Κεφάλαιο 4

Εξαγωγή Υφολογικών Δεικτών

4.1 Τάσεις στην Σύγχρονη Υφομετρία

Η έρευνα στην υπολογιστική υφολογία είναι αρκετά περιορισμένη. Δύο είναι οι βασικοί παράγοντες που ευθύνονται γι' αυτό:

- Η έλλειψη ενός κοινά αποδεκτού τυπικού ορισμού του ύφους ενός κειμένου.
- Η αποτυχία των σύγχρονων συστημάτων επεξεργασίας φυσικής γλώσσας να ενσωματώσουν υφολογικές θεωρίες, λόγω της συνήθως πολύπλοκης πληροφορίας στην οποία βασίζονται (σημασιολογικού ή/και πραγματολογικού επιπέδου).

Σε αντίθεση με την παραδοσιακή υφολογία, η χρησιμοποίηση στατιστικών μεθόδων στην ανάλυση του ύφους έχει δώσει μέχρι τώρα τα πιο αξιόλογα αποτελέσματα [11] και έχει οδηγήσει στην ανάπτυξη της *υφομετρίας* (βλ. § 1.2.3). Στόχος της υφομετρίας είναι η αναπαράσταση του ύφους ενός κειμένου βάσει ενός συνόλου μετρήσιμων παραμέτρων που συχνά καλούνται *δείκτες ύφους* (style markers). Με άλλα λόγια, η υφομετρία προσπαθεί να ποσοτικοποιήσει, μέσω ενός διανύσματος παραμέτρων, τη

γλώσσα που χρησιμοποιείται σε ένα κείμενο. Από τη στιγμή που αυτό γίνεται εφικτό μπορούν να ανιχνευτούν διάφορα χαρακτηριστικά του κειμένου. Τυπικοί στόχοι της έρευνας στην υπολογιστική υφολογία είναι οι εξής:

- **Αναγνώριση είδους κειμένου (text-genre detection):** αφορά στην αναγνώριση του λειτουργικού ρόλου του κειμένου. Ένα είδος κειμένου περιλαμβάνει κείμενα που έχουν γραφτεί για τον ίδιο επικοινωνιακό σκοπό και εκπληρώνουν την ίδια κοινωνική αποστολή ανεξαρτήτως συγγραφέα (π.χ. επίσημα έγγραφα, επιστημονικά άρθρα, άρθρα εφημερίδας κ.α.).
- **Προσδιορισμός συγγραφέα (authorship attribution):** αφορά στην αναγνώριση του συγγραφέα του κειμένου.

Αυτοί οι στόχοι έχουν αντιμετωπιστεί έως τώρα σαν εντελώς ξεχωριστά προβλήματα. Γενικά, επικρατεί η αντίληψη ότι το σύνολο των υφολογικών δεικτών που κατορθώνει να αναπαραστήσει το ύφος ενός είδους κειμένου και το σύνολο των υφολογικών δεικτών που κατορθώνει να αναπαραστήσει το προσωπικό ύφος ενός συγγραφέα δεν μπορεί να είναι το ίδιο. Έτσι, επικρατούν διαφορετικές τάσεις για τους δύο αυτούς στόχους όσον αφορά την επιλογή των υφολογικών παραμέτρων.

Οι πιο σημαντικές προσεγγίσεις στην αυτόματη αναγνώριση είδους κειμένου επικεντρώνουν το ενδιαφέρον τους στην χρήση όσο το δυνατόν πιο απλών και εύκολα μετρήσιμων παραμέτρων με στόχο την εύκολη υλοποίηση ενός υπολογιστικού συστήματος [49, 54]. Οι υφολογικοί δείκτες που χρησιμοποιούν αναφέρονται έως και στο συντακτικό επίπεδο ανάλυσης. Στην ουσία όμως, προσπαθούν να αποφύγουν την πραγματική υπολογιστική ανάλυση του κειμένου παρά να την εκμεταλλευτούν.

Οι μελέτες προσδιορισμού συγγραφέα, απ' την άλλη, επικεντρώνουν το ενδιαφέρον τους στην αναγνώριση της πατρότητας ανώνυμων ή αμφισβητούμενων κειμένων [45]. Σχεδόν όλες οι προτεινόμενες προσεγγίσεις βασίζονται αποκλειστικά σε λεξιλογικούς δείκτες ύφους (δηλ. δείκτες που σχετίζονται είτε με συχνότητες εμφάνισης συγκεκριμένων λέξεων είτε με συναρτήσεις αναπαράστασης του πλούτου του λεξιλογίου). Είναι χαρακτηριστικό ότι σε μία εργασία ανασκόπησης της έρευνας στον προσδιορισμό συγγραφέα ο Holmes [44] σημειώνει:

> … έως σήμερα, καμία υφομετρική πρόταση δεν κατάφερε να καθιερώσει μία μεθοδολογία που να καταφέρνει να συλλάβει το ύφος ενός κειμένου καλύτερα από αυτήν που βασίζεται σε λεξιλογικά στοιχεία.

Πρόσφατα, αποδείχτηκε ότι η χρήση υφολογικών παραμέτρων συντακτικού επιπέδου έχει τουλάχιστον το ίδιο καλά αποτελέσματα με τις λεξιλογικές παραμέτρους. Ο Baayen [6] χρησιμοποίησε συχνότητες χρήσης κανόνων επανεγγραφής (rewrite rules) όπως εμφανίζονται σε ένα συντακτικά σχολιασμένο σώμα κειμένων. Η σύγκριση της μεθόδου αυτής με τις λεξιλογικές προσεγγίσεις έδειξε οι συχνότητες εμφάνισης κανόνων επανεγγραφής είναι πιο σημαντικοί από τις συχνότητες εμφάνισης λέξεων. Όμως, δεν παραλείπει να σημειώσει:

> Δεν είμαστε πολύ αισιόδοξοι σχετικά με τη χρήση πλήρως αυτοματοποιημένων συντακτικών αναλυτών, άλλα η μελλοντική έρευνα δεν θα πρέπει να αποκλείσει αυτήν τη δυνατότητα.

Μία τυπική προσέγγιση στον προσδιορισμό συγγραφέα περιλαμβάνει αρχικά τον ορισμό ενός συνόλου υφολογικών δεικτών και στην συνέχεια τον υπολογισμό αυτών των μεγεθών, είτε χειρονακτικά είτε με την εύρεση κατάλληλων υπολογιστικών εργαλείων ικανά να παρέχουν αυτές τις μετρήσεις αυτόματα. Στην δεύτερη περίπτωση, συχνά οι αυτόματα εξαγόμενες μετρήσεις διορθώνονται, αν χρειάζεται, από κάποιον άνθρωπο-ελεγκτή. Έτσι, η χρήση υπολογιστών στον προσδιορισμό συγγραφέα, όσον αφορά την εξαγωγή των υφολογικών δεικτών, έχει περιοριστεί έως σήμερα σε βοηθητικά εργαλεία που μετράνε συχνότητες εμφάνισης λέξεων γρήγορα και αξιόπιστα. Επομένως, οι μελέτες προσδιορισμού συγγραφέα μπορούν να θεωρηθούν ως *βοηθούμενες από υπολογιστή* (computer-assisted) παρά ως *βασιζόμενες σε υπολογιστή* (computer-based).

Μία εναλλακτική προσέγγιση που στοχεύει στην αυτόματη επιλογή των δεικτών ύφους έχει προταθεί από τους Forsyth και Holmes [37]. Πιο συγκεκριμένα, εκτελέστηκαν πειράματα κατηγοριοποίησης κειμένων (συμπεριλαμβανομένου του προσδιορισμού συγγραφέα) αφήνοντας στον υπολογιστή την εύρεση των αλφαριθμητικών εκείνων (όχι απαραίτητα ολοκληρωμένες λέξεις) που διαχωρίζουν καλύτερα τις κατηγορίες ενός δεδομένου σώματος κειμένων έχοντας ως βάση την διαδικασία εύρεσης χαρακτηριστικών Monte-Carlo. Τα αποτελέσματα που αναφέρονται δείχνουν ότι οι συχνότητες των αυτόματα εξαγόμενων αλφαριθμητικών είναι πιο αποτελεσματικές από τις συχνότητες των γραμμάτων ή των λέξεων. Αυτή η

μέθοδος απαιτεί ελάχιστο υπολογιστικό κόστος καθώς χειρίζεται πληροφορία χαμηλού επιπέδου. Όμως, παρά το ότι υποστηρίζεται πως αυτή η πληροφορία μπορεί να συνδυαστεί με συντακτικούς ή/και σημασιολογικούς δείκτες ύφους, δεν είναι καθόλου ξεκάθαρος ο τρόπος με τον οποίο μπορούν να χρησιμοποιηθούν ήδη υπάρχοντα εργαλεία επεξεργασίας φυσικής γλώσσας προς αυτήν την κατεύθυνση.

Η δική μας πρόταση για τον καθορισμό ενός συνόλου υφολογικών δεικτών δεν κάνει διάκριση ανάμεσα στην αναγνώριση είδους κειμένου και στον προσδιορισμό συγγραφέα. Αντίθετα, όπως θα δείξουμε στα επόμενα κεφάλαια, το σύνολο δεικτών ύφους που θα περιγράψουμε στην συνέχεια είναι ικανό να αναγνωρίσει αξιόπιστα κάθε υφολογικά ομοιογενή κατηγορία (δηλ. είτε κείμενα του ίδιου είδους κειμένου, είτε κείμενα του ίδιου συγγραφέα). Επιπλέον, η επιλογή του συνόλου των υφολογικών δεικτών σχετίζεται άμεσα με τα υπολογιστικά εργαλεία που χρησιμοποιούνται για τη μέτρησή τους, με αποτέλεσμα την επίτευξη ενός πλήρως αυτοματοποιημένου συστήματος. Επίσης, πρέπει να σημειωθεί ότι δεν χρησιμοποιείται καμιά λεξιλογική πληροφορία και έτσι αποφεύγεται κάθε είδους εξάρτηση από συγκεκριμένα είδη κειμένου ή συγκεκριμένους συγγραφείς.

Στο επόμενο τμήμα δίνεται μία συνοπτική περιγραφή των σημαντικότερων μεθοδολογιών προσδιορισμού ενός συνόλου υφολογικών δεικτών. Στο τμήμα 4.3 περιγράφεται αναλυτικά η πρότασή μας και στο τμήμα 4.4 η περίληψη του κεφαλαίου καθώς και τα συμπεράσματα που αποκομίστηκαν.

4.2 Προηγούμενες Υφομετρικές Προτάσεις

Σε αυτό το τμήμα θα προσπαθήσουμε να κατατάξουμε τις διάφορες υφολογικές παραμέτρους που έχουν προταθεί. Ως μέσο κατάταξης επιλέξαμε την πολυπλοκότητα της πληροφορίας στην οποία βασίζεται ο υπολογισμός τους και όχι το πρόβλημα στο οποίο εφαρμόστηκαν.

4.2.1 Επίπεδο δείγματος

Η πιο απλή προσέγγιση είναι να θεωρήσουμε το κείμενο ως ένα σύνολο δειγμάτων (ή λέξεων) ομαδοποιημένα σε περιόδους. Χαρακτηριστικά παραδείγματα αυτής της κατηγορίας υφολογικών δεικτών είναι: ο αριθμός των λέξεων, ο αριθμός των

περιόδων, ο αριθμός των συλλαβών, ο αριθμός των σημείων στίξης κτλ. Αυτές οι παράμετροι χρησιμοποιούνται ευρέως και στην αναγνώριση είδους κειμένου και στον προσδιορισμό συγγραφέα αφού είναι πολύ εύκολο (υπολογιστικά) να μετρηθούν. Μάλιστα, οι πρώτες προσπάθειες έρευνας στον προσδιορισμό συγγραφέα, όταν δεν ήταν διαθέσιμα ισχυρά υπολογιστικά συστήματα, βασιζόταν εξολοκλήρου σε τέτοιες μετρήσεις. Για παράδειγμα, ο Morton [65] χρησιμοποίησε το μήκος των περιόδων για να ελέγξει την πατρότητα κειμένων των Αρχαίων Ελληνικών ενώ ο Brinegar [15] υιοθέτησε το μήκος των λέξεων ως παράμετρο για να αποδείξει ότι ο Mark Twain δεν έγραψε το έργο *"The Qunitus Curtius Snodgrass Letters"*. Παρά την αναμφισβήτητη σπουδαιότητά τους, όμως, αυτές οι παράμετροι δεν μπορούν να οδηγήσουν σε αξιόπιστα αποτελέσματα από μόνες τους. Έτσι, πρέπει να χρησιμοποιήσουν ως συμπλήρωμα άλλες, πιο περίπλοκες παραμέτρους,

4.2.2 Συντακτικός σχολιασμός

Οι υφολογικοί δείκτες που βασίζονται σε συντακτική πληροφορία είναι πολύ σημαντικοί για τη διερεύνηση των χαρακτηριστικών του ύφους [11]. Τυπικά παραδείγματα αυτής της κατηγορίας είναι: μετρήσεις συχνότητας διαφόρων μερών-του-λόγου (ουσιαστικά, ρήματα, κτλ.), μετρήσεις παθητικής φωνής, ουσιαστικοποιήσεων, κ.ά. Η χρησιμοποίηση τέτοιων δεικτών είναι πολύ συνηθισμένη στην αναγνώριση είδους κειμένου, ενώ πρόσφατα ο Baayen [6] απέδειξε ότι η συντακτική πληροφορία μπορεί να βρει εφαρμογή και στον προσδιορισμό συγγραφέα. Όμως, ο υπολογισμός τους απαιτεί κείμενο συντακτικά αναλυμένο (parsed text) ή σχολιασμένο (tagged text). Επιπλέον, τα σύγχρονα εργαλεία επεξεργασίας κειμένου δεν είναι ικανά να παρέχουν ακριβείς και αξιόπιστες μετρήσεις για πολλές από τις παραμέτρους που έχουν προταθεί. Η περίπτωση της μελέτης του Biber [11] είναι χαρακτηριστική: μετά τον ορισμό ενός μεγάλου συνόλου υφολογικών δεικτών στους οποίους συμπεριλαμβάνονται και αρκετοί σχετικοί με συντακτική πληροφορία χρησιμοποίησε υπολογιστικά εργαλεία για να μετρήσει τις τιμές μερικών παραμέτρων (των πιο απλών) σε διάφορα κείμενα και οι τιμές των υπόλοιπων μετρήθηκαν χειρονακτικά. Επιπλέον, το υποσύνολο των παραμέτρων που μετρήθηκαν αυτόματα ελέγχθηκε χειρονακτικά για να διαπιστωθεί αν περιέχουν λάθη. Πολλοί ερευνητές, λοιπόν, προσπαθούν να αποφύγουν τη χρησιμοποίηση

υφολογικών δεικτών σχετικών με τη σύνταξη του κειμένου για να παρακάμψουν τέτοιου είδους δυσκολίες [54].

4.2.3 Πλούτος λεξιλογίου

Διάφοροι τρόποι έχουν προταθεί για την αναπαράσταση του πλούτου (ή της ποικιλίας) του λεξιλογίου ενός κειμένου (vocabulary richness) και έχουν εφαρμοστεί, ως επί το πλείστον, σε μελέτες προσδιορισμού συγγραφέα. Το πιο χαρακτηριστικό παράδειγμα αυτής της κατηγορίας είναι ο *λόγος τύπου-δείγματος* (type-token ratio) *V/N* όπου το *V* είναι το μέγεθος του λεξιλογίου του κειμένου και *N* είναι ο αριθμός των δειγμάτων που αποτελούν το κείμενο. Ένας άλλος τρόπος αναπαράστασης του πλούτου του λεξιλογίου είναι η μέτρηση του αριθμού των λέξεων που εμφανίζονται μία φορά στο κείμενο (άπαξ λεγόμενα), η μέτρηση του αριθμού των λέξεων που εμφανίζονται δύο φορές (δισλεγόμενα), κ.ο.κ. Αυτές οι μετρήσεις εξαρτώνται σε μεγάλο βαθμό από το μέγεθος του κειμένου. Για παράδειγμα, ο Sichel [83] αποδεικνύει ότι ο λόγος των δισλεγομένων προς το λεξιλόγιο του κειμένου (*V*) είναι πολύ ασταθές για $N < 1.000$. Για να αποφευχθεί αυτή η εξάρτηση, πολλοί ερευνητές έχουν προτείνει συναρτήσεις που υποστηρίζουν ότι είναι σταθερές σε σχέση με το μέγεθος του κειμένου. Χαρακτηριστικά παραδείγματα είναι οι συναρτήσεις *R* και *K* που έχουν προταθεί από τους Honore [46] και Yule [103], αντίστοιχα:

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

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

όπου το V_i είναι ο αριθμός των λέξεων που εμφανίζονται ακριβώς *i* φορές στο κείμενο (άρα V_1 είναι τα άπαξ λεγόμενα). Επιπλέον, υπάρχουν προσεγγίσεις βασιζόμενες σε ένα σύνολο από τέτοιες συναρτήσεις και σε τεχνικές της *πολυπαραγοντικής στατιστικής* (multivariate statistics), σε μία προσπάθεια να επιτύχουν πιο ακριβή αποτελέσματα [43]. Παρ' όλα αυτά, πρόσφατες μελέτες έδειξαν ότι η πλειοψηφία αυτών των συναρτήσεων δεν είναι στην πραγματικότητα ανεξάρτητες του μεγέθους του κειμένου [95]. Ειδικότερα, για σχετικά μικρά μεγέθη

κειμένων ($N < 1.000$ λέξεις), οι συναρτήσεις πλούτου λεξιλογίου είναι πολύ ασταθείς και ιδιαιτέρως αναξιόπιστες.

4.2.4 Συχνότητα λέξεων

Αντί της μέτρησης των λέξεων που εμφανίζονται συγκεκριμένες φορές σε ένα κείμενο, μία εναλλακτική προσέγγιση ερευνά πόσες φορές εμφανίζονται στο κείμενο συγκεκριμένες λέξεις. Κατάλληλες για διαχωρισμό υφολογικών κατηγοριών θεωρούνται οι λέξεις που είναι ανεξάρτητες από τα συμφραζόμενα, οι οποίες συχνά καλούνται *λειτουργικές λέξεις* (function words). Τέτοιες μετρήσεις έχουν εφαρμοστεί σε σχεδόν όλες τις μελέτες αναγνώρισης είδους κειμένου και προσδιορισμού συγγραφέα αφού αποτελούν ένα αξιόπιστο παράγοντα διαχωρισμού. Ο υπολογισμός τους σε ένα κείμενο είναι μεν απλός αλλά απαιτείται πολύς κόπος και αρκετοί πειραματισμοί για την επιλογή των πιο κατάλληλων λειτουργικών λέξεων για κάποιο συγκεκριμένο πρόβλημα [66]. Επιπλέον, εφόσον βρεθεί ένα συγκεκριμένο σύνολο λέξεων, των οποίων η συχνότητα εμφάνισης μπορεί να χρησιμοποιηθεί για τον αξιόπιστο διαχωρισμό ενός συνόλου υφολογικών κατηγοριών, δεν συνεπάγεται ότι μπορεί να εφαρμοστεί σε κάποιο διαφορετικό σύνολο κατηγοριών με την ίδια επιτυχία. Επίσης, έχει προταθεί η ομαδοποίηση συγκεκριμένων λέξεων σε κατηγορίες όπως ιδιωματικές εκφράσεις, επιστημονική ορολογία, «επίσημες» λέξεις, κτλ. [63]. Αν και αυτή η λύση είναι αρκετά πιο γενική, απαιτεί την κατασκευή ενός περίπλοκου υπολογιστικού μηχανισμού για την αυτόματη ανίχνευση και την μέτρηση τέτοιων παραμέτρων σε ένα κείμενο.

Η θεώρηση ότι οι λειτουργικές λέξεις είναι ανεξάρτητες από τα συμφραζόμενα έχει αμφισβητηθεί από αρκετούς ερευνητές. Ο Damerau [27] βασίστηκε στην υπόθεση ότι για να είναι μία λέξη ανεξάρτητη των συμφραζομένων, η κατανομή της μέσα σε οποιοδήποτε κείμενο πρέπει να ακολουθεί την κατανομή Poisson. Τα πειράματα που έκανε προς αυτήν την κατεύθυνση έδειξαν ότι οι λέξεις που ακολουθούν την κατανομή Poisson στα κείμενα ορισμένων συγγραφέων δεν την ακολουθούν στα κείμενα άλλων συγγραφέων. Ακόμη, σε μερικούς συγγραφείς ακολουθούν την κατανομή Poisson πολλές λέξεις και σε άλλους συγγραφείς ελάχιστες. Ο Oakman [69] μάλιστα υποστηρίζει:

> Το μάθημα φαίνεται να είναι σαφές όχι μόνο για τις λειτουργικές λέξεις αλλά γενικά για τις μελέτες προσδιορισμού συγγραφέα που βασίζονται σε λέξεις: συγκεκριμένες λέξεις μπορεί να δουλεύουν για κάποιες περιπτώσεις, όπως η υπόθεση *The Federalist Papers,* ωστόσο δεν μπορούν να ληφθούν υπ' όψιν σε άλλες αναλύσεις.

Τελευταία, αλλά όχι λιγότερο σημαντική, αναφέρουμε την προσέγγιση που έχει προταθεί από τον Burrows [18, 19], σύμφωνα με την οποία, ως σύνολο υφολογικών δεικτών για τον προσδιορισμό συγγραφέα χρησιμοποιούνται οι συχνότητες εμφάνισης των πιο συχνά χρησιμοποιούμενων λέξεων (συνήθως σύνολα των 30 ή 50 λέξεων), χωρίς να γίνεται διάκριση ανάμεσα σε λειτουργικές ή όχι λέξεις. Αυτή φαίνεται να είναι η πιο πολλά υποσχόμενη προσέγγιση αφού απαιτεί ελάχιστο υπολογιστικό κόστος και έχει επιτύχει αξιόλογα αποτελέσματα για ένα ευρύ φάσμα συγγραφέων. Η διάκριση μεταξύ κοινών ομογραφικών μορφών, δηλ. λέξεων που μπορεί να έχουν δύο ή περισσότερες μορφές (π.χ. η λέξη *που* είναι είτε αναφορική αντωνυμία είτε ειδικός σύνδεσμος) βελτιώνει την ακρίβεια. Όμως, όσον αφορά ένα πλήρως αυτοματοποιημένο σύστημα, αυτός ο διαχωρισμός απαιτεί την ανάπτυξη αξιόπιστων εργαλείων επεξεργασίας φυσικής γλώσσας ικανά να αναγνωρίζουν αυτές τις διαφορετικές μορφές σε οποιοδήποτε κείμενο. Επιπλέον, στην περίπτωση που τα κύρια ονόματα εξαιρούνται από τη λίστα των πιο συχνά χρησιμοποιούμενων λέξεων, το σύστημα πρέπει να είναι εφοδιασμένο με έναν ανιχνευτή κυρίων ονομάτων.

4.3 Η Πρότασή μας

Τα κύρια χαρακτηριστικά της πρότασής μας, όσον αφορά τους υφολογικούς δείκτες, είναι τα εξής:

- **Συνδυασμός επιπέδου-δείγματος και συντακτικής πληροφορίας:** οι προτεινόμενοι δείκτες εκμεταλλεύονται τα πλεονεκτήματα των δεικτών επιπέδου-δείγματος και σύνταξης.
- **Μη-χρήση λεξιλογικής πληροφορίας:** δεν χρησιμοποιείται κανένας από τους παραδοσιακούς λεξιλογικούς δείκτες (ούτε συναρτήσεις πλούτου λεξιλογίου ούτε συχνότητες εμφάνισης λέξεων).
- **Εκμετάλλευση εργαλείων επεξεργασίας κειμένου:** το σύνολο των υφολογικών δεικτών σχεδιάστηκε με στόχο την πλήρη αξιοποίηση της

αυτόματης επεξεργασίας κειμένου μέσω των ανιχνευτών ορίων περιόδων και φράσεων που παρουσιάστηκαν στα προηγούμενα κεφάλαια.

- **Πλήρης αυτοματοποίηση:** ως αποτέλεσμα της χρήσης των ανιχνευτών ορίων και φράσεων, η διαδικασία την εξαγωγής της υφολογικής πληροφορίας είναι πλήρως αυτοματοποιημένη. Δεν απαιτείται καμία χειρονακτική προεπεξεργασία κειμένου (χωρισμός ή δειγματοληψία κειμένου).

Το προτεινόμενο σύνολο δεικτών ύφους μπορεί να διακριθεί σε τρία υφομετρικά επίπεδα. Τα πρώτα δύο ασχολούνται με την έξοδο του ανιχνευτή ορίων περιόδων και φράσεων. Να σημειωθεί ότι ο κάθε υφολογικός δείκτης προκύπτει από το λόγο δύο σχετικών μεγεθών για να επιτευχθεί ομαλοποίηση των τιμών σχετικά με το μέγεθος του κειμένου.

1. **Επίπεδο δείγματος:** Αυτό το επίπεδο δεικτών βασίζεται στην έξοδο του ανιχνευτή ορίων περιόδων. Το κείμενο εισόδου θεωρείται ως μία ακολουθία δειγμάτων ομαδοποιημένα σε περιόδους. Υπάρχουν τρεις τέτοιοι δείκτες ύφους:

<u>Κωδικός</u>	<u>Περιγραφή</u>
Δ01	*περίοδοι / λέξεις*
Δ02	*σημεία στίξης / λέξεις*
Δ03	*περίοδοι / υποψήφια όρια περιόδων*

2. **Επίπεδο φράσης:** Αυτό το επίπεδο δεικτών βασίζεται στην έξοδο του ανιχνευτή ορίων φράσεων. Το κείμενο εισόδου θεωρείται ως μία ακολουθία φράσεων. Κάθε φράση περιέχει τουλάχιστον μία λέξη. Υπάρχουν δέκα τέτοιοι δείκτες ύφους:

<u>Κωδικός</u>	<u>Περιγραφή</u>
Δ04	*αριθμός ΟΦ / συνολικές φράσεις*
Δ05	*αριθμός ΡΦ / συνολικές φράσεις*
Δ06	*αριθμός ΕΦ / συνολικές φράσεις*
Δ07	*αριθμός ΠΦ / συνολικές φράσεις*
Δ08	*αριθμός ΣΦ / συνολικές φράσεις*
Δ09	*λέξεις που περιέχονται σε ΟΦ / αριθμός ΟΦ*

Δ10	*λέξεις που περιέχονται σε ΡΦ / αριθμός ΡΦ*
Δ11	*λέξεις που περιέχονται σε ΕΦ / αριθμός ΕΦ*
Δ12	*λέξεις που περιέχονται σε ΠΦ / αριθμός ΠΦ*
Δ13	*λέξεις που περιέχονται σε ΣΦ / αριθμός ΣΦ*

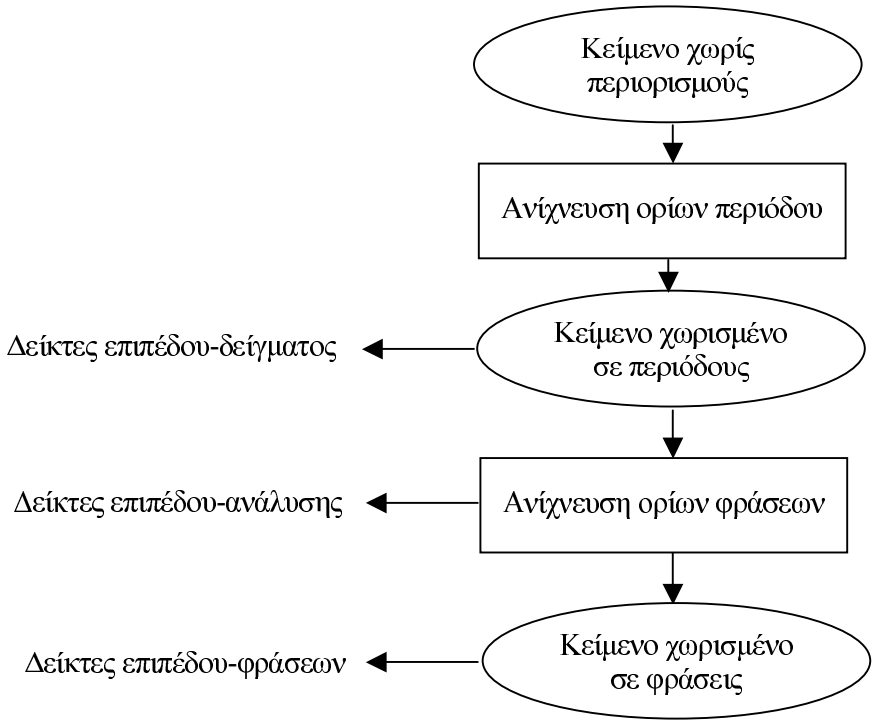
Καθώς ο συνδυασμός των ανιχνευτών ορίων περιόδων και φράσεων είναι ένα αυτόματο υπολογιστικό εργαλείο, οι παραπάνω υφολογικοί δείκτες μετρούνται προσεγγιστικά. Ανάλογα με τη συντακτική ασάφεια του κειμένου, οι αυτόματα εξαγόμενες μετρήσεις μπορεί να διαφέρουν από αυτές που θα προέκυπταν από την χειρονακτική μέτρηση αυτών των τιμών. Για να αντιμετωπίσουμε αυτό το πρόβλημα, προσπαθήσαμε να χρησιμοποιήσουμε ένα εναλλακτικό τρόπο σύλληψης της υφολογικής πληροφορίας. Έτσι, προτείνουμε ένα νέο επίπεδο υφολογικών δεικτών:

3. **Επίπεδο ανάλυσης:** Περιλαμβάνει δείκτες ύφους που αναπαριστούν τον τρόπο με τον οποίο έγινε η ανάλυση του κειμένου από τον ανιχνευτή ορίων φράσεων. Υπάρχουν εννιά τέτοιοι δείκτες:

Κωδικός	*Περιγραφή*
Δ14	*λέξεις-κλειδιά του κειμένου / λέξεις*
Δ15	*λέξεις χωρίς κοινές καταλήξεις του κειμένου / λέξεις*
Δ16	*μορφολογικές περιγραφές λέξεων / λέξεις*
Δ17	*μορφολογικές περιγραφές φράσεων / συνολικές φράσεις*
Δ18	*λέξεις που έμειναν χωρίς ανάλυση μετά το 1ο πέρασμα / λέξεις*
Δ19	*λέξεις που έμειναν χωρίς ανάλυση μετά το 2ο πέρασμα / λέξεις*
Δ20	*λέξεις που έμειναν χωρίς ανάλυση μετά το 3ο πέρασμα / λέξεις*
Δ21	*λέξεις που έμειναν χωρίς ανάλυση μετά το 4ο πέρασμα / λέξεις*
Δ22	*λέξεις που έμειναν χωρίς ανάλυση μετά το 5ο πέρασμα / λέξεις*

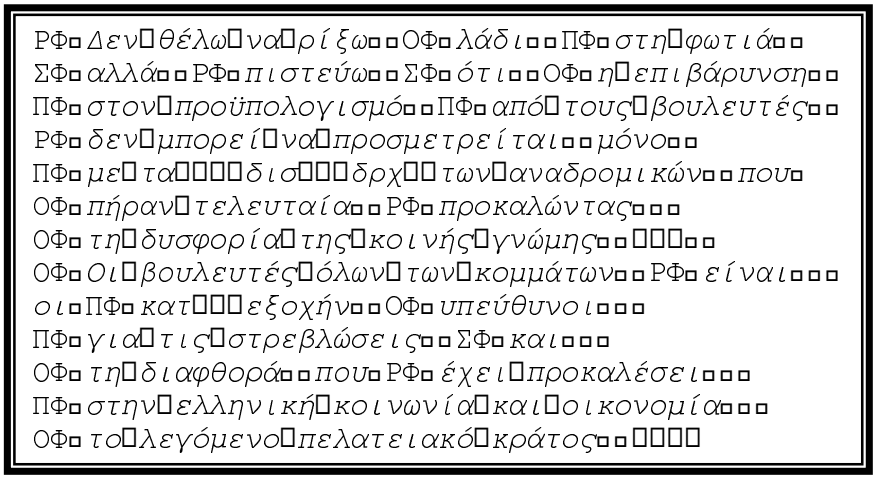
Οι δείκτες Δ14 και Δ15 αποτελούν ένα εναλλακτικό τρόπο μέτρησης της λεξιλογικής πολυπλοκότητας του κειμένου καθώς φανερώνουν πόσες κοινές λέξεις και πόσες ξένες ή μη-κοινές λέξεις εμφανίζονται στο κείμενο, αντίστοιχα. Οι δείκτες Δ16 και Δ17 αναπαριστούν το βαθμό της μορφολογικής αμφισημίας των λέξεων και των

παραγόμενων φράσεων αντίστοιχα. Τέλος, οι δείκτες Δ18-Δ22 υποδηλώνουν τη συντακτική πολυπλοκότητα του κειμένου. Εφόσον τα πρώτα περάσματα ανάλυσης αναγνωρίζουν τις πιο απλές περιπτώσεις, είναι εύκολο να αντιληφθούμε ότι ένα μεγάλο μέρος ενός συντακτικά πολύπλοκου κειμένου θα παρέμενε χωρίς ανάλυση από αυτά. Επομένως, μεγάλες τιμές των δεικτών Δ18 και Δ19 σε συνδυασμό με χαμηλές τιμές των Δ21 και Δ22 σηματοδοτούν συντακτικά πολύπλοκο κείμενο και αντίστροφα. Στο σχήμα 4.1 φαίνεται η διαδικασία εξαγωγής των υφολογικών δεικτών.



Σχήμα 4.1. Τα τρία επίπεδα υφολογικών δεικτών.

Για να γίνει πιο κατανοητός ο τρόπος υπολογισμού των 22 παραμέτρων που προτείνουμε, παραθέτουμε το ακόλουθο παράδειγμα: Στο σχήμα 4.2 φαίνεται η ανάλυση ενός δείγματος κειμένου από τον ανιχνευτή ορίων περιόδων και φράσεων. Οι τιμές του συνόλου των υφολογικών δεικτών για αυτό το κείμενο δίνονται στον πίνακα 4.1. Οι λέξεις που δεν ταιριάζουν με κάποια από τις κοινές καταλήξεις (ειδικές λέξεις) σημειώνονται με ένα αστεράκι (*).



Σχήμα 4.2. Ανάλυση δείγματος κειμένου από τον ανιχνευτή ορίων περιόδων και φράσεων.

Κωδικός	Τιμή	Κωδικός	Τιμή	Κωδικός	Τιμή
Δ01	0.03 (2/66)	Δ09	2.75 (22/8)	Δ17	1.83 (44/24)
Δ02	0.08 (5/66)	Δ10	2.17 (13/6)	Δ18	0.29 (19/66)
Δ03	0.50 (2/4)	Δ11	0.00	Δ19	0.20 (13/66)
Δ04	0.33 (8/24)	Δ12	3.43 (24/7)	Δ20	0.20 (13/66)
Δ05	0.25 (6/24)	Δ13	1.00 (3/3)	Δ21	0.05 (3/66)
Δ06	0.00 (0/24)	Δ14	0.54 (36/66)	Δ22	0.05 (3/66)
Δ07	0.29 (7/24)	Δ15	0.05 (3/66)		
Δ08	0.12 (3/24)	Δ16	1.62 (107/66)		

Πίνακας 4.1. Οι τιμές των υφολογικών δεικτών για το δείγμα κειμένου του σχήματος 4.2.

4.4 Περίληψη - Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε την προσέγγισή μας για την εξαγωγή υφολογικών δεικτών, δηλ. μετρήσιμων παραμέτρων που ποσοτικοποιούν και αναπαριστούν το ύφος ενός κειμένου. Οι προτεινόμενοι δείκτες μπορούν να χρησιμοποιηθούν για την αναπαράσταση του ύφους οποιουδήποτε κειμένου καθώς βασίζονται σε αυτόματα εργαλεία επεξεργασίας κειμένου. Όπως θα δείξουμε στα επόμενα κεφάλαια, μπορούν να χρησιμοποιηθούν και για αναγνώριση είδους κειμένου και για προσδιορισμό συγγραφέα και γενικά για αναγνώριση κάθε υφολογικά ομοιογενούς κατηγορίας.

Το προτεινόμενο σύνολο δεικτών ύφους εκμεταλλεύεται τα πλεονεκτήματα των παραμέτρων που σχετίζονται με το επίπεδο δείγματος και το επίπεδο σύνταξης.

Αντίθετα, δεν εμπλέκει καμία από τις ευρέως χρησιμοποιούμενες λεξιλογικές παραμέτρους (συναρτήσεις πλούτου λεξιλογίου ή συχνότητες εμφάνισης συγκεκριμένων λέξεων). Έτσι αποφεύγεται οποιαδήποτε εξάρτηση από συγκεκριμένες ομάδες συγγραφέων ή ειδών κειμένου. Επίσης, με αυτόν τον τρόπο το σύνολο των 22 δεικτών ύφους που προτείνουμε είναι πιο ανεξάρτητο-γλώσσας.

Αξίζει να σημειωθεί ότι δεν υποστηρίζουμε πως το προτεινόμενο σύνολο υφολογικών δεικτών είναι πλήρες ούτε το ιδανικό για όλες τις περιπτώσεις. Ο δείκτης Δ02, για παράδειγμα, θα μπορούσε να διακριθεί σε επιμέρους δείκτες όπως τελείες προς λέξεις, κόμματα προς λέξεις, ερωτηματικά προς λέξεις κτλ. Ο στόχος αυτής της διατριβής είναι η παρουσίαση ενός αξιόπιστου συνόλου υφολογικών δεικτών που να οδηγεί σε ικανοποιητικά αποτελέσματα ταξινόμησης κειμένων ως προς το είδος και το συγγραφέα τους αλλά και να υποδείξει τον τρόπο με τον οποίο ήδη υπάρχοντα εργαλεία φυσικής γλώσσας μπορούν να χρησιμοποιηθούν για την εξαγωγή υφολογικής πληροφορίας.

Συνήθως, οι ερευνητές της υφομετρίας προσπαθούν να προκαθορίσουν ένα αξιόπιστο σύνολο δεικτών ύφους και στην συνέχεια ψάχνουν για υπολογιστικά εργαλεία κατάλληλα να παρέχουν αυτόματα τις αντίστοιχες μετρήσεις. Αντίθετα, εμείς προσαρμόσαμε το σύνολο των δεικτών ύφους στα διαθέσιμα εργαλεία επεξεργασίας φυσικής γλώσσας, με γνώμονα την πλήρη αξιοποίηση της υφολογικής πληροφορίας που μπορούσαν να εξάγουν, με οποιοδήποτε τρόπο. Προς αυτήν την κατεύθυνση, ορίσαμε ένα νέο είδος υφολογικών δεικτών, τις παραμέτρους επιπέδου-ανάλυσης, που είναι ένας εναλλακτικός τρόπος σύλληψης της υφολογικής πληροφορίας. Έτσι, ακόμα και το ποσοστό του κειμένου που δεν μπόρεσε να αναλύσει το σύστημα χρησιμοποιείται ως υφολογική παράμετρος.

Οι δείκτες ύφους επιπέδου-ανάλυσης μπορούν να υπολογιστούν μόνο όταν χρησιμοποιείται ο συγκεκριμένος ανιχνευτής ορίων περιόδων και φράσεων για την ανάλυση του κειμένου. Όμως, ο ανιχνευτής ορίων περιόδων και φράσεων που παρουσιάστηκε στα προηγούμενα κεφάλαια είναι ένα υπολογιστικό εργαλείο γενικού σκοπού και δεν σχεδιάστηκε ειδικά για να χρησιμοποιηθεί στην αυτόματη εξαγωγή υφολογικών δεικτών. Έτσι, κάθε εργαλείο επεξεργασίας φυσικής γλώσσας (π.χ. σχολιαστής μέρους-του-λόγου, συντακτικός αναλυτής κ.ά.) μπορεί να χρησιμοποιηθεί για εξαγωγή υφολογικής πληροφορίας με τον κατάλληλο ορισμό των δεικτών ύφους

επιπέδου-ανάλυσης. Ο τρόπος ορισμού τους, βέβαια, εξαρτάται άμεσα από τη μέθοδο στην οποία βασίζεται το εργαλείο για να πραγματοποιήσει την ανάλυση. Πρέπει να σημειωθεί ότι αυτή η διαδικασία είναι ανεξάρτητη-γλώσσας.

Για να γίνει αυτό πιο αντιληπτό αξίζει να παραθέσουμε το ακόλουθο παράδειγμα. Κάποιες παρόμοιες μετρήσεις έχουν χρησιμοποιηθεί σε υφολογικά πειράματα πάνω στην ανάκτηση πληροφορίας. Για τον σκοπό αυτό χρησιμοποιήθηκε ένας εύρωστος συντακτικός αναλυτής που δημιουργήθηκε ειδικά για εφαρμογές ανάκτησης πληροφορίας [91]. Ο αναλυτής αυτός παράγει δέντρα ανάλυσης για να αναπαραστήσει την δομή των περιόδων που συνθέτουν το κείμενο. Επίσης, όταν η χρονική διάρκεια ανάλυσης μιας πρότασης μιας περιόδου ξεπεράσει κάποιο προκαθορισμένο κατώφλι σταματά την ανάλυση και συνεχίζει με την επόμενη περίοδο. Όταν σταματά κάποια προσπάθεια, ο αναλυτής το σημειώνει στο δέντρο ανάλυσης. Οι μετρήσεις που χρησιμοποιούνται ως δείκτες της πολυπλοκότητας των προτάσεων είναι ο μέσος όρος του βάθους του δέντρου ανάλυσης και ο αριθμός των φορών που σταμάτησε η ανάλυση προς τον αριθμό των περιόδων [51]. Πρόκειται δηλαδή για μετρήσεις επιπέδου ανάλυσης που έχουν να κάνουν με τον τρόπο που λειτουργεί το συγκεκριμένο εργαλείο επεξεργασίας φυσικής γλώσσας.