Επιφανειακή Συντακτική Ανάλυση Ελεύθερου Κειμένου

Σωτήρης Μπούτσης[1,2], **Προκόπης Προκοπίδης**[1], **Βούλα Γιούλη**[1], **Στέλιος Πιπερίδης**[1,2]

[1]Ινστιτούτο Επεξεργασίας του Λόγου,
Αρτέμιδος 6 & Επιδαύρου, 151 25 Μαρούσι
{sboutsis, prokopis, voula, spip}@ilsp.gr

[2]Εθνικό Μετσόβιο Πολυτεχνείο

Abstract

In this paper we describe a method for the efficient parsing of real-life Greek texts at the surface syntactic level. A grammar consisting of non-recursive regular expressions depicting Greek phrase structure has been compiled into a cascade of finite state transducers used to recognize syntactic constituents. The implemented parser is used in applications where large scale text processing is involved, and fast, robust, and relatively accurate syntactic analysis is necessary. The parser has been evaluated against a ca 32000 word corpus of financial and news texts and achieved promising precision and recall scores.

1. Εισαγωγή

Η πρόοδος στην τεχνολογία συντακτικής ανάλυσης ανοίγει νέες δυνατότητες για την εφαρμογή της επεξεργασίας φυσικής γλώσσας σε διεργασίες όπου περιορισμοί ακρίβειας, αποδοτικότητας και ταχύτητας την έκαναν αδύνατη στο παρελθόν ή την περιόριζαν σε μικρούς όγκους κειμένων. Η κατασκευή εύρωστων συντακτικών αναλυτών οδηγεί στην δημιουργία νέων εφαρμογών και επιτρέπει στις υπάρχουσες να αξιοποιήσουν μεγάλους όγκους κειμενικών δεδομένων.

Στον τομέα της λεξικογραφίας γίνεται εκτεταμένη χρήση μεθόδων επιφανειακής συντακτικής επεξεργασίας. Για παράδειγμα, στο LEXTER, ένα σύστημα λογισμικού για την εξαγωγή ορολογίας που ανέπτυξε ο Bourigault (1992), πραγματοποιείται αρχικά επιφανειακή συντακτική ανάλυση και ακολούθως αναγνωρίζονται στο κείμενο ορολογικές μονάδες με βάση την υπόθεση ότι η γραμματική δομή των όρων είναι προβλέψιμη. Σε άλλες εργασίες (Boutsis et al. 1999) περιγράφεται μία μέθοδος για την εξαγωγή δίγλωσσης ορολογίας από μεταφρασμένα κείμενα τα οποία έχουν προηγουμένως παραλληλοποιηθεί στο επίπεδο της πρότασης. Η μέθοδος εφαρμόζει γραμματικές συντακτικών προτύπων στο κείμενο των δύο γλωσσών και εξάγει δίγλωσσες αντιστοιχίες μεταξύ λέξεων και φράσεων των κειμένων.

Χαρακτηριστικό επίσης παράδειγμα εφαρμογών που ενσωματώνουν γλωσσική επεξεργασία και συγκεκριμένα συντακτική ανάλυση είναι οι νέες υπηρεσίες πληροφόρησης του Διαδικτύου. Στοχεύουν στην παροχή στο χρήστη αποτελεσματικής πρόσβασης στο μεγάλο όγκο των διαθέσιμων πληροφοριών. Για παράδειγμα, η εξαγωγή πληροφοριών αποσκοπεί στην αναγνώριση και εξαγωγή συγκεκριμένων γεγονότων από κείμενα και στην καταχώρισή τους σε μία βάση δεδομένων. Παρέχει έναν πρακτικό τρόπο αξιοποίησης των πληροφοριών του Διαδικτύου για επιχειρηματικούς και άλλους σκοπούς καθώς τα γεγονότα που εξάγονται αυτόματα μπορούν να χρησιμοποιηθούν για να καθοδηγήσουν στρατηγικές αποφάσεις στον επιχειρηματικό τομέα και αλλού. Για την υλοποίηση των συστημάτων εξαγωγής πληροφοριών η τεχνολογία συντακτικής ανάλυσης είναι πρωταρχικής σημασίας. Είναι αξιοσημείωτο ότι οι γραμματικές συντακτικών προτύπων που υλοποιούνται με τεχνικές πεπερασμένων αυτομάτων αποδίδουν εξαιρετικά καλά, και πολλά συστήματα εξαγωγής πληροφοριών έχουν αντικαταστήσει τα υποσυστήματα πλήρους συντακτικής ανάλυσης που χρησιμοποιούνταν στο παρελθόν, με υποσυστήματα επιφανειακής συντακτικής ανάλυσης (Grishman, 1995· Appelt and Hobbs, 1995).

Μία άλλη εφαρμογή του Διαδικτύου που μπορεί να αξιοποιήσει την επιφανειακή συντακτική ανάλυση είναι η ανάκτηση πληροφοριών. Στόχος της ανάκτησης πληροφοριών είναι η διευκόλυνση του χρήστη στην ανεύρεση εγγράφων που καλύπτουν συγκεκριμένες πληροφοριακές του ανάγκες. Μάλιστα, η ανάκτηση πληροφοριών μπορεί να συνδυαστεί με την εξαγωγή πληροφοριών, ώστε σε ένα πρώτο στάδιο να ανακληθούν έγγραφα σχετικά με την ερώτηση του χρήστη και σε ένα δεύτερο στάδιο να

εξαχθεί η απάντηση στην ερώτηση με τεχνικές εξαγωγής πληροφοριών. Όσον αφορά το ρόλο της συντακτικής ανάλυσης στην ανάκτηση πληροφοριών, έχουν προταθεί διάφορες προσεγγίσεις οι οποίες στη βάση της συντακτικής επεξεργασίας ανυψώνουν το επίπεδο δεικτοδότησης από τη λέξη στη φράση. Για παράδειγμα, στα (Zhai, 1997) και (Evans and Zhai, 1996) προτείνονται μέθοδοι για την αποδοτική αναγνώριση ονοματικών φράσεων που λειτουργούν ως μονάδες δεικτοδότησης στο εμπορικό σύστημα CLARIT (Evans et al., 1995). Παρόμοια, οι Stralkowski και Carballo (1995) μέσω της συντακτικής ανάλυσης αναγνωρίζουν όρους και εξάγουν σχέσεις μεταξύ αυτών με σκοπό την κατασκευή δομών που μπορούν να χρησιμοποιηθούν για τον αποδοτικότερο σχηματισμό ερωτήσεων κατά την ανάκτηση κειμένων.

Τέλος, η αυτόματη μετάφραση χρησιμοποιεί μεθόδους συντακτικής ανάλυσης. Πολλά συστήματα υποβοήθησης μετάφρασης χρησιμοποιούν συντακτικούς αναλυτές για την κατάτμηση του προς μετάφραση κειμένου σε τμήματα τα οποία μεταφράζονται σχεδόν αυτόνομα στην άλλη γλώσσα.

Από τα παραπάνω προκύπτει ότι ένας συντακτικός αναλυτής πρέπει να πληροί ορισμένα κριτήρια προκειμένου να μπορεί να ενσωματωθεί σε συστήματα στο πλαίσιο των νέων υπηρεσιών πληροφόρησης. Στις εφαρμογές που αναφέρθηκαν, ο συντακτικός αναλυτής επεξεργάζεται πραγματικά δεδομένα, δηλαδή ελεύθερο κείμενο, και συνεπώς πρέπει να είναι εύρωστος ως προς το χειρισμό φαινομένων του ελεύθερου κειμένου. Η ακρίβεια του συντακτικού αναλυτή είναι σημαντική για την ενσωμάτωσή του σε περιβάλλοντα με ιδιαίτερες απαιτήσεις. Έχει αποδειχθεί ότι η ακρίβεια της συντακτικής ανάλυσης είναι πρωταρχικής σημασίας για εφαρμογές όπως η εξαγωγή πληροφοριών και η ανάκτηση πληροφοριών (Grishman, 1995). Επίσης, η ύπαρξη αμφισημιών στην έξοδο ίσως εμποδίσει την ολοκλήρωση του συντακτικού αναλυτή σε μεγαλύτερα συστήματα, καθώς μεταθέτει το πρόβλημα επίλυσης της αμφισημίας σε κάποιο επόμενο στάδιο το οποίο συνήθως δεν είναι διαθέσιμο.

Στο πλαίσιο αυτών των παρατηρήσεων, η παρούσα εργασία παρουσιάζει ένα σύστημα επιφανειακής συντακτικής επεξεργασίας για τα Ελληνικά. Απευθύνεται σε εφαρμογές για τις οποίες απαιτείται επεξεργασία κειμένων σε μεγάλη κλίμακα, ενώ η πλήρης συντακτική ανάλυση δεν είναι αναγκαία. Η ανάλυση είναι ντετερμινιστική με την έννοια ότι οι αμφίσημες δομές παραμένουν μερικώς χαρακτηρισμένες, αλλά περικλείονται σε μεγαλύτερες δομές τα όρια των οποίων είναι καλά ορισμένα. Η ταχύτητα της επεξεργασίας, η ευρωστία του συστήματος και η σχετική ακρίβεια των αποτελεσμάτων είναι οι βασικές παράμετροι που καθόρισαν την ανάπτυξη του συστήματος και ικανοποιήθηκαν με την εφαρμογή τεχνικών πεπερασμένων καταστάσεων.

2. Βιβλιογραφική ανασκόπηση

Ένας από τους πρώτους και πιο πετυχημένους ντετερμινιστικούς συντακτικούς αναλυτές που χρησιμοποιήθηκε σε πραγματικές εφαρμογές είναι ο Fidditch (Hindle, 1983a· Hindle, 1983b), ο οποίος βασίζεται στη D-theory του Marcus (1980). Ο Fidditch σχεδιάστηκε όχι ως μερικός συντακτικός αναλυτής αλλά προκειμένου να αναλύει ελεύθερο κείμενο, συμπεριλαμβανομένου κειμένου που προέρχεται από μεταγραφή προφορικού λόγου. Ένα βασικό χαρακτηριστικό του είναι ότι μία φράση της οποίας ο ρόλος στην πρόταση δεν μπορεί να αναγνωριστεί, δεν προσαρτάται πουθενά και η περαιτέρω επεξεργασία δεν επηρεάζεται από αυτήν. Με αυτόν τον τρόπο ο αναλυτής περικλείει την αμφισημία σε μεγαλύτερες δομές, π.χ. προτάσεις, για τις οποίες αναγνωρίζονται βασικά χαρακτηριστικά όπως τα όρια της δομής, το ρήμα, το υποκείμενο, το αντικείμενο κλπ.

Ο Abney (1990) περιγράφει τον CASS, έναν συντακτικό αναλυτή που αποτελείται από μία ακολουθία φίλτρων. Κάθε ένα από αυτά λαμβάνει ντετερμινιστικές αποφάσεις για ένα συγκεκριμένο πρόβλημα. Αν και κάθε φίλτρο μπορεί να έχει να επιλέξει μεταξύ διαφορετικών δυνατοτήτων, δεν προωθεί την αμφισημία στα επόμενα στάδια, τα οποία, στο φως περαιτέρω ανάλυσης, ίσως αναθεωρήσουν κάποια απόφαση που λήφθηκε νωρίτερα. Ένα άλλο χαρακτηριστικό αυτής της προσέγγισης είναι η αναγνώριση των ορίων των φράσεων, προτού επιχειρηθεί αναγνώριση της εσωτερικής τους δομής.

Οι Karlsson (1990), Voutilainen (1993), και Karlsson et al. (1995) προτείνουν ένα σύστημα συντακτικής ανάλυσης που βασίζεται στη χρήση κανόνων που έχουν τη μορφή περιορισμών. Οι κανόνες αυτοί δεν επιτελούν τη συνηθισμένη λειτουργία του ορισμού των ορθών δομών της γλώσσας,

αλλά προδιαγράφουν ακολουθίες χαρακτηριστικών που είναι γραμματικά και συντακτικά μη αποδεκτές. Οι κανόνες προκύπτουν από εκτενείς μελέτες σωμάτων κειμένων και διακρίνονται σε απόλυτους περιορισμούς που ισχύουν πάντοτε, και περιορισμούς που είναι λειτουργικοί στη συντριπτική πλειοψηφία των περιπτώσεων, οπότε το ποσοστό λάθους που εισάγουν είναι πολύ μικρό. Αρχικά, κάθε λέξη λαμβάνει όλα τα δυνατά γραμματικά και συντακτικά χαρακτηριστικά της με τη βοήθεια λεξικού. Στη συνέχεια, εφαρμόζεται στο κείμενο ένα σύνολο από γραμματικούς και συντακτικούς περιορισμούς προκειμένου να απαλειφθούν τα μορφολογικά και στη συνέχεια τα συντακτικά χαρακτηριστικά κάθε λέξης που είναι ασύμβατα με το περιβάλλον της. Κατ΄ αυτόν τον τρόπο, προκύπτουν στην έξοδο συντακτικές δομές με μεγάλα ποσοστά ακρίβειας και ανάκλησης. Για παράδειγμα, στο (Voutilainen, 1993) περιγράφεται το σύστημα NpTool το οποίο αναγνωρίζει ονοματικές φράσεις με ποσοστά ακρίβειας και ανάκλησης μεγαλύτερα από 95%.

Μία μέθοδος συντακτικής ανάλυσης που βασίζεται σε τεχνικές μηχανικής μάθησης παρουσιάζεται στα (Brill, 1993) και (Satta and Brill, 1996). Στο στάδιο της εκπαίδευσης, η μέθοδος συνίσταται στην επεξεργασία κειμένου το οποίο έχει προηγουμένως χαρακτηριστεί συντακτικά. Συγκρίνει το σωστό συντακτικό χαρακτηρισμό του κειμένου με έναν χαρακτηρισμό που παράγεται σχεδόν τυχαία και μαθαίνει τους μετασχηματισμούς που, αν εφαρμοστούν στον τυχαίο μετασχηματισμό, τον μετατρέπουν στον συντακτικά ορθό. Στο στάδιο της συντακτικής ανάλυσης νέου κειμένου, η μέθοδος παράγει έναν τυχαίο συντακτικό χαρακτηρισμό του κειμένου που δίνεται στην είσοδο και εφαρμόζει σε αυτόν τους μετασχηματισμούς που έχει μάθει κατά το στάδιο της εκπαίδευσης.

Χρήση πεπερασμένων μεταγραφέων για συντακτική ανάλυση κάνουν επίσης οι Grefenstette (1996), Ait-Mokhtar και Channod (1997). Η ανάλυση πραγματοποιείται με μεταγραφείς που εισάγουν στο κείμενο ή διαγράφουν από αυτό ειδικά σύμβολα που οριοθετούν την αρχή και το τέλος των φράσεων. Η ανάλυση πραγματοποιείται με πολλαπλά περάσματα κάθε ένα από τα οποία αναγνωρίζει ένα νέο συστατικό ή διορθώνει τον χαρακτηρισμό που παρήγαγε κάποιο προηγούμενο στάδιο στο φως νέων δεδομένων.

Μία επιπρόσθετη προσέγγιση στην επιφανειακή συντακτική ανάλυση από τον Abney είναι ο CASS2 (Abney 1996; Abney 1997), που βασίζεται σε τεχνικές πεπερασμένων μεταγραφέων. Η συντακτική ανάλυση πραγματοποιείται σε μία σειρά από επίπεδα, σε κάθε ένα από τα οποία αναγνωρίζονται συγκεκριμένα φαινόμενα. Οι φράσεις που αναγνωρίζονται σε ένα επίπεδο σχηματίζονται από φράσεις κατώτερων επιπέδων και ποτέ από φράσεις του ίδιου ή ανώτερου επιπέδου. Κάθε επίπεδο επεξεργασίας υλοποιείται μέσω ενός πεπερασμένου μεταγραφέα ο οποίος εισάγει συντακτικά χαρακτηριστικά στο κείμενο. Η βασική διαφορά από άλλες προσεγγίσεις που χρησιμοποιούν πεπερασμένους μεταγραφείς (Koskienniemi, 1992; Roche, 1993) έγκειται στο ότι κάθε φράση που αναγνωρίζεται αντικαθιστά πλήρως τις υποκείμενες φράσεις, όπως και στις παραδοσιακές μεθόδους συντακτικής ανάλυσης. Σύμφωνα με αυτή την προσέγγιση, μία ακολουθία μεταγραφέων συνδέονται έτσι ώστε η έξοδος του ενός να δίνεται ως είσοδος στον επόμενο. Όπως και στον CASS, κάθε στάδιο είναι υπεύθυνο για την αναγνώριση ενός συγκεκριμένου συντακτικού φαινομένου και λαμβάνει μία ντετερμινιστική απόφαση σχετικά με την ύπαρξη ή μη ενός φαινομένου, κάτι που έχει ως αποτέλεσμα να παράγεται στην έξοδο μία μη αμφίσημη συντακτική αναπαράσταση του κειμένου. Οι κανόνες έχουν σχεδιαστεί έτσι ώστε να είναι πιο αξιόπιστοι, όταν εφαρμόζονται σύμφωνα με την αρχή του μέγιστου ταιριάσματος.

3. Μεθοδολογία

Η μεθοδολογία που παρουσιάζεται σε αυτή την ενότητα αποσκοπεί στη μερική συντακτική ανάλυση με χρήση τεχνικών πεπερασμένων καταστάσεων και πλαισίων υποκατηγοριοποίησης. Σύμφωνα με αυτή την προσέγγιση, ένα κείμενο αναλύεται συντακτικά από κανόνες που έχουν τη μορφή κανονικών εκφράσεων οι οποίες μπορούν να μεταφραστούν σε πεπερασμένα αυτόματα ή πεπερασμένους μεταγραφείς με γνωστές τεχνικές (Roche and Schabes, 1997). Οι κανόνες είναι αριθμημένοι, ώστε να εφαρμόζονται στο κείμενο με συγκεκριμένη σειρά και να αναγνωρίζουν συστατικά μιας κατηγορίας με βάση τα συστατικά που έχουν ήδη αναγνωριστεί και ειδικούς οριοθέτες που έχουν παρεμβληθεί από προηγούμενους κανόνες και αφαιρούνται πριν το τέλος της επεξεργασίας. Η επεξεργασία είναι ντετερμινιστική και δε λαμβάνει χώρα επαναληπτικός έλεγχος (backtraking), όπως συμβαίνει σε άλλα

είδη αναλυτών. Όσον αφορά την αναμφίσημη εφαρμογή των κανονικών εκφράσεων στο κείμενο, αυτή είναι δυνατή εν μέρει χάρη στην αρχή του μέγιστου ταιριάσματος (Kartunnen , 1997), σύμφωνα με την οποία οι κανόνες σχεδιάζονται έτσι ώστε να είναι περισσότερο αξιόπιστοι, όταν επιδρούν στο μεγαλύτερο δυνατό ταίριασμά τους στο κείμενο. Η αρχή του μέγιστου ταιριάσματος είναι συμβατή με το ψυχογλωσσικό κριτήριο του right association ή late closure σύμφωνα με το οποίο φυσικοί ομιλητές της γλώσσας προσαρτούν τα νέα συστατικά στη φράση που τελευταία υπόκειται σε επεξεργασία (Allen, 1995; Papadopoulou, 1998). Δύο επιπρόσθετοι παράγοντες που συμβάλλουν στη μη αμφίσημη ανάλυση της εξόδου είναι οι εξής:

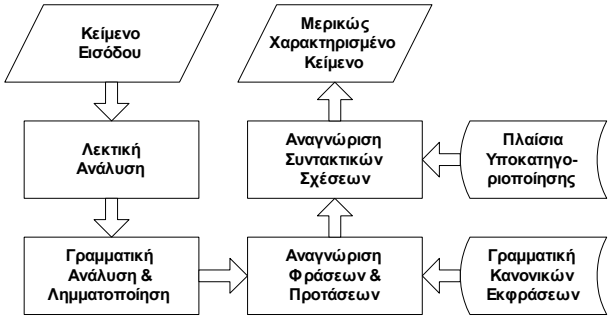
- Ορισμένες δομές δεν αναλύονται στον μέγιστο δυνατό βαθμό, δηλαδή κάποια είδη αμφισημίας δεν επιλύονται,
- Λαμβάνονται ορισμένες αποφάσεις που ευνοούν μία ανάλυση σε σχέση με τις υπόλοιπες, όταν η ανάλυση που προτιμάται έχει μεγαλύτερη πιθανότητα να είναι η σωστή.

Οι κανόνες μεταγλωττίζονται σε πεπερασμένους μεταγραφείς με τη χρήση του πακέτου FSA6.2 (Van Noord and Gerdemann, 1999) ενώ ο ίδιος ο parser είναι σχεδιασμένος στη γλώσσα προγραμματισμού C για λόγους ταχύτητας και απόδοσης.

4. Αρχιτεκτονική

Η αρχιτεκτονική του συστήματος απεικονίζεται στο Σχήμα 1. Η συντακτική ανάλυση πραγματοποιείται σε δύο στάδια 'Αναγνώριση Φράσεων & Προτάσεων' και 'Αναγνώριση Γραμματικών Σχέσεων', ενώ προηγούνται δύο στάδια προεπεξεργασίας, 'Λεκτική Ανάλυση' και 'Γραμματική Ανάλυση & Λημματοποίηση'. Η προσέγγιση που ακολουθείται είναι συμβατή με το πρότυπο EAGLES (Leech et al., 1996) και έγκειται στην εξής ακολουθία ενεργειών: αναγνώριση των ορίων των φράσεων, απόδοση χαρακτηριστικών υποκατηγοριοποίησης (subcategorisation information), και αναγνώριση γραμματικών σχέσεων. Για την αναγνώριση γραμματικών σχέσεων χρησιμοποιούνται πλαίσια υποκατηγοριοποίησης τα οποία απαριθμούν για κάθε κεφαλή φράσης τις γραμματικές σχέσεις στις οποίες μπορεί αυτή να συμμετάσχει. Σύμφωνα με το παραπάνω σχήμα, τα πλαίσια εφαρμόζονται στο κείμενο μετά την αναγνώριση των βασικών φραστικών και προτασιακών δομών.

Η λεκτική ανάλυση πραγματοποιείται σύμφωνα με τη μεθοδολογία που υιοθετήθηκε στο πρόγραμμα Multext (Di Christo et. al, 1995). Πρώτα λαμβάνει χώρα η αναγνώριση των ορίων των λέξεων, των ημερομηνιών, των συντμήσεων κλπ. Γι' αυτό το σκοπό πρέπει να επιλυθεί η αμφισημία όσον αφορά τη χρήση των σημείων στίξης καθώς βασικές δομικές μονάδες, π.χ. αριθμοί, αλφαριθμητικές αναφορές, ημερομηνίες, αρκτικόλεξα, συντμήσεις κλπ., είναι δυνατό να εμπερικλείουν σημεία στίξης. Ακολουθώντας την κοινή πρακτική, ο λεκτικός αναλυτής αυτού του σταδίου κάνει χρήση κανονικών εκφράσεων για τον ορισμό των λέξεων κλπ. σε συνδυασμό με λίστες για τη γλώσσα των υπό ανάλυση κειμένων και ευριστικούς κανόνες άρσης της αμφισημίας κατά την εφαρμογή των κανόνων.



Σχήμα 1: Αρχιτεκτονική συστήματος μερικής συντακτικής ανάλυσης

Μετά την αναγνώριση των επιφανειακών δομών του, το κείμενο προωθείται στο υποσύστημα γραμματικής ανάλυσης και λημματοποίησης (Papageorgiou et al., 2000). Χρησιμοποιείται μία έκδοση του γραμματικού αναλυτή του Brill (1993b) που έχει εκπαιδευτεί σε ελληνικό κείμενο χαρακτηρισμένο σύμφωνα με το σύνολο χαρακτηριστικών ILSP PAROLE (Labropoulou et al., 1997). Υπάρχουν 584 διαφορετικά χαρακτηριστικά. Η ακρίβεια του χαρακτηρισμού είναι περίπου 90%, όταν όλα τα χαρακτηριστικά λαμβάνονται υπόψη, και 96%, όταν μόνο το μέρος του λόγου λαμβάνεται υπόψη. Η διαδικασία έχει ως εξής: Αρχικά ο αναλυτής δίνει σε κάθε λέξη ένα χαρακτηριστικό με βάση ένα λεξικό που έχει σχηματιστεί κατά το στάδιο της εκπαίδευσης. Ένα λεξικό καταλήξεων χρησιμοποιείται για την απόδοση χαρακτηριστικών σε άγνωστες λέξεις. Στη συνέχεια 799 κανόνες συμφραζομένων μετασχηματίζουν και βελτιώνουν τους αρχικούς χαρακτηρισμούς. Αφού κάθε λέξη πάρει τον τελικό γραμματικό χαρακτηρισμό της, ένα λεξικό 70.000 λημμάτων χρησιμοποιείται για την απόδοση λήμματος σε κάθε λέξη.

Πριν λάβει χώρα η συντακτική ανάλυση, το σύνολο των γραμματικών χαρακτηριστικών μετασχηματίζεται και προσαρμόζεται στις ανάγκες της επεξεργασίας που ακολουθεί. Συγκεκριμένα, πληροφορίες οι οποίες δε χρησιμοποιούνται στα επόμενα στάδια (π.χ. γένος ουσιαστικών, επιθέτων κλπ.) απαλείφονται. Κατά συνέπεια, ο αριθμός των διαφορετικών χαρακτηριστικών μειώνεται, γεγονός που οδηγεί σε πιο συμπαγείς μεταγραφείς που απαιτούν λιγότερο χρόνο για να μεταγλωττιστούν και να εφαρμοστούν στο κείμενο. Παράλληλα, ειδικές περιπτώσεις λέξεων χαρακτηρίζονται με στοιχεία πληροφορίας όπως το πλαίσιο υποκατηγοριοποίησής τους, το λήμμα τους, ή το είδος των συστατικών τα οποία συνήθως εισάγουν. Για παράδειγμα, το επίρρημα *μαζί* χαρακτηρίζεται *ad_me_gen* αφού μπορεί να ακολουθείται από προθετική φράση που εισάγει η πρόθεση *με*, ή από τον αδύνατο τύπο της προσωπικής αντωνυμίας σε γενική. Τα `όλος` και `ολόκληρος` λαμβάνουν διαφορετικά χαρακτηριστικά δεδομένης της διαφορετικής τους σύνταξης σε σχέση με άλλα επίθετα (*όλος ο κόσμος* αλλά **ο όλος ο κόσμος*). Τα χαρακτηριστικά αυτά έχουν το πρόθημα *olos_* που ακολουθείται από τον αριθμό και την πτώση του επιθέτου: *olos_sgnm*, κλπ. Επίσης οι σύνδεσμοι *αν*, και *εάν* παίρνουν το χαρακτηρισμό *conj_cond*, ο οποίος θα χρησιμοποιηθεί αργότερα για την αναγνώριση υποθετικών προτάσεων.

Τέλος, το γραμματικά χαρακτηρισμένο κείμενο περνάει στο συντακτικό αναλυτή. Η συντακτική ανάλυση λαμβάνει χώρα σε δύο στάδια. Στο πρώτο, οι φράσεις και οι προτάσεις αναγνωρίζονται με τη χρήση μιας γραμματικής πεπερασμένων μεταγραφέων ενώ το δεύτερο στάδιο αναλαμβάνει τη σηματοδότηση των γραμματικών σχέσεων μεταξύ των αναγνωρισμένων συστατικών βασισμένο σε ένα λεξικό υποκατηγοριοποίησης και ένα μηχανισμό ταιριάσματος προτύπων (pattern matching).

```
%% insert 'stop' symbol before 'as' tags (prepositions)
markup(as, stop, []) o

%% group np's in accusative with post modifying np's in genitive
markup([const('[np_ac', '/np_ac]'), [const('[np_ge', '/np_ge]'), [{cjco,
punct_cm}^, const('[np_ge', '/np_ge]')]*]*], '[npacmax', '/npacmax]') o

%% mark pp with possibly coordinated np's in accusative
markup([stop, as+, {ptng, cjco, const('[np_ge', '/np_ge]'), const('[advp',
'/advp]')}*, const('[npacmax', '/npacmax]'), [{cjco, punct_cm},
const('[npacmax', '/npacmax]')]*], '[pp', '/pp]').

<EOR>    %% End Of Rule (stage 1)

%% remove stop symbol before prepositions already participating in pp's
['[pp', stop] => '[pp' o

%% mark pp with possibly coordinated np's in genitive
markup([stop, as+, {ptng, cjco, const('[advp', '/advp]')}*, const('[np_ge',
'/np_ge]'), [{cjco, punct_cm}^, const('[np_ge', '/np_ge]')]*], '[pp', '/pp]') o

%% delete temporary markers
{'[npacmax', '/npacmax]', stop} => [] o

<EOR>    %% End Of Rule (stage 2)
```

Σχήμα 2: Ακολουθία κανόνων για την αναγνώριση προθετικών φράσεων

5. Συγκρότηση του Σώματος Κειμένων

Η αξιολόγηση του συντακτικού αναλυτή έγινε σε ένα σώμα κειμένων που απαρτίζεται από ειδησεογραφικά και οικονομικά άρθρα από τις δικτυακές εκδόσεις ελληνικών περιοδικών και εφημερίδων. Το συνολικό μέγεθος της συλλογής είναι περίπου 32000 λέξεις. Τα κείμενα χαρακτηρίστηκαν ως προς τη συντακτική τους πληροφορία από δύο γλωσσολόγους που χρησιμοποίησαν ένα περιβάλλον που σχεδιάστηκε στη γλώσσα προγραμματισμού Java για αυτό το σκοπό. Ένας αριθμός κειμένων χαρακτηρίστηκε και από τους δύο γλωσσολόγους για να εξασφαλιστεί η συνέπεια στην αντιμετώπιση των φαινομένων. Στο 95% των περιπτώσεων υπήρξε ταύτιση απόψεων.

6. Γραμματική

Η γραμματική περιέχει κανόνες που αναγνωρίζουν τις ακόλουθες φραστικές κατηγορίες: επιθετική φράση, ονοματική φράση, ρηματική ομάδα, προθετική φράση, επιρρηματική φράση. Στο επίπεδο της πρότασης αναγνωρίζονται κύριες και δευτερεύουσες προτάσεις. Η αντιμετώπιση αυτή ακολουθεί κατά μεγάλο βαθμό τις προτάσεις για τα σχήματα σχολιασμού του EAGLES (Leech et al., 1996).

Στο Σχήμα 2, παρατίθεται η ομάδα των κανόνων που αναγνωρίζουν προθετικές φράσεις. Ένας κανόνας *markup(X, y, z)* περικλείει μέγιστα ταιριάσματα της κανονικής έκφρασης *X* μέσα στα *y* και *z*, ενώ ένας κανόνας *X => y* αντικαθιστά μέγιστα ταιριάσματα της κανονικής έκφρασης *X* με το *y*, (Kartunnen, 1997). Μία ή περισσότερες μακροεντολές ομαδοποιούνται με τον τελεστή (○), ο οποίος τις συνθέτει σε έναν κανόνα που μεταφράζεται σε έναν πεπερασμένο μεταγραφέα. Όπως μπορούμε να δούμε, η αναγνώριση των προθετικών φράσεων λαμβάνει χώρα σε δύο στάδια. Κατά το πρώτο, αναγνωρίζονται οι φράσεις που αποτελούνται από μια πρόθεση που ακολουθείται από μία ή περισσότερες ονοματικές φράσεις σε αιτιατική. Κατά το δεύτερο στάδιο, αναγνωρίζονται παρόμοιες δομές με ονοματικές φράσεις σε γενική. Σε κάθε περίπτωση λαμβάνεται υπόψη η πιθανή εμφάνιση επιρρημάτων ή αρνητικών μορίων.

Ακολουθεί μια περιληπτική περιγραφή των φαινομένων που αντιμετωπίζει η γραμματική στα πλαίσια κάθε φραστικής και προτασιακής κατηγορίας.

6.1. Επιρρηματικές Φράσεις

Οι επιρρηματικές φράσεις σχηματίζονται από επιρρήματα που πιθανώς προσδιορίζονται από άλλα επιρρήματα. Επίσης, μπορεί να υπάρχουν ονοματικές και προθετικές φράσεις που λειτουργούν ως συμπληρώματα. Η αναγνώριση των επιρρηματικών φράσεων γίνεται σε δύο φάσεις. Η πρώτη φάση λαμβάνει χώρα πριν την αναγνώριση των επιθετικών και των ονοματικών φράσεων, ώστε οι επιρρηματικές φράσεις να μπορούν να αναγνωριστούν ως συστατικά των φράσεων αυτών. Η δεύτερη φάση διορθώνει τις επιρρηματικές φράσεις που αναγνωρίστηκαν στην πρώτη φάση, ώστε να συμπεριληφθούν σε αυτές ορίσματα σύμφωνα με το πλαίσιο υποκατηγοριοποίησης του κάθε επιρρήματος. Αυτή είναι η μόνη περίπτωση χρήσης πλαισίων υποκατηγοριοποίησης κατά την αναγνώριση φράσεων και προτάσεων. Η εξαίρεση αυτή είναι απαραίτητη, επειδή η μερική αναγνώριση των επιρρηματικών φράσεων κατά την αναγνώριση φράσεων και προτάσεων θα μπορούσε να εμποδίσει την αναγνώριση μεγαλύτερων δομών που τις εμπεριέχουν.

```
[advp Δυστυχώς advp] καμία πρόταση δεν έγινε  ...
[advp εκτός [pp από [np_ac το Νίκο np_ac] pp] advp] ...
[advp ανεξαρτήτως [np_ge αποτελέσματος np_ge] advp] ...
```

6.2. Επιθετικές Φράσεις

Η επιθετική φράση περιέχει ένα ή περισσότερα επίθετα, αριθμητικά ή/και παθητικές μετοχές που πιθανόν προσδιορίζονται από ένα ή περισσότερα επιρρήματα. Επίθετα και μετοχές που χωρίζονται (προαιρετικά) με κόμμα ή συμπλεκτικούς συνδέσμους περικλείονται στην ίδια φράση υπό τον όρο ότι συμφωνούν σε πτώση και αριθμό. Οι επιθετικές φράσεις κατηγοριοποιούνται σύμφωνα με τη γραμματική πτώση της κεφαλής τους. Έτσι, διακρίνονται επιθετικές φράσεις σε ονομαστική, γενική και αιτιατική που μαρκάρονται ως `adjp_nm`, `adjp_ge` και `adjp_ac` αντίστοιχα.

```
Η [adjp_nm καθημερινή adjp_nm] ενημέρωση ...
του [adjp_ge πασίγνωστου πια /adjp_ge] ...
Η [adjp_nm πολύ γρήγορη και αποτελεσματική adjp_nm] απάντηση ...
```

6.3. Ονοματικές Φράσεις

Εκτός από ουσιαστικά, η ονοματική φράση μπορεί να έχει ως κεφαλή αντωνυμία, επίθετο (σε περίπτωση έλλειψης ή ονοματοποιημένου επιθέτου), μετοχή ή αριθμητικό. Τυχόν προσδιορισμοί πριν από την κεφαλή, όπως αντωνυμίες, αριθμητικά, επιθετικές φράσεις κλπ., περιλαμβάνονται στην ονοματική φράση, ενώ ως προσδιορισμοί μετά την κεφαλή μπορούν να λειτουργούν επίθετα, δεικτικές αντωνυμίες και κτητικές αντωνυμίες. Ονοματικές φράσεις σε γενική, προθετικές και επιρρηματικές φράσεις καθώς και δευτερεύουσες προτάσεις που προσδιορίζουν ή αποτελούν συμπληρώματα στην ονοματική φράση αναγνωρίζονται ως ανεξάρτητα συστατικά. Αυτό οφείλεται στο ότι η προσάρτηση των συστατικών αυτών δεν είναι δυνατό να γίνει με αναμφίσημο τρόπο στο επίπεδο της επιφανειακής συντακτικής ανάλυσης. Διακρίνονται ονοματικές φράσεις σε ονομαστική, γενική, αιτιατική, δοτική και κλητική, αν και στις τελευταίες δύο πτώσεις αναγνωρίζονται στοιχειώδεις μόνο δομές.

```
[np_nm ο [adjp_nm [advp πρώην /advp] επιστημονικός /adjp_nm]
συνεργάτης /np_nm]
[np_nm το [adjp_nm οικονομικό /adjp_nm] ινστιτούτο IFW /np_nm]
[np_nm ο Χέρμπερτ Χαξ /np_nm]
[np_nm οι τάξεις 2, 3 και 4 /np_nm]
```

6.4. Προθετικές Φράσεις

Η αναγνώριση των προθετικών φράσεων βασίζεται στην αναγνώριση των βασικών δομών των ονοματικών φράσεων του κειμένου. Σε γενικές γραμμές, οι προθετικές φράσεις αποτελούνται από μία πρόθεση που ακολουθείται από μία ή περισσότερες ονοματικές φράσεις σε αιτιατική ή γενική, οι οποίες βρίσκονται σε σχέση παράταξης. Οι ονοματικές φράσεις που κατέχουν τη θέση συμπληρώματος της πρόθεσης μπορούν να προσδιορίζονται από άλλες ονοματικές φράσεις σε γενική, οι οποίες αναγνωρίζονται μέσα στα όρια της προθετικής φράσης.

```
[pp για [np_ac τη δημιουργία /np_ac] [np_ge ενός νέου τύπου
np_ge][np_ge κομματικού οργανισμού /np_ge] /pp]
[pp με [np_ac το τυχόν Γενικό Πολεοδομικό Σχέδιο /np_ac] [np_ge της
περιοχής /np_ge] ή [np_ac τα όρια /np_ac] [np_ge οικισμού /np_ge]
/pp]
```

6.5. Ρηματικά Σύνολα

Η ρηματική ομάδα περιέχει το ρήμα μαζί με το τυχόν βοηθητικό για το σχηματισμό περιφραστικών χρόνων. Μόρια που εκφράζουν άρνηση, μέλλοντα χρόνο ή υποτακτική έγκλιση περιλαμβάνονται στη ρηματική ομάδα. Επιρρηματικές φράσεις και κλιτικά τα οποία παγιδεύονται μέσα στη ρηματική ομάδα από τα παραπάνω στοιχεία περικλείονται επίσης. Η ρηματική ομάδα υποκατηγοριοποιείται περαιτέρω και μαρκάρεται αντιστοίχως με ειδικούς δείκτες ως εξής: ρηματική ομάδα με ρήμα οριστικής (vg), με ρήμα υποτακτικής (vg_s) και με μετοχή ενεργητικού ενεστώτα (vg_g).

```
Η Επιτροπή [vg  ενέκρινε vg] το έργο.
Οι γιατροί [vg δεν [np_ge τους np_ge] [np_ac το np_ac] έχουν [advp
ακόμη advp] πει vg].
Οι Ευρωπαίοι προσπάθησαν [vg_s να μποϋκοτάρουν vg_s] τη συγχώνευση.
[vg_g Αναγνωρίζοντας vg_g]  την ήττα του ...
```

6.6. Προτάσεις

Μετά την αναγνώριση των φράσεων, η συντακτική ανάλυση περιλαμβάνει την αναγνώριση των προτάσεων. Η αναγνώριση των προτάσεων βασίζεται μεταξύ άλλων σε μία λίστα λέξεων και συμβόλων για την αναγνώριση της αρχής και του τέλους των προτάσεων. Υποτακτικοί σύνδεσμοι, αντωνυμίες, επιρρηματικές φράσεις κλπ. χρησιμοποιούνται για την αναγνώριση των ορίων των προτάσεων. Η ύπαρξη μόνο μιας ρηματικής ομάδας σε κάθε πρόταση είναι ένα ισχυρό κριτήριο για την αναγνώριση προτάσεων.

Αναγνωρίζονται τόσο κύριες όσο και δευτερεύουσες προτάσεις. Οι τελευταίες διακρίνονται σε αναφορικές, αόριστες αναφορικές, χρονικές, υποθετικές και ερωτηματικές. Οι υπόλοιπες δευτερεύουσες προτάσεις, όπως και εκείνες που δε μπορούν με αξιοπιστία να καταχωριστούν σε μία από τις παραπάνω κατηγορίες, χαρακτηρίζονται ως *άλλες*, `cl_other`, δευτερεύουσες προτάσεις. Παρακάτω δίνονται μερικά παραδείγματα προτάσεων που αναγνωρίζονται:

Κύριες και Αναφορικές `[cl Για εκείνους [cl_re που υποφέρουν υπό το βάρβαρο καθεστώς cl_re] ... cl]`

Αόριστες Αναφορικές `[cl [cl_ri Όποιοι εξασφάλισαν την ελευθερία τους cl_ri] κέρδισαν ... cl]`

Χρονικές `[cl_t Όταν η αγορά κατέρρευσε cl_t], [cl οι μέτοχοι ... cl]`

[cl_c Αν η κυβέρνηση αποφασίσει την καταστολή της απεργίας cl_c], [cl οι εργαζόμενοι ... cl]

Ερωτηματικές [cl Συνειδητοποίησα cl][cl_ir πόσο θα βοηθήσει ο νόμος ... cl_ir]

Άλλες Δευτερεύουσες [cl Η εφημερίδα αποφάσισε cl] [cl_o να ενημερώσει τους αναγνώστες της ... cl_o]

Οι αναφορικές και οι αναφορικοαόριστες προτάσεις περικλείονται πάντοτε σε άλλες προτάσεις. Οι προτάσεις άλλων τύπων περικλείονται σε άλλες προτάσεις, μόνο αν είναι παγιδευμένες σε αυτές.

[cl_c Αν, [cl_o αφού ολοκληρωθεί η εξαγορά cl_o], δεν υπάρχουν διαθέσιμα κεφάλαια cl_c], [cl θα προχωρήσουμε ... cl].

6.7. Γραμματικές Σχέσεις

Η αναγνώριση γραμματικών σχέσεων είναι το τελευταίο στάδιο επεξεργασίας του συστήματος επιφανειακής συντακτικής ανάλυσης. Στο στάδιο αυτό, κείμενο στο οποίο έχουν σημειωθεί φράσεις και προτάσεις υφίσταται επεξεργασία με στόχο την αναγνώριση γραμματικών σχέσεων μεταξύ των συστατικών αυτών. Συγκεκριμένα, αναγνωρίζονται οι εξής σχέσεις: υποκείμενα, άμεσα αντικείμενα, κατηγορούμενα, προτασιακά και προθετικά συμπληρώματα ρημάτων. Για την αναγνώριση αυτών των φαινομένων το σύστημα χρησιμοποιεί πληροφορία συντακτικού επιπέδου του λεξικού PAROLE (Gavrilidou et al., 1998), το οποίο για τα ρήματα περιέχει 5927 εγγραφές. Τα ρήματα που περιλήφθηκαν στο λεξικό έχουν επιλεγεί με βάση τη μεγάλη συχνότητα εμφάνισής τους σε κείμενα της Νέας Ελληνικής.

Παραδείγματα πλαισίων υποκατηγοριοποίησης ρήματος είναι τα εξής:

δίνω #subj_np_nm##obj_np_ac##ind_obj_np_ge
δίνω #subj_np_nm##obj_np_ac#ind_pp_se#

Παρατηρούμε ότι τα πλαίσια προβλέπουν ονοματική φράση σε ονομαστική για υποκείμενο, ονοματική φράση σε αιτιατική για αντικείμενο, ονοματική φράση σε γενική για έμμεσο αντικείμενο ή εναλλακτικά έμμεσο αντικείμενο που εισάγεται με την πρόθεση σε. Ορισμένα ρήματα, δηλαδή, έχουν περισσότερα από ένα πλαίσια.

Όσον αφορά την εφαρμογή των πλαισίων υποκατηγοριοποίησης, αυτή περιλαμβάνει τα εξής: Αρχικά, σε κάθε πρόταση που αναγνωρίστηκε στο προηγούμενο στάδιο εντοπίζεται το ρήμα. Με βάση το ρηματικό λήμμα, που έχει υπολογιστεί στο στάδιο της γραμματικής ανάλυσης και λημματοποίησης, αναζητούνται τα πλαίσια υποκατηγοριοποίησης που αντιστοιχούν σε αυτό. Αν δεν βρεθεί πλαίσιο, η επεξεργασία συνεχίζει με την επόμενη πρόταση. Αν βρεθούν περισσότερα από ένα πλαίσια, εφαρμόζονται όλα στην πρόταση και, τελικά, γίνεται δεκτό εκείνο με τη μεγαλύτερη κάλυψη της πρότασης, δηλαδή εκείνο που αναγνώρισε τα περισσότερα συμπληρώματα.

Παραδείγματος χάριν, σχετικά με τα υποκείμενα και τα κατηγορούμενα, αρχικά, εντοπίζονται όλες οι ονοματικές φράσεις σε ονομαστική εντός των ορίων της πρότασης. Αν δύο ή περισσότερες από αυτές βρίσκονται σε παρατακτική σύνδεση, λαμβάνονται ως μία ενιαία μονάδα. Διακρίνονται τρεις περιπτώσεις: το πλαίσιο να προβλέπει μόνο υποκείμενο, μόνο κατηγορούμενο ή και τα δύο. Αν αναζητείται μόνο υποκείμενο, τότε επιλέγεται η πλησιέστερη στο ρήμα ονοματική φράση από αριστερά του ρήματος και, αν δε βρεθεί, η αναζήτηση συνεχίζει στα δεξιά του ρήματος. Αν αναζητείται μόνο κατηγορούμενο, τότε λαμβάνεται η πρώτη ονοματική φράση που κατά προτίμηση δεν περιέχει οριστικό άρθρο από δεξιά του ρήματος. Σε αντίθετη περίπτωση αναζητούνται ονοματικές φράσεις στα αριστερά του ρήματος. Αν αναζητούνται υποκείμενο και κατηγορούμενο, τότε χρησιμοποιούνται ορισμένοι απλοί ευριστικοί κανόνες. Για παράδειγμα, αν μία ονοματική φράση σε

ονομαστική περιέχει οριστικό άρθρο ή αόριστη αναφορική αντωνυμία (στην περίπτωση αόριστης αναφορικής πρότασης), τότε αυτή λαμβάνεται ως υποκείμενο. Αν το υποκείμενο είναι αριστερά του ρήματος, το κατηγορούμενο αναζητείται στα δεξιά. Παρόμοιες ευριστικές μέθοδοι ακολουθούνται για την αναγνώριση των άλλων ορισμάτων του ρήματος (αντικείμενα σε αιτιατική ή προθετική φράση, προτασιακά ορίσματα κλπ).

7. Δείγμα Εξόδου

Θα χρησιμοποιήσουμε την πρόταση που ακολουθεί, καθώς και την πλήρη αναπαράστασή της στο Παράρτημα, ως δείγμα της εξόδου του συστήματος

```
Η Αγροτική Τράπεζα ενημέρωσε τους αρμόδιους ερευνητές της υπόθεσης
ότι ανευρέθησαν 33 στελέχη που είχαν προχωρήσει σε εικονικές
προεγγραφές.
```

Παρατηρούμε ότι έχουν αναγνωριστεί μία κύρια και δύο δευτερεύουσες προτάσεις. Ο συντακτικός αναλυτής έχει συμπεριλάβει την αναφορική πρόταση η οποία αρχίζει με την αντωνυμία *που* στην πρόταση που αρχίζει με το σύνδεσμο *ότι*. Το ρήμα *προχωρώ* στην αναφορική έχει το πλαίσιο υποκατηγοριοποίησης `#subj_np_nm##advp##pp_se#`. Έτσι, η προθετική φράση *σε εικονικές προεγγραφές* έχει αναγνωριστεί ως όρισμα του ρήματος. Επίσης η αναφορική αντωνυμία *που* έχει αναγνωριστεί ως το υποκείμενο του ρήματος αφού έχει ληφθεί υπόψη η γραμματική πληροφορία της πτώσης `PnReNe03Pl`**`Nm`**`Xx` που έχει αποδοθεί από το μορφολογικό σχολιαστή.

8. Αποτελέσματα και Συμπεράσματα

Η απόδοση του συστήματος υπολογίστηκε με βάση τις γνωστές μετρικές ακρίβειας και ανάκλησης. Οι τιμές των μετρικών αυτών για την αναγνώριση κάθε κατηγορίας φράσεων και προτάσεων δίνονται παρακάτω.

$$\text{Ακρίβεια} = \frac{\text{Αριθμός ορθών φράσεων που αναγνωρίστηκαν}}{\text{Αριθμός φράσεων που αναγνωρίστηκαν}}$$

$$\text{Ανάκληση} = \frac{\text{Αριθμός ορθών φράσεων που αναγνωρίστηκαν}}{\text{Αριθμός φράσεων που έπρεπε να αναγνωριστούν}}$$

Για κάθε περίπτωση, δίνονται στον Πίνακα 1 οι τιμές της ακρίβειας και της ανάκλησης όταν η είσοδος στα υποσυστήματα συντακτικής ανάλυσης είναι διορθωμένη, δηλαδή τα λάθη της λεκτικής και της γραμματικής ανάλυσης έχουν διορθωθεί. Με αυτό τον τρόπο, μπορούμε να υπολογίσουμε τα λάθη που οφείλονται καθαρά στη συντακτική ανάλυση. Τα ποσοστά ακρίβειας και ανάκλησης για την αναγνώριση των φράσεων είναι πολύ καλά, καθώς κυμαίνονται στις περισσότερες περιπτώσεις μεταξύ 95% και 100%. Επίσης, η ακρίβεια αναγνώρισης των προτάσεων είναι σταθερά πάνω από 70%. Η ακρίβεια της ανάλυσης είναι ικανοποιητική για τις περισσότερες εφαρμογές όπου απαιτείται ανάλυση μεγάλου όγκου κειμένων, δίχως να είναι απαραίτητη η πλήρης συντακτική ανάλυση. Πέρα από τη σχετική ακρίβεια των αποτελεσμάτων, η ευρωστία και η αποδοτικότητα της επεξεργασίας είναι επιπλέον χαρακτηριστικά της μεθόδου που διευκολύνουν την ενσωμάτωσή της σε πραγματικά συστήματα. Χαρακτηριστικά αναφέρουμε ότι η ταχύτητα της συντακτικής ανάλυσης είναι (χωρίς τα στάδια της λεξικής και της μορφολογικής ανάλυσης) είναι ~260 λέξεις/δευτερόλεπτο.

Είδος Συστατικού	Ακρίβεια (Διορθωμένη είσοδος)	Ανάκληση (Διορθωμένη είσοδος)
adjp_nm	0.95	0.96
adjp_ge	0.97	0.96
adjp_ac	0.96	0.97
np_nm	0.93	0.93
np_ge	0.94	0.94
np_ac	0.95	0.95
advp	0.92	0.91
pp	0.87	0.86
vg	0.94	0.97
cl	0.70	0.81
cl_r	0.89	0.85
cl_ri	0.86	0.92
cl_ir	0.90	0.75
cl_c	0.77	0.73
cl_t	0.92	0.72
cl_o	0.72	0.75
Είδος Γραμματικής Σχέσης	**Ακρίβεια (Διορθωμένη είσοδος)**	**Ανάκληση (Διορθωμένη είσοδος)**
Υποκείμενα	0.95	0.75
Κατηγορούμενα	0.81	0.84
Άμεσα Αντικείμενα	0.79	0.82
Προθ. Φράσεις -Συμπληρώματα	0.76	0.72
Προτασιακά Ορίσματα	0.84	0.80

Σχήμα 3: Απόδοση συστήματος

Οι προσπάθειες μας αυτή τη στιγμή εστιάζονται στη βελτίωση των αποτελεσμάτων όσον αφορά τα όρια και το είδος των προτάσεων, καθώς και την αναγνώριση των γραμματικών σχέσεων. Σε αυτό το πλαίσιο, σχεδιάζουμε να μεγαλώσουμε τη βάση των πλαισίων υποκατηγοριοποίησης και να εξετάσουμε την πιθανότητα αυτόματης εξαγωγής αυτών των πλαισίων από κατάλληλα χαρακτηρισμένα σώματα κειμένων

Formatted: Bullets and Numbering

~~9.~~ Βιβλιογραφία

Abney, S., 1990. Rapid Incremental Parsing with Repair. In *Proceedings of the 6th New OED Conference*, Electronic Text Research.

Abney, S., 1996. Partial Parsing via Finite-State Cascades. In *Proceedings of the Robust Parsing Workshop*, ESSLLI.

Abney, S., 1997. Part of Speech Tagging and Partial Parsing. In *Corpus-Based Methods in Language and Speech Processing*, Steve Young and Gerrit Bloothooft (eds.), Kluwer Academic Publishers, pp. 118-136.

Ait-Mokhtar, S. and J.P. Channod, 1997. Incremental Finite State Parsing. In *Proceedings of ANLP*, pp. 72-79.

Allen, James, 1995. *Natural Language Understanding*, Benjamin Cummings Publishing.

Appelt, D. and J. Hobbs, 1995. SRI International FASTUS System - MUC6 Test Results and Analysis. In *Proceedings of MUC6*.

Bourigault, D., 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of the 14th International Conference on Computational Linguistics*.

Boutsis, S., S. Piperidis and I. Demiros, 1999. Generating Bilingual Lexical Equivalences from Parallel Texts, *Applied Artificial Intelligence.*

Brill, E., 1993a. Transformation-Based Error-Driven Parsing. In *Proceedings of the 3rd International Workshop on Parsing Technologies.*

Brill, E., 1993b. *A Corpus-based Approach to Language Learning*, Doctoral Dissertation, University of Pennsylvania.

Di Christo, P., S. Harie, C. de Loupy, N. Ide, and J. Veronis, 1995. Set of Programs for Segmentation and Lexical Look up, Deliverable 2.2.1, MULTEXT, LRE 62-050.

Evans, D. and C. Zhai, 1996. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics.*

Evans, D., N. Milic-Frayling, and R. G. Lefferts, 1995. CLARIT TREC-4 Experiments. In *Proceedings of the 4th Text Retrieval Conference (TREC-4).*

Gavrilidou, M., Labropoulou, P., Mantzari, E. and S. Roussou. G*reek Lexicon Documentation.* Parole LE2-4017/10369, WP3.9-WP-ATH-1. 1998.

Grefenstette, G., 1996. Light Parsing as Finite State Filtering*.* In *Proceedings of Workshop on Extended Finite State Models of Language, ECAI.*

Grishman, R., 1995. The NYU System for MUC-6 or Where's the Syntax? In *Proceedings of MUC6.*

Hindle, D., 1983a. *User Manual for Fidditch.* Technical Memorandum #7590-142, Naval Research Laboratory.

Hindle, D., 1983b. Deterministic Parsing of Syntactic Non-Fluences. In *Proceedings of the 21st Annual Meeting of the Association of Computational Linguistics.*

Karlsson, F., A. Voutilainen, J. Hekkila, and A. Anttila, (eds.), 1995. *Constraint Grammar: a Language Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Karlsson, F., 1990. Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of 10th International Conference in Computational Linguistics*.

Karttunnen L. The Replace Operator. In *Finite State Language Processing*, ed. Roche Em. and Schabes Yv., MIT Press. 1997

Labropoulou, P., E. Mantzari and M. Gavrilidou. *Codification manual for the ILSP/LE-Parole tagset*. ILSP Internal Report. 1997.

Leech, G., R. Barnett, and P. Kahrel, 1996. *Provisional Recommendations and Guidelines for the Syntactic Annotation of Corpora*, EAGLES DOCUMENT EAG—TCWG—SASG/1.8.

Marcus, M., 1980. *A Theory of Syntactic Recognition for Natural Language*, MIT Press

Papageorgiou, H, 1996. Part of Speech Disambiguation. In *Hybrid Techniques for Bilingual Corpus Processing*, PhD dissertation, National Technical University of Athens.

Roche, E. and Y. Schabes (eds.), 1997. *Finite State Language Processing*. MIT Press

Satta, G. and E. Brill, 1996. Efficient Transformation-Based Parsing. In *Proceedings of the 34st Annual Meeting of the Association of Computational Linguistics.*

Stralkowski, T. and J. P. Carballo, 1995. Natural Language Information Retrieval: TREC-4 Report. In *Proceedings of the 4th Text Retrieval Conference (TREC-4).*

Van Noord, G. and Gerdemann D., 1999. *An Extendible Regular Expression Compiler for Finite-state Approaches in Natural Language Processing*. WIA, Potsdam, Germany

Voutilainen, A., 1993. NPtool, a Detector of English Noun Phrases. In *Proceedings of the Workshop on Very Large Corpora*.

Zhai, C., 1997. Fast Statistical Parsing of Noun Phrases for Document Indexing. In *Proceedings of the 5th Conference on Applied Natural Language Processing.*

Λέξεις - κλειδιά

εύρωστος συντακτικός αναλυτής, γραμματικές σχέσεις, ντετερμινιστική ανάλυση, αμφίσημες δομές, κανονικές εκφράσεις, πεπερασμένα αυτόματα ή πεπερασμένοι μεταγραφείς, αρχή του μέγιστου ταιριάσματος, χαρακτηριστικά υποκατηγοριοποίησης, πλαίσια υποκατηγοριοποίησης, σώμα κειμένων.

~~10.~~Παράρτημα

Formatted: Bullets and Numbering

	)SENT	<S>				
REF<948>	SYN	[cl				
REF<949>	SYN	[np_nm				
REF<950>	TOK		Η	ο	AtDfFeSgNm	atdfsgnm
REF<951>	SYN	[adjp_nm				
REF<952>	TOK		Αγροτική	αγροτικός	AjBaFeSgNm	ajbasgnm
REF<953>	SYN	/adjp_nm]				
REF<954>	TOK		Τράπεζα	τράπεζα	NoCmFeSgNm	nosgnm
REF<955>	SYN	/np_nm]				
REF<956>	SYN	[vg				
REF<957>	TOK		ενημέρωσε	ενημερώνω	VbMnIdPa03SgXxPeAvXx	vb
STRUCT<subj_np_nm,949,955>			STRUCT<cl_arg,971,1010>		STRUCT<compl_np_ac,959,965>	
REF<958>	SYN	/vg]				
REF<959>	SYN	[np_ac				
REF<960>	TOK		τους	ο	AtDfMaPlAc	atdfplac
REF<961>	SYN	[adjp_ac				
REF<962>	TOK		αρμόδιους	αρμόδιος	AjBaMaPlAc	ajbaplac
REF<963>	SYN	/adjp_ac]				
REF<964>	TOK		ερευνητές	ερευνητής	NoCmMaPlAc	noplac
					STRUCT<arg_np_ge,966,969>	
REF<965>	SYN	/np_ac]				
REF<966>	SYN	[np_ge				
REF<967>	TOK		της	ο	AtDfFeSgGe	atdfsgge
REF<968>	TOK		υπόθεσης	υπόθεση	NoCmFeSgGe	nosgge
REF<969>	SYN	/np_ge]				
REF<970>	SYN	/cl]				
REF<971>	SYN	[cl_o				
REF<972>	TOK		ότι	ότι	CjSb	cjsb_other
REF<973>	CHUNK		_	_		
REF<974>	SYN	[vg				
REF<975>	TOK		ανευρέθησαν	ανευρίσκω	VbMnIdPa03PlXxPePvXx	vb
					STRUCT<subj_np_nm,977,980>	
REF<976>	SYN	/vg]				
REF<977>	SYN	[np_nm				
REF<978>	DIG		33	33	DIG	dig
REF<979>	TOK		στελέχη	στέλεχος	NoCmNePlNm	noplnm
REF<980>	SYN	/np_nm]				
REF<981>	SYN	[cl_r				
REF<982>	TOK		που	που	PnReNe03PlNmXx	pn_pou
REF<983>	SYN	[vg				
REF<984>	TOK		είχαν	έχω	VbMnIdPa03PlXxIpAvXx	vb_exw
REF<985>	TOK		προχωρήσει	προχωρώ	VbMnNfXxXxXxXxPeAvXx	vb_inf
			STRUCT<pp_arg,987,995>		STRUCT<subj_np_nm,982,982>	
REF<986>	SYN	/vg]				
REF<987>	SYN	[pp				
REF<988>	TOK		σε	σε	AsPpSp	as_se
REF<989>	SYN	[np_ac				
REF<990>	SYN	[adjp_ac				
REF<991>	TOK		εικονικές	εικονικός	AjBaFePlAc	ajbaplac
REF<992>	SYN	/adjp_ac]				
REF<993>	TOK		προεγγραφές	προεγγραφή	NoCmFePlAc	noplac
REF<994>	SYN	/np_ac]				
REF<995>	SYN	/pp]				
REF<996>	SYN	/cl_r]				
REF<997>	SYN	/cl_o]				
REF<998>	PTERM_P		.	.	PTERM_P	punct_fs
	)SENT	</S>				