Κεφάλαιο 3

Ανίχνευση Ορίων Φράσεων

> It will be bags of tricks and not theory that would advance computational linguistics in the future [M. Bar-Hillel]

3.1 Εισαγωγή

Τα τελευταία χρόνια, με την ανάπτυξη μεγάλων βάσεων δεδομένων έχει γίνει διαθέσιμος μεγάλος όγκος κειμένων σε ηλεκτρονική μορφή. Ως επί το πλείστον, οι βάσεις αυτές περιλαμβάνουν *κείμενα χωρίς περιορισμούς* στο μέγεθος, στη μορφή ή στην πολυπλοκότητα (unrestricted text). Τέτοια κείμενα συνήθως περιέχουν τίτλους και άλλες μη-προτασιακές φόρμες, διαλέκτους και ιδιωματικές εκφράσεις, καθώς και πληθώρα λέξεων που δε βρίσκονται ούτε στα πιο ογκώδη ηλεκτρονικά λεξικά. Επίσης, τα κείμενα αυτά μπορεί να περιέχουν λάθη, ειδικά στην περίπτωση που προέρχονται από εργαλεία οπτικής αναγνώρισης χαρακτήρων (optical character recognition tools).

Πολλές εφαρμογές της επεξεργασίας φυσικής γλώσσας, όπως η *ανάκτηση πληροφορίας* (information retrieval) [60] και η *εξαγωγή πληροφορίας* (information extraction) [25], απαιτούν γρήγορη και εύρωστη ανάλυση μεγάλου όγκου κειμένων

χωρίς περιορισμούς. Σε τέτοιες εφαρμογές το πρώτιστο μέλημα είναι η πολύ γρήγορη ανάλυση κειμένων έστω και αν τα αποτελέσματα της ανάλυσης δεν είναι τα βέλτιστα.

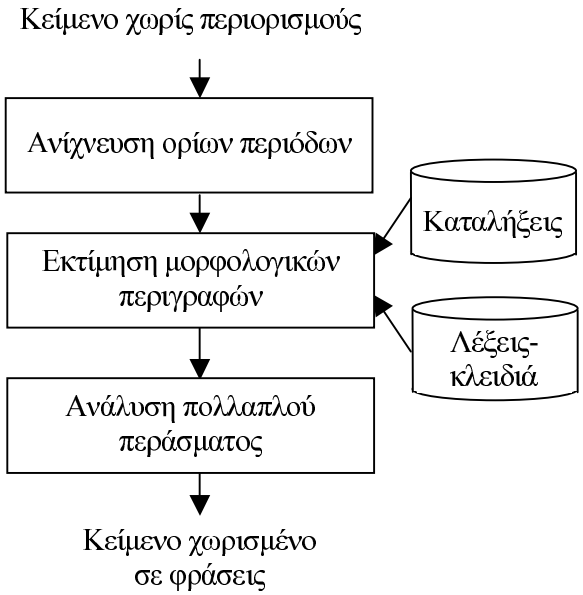
Γενικά, ο στόχος ενός *αναλυτή* (parser) είναι να αντιστοιχήσει μία κατάλληλη δομή στο κείμενο εισόδου. Για παράδειγμα, ο συντακτικός αναλυτής αντιστοιχεί μία τουλάχιστον συντακτική δομή σε ένα κείμενο. Σε αυτό το κεφάλαιο παρουσιάζουμε έναν αναλυτή που προσδιορίζει τα όρια των ενδοπεριοδικών, μη-επικαλυπτόμενων φράσεων (text chunking) σε κείμενο χωρίς περιορισμούς. Σχεδόν όλες οι προηγούμενες προσεγγίσεις σε αυτό το πρόβλημα βασίζονται σε ογκώδεις και πολύπλοκους πόρους, όπως λεξικά που περιέχουν δεκάδες χιλιάδες λήμματα και γραμματικές που αποτελούνται από εκατοντάδες ή ακόμα και χιλιάδες κανόνες [52]. Αυτή η λύση απαιτεί τεράστιο υπολογιστικό κόστος και η κατασκευή ενός τέτοιου συστήματος καθώς και η τροποποίησή του είναι πολύ δύσκολη και χρονοβόρα διαδικασία.

Ένα από τα μεγαλύτερα προβλήματα των αναλυτών που βασίζονται σε ογκώδη λεξικά είναι οι άγνωστες λέξεις. Μερικοί ερευνητές χρησιμοποιούν ευρετικές μεθόδους (όπως η αναγνώριση συγκεκριμένων καταλήξεων ή ο έλεγχος για το αν το πρώτο γράμμα της λέξης είναι κεφαλαίο για την περίπτωση κυρίων ονομάτων) και άλλοι απλά αγνοούν όλες τις άγνωστες λέξεις και προσπαθούν να αναλύσουν το υπόλοιπο μέρος του κειμένου [42]. Πρόσφατα, αναπτύχθηκαν στατιστικές μέθοδοι που παρέχουν την πιο πιθανή εκτίμηση σχετικά με τη μορφο-συντακτική πληροφορία των λέξεων που δε βρέθηκαν στο λεξικό [29, 64].

Η εργασία που παρουσιάζεται εδώ, προσπαθεί να ελαχιστοποιήσει τους πόρους στους οποίους βασίζεται η ανάλυση, εκμεταλλευόμενη πλήρως τα χαρακτηριστικά της Νέας Ελληνικής γλώσσας. Πιο συγκεκριμένα, τα Νέα Ελληνικά είναι μία γλώσσα χωρίς προκαθορισμένη διάταξη των όρων της πρότασης (quasi-free word order) και εξαιρετικά περίπλοκη μορφολογικά, περιλαμβάνοντας πληθώρα κλιτικών κατηγοριών [107]. Η εμπειρική παρατήρηση έδειξε ότι οι καταλήξεις των λέξεων, και ιδίως των ρημάτων, είναι πολύ χαρακτηριστικές και σε πάρα πολλές περιπτώσεις μπορούν να οδηγήσουν σε ασφαλή συμπεράσματα για τη μορφολογική περιγραφή της κάθε λέξης. Επιπλέον, η χρήση άρθρων, προθέσεων, και μορίων είναι υποχρεωτική ακόμα και μπροστά από κύρια ονόματα και συνήθως τέτοιες λέξεις σηματοδοτούν την έναρξη κάποιας ονοματικής, προθετικής ή ρηματικής φράσης αντίστοιχα. Πρέπει να

επισημανθεί ότι παρόμοιες ιδιότητες χαρακτηρίζουν και άλλες γλώσσες όπως τα Ισπανικά, τα Γερμανικά, και τα Ιταλικά.

Οι πόροι που χρησιμοποιούνται, λοιπόν, για την εύρεση της μορφολογικής πληροφορίας των λέξεων είναι: ένα λεξικό περίπου 450 λέξεων-κλειδιών (keywords), ή αλλιώς λέξεων κλειστής τάξης (closed-class), όπως άρθρα, προθέσεις, αντωνυμίες κ.ά., και ένα λεξικό περίπου 300 κοινών καταλήξεων που περιέχει τις πιο συνηθισμένες καταλήξεις των Νέων Ελληνικών μαζί με τις μορφολογικές πληροφορίες που μπορεί να υποδηλώνουν (π.χ. η κατάληξη *-ει* υποδηλώνει ρήμα στο τρίτο πρόσωπο). Έτσι, σύμφωνα με την προσέγγισή μας, και σε αντίθεση με τη φιλοσοφία των άλλων προσεγγίσεων, η διαδικασία εξαγωγής της πιο πιθανής εκτίμησης για τη μορφολογική πληροφορία των λέξεων αντικαθιστά πλήρως τα ογκώδη λεξικά των χιλιάδων λημμάτων.



Σχήμα 3.1. Δομή του συστήματος ανίχνευσης ορίων φράσεων.

Η αναγνώριση των ορίων των φράσεων στο κείμενο πραγματοποιείται μέσω *ανάλυσης πολλαπλού περάσματος* (multiple-pass parsing), μία τεχνική που έχει εφαρμοστεί κυρίως σε στατιστικούς αναλυτές με στόχο την επίτευξη βελτιωμένων αποτελεσμάτων ανάλυσης [39] καθώς και στην περιοχή αναγνώρισης ομιλίας ως μέσο ουσιαστικής μείωσης της υπολογιστικής πολυπλοκότητας, χωρίς αντίστοιχη αύξηση του σφάλματος αναγνώρισης [80]. Ακόμη, δίνει τη δυνατότητα προσαρμογής της ανάλυσης στις ανάγκες μιας συγκεκριμένης εφαρμογής, αφού η πολυπλοκότητα

της ανάλυσης μπορεί να ρυθμιστεί εύκολα με την επιλογή των κατάλληλων περασμάτων.

Ο ανιχνευτής ορίων φράσεων συνεργάζεται με τον ανιχνευτή ορίων περιόδων που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Έτσι η είσοδος του ανιχνευτή ορίων φράσεων είναι κείμενο χωρισμένο σε περιόδους. Η δομή του ολοκληρωμένου συστήματος φαίνεται στο σχήμα 3.1.

Στο επόμενο τμήμα περιγράφονται εν συντομία οι προηγούμενες προσεγγίσεις στην ανίχνευση ορίων φράσεων. Στα τμήματα 3.3 και 3.4 περιγράφεται αναλυτικά η διαδικασία εκτίμησης της μορφολογικής πληροφορίας της κάθε λέξης και η ανάλυση πολλαπλού περάσματος που πραγματοποιείται στο κείμενο εισόδου, αντίστοιχα. Στο τμήμα 3.5 γίνεται η αξιολόγηση της προτεινόμενης μεθόδου ενώ στο τμήμα 3.6 παρουσιάζεται μία εναλλακτική πρόταση βασιζόμενη σε ένα μεγάλο λεξικό λημμάτων και δίνονται συγκριτικά αποτελέσματα. Τέλος, στο τμήμα 3.7 περιλαμβάνονται η περίληψη του κεφαλαίου και τα συμπεράσματα που αποκομίστηκαν από αυτήν την έρευνα.

3.2 Σχετική Έρευνα

Η αναγνώριση ορίων ενδοπεριοδικών φράσεων (text chunking) έχει ως στόχο το χωρισμό των περιόδων σε σχετικά απλές συντακτικές δομές, όπως ονοματικές φράσεις και ρηματικές φράσεις. Προτάθηκε από τον Abney [1] ως ένα βήμα προεπεξεργασίας κειμένου πριν από την πλήρη συντακτική ανάλυση (full parsing).

Ο Μίχος [62] παρουσιάζει έναν αναλυτή για τη Νέα Ελληνική γλώσσα που προσδιορίζει τον τύπο των προτάσεων που αποτελούν μία περίοδο καθώς και τις φράσεις που αποτελούν την κάθε πρόταση, με χρήση ενός συνόλου λέξεων-κλειδιών και ενός συνόλου ευρετικών κανόνων. Αναφέρεται ακρίβεια της τάξης του 84% (η ακρίβεια ανεβαίνει στο 96% μετά τη χρήση κάποιων επιπρόσθετων ευρετικών τεχνικών). Αυτή η προσέγγιση απαιτεί πλήρη και αποσαφηνισμένη μορφολογική ανάλυση της κάθε λέξης του κειμένου. Επιπλέον, σε περίπτωση που το κείμενο περιέχει άγνωστες λέξεις ή είναι ιδιάζουσας σύνταξης αποτυγχάνει και δεν επιστρέφει καμιά χρήσιμη πληροφορία.

Ένα σύστημα ανεξάρτητο-γλώσσας για την ανάλυση κειμένου χωρίς περιορισμούς βάσει του φορμαλισμού της *Γραμματικής Περιορισμών* (Constraint Grammar) παρουσιάζεται από τον Karlsson [52]. Όπως υποστηρίζεται, αυτή η προσέγγιση παρέχει μορφολογική και συντακτική αποσαφήνιση οποιουδήποτε τμήματος κειμένου. Όμως, απαιτεί ένα πολύ ογκώδες κύριο λεξικό και μερικά λεξικά εξαρτημένα από την εφαρμογή καθώς και μία πολύ μεγάλη γραμματική αποτελούμενη από χιλιάδες κανόνες.

Ως εναλλακτική πρόταση στους συμβατικούς αναλυτές, οι *επιφανειακοί αναλυτές* (shallow parsers) παρέχουν αναλύσεις που είναι λιγότερο πλήρεις. Γενικά, ένας επιφανειακός αναλυτής αναγνωρίζει μερικές φραστικές δομές, όπως τις ονοματικές φράσεις, χωρίς να προσδιορίζει την εσωτερική δομή τους και το ρόλο τους μέσα στην περίοδο [23].

Ένας ανιχνευτής ορίων φράσεων που χρησιμοποιεί την *εκμάθηση βάσει μετασχηματισμών* (transformation-based learning) [13] περιγράφεται από τους Ramshaw και Marcus [73]. Αυτή η προσέγγιση καταφέρνει να δώσει ανάκληση και ακρίβεια της τάξης του 92% περίπου για απλές ονοματικές φράσεις και της τάξης του 88% για κάπως πιο περίπλοκα είδη φράσεων. Επίσης, μία στοχαστική προσέγγιση στην αναγνώριση ορίων φράσεων με χρήση μοντέλων Μαρκώφ περιγράφεται από τους Skut και Brants [85]. Και οι δύο αυτές προσεγγίσεις απαιτούν ως προεπεξεργασία ένα *σχολιαστή μέρους-του-λόγου* (part-of-speech tagger), από την ακρίβεια του οποίου εξαρτάται άμεσα και η δική τους ακρίβεια.

Το σύστημα *LEXTER* [16] είναι ένας επιφανειακός συντακτικός αναλυτής που εξάγει ονοματικές φράσεις μέγιστου μήκους από Γαλλικά κείμενα για εφαρμογές *αυτόματης εξαγωγής ορολογίας* (terminology acquisition). Υποστηρίζεται ότι το 95% όλων των ονοματικών φράσεων μέγιστου μήκους αναγνωρίζεται σωστά αλλά η αντίστοιχη ακρίβεια δεν αναφέρεται. Ένα άλλο σύστημα εξαγωγής ονοματικών φράσεων μέγιστου μήκους είναι το *NPTool* [98]. Το εργαλείο αυτό βασίζεται σε ένα χειροποίητο λεξικό και σε δύο αναλυτές πεπερασμένης κατάστασης (finite state parsers), έναν εχθρικό και ένα φιλικό προς τις ονοματικές φράσεις. Ο συνδυασμός αυτών των αναλυτών παράγει μία λίστα από αποδεκτές ονοματικές φράσεις που μπορούν να χρησιμοποιηθούν σε εφαρμογές εξαγωγής ορολογίας. Το σύστημα αξιολογήθηκε σε ένα σώμα κειμένων 20.000 λέξεων που περιλαμβάνει κείμενα από

διάφορους τομείς επιτυγχάνοντας ανάκληση και ακρίβεια της τάξης του 98,5-100% και 95-98% αντίστοιχα. Ένας άλλος αποτελεσματικός *μερικός αναλυτής* (partial parser) που συνδυάζει ενισχυμένες ετικέτες μέρους-του-λόγου (part-of-speech tags), οι οποίες καλούνται *υπερετικέτες* (supertags), με έναν αναλυτή ελαφριάς εξάρτησης (lightweight dependency analyzer) προτείνεται από τον Srinivas [89]. Ο αναλυτής αυτός επιτυγχάνει ανάκληση και ακρίβεια για τις ονοματικές φράσεις της τάξης του 93% και 91,8% αντίστοιχα.

Τέλος, το *FASTUS* [42] είναι ένα σύστημα για την εξαγωγή πληροφορίας από Αγγλικά κείμενα το οποίο δουλεύει σαν ένα συνεχές μη-ντετερμινιστικό αυτόματο (cascaded non-deterministic automaton). Αυτό το σύστημα αρχικά προσπαθεί να αναγνωρίσει τις βασικές ονοματικές και ρηματικές φράσεις, βασισμένο σε μια γραμματική πεπερασμένης κατάστασης και στην συνέχεια αναγνωρίζει πιο περίπλοκες φράσεις συνδυάζοντας τις βασικές. Οι άγνωστες λέξεις αγνοούνται από την περαιτέρω ανάλυση εκτός και αν βρίσκονται σε περιβάλλον που υποδηλώνει ότι μπορεί να είναι κύρια ονόματα. Η σύγκριση του *FASTUS* με πιο εξελιγμένα και περίπλοκα συστήματα δείχνει ότι είναι δυνατή μια αρκετά ικανοποιητική ανάλυση κειμένου με χρήση σχετικά απλών τεχνικών αλλά και η επίτευξη πολύ καλών αποτελεσμάτων ανάλυσης πολύ γρήγορα [106].

3.3 Εκτίμηση Μορφολογικής Πληροφορίας

Με βάση το λεξικό λέξεων-κλειδιών και το λεξικό καταλήξεων σε κάθε λέξη της περιόδου αντιστοιχίζεται ένα σύνολο από μορφολογικές περιγραφές. Πιο αναλυτικά, το λεξικό λέξεων-κλειδιών περιέχει 432 λέξεις που περιλαμβάνουν άρθρα, μόρια, προθέσεις, αντωνυμίες, αριθμητικά και μερικά ειδικά επιρρήματα (όπως τα πάνω, κάτω κτλ.). Οι καταχωρήσεις σε αυτό το λεξικό είναι της ακόλουθης μορφής (συνάρτηση Prolog):

keyword(ΛΕΞΗ, ΕΚΚΙΝΗΣΗ, ΠΕΡΙΓΡΑΦΕΣ)

όπου η *ΛΕΞΗ* είναι η λέξη-κλειδί, η *ΕΚΚΙΝΗΣΗ* δηλώνει αν η λέξη αυτή είναι συνήθως η πρώτη λέξη μιας ονοματικής, προθεματικής ή ρηματικής φράσης, και οι *ΠΕΡΙΓΡΑΦΕΣ* είναι μία λίστα από μορφολογικές περιγραφές. Πιο κάτω δίνονται ως παράδειγμα οι καταχωρήσεις για τις λέξεις *θα, των, την* και *αύριο*:

keyword("θα", "ΕΚΚΙΝΗΣΗ_ΡΦ", [μόριο])

keyword("των", "ΕΚΚΙΝΗΣΗ_ΟΦ", [άρθρο("ΑΡΣ-ΘΗΛ-ΟΥΔ, ΠΛΘ, ΓΕΝ")])

keyword("την", "ΕΚΚΙΝΗΣΗ_ΟΦ", [άρθρο("ΘΗΛ", "ΕΝΚ", "ΑΙΤ"),
πρ_αντ("ΘΗΛ", "ΕΝΚ", "ΑΙΤ")])

keyword("αύριο", "", [επίρρημα])

Να σημειωθεί ότι οι μορφολογικές πληροφορίες σχετίζονται άμεσα με το μέρος-του-λόγου της λέξης. Έτσι το άρθρο *των* μπορεί να είναι αρσενικού ή θηλυκού ή ουδέτερου γένους, πληθυντικού αριθμού και πτώσης γενικής. Επίσης, η λέξη *την* μπορεί να είναι είτε άρθρο είτε προσωπική αντωνυμία θηλυκού γένους, ενικού αριθμού και πτώσης αιτιατικής. Τέλος, οι λέξεις *θα, των* και *την* δηλώνουν έναρξη ρηματικής και ονοματικής φράσης αντίστοιχα, ενώ η λέξη *αύριο* όχι.

Το λεξικό λέξεων-κλειδιών κατασκευάστηκε χειρονακτικά και δε βασίστηκε σε κάποιο προϋπάρχον λεξικό. Πρέπει πάντως να σημειωθεί ότι οποιοδήποτε μεγάλο λεξικό γενικού σκοπού που καλύπτει λέξεις κλειστού τύπου μπορεί να χρησιμοποιηθεί για την εξαγωγή ενός λεξικού λέξεων-κλειδιών.

Για όποια λέξη της περιόδου δε βρεθεί αντίστοιχη καταχώρηση στο λεξικό λέξεων-κλειδιών, γίνεται μία εκτίμηση της μορφολογικής της πληροφορίας σύμφωνα με την κατάληξή της. Το λεξικό καταλήξεων περιέχει 282 κοινές καταλήξεις των Νέων Ελληνικών. Η συντριπτική πλειοψηφία αυτών των καταλήξεων αποκομίστηκε από μία ήδη υπάρχουσα μορφολογική περιγραφή δύο επιπέδων της Νέας Ελληνικής γλώσσας [81] που βασίζεται στο φορμαλισμό του *PC-KIMMO* [5]. Οι καταχωρήσεις σ' αυτό το λεξικό είναι της ακόλουθης μορφής (συνάρτηση Prolog):

ending(ΚΑΤΑΛΗΞΗ, ΠΕΡΙΓΡΑΦΕΣ)

όπου οι *ΠΕΡΙΓΡΑΦΕΣ* είναι μία λίστα μορφολογικών περιγραφών που καταχωρούνται σε μία λέξη της οποίας η κατάληξη ταιριάζει με την *ΚΑΤΑΛΗΞΗ*. Για κάθε λέξη επιλέγεται η κατάληξη με το μεγαλύτερο μήκος. Πιο κάτω δίνονται μερικά παραδείγματα καταχωρήσεων στο λεξικό καταλήξεων.

ending("ιμους", [επίθετο("ΑΡΣ","ΠΛΘ","ΑΙΤ")])

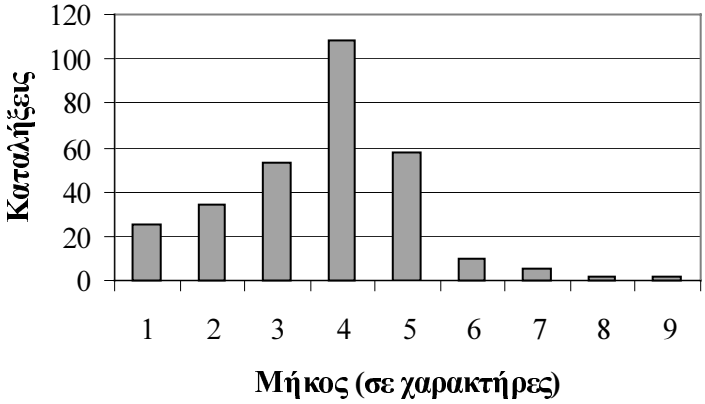
ending("άκι", [ουσιαστικό("ΟΥΔ","ΕΝΚ","ΟΝΜ-ΑΙΤ")])

ending("άει", [ρήμα])

ending("ας", [ουσιαστικό("ΑΡΣ","ΕΝΚ","ΟΝΜ"),
επίθετο("ΑΡΣ","ΕΝΚ","ΟΝΜ"),
ουσιαστικό("ΘΗΛ","ΕΝΚ","ΓΕΝ"),
επίθετο("ΘΗΛ","ΕΝΚ","ΓΕΝ")])

ending("είς", [ρήμα,
ουσιαστικό("ΑΡΣ-ΘΗΛ","ΠΛΘ","ΟΝΜ-ΑΙΤ"),
επίθετο("ΑΡΣ-ΘΗΛ","ΠΛΘ","ΟΝΜ-ΑΙΤ")])

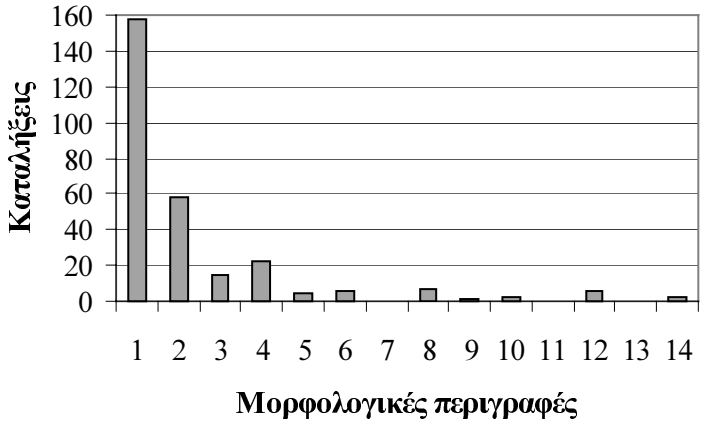
Οι καταλήξεις *–ιμους, –άκι,* και *–άει* αντιστοιχούν σε ένα μόνο μέρος-του-λόγου (επίθετο, ουσιαστικό και ρήμα αντίστοιχα). Αντίθετα η κατάληξη *–ας* μπορεί να είναι ουσιαστικό ή επίθετο, είτε αρσενικού γένους και πτώσης ονομαστικής είτε θηλυκού γένους και πτώσης γενικής (π.χ. *ο πατέρας*, *της μητέρας*). Τέλος, η κατάληξη *–είς* μπορεί να δηλώνει ρήμα (π.χ. *καταστραφείς*) ή ουσιαστικό (π.χ. *αμφορείς*) ή επίθετο (π.χ. *ειλικρινείς*).



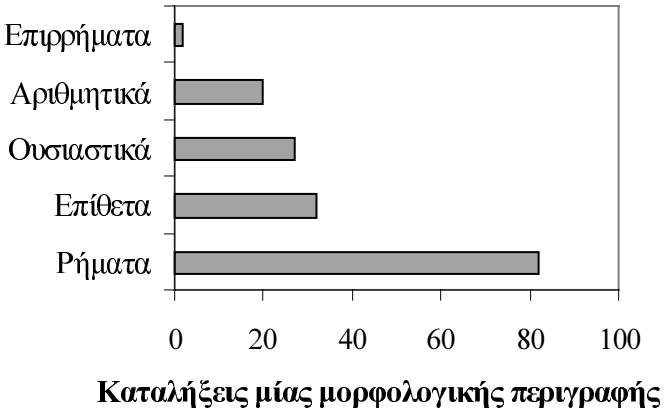
Σχήμα 3.2. Κατανομή των καταλήξεων συναρτήσει του μήκους τους.

Το σχήμα 3.2 δείχνει την κατανομή των καταλήξεων συναρτήσει του μήκους τους σε χαρακτήρες. Όπως φαίνεται, η κατανομή αυτή προσεγγίζει την κανονική. Στο σχήμα 3.3 φαίνεται ο αριθμός των καταλήξεων σε σχέση με τον αριθμό των μορφολογικών περιγραφών που καταχωρούν στις λέξεις. Πάνω από το 56% (158 στις 282) των καταλήξεων καταχωρούν μόνο μία μορφολογική περιγραφή στις λέξεις που ταιριάζουν. Παρατηρούμε ότι οι καταλήξεις που αντιστοιχούν σε άρτιο αριθμό μορφολογικών περιγραφών υπερτερούν εκείνων που αντιστοιχούν σε περιττό αριθμό

μορφολογικών περιγραφών. Αυτό οφείλεται στην μεγάλη ομοιότητα που παρουσιάζουν οι καταλήξεις των ουσιαστικών με τις καταλήξεις των επιθέτων. Περίπου το 52% (82 στις 158) αυτών των καταλήξεων (δηλ. των καταλήξεων που καταχωρούν μόνο μία μορφολογική περιγραφή) αφορούν καταλήξεις ρημάτων όπως φαίνεται και στο σχήμα 3.4. Επιβεβαιώνεται έτσι η αρχική εμπειρική παρατήρηση για τις χαρακτηριστικές καταλήξεις των ρημάτων.



Σχήμα 3.3. Κατανομή των καταλήξεων συναρτήσει



των μορφολογικών περιγραφών που υποδηλώνουν.

Σχήμα 3.4. Καταλήξεις μιας μορφολογικής περιγραφής ανά μέρος-του-λόγου.

Σε περίπτωση που κάποια λέξη της περιόδου δεν ταιριάζει με καμία από τις καταχωρήσεις του λεξικού καταλήξεων, δεν καταχωρείται σε αυτήν καμία μορφολογική περιγραφή και μαρκάρεται ως **ειδική λέξη**. Συνήθως αυτό συμβαίνει για ξένες λέξεις (π.χ. *απεριτίφ*), κύρια ονόματα (π.χ. *Σμιθ*) και ελληνικές λέξεις με

αρχαΐζουσα κλίση ή λέξεις της καθαρεύουσας (π.χ. *ήπαρ*). Όμως, τέτοιες λέξεις λαμβάνουν κανονικά μέρος στην περαιτέρω ανάλυση. Επιπλέον, εφόσον ο πρώτος χαρακτήρας μιας λέξης που δεν ταιριάζει με καμία κατάληξη είναι κεφαλαίος, τότε η λέξη αυτή μαρκάρεται ως πιθανό κύριο όνομα. Τέλος, εφόσον εντοπιστεί κάποιο εμπρόθετο άρθρο, χωρίζεται στην πρόθεση και το άρθρο που το συνθέτουν (π.χ. *στον* = *σε* + *τον*).

3.4 Ανάλυση Πολλαπλού Περάσματος

Ο στόχος του προτεινόμενου αναλυτή είναι η αναγνώριση των ορίων των κυρίων φράσεων που περιλαμβάνονται σε κάθε περίοδο χωρίς την ανάλυση της εσωτερικής τους δομής ή της λειτουργίας τους στην περίοδο. Παρ' όλα αυτά, πραγματοποιείται και κάποια απλή μορφολογική αποσαφήνιση εφαρμόζοντας *επιλεκτικούς περιορισμούς* (selectional restrictions) [3] στα μέλη της ίδιας φράσης (π.χ. συμφωνία αριθμού, γένους, πτώσης μέσα σε μία ονοματική φράση).

Πιο συγκεκριμένα, τα είδη των φράσεων που ανιχνεύονται είναι: ονοματικές φράσεις (ΟΦ), προθετικές φράσεις (ΠΦ), ρηματικές φράσεις (ΡΦ) και επιρρηματικές φράσεις (ΕΦ). Επιπλέον, δύο φράσεις μπορεί να συνδέονται μέσω μιας ακολουθίας συνδέσμων, τις οποίες καλούμε συνδετικές φράσεις (ΣΦ).

Ο προσδιορισμός των ορίων των φράσεων γίνεται μέσω ανάλυσης πολλαπλού περάσματος. Κάθε πέρασμα αναλύει ένα τμήμα της περιόδου, στηριζόμενο στα αποτελέσματα των προηγούμενων περασμάτων, ενώ το τμήμα της περιόδου που απομένει αναλύεται από τα επόμενα περάσματα. Σε γενικές γραμμές, η φιλοσοφία της σχεδίασης των περασμάτων ανάλυσης ήταν τα πρώτα περάσματα να ασχολούνται με τις απλές περιπτώσεις, που ανιχνεύονται πολύ εύκολα, και τα τελευταία περάσματα να ασχολούνται με τις πιο περίπλοκες και σύνθετες περιπτώσεις. Έτσι, τα τελευταία περάσματα ανάλυσης είναι λιγότερο ακριβή από τα πρώτα εξαιτίας του μεγαλύτερου βαθμού δυσκολίας και ασάφειας που έχουν να αντιμετωπίσουν. Σε περίπτωση που και το τελευταίο πέρασμα ανάλυσης δεν καταφέρει να αναλύσει κάποιο κομμάτι της περιόδου, τότε αυτό το κομμάτι μένει χωρίς ανάλυση (δηλ. χωρίς να περικλείεται στα όρια κάποιας φράσης).

Η προσέγγισή μας πραγματοποιεί πέντε περάσματα ανάλυσης. Στην συνέχεια περιγράφεται η λειτουργία του κάθε περάσματος.

Πέρασμα 1: Ανιχνεύονται απλές ΟΦ, ΠΦ και ΡΦ βάσει των λέξεων-κλειδιών που δηλώνουν εκκίνηση φράσης. Η αναγνώριση των ορίων αυτών των φράσεων γίνεται με χρήση απλών εμπειρικών κανόνων. Για παράδειγμα, για την ανίχνευση των απλών ΟΦ εφαρμόζεται ο ακόλουθος κανόνας:

> *Μία ΟΦ περιλαμβάνει ουσιαστικά, επίθετα, επιρρήματα, συνδέσμους, αριθμητικά και ειδικές λέξεις που ακολουθούν μία λέξη εκκίνησης ΟΦ (συνήθως άρθρα ή αντωνυμίες). Επιπλέον, τα ουσιαστικά και τα επίθετα πρέπει να ταιριάζουν με τη λέξη εκκίνησης όσον αφορά το γένος, τον αριθμό και την πτώση. Η τελευταία λέξη μιας ΟΦ δεν μπορεί να είναι επίρρημα ή σύνδεσμος.*

Αυτό το πέρασμα ανάλυσης μπορεί να ανιχνεύσει σωστά φράσεις όπως οι ακόλουθες:

Φράση	Είδος
την αναγκαία γνώση και ευαισθησία	ΟΦ
ο 20ός αιώνας	ΟΦ
της κ. Ελένης Παπαδοπούλου	ΟΦ
οι αραιοκατοικημένες και γεωλογικά σχεδόν απομονωμένες περιοχές	ΟΦ
με μεγάλη δύναμη	ΠΦ
με συγκριτικά ελάχιστο κόστος	ΠΦ
δεν έχουν περάσει	ΡΦ
αναρωτήθηκα	ΡΦ
να δώσεις	ΡΦ

Πέρασμα 2: Ανιχνεύονται ΟΦ γενικής πτώσης χωρίς άρθρο που συνήθως ακολουθούν άλλες ΟΦ, καθώς και απλές ΠΦ. Για παράδειγμα οι ακόλουθες φράσεις ανιχνεύονται σωστά (Να σημειωθεί ότι οι φράσεις που μαρκάρονται μέσα σε αγκύλες έχουν ανιχνευτεί στο προηγούμενο πέρασμα ανάλυσης):

Φράση	Είδος
ΟΦ[*τη χρήση*] *δορυφορικών συστημάτων*	ΟΦ
ΟΦ[*τα μέσα*] *μαζικής ενημέρωσης*	ΟΦ
με ΟΦ[*τον κ. Μπιλ Σμιθ*]	ΠΦ
από ΟΦ[*το περιπολικό*]	ΠΦ
για ΟΦ[*το μηχανικό*]	ΠΦ

Πέρασμα 3: Οι αντωνυμίες που απομένουν εκτός ορίων φράσεων είτε ενώνονται με γειτονικές ΟΦ είτε σχηματίζουν από μόνες τους καινούργιες και ανιχνεύονται κατηγορούμενα των ρημάτων. Για παράδειγμα, δίνονται οι ακόλουθες περιπτώσεις (υπενθυμίζεται ότι οι φράσεις που εσωκλείονται σε αγκύλες έχουν ανιχνευτεί στα προηγούμενα περάσματα ανάλυσης):

Φράση	Είδος
όλα ΟΦ[*τα κλειδιά*]	ΟΦ
ΟΦ[*η μητέρα*] *μας*	ΟΦ
αυτό	ΡΦ
ΡΦ[*είναι*] *σημαντικά αλλά περίπλοκα*	ΡΦ

Πέρασμα 4: Ανιχνεύονται ΕΦ, ΣΦ καθώς και ΟΦ χωρίς λέξεις-κλειδιά που να υποδηλώνουν εκκίνηση ΟΦ. Επίσης ΠΦ που περιέχουν περίπλοκες ΟΦ. Αυτό το πέρασμα ανάλυσης μπορεί να ανιχνεύσει σωστά τις ακόλουθες φράσεις:

Φράση	Είδος
σχεδόν τελείως	ΕΦ
αν και	ΣΦ
εθνικό έργο	ΟΦ
ζωή και ελπίδα	ΟΦ
σε ΟΦ[*όλες τις περιπτώσεις*]	ΠΦ

Πέρασμα 5: Σε αυτό το πέρασμα ανάλυσης λαμβάνει χώρα συνδυασμός των φράσεων που ανιχνεύτηκαν στα προηγούμενα περάσματα με στόχο το σχηματισμό

των μεγαλύτερων δυνατών φράσεων. Ακόμη ανιχνεύονται σύνθετες ΕΦ. Μερικές συνδυασμοί φράσεων που πραγματοποιούνται φαίνονται παρακάτω:

Φράση	Είδος
ΟΦ[*η ανάπτυξη*] ΟΦ[*της νέας τεχνολογίας*]	ΟΦ
ΠΦ[*στις ανάγκες*] ΟΦ[*των νέων ανθρώπων*]	ΠΦ
ΡΦ[*τρέχει*] ΡΦ[*να σωθεί*]	ΡΦ
ΕΦ[*πολύ*] *προσεκτικά*	ΕΦ
ΕΦ[*εντελώς*] *αντικειμενικά*	ΕΦ

Πρέπει να σημειωθεί ότι τα σημεία στίξης παίρνουν μέρος στην διαδικασία ανάλυσης και μεταχειρίζονται ως ειδικά σύμβολα. Για παράδειγμα, το κόμμα μεταχειρίζεται ως ένας σύνδεσμος. Το πλήρες σύνολο των κανόνων ανίχνευσης φράσεων δίνεται στο Παράρτημα Α.

Για να γίνει πιο κατανοητή η διαδικασία ανάλυσης μέσω πολλαπλού περάσματος παραθέτουμε ένα παράδειγμα ανάλυσης ενός δείγματος κειμένου στο σχήμα 3.5. Οι φράσεις που ανιχνεύονται σε κάθε πέρασμα φαίνονται με έντονα γράμματα. Το σύμβολο # υποδεικνύει το τέλος μιας περιόδου. Παρατηρούμε ότι το σύστημα απέτυχε να αναλύσει τη λέξη *Σύμφωνα* αφού η κατάληξή της (*-α*) ήταν πολύ πιθανό να δηλώνει τόσο επίρρημα όσο και ουσιαστικό και τα συμφραζόμενα δεν βοήθησαν να επιλυθεί αυτή η ασάφεια.

3.5 Αξιολόγηση

Ο ανιχνευτής ορίων φράσεων είναι πλήρως υλοποιημένος στη γλώσσα *Visual Prolog v. 5.0*. Η απόδοσή του ελέγχθηκε σε ένα σώμα 200.829 λέξεων αποτελούμενο από κείμενα της εφημερίδας *Το Βήμα*. Να σημειωθεί ότι το σώμα αυτό είναι το ίδιο με το σώμα ελέγχου του ανιχνευτή ορίων περιόδων (βλ. § 2.4.1).

Κείμενο:
Το άλλο, τραγικό θύμα αυτής της ιστορίας, η 25άχρονη Αμαλία Παπαδοπούλου, συνεχίζει να δίνει από την εντατική μονάδα του Ερυθρού Σταυρού, τον αγώνα της να κρατηθεί στη ζωή. Σύμφωνα με το σημερινό ιατρικό ανακοινωθέν, οι θεράποντες ιατροί, διαπιστώνουν μικρή βελτίωση της κατάστασης, η οποία ωστόσο παραμένει ιδιαιτέρως κρίσιμη.

Μετατροπή σε λέξεις-κλειδιά και κοινές καταλήξεις:
το -ο , -ικό -μα αυτής της -ίας , η -η -ία -ου , -ει να -ει από την -ική -α του -ού -ού , τον -α της να -εί σε τη -ή . -α με το -ό -ικό -έν , οι -οντες -οί , -ουν -ή -η της -ης , η οποία ωστόσο -ει -ως –ιμη.

Πέρασμα 1:
ΟΦ[Το άλλο , τραγικό θύμα] αυτής **ΟΦ[της ιστορίας]** , **ΟΦ[η 25άχρονη Αμαλία]** Παπαδοπούλου , **ΡΦ[συνεχίζει] ΡΦ[να δίνει]** από **ΟΦ[την εντατική μονάδα] ΟΦ[του Ερυθρού Σταυρού]** , **ΟΦ[τον αγώνα]** της **ΡΦ[να κρατηθεί]** σε **ΟΦ[τη ζωή]** . #
Σύμφωνα με **ΟΦ[το σημερινό ιατρικό ανακοινωθέν]** , **ΟΦ[οι θεράποντες ιατροί]** , **ΡΦ[διαπιστώνουν]** μικρή βελτίωση **ΟΦ[της κατάστασης]** , **ΟΦ[η οποία]** ωστόσο **ΡΦ[παραμένει]** ιδιαιτέρως κρίσιμη . #

Πέρασμα 2:
ΟΦ[Το άλλο , τραγικό θύμα] αυτής ΟΦ[της ιστορίας] , ΟΦ[η 25άχρονη Αμαλία] **ΟΦ[Παπαδοπούλου]** , ΡΦ[συνεχίζει] ΡΦ[να δίνει] **ΠΦ[από την εντατική μονάδα]** ΟΦ[του Ερυθρού Σταυρού] , ΟΦ[τον αγώνα] της ΡΦ[να κρατηθεί] **ΠΦ[στη ζωή]** . #
Σύμφωνα **ΠΦ[με το σημερινό ιατρικό ανακοινωθέν]** , ΟΦ[οι θεράποντες ιατροί] , ΡΦ[διαπιστώνουν] μικρή βελτίωση ΟΦ[της κατάστασης] , ΟΦ[η οποία] ωστόσο ΡΦ[παραμένει] ιδιαιτέρως κρίσιμη . #

Πέρασμα 3:
ΟΦ[Το άλλο , τραγικό θύμα] **ΟΦ[αυτής της ιστορίας]** , ΟΦ[η 25άχρονη Αμαλία] ΟΦ[Παπαδοπούλου] , ΡΦ[συνεχίζει] ΡΦ[να δίνει] ΠΦ[από την εντατική μονάδα] ΟΦ[του Ερυθρού Σταυρού] , **ΟΦ[τον αγώνα της]** ΡΦ[να κρατηθεί] ΠΦ[στη ζωή] . #
Σύμφωνα ΠΦ[με το σημερινό ιατρικό ανακοινωθέν] , ΟΦ[οι θεράποντες ιατροί] , ΡΦ[διαπιστώνουν] μικρή βελτίωση ΟΦ[της κατάστασης] , ΟΦ[η οποία] ωστόσο **ΡΦ[παραμένει ιδιαιτέρως κρίσιμη]** . #

Πέρασμα 4:
ΟΦ[Το άλλο , τραγικό θύμα] ΟΦ[αυτής της ιστορίας] , ΟΦ[η 25άχρονη Αμαλία] ΟΦ[Παπαδοπούλου] , ΡΦ[συνεχίζει] ΡΦ[να δίνει] ΠΦ[από την εντατική μονάδα] ΟΦ[του Ερυθρού Σταυρού] , ΟΦ[τον αγώνα της] ΡΦ[να κρατηθεί] ΠΦ[στη ζωή] . #
Σύμφωνα ΠΦ[με το σημερινό ιατρικό ανακοινωθέν] , ΟΦ[οι θεράποντες ιατροί] , ΡΦ[διαπιστώνουν] **ΟΦ[μικρή βελτίωση]** ΟΦ[της κατάστασης] , ΟΦ[η οποία] **ΣΦ[ωστόσο]** ΡΦ[παραμένει ιδιαιτέρως κρίσιμη] . #

Πέρασμα 5:
ΟΦ[Το άλλο , τραγικό θύμα αυτής της ιστορίας] , **ΟΦ[η 25άχρονη Αμαλία Παπαδοπούλου]** , **ΡΦ[συνεχίζει να δίνει] ΠΦ[από την εντατική μονάδα του Ερυθρού Σταυρού]** , ΟΦ[τον αγώνα της] ΡΦ[να κρατηθεί] ΠΦ[στη ζωή] . #
Σύμφωνα ΠΦ[με το σημερινό ιατρικό ανακοινωθέν] , ΟΦ[οι θεράποντες ιατροί] , ΡΦ[διαπιστώνουν] **ΟΦ[μικρή βελτίωση της κατάστασης]** ,ΟΦ[η οποία] ΣΦ[ωστόσο] ΡΦ[παραμένει ιδιαιτέρως κρίσιμη] . #

Σχήμα 3.5. Ανάλυση δείγματος κειμένου μέσω πολλαπλού περάσματος.

Το σώμα ελέγχου αναλύθηκε από τον ανιχνευτή ορίων φράσεων και στην συνέχεια ένας κριτής αξιολόγησε χειρονακτικά την έξοδό του. Τα αποτελέσματα αυτής της αξιολόγησης φαίνονται στον πίνακα 3.1. Για κάθε είδος φράσης δίνονται αποτελέσματα για την *ανάκληση* (recall) και την *ακρίβεια* (precision). Τα δύο αυτά μεγέθη έχουν καθιερωθεί για τον έλεγχο συστημάτων εξαγωγής πληροφορίας [106] και στην περίπτωση της ανίχνευσης ορίων φράσεων ορίζονται ως εξής:

Ανάκληση = *αριθμός των σωστά προσδιορισμένων φράσεων από το σύστημα προς τον αριθμό όλων των φράσεων που περιέχονται στο κείμενο*

Ακρίβεια = *αριθμός των σωστά προσδιορισμένων φράσεων από το σύστημα προς τον αριθμό των συνολικών φράσεων που ανίχνευσε το σύστημα*

Η χαμηλή τιμή της ανάκλησης των ΕΦ οφείλεται κυρίως στην ομοιότητα των καταλήξεων της πλειοψηφίας των επιρρημάτων των Νέων Ελληνικών με τις καταλήξεις των επιθέτων πληθυντικού αριθμού και ουδέτερου γένους (π.χ. *πέρασα καταπληκτικά - είναι καταπληκτικά παιδιά*). Η χαμηλή τιμή της ακρίβειας των ΟΦ οφείλεται, ως επί το πλείστον, στην ανάλυση που πραγματοποιείται στο τέταρτο πέρασμα όπου αναγνωρίζονται οι ΟΦ που δεν εισάγονται με κάποιο άρθρο ή αντωνυμία. Επίσης, οι ξένες λέξεις των οποίων οι καταλήξεις μοιάζουν με λέξεις των Νέων Ελληνικών, μετά τη μεταγραφή τους με ελληνικούς χαρακτήρες (π.χ. *down - ντάουν*) δημιούργησαν διάφορα λάθη.

Λέξεις			200.829
Περίοδοι			8.736
Χρονικό κόστος ανάλυσης (λέξεις/δευτ.)			514
Κείμενο που δεν αναλύθηκε (σε λέξεις)			7.323
Φράση	**Σύνολο**	**Ανάκληση (%)**	**Ακρίβεια (%)**
ΟΦ	31.339	91,18	88,72
ΠΦ	17.947	93,35	99,36
ΡΦ	25.538	91,32	98,19
ΕΦ	9.651	72,47	96,27
Σύνολο	**84.475**	**89,55**	**94,45**

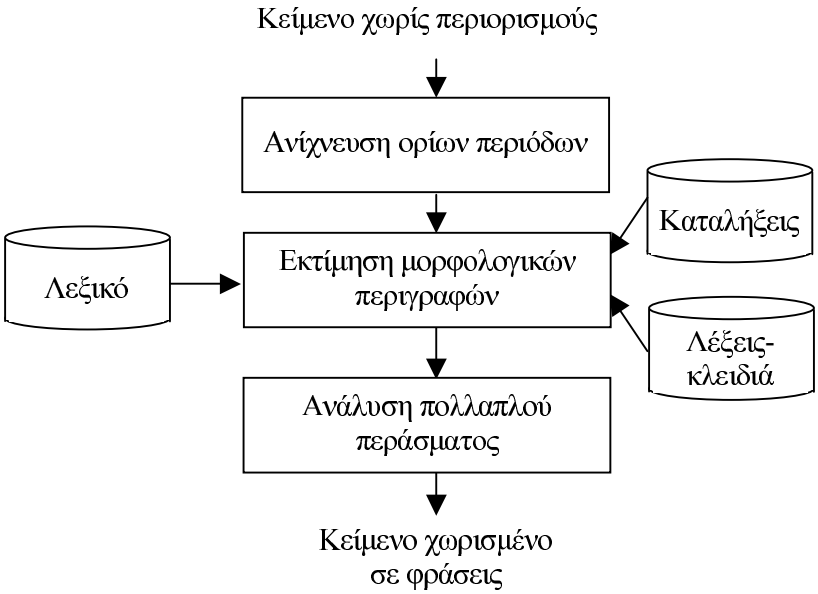
Πίνακας 3.1. Η απόδοση του ανιχνευτή ορίων φράσεων.

Για την ανάλυση του σώματος ελέγχου χρησιμοποιήθηκε ένας Pentium στα 133 MHz. Πρέπει να επισημανθεί ότι η διαδικασία ανάλυσης δεν ήταν η μοναδική που απασχολούσε τον υπολογιστή κατά τη μέτρηση του χρονικού κόστους ανάλυσης.

Επιπλέον, το σύστημα απέτυχε να αναλύσει το 3,6% των συνολικών λέξεων του σώματος (7.323 από τις 200.829 λέξεις) κυρίως λόγω περίπλοκης σύνταξης. Ένα παράδειγμα αναλυμένου κειμένου από το προτεινόμενο σύστημα δίνεται στο Παράρτημα Β.

3.6 Προσέγγιση Βασιζόμενη σε Λεξικό

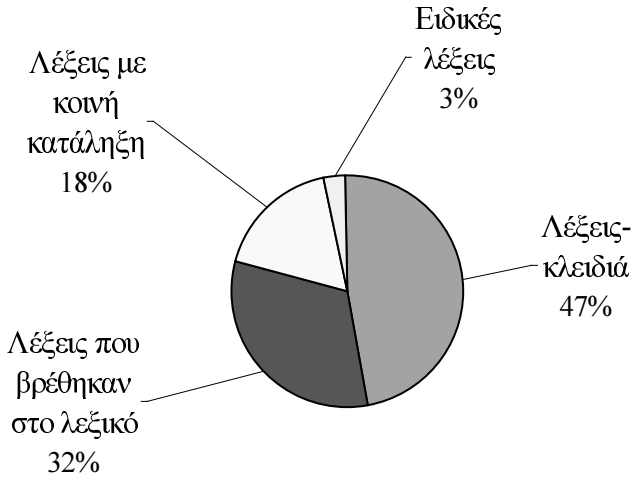
Για να μπορέσουμε να συγκρίνουμε τα αποτελέσματα του προτεινόμενου ανιχνευτή ορίων φράσεων αναπτύξαμε ένα σύστημα που χρησιμοποιεί πιο περίπλοκους πόρους. Πιο συγκεκριμένα, ακολουθήσαμε την παραδοσιακή προσέγγιση και ενσωματώσαμε ένα ογκώδες λεξικό λημμάτων στο σύστημα που περιγράψαμε στα προηγούμενα τμήματα. Το λεξικό αυτό αναπτύχθηκε στα πλαίσια ενός μορφολογικού αναλυτή δύο επιπέδων της Νέας Ελληνικής γλώσσας [81] και περιέχει περίπου 30.000 λήμματα που καλύπτουν ουσιαστικά, επίθετα, ρήματα και επιρρήματα. Η διαδικασία εκτίμησης της μορφολογικής πληροφορίας βάσει της κατάληξής της εφαρμόστηκε σε αυτήν την προσέγγιση μόνο για τις λέξεις που δεν καλυπτόταν από το λεξικό. Αντίθετα, η διαδικασία ανάλυσης πολλαπλού περάσματος παρέμεινε η ίδια. Η δομή της προσέγγισης βάσει λεξικού φαίνεται στο σχήμα 3.6.



Σχήμα 3.6. Δομή του συστήματος ανίχνευσης ορίων φράσεων βάσει λεξικού.

Λέξεις			200.829
Περίοδοι			8.736
Χρονικό κόστος ανάλυσης (λέξεις/δευτ.)			238
Κείμενο που δεν αναλύθηκε (σε λέξεις)			5.833
Φράση	**Σύνολο**	**Ανάκληση (%)**	**Ακρίβεια (%)**
ΟΦ	31.339	94,46	85,58
ΠΦ	17.947	93,96	99,12
ΡΦ	25.538	93,63	97,57
ΕΦ	9.651	85,28	96,60
Σύνολο	**84.475**	**93,05**	**92,35**

Πίνακας 3.2. Η απόδοση του ανιχνευτή ορίων φράσεων με χρήση λεξικού.



Σχήμα 3.7. Μορφολογική ανάλυση του σώματος ελέγχου από την προσέγγιση με χρήση λεξικού.

Το σώμα ελέγχου 200.829 λέξεων αναλύθηκε από αυτόν τον ανιχνευτή και η απόδοσή του φαίνεται στον πίνακα 3.2. Σε σύγκριση με την προσέγγιση ελάχιστων πόρων η ανάκληση είναι βελτιωμένη ειδικά στην περίπτωση των ΟΦ και των ΕΦ. Απ' την άλλη, η ακρίβεια είναι πιο χαμηλή με τη χρήση λεξικού. Αυτό οφείλεται κυρίως στο γεγονός ότι το λεξικό σε μερικές περιπτώσεις δεν παρείχε όλες τις δυνατές μορφολογικές περιγραφές για κάποιες λέξεις. Για παράδειγμα, η λέξη *μετρήσεις* μπορεί να είναι ουσιαστικό καθώς και ρήμα. Όμως, η μορφολογική ανάλυση που προήλθε βάσει του λεξικού παρείχε μόνο τη δεύτερη μορφή. Τέτοιου είδους λάθη επηρεάζουν και την ανάκληση των ΟΦ και την ακρίβεια των ΡΦ.

Ο ανιχνευτής ορίων φράσεων με χρήση λεξικού απέτυχε να αναλύσει 5.833 λέξεις του σώματος ελέγχου, δηλ. 20% λιγότερες λέξεις από την προσέγγιση με χρήση

ελάχιστων πόρων. Όμως, το χρονικό κόστος της ανάλυσης με χρήση λεξικού είναι περίπου κατά 50% μεγαλύτερο του αντίστοιχου χρόνου της προσέγγισης ελάχιστων πόρων.

Επίσης, είναι πολύ ενδιαφέρον να δούμε για πόσες λέξεις του σώματος ελέγχου μπόρεσε το λεξικό να δώσει χρήσιμη πληροφορία. Όπως φαίνεται στο σχήμα 3.7 το λεξικό παρείχε μορφολογική πληροφορία για το 32% των συνολικών λέξεων του σώματος ελέγχου. Πρέπει να σημειωθεί ότι το λεξικό παρείχε κατά μέσο όρο 1,9 μορφολογικές περιγραφές ανά λέξη. Η εφαρμογή της διαδικασίας εκτίμησης της μορφολογικής πληροφορίας, σύμφωνα με την κατάληξη, στις ίδιες αυτές λέξεις (δηλ. το 32% του σώματος ελέγχου που αναλύθηκαν από το λεξικό) έδειξε ότι ο αντίστοιχος μέσος όρος είναι 3,6, δηλ. περίπου διπλάσιος. Το 47% των συνολικών λέξεων του σώματος ελέγχου αναλύθηκε βάσει του λεξικού λέξεων-κλειδιών ενώ το 18% των λέξεων αναλύθηκε με χρήση του λεξικού καταλήξεων. Τέλος, το 3% των λέξεων χαρακτηρίστηκαν ως ειδικές λέξεις αφού δεν αναλύθηκαν από το λεξικό λημμάτων και δε βρέθηκε κάποια κατάληξη στο λεξικό καταλήξεων που να ταιριάζει με αυτές.

3.7 Περίληψη - Συμπεράσματα

Στο κεφάλαιο αυτό περιγράψαμε έναν ανιχνευτή ορίων φράσεων για κείμενα χωρίς περιορισμούς της Νέας Ελληνικής γλώσσας. Η προσέγγισή μας διαφέρει από τις σύγχρονες μεθόδους ανάλυσης αφού δεν βασίζεται ούτε στην έξοδο κάποιου σχολιαστή μέρους-του-λόγου ούτε σε λεξικά χιλιάδων λημμάτων και περίπλοκες γραμματικές. Αντίθετα, οι χρησιμοποιούμενοι πόροι είναι δύο πολύ μικρού μεγέθους λεξικά, η πληροφορία των οποίων μπορεί να ανανεωθεί πολύ σύντομα χωρίς περίπλοκες διαδικασίες. Τα ποσοστά ανάκλησης και ακρίβειας σε ένα σώμα κειμένων περίπου 200.000 λέξεων κρίνονται πολύ ικανοποιητικά και αποδεικνύεται ότι είναι δυνατή η επίτευξη αρκετά καλών αποτελεσμάτων με χρήση σχετικά απλών τεχνικών.

Η διαδικασία εύρεσης της πιο πιθανής μορφολογικής πληροφορίας της κάθε λέξης σύμφωνα με την κατάληξή της αντικαθιστά πλήρως τα ογκώδη λεξικά λημμάτων σε αντίθεση με τη φιλοσοφία των περισσότερων συστημάτων που έχουν αναπτυχθεί έως σήμερα. Επιπλέον, η σύγκριση του συστήματός μας με μία παρόμοια προσέγγιση που

βασίζεται σε λεξικό λημμάτων έδειξε ότι η διαφορά στην απόδοση των δύο συστημάτων είναι πολύ μικρή. Μάλιστα το σύστημα ελάχιστων πόρων επιτυγχάνει υψηλότερα ποσοστά ακρίβειας. Η πιο σημαντική διαφορά τους όμως είναι στο θέμα της χρονικής απόκρισης. Έτσι ο ανιχνευτής ορίων φράσεων βάσει ελάχιστων πόρων απαιτεί σχεδόν το μισό χρόνο απόκρισης του ανιχνευτή που βασίζεται σε λεξικό και αυτή η διαφορά είναι πολύ σημαντική ειδικά σε περιπτώσεις εφαρμογών που απαιτούν ακαριαία απόκριση (π.χ. εξαγωγή πληροφορίας).

Ο συνδυασμός των ανιχνευτών περιόδων και φράσεων αποτελεί ένα πολύ αξιόπιστο εργαλείο προεπεξεργασίας κειμένου για τη Νέα Ελληνική γλώσσα και μπορεί να χρησιμοποιηθεί σε πλειάδα εφαρμογών. Ενδεικτικά αναφέρουμε την ανάκτηση και την εξαγωγή πληροφορίας καθώς και την εξαγωγή ορολογίας. Ασφαλώς, για την επίτευξη των καλύτερων δυνατών αποτελεσμάτων πρέπει να γίνει προσαρμογή των συστημάτων αυτών στο είδος κειμένων που πρόκειται να αναλύσουν. Όσον αφορά τον ανιχνευτή φράσεων αυτό μπορεί να γίνει με την επιλογή των κατάλληλων περασμάτων ανάλυσης με στόχο την προσαρμογή του μεγέθους και της ακρίβειας των εξαγόμενων φράσεων. Για παράδειγμα, εφόσον τα πρώτα περάσματα ανάλυσης είναι πολύ πιο ακριβή από τα τελευταία, μία εφαρμογή που απαιτεί ελάχιστα λάθη στις εξαγόμενες φράσεις θα μπορούσε να βασιστεί αποκλειστικά σε αυτά και να αγνοήσει τα επόμενα. Βέβαια, αυτό θα είχε ως συνέπεια να παραμείνει χωρίς ανάλυση πολύ μεγαλύτερο τμήμα κειμένου. Ακόμη, οι κανόνες σχηματισμού φράσεων θα μπορούσαν να τροποποιηθούν με στόχο να καλύπτουν κάποιες ιδιαιτερότητες του σώματος κειμένου που πρόκειται να αναλυθεί.

Ο ανιχνευτής ορίων φράσεων σχεδιάστηκε με την προοπτική να εφαρμοστεί σε κείμενο χωρίς περιορισμούς. Στον σχεδιασμό του προσπαθήσαμε να εκμεταλλευτούμε στον μέγιστο βαθμό τις ιδιαιτερότητες της Νέας Ελληνικής γλώσσας. Παρ' όλα αυτά πιστεύουμε ότι παρόμοιες τεχνικές μπορούν να εφαρμοστούν με επιτυχία και σε άλλες γλώσσες που χαρακτηρίζονται από κοινές ιδιότητες (δηλ. μορφολογική πολυπλοκότητα, συχνή χρήση άρθρων, μορίων κτλ.). Πιο συγκεκριμένα, γλώσσες όπως τα Ιταλικά, τα Ισπανικά και τα Γερμανικά είναι περισσότερο πιθανό να επωφεληθούν από τις προτεινόμενες μεθόδους.

Ως μελλοντικές βελτιώσεις του συστήματος που παρουσιάστηκε προτείνουμε την αυτόματη εξαγωγή των κανόνων αναγνώρισης ορίων φράσεων που εφαρμόζονται στα

περάσματα ανάλυσης, μέσω της επεξεργασίας μεγάλων σωμάτων κειμένων. Μια τέτοια διαδικασία θα αυτοματοποιούσε και την εκπαίδευση του συστήματος για κάποιο συγκεκριμένο είδος κειμένων. Επίσης, πιστεύουμε ότι η αξιοπιστία του συστήματος θα βελτιωθεί κατά πολύ με την ενσωμάτωση τεχνικών που θα επεξεργάζονται με επιτυχία ειδικές πτυχές της χρησιμοποίησης των σημείων στίξης. Για παράδειγμα, τεχνικές που θα επιλύουν την αμφισημία που προκύπτει από τα περιεχόμενα μιας παρένθεσης. Τέλος, αρκετή έρευνα απομένει να γίνει για την αξιόπιστη προσάρτηση μιας ΠΦ στην κατάλληλη ΟΦ (PP-attachment). Για παράδειγμα, στην πρόταση *ο άνθρωπος με το γαρύφαλλο γέλασε*, οι φράσεις ΟΦ[*ο άνθρωπος*] και ΠΦ[*με το γαρύφαλλο*] πρέπει να ενωθούν σε μία ΟΦ.