Κεφάλαιο 1

Εισαγωγή

> We must find out what words are and how they function. They become images when written down, but images of words repeated in the mind and not of the image of the thing itself. [W.S. Burroughs]

1.1 Τάσεις στη Σύγχρονη Υπολογιστική Γλωσσολογία

Το αντικείμενο της *υπολογιστικής γλωσσολογίας* (computational linguistics) είναι η αυτοματοποίηση της ανάλυσης κειμένων που βρίσκονται σε ηλεκτρονική μορφή. Συχνά καλείται και *επεξεργασία φυσικής γλώσσας* (natural language processing), σε αντίθεση με τις τεχνητές γλώσσες των υπολογιστών. Η ιστορία της καλύπτει το δεύτερο μισό του 20ού αιώνα και τυπικές εφαρμογές της είναι η μηχανική μετάφραση, η αυτόματη απάντηση ερωτήσεων, η αυτόματη σύνθεση κειμένου, κ.ά. Ωστόσο, παρά το πλήθος των ερευνητών που έχουν ασχοληθεί με την υπολογιστική γλωσσολογία, τα αποτελέσματα που έχουν επιτευχθεί έως σήμερα δεν είναι ανάλογα άλλων, συγγενών επιστημονικών περιοχών, όπως της επεξεργασίας ομιλίας.

Μέχρι πρόσφατα, η έρευνα στην υπολογιστική γλωσσολογία έδινε έμφαση σε συστήματα που προσπαθούσαν να εφαρμόσουν τις παραδοσιακές θεωρίες της τεχνητής νοημοσύνης και να επιτύχουν πλήρη κατανόηση του κειμένου [3, 4, 31, 99].

Όμως, ένα κείμενο περιλαμβάνει πλήθος αφηρημένων εννοιών και αμέτρητους συνδυασμούς αφηρημένων σχέσεων μεταξύ των εννοιών που είναι πολύ δύσκολο να αναπαρασταθούν με χρήση των κλασσικών τεχνικών αναπαράστασης γνώσης. Τα προγράμματα που γράφτηκαν εκείνη την περίοδο μπορούσαν να επεξεργαστούν πολύ λίγες προτάσεις αν και το γεγονός αυτό συχνά δεν αναφερόταν στις σχετικές δημοσιεύσεις (διατριβές, βιβλία κτλ.). Η περίπτωση που αναφέρεται στην συνέχεια είναι χαρακτηριστική [41]: Ένας μελετητής της ηλεκτρονικής λεξικογραφίας (Bran Boguraev) κάλεσε αρκετούς ερευνητές εκείνης της περιόδου να του δώσουν στοιχεία σχετικά με το πραγματικό μέγεθος των λεξικών που χρησιμοποιούσαν στα προγράμματά τους. Από τις απαντήσεις που έλαβε διαπίστωσε ότι τα λεξικά περιείχαν κατά μέσο όρο μόλις 36 λέξεις!

Την τελευταία δεκαετία παρατηρείται μία τάση των ερευνητών προς την ανάπτυξη προγραμμάτων ευρείας κλίμακας. Προς αυτήν την κατεύθυνση, άρχισαν να εφαρμόζονται εμπειρικές μέθοδοι [14, 20] και επανήλθαν στην επιφάνεια οι στατιστικές τεχνικές [2, 21]. Τέτοιες προσεγγίσεις βασίζονται σε μεγάλα σώματα εκπαίδευσης και όχι σε κάποια γλωσσολογική θεωρία χωρίς να υποστηρίζουν κατανόηση του κειμένου σε βάθος. Ωστόσο, έχει αποδειχτεί ότι αποτελούν πολύ αξιόπιστη λύση. Επίσης, μεγάλη ανάπτυξη γνωρίζει και η εφαρμογή θεωριών της μηχανικής εκμάθησης (machine learning) στην επεξεργασία φυσικής γλώσσας [13, 58]. Σ' αυτήν την περίπτωση, η γλωσσολογική γνώση που απαιτείται για την ανάλυση του κειμένου εξάγεται αυτόματα με βάση ένα σώμα κειμένων εκπαίδευσης. Όλες αυτές οι τεχνικές συχνά καλούνται μέθοδοι βάσει σώματος κειμένου (corpus-based methods). Ενδεικτικό της τάσης που επικρατεί στην σύγχρονη υπολογιστική γλωσσολογία είναι το γεγονός ότι στα πρακτικά του ετήσιου συνεδρίου της *Εταιρείας Υπολογιστικής Γλωσσολογίας* (annual meeting of the Association for Computational Linguistics) του 1991 [104] συμπεριλήφθηκαν μόλις 3 εργασίες βασισμένες σε σώμα κειμένου, ενώ στα αντίστοιχα πρακτικά του 1996 [105] τουλάχιστον οι μισές εργασίες αφορούσαν μελέτες βάσει σώματος κειμένου!

Ένας άλλος βασικός παράγοντας, που επηρέασε την πορεία της έρευνας στην υπολογιστική γλωσσολογία τα τελευταία χρόνια, ήταν η ραγδαία ανάπτυξη βάσεων δεδομένων με κείμενα σε ηλεκτρονική μορφή και κυρίως η ανάπτυξη του *Διαδικτύου* και των υπηρεσιών του (World-Wide Web). Ο όγκος των κειμένων έγινε πλέον

δυσθεώρητος και πολλαπλασιάστηκαν οι ανάγκες για αυτόματη εξαγωγή πληροφορίας από κείμενα, αυτόματη ταξινόμηση κειμένων, αυτόματη εξαγωγή περίληψης από μεγάλα κείμενα κτλ. Έτσι, οι ερευνητές της επεξεργασίας φυσικής γλώσσας έστρεψαν την προσοχή τους σε πρακτικές εφαρμογές προσπαθώντας να επιτύχουν συστήματα που θα συνδύαζαν ελάχιστο χρόνο απόκρισης και ικανοποιητική (όχι απαραίτητα τέλεια) ακρίβεια [106]. Ακόμη, απέκτησε μεγάλη σημασία η σωστή αξιολόγηση των συστημάτων που προέκυψαν. Έγινε, λοιπόν, προσπάθεια ώστε να καθοριστούν αντικειμενικές και αξιόπιστες μέθοδοι αξιολόγησης [56].

Οι βάσεις κειμένων και το Διαδίκτυο περιέχουν κείμενα σε ακατέργαστη μορφή που πιθανόν να περιέχουν και λάθη. Αυτό το γεγονός δημιούργησε την ανάγκη ανάπτυξης εύρωστων συστημάτων που να δίνουν έμφαση σε εργασίες χαμηλού επιπέδου, όπως ο χωρισμός του κειμένου σε λέξεις, η αναγνώριση κύριων ονομάτων, η αναγνώριση ορίων περιόδων, κ.ά. [12, 71, 68]. Μέχρι πρόσφατα, η έρευνα δεν είχε δώσει μεγάλη προσοχή σε τέτοιου είδους εργασίες (η λύση τους θεωρούταν δεδομένη) με αποτέλεσμα να αντιμετωπίζονται ανεπαρκώς και να δημιουργείται πρόβλημα στα υψηλότερα επίπεδα ανάλυσης (π.χ. συντακτικό, σημασιολογικό).

Τέλος, τα τελευταία χρόνια άρχισαν δειλά-δειλά να κάνουν την εμφάνισή τους μελέτες όπου το κείμενο δεν αντιμετωπίζεται πλέον με βάση μόνο το σημασιολογικό του περιεχόμενο, αλλά λαμβάνεται υπ' όψιν και η δομή του ή το ύφος του [47, 50, 90]. Παρά το πολύ πρώιμο στάδιο της αυτόματης υφολογικής ανάλυσης, θεωρείται πλέον δεδομένο ότι το ύφος είναι ένας πολύ σημαντικός παράγοντας και πρέπει να λαμβάνεται υπ' όψιν από κάθε σύστημα επεξεργασίας κειμένου που στοχεύει στην επίτευξη των καλύτερων δυνατών αποτελεσμάτων.

1.2 Γλώσσα και Ύφος

Η έννοια του ύφους απασχολεί το ανθρώπινο πνεύμα για περισσότερο από δύο χιλιετίες. Σχεδόν όλοι οι μελετητές του ύφους, οι οποίοι προέρχονται από διάφορες επιστήμες (γλωσσολόγοι, λογοτέχνες, φιλόσοφοι κτλ.), έχουν μπει στον πειρασμό να

δώσουν το δικό τους ορισμό ο οποίος κατά κανόνα είναι είτε πολύ γενικός είτε ανεπαρκής. Τα παρακάτω παραδείγματα είναι χαρακτηριστικά[1]:

- Ύφος λέγεται η γλώσσα που χρησιμοποιείται ως τέχνη. [Spitzer]
- Το ύφος είναι για το συγγραφέα όπως το χρώμα για το ζωγράφο, όχι θέμα τεχνικής, αλλά τρόπος να βλέπει τα πράγματα. [Proust]
- Το ύφος είναι η φυσιογνωμία του πνεύματος. [Schopenhauer]
- Το ύφος είναι το ένδυμα της σκέψης. [Chesterfield]
- Η γλώσσα είναι το ένδυμα της σκέψης και το ύφος το ιδιαίτερο κόψιμο και η μόδα του ενδύματος. [Hough]
- Ύφος είναι το σύνολο των ιδιαιτεροτήτων που χαρακτηρίζουν το άτομο στον γραπτό και προφορικό λόγο. [Bruneau]

Γενικά, το ύφος σχετίζεται με τη μορφή ενός κειμένου παρά με το θεματικό του περιεχόμενο. Πολλοί φιλόσοφοι και γλωσσολόγοι δεν δέχονται ότι αυτός ο διαχωρισμός είναι δυνατός και φτάνουν στο σημείο να αμφισβητούν ακόμα και την ύπαρξη του ύφους [26, 40]. Όμως, όσο ανεπαρκείς είναι οι ορισμοί του ύφους, άλλο τόσο ανεπαρκείς είναι οι αφοριστικές διατυπώσεις εναντίον του.

Επίσης, πολλοί ερευνητές υποστηρίζουν ότι μόνο τα λογοτεχνικά έργα έχουν ύφος ενώ, για παράδειγμα, τα επιστημονικά άρθρα χαρακτηρίζονται από «αντι-ύφος» (anti-style) [28]. Όμως, είναι αμφίβολο αν μπορούμε να ισχυριστούμε ότι ένα κείμενο δεν έχει καθόλου ύφος. Ακόμα και η απουσία κάθε γνωστού τύπου ύφους από ένα κείμενο είναι αυτή καθ' αυτή ένας ιδιαίτερος τύπος ύφους. Κάθε γλωσσολογική εξωτερίκευση, ακόμα και μία μαγειρική συνταγή, έχει το ύφος της [100].

Καθώς δεν υπάρχει ένας κοινά αποδεκτός ορισμός του ύφους επικρατεί πλήρης σύγχυση σχετικά με τον προσδιορισμό των γνωρισμάτων του. Μερικά μόνο από τα στερεότυπα επίθετα που χρησιμοποιούνται είναι τα ακόλουθα: γλαφυρό, λιτό, μικτό, πλούσιο, ποικίλο, χαλαρό κτλ. Σε μία έρευνα που έκανε η Tufte [93] ζήτησε από 44 καθηγητές Μέσης Εκπαίδευσης να χαρακτηρίσουν με πέντε ως έξι λέξεις το ύφος ενός σύντομου κειμένου. Το αποτέλεσμα ήταν να χρησιμοποιηθούν συνολικά 222 διαφορετικά επίθετα από τα οποία μόνο ένα είχε χρησιμοποιηθεί από 12 εκπαιδευτικούς!

[1] Μετάφραση Χ. Χαραλαμπάκη [100].

Στην συνέχεια περιγράφονται οι πιο σημαντικές γλωσσολογικές προσεγγίσεις για τον προσδιορισμό του ύφους.

1.2.1 Το ύφος ως απόκλιση

Μία από τις πιο δημοφιλείς υφολογικές προσεγγίσεις είναι ο ορισμός του ύφους ως απόκλιση από μία νόρμα (deviation form a norm), όπου νόρμα θεωρείται μία συγκεκριμένη χρήση της γλώσσας [34]. Οι αποκλίσεις μπορεί να αναφέρονται σε επίπεδο λεξιλογικό, συντακτικό, σημασιολογικό, φωνολογικό, γραφολογικό κτλ. Οι Leech και Short [59] υποστηρίζουν ότι η στατιστική απόκλιση κάποιων γλωσσολογικών γνωρισμάτων οδηγεί στην υποκειμενική αναγνώριση υφολογικών στοιχείων και στην συνέχεια στην απόκλιση με αισθητικά κριτήρια. Η στατιστική απόκλιση από μόνη της δεν οδηγεί σε υφολογικά συμπεράσματα.

Τα βασικά προβλήματα της θεώρησης του ύφους ως απόκλισης από μία νόρμα συνοψίζονται στις παρακάτω ερωτήσεις:

- Με ποια κριτήρια ορίζονται οι αποκλίσεις;
- Πότε θεωρείται μία απόκλιση σημαντική;
- Πώς γίνεται η ιεράρχηση των αποκλίσεων;
- Πώς καθορίζεται η νόρμα μιας γλώσσας;
- Τι γίνεται στην περίπτωση όπου ένα κείμενο δεν παρουσιάζει καμία απόκλιση από τη νόρμα;

Ασφαλώς, ο καθορισμός της νόρμας αποτελεί το σπουδαιότερο πρόβλημα. Αν θεωρηθεί ως νόρμα ο μέσος όρος της στατιστικής ανάλυσης όλων των γλωσσικών γνωρισμάτων του συνόλου των υπαρχόντων κειμένων τότε προκύπτουν και πρακτικά και θεωρητικά προβλήματα. Πιο συγκεκριμένα, είναι αδύνατον να συγκεντρωθούν όλα τα κείμενα μιας γλώσσας και να αναλυθούν. Επίσης, όπως τονίστηκε πιο πριν, η στατιστική απόκλιση κάποιων γλωσσικών γνωρισμάτων από μόνη της δεν μπορεί να οδηγήσει σε ασφαλή συμπεράσματα.

1.2.2 Το ύφος ως επιλογή

Η θεώρηση του ύφους ως επιλογής είναι μία από τις πιο ενδιαφέρουσες προσεγγίσεις της έννοιας του ύφους. Η επιλεκτική υφολογία βασίζεται στο γεγονός ότι η γλώσσα προσφέρει στον ομιλητή/συγγραφέα μία σειρά δυνατοτήτων για να εκφράσει αυτό που θέλει. Έτσι, το ύφος ορίζεται ως το αποτέλεσμα επιλογής ανάμεσα στις προαιρετικές γλωσσικές δυνατότητες που βρίσκονται σε σχέση παράφρασης, δηλ. είναι ισοδύναμες σημασιολογικά. Το μεγαλύτερο πρόβλημα αυτής της προσέγγισης είναι ότι το πλήθος των επιλογών μεταξύ των οποίων καλείται να διαλέξει ο ομιλητής/συγγραφέας είναι τεράστιο (για μία μόνο πρόταση μπορεί να είναι αρκετές χιλιάδες). Επομένως, από πρακτική άποψη δεν είναι δυνατόν να υλοποιηθεί ένα σύστημα που να βασίζεται στην επιλεκτική υφολογία.

Η σχέση ύφους-επιλογής αποκτά μεγαλύτερη σημασία εφόσον συνδυαστεί με τη γενετική-μετασχηματιστική γραμματική [22] η οποία κάνει διάκριση ανάμεσα σε υποχρεωτικούς και προαιρετικούς μετασχηματιστικούς κανόνες. Το ύφος, λοιπόν, μπορεί να θεωρηθεί ως το αποτέλεσμα επιλογής των προαιρετικών μετασχηματιστικών κανόνων [70]. Ωστόσο, τα αποτελέσματα των μελετών που βασίστηκαν σε αυτήν τη θεωρία δεν ήταν καθόλου ικανοποιητικά [78]. Κάποιοι ερευνητές θεώρησαν ότι αυτό οφείλεται στο ότι η γενετική-μετασχηματιστική γραμματική δεν παρέχει το κατάλληλο πλαίσιο για την ανάπτυξη υφολογικών θεωριών [38]. Επίσης, επικρίθηκε η υπόθεση ότι προτάσεις σημασιολογικά περίπου ισοδύναμες μπορούν να χρησιμοποιηθούν σε κάθε περίπτωση.

1.2.3 Στατιστική υφολογία

Η πιο πολλά υποσχόμενη προσέγγιση του ύφους είναι η θεώρησή του ως ένα σύνολο μετρήσιμων παραμέτρων. Έτσι, η γλώσσα ενός κειμένου ποσοτικοποιείται και το ύφος προκύπτει από την ανάλυση του διανύσματος των παραμέτρων βάσει στατιστικών τεχνικών. Η στατιστική υφολογία αντιμετωπίζει το ύφος από πρακτική άποψη προσφέροντας τη δυνατότητα ανάπτυξης υπολογιστικών συστημάτων ανάλυσης ύφους. Γι' αυτό το λόγο συχνά καλείται και *υπολογιστική υφολογία* (computational stylistics).

Ως παράμετροι χρησιμοποιούνται γλωσσολογικά γνωρίσματα κυρίως λεξιλογικού και συντακτικού επιπέδου τα οποία ο συγγραφέας δεν χρησιμοποιεί συνειδητά και επομένως μένουν σχετικά αμετάβλητα σε υφολογικά παρόμοια κείμενα (π.χ. μήκος περιόδων, συχνότητα εμφάνισης μερών του λόγου κ.ά.). Η εύρεση των πιο κατάλληλων παραμέτρων είναι το αντικείμενο της *υφομετρίας* (stylometry) (βλ. § 4.1).

Πολλοί ερευνητές είναι επιφυλακτικοί σχετικά με τη χρήση στατιστικών μεθόδων για την ανάλυση του ύφους [96]. Τα πιο σημαντικά προβλήματα είναι τα ακόλουθα:

- Η επίδραση του γλωσσικού περιβάλλοντος δεν λαμβάνεται υπ' όψιν. Για παράδειγμα, η απλή μέτρηση της συχνότητας μιας λέξης παραβλέπει το γεγονός ότι η λέξη αυτή μπορεί να χρησιμοποιείται σε ιδιωματικές εκφράσεις ή μεταφορές. Κατά συνέπεια, παραβλέπονται λεπτές υφολογικές αποχρώσεις.
- Η ερμηνεία των αποτελεσμάτων της στατιστικής ανάλυσης είναι κατά κανόνα δύσκολη. Δεν υπάρχει κάποια γλωσσολογική θεωρία πίσω από την στατιστική επεξεργασία. Έτσι, η κατανόηση της έννοιας του ύφους είναι επίπονη και υποκειμενική εργασία [72].

Ένα άλλο πρόβλημα που έχει προκύψει από την έρευνα έως σήμερα έχει να κάνει με την έλλειψη βασικών γνώσεων της στατιστικής θεωρίας από πολλούς μελετητές (κυρίως γλωσσολόγους και φιλολόγους) με αποτέλεσμα τη χαμηλή ποιότητα των μελετών [86]. Επίσης, πρέπει να σημειωθεί ότι μέχρι πρόσφατα οι δυνατότητες των υπολογιστικών συστημάτων δεν ήταν τόσο μεγάλες ώστε να επιτρέπουν αξιόλογες στατιστικές αναλύσεις μεγάλων σωμάτων κειμένων. Με την πρόοδο της υπολογιστικής ισχύος τα τελευταία χρόνια αυτό το πρόβλημα δεν υφίσταται πλέον. Επομένως, μπορούμε να πούμε ότι η στατιστική προσέγγιση αποτελεί την πιο ελπιδοφόρο προοπτική της υφολογίας.

1.3 Το Ύφος στην Επεξεργασία Φυσικής Γλώσσας

Μέχρι πρόσφατα, το ύφος συνδεόταν με την επεξεργασία φυσικής γλώσσας μόνο μέσω των λεγόμενων *ελεγκτών ύφους* (style checkers). Πρόκειται για βοηθητικά προγράμματα επεξεργαστών κειμένου που έχουν ως στόχο τη σύνταξη κειμένων με

όσο το δυνατόν μεγαλύτερη σαφήνεια, απλότητα και ακρίβεια. Τα λάθη που εντοπίζουν είναι πολύ απλά, όπως μη-αρμονικά σημεία στίξης, υπερβολικά μεγάλο μήκος περιόδων, χρήση παθητικής σύνταξης κ.ά. Έχει αναπτυχθεί πλήθος τέτοιων προγραμμάτων [74]. Ενδεικτικά αναφέρουμε τα *PC-Style, RightWriter* και *CRITIQUE*.

Η πρώτη συστηματική προσπάθεια να ενσωματωθεί το ύφος σε ένα σύστημα επεξεργασίας φυσικής γλώσσας έγινε από τον Hovy στο σύστημα *PAULINE* [47], ένα συνθέτη φυσικής γλώσσας που λαμβάνει υπ' όψιν του πραγματολογικούς περιορισμούς. Σε μία υποθετική συνομιλία μεταξύ ενός συστήματος και ενός ανθρώπου, οι περίπλοκες διαπροσωπικές σχέσεις μεταξύ των ομιλητών επηρεάζουν άμεσα τόσο το περιεχόμενο της συζήτησης όσο και το ύφος του παραγόμενου κειμένου. Για παράδειγμα, αλλιώς θα περιγράψει κάποιος ένα γεγονός σε μία δημόσια ομιλία του, αλλιώς όταν μιλάει με συνεργάτες του, αλλιώς όταν το περιγράφει σε ένα φιλικό του πρόσωπο κτλ. Για να κατορθώσει να αναπαραστήσει τέτοιες καταστάσεις ο Hovy βασίστηκε στον καθορισμό ενός συνόλου *ρητορικών στόχων του ύφους* (π.χ. επισημότητα, δυναμικότητα, βιασύνη, κ.ά.), οι οποίοι επηρεάζουν πιο ευρείς πραγματολογικούς στόχους. Η επίτευξη ενός ρητορικού στόχου γίνεται μέσω των λεξιλογικών και συντακτικών χαρακτηριστικών του παραγόμενου κειμένου με βάση ένα σύνολο ευρετικών κανόνων. Η συμπεριφορά του *PAULINE* είναι πολύ καλή έως εντυπωσιακή σε ορισμένες περιπτώσεις.

Μία παρόμοια προσέγγιση υιοθετήθηκε και από τους DiMarco και Hirst [30] στην προσπάθεια βελτίωσης της εξόδου ενός συστήματος μηχανικής μετάφρασης. Κατά την μετάφραση ενός κειμένου από μία γλώσσα σε μία άλλη πρέπει να διατηρηθούν ακέραιοι οι υφολογικοί στόχοι του αρχικού κειμένου έστω και αν αυτό οδηγήσει σε διαφορετικό συντακτικό σχήμα στο κείμενο εξόδου. Για την αναπαράσταση αυτής της γλωσσολογικής γνώσης οι DiMarco και Hirst εισήγαγαν την ιδέα της *γραμματικής του ύφους* (style grammar) η οποία συσχετίζει τις συντακτικές δομές μιας γλώσσας με ένα σύνολο γλωσσικά-ανεξάρτητων, υφολογικών στόχων. Έτσι, κατά την μετάφραση, οι στόχοι αυτοί εντοπίζονται στο κείμενο εισόδου και χρησιμοποιούνται για την σύνθεση του κειμένου εξόδου. Εύκολα γίνεται αντιληπτό πως η θεώρηση του ύφους ως αποτέλεσμα συγκεκριμένων στόχων βρίσκει εφαρμογή πιο εύκολα σε συστήματα σύνθεσης παρά ανάλυσης φυσικής γλώσσας.

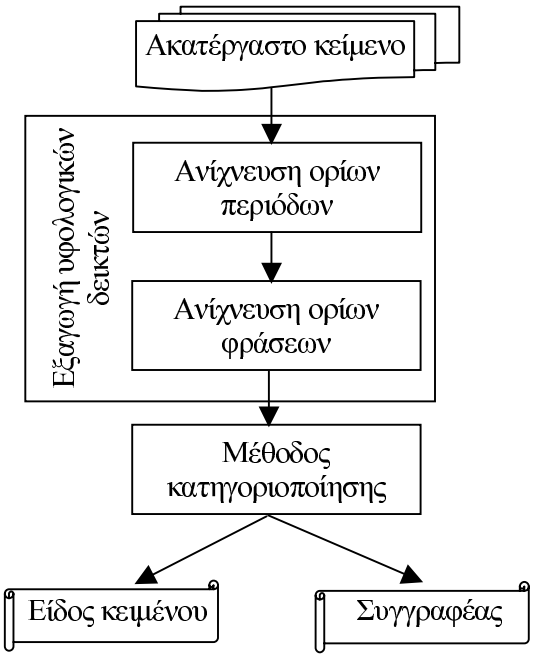
Οι πιο σημαντικές προσπάθειες εκμετάλλευσης του ύφους, όσον αφορά την ανάλυση κειμένων, έχουν γίνει στην αναγνώριση του είδους κειμένου, μία εφαρμογή που αποκτά όλο και μεγαλύτερη σημασία τα τελευταία χρόνια. Το είδος ενός κειμένου δεν καθορίζεται από το θεματικό του περιεχόμενο αλλά από το λειτουργικό του ρόλο, το κοινό στο οποίο απευθύνεται, τον τρόπο κατασκευής του κ.ά. Χαρακτηριστικά παραδείγματα ειδών κειμένων είναι τα επιστημονικά άρθρα, οι συνεντεύξεις, τα δημόσια έγγραφα, κτλ. Ουσιαστικά, ένα σύστημα αναγνώρισης είδους κειμένου χρησιμοποιεί το ύφος ως μέσο για να ταξινομήσει ένα κείμενο σε μία υφολογική κατηγορία. Έτσι, δίνεται η δυνατότητα κατηγοριοποίησης κειμένων όχι μόνο βάσει του περιεχομένου τους αλλά και του ύφους τους. Μία εκτενής αναφορά στα συστήματα αναγνώρισης κειμένου που έχουν προταθεί έως σήμερα δίνεται στο τμήμα 5.1.

1.4 Προδιαγραφές - Καινοτομίες της Διατριβής

Το αντικείμενο αυτής της διατριβής είναι η στατιστική υφολογική επεξεργασία κειμένων της Νέας Ελληνικής γλώσσας με στόχο την αυτόματη ταξινόμησή τους είτε ως προς το είδος είτε ως προς το συγγραφέα τους. Οι προδιαγραφές που καθόρισαν το σχεδιασμό του προτεινόμενου συστήματος συνοψίζονται παρακάτω:

- **Ελάχιστος χρόνος απόκρισης:** Το σύστημα θα πρέπει να αποκρίνεται σε σχεδόν πραγματικό χρόνο. Αυτή η απαίτηση είναι αυτονόητη για μία εφαρμογή που στοχεύει στην αυτόματη ταξινόμηση μεγάλων όγκων κειμένων.
- **Ελάχιστοι πόροι:** Η όλη διαδικασία ανάλυσης του κειμένου δεν πρέπει να βασίζεται σε ογκώδεις και εξειδικευμένους πόρους (π.χ. λεξικά εκατοντάδων χιλιάδων λημμάτων με εξειδικευμένη πληροφορία). Η αλλαγή και η ενημέρωση των πόρων αυτών πρέπει να είναι πολύ εύκολη.
- **Δυνατότητα επεξεργασίας κειμένου χωρίς περιορισμούς:** Κανένας περιορισμός δεν τίθεται στο προς ανάλυση κείμενο εκτός του ότι πρέπει να είναι ήδη σε ηλεκτρονική μορφή. Το κείμενο πρέπει να είναι διαθέσιμο όπως ακριβώς εμφανίζεται στην πηγή του.

- **Αυτοματοποίηση:** Η διαδικασία επεξεργασίας-ταξινόμησης κειμένου πρέπει να είναι εντελώς αυτόματη (καμία χειρονακτική προ-επεξεργασία ή επιλογή συγκεκριμένων τμημάτων του κειμένου).
- **Δυνατότητα επέκτασης:** Το σύνολο των αναγνωρίσιμων ειδών κειμένων/συγγραφέων πρέπει να μπορεί να επεκταθεί-τροποποιηθεί εύκολα.
- **Γενικότητα:** Η διαδικασία ταξινόμησης του κειμένου σε ένα είδος κειμένου/συγγραφέα δεν πρέπει να βασίζεται σε κάποιο ιδιαίτερο χαρακτηριστικό ενός συγκεκριμένου είδους κειμένου/συγγραφέα αλλά να είναι όσο το δυνατόν πιο γενική.
- **Ανεξαρτησία γλώσσας:** Ως επέκταση της προηγούμενης απαίτησης, η διαδικασία ταξινόμησης πρέπει να είναι όσο το δυνατόν πιο ανεξάρτητη γλώσσας.



Σχήμα 1.1. Η προτεινόμενη λύση.

Η λύση που προτείνουμε για την αντιμετώπιση αυτού του προβλήματος φαίνεται στο σχήμα 1.1. Η όλη διαδικασία μπορεί να διακριθεί σε δύο φάσεις. Η πρώτη φάση περιλαμβάνει την εξαγωγή των υφολογικών παραμέτρων από το κείμενο με βάση την αυτόματη ανίχνευση των ορίων περιόδων και φράσεων που το αποτελούν. Στην συνέχεια, το διάνυσμα των υφολογικών παραμέτρων ταξινομείται σε μία υφολογική κατηγορία (είδος κειμένου ή συγγραφέας) με βάση μία μέθοδο κατηγοριοποίησης.

Τα πρωτότυπα σημεία αυτής της διατριβής είναι τα ακόλουθα:

- Ανιχνευτής ορίων περιόδων για κείμενα της Νέας Ελληνικής γλώσσας. Πρόκειται για ένα εργαλείο που στοχεύει στον χωρισμό οποιουδήποτε κειμένου σε περιόδους με μεγάλη ακρίβεια και πολύ γρήγορα. Το πιο σημαντικό χαρακτηριστικό του είναι ότι, σε αντίθεση με τις σύγχρονες προσεγγίσεις, βασίζεται σε ελάχιστους πόρους, όπως μήκος λέξεων, τύποι χαρακτήρων κ.ά. και όχι σε λίστες συντομογραφιών και λεξικά με περίπλοκη πληροφορία.
- Αυτόματη εξαγωγή κανόνων. Ο ανιχνευτής ορίων περιόδων μπορεί να εκπαιδευτεί και να προσαρμοστεί σε ένα συγκεκριμένο τύπο κειμένου αφού η εξαγωγή της απαιτούμενης γνώσης γίνεται αυτόματα. Για το σκοπό αυτό αναπτύχθηκε μία νέα μέθοδος μηχανικής εκμάθησης.
- Ανιχνευτής ορίων φράσεων για κείμενα της Νέας Ελληνικής γλώσσας. Πρόκειται για ένα εργαλείο που χωρίζει την κάθε περίοδο σε φράσεις (π.χ. ονοματικές, ρηματικές, κτλ.) και είναι ικανό να χειριστεί οποιοδήποτε τμήμα κειμένου με ελάχιστο χρονικό κόστος. Σε αντίθεση με τις σύγχρονες προσεγγίσεις, η διαδικασία εκτίμησης της πιο πιθανής μορφολογικής ανάλυσης της κάθε λέξης, βάσει της κατάληξής της, αντικαθιστά πλήρως τα ογκώδη λεξικά λημμάτων. Η ανάλυση του κειμένου βασίζεται σε πολλαπλά περάσματα ανάλυσης που βασίζονται σε εμπειρικούς κανόνες. Οι μόνοι πόροι που χρησιμοποιούνται είναι ένα λεξικό περίπου 450 λέξεων-κλειδιών και ένα λεξικό περίπου 300 κοινών καταλήξεων της Νέας Ελληνικής γλώσσας.
- Ένα σύνολο υφολογικών παραμέτρων που αποτελείται από τρία επίπεδα: δείγματος, φράσης και ανάλυσης. Τα δύο πρώτα έχουν να κάνουν να με την έξοδο των εργαλείων ανίχνευσης ορίων περιόδων και φράσεων αντίστοιχα. Το τελευταίο επίπεδο υφολογικών παραμέτρων είναι ένας εναλλακτικός τρόπος σύλληψης της υφολογικής πληροφορίας και βασίζεται στον τρόπο που έγινε η ανάλυση του κειμένου από τον ανιχνευτή ορίων περιόδων-φράσεων. Έτσι, σε αντίθεση με τις προηγούμενες προσεγγίσεις, το σύνολο των υφολογικών παραμέτρων δεν είναι προκαθορισμένο αλλά προσαρμόζεται στον τρόπο μέτρησης των παραμέτρων μέσω ενός υπολογιστικού εργαλείου. Ως αποτέλεσμα για την εξαγωγή υφολογικής πληροφορίας μπορούν να

χρησιμοποιηθούν ήδη υπάρχοντα εργαλεία επεξεργασίας φυσικής γλώσσας χωρίς επιπρόσθετο κόστος. Ένα άλλο σημείο που διαφοροποιεί το προτεινόμενο σύνολο από τις προηγούμενες προσεγγίσεις είναι το γεγονός ότι δεν περιλαμβάνει καμία υφολογική παράμετρο σχετική με λεξιλογική πληροφορία (π.χ. συχνότητες εμφάνισης συγκεκριμένων λέξεων). Έτσι, επιτυγχάνεται μεγαλύτερη γενικότητα.

- Κοινή αντιμετώπιση της αναγνώρισης είδους κειμένου και προσδιορισμού συγγραφέα. Το προτεινόμενο σύνολο υφολογικών δεικτών είναι ικανό να αναγνωρίσει οποιαδήποτε υφολογικά ομοιογενή κατηγορία.
- Ένα ολοκληρωμένο σύστημα αναγνώρισης είδους κειμένου. Το σύστημα αυτό είναι εκπαιδεύσιμο και μπορεί να εφαρμοστεί σε οποιοδήποτε πλήθος ειδών κειμένου. Η ακρίβειά του είναι πολύ υψηλότερη από αντίστοιχα συστήματα που έχουν προταθεί για την Αγγλική γλώσσα.
- Ένα ολοκληρωμένο σύστημα αναγνώρισης και επιβεβαίωσης συγγραφέα. Το σύστημα μπορεί να εκπαιδευτεί για μεγάλο αριθμό συγγραφέων και η ακρίβειά του είναι πολύ ικανοποιητική. Να σημειωθεί ότι στην διεθνή βιβλιογραφία δεν αναφέρεται άλλο ολοκληρωμένο σύστημα αυτόματου προσδιορισμού συγγραφέα.

1.5 Διάρθρωση της Διατριβής

Κατά τη συγγραφή αυτής της διατριβής έγινε προσπάθεια ώστε το κάθε κεφάλαιο να είναι όσο το δυνατόν πιο αυτόνομο. Πιο αναλυτικά, η διάρθρωση της διατριβής έχει ως εξής:

Το Κεφάλαιο 2 ασχολείται με την ανίχνευση ορίων περιόδων σε κείμενο χωρίς περιορισμούς της Νέας Ελληνικής γλώσσας. Αφού περιγραφούν συνοπτικά οι προηγούμενες προσεγγίσεις στην ανίχνευση ορίων περιόδων παρουσιάζεται η προτεινόμενη μέθοδος και τα αποτελέσματα που επιτυγχάνει. Επίσης, συγκρίνεται η απόδοση της προτεινόμενης μεθόδου εξαγωγής γνώσης με μία γνωστή θεωρία μηχανικής εκμάθησης.

Το Κεφάλαιο 3 περιγράφει τον ανιχνευτή ορίων φράσεων. Οι προηγούμενες προσεγγίσεις περιγράφονται συνοπτικά. Στη συνέχεια, παρουσιάζεται η μέθοδος που ακολουθήθηκε και ελέγχεται η απόδοσή της σε ένα σώμα κειμένων. Ακόμη, γίνεται σύγκριση του προτεινόμενου συστήματος με μία παρόμοια προσέγγιση που βασίζεται σε ένα ογκώδες λεξικό λημμάτων.

Το Κεφάλαιο 4 αφορά στην διαδικασία εξαγωγής υφολογικών δεικτών από το προς ανάλυση κείμενο. Η σχετική έρευνα και οι τάσεις στη σύγχρονη υφομετρία περιγράφονται επαρκώς, και στην συνέχεια παρουσιάζεται αναλυτικά η πρότασή μας.

Το Κεφάλαιο 5 περιγράφει τη διαδικασία αυτόματης αναγνώρισης είδους κειμένου. Αρχικά, περιγράφονται οι προηγούμενες προσεγγίσεις σε αυτό το πρόβλημα. Στην συνέχεια, παρουσιάζεται αναλυτικά η μέθοδος κατηγοριοποίησης, το σώμα κειμένων που χρησιμοποιήθηκε στα πειράματα και τα αποτελέσματα της εφαρμογής της προτεινόμενης μεθόδου. Επίσης περιγράφονται συγκριτικά πειράματα για άλλες δύο σύγχρονες υφομετρικές προσεγγίσεις και εξετάζεται η σχέση ακρίβειας και μεγέθους του σώματος εκπαίδευσης καθώς και η σημαντικότητα των υφολογικών δεικτών.

Το Κεφάλαιο 6 αφορά στον αυτόματο προσδιορισμό συγγραφέα. Αρχικά, περιγράφεται η τρέχουσα κατάσταση της σχετικής έρευνας. Στην συνέχεια παρουσιάζεται το σώμα κειμένων που χρησιμοποιήθηκε στα πειράματα και αναλυτικά αποτελέσματα τόσο για την αναγνώριση συγγραφέα όσο και για την επιβεβαίωση συγγραφέα. Γίνεται σύγκριση της προτεινόμενης μεθόδου με υφομετρικές προσεγγίσεις που βασίζονται αποκλειστικά σε λεξιλογική πληροφορία και εξετάζεται ο συνδυασμός της μεθόδου μας με μία από αυτές τις προσεγγίσεις. Ακόμη, περιγράφονται πειράματα σχετικά με το μέγεθος του σώματος εκπαίδευσης και τη σημαντικότητα των υφολογικών δεικτών.

Τέλος, το Κεφάλαιο 7 περιλαμβάνει τα συμπεράσματα που αποκομίστηκαν από αυτήν τη διατριβή. Επίσης, προτείνονται πιθανές εφαρμογές και κατευθύνσεις για μελλοντική έρευνα.