirit and rely on PoS and structural measures (see Tables 1 and 2).

Tables 1 and 2 show that the accuracy of clustering reaches approximately 99% while the seeded clustering approach had a high degree of accuracy, reaching 100% when using 5 clusters. For the 27-element vector with both morphological and PoS information, perfect clustering has been achieved even with 4 clusters. On the other hand, a successful clustering (albeit with a lower level of accuracy) is achieved using only structural and PoS information.

It should be noted that the lexical variables used, that is the negation particles, did not contribute at all (Markantonatou et al., 2000). Furthermore, the system performed almost as well with and without macro-structure features, the difference in accuracy being less than 5%.

The parliament texts can be claimed to form a register whose patterns are closely positioned in the pattern space. Of the three registers, the literature one presented the highest degree of variance, with more than one sub-clusters existing as well as outlier elements. This may be explained by the fact that the parliament proceedings, contrary to literature, undergo intensive editing by a small group of specialised public servants.

3. Distinguishing Styles within One Register

In this section, we report on our efforts to distinguish among individual styles within one register. In particular, we intend to distinguish among speakers of the Parliament by studying the transcripts of the speeches of five parliament members over the period 1997-2000. Each of these speakers belongs to one of the five political parties that were represented in the Greek parliament over that period. Up to date, the experiments have been limited to the period 1999-2000.

3.1 Method of work

The number of variables (46 in total) calculated for each of the five speakers can be grouped as follows:

	12-elem.	16-elem.	27-elem.	15-elem
6 clust.	95.5%	100.0%	100.0%	100.0%
5 clust.	95.5%	100.0%	100.0%	89.6%
4 clust.	94.1%	98.5%	100.0%	83.4%
3 clust.	94.1%	98.5%	98.5%	83.4%

Table 2 - Unseeded clustering accuracy as a function of the cluster number and vector size used.

1. *Morphological variables* (20 variables):
   - Verbal endings expressing the Katharevousa / Demotiki contrast giving rise to 12 variables.
   - the use of infixes (2 variables) in the past tense forms.
   - the person and number of the verb form (6 variables).

   The last two types of variable are expressed as percentages normalised over the number of verb forms.
2. *Lexical variables* (6 variables):
   - Negation particles (*όχι, δεν, μην*).
   - Negative words of Katharevousa (*ουδείς, άνευ*).
   - Other words which also express the contrast Katharevousa / Demotiki (the anaphoric pronouns '*οποίος*' (Kath) and '*που*' (Dem)), currently resulting in a single variable.
3. *Structural macro-features*: average sentence and word length, number of commas, question marks, dashes and brackets, resulting in a total of 6 variables.
4. *Structural micro-features* (other than lexical):
   - Part-of-Speech counts (10 variables).
   - Use of grammatical categories such as the genitive case with nouns and adjectives (2 variables).
5. The year when the speech was presented in the Parliament and the order of the speech in the daily schedule, that is whether it was the first speech of the speaker that day (hereafter denoted as "protoloyia") or the second, third etc. (resulting in a total of 2 variables).
6. The identity of the speaker, denoted as the speaker Signature (1 variable), which was used to determine the desired classification.

Similarly to the clustering experiments, a set of C programs was used to extract automatically the values of the aforementioned variables from the transcripts. Most of these programs rely on measuring the occurrence of di-grams, and more generally n-grams, for letters, words and tagsets, thus being straight-forward. In the case of speaker identification, Discriminant Analysis was used, as the clustering approach did not give very good results, indicating that the distinction among personal styles is weaker than that among registers. Even when only 2 speakers were used, the clusters formed involved patterns from both speaker classes.

We experimented with two corpora, Corpus I and Corpus II, as described in Table 3. Corpus II is a subset of Corpus I. Each of the speeches included in Corpus II was delivered as an opening speech ("protoloyia") at a parliament session when at least two of the studied speakers delivered speeches.

An important issue is whether the selected variables are strongly correlated. If indeed strong correlations do exist, these might be used to reduce the dimensionality of the pattern space. For the purposes of this analysis, the 46 independent variables were used (45 in the case of Corpus II where only "protoloyiai" exist, since then the order variable is constantly equal to 1). The number of correlations of all variable pairs exceeding given thresholds is depicted in Figure 1, for both Corpus I and Corpus II. According to this study, in Corpus II, the percentage of variable pairs with an absolute value of correlation exceeding 0.5 is approximately 3%, indicating a low correlation between the parameters. Additionally, out of 990 pairs of Corpus II, only a single one has a correlation exceeding 0.8. The correlations for the same parameter pairs over the two corpora are similar, though as a rule the correlation for Corpus I is less than that for Corpus II, reflecting the larger variability of texts in Corpus I. The correlation study indicated that most of the parameters are not strongly correlated. Thus, a factor analysis step is not necessary and the application of the discriminant analysis directly on the original variables is justified.

Initially, Corpus I (see Table 3) was processed. The 46 aforementioned variables were used to generate discriminant functions accurately recognising the identity of the speaker. To that end, three different approaches were used:

(i) the full model: all variables were used to determine the discriminant functions;
(ii) the forward model: starting from an empty model, variables were introduced in order to create a reduced model, with a small number of variables;
(iii) the backward model: starting from the full model, variables were eliminated to create a reduced model.

In the cases of the forward and backward models, the values of the F parameter to both enter and delete a variable were set to 4 while the maximum number of steps to generate the model was set to 50.

The performance of this model is improved if:

Year	1999-2000	
Speaker	Corpus I	Corpus II
A	92	30
B	45	24
C	33	21
D	21	16
E	150	36

Table 3 – Comparative composition of Corpus I and Corpus II.

1. the order in which each particular speech was delivered is taken into account: the subset of "protoloyiai" is well-defined and presents a low variance while the speeches of second or lower order have a higher variance.
2. the corpus comprises only sessions where more than one speaker has delivered speeches. Thus, the more balanced Corpus II (Table 3) presents an improved discrimination performance.

For these two corpora, the results of the discriminant analysis are shown in Table 4. The discrimination rate obtained with Corpus II is much higher than that for Corpus I. In addition, smaller models, with 8 variables, may be created that correctly classify at least 75% of Corpus II. An example of the factors generated and the manner in which they separate the pattern space is shown in the diagrams of Figure 2.

3.2 Comments on the Results

Though this research is continuing, certain facts can be reported with confidence.

Within the Greek Parliament Proceedings register, individual styles can not be classified on the basis of morphological features expressing the contrast Katharevousa / Demotiki. This may be explained by the fact that these texts undergo intensive editing towards a well-established sub-language. This editing homogenises the morphological profile of the texts but, of course, does not go as far as homogenising the lexical preferences of the various speakers. That is why, contrary to the register-clustering experiments, lexical variables expressing the particular contrast seem to play a role in discriminating between speakers and why the use of Katharevousa-oriented negative particles, which was not important in register discrimination, seems to be of some importance in style discrimination. The observation that negative words play a role in style identification is in agreement with the observations of Labbé (1983) on the French political speech.

Structural features have turned out to be important: the average word length, the use of punctuation and question marks and the use of certain parts-of-speech such as articles, conjunctions, adjunctions and - especially - verbs. Furthermore, the distribution of verbs into persons and numbers seems to be important, though the exact variables selected differ depending on the exact set of speeches used (these variables are of course complementary).

One of the most interesting findings of this research is that it is important whether the speaker delivers a "protoloyia" or not. "Protoloyiai" can be classified at a rate of 95% while mixed deliveries result in a lower rate, as low as 75%. This may be caused by two factors:

1. "Protoloyiai" represent longer stretches of text, which are more characteristic of a given speaker.
2. Speakers prepare meticulously for their "protoloyiai" while their other deliveries represent a more spontaneous type of speech, which tends to contain patterns shared by all the parliament members.

Finally, certain additional patterns are emerging for each of the speakers. Certain speakers (e.g. speaker A) are more consistently recognised than others (e.g. speaker B) while speaker B is similar to speaker C and speaker D is similar to speaker E. This indicates that additional variables may be required to improve the classification accuracy for all speakers.

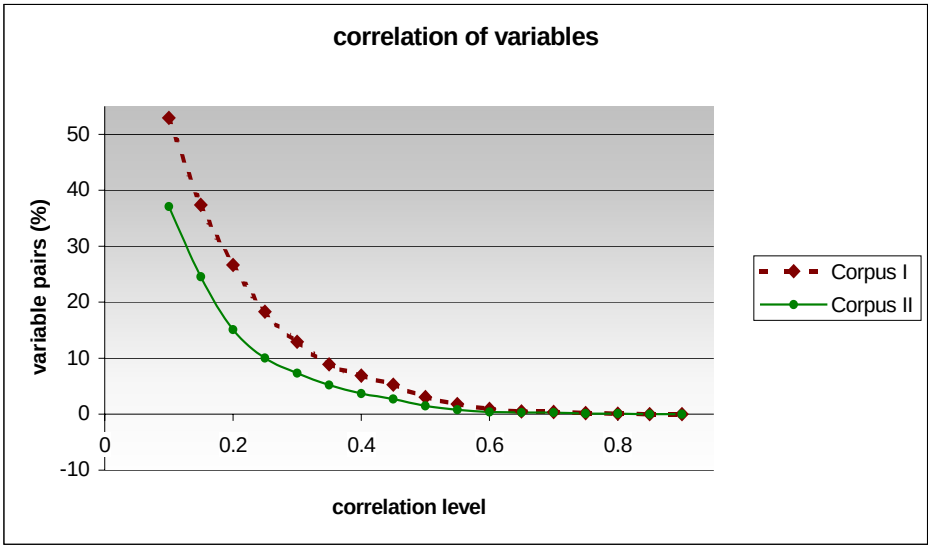


Figure 1 – Percentage of variable pairs exceeding a given level of absolute correlation.

Dataset	Model			observations
	full	forward	backward	
Corpus I	93.79 % *(46)*	75.37 % *(46)*	78.30 % *(46)*	341
Corpus II	97.64 % *(45)*	94.49 % *(13)*	92.91 % *(20)*	127
Corpus II (reduced model)	97.64 % *(45)*	87.40 % *(8)*	79.53 % *(8)*	127

Table 4 – Discrimination rate (the corresponding model size is shown in italics).

4. Future Plans

As a next step, frequency of use of certain lemmata shall be introduced since visual inspection indicates that they may provide good discriminatory features. We also plan to substitute average lengths (of both words and sentences) with the distribution of lengths. Furthermore, we intend to introduce certain structural measurements such as repetition of structures, chains of nominals and the occurrence of negation within NP phrasal constituents. Another possible extension involves the inclusion of the speech topic. As certain speakers' characteristics seem to change through time, we plan to process the entire corpus of speeches for the target period 1997-2000. Finally, an important issue is the comparison of the results obtained in our experiments to these generated by alternative techniques proposed by other researchers. This will allow the deduction of more accurate conclusions regarding the strengths and the weaknesses of the research strategies.

5. Conclusions

In this article, ongoing research on register and individual style categorisation of written Modern Greek has been reported. A system has been proposed for the automatic register categorisation of corpora in Modern Greek exploiting the highly inflectional nature of the language. The results have been obtained with a relatively constrained set of registers; however their recognition accuracy is remarkably high, exceeding 98% with an unseeded clustering approach using between 3 and 6 clusters.

On the front of individual style categorisation, a discrimination rate of over 80% was achieved for five speakers within the Greek Parliament register. Morphological variables were shown to be of less importance to this task, while lexical and structural variables seemed to take over. We are planning to introduce several new lexical and structural variables in order to achieve better

discrimination rates and to determine discriminating features of the different styles.

Acknowledgements

The authors wish to thank the Dept. of Language Technology Applications and specifically Dr. Harris Papageorgiou and Mr. Prokopis Prokopidis in obtaining the lemmatised versions of the parliament transcripts. Additionally, the authors wish to acknowledge the assistance of the Secretariat of the Hellenic Parliament in obtaining the session transcripts.

References

Baayen, R. H., van Halteren, H. and Tweedie, F. J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, Vol. 11, No. 3, pp. 121-131.

Biber, D. (1995) *Dimensions of Register Variation: A cross-linguistic comparison.* Cambridge University Press.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press.

Clairis, C. & Babiniotis, G. (1999) Grammar of Modern Greek – II Verbs. *Ellinika Grammata*, Athens (in Greek).

Gavrilidou, M., Labropoulou P., Papakostopoulou N., Spiliotopoulou S., Nassos N. (1998) Greek Corpus Documentation, Parole LE2-4017/10369, WP2.9-WP-ATH-1.

Holton D., Mackridge, P. & Philippaki-Warburton, I. (1997) *Greek: A Comprehensive Grammar of the Modern Language.* Routledge, London and New York.

Karlgren, J., (1999) Stylistic Experiments in Information Retrieval. In T. Strzalkowski (ed.), *Natural Language Information Retrieval,* pp. 147-166. Dordrecht: Kluwer.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition,* Sussex University, April , pp. 33-40.

Labbé, D. (1983). *François Mitterrand. Essai sur le discours*. La Pensée Sauvage, Grenoble.

Markantonatou, S. & Tambouratzis, G. (2000) Some quantitative observations regarding the use of grammatical negation in Modern Greek. *Proceedings of the 21st Annual Meeting of the Department of Linguistics*, Faculty of Philosophy of the Aristotelian University of Thessaloniki, May 2000 (in print/in Greek).

Papageorgiou, H., Prokopidis, P., Giouli, V. & Piperidis, S. (2000) A Unified PoS Tagging Architecture and its application to Greek. *Proceedings of the 2nd International Conference on Language Resources and Evaluations*, Athens, Greece, 31 May - 2 June, Vol. 3, pp. 1455-1462.

Tambouratzis, G., Markantonatou, S., Hairetakis, N. & Carayannis, G. (2000) Automatic Style Categorisation of Corpora in the Greek Language. *Proceedings of the 2nd International Conference on Language Resources and Evaluations*, Athens, Greece, 31 May - 2 June, Vol. 1, pp. 135-140.

APPENDIX

Characteristics of Katharevousa and Demotiki

Diglossia in Modern Greek is due to the contrast between *Katharevousa* and *Demotiki* and is well-manifested on the morphological level. Here we concentrate on verb morphology.

Demotiki tends to have words ending with an 'open' syllable. So, *3rd Plural* verbal endings in *–n* (1) are augmented to *–ne* (2).

(1) *έλεγαν* [e'leγan] *(Kath)* (=they said)

(2) *λέγανε* [le'γane] *(Dem)* (=they said)

In *Demotiki*, *Katharevousa's* consonant clusters of two fricatives or two plosives are converted into clusters of one fricative and one plosive (3) – (4) (Holton et al., 1997, pp. 14).

(3) *πεισθώ* [pisθo'] */πειστώ* [pisto'] (=to be convinced)

(4) *καλυφθώ* [kalifθo'] */καλυφτώ* [kalifto'] (=to be covered)

Certain verb classes exhibit thematic vowel alternations either following the inflectional paradigm of Ancient Greek or Demotiki (5) (Clairis and Babiniotis, 1999).

(5) εξαρτάται [eksarta'te] *(Kath) / εξαρτιέται* [eksartiete] *(Dem)* (=depends)

Sometimes *Demotiki* uses a verbal root, which is similar though not identical to the *Katharevousa* one (6).

(6) λύω [li'o] *(Kath) / λύνω* [li'no] *(Dem)* (=to solve)

Finally, many verbs inherited from *Katharevousa* survive in *Demotiki*, either having an equivalent –mainly colloquial- (7) or not (8) (Clairis and Babiniotis, 1999).

(7) προτίθεμαι [proti'θeme] *(Kath) / σκοπεύω* [skope'vo] *(Dem)* (=I intend to)

(8) προΐσταμαι [proi'stame] (=supervise)

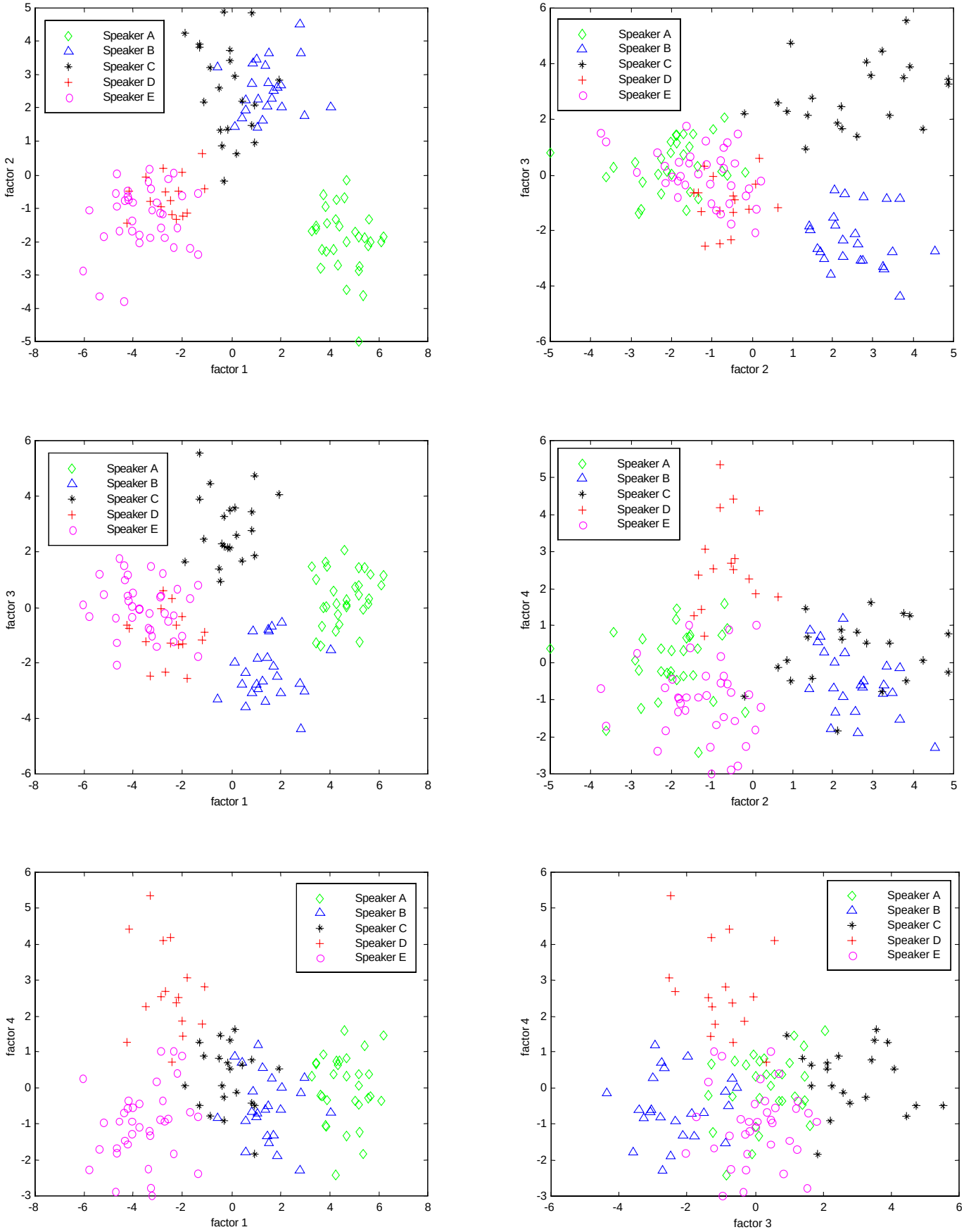


Figure 2 – Discriminant factors plotted against the patterns for corpus II.

Παράρτημα ΙΙΙ

Οι 230 (145 καταλήξεις της λόγιας-αρχαΐζουσας και 85 της δημοτικής) χαρακτηριστικές καταλήξεις ρημάτων οι οποίες χρησιμοποιήθηκαν για να μπορέσουμε να διακρίνουμε την ρηματική πολυτυπία είναι οι ακόλουθες :

Λόγια	***Αρχαΐζουσα***	***Δημοτική***
χθηκα	χθην	χτηκα
χθηκες	χθης	χτηκες
χθηκε	χθη	χτηκε
χθήκαμε	χθημεν	χτήκαμε
χθήκατε	χθητε	χτήκατε
χθηκαν χθήκαν χθήκανε	χθησαν	χτηκαν χτήκαν χτήκανε

Λόγια	***Αρχαΐζουσα***	***Δημοτική***
φθηκα	φθην	φτηκα
φθηκες	φθης	φτηκες
φθηκε	φθη	φτηκε
φθήκαμε	φθημεν	φτήκαμε
φθήκατε	φθητε	φτήκατε
φθηκαν φθήκαν φθήκανε	φθησαν	φτηκαν φτήκαν φτήκανε

Λόγια	***Αρχαΐζουσα***	***Δημοτική***
σθηκα	σθην	στηκα
σθηκες	σθης	στηκες
σθηκε	σθη	στηκε
σθήκαμε	σθημεν	στήκαμε
σθήκατε	σθητε	στήκατε
σθηκαν σθήκαν σθήκανε	σθησαν	στηκαν στήκαν στήκανε

Λόγια		***Δημοτική***
εικνύω		είχνω
εικνύεις		είχνεις
εικνύει		είχνει
εικνύουμε		είχνουμε είχνομε
εικνύετε		είχνετε
εικνύουν		είχνουν
εικνύομαι		είχνομαι
εικνύεσαι		είχνεσαι
εικνύεται		είχνεται
εικνυόμαστε		ειχνόμαστε
εικνύεστε είχνεσθε		είχνεστε
εικνύονται		είχνονται

Λόγια		
ιγνύω		
ιγνύεις		
ιγνύει		
ιγνύουμε		
ιγνύετε		
ιγνύουν		
ιγνύομαι		
ιγνύεσαι		
ιγνύεται		
ιγνυόμαστε		
ιγνύεσθε		
ιγνύονται		

Λόγια		***Δημοτική***
ώμαι		ιέμαι
άσαι		ιέσαι
άται		ιέται
ώμεθα		ιόμαστε
άσθε		ιέστε
ώνται		ιόνται ιώνται

Λόγια		***Δημοτική***
ώ		ώνω
οίς		ώνεις
οί		ώνει
ούμε		ώνουμε ώνομε
ούτε		ώνετε
ούν		ώνουν

Λόγια		***Δημοτική***
ούμην		ούμουν
είσο		ούσουν ούσουνα
είτο		ούνταν
ούμεθα		ούμασταν ούμαστε
είσθε ούσθε		ούσασταν είστε ούστε
ούντο		ούνταν

Λόγια	***Αρχαΐζουσα***	
σταμαι	στάμην	
στασαι	στασο	
σταται	στατο	
στάμεθα	στάμεθα	
στασθε	στασθε	
στανται	σταντο	

Λόγια	***Αρχαΐζουσα***	
θεμαι	θεμην	
θεσαι	θεσο	
θεται	θετο	
θέμεθα	θέμεθα	
θεσθε	θεσθε	
θενται	θεντο	

Λόγια		
κειμαι		
κεισαι		
κειται		
κείμεθα		
κεισθε		
κεινται		

Λόγια		
θημαι		
θησαι		
θηται		
θήμεθα		
θησθε		
θηνται		

Λόγια		***Δημοτική***
φθώ		φτώ
φθείς		φτείς
φθεί		φτεί
φθούμε		φτούμε
φθείτε		φτείτε
φθούν φθούνε		φτούν φτούνε

Λόγια		***Δημοτική***
χθώ		χτώ
χθείς		χτείς
χθεί		χτεί
χθούμε		χτούμε
χθείτε		χτείτε
χθούν χθούνε		χτούν χτούνε

Λόγια		***Δημοτική***
σθώ		στώ
σθείς		στείς
σθεί		στεί
σθούμε		στούμε
σθείτε		στείτε
σθούν σθούνε		στούν στούνε

Λόγια		***Δημοτική***
όμεθα		όμαστε
εσθε		εστε
		όσαστε

Παράρτημα IV

Οι εντολές που χρησιμοποιήσαμε στο περιβάλλον της Matlab για την απεικόνιση των χαρτών είναι οι ακόλουθες:

```
sD=som_read_data('data.txt');
sD=som_normalize(sD,'var');
sM=som_make(sD,'msize',[k1 k2]);
sM=som_autolabel(sM,sD,'add1');
som_show(sM,'umat','all','comp',1:x,'empty','Labels','norm','d');
som_show_add('label',sM,'subplot',x+2);
som_show(sM,'umat','all','empty','Labels');
som_show_add('label',sM,'Textsize',8,'Textcolor','r','subplot',2);
```

Η εντολή som_read_data διαβάζει τα δεδομένα από το ASCII αρχείο data.txt και στη συνέχεια η som_normalize τα κανονικοποιεί. Η κανονικοποίηση αυτή γίνεται μέσω του γραμμικού μετασχηματισμού $x' = \frac{x - \overline{x}}{\sigma_x}$, όπου $\overline{x}$ είναι ο μέσος της μεταβλητής x και σ_x η τυπική απόκλιση της. Η εντολή som_make αρχικοποιεί και εκπαιδεύει τον χάρτη μεγέθους k_1 x k_2 ενώ η εντολή som_autolabel δίνει αυτόματα ετικέτες στο χάρτη και τα στοιχεία του. Οι εντολές som_show και som_show_add χρησιμοποιούνται για την απεικόνιση του χάρτη. Τέλος, στις περιπτώσεις που θελήσαμε να κάνουμε τυχαία αρχικοποίηση του SOM χρησιμοποιήσαμε και την εντολή som_randinit.

Παράρτημα V

Οι 29 μεταβλητές οι οποίες παρουσιάζονται στο σχήμα 3.1 είναι κατά σειρά εμφάνιση τους οι εξής :

1. Προτάσεις
2. Κόμματα
3. Παρενθέσεις
4. Παύλες
5. 1ο Ενικό Λόγιο
6. 2ο Ενικό Λόγιο
7. 3ο Ενικό Λόγιο
8. 1ο Πληθυντικό Λόγιο
9. 2ο Πληθυντικό Λόγιο
10. 3ο Πληθυντικό Λόγιο
11. 1ο Ενικό Δημοτική
12. 2ο Ενικό Δημοτική
13. 3ο Ενικό Δημοτική
14. 1ο Πληθυντικό Δημοτική
15. 2ο Πληθυντικό Δημοτική
16. 3ο Πληθυντικό Δημοτική
17. Σύνολο Λόγιας
18. Σύνολο Δημοτικής
19. Επίθετα
20. Προθέσεις
21. Επιρρήματα
22. Άρθρα
23. Σύνδεσμοι
24. Ουσιαστικά
25. Αριθμητικά
26. Μόρια
27. Αντωνυμίες
28. Υπόλοιπα
29. Ρήματα

Παράρτημα VI

Οι 85 μεταβλητές οι οποίες παρουσιάζονται στο σχήμα 4.5 είναι κατά σειρά εμφάνισης οι εξής :

1. Γράμματα ανά Λέξη
2. Λέξεις ανά Πρόταση
3. Κόμματα ανά Πρόταση
4. Παρενθέσεις
5. Παύλες
6. 1ο Ενικό Λόγιο
7. 2ο Ενικό Λόγιο
8. 3ο Ενικό Λόγιο
9. 1ο Πληθυντικό Λόγιο
10. 2ο Πληθυντικό Λόγιο
11. 3ο Πληθυντικό Λόγιο
12. 1ο Ενικό Δημοτική
13. 2ο Ενικό Δημοτική
14. 3ο Ενικό Δημοτική
15. 1ο Πληθυντικό Δημοτική
16. 2ο Πληθυντικό Δημοτική
17. 3ο Πληθυντικό Δημοτική
18. Σύνολο Λόγιας
19. Σύνολο Δημοτικής
20. Επίθετα
21. Προθέσεις
22. Επιρρήματα
23. Άρθρα
24. Σύνδεσμοι
25. Ουσιαστικά
26. Αριθμητικά
27. Μόρια
28. Αντωνυμίες
29. Υπόλοιπα
30. Ρήματα
31. Γενική Επιθέτων
32. Γενική Ουσιαστικών
33. 1ο Ενικό Ρημάτων
34. 2ο Ενικό Ρημάτων
35. 3ο Ενικό Ρημάτων
36. 1ο Πληθυντικό Ρημάτων
37. 2ο Πληθυντικό Ρημάτων
38. 3ο Πληθυντικό Ρημάτων
39. όχι
40. ουδείς
41. ουδέποτε
42. ουδαμού
43. άνευ
44. δε(ν)
45. μη(ν)
46. Σημεία στίξης
47. Ερωτηματικά
48. οποίος/που
49. α
50. β
51. ουδ*********
52. 1-3 γράμματα ανά λέξη
53. 4-8 γράμματα ανά λέξη
54. 9-15 γράμματα ανά λέξη
55. 16-20 γράμματα ανά λέξη
56. 21-30 γράμματα ανά λέξη
57. 1 λέξη ανά πρόταση
58. 2-5 λέξεις ανά πρόταση
59. 6-10 λέξεις ανά πρόταση
60. 11-15 λέξεις ανά πρόταση
61. 16-20 λέξεις ανά πρόταση
62. 21-25 λέξεις ανά πρόταση
63. 26-30 λέξεις ανά πρόταση
64. 31-40 λέξεις ανά πρόταση
65. 41-50 λέξεις ανά πρόταση
66. 51-75 λέξεις ανά πρόταση
67. 76-100 λέξεις ανά πρόταση
68. 101-150 λέξεις ανά πρόταση
69-85 Λήμματα

Ευρετήριο

Automated Morphological Processor, 13
Backward Stepwise Analysis, 18
Cluster analysis, 16, 32, 33, 56
Discriminant analysis, 9, 16, 17, 57, 59, 60, 99
Empirical learning of term-category associations, 9
Factor analysis, 9
Forward Stepwise Analysis, 17, 18
Information retrieval, 3
k-means, 19
Learning Vector Quantization, 19
Linear regression, 20
Matching and masking technique, 14
Multiple Cause Mixture Model, 8, 103
Non-seeded analysis, 33
Pattern recognition, 14
Precision, 2
Recall, 2
Seeded analysis, 33
Self Organizing Map, 18
Style, 4
Tagger, 15, 16, 27, 43
Thesaurus-based matching, 8
U-matrix, 37, 63
Word-matching, 8
Αναγνώριση προτύπων, 14, 21
Αναζήτηση πληροφοριών, 2, 3, 4, 7
Ανάλυση με αρχικοποίηση, 33
Ανάλυση σε ομάδες, 16, 33
Ανάλυση χωρίς αρχικοποίηση, 33
Αρνήσεις, 7
Γραμμική παλινδρόμηση, 20
Δημοτική, 6, 24, 25, 26, 46, 89, 90, 91, 92, 94, 95
Διακριτική ανάλυση, 16, 17
Διγράμματα, 46
Δομικά χαρακτηριστικά, 6, 26, 31, 34, 36, 50, 55, 66
Εμπρός Βηματική Ανάλυση, 17
Ευρετήριο, 1, 2
Ιστορικός λόγος, 10, 32
Καθαρεύουσα, 7, 24, 25, 28, 46, 55
Κανόνας εκμάθησης Kohonen, 20
Λημματοποιητής, 16
Μορφολογικά χαρακτηριστικά, 25, 34, 43, 66
Πίσω Βηματική Ανάλυση, 18
Πολιτικός λόγος, 10, 32
Πολυτυπία, 5, 6, 23, 24, 55, 89
Συσχέτιση, 57
Ταίριασμα με βάση λέξεις, 8
Ταίριασμα με βάση λεξικό συνωνύμων, 8
Ταξινόμηση κειμένων, 4
Ταύτιση και απόκρυψη, 14
Ύφος, 4, 5, 6, 26, 29

Βιβλιογραφία

[AM95] Αγλαμίσης, Ι. και Μάντζαρη, Ε. 1995. "*Μορφολογικό Λεξικό – Γραμματικός Χαρακτηριστής*". Κείμενο Εργασίας ELL/STD/DEPT/9501, Ινστιτούτο Επεξεργασίας του Λόγου, Αθήνα.

[CG95] Kenneth Church, William Gale. 1995. "Inverse Document Frequency (IDF): A Measure of Deviations from Poisson". In *Proceedings of the Third Workshop on Very Large Corpora (ACL-95)*, pp. 121-130, MIT, Cambridge, Massachusetts, 30 June.

[DKM+98] Stephen D'Alessio, Aaron Kershenbaum, Keitha Murray and Robert Schiaffino. 1998. "Hierarchical Text Categorization".

[DM96] Venu Dasigi and Reinhold C. Mann. 1996. "Neural Net Learning Issues in Classification of Free Text Documents". In *Machine Learning in Information Access, Papers from the AAAI Spring Symposium,* pp.101-103, March 25-27, Stanford.

[Gas94] Michael Gasser. 1994. "Modularity in a Connectionist Model of Morphology Acquisition". In *Coling 94 / The 15th International Conference on Computational Linguistics*, Vol. 1, August 5-9, Kyoto, Japan.

[Gre93] Gregory Grefenstette. 1993. "Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques". *Technical Report MLTT-001*, Rank Xerox Research Center.

[HG96] David A. Hull and Gregory Grefenstette. 1996. "A Detailed Analysis of English Stemming Algorithms". *Technical Report MLTT-023*, Rank Xerox Research Center.

[Hud94] Richard Hudson. 1994. "About 37% of word-tokens are nouns". In *LANGUAGE,* Vol. 70, No 2, pp. 331-339.

[Hyo96] Heikki Hyotyniemi. 1996. "Text Document Classification with Self-Organizing maps". In SteP '96-Genes, Nets and Symbols, edited by Alander, J., Honkela, T., and Jakobsson, M., Finnish Artificial Intelligence Society, pp. 64-72.

[Kar96] Jussi Karlgren. 1996. "Stylistic Variation in an Information Retrieval Experiment". In *Proceedings of the 2nd International Conference on New Methods in Natural Language Processing*, Bilkent University, Ankara, Turkey, September.

[Kar99] Jussi Karlgren. 1999. "Stylistic Experiments in Information Retrieval". In Tomek Strzalkowski (editor), *Natural Language Information Retrieval*, pp. 147-166, Kluwer, Boston.

[Kar00] Jussi Karlgren. 2000. "Stylistic Experiments for Information Retrieval". *PhD Dissertation*, Stockholm University, Department of Linguistics.

[KBD+98] Jussi Karlgren, Ivan Bretan, Johan Dewe, Anders Hallberg and Niklas Wolkert. 1998. "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres". In Preben Hansen (editor), *Proceedings of Eighth DELOS Workshop on User Interfaces in Digital Libraries*, pp. 85-92, Langholmen. ERCIM.

[KC94] Jussi Karlgren and Douglass Cutting. 1994. "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In *Proceedings of the 15th International Conference on Computational Linguistics*, Vol. 2, pp. 1071-1075, Kyoto, Japan, August. ICCL.

[Kil96] Adam Kilgariff. 1996. "Which words are particularly characteristic of a text ? A survey of statistical approaches". *In Language Engineering for Document Analysis and Recognition*, pp. 33-40, Brighton, England, April. AISB Workshop Series.

[KK96] Samuel Kaski and Teuvo Kohonen. 1996. "Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World". In Apostolos-Paul N. Refenes, Yaser Abu Mostafa, John Moody and Andreas Weigend (editors), *Neural Networks in Financial Engineering. Proceedings of the 3rd International Conference on neural Networks in the Capital Markets*, London, England, 11-13 October 1995, pp. 498-507. World Scientific, Singapore.

[KM99] Κλαίρης Χρήστος και Μπαμπινιώτης Γιώργος. 1999. "*Γραμματική της Ελληνικής. Δομολειτουργική-επικοινωνιακή. ΙΙ. Το Ρήμα. Η οργάνωση του μηνύματος*". Ελληνικά Γράμματα, Αθήνα.

[Koh82] Teuvo Kohonen. 1982. "Self-organized formation of topologically correct feature maps". In *Biological Cybernetics*, vol. 43, No 1, pp. 59-69.

[Koh90] Teuvo Kohonen. 1990. "The Self-Organizing Map". In *Proceedings of the IEEE*, Vol. 78, No. 9, pp. 1464-1480.

[KOS+96] Teuvo Kohonen, Erkki Oja, Olli Simula, Ari Visa and Jari Kangas. 1996. "Engineering Applications of the Self-Organizing Map". In *Proceedings of the IEEE*, Vol. 84, No. 10, pp. 1358-1384.

[KR98] Adam Kilgariff and Tony Rose. 1998. "Measures for corpus similarity and homogeneity". In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, pp. 46-52.

[KS91] Γ. Καραγιάννης και Γ. Σταϊνχάουερ. 1991. "Αναγνώριση προτύπων και μηχανές που μαθαίνουν". Εκδόσεις Συμεών.

[Lip87] Richard P. Lippmann. 1987. "An Introduction to Computing with Neural Nets". In *IEEE ASSP MAGAZINE*, Vol. 4, No 2, pp. 4-22.

[LT96] Ray Liere and Prasad Tadepalli. 1996. "The Use of Active Learning in Text Categorization". In *Working Notes of the AAAI 1996 Spring Symposium on Machine Learning in Information Access*, Stanford, CA.

[Man86] Bryan F.J. Manly. 1986. "Multivariate Statistical Methods – A primer". Chapman and Hall. ISBN 0-412-28620-3.

[Mer97] Dieter Merkl. 1997. "Lessons Learned in Text Document Classification". In *Proceedings of the Workshop on Self-Organizing Maps (WSOM '97)*, Espoo, Finland, June 4-6, pp. 316-321.

[Mer99] Dieter Merkl. 1999. "Document Classification with Self-Organizing Maps". In *Kohonen Maps*, E. Oja and S. Kaski (Eds.), Elsevier, Amsterdam, The Netherlands.

[MK00] George Mikros and George Karayannis. 2000. "Modern Greek Corpus Taxonomy". In *Proceedings of the 2nd Conference on Language Resources and Evaluation*, Athens, Greece, 31 May-2 June, Vol. 1, pp. 129-134.

[Mou96] Isabelle Moulinier. 1996. "A Framework for Comparing Text Categorization Approaches". AAAI Spring Symposium on Machine Learning and Information Access, Stanford University, March.

[Μπα72] Γεώργιος Μπαμπινιώτης. 1972. "Το ρήμα της Ελληνικής : Δομικαί Εξελίξεις και Συστηματοποίησις του ρήματος της Ελληνικής (Αρχαίας και Νέας)". Βιβλιοθήκη Σοφίας Σαριπόλου 20, Πανεπιστήμιον Αθηνών, Αθήναι.

[MP69] Marvin Minsky and Seymour Papert. 1969. "Perceptrons: An Introduction to Computational Geometry". In *Cambridge, MA:MIT Press, Introduction*, pp.1-20 and p. 73.

[MT00] Στέλλα Μαρκαντωνάτου και Γεώργιος Ταμπουρατζής. 2000. "Μερικές ποσοτικές παρατηρήσεις για την χρήση της γραμματικής άρνησης στην Νέα Ελληνική". In *Proceedings of the 21st Annual Meeting of the Department of Linguistics*, Τμήμα Φιλοσοφίας του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, Μάιος. (υπό έκδοση).

[PPG+00] Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, Stelios Piperidis. 2000. "A Unified PoS Tagging Architecture and its Application to Greek". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May-2 June, Vol.3, pp.1455-1462.

[PW93] Joseph Pentheroudakis and Lucy Wanderwende. 1993. "Automatically Identifying Morphological Relations in Machine-Readable Dictionaries". *Technical Report MSR-TR-93-06*, Microsoft Research Advanced Technology Division, Microsoft Corporation.

[Ral86] Αγγελική Ράλλη. 1986. "Κλίση και Παραγωγή". *Πρακτικά της 7ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης*, 12-14 Μαΐου 1986, pp. 29-48.

[Ril95] Ellen Riloff. 1995. "Little Words can make a Big Difference for Text Classification". In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 130-136.

[Ril96a] Ellen Riloff. 1996. "Automatically Generating Extraction Patterns from Untagged Text". In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044-1049.

[Ril96b] Ellen Riloff. 1996. "Using Learned Extraction Patterns for Text Classification". In S. Wermter, E. Riloff and G. Scheler (editors) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Springer-Verlag, Berlin, pp. 275-289.

[SHS96] Mehran Sahami, Marti Hearst and Eric Saund. 1996. "Applying the Multiple Cause Mixture Model to Text Categorization". In Lorenza Saitta, ed., *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 435-443, Bari, Italy, July 3-6. Morgan Kaufmann, ISBN 1-55860-419-7.

[TK99] Γεώργιος Ταμπουρατζής και Γεώργιος Καραγιάννης. 1999. "Αυτόματη δημιουργία μορφολογικών λεξικών με ορολογικό πλούτο βάσει κειμένων εντάσεως όρων". In *Proceedings of the 2nd Conference of Hellenic Language and Terminology*, Αθήνα , 21-23 Οκτωβρίου 1999, pp. 149-155.

[TMH+00a] George Tambouratzis, Stella Markantonatou, Nikolaos Hairetakis and George Carayannis. 2000. "Automatic Style Categorisation of Corpora in the Greek Language". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May - 2 June, Vol. 1, pp. 135-140.

[TMH+00b] George Tambouratzis, Stella Markantonatou, Nikolaos Hairetakis, Marina Vassiliou, Dimitrios Tambouratzis and George Carayannis. 2000. "Discriminating the registers and styles in the Modern Greek Language". *In Proceedings of the Workshop on Comparing Corpora*, held in conjunction with the 38th Annual meeting of the Association for Computational Linguistics, 7 October, Hong Kong, pp. 35-42.

[VHA+00] Juha Vesanto, Johan Himberg, Esa Alhoniemi and Juha Parhankangas. 2000. "SOM Toolbox for Matlab 5". *Report A57.* SOM Toolbox Team, Helsinki University of Technology, Finland (*http://www.cis.hut.fi/projects/somtoolbox*).

[Yan96] Yiming Yang. 1996. "Sampling Strategies and Learning Efficiency in Text Categorization". In *AAAI Spring Symposium on Machine Learning in Information Access,* pp. 88-95.

[Yan99] Yiming Yang. 1999. "An Evaluation of Statistical Approaches to Text Categorization". In *Journal of Information Retrieval*, 1999, Vol 1, No. 1/2, pp. 67-88.