ΕΛΕΤΟ – 7ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία»
Αθήνα, 22-24 Οκτωβρίου 2009

Κατασκευή σώματος κειμένων οικονομικού περιεχομένου και συμφραστικών πινάκων

Τζιάφα Ελένη

ΠΕΡΙΛΗΨΗ

Η έρευνα συνίσταται στη δημιουργία ενός ηλεκτρονικού σώματος κειμένων που προέρχονται από το χώρο της οικονομίας και ειδικότερα του χρηματιστηρίου, με στόχο την κατασκευή συμφραστικών πινάκων και την προσθήκη σημασιολογικών χαρακτηριστικών στα λήμματα του λεξικού των όρων του χρηματιστηρίου. Πιο συγκεκριμένα το Σώμα Κειμένων Χρηματιστηριακού Περιεχομένου, η έκταση του οποίου ανέρχεται σε δύο περίπου εκατομμύρια λέξεις, περιλαμβάνει κείμενα που αντλήθηκαν από τον ημερήσιο και περιοδικό τύπο, καθώς και από κείμενα που βρίσκονται σε ηλεκτρονική μορφή και αφορούν το χρηματιστήριο (π.χ. ενημερωτικά δελτία κ.ά.), σε μια χρονική περίοδο τριών ετών (1999-2001). Για τη δημιουργία του σώματος κειμένων και των συμφραστικών πινάκων χρησιμοποιήθηκαν τα προγράμματα κωδικοποίησης και αυτόματης κλίσης που έχουν χρησιμοποιηθεί για την κατασκευή ηλεκτρονικών λεξικών από τη Μονάδα Αυτόματης Επεξεργασίας Φυσικών Γλωσσών (http://linginfo.frl.auth.gr) του Εργαστηρίου Μετάφρασης και Επεξεργασίας του Λόγου (Ε.Μ.Ε.Λ.) του Α.Π.Θ. σε συνεργασία με το Institut Gaspard-Monge του Πανεπιστημίου της Marne-la-Vallée, Paris-Est.

Construction of an electronic corpus of financial texts and concordances

Tziafa Eleni

ABSTRACT

This paper is about the construction of an electronic corpus of financial and stock market texts, aiming at building a concordances table and adding semantic characteristics to the entries of the dictionary of stock market terms. In particular, the Stock Exchange Text Corpus, with more than two million words, includes articles, which are derived from daily and weekly press and also financial texts in electronic form, such as annual reports, etc, during a period of three years (1999-2001). For the construction of the corpus and the concordances, the computational linguistics team of the Natural Language Processing Unit (http://linginfo.frl.auth.gr) of the Laboratory of Translation and Language Processing, Aristotle University of Thessaloniki, in association with the Gaspard-Monge Institute, University of Marne-la-Vallée, Paris-Est, used the same codification and automatic inflection programs that were used for the creation of electronic dictionaries. The research was conducted by the said team, whose studies aim at a detailed and formalized description of Modern Greek, the final objective being the recognition of linguistic data by natural language processing systems.

0 Εισαγωγή

Η εργασία εντάσσεται στο γενικότερο πλαίσιο της έρευνας για την ανάλυση της γλώσσας με τη βοήθεια των ηλεκτρονικών υπολογιστών. Πιο συγκεκριμένα, εντάσσεται στην προσπάθεια της γλωσσολογικής ανάλυσης κειμένων κατά τρόπο τέτοιο, ώστε τα αποτελέσματά της να μπορούν να εφαρμοστούν σε οποιοδήποτε λογισμικό σύστημα αυτόματης ανάλυσης και επεξεργασίας φυσικών γλωσσών και κατ' επέκταση αυτόματης μετάφρασης. Για τη διεξαγωγή της έρευνας ακολουθήσαμε τις μεθοδολογικές αρχές και το θεωρητικό πλαίσιο του Maurice Gross (1975) που εφαρμόστηκε στο Laboratoire d'Automatique et Documentaire Linguistique (L.A.D.L.[1]). Η έρευνα αυτή εντάσσεται στο πρόγραμμα για την ανάπτυξη και τη σταδιακή ολοκλήρωση της ελληνικής έκδοσης του Unitex[2] - συστήματος ανάλυσης κειμένων που ήδη λειτουργεί σε πολλές ευρωπαϊκές γλώσσες.

Ο σχεδιασμός ενός εξειδικευμένου σώματος κειμένων για μία από τις λιγότερο ομιλούμενες γλώσσες, όπως είναι η ελληνική, αποτελεί σημαντικό παράγοντα για την ανάπτυξη της έρευνας γλωσσικής τεχνολογίας. Στην παρούσα ανακοίνωση θα παρουσιάσουμε το σχεδιασμό και την κατασκευή ενός σώματος κειμένων οικονομικού και συγκεκριμένα χρηματιστηριακού περιεχομένου. Η εργασία αυτή αποτελεί συνέχεια της κατασκευής ενός ελληνικού μορφολογικού λεξικού χρηματιστηριακών όρων 10.000 περίπου λέξεων (Τζιάφα, 2007).

Ειδικότερος στόχος της δημιουργίας ενός σώματος κειμένων χρηματιστηριακού περιεχομένου είναι να διερευνηθούν και να καταγραφούν τα ιδιαίτερα συντακτικά και σημασιολογικά χαρακτηριστικά των όρων αυτών και να οργανωθούν σε λεξικά, έτσι ώστε να συνδέονται αυτόματα με τους όρους που περιέχονται στα κείμενα.

Κατά τη διάρκεια των τελευταίων ετών, η ορολογική έρευνα έχει στραφεί στα σώματα κειμένων για τον εντοπισμό και την εξαγωγή όρων, καθώς και την ανάπτυξη υπολογιστικών εργαλείων για την επεξεργασία, την τεκμηρίωση και την αξιολόγηση των όρων. Ως σώμα κειμένων ορίζεται «κάθε συλλογή τμημάτων μιας συγκεκριμένης γλώσσας σε ηλεκτρονική μορφή, τα οποία έχουν επιλεγεί με εξωτερικά κριτήρια έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα μιας γλώσσας ή μιας γλωσσικής ποικιλίας» (Sinclair, 2005).

---

[1] Βλ. ιστοσελίδα: http://ladl.univ-mlv.fr/index.html.

[2] Το Unitex είναι ένα γλωσσικό περιβάλλον ανάπτυξης, ένα σύνολο λογισμικών προγραμμάτων, που χρησιμοποιείται για την εκτενή περιγραφή των ζωντανών γλωσσών και επιτρέπει τη διαχείριση κειμένων με τη χρήση γλωσσικών πόρων. Αυτές οι πηγές αποτελούνται από ηλεκτρονικά λεξικά, γραμματικές και λεξικά-γραμματικές. Κατασκευάστηκε από τον S.Paumier το 2002 στο Ινστιτούτο Gaspard-Monge του Πανεπιστημίου της Marne-la-Vallée στη Γαλλία.

Το μεγαλύτερο πλεονέκτημα της χρήσης σωμάτων κειμένων στη λεξικογραφία αποτελεί η φύση τους ως ηλεκτρονικών κειμένων που επιτρέπει την εξαγωγή αυθεντικών παραδειγμάτων χρήσης των όρων ενός λεξικού μέσα σε λίγα μόνο λεπτά. Είναι επίσης δυνατή η εξαγωγή ποσοτικών στοιχείων που σχετίζονται με τις συχνότητες εμφάνισης όρων. Για παράδειγμα, μπορούμε άμεσα να παρατηρήσουμε ότι στο σώμα κειμένων η συχνότητα εμφάνισης των ξένων χρηματιστηριακών όρων (με λατινικούς χαρακτήρες, χωρίς να υπολογίζουμε δηλαδή τους ξένους όρους που μεταγράφονται με ελληνικούς χαρακτήρες) ανέρχεται στο 4% του συνολικού αριθμού εμφάνισης των όρων στο σώμα κειμένων.

1 Περιγραφή του σώματος κειμένων χρηματιστηριακού περιεχομένου

Το σώμα κειμένων χρηματιστηριακού περιεχομένου στο οποίο αναφερόμαστε είναι ένα ηλεκτρονικό σώμα κειμένων ειδικό, μονόγλωσσο (δεν περιέχει κείμενα σε άλλη γλώσσα εκτός της ελληνικής) και συγχρονικό, καθώς περιέχει κείμενα από μια συγχρονία της ελληνικής και συγκεκριμένα από το 1999 έως το 2001. Η επιλογή της περιόδου δεν ήταν τυχαία. Καθοριστική για τη διαμόρφωση του ειδικού λεξιλογίου του χρηματιστηρίου υπήρξε η περίοδος 1993-2001, κατά την οποία αυξήθηκε εντυπωσιακά το ενδιαφέρον και η δραστηριοποίηση του εγχώριου πληθυσμού. Η άνοδος του χρηματιστηρίου δημιούργησε νέα δεδομένα στην αγορά από τα οποία προέκυψε η ανάγκη χρήσης νέων λέξεων, που εισήλθαν στο χώρο ως *νεολογισμοί* ή *νεώνυμα*[3]. Ειδικότερα κατά την περίοδο 1999-2001 εντός της οποίας εμφανίζεται η μεγαλύτερη άνοδος και η μεγαλύτερη πτώση στην ιστορία του ελληνικού χρηματιστηρίου, παρατηρείται και ο μεγαλύτερος αριθμός δημοσιευμάτων που αφορούν το χρηματιστήριο.

Το σώμα κειμένων αποτελείται από 2112 άρθρα, μόνο ολόκληρα κείμενα, γραμμένα κατά την περίοδο 1999-2001, ενημερωτικά εγχειρίδια για τους επενδυτές από την ιστοσελίδα του Χρηματιστηρίου Αξιών Αθηνών (κείμενα ειδικής πληροφόρησης), καθώς και τα ενημερωτικά δελτία του ΧΑΑ των αντίστοιχων ετών. Τα άρθρα του σώματος κειμένων αντλήθηκαν από τις εξής εφημερίδες: *Αδέσμευτος Τύπος, Ακρόπολις, Αξία, Απογευματινή της Κυριακής, Δείκτης, Έθνος της Κυριακής, Ελευθεροτυπία, Εξουσία, Εξπρές, Έξυπνο Χρήμα, Επενδυτής, Ημερησία, Ισοτιμία, Καθημερινή, Κέρδος, Μακεδονία, Μέτοχος, Ναυτεμπορική, Οικονομία, Οικονομικός Ταχυδρόμος, Σύμβουλος, Τα Νέα, Το Βήμα της Κυριακής, Το Βήμα, Το Καρφί, Τύπος της Κυριακής, Χρήμα και Αγορά, Χρηματιστήριο, Retail Business.*

[3] Νεολογισμοί ονομάζονται τα αποτελέσματα της διαδικασίας της λεξιλογικής ανανέωσης, δηλαδή της νεολογίας (Αναστασιάδη-Συμεωνίδη, 1986: 26-27), ενώ σύμφωνα με τον Kocourek (1982: 174), μπορούμε να χρησιμοποιήσουμε τον όρο νεωνυμία για να δηλώσουμε τη "νεολογία στο ειδικό λεξιλόγιο", και τον όρο νεώνυμα για τα αποτελέσματα της διαδικασίας της νεωνυμίας.

Ένα σώμα κειμένων θεωρείται αντιπροσωπευτικό με βάση το μέγεθος, την αυθεντικότητα και τις αναλογίες του, δηλαδή τη σχετική ισορροπία μεταξύ των κειμενικών ειδών που το απαρτίζουν (Γούτσος, 2003). Το εν λόγω σώμα κειμένων μπορεί να θεωρηθεί αντιπροσωπευτικό καταρχήν όσον αφορά το μέγεθός του (δύο εκατομμύρια λέξεις), καθώς για ένα εξειδικευμένο σώμα κειμένων κρίνεται ικανοποιητικό ένα μέγεθος από 500.000 έως 5.000.000 λέξεις[4]. Κατά τους Friedbichler & Friedbichler (2000) σώματα κειμένων ανάλογου μεγέθους ανά γλώσσα αποδίδουν επαρκείς πληροφορίες σε ποσοστό 97% των γλωσσικών ερευνών. Σύμφωνα με τους Bowker & Pearson (2002) ακόμη και σώματα κειμένων μερικών χιλιάδων ή μερικών εκατοντάδων χιλιάδων λέξεων αποδείχθηκαν χρήσιμα για την έρευνα ειδικών γλωσσών.

Στην αντιπροσωπευτικότητα του σώματος κειμένων συμβάλλει επίσης η αυθεντικότητά του. Τα κείμενα δεν είναι κατασκευασμένα και έχουν δημιουργηθεί με φυσικό τρόπο (όχι κάτω από πειραματικές συνθήκες).

Τέλος, είναι αντιπροσωπευτικό ως προς τις αναλογίες του, καθώς τα κειμενικά είδη που το απαρτίζουν βρίσκονται σε άμεση συνάρτηση με τους ερευνητικούς στόχους, δηλαδή την κατά το δυνατόν εξαντλητική περιγραφή της χρηματιστηριακής ορολογίας. Έτσι περιλαμβάνονται κείμενα με διαφορετικά επίπεδα ύφους, π.χ.

Τα λαμόγια συνεχίζουν τη σπέκουλα! Τα... πρόβατα φεύγουν ατάκτως από το μαντρί. (Δυστυχώς γι' αυτούς, δεν διαβάζουν όλοι την αφεντιά μου...). Τους τρομοκράτησαν οι ...ποντικοί. Οι γάτες ορμάνε. Τσιμπάνε ό,τι βρουν, όπου το πετύχουν. Οι τρομαγμένοι χώνονται στις Τράπεζες. Για να εισπράξουν τόκο ένα βάρι, όσο ένα λίμιτ απ μιας μέρας στο Χρηματιστήριο. Οι Τράπεζες, βέβαια, θα... επενδύσουν στη Σοφοκλέους.

και

Σε αντίθεση με τα Συμβόλαια Μελλοντικής Εκπλήρωσης, που είναι δεσμευτικά τόσο για τον αγοραστή όσο και για τον πωλητή, ένα συμβόλαιο Δικαιώματος δίνει στον αγοραστή του το δικαίωμα (αλλά όχι την υποχρέωση) να επιλέξει αν τελικά θα το εξασκήσει. Έτσι, ο αγοραστής ανάλογα με τις συνθήκες που διαμορφώνονται στην αγορά αποφασίζει αν τελικά τον συμφέρει να προβεί σε χρήση αυτού του δικαιώματος ή όχι. Η υποκείμενη αξία των Δικαιωμάτων Προαίρεσης, όπως και για τα Σ.Μ.Ε., μπορεί να είναι μία ποικιλία προϊόντων ή αγαθών. Οι συναλλασσόμενοι χρησιμοποιούν τα Δικαιώματα Προαίρεσης για να αγοράσουν ή να πουλήσουν σιτάρι, πολύτιμα μέταλλα, συνάλλαγμα, μετοχές, δείκτες και άλλα

[4] Τα σώματα κειμένων γενικής γλώσσας κυμαίνονται από 10 έως 329 εκατομμύρια λέξεις (για την αγγλική γλώσσα).

αξιόγραφα, τα οποία αποτελούν ένα μικρό μέρος από το σύνολο των προϊόντων που σήμερα συναλλάσσονται. Η υποκείμενη αξία για τα Δικαιώματα στο δείκτη FTSE/ASE-20 που διαπραγματεύεται στην αγορά παραγώγων είναι ο χρηματιστηριακός δείκτης FTSE/ASE-20.

Η συλλογή των δεδομένων έγινε με ηλεκτρονική σάρωση ή χρήση του διαδικτύου. Τα άρθρα που προέρχονταν από εφημερίδες σαρώθηκαν αρχικά ως εικόνες. Για τη μετατροπή τους σε κείμενα χρησιμοποιήθηκε το πρόγραμμα σάρωσης FineReader optical character recognition (OCR). Ακολούθησε καθαρισμός των αρχείων από περιττά μη λεκτικά στοιχεία (π.χ. εικόνες, γραμμές, κενά, κτλ.) και αποθήκευση με τη μορφή απλού κειμένου και κωδικοποίηση σε Unicode[5]. Το αρχικό αποτέλεσμα ήταν ένα κατακερματισμένο κείμενο το οποίο έχρηζε πολλών διορθώσεων.

Η σάρωση των άρθρων και η μετατροπή τους σε έγγραφα απλού κειμένου αποδείχθηκε δύσκολη καθώς πέρα από τα τυπογραφικά λάθη που προέκυψαν, παρατηρήθηκαν μετατοπίσεις στη θέση των παραγράφων. Παρόλο που αυτό δεν δημιουργεί πρόβλημα κατά τη δημιουργία των συμφραστικών πινάκων, η σειρά των παραγράφων αποκαταστάθηκε ώστε να ανταποκρίνεται πλήρως στην έντυπη μορφή των κειμένων. Πρόβλημα δημιούργησαν επίσης οι στήλες, καθώς γραμμές της μίας στήλης εμπλέκονταν με αυτές της διπλανής στήλης.

Ορισμένα άρθρα παρουσίασαν αυξημένη δυσκολία κατά τη μετατροπή τους σε απλό κείμενο λόγω εικόνων και γραφικών στο σχεδιασμό τους. Ένα πρόβλημα επίσης ήταν ότι πολλές φορές οι εικόνες που συνοδεύουν τα άρθρα (στο συγκεκριμένο σώμα κειμένων κυρίως γελοιογραφίες) αποτελούν ουσιώδες μέρος του κειμένου. Αυτά τα εξωκειμενικά στοιχεία χάθηκαν κατά τη μετατροπή σε απλό κείμενο.

Μετά τη διόρθωσή του, το σώμα κειμένων ενσωματώθηκε στο σύστημα αυτόματης ανάλυσης και πραγματοποιήθηκε η προεπεξεργασία του ώστε να γίνει ο τεμαχισμός του σε προτάσεις. Στη συνέχεια κωδικοποιήθηκαν τα στοιχεία που περιλαμβάνουν προέλευση του κειμένου, όνομα συγγραφέα, ημερομηνία παραγωγής και τίτλο.

2 Συμφραστικοί πίνακες

Τα προγράμματα επεξεργασίας σωμάτων κειμένων αναζητούν στα κείμενα δεδομένες

[5] Η κωδικοποίηση σε μορφή Unicode είναι απαραίτητη για την επεξεργασία των κειμένων μέσω του προγράμματος Unitex.

μορφοσυντακτικές δομές[6] οι οποίες, μετά την εφαρμογή του ηλεκτρονικού λεξικού σε σώματα κειμένων, παρουσιάζονται με τη μορφή συμφραστικών πινάκων. Με βάση συγκεκριμένο κείμενο, το πρόγραμμα των συμφραστικών πινάκων κατασκευάζει έναν κατάλογο με όλες τις λέξεις του κειμένου που έχουν επιλεγεί με βάση το λεξικό ταξινομημένες σε αλφαβητική σειρά, παρουσιάζοντας κάθε λέξη με το συγκείμενο που βρίσκεται δεξιά και αριστερά. Π.χ.

Τα Πετρέλαια απέκτησαν πλέον πολλαπλασιαστή κερδών ανάλογο με αυτόν της αγοράς

Διαθέτει αξιοπρεπές P/E, λίγες μετοχές (3,1 εκατ.) και... σεμνή κεφαλαιοποίηση 33,5 δισ. δρχ.

Έκαναν αύξηση μετοχικού κεφαλαίου, με κεφαλαιοποίηση αποθεματικών υπέρ το άρτιο.

Η χρησιμότητα των συμφραστικών πινάκων είναι μεγάλη γιατί επιτρέπουν τη μελέτη του τρόπου χρήσης μιας λέξης στα κείμενα, το είδος του συγκειμένου όπου χρησιμοποιείται μια λέξη του κειμένου που μελετούμε, κ.ά. Τα προγράμματα συμφραστικών πινάκων αποτελούν επομένως βασικά εργαλεία για την επεξεργασία σωμάτων κειμένων.

Αμέσως μετά την κατάρτιση του σώματος κειμένων χρηματιστηριακού περιεχομένου πραγματοποιήθηκε το ακόλουθο πείραμα, η αναζήτηση του όρου «άλογο» (στο χρηματιστήριο η μετοχή υψηλού ρίσκου, η αγορά της οποίας ισοδυναμεί με στοίχημα σε άλογο του ιππόδρομου) και τα συμφραζόμενά του στο σώμα κειμένων που αποτελείται από την ηλεκτρονική έκδοση της εφημερίδας ΤΑ ΝΕΑ (1997-2003) με 115.000.000 λέξεις[7]. Βρέθηκαν 4824 εμφανίσεις του όρου. Παρόλο που η θεματολογία της εφημερίδας *Τα Νέα*, περιλαμβάνει πληθώρα πληροφοριών από διάφορα γνωστικά πεδία, οι συμφραστικοί πίνακες που προέκυψαν αφορούσαν τη σημασία της λέξης ως ζώου. Παρόλο που ορισμένες φορές η λέξη μπορεί να είχε την έννοια του αλόγου ιπποδρομιών και συνειρμικά του τζόγου, η έννοια ήταν δύσκολο να εντοπιστεί, λόγω της ύπαρξης "θορύβου", δηλαδή λανθασμένου εντοπισμού όρων κατά την αναζήτηση. Το ίδιο αποτέλεσμα προέκυψε και κατά την αναζήτηση το ίδιου όρου στο διαδίκτυο (32.000.000 αποτελέσματα, 14.900 σε συνδυασμό

[6] Οι όροι που περιλαμβάνονται στο ηλεκτρονικό λεξικό χρηματιστηριακών όρων εμφανίζουν τουλάχιστον 200 διαφορετικές μορφοσυντακτικές δομές, όσον αφορά τα συστατικά τους στοιχεία, π.χ. επίθετα, ουσιαστικά, προσδιοριστές, επιρρήματα, προθέσεις, μόρια. Οι δομές με τη μεγαλύτερη συχνότητα είναι η δομή Επίθετο + Ουσιαστικό (ΑΝ) (30%), η δομή Ουσιαστικό + Ουσιαστικό (ΝΝ) (22%) και η δομή Ουσιαστικό + Επίθετο + Ουσιαστικό (ΝΑΝ) (8,4%).

[7] Πρόκειται για ένα αντιπροσωπευτικό σώμα κειμένων καθώς τα κείμενα καλύπτουν αναλογικά όλα τα επιμέρους σώματα της εφημερίδας, δηλαδή τις θεματικές στήλες της (ρουμπρίκες). Το σώμα κειμένων, αφού κωδικοποιήθηκε μας παραχωρήθηκε από τον Cédric Fairon του πανεπιστημίου UCL (Université Catholique de Louvain) του Βελγίου.

με τη λέξη «χρηματιστήριο»)[8]. Κατά την αναζήτηση στο σώμα κειμένων χρηματιστηριακού περιεχομένου προέκυψαν μόλις 36 εμφανίσεις του όρου «άλογο» με τη σημασία της μετοχής.

Το αποτέλεσμα της έρευνας καταδεικνύει τη σημασία της κατάρτισης σωμάτων κειμένων με ειδικό περιεχόμενο.

3 Σημασιολογικά χαρακτηριστικά

Έχοντας ως στόχο την πλήρη περιγραφή της Νέας Ελληνικής με απώτερο σκοπό την αυτόματη ανάλυση της ελληνικής γλώσσας, η μορφολογική περιγραφή των λέξεων πρέπει να ολοκληρωθεί με την συντακτικο-σημασιολογική ανάλυση. Για την ανάλυση αυτή, υιοθετούμε το μεθοδολογικό μοντέλο που πρότεινε ο Gross (1975), με τη χρήση πινάκων λεξικού-γραμματικής[9]. Η προσθήκη σημασιολογικών χαρακτηριστικών στα ουσιαστικά του λεξιλογίου των όρων του χρηματιστηρίου θα επιτρέψει τη σύνδεση των ηλεκτρονικών μορφολογικών λεξικών με τους πίνακες του λεξικού-γραμματικής. Η προσθήκη αυτή διευκολύνθηκε με τη δημιουργία του σώματος κειμένων χρηματιστηριακού περιεχομένου και την κατάρτιση συμφραστικών πινάκων.

Για την εκπόνηση αυτού του τμήματος της εργασίας, βασιστήκαμε στην αντίστοιχη έρευνα της Courtois (1994) για τη γαλλική γλώσσα, την οποία η Φούφη (2005) προσάρμοσε στις ιδιαιτερότητες των ελληνικών[10]. Υιοθετήσαμε τη μεθοδολογία τους και τον τρόπο παρουσίασης για την κωδικοποίηση των σημασιολογικών χαρακτηριστικών.

Τα βασικά σημασιολογικά χαρακτηριστικά που προστέθηκαν στο ηλεκτρονικό λεξικό των

[8] Ανάλογη έρευνα πραγματοποίησαν οι Bowker & Pearson (2002), αναζητώντας τον όρο «nut» («παξιμάδι») σε τεχνικά κείμενα. Η πρώτη τους αναζήτηση στο σώμα κειμένων British National Corpus (BNC) των 100 εκατομμυρίων λέξεων έφερε 670 αποτελέσματα, τα οποία ωστόσο δεν είχαν σχέση με το μηχανολογικό όρο. Η αναζήτηση σε ένα σώμα τεχνικών κειμένων μόλις 10.000 λέξεων έφερε 49 αποτελέσματα, τα οποία αν και πολύ λιγότερα, αφορούσαν την τεχνική έννοια του όρου, καθώς μειώθηκε ο «θόρυβος», ενώ διευκολύνθηκε και ο εντοπισμός των διαφορετικών τύπων του εξαρτήματος «nut» που χρησιμοποιούνται στις κατασκευές, καθώς και των ρημάτων που συνδέονται με τον όρο.

[9] Πρόκειται για πίνακες συντακτικής περιγραφής που συνδέουν τα εκάστοτε κατηγορήματα με τα υποκείμενα και τα συμπληρώματα που δέχονται. Πιο συγκεκριμένα, τα ονοματικά σύνολα ορίζονται με βάση μορφολογικά και σημασιολογικά χαρακτηριστικά.

[10] Μια πρώτη προσέγγιση για τα ελληνικά έχει γίνει από την Κωνσταντάρα (2003).

χρηματιστηριακών όρων ήταν: το *Hum* για ανθρώπινα ουσιαστικά[11] π.χ. *επενδυ"τής.N44500,N+Hum+[Eco]*, το *Conc* για τα συγκεκριμένα ουσιαστικά, π.χ. *ενεχυρόγρα"φο.N327,N+Conc+[Eco]* και το *Abst* για τα αφηρημένα ουσιαστικά (Abst), π.χ. *αρνητική.A10 ελαστικότη"τα.N236 ζήτη"σης.N265,N+Abst+[Eco],-GS3*. Σε όλα προστέθηκε το ειδικό σημασιολογικό χαρακτηριστικό *[Eco]* που χαρακτηρίζει τους όρους από το χώρο της οικονομίας.

Αξίζει να σημειωθεί ότι σε σύγκριση με το γενικό λεξιλόγιο, εδώ δεν υπάρχει η γενική κατηγορία *Ζώα [Anl]*. Έτσι, το *άλογο* ή *αλογάκι*, δηλώνει τη μετοχή υψηλού ρίσκου που προσφέρεται για κερδοσκοπία, ο *παπαγάλος* ή το *παπαγαλάκι* τον καλοθελητή που διαδίδει ψευδείς ειδήσεις, η *αρκούδα* τον απαισιόδοξο επενδυτή ή την απαισιόδοξη αγορά γενικότερα (*αγορά-αρκούδα*), ο *ταύρος* τον αισιόδοξο επενδυτή ή την αισιόδοξη αγορά, η *γαλοπούλα* σημαίνει μια άστοχη, άνευ σημασίας επένδυση.

Επίσης, στο λεξικό των οικονομικών όρων δεν υπάρχουν λήμματα που περιγράφουν μέρη του σώματος. Η ορολογική πολυλεκτική σύνθετη μονάδα *κεφάλι και ώμοι* αναφέρεται σε σήμα αντιστροφής της ανοδικής τάσης.

Σημαντικό είναι ότι με την προσθήκη των σημασιολογικών χαρακτηριστικών επιλύονται αμφισημίες σε περιπτώσεις όπως π.χ. *μάνες* [όρος ο οποίος δεν φέρει το σημασιολογικό χαρακτηριστικό του ανθρώπινου και σε χρηματιστηριακά κείμενα αναφέρεται σε παλιές μετοχές της εταιρείας, πριν αυτή μοιράσει νέες (*παιδιά*)]. Ομοίως, στα χρηματοοικονομικά η λέξη *μετάφραση* έχει την έννοια της μετατροπής μιας αξίας από ένα νόμισμα σε κάποιο άλλο, χρηματιστές και επενδυτές ονομάζουν *ναό* το Χρηματιστήριο Αξιών Αθηνών, *φούσκα* τη μετοχή ευτελούς αξίας με διογκωμένη τιμή κ.ο.κ. Η επίλυση αυτών των αμφισημιών δεν ήταν μέχρι τώρα εφικτή, εφόσον η επεξεργασία περιοριζόταν στο μορφολογικό επίπεδο.

6 Συμπεράσματα – Προοπτικές

Η χρήση των ηλεκτρονικών σωμάτων κειμένων έχει γενικευτεί τα τελευταία χρόνια, με αποτέλεσμα την καθιέρωσή τους ως εργαλείων αναφοράς μαζί με τα παραδοσιακά εξειδικευμένα λεξικά και τις ορολογικές βάσεις δεδομένων.

[11] Για το χαρακτηρισμό των σημασιολογικών χαρακτηριστικών χρησιμοποιείται ως κοινή γλώσσα η γαλλική στο πλαίσιο συνεργασίας του Εργαστηρίου Μετάφρασης και Επεξεργασίας του Λόγου του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης και του Institut Gaspard-Monge του Université de Paris-Est που συμμετέχουν στο ερευνητικό δίκτυο RELEX, το οποίο απαρτίζεται από εργαστήρια σε όλο τον κόσμο που ασχολούνται με την τυπική περιγραφή φυσικών γλωσσών, όπως η αγγλική, η γερμανική, η ισπανική, η ιταλική, η πορτογαλική, η νορβηγική, η κορεατική, η ταϋλανδέζικη κ.ά. Έτσι, το N αναφέρεται σε ουσιαστικά, το A σε επίθετα, οι αριθμοί δίπλα στις γραμματικές κατηγορίες αναφέρονται σε διαφορετικούς τρόπους κλίσης.

Το σώμα κειμένων αποδεικνύεται ένα εξαιρετικά πολύτιμο εργαλείο. Οι συμφραστικοί πίνακες που προκύπτουν από την επεξεργασία του επιλύουν προβλήματα που η διαίσθηση του φυσικού ομιλητή αδυνατεί να επιλύσει (για παράδειγμα δεν μπορεί να μας δώσει πληροφορίες για τη συχνότητα χρήσης). Γίνεται δυνατή η σωστή χρήση των όρων και η εξαγωγή συμπερασμάτων από τη χρήση ποσοτικών στοιχείων.

Το σώμα κειμένων χρηματιστηριακού περιεχομένου μπορεί να αποτελέσει ένα βασικό εργαλείο έρευνας με πολλαπλές εφαρμογές. Μπορεί να προσφέρει στην έρευνα με τη χρήση ηλεκτρονικών υπολογιστών για το λεξιλόγιο και τη σύνταξη ενός ιδιολέκτου της Ελληνικής. Στη λεξικογραφία μπορεί να αποτελέσει βάση για τον αυτόματο εντοπισμό και εξαγωγή όρων, για παράδειγμα με τη χρήση λέξεων-κλειδιών, π.χ.

Αναπροσαρμόστηκε η μέθοδος υπολογισμού της τιμής κλεισίματος των μετοχών. Ως τιμή κλεισίματος ***θεωρείται*** *ο σταθμικός μέσος όρων των τιμών που είχε η μετοχή κατά το τελευταίο δεκάλεπτο της συνεδρίασης με στάθμιση τον όγκο της κάθε συναλλαγής.*

Το γεγονός ότι ελεύθερα οι κεφαλαιαγορές αυτό που φτιάχνουν είναι «φούσκα», που στην οικονομική ορολογία ***σημαίνει*** *διόγκωση των αξιών που δεν ανταποκρίνεται στα fundamentals.*

Η συμβολή ενός σώματος κειμένων είναι καθοριστική για την παραγωγή έγκυρων και σύγχρονων ορολογικών έργων αναφοράς όπως γλωσσάρια, λεξικά υπογλωσσών, ορολογικές βάσεις δεδομένων, τα οποία μπορούν να χρησιμοποιηθούν από συγγραφείς τεχνικών κειμένων, μεταφραστές κ.ά. Η προσθήκη συντακτικών χαρακτηριστικών που θα ολοκληρωθεί σε επόμενο στάδιο της έρευνας θα εμπλουτίσει περαιτέρω το λεξικό χρηματιστηριακών όρων.

Μπορεί να αποτελέσει πηγή αυθεντικού διδακτικού υλικού για τη διδασκαλία ειδικού λεξιλογίου. Η γλωσσική μελέτη και διδασκαλία εμπλουτίζεται με μη κατασκευασμένα παραδείγματα. Μπορούμε επίσης να παρατηρήσουμε τα χαρακτηριστικά των ποικιλιών και να συγκρίνουμε μεταξύ ποικιλιών και επιπέδων ύφους.

Η χρησιμότητά του αποδεικνύεται κυρίως στη μετάφραση, καθώς συμβάλλει στην ορθότητα του μεταφράσματος και στην κατανόηση όρων με τη βοήθεια των συμφραζομένων. Αποτελεί μάλιστα ένα από τα πλέον απαραίτητα βοηθήματα καθώς συχνά οι όροι στα εξειδικευμένα λεξικά και τις ορολογικές βάσεις εμφανίζονται συχνά δίχως τα συμφραζόμενά τους και με ελλιπείς πληροφορίες για τη χρήση τους. Μπορεί να συμβάλλει στην ανάπτυξη των μεταφραστικών σπουδών σε σύνδεση με συλλογές κειμένων σε άλλες γλώσσες καθώς και με πολύγλωσσα και παράλληλα σώματα κειμένων. Μία από τις προοπτικές της έρευνας στο

μέλλον είναι η δημιουργία ενός σώματος παράλληλων κειμένων από το χώρο του χρηματιστηρίου καθώς νέες μεταφράσεις διατίθενται πλέον στο διαδίκτυο.

Βιβλιογραφικές παραπομπές

[1] Αναστασιάδη-Συμεωνίδη, Α. *Η Νεολογία στην Κοινή Νεοελληνική*: Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής του Α.Π.Θ., Παράρτημα 65, Θεσσαλονίκη, 1986.

[2] Bowker, L. & J. Pearson, *Working with Specialize Language: a practical guide to using corpora*, Routledge, London, 2002.

[3] Γούτσος, Δ. «Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση». *Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας*, Πανεπιστήμιο Κρήτης, 18-21 Σεπτεμβρίου 2003.

[4] Courtois, B. *Marques Lexicales, Syntaxiques, Sémantiques dans le DELAS V08E*, rapporte technique, LADL, Université Paris 7, 1994.

[5] Friedbichler, I. & M. Friedbichler, "The Potential of Domain-Specific Target-Language Corpora for the Translator's Workbench", *I corpora nella didattica della traduzione: Corpus Use and Learning to Translate*, Bologna, 2000, σσ. 107-116.

[6] Gross, M. *Méthodes en syntaxe*, Hermann, Paris, 1975.

[7] Kocourek, R. *La Langue française de la technique et de la science*, Brandstetter, 1982.

[8] Κωνσταντάρα Ζ. «Ταξινόμηση των μονολεκτικών ουσιαστικών της ΝΕ με βάση τα σημασιοσυντακτικά χαρακτηριστικά τους», *Μελέτες για την ελληνική γλώσσα-Πρακτικά της 24ης ετήσιας συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής Α.Π.Θ.,* Θεσσαλονίκη, 2003, σσ. 334-344.

[9] Kyriacopoulou, T. *L'analyse automatique des textes écrits : le cas du grec moderne*. University Studio Press, Θεσσαλονίκη, 2005.

[10] Paumier, S. Unitex. *Manuel d' utilisation*, Université de Marne-la-Vallée, 2002.

[11] Silberztein, M. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX, sous la direction de Maurice Gross*, Collection Informatique Linguistique, Masson, Paris, 1993.

[12] Sinclair, J. *Corpus and Text – Basic Principles, in Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books, 2005.

[13] Τζιάφα, Ε. «Κατασκευή ηλεκτρονικού λεξικού οικονομικής-χρηματιστηριακής ορολογίας – Γενικές παρατηρήσεις», *Ελληνική Γλώσσα και Ορολογία. Ανακοινώσεις 6ου Συνεδρίου,* Αθήνα, 2007, σσ. 289-298.

[14] Φούφη, Β. *Μέθοδοι και Κριτήρια Σημασιολογικής Κωδικοποίησης των Ουσιαστικών της Νέας Ελληνικής με σκοπό την Αυτόματη Ανάλυση Κειμένων*, αδημ. μεταπτυχιακή εργασία, 2005.

Τζιάφα Ελένη

Υποψήφια Διδάκτωρ, Τμήμα Φιλολογίας
Ταχ. Διεύθυνση: Α.Π.Θ. Τμήμα Γαλλικής Γλώσσας και Φιλολογίας
GR-54124, Θεσσαλονίκη
Τηλέφωνο: +30 2310 997516
Ηλ-διεύθυνση: etziafa@lit.auth.gr