Π. Γιούλη, Ι. Δεμοίρος, Β. Αντωνόπουλος, Χ. Παπαγεωργίου, Σ. Πιπερίδης, Π. Αρβανίτης, Ε. Παπαναστασίου, Δ. Κακλαμανίδου, Ά. Χιδίρογλου, Α. Αλεξανδροπούλου, Σ. Κυλιντηρέα, Μ. Κοπιδάκης

"οικΟΝΟΜίΑ": ΕΝΑ ΣΩΜΑ ΣΧΟΛΙΑΣΜΕΝΩΝ ΟΙΚΟΝΟΜΙΚΩΝ ΚΕΙΜΕΝΩΝ

Abstract

This paper presents a corpus of financial texts which was collected and annotated at various linguistic levels of analysis within the framework of the EPET "oikONOMiA" project undertaken by the Institute for Language and Speech Processing, the University of Athens and the Aristotle University of Thessaloniki. The aim of this project was the development of an Information Extraction system involving large scale text processing. The annotated corpus was used for both development and evaluation purposes.

1. Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η παρουσίαση του σώματος κειμένων το οποίο συνελέγη και σχολιάστηκε σε διάφορα επίπεδα γλωσσικής ανάλυσης στα πλαίσια του προγράμματος ΕΠΕΤ "οικΟΝΟΜίΑ" για την ανάπτυξη και τον έλεγχο συστήματος εξαγωγής πληροφορίας. Το παρόν άρθρο περιγράφει τη διαδικασία συλλογής των κειμένων , τη μεθοδολογία και τις προδιαγραφές που τέθηκαν κατά το σχολιασμό ανά επίπεδο, καθώς επίσης και τα υπολογιστικά εργαλεία που χρησιμοποιήθηκαν για την επεξεργασία των κειμένων.

2. Το έργο "οικΟΝΟΜίΑ"

Αποτελεί γενική διαπίστωση ότι η τρέχουσα τεχνολογία στους τομείς της ανάκτησης και εξαγωγής πληροφοριών, δεν επιτρέπει την πλήρη εκμετάλλευση του όγκου των διαθέσιμων μέσω του διαδικτύου δεδομένων. Παρά το γεγονός ότι οι μηχανές αναζήτησης, που υπάρχουν αυτή τη στιγμή, επιτρέπουν την ανάκτηση πληροφοριών με απλοϊκό τρόπο, πολλές φορές αποτυγχάνουν να παρουσιάσουν στο χρήστη κείμενα σχετικά με το αντικείμενο της αναζήτησης ή «απαντούν» με κείμενα άσχετα με αυτό. Τα συστήματα ανάκτησης πληροφοριών, στην καλύτερη περίπτωση, βρίσκουν κείμενα σχετικά με μία ερώτηση πιθανών χρηστών τους, τα οποία, όμως πρέπει να διαβαστούν προκειμένου για την εξαγωγή της πληροφορίας. Αντίθετα, τα συστήματα εξαγωγής πληροφορίας βρίσκουν γεγονότα: ποιος προσλήφθηκε/ πού; ποια εταιρεία παράγει / ποιο προϊόν; Επομένως, απαιτούνται εργαλεία έξυπνης ανάκτησης και εξαγωγής της πληροφορίας, για τη βελτίωση της απόδοσης των μηχανών αναζήτησης από το διαδίκτυο και από τοπικές βάσεις δεδομένων.

Στην κατεύθυνση αυτή, το έργο "οικΟΝΟΜίΑ" υλοποιήθηκε από το Ινστιτούτο Επεξεργασίας Λόγου (ΙΕΛ) σε συνεργασία με το Εθνικό και Καποδιστριακό Παν/μιο Αθηνών (Τμήμα Μεθοδολογίας, Ιστορίας και Θεωρίας της Επιστήμης) και το Αριστοτέλειο Παν/μιο Θεσσαλονίκης (Τμήμα Γαλλικής Γλώσσας και Φιλολογίας), με

στόχο την ανάπτυξη τεχνικών που βρίσκουν εφαρμογή στην «ευφυή» αναζήτηση και εξαγωγή πληροφοριών. Ειδικότερα, το έργο ανέπτυξε εύρωστες μεθόδους που αποσκοπούν στην επιφανειακή κατανόηση ελεύθερου κειμένου, η οποία πραγματοποιείται με τη βοήθεια συστήματος αυτόματης ανάλυσης για την παραγωγή από ελεύθερο κείμενο μιας επιφανειακής σημασιολογικής αναπαράστασης, που περιλαμβάνει τις ακόλουθες πληροφορίες:

(i) μορφοσυντακτική κατηγοριοποίηση και λήμμα για κάθε λέξη,
(ii) ονοματικές οντότητες και κατηγοριοποίηση αυτών,
(iii) ανάλυση της επιφανειακής συντακτικής δομής κάθε πρότασης,
(iv) γραμματικές σχέσεις μεταξύ των συντακτικών δομών και εντός αυτών,
(v) συναναφορές μεταξύ ονομάτων, ονοματικών φράσεων και αντωνυμικών τύπων, χειρισμός ελλειπτικών φαινομένων.

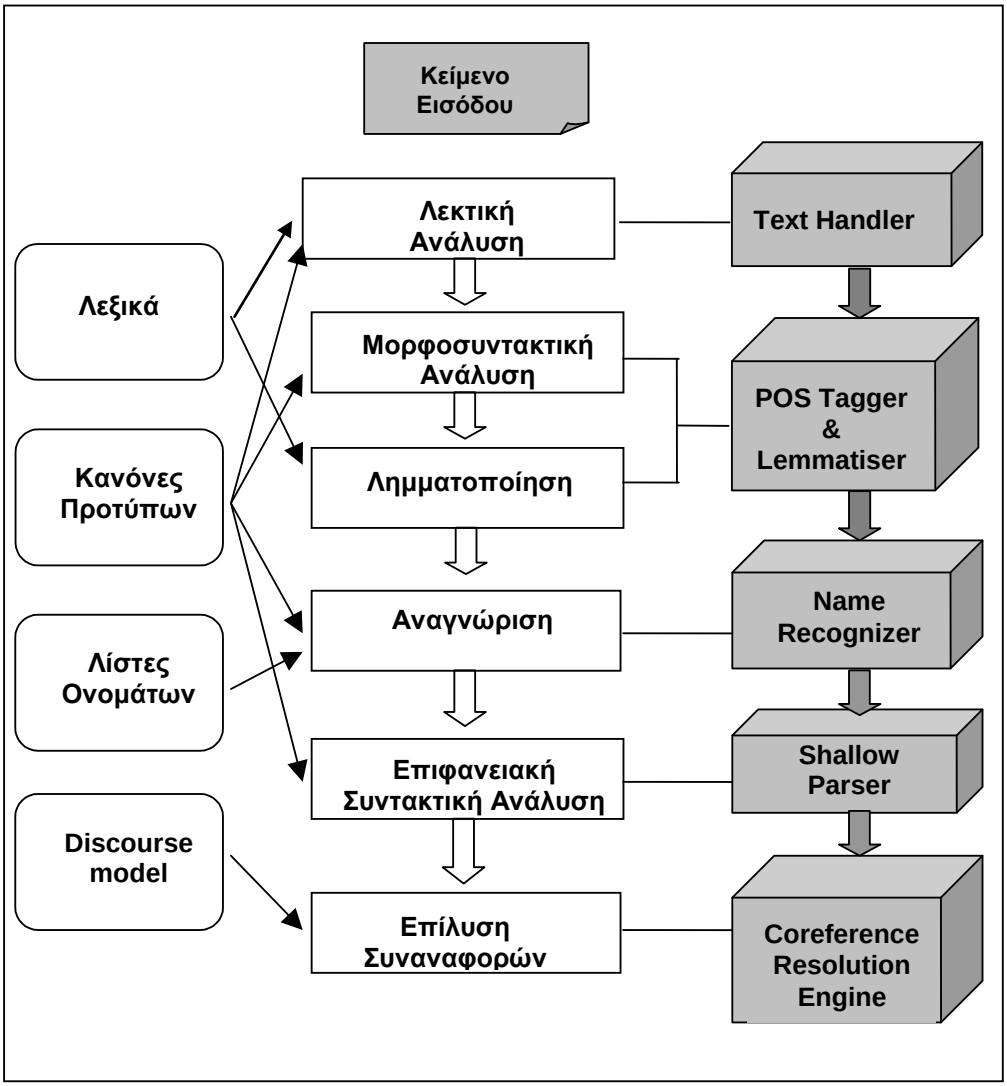
Η αναπαράσταση αυτή μπορεί να χρησιμοποιηθεί για την αποδοτική δεικτοδότηση κειμένων, η οποία βασίζεται όχι μόνο στην «επίπεδη» πληροφορία που παρέχουν οι συχνότητες των λέξεων-κλειδιών, αλλά και στη σύνθετη επεξεργασία σε γραμματικό και συντακτικό επίπεδο, της αναζήτησης του χρήστη. Παράλληλα, περιέχει πληροφορίες για την κατηγοριοποίηση των ονοματικών οντοτήτων και την επίλυση των συναναφορών.

3. Μεθοδολογία και Αρχιτεκτονική του Συστήματος

Οι εφαρμογές ευφυούς ανάκτησης πληροφοριών και εξαγωγής πληροφοριών αποτελούν εφαρμογές πραγματικού χρόνου και έντασης δεδομένων. Γιαυτό το λόγο, στη σχεδίαση και ανάπτυξη του συστήματος ελήφθησαν υπόψη παράγοντες, όπως: η ταχύτητα επεξεργασίας, η ευρωστία και η ακρίβεια των αποτελεσμάτων. Προκειμένου το σύστημα να μπορεί να εφαρμοστεί σε πραγματικά δεδομένα, δηλαδή ελεύθερα κείμενα, θα έπρεπε να χαρακτηρίζεται από ευρωστία, σε σχέση με τα φαινόμενα που εμφανίζονται στα ελεύθερα κείμενα του οικονομικού πεδίου. Προς αυτήν την κατεύθυνση, το σώμα κειμένων χρησιμοποιήθηκε τόσο για την ανάπτυξη και προσαρμογή των επιμέρους συστημάτων, όσο και για τον τελικό έλεγχο της αξιοπιστίας αυτών. Ο σχολιασμός των κειμένων στα διάφορα επίπεδα πραγματοποιήθηκε με ημι-αυτόματο τρόπο, δηλαδή με διόρθωση και χειρωνακτικό έλεγχο του σχολιασμού που παρήχθη αυτομάτως με τη βοήθεια των υποσυστημάτων.

Η ροή δεδομένων στο σύστημα φαίνεται στο σχήμα 1. Η επιφανειακή κατανόηση κειμένου πραγματοποιείται μέσα από μία ακολουθία αυτόματων διεργασιών. Η έξοδος μίας διεργασίας οδηγείται στην είσοδο της επόμενης. Η μέθοδος παίρνει στην είσοδο κείμενα και δίνει στην έξοδο μία επιφανειακή σημασιολογική αναπαράσταση του κειμένου, στην οποία σημειώνονται τα αποτελέσματα της ανάλυσης σε γραμματικό επίπεδο, συντακτικό επίπεδο και επίπεδο λόγου. Στο πρώτο στάδιο επεξεργασίας, αναγνωρίζονται τα επιφανειακά χαρακτηριστικά των κειμένων, όπως τα όρια των λέξεων, των προτάσεων, κλπ. Στη συνέχεια, η γραμματική ανάλυση και η λημματοποίηση υπολογίζουν το μέρος του λόγου και το λήμμα κάθε λέξης και η μερική συντακτική ανάλυση, αναγνωρίζει τα συντακτικά δομικά στοιχεία του κειμένου. Στο τελευταίο στάδιο επιλύονται οι

συναναφορές. Στο σχήμα 1 παρουσιάζεται η αντιστοιχία μεταξύ επιπέδων ανάλυσης, γλωσσικών πόρων και υπολογιστικών εργαλείων ανά επίπεδο ανάλυσης.



Σχήμα 1. Επίπεδα ανάλυσης: Πόροι και Εργαλεία

4. Το σώμα κειμένων

Η συλλογή των κειμένων πραγματοποιήθηκε είτε από διάφορες πηγές του διαδικτύου που αφορούν στον οικονομικό τομέα (κείμενα γραπτού λόγου) είτε με απομαγνητοφωνήσεις ραδιοφωνικών και τηλεοπτικών εκπομπών (προφορικός λόγος) οικονομικού, επίσης,

ενδιαφέροντος. Το σώμα κειμένων έχει συνολικό μέγεθος 425.000 περίπου μορφολογικά χαρακτηρισμένων λέξεων (morphological words), εκ των οποίων οι 304.130 αφορούν στον γραπτό λόγο, και οι υπόλοιπες 121.042 στον προφορικό. Εν συνεχεία πραγματοποιήθηκε η ταξινόμηση των κειμένων σε υποκατηγορίες αφενός με βάση το μέσο μετάδοσης (written, audio, visual) αλλά και το περιεχόμενο (εξαγορά εταιρείας, συγχώνευση εταιρειών, ίδρυση εταιρείας, σύναψη συμφωνίας, προκήρυξη γενικής συνέλευσης, αύξηση μετοχικού κεφαλαίου, ανακοίνωση κερδών χρήσης, αγορές χρήματος, μεταβολή επιτοκίων, ισοτιμιών, κλπ).

5. Προδιαγραφές Σχολιασμού

Ακολουθεί συνοπτική αναφορά στις προδιαγραφές που τέθηκαν ανά επίπεδο ανάλυσης. Να σημειωθεί ότι οι προδιαγραφές αυτές είναι καταρχήν συμβατές με αντίστοιχα διεθνή πρότυπα (Multext, Parole, EAGLES, MUC και MATE), με τις αναγκαίες τροποποιήσεις προκειμένου για την επαρκή περιγραφή των φαινομένων της ελληνικής.

5.1 Λεκτική Ανάλυση

Η λεκτική ανάλυση πραγματοποιήθηκε σύμφωνα με τη μεθοδολογία που υιοθετήθηκε στο πρόγραμμα Multext (Di Christo et. al, 1995), και αφορά στην αναγνώριση και το σχολιασμό των επιφανειακών βασικών δομικών συστατικών του κειμένου. Με άλλα λόγια, στο στάδιο αυτό λαμβάνει χώρα η αναγνώριση των ορίων των λέξεων και των περιόδων, και αναγνωρίζονται αλφαριθμητικά, ημερομηνίες και συντμήσεις. Γι' αυτό, είναι απαραίτητη η επίλυση της αμφισημίας όσον αφορά στο είδος των σημείων στίξης (τερματικό – μη τερματικό), 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 (Text Handler) 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. Το εισαγόμενο στο εργαλείο αυτό ήταν απλό κείμενο, ενώ ακολούθησε χειρωνακτικός έλεγχος, διόρθωση και προσθήκη του εξαγομένου, δηλ. του δομικά σχολιασμένου κειμένου.

5.2 Μορφοσυντακτική ανάλυση και Λημματοποίηση

Δεύτερο στάδιο επεξεργασίας του corpus αποτέλεσε ο μορφοσυντακτικός σχολιασμός του αρχικού σώματος κειμένων (Test or Raw Corpus), με τη βοήθεια ενός εργαλείου υποδομής του ΙΕΛ (Part-of-Speech Tagger & Lemmatizer). Στο στάδιο αυτό, για κάθε λέξη δίδεται το αντίστοιχο λήμμα, καθώς επίσης και τα (κατά περίπτωση) γραμματικά χαρακτηριστικά της, όπως: μέρος του λόγου, χαρακτηριστικά συμφωνίας, διάθεση, φωνή, κλπ. Η γραμματική ανάλυση για την ελληνική γλώσσα βασίζεται σε ένα μορφολογικό λεξικό και σε ένα σύστημα κανόνων για άρση της αμφισημίας. Τα γραμματικά χαρακτηριστικά δίδονται σε κάθε λέξη από την αντίστοιχη καταχώρηση του λεξικού. Σε περίπτωση που υπάρχουν περισσότερες της μίας γραμματικές αναλύσεις για μία λέξη, ο γραμματικός αναλυτής επιλέγει μία ανάλυση με βάση τα συμφραζόμενα. Στην περίπτωση που η λέξη δεν υπάρχει στο λεξικό, μία διαδικασία μορφολογικής ανάλυσης παράγει μία εκτίμηση

πιθανών χαρακτηρισμών. Η περίπτωση αμφισημίας αντιμετωπίζεται από ένα σύστημα άρσης της αμφισημίας με τη χρήση στατιστικών κανόνων. Συνολικά χρησιμοποιήθηκε ένα σύνολο ~670 γραμματικών χαρακτηριστικών (tagset) το οποίο είναι συμβατό με το αντίστοιχο πρότυπο PAROLE. Και σε αυτό το στάδιο, η εργασία πραγματοποιήθηκε με ημιαυτόματο τρόπο, καθώς το εξαγόμενο της αυτόματης διαδικασίας ελέγχθηκε εν συνεχεία από γλωσσολόγους.

5.3 Αναγνώριση Ονοματικών Οντοτήτων

Με τον όρο «ονοματικές οντότητες» (Named Entities) εννοούμε τα κύρια ονόματα προσώπων, τοπωνυμίων και οργανισμών, τις χρονικές εκφράσεις (Time Expressions) και τις αριθμητικές εκφράσεις (Numeric Expressions), που αντιστοιχούν στις κατηγορίες του MUC[1] **ENAMEX, TIMEX και NUMEX**. Οι κατηγορίες αυτές περιλαμβάνουν επιμέρους υποκατηγορίες, οι οποίες είναι και το ζητούμενο κατά την αναγνώριση και κατηγοριοποίηση των Ονοματικών Οντοτήτων. Ειδικότερα, η κατηγορία **ENAMEX** περιλαμβάνει τις εξής κατηγορίες:

PERSON, στην οποία εμπίπτουν κύρια ονόματα προσώπων
(π.χ. [person Ιωάννης Καποδίστριας/ person])
ORGANISATION, η οποία περιλαμβάνει ονόματα οργανισμών
(π.χ. [org Υπουργείο Εθνικής Οικονομίας/org])
LOCATION, που περιλαμβάνει τοπωνύμια
(π.χ. [loc Παράδεισος Αμαρουσίου/loc])

Ως Ονοματικές Οντότητες της κατηγορίας **TIMEX** χαρακτηρίζονται τόσο οι απόλυτες χρονικές εκφράσεις (absolute temporal expressions) (π.χ. 'Παρασκευή 23 Ιουλίου') όσο και οι σχετικές χρονικές εκφράσεις (relative temporal expressions) (π.χ. 'χθες') που δηλώνουν είτε συγκεκριμένη ώρα της ημέρας και ανήκουν στην υποκατηγορία **TIME**, είτε ημερομηνία και ανήκουν στην υποκατηγορία **DATE.**

Οι αριθμητικές εκφράσεις **(NUMEX)** μπορούν να είναι γραμμένες είτε αριθμητικά, είτε ολογράφως και διακρίνονται αφενός μεν σε εκφράσεις που δηλώνουν χρηματικό ποσό (monetary expressions) και χαρακτηρίζονται ως **MONEY**, αφετέρου δε σε εκφράσεις που δηλώνουν ποσοστό και χαρακτηρίζονται ως **PERCENT**.

Στα πλαίσια του έργου αναπτύχθηκε εργαλείο αναγνώρισης Ονοματικών Οντοτήτων (Named Entity Recognizer). Το σύστημα αυτό δέχεται στην είσοδο το κείμενο, το οποίο έχει περάσει από τα στάδια λεκτικού σχολιασμού και μορφοσυντακτικού σχολιασμού. Στο πρώτο στάδιο της αναγνώρισης, στο κείμενο σημειώνονται ονόματα που αντιστοιχούν σε γνωστές καταχωρήσεις καταλόγων ονομάτων. Στο δεύτερο στάδιο, το μερικώς χαρακτηρισμένο κείμενο διοχετεύεται σε υποσύστημα αναγνώρισης και κατηγοριοποίησης ονομάτων το οποίο κάνει χρήση κανόνων προτύπων γραμμένων στη μορφή μη

[1] Η αναγνώριση ονοματικών οντοτήτων αντιμετωπίζεται διεθνώς ως επί μέρους εργασία στην διαδικασία εξαγωγής πληροφορίας, κυρίως στο πλαίσιο των Διεθνών Συνεδρίων Αξιολόγησης Τεχνολογίας Εξαγωγής Πληροφορίας (Message Understanding Conference: MUC).

αναδρομικών κανονικών εκφράσεων και τεχνικών πεπερασμένων αυτομάτων (finite state techniques). Οι κανόνες είναι αριθμημένοι για να εφαρμόζονται με συγκεκριμένη σειρά και μεταφράζονται σε αυτόματα (finite state automata and tranducers) με γνωστές τεχνικές. Τα αυτόματα κάθε γραμματικής συνδέονται μεταξύ τους για να σχηματίσουν μία σειριακή ακολουθία (pipeline), που χαρακτηρίζει το κείμενο εισόδου με αυξητικό τρόπο. Κάθε κανόνας (κανονική έκφραση) περιγράφει ένα καθορισμένο φαινόμενο και κανόνες στα υψηλότερα επίπεδα περιγράφουν φαινόμενα με βάση τους ήδη διατυπωμένους κανόνες. Οι κανόνες είναι σχεδιασμένοι να είναι αξιόπιστοι όταν εφαρμόζονται με το μεγαλύτερο δυνατό ταίριασμά τους στο κείμενο (longest match), ώστε να μην υπάρχει ανάγκη για άρση της σχετικής αμφισημίας.

Ο σχολιασμός στο επίπεδο αυτό είναι συμβατός με τις προδιαγραφές του MUC-7 με προσαρμογή των σχετικών οδηγιών στα ελληνικά δεδομένα, και πραγματοποιήθηκε επίσης με ημι-αυτόματο τρόπο

5.4 Επιφανειακή Συντακτική Ανάλυση

Το επόμενο επίπεδο ανάλυσης συνίσταται στην αναγνώριση φραστικών κατηγοριών σύμφωνα με το πρότυπο EAGLES. Σε αυτό, αναγνωρίστηκαν και χαρακτηρίστηκαν ονοματικές, επιθετικές, προθετικές, επιρρηματικές και ρηματικές φράσεις, οι κεφαλές τους και τυχόν προσδιορισμοί. Επίσης σημειώθηκαν τα όρια των προτάσεων (clauses) και οι κατηγορίες τους. Γι αυτό το σκοπό χρησιμοποιήθηκε ο επιφανειακός συντακτικός αναλυτής του ΙΕΛ (Shallow Parser) με γραμματική για τα ελληνικά. Η ανάλυση λαμβάνει χώρα με τη χρήση τεχνικών πεπερασμένων αυτομάτων, με τη μέθοδο που περιγράφηκε στην προηγούμενη παράγραφο (βλ. Αναγνώριση και Κατηγοριοποίηση Ονοματικών Οντοτήτων). Η ανάλυση που πραγματοποιήθηκε είναι επιφανειακή και δεν ασχολήθηκε με φαινόμενα επαναδρομής (recursion). Αναγνωρίστηκαν οι παρακάτω φραστικές κατηγορίες:

np_nm: ΟΦ σε ονομαστική (nominative NP)
np _ge: ΟΦ σε γενική (genitive NP)
np _ac: ΟΦ σε αιτιατική (accusative NP)
np _vo: ΟΦ σε κλητική (vocative NP)
np _da: ΟΦ σε δοτική (dative NP)
adjp_nm: ΕΠΙΘ_ΦΡ σε ονομαστική (nominative ADJP)
adjp_ge: ΕΠΙΘ_ΦΡ σε γενική (genitive ADJP)
adjp_ac: ΕΠΙΘ_ΦΡ σε αιτιατική (accusative ADJP)
adjp_vo: ΕΠΙΘ_ΦΡ σε κλητική (vocative ADJP)
adjp_da: ΕΠΙΘ_ΦΡ σε δοτική (dative ADJP)
pp: ΠΡΟΘ_ΦΡ (prepositional phrase)
advp: ΕΠΙΡΡ_ΦΡ (adverb phrase)**.**
vg: Ρηματική Ομάδα με θέμα οριστικής (Verb Group base)
vg_s: (Verb Group in Subjunctive)
vg_g: (Verbal Group Gerund)

Επίσης αναγνωρίστηκαν τα εξής είδη προτάσεων (clauses):

cl : κύρια πρόταση (main clause)
cl_c : δευτερεύουσα υποθετική πρόταση (conditional clause)
cl_g : δευτερεύουσα γερουνδιακή πρόταση (gerund clause)
cl_ir : δευτερεύουσα ερωτηματική πρόταση (interrogative clause)
cl_q : κύρια ερωτηματική πρόταση (main clause in question form)
cl_r : δευτερεύουσα αναφορική πρόταση (relative clasuse)
cl_ri : δευτερεύουσα αόριστη αναφορική (relative indefinite)
cl_t : δευτερεύουσα χρονική πρόταση (temporal clause)
cl_o : άλλη πρόταση (other types of clauses)

Δείγμα σχολιασμένου κειμένου μέχρι και το στάδιο αυτό δίνεται στο Παράρτημα.

5.5 Σχολιασμός αναφορικών σχέσεων

Τέλος, στο σώμα κειμένων πραγματοποιήθηκε ο σχολιασμός συναναφοράς, ο οποίος συνίσταται αφενός μεν στην αναγνώριση των δομικών στοιχείων του κειμένου που εισάγουν οντότητες στο λόγο (discourse entities) και που δυνάμει μετέχουν σε σχέσεις συναναφοράς, και αφετέρου στον εντοπισμό των στοιχείων που όντως μετέχουν σε σχέσεις συναναφοράς και στη δήλωση των μεταξύ τους σχέσεων. Η κατ΄ αυτόν τον τρόπο επεξεργασία των κειμένων προϋποθέτει την ύπαρξη ενός *μοντέλου του λόγου* (discourse model) στο οποίο περιλαμβάνονται *οι οντότητες του λόγου* (discourse entities) και *οι αναφορικές σχέσεις* μεταξύ αυτών. Το μοντέλο του λόγου περιλαμβάνει στοιχεία τα οποία στο στάδιο της συντακτικής ανάλυσης έχουν λάβει το χαρακτηρισμό ΟΦ, με εξαίρεση τις προσωπικές αντωνυμίες πρώτου και δευτέρου προσώπου, ερωτηματικές και αόριστες αντωνυμίες, δεδομένου ότι ούτε εισάγουν, ούτε αναφέρονται ρητά σε κάποια οντότητα του λόγου, αλλά και ΟΦ με ποσοδείκτη, δεδομένου ότι δεν αναφέρονται σε κάποια συγκεκριμένη οντότητα του λόγου. Επομένως, στοιχεία τα οποία είναι μεν αναφορικά - με την έννοια ότι λειτουργούν ως σημείο αναφοράς κάποιου αναφορικού – αλλά δεν είναι ΟΦ (όπως για παράδειγμα μία ολόκληρη πρόταση ή φράση) δεν μετέχουν του discourse model. Το σχήμα σχολιασμού βασίστηκε στο πρότυπο MATE προσαρμοσμένο στα ελληνικά δεδομένα. Σύμφωνα με το ακολουθούμενο πρότυπο, ο σχολιασμός συνίσταται σε γλωσσολογικά *μεταδεδομένα* με τη μορφή δομημένων ζευγών μεταβλητών/τιμών (attribute/value pairs), η συνοπτική περιγραφή των οποίων ακολουθεί. Συγκεκριμένα, στο στάδιο του χαρακτηρισμού των υποψήφιων να μετέχουν σε σχέσεις συναναφοράς οντοτήτων για κάθε *επιλέξιμο στοιχείο* (markable) δίνονται τα ακόλουθα μεταγλωσσικά δεδομένα:

<markable>: χρησιμοποιείται για το χαρακτηρισμό κάθε αλφαριθμητικού στοιχείου το οποίο κρίνεται επιλέξιμο προς περαιτέρω σχολιασμό στο παρόν επίπεδο. Προστίθεται αυτομάτως, με την επιλογή του εν λόγω κειμενικού στοιχείου, και παραμένει χωρίς τιμή.

Ο χαρακτηρισμός, επιπλέον, περιλαμβάνει τις τιμές των εξής μεταβλητών :

id: ο μοναδικός αριθμός κάθε στοιχείου στο οποίο δίδεται το χαρακτηριστικό <markable>.
type: το είδος του επιλεγμένου <markable>. Μπορεί να λάβει μία από τις ακόλουθες τιμές:
Np-def (ΟΦ με οριστικό άρθρο)
Np-indef (ΟΦ με αόριστο άρθρο)

Np-bare (ΟΦ χωρίς οριστικό/αόριστο άρθρο)
Empty-Np (ΟΦ-Κενός αντωνυμικός τύπος)
Np-pn (ΟΦ αντωνυμικός τύπος)
Np-conj (ΟΦ παρατακτικά συνδεδεμένες)

subtype: η υποκατηγορία στην οποία ανήκει το επιλεγμένο <markable>. Η μεταβλητή Subtype περιλαμβάνει πληροφορία σχετικά με την εσωτερική *δομή* της επιλεγμένης ενότητας, και δέχεται τιμή μόνο στην περίπτωση που η μεταβλητή Type έχει λάβει μία από τις ακόλουθες τιμές: np_def, np_indef, np_bare. Στις περιπτώσεις αυτές, πιθανές τιμές της μεταβλητής μπορεί να είναι :

Np_base (ΟΦ χωρίς συμπλήρωμα- προσδιορισμό ετερόπτωτο ή εμπρόθετο)
Np_num (ΟΦ με αριθμητικό)
Np_dem (ΟΦ με δεικτική αντωνυμία)
Np_app (ΟΦ με προσδιορισμό παραθετική δομή)
Np_compl (ΟΦ με συμπλήρωμα και/ή προσδιορισμό ετερόπτωτο ή εμπρόθετο)
Np_pred (ΟΦ σε θέση κατηγορουμένου)
Np_rel (ΟΦ με συμπλήρωμα αναφορική πρόταση)
Np_null

5.5.1 Σχολιασμός αναφορικών σχέσεων: Σύνδεση αναφορικών με το σημείο αναφοράς τους

Την επιλογή και χαρακτηρισμό των κειμενικών στοιχείων τα οποία δυνάμει μετέχουν σε σχέσεις συναναφοράς, ακολούθησε ο εντοπισμός και η διασύνδεση εκείνων των στοιχείων που μετέχουν σε τέτοιες σχέσεις, καθώς επίσης και η δήλωση της σχέσης μεταξύ αναφορικού και σημείου αναφοράς. Να σημειώσουμε στο σημείο αυτό ότι τα βασικά είδη των σχέσεων συναναφοράς που αναφέρονται στη βιβλιογραφία είναι τα ακόλουθα:

Σχέση ταυτότητας *(identity relation):* πραγματώνεται μεταξύ δύο οντοτήτων του λόγου ο οποίες ορίζουν το ίδιο αντικείμενο στο μοντέλο του λόγου. Είναι δυνατόν σε ένα κείμενο να γίνονται διαδοχικές αναφορές σε ένα αντικείμενο, και οι αναφορές αυτές να πραγματώνονται φωνολογικά είτε με την επανάληψη της ίδιας λεξικής μονάδας ή με διαφορετική. Στην δεύτερη περίπτωση αυτό επιτυγχάνεται με τη χρήση αντωνυμιών (ρητώς εκπεφρασμένων ή pro), ή με άλλη ΟΦ η οποία εκφράζει ιδιότητα, τίτλο, μετωνυμική αναφορά, κλπ.

Σχέση στοιχείου-συνόλου *(set-member relation)*: πραγματώνεται όταν μία οντότητα του λόγου αποτελεί στοιχείο ενός συνόλου, το οποίο με τη σειρά του αντιστοιχεί σε μια άλλη οντότητα του λόγου. Όταν η σχέση μαρκάρεται η σειρά των όρων έχει ιδιαίτερη σημασία Πρέπει να σημειωθεί ότι η σχέση αυτή δηλώνεται είτε εμφανίζεται πρώτο στο λόγο το στοιχείο του συνόλου είτε το ίδιο το σύνολο. Στο παράδειγμα που ακολουθεί, **η Fiat** και **η Alcatel** αποτελούν στοιχεία του ευρύτερου συνόλου **πελάτες**:

<markable id="1_01">**οι πελάτες της**</markable>, όπως
<markable id="1_02"> **η Fiat** </markable>,
<markable id="1_03"> **η Alcatel** </markable>

Σχέση συνόλου-υποσυνόλου ή μέρους-όλου (whole-part relation): πραγματώνεται όταν μία οντότητα του λόγου αντιστοιχεί σε ένα υποσύνολο ενός συνόλου, που ορίζεται από μία

άλλη οντότητα του λόγου. Στο παράδειγμα που ακολουθεί, **η Δ/νση Διαχείρισης Ειδικών Πελατών** αποτελεί ένα σύνολο, το οποίο αποτελεί υποσύνολο του ευρύτερου συνόλου **Ο ΟΤΕ:**

<markable id="1_01">**Ο ΟΤΕ** </markable> αναλαμβάνει μέσω
<markable id="1_02"> **της Δ/νσης Διαχείρισης Ειδικών Πελατών**
</markable>,

Σχέση κτήσης – κτήτορα (possession relation): πραγματώνεται όταν μια οντότητα του λόγου αποτελεί τον κτήτορα μιας άλλης οντότητας. Ο κτήτορας συνήθως δηλώνεται με κτητική αντωνυμία ή με γενική. Παράδειγμα:

Η πλειοψηφία <markable id="1_02"> **των εσόδων**
<markable id="1_03"> **της** </markable> </markable>

Στα πλαίσια του έργου ασχοληθήκαμε με τις αναφορικές σχέσεις ταυτότητας. Ικανή και αναγκαία συνθήκη για τον χαρακτηρισμό μιας σχέσης ως σχέσης ταυτότητας είναι να ικανοποιούνται οι ακόλουθες ιδιότητες:

η **μεταθετική ιδιότητα:** αν μια οντότητα Α συνδέεται με σχέση ταυτότητας με μια οντότητα Β και η Β με μια οντότητα Γ, τότε και η Α συνδέεται με σχέση ταυτότητας με τη Γ. Για το λόγο αυτό, δεν έχει ιδιαίτερη σημασία ποιο στοιχείο θα επιλεγεί ως σημείο αναφοράς, και, επιπλέον

η **συμμετρική ιδιότητα,** δηλαδή αν μία οντότητα Α συναναφέρεται με μία οντότητα Β, τότε και η Β συναναφέρεται με την Α. Επομένως, το παρόν σχήμα σχολιασμού δεν προβλέπει μεταβλητή για τη δήλωση του στοιχείου που αποτελεί το σημείο αναφοράς.

Επιπλέον, σχέση ταυτότητας ορίζει **κλάσεις ισοδυναμίας (chains)**, κάθε μία από τις οποίες περιλαμβάνει όλες τις περιπτώσεις συναναφοράς για κάθε οντότητα, ενώ δεν επιτρέπεται δύο διαφορετικές κλάσεις να έχουν κοινά στοιχεία μεταξύ τους. Η δήλωση των σχέσεων συναναφοράς μεταξύ κειμενικών στοιχείων που ορίζουν οντότητες του λόγου επιτυγχάνεται με την διασύνδεση των στοιχείων αυτών και την υπαγωγή τους σε μία διακριτή κλάση ισοδυναμίας. Ο χαρακτηρισμός στο επίπεδο αυτό συνίσταται στην επισήμανση των κλάσεων ισοδυναμίας που εντοπίζονται σε ένα κείμενο, και στην δήλωση των **ζευγών οντοτήτων** τα οποία περιλαμβάνονται σε κάθε κλάση ισοδυναμίας του κειμένου και τα οποία αποτελούνται από οντότητες του λόγου οι οποίες συναναφέρονται. Τα γλωσσικά μεταδεδομένα που δίδονται στο επίπεδο αυτό είναι:

<chain>: χρησιμοποιείται για τη δήλωση ύπαρξης κλάσης ισοδυναμίας. Δημιουργείται αυτομάτως και συνοδεύεται από ζεύγη μεταβλητών/τιμών:

id: ο αριθμός που δίνεται σε κάθε σχέση ισοδυναμίας. Είναι μοναδικός για κάθε μία και προστίθεται αυτομάτως

type: ο τύπος της σχέσης μεταξύ του αναφορικού και του σημείου αναφοράς του. Πιθανές τιμές είναι: σχέση ταυτότητας, συνόλου-υποσυνόλου και κτήσης κτήτορα. Στα πλαίσια του

έργου σημειώθηκαν μόνο στοιχεία που συνδέονται με σχέση ταυτότητας, παίρνοντας πάντα την τιμή **"Identity".**

<link>: χρησιμοποιείται για τη δήλωση των στοιχείων ενός ζεύγους οντοτήτων (ενός αναφορικού και του σημείο αναφοράς του) τα οποία συνδέονται με σχέση ισοδυναμίας. Περιέχει την εξής μεταβλητή:

id: ο αριθμός που δίνεται σε κάθε ζεύγος οντοτήτων μιας κλάσης ισοδυναμίας. Είναι μοναδικός για το κάθε ζεύγος και προστίθεται αυτομάτως

Παράδειγμα Σχολιασμού:

<markable> id="1_01">
<markable> id="1_02">**ο ΟΤΕ** </markable> **και**
<markable> id="1_03">**η KPN** </markable></markable> δημιούργησαν κοινοπραξία….
<markable> id="2_01"> **Η KPN** </markable> είναι μια από τις μεγαλύτερες εταιρείες τηλεπικοινωνιών…….
<markable> id="4_01"> **Οι δύο εταιρείες** </markable> έχουν εκφράσει επανειλημμένως την πάγια πρόθεσή <markable> id="4_03">**τους** </markable>

6. Συμπεράσματα

Στο άρθρο αυτό παρουσιάστηκε το σώμα οικονομικών κειμένων του έργου "οικΟΝΟΜίΑ", και ο σχολιασμός που πραγματοποιήθηκε σε διάφορα επίπεδα γλωσσικής ανάλυσης για την ανάπτυξη και τον έλεγχο συστήματος εξαγωγής πληροφορίας. Παράλληλα με την ανάπτυξη υποσυστημάτων επεξεργασίας φυσικής γλώσσας, το έργο ανέπτυξε υπολογιστικά εργαλεία φιλικά προς το χρήστη για τον σχολιασμό των κειμένων σε όλα τα επίπεδα που περιγράφηκαν.

Βιβλιογραφία

Aone Chinatsu, Halverson Lauren, Hampton Tom, Ramos-Santacruz Mila *SRA: Description of the IE^2 system used for MUC-7*, MUC-7 Proceedings, 1998.
Black W., Rinaldi F., Mowatt D. Facile: description of the NE system used for MUC-7. Proceedings of Seventh Message Understanding Conference, 1998.
Borthwick A., Sterling J., Agichtein E., Grishman R. Description of the MENE Named Entity System as used in MUC-7. Proceedings of Seventh Message Understanding Conference, 1998.
Brill E. A corpus-based approach to language learning. Doctoral Dissertation, Univ. of Pennsylvania, 1993.
Chinchor N., MUC-7 Named Entity Task Definition, Version 3.5, 1997.
Ido Dagan, Alon Itai, *Automatic Processing of Large Corpora for the Resolution of Anaphoric References*. In: Proceedings of the 13th International Conference on Computational Linguistics (COLING), Vol. 3, Helsinki, 1990, 330-332.

Di Christo, P., S. Harie, C. De Loupy, N. Ide, and J. Veronis. Set of programs for segmentation and lexical look up, MULTEXT LRE 62-050 project Deliverable 2.2.1, 1995.
Gallippi A., Learning to recognize names across languages. Proceedings of the 16th International Conference on Computational Linguistics, 1996.
Garigliano Roberto, Urbanowicz Agnieszka and Nettleton David J. *University of Durham: Description of the LOLITA system as used in MUC-7*, MUC-7 Proceedings, 1998.
Gavrilidou, M., Labropoulou, P., Mantzari, E. and S. Roussou. G*reek Lexicon* Documentation. Parole LE2-4017/10369, WP3.9-WP-ATH-1. 1998.
Grishman R., Tipster architecture design document version 2.3. Technical report, DARPA, 1997.
Barbara J. Grosz, Aravind K. Joshi, Scott Weinstein. *Providing a unified Account of definite Noun Phrase in Discourse*. In: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics. Cambridge, Mass., Juni 1983, 44-50.
Barbara J. Grosz, Candace L. Sidner. *Attention, Intentions, and the Structure of Discourse.* In: Computational Linguistics, Vol. 12, Number 3, July-September 1986, 175-204.
Barbara J. Grosz, Aravind K. Joshi, Scott Weinstein. *Centering: A Framework for Modeling the Local Coherence of Discourse.* In: Computational Linguistics. Vol. 21, Nr. 2, Juni 1995. 203-225.
Hirschman Lynette, MUC-7 Conference Task Definition, Version 3.0.
Humphreys K., Gaizauskas R., Azzam S., Huyck C., Mitchell B., Cunningham H., Wilks Y. *University of Sheffield: Description of the LaSIE System as used for MUC-7*, MUC-7 Proceedings, 1998.
Karkaletsis V., Spyropoulos C., Petasis G. Named entity recognition from Greek texts: the GIE project, 1999.
Labropoulou, P., E. Mantzari and M. Gavrilidou. *Codification manual for the ILSP/LE-Parole tagset*. ILSP Internal Report. 1997.
Pazienza M.T. Ed., *Information Extraction: Multidisciplinary contributions to an emerging Information Technology*, Lecture Notes in Artificial Intelligence 1299, Springer-Verlag, Berlin Heidelberg, 1997.
Van Noord Gertjan and Dale Gerdemann. An Extendible Regular Expression Compiler for Finite-state Approaches in Natural Language Processing. WIA, Potsdam, Germany, 1999.
Yangarber Roman and Grishman Ralph *NYU: Description of the Proteus/PET system as used for MUC-7*, MUC-7 Proceedings, 1998.
Yu S., Bai S., Wu P. Description of the Kent Ridge Digital Labs system used for MUC-7. Proceedings of Seventh Message Understanding Conference, 1998.
Ελίνα Δεσύπρη, Ειρήνη Σπυροπούλου, *Επίλυση Συναναφορών*, Διπλωματική Εργασία, Τεχνογλωσσία, Πανεπιστήμιο Αθηνών και ΕΜΠ, 2000.

Λέξεις – Κλειδιά

Σώμα οικονομικών κειμένων – λεκτική ανάλυση – μορφοσυντακτικός σχολιασμός – συντακτική ανάλυση – συναναφορά – αναφορικές σχέσεις.

Παράρτημα: Δείγμα Σχολιασμένου Κειμένου

(SENT	<S>				
SYN	[cl				
SYN	[np_nm				
TOK		Η	ο	AtDfFeSgNm	atdfsgnm
NE	*[org*				
SYN	[adjp_nm				
TOK		Αγροτική	αγροτικός	AjBaFeSgNm	ajbasgnm
SYN	/adjp_nm]				
TOK		Τράπεζα	τράπεζα	NoCmFeSgNm	nosgnm
SYN	/np_nm]				
NE	*/org]*				
SYN	[vg				
TOK		ενημέρωσε	ενημερώνω	VbMnIdPa03SgXxPeAvXx	Vb
SYN	/vg]				
SYN	[np_ac				
TOK		τους	ο	AtDfMaPlAc	atdfplac
SYN	[adjp_ac				
TOK		αρμόδιους	αρμόδιος	AjBaMaPlAc	
SYN	/adjp_ac]				
TOK		ερευνητές	ερευνητής	NoCmMaPlAc	noplac
SYN	/np_ac]				
SYN	[np_ge				
TOK		της	ο	AtDfFeSgGe	atdfsgge
TOK		υπόθεσης	υπόθεση	NoCmFeSgGe	nosgge
SYN	/np_ge]				
SYN	/cl]				
SYN	[cl_o				
TOK		ότι	ότι	CjSb	cjsb_other
CHUNK	_	_			
SYN	[vg				
TOK		ανευρέθησαν	ανευρίσκω	VbMnIdPa03PlXxPePvXx	vb
SYN	/vg]				
SYN	[np_nm				
DIG		33	33	DIG	dig
TOK		στελέχη	στέλεχος	NoCmNePlNm	noplnm
SYN	/np_nm]				
SYN	[cl_r				
TOK		που	που	PnReNe03PlNmXx	pn_pou
SYN	[vg				
TOK		είχαν	έχω	VbMnIdPa03PlXxIpAvXx	vb_exw
TOK		προχωρήσει	προχωρώ	VbMnNfXxXxXxXxPeAvXx	vb_inf
SYN	/vg]				
SYN	[pp				
TOK		σε	σε	AsPpSp	as_se
SYN	[np_ac				
TOK		προεγγραφές	προεγγραφή	NoCmFePlAc	noplac
SYN	/np_ac]				
SYN	/pp]				
SYN	/cl_r]				
SYN	/cl_o]				
PTERM_P	.	.	PTERM_P	punct_fs	
)SENT	</S>				