Καθορισμός Βασικού Λεξιλογίου μέσω ΗΣΚ για τη διδασκαλία της ΝΕ ως ΞΓ

Μαρία Ιακώβου, Γιώργος Μαρκόπουλος, Γιώργος Μικρός
Πανεπιστήμιο Αθηνών

1 Κριτήρια καθορισμού Βασικού Λεξιλογίου

Σκοπός της ανακοίνωσής μας είναι να παρουσιαστούν παρατηρήσεις και αποτελέσματα από μια προσπάθεια καθορισμού του Βασικού Λεξιλογίου (ΒΛ) της Νέας Ελληνικής στην προσέγγισή της ως Ξένης Γλώσσας μέσω Ηλεκτρονικών Σωμάτων Κειμένων (ΗΣΚ).

Ο όρος «Βασικό Λεξιλόγιο» (Core Vocabulary) αποδίδεται στον Carter (1987: 33-46), όπως και μια σειρά σημασιοσυντακτικών κριτηρίων για τη διάκριση των βασικών λέξεων. Με δεδομένο, όμως, τον ασαφή και δυσδιάκριτο χαρακτήρα των ορίων του, είναι εμφανές ότι δεν υπάρχει ένα παρά πολλά βασικά λεξιλόγια σε μια γλώσσα και σε έναν ομιλητή ανάλογα προς τις επικοινωνιακές καταστάσεις που καλείται να ντύσει γλωσσικά με αυτά. Για τον λόγο αυτό καταλήγουμε στο συμπέρασμα ότι πίσω από την έννοια της «βασικότητας» των λεξιλογικών στοιχείων καλύπτεται η ανάγκη μιας απλοποιημένης και ταυτόχρονα επαρκούς δυνατότητας επικοινωνίας σε συγκεκριμένα περιβάλλοντα της γλώσσας-στόχου. Η ανάγκη αυτή είναι κατ'εξοχήν εμφανής στα πρώτα στάδια εκμάθησης μιας ξένης γλώσσας όταν ο μαθητής επιζητώντας τη γλωσσική πραγμάτωση λειτουργιών ανάλογων προς τις γνωστικές δομές που διαθέτει από τη μητρική του γλώσσα, ανατρέχει στην επικοινωνιακή αξιοποίηση των λεξιλογικών στοιχείων της γλώσσας-στόχου με τα οποία αισθάνεται τη μεγαλύτερη εξοικείωση και συνιστούν το δικό του «βασικό λεξιλόγιο». Στο πλαίσιο αυτό η έννοια του ΒΛ επαναπροσδιορίζεται βάσει του ποσοτικού κριτηρίου της συχνότητας: το ΒΛ συνδέεται με την υψηλή συχνότητα εμφάνισης των λέξεων ενός σημασιολογικού πεδίου[1], καθώς οι λέξεις αυτές απαντούν σε όλες τις χρήσεις της γλώσσας, καλύπτουν ένα σημαντικό ποσοστό κειμένων γραπτού και προφορικού λόγου και λειτουργούν ικανοποιητικά για την κάλυψη βασικών επικοινωνιακών αναγκών (Nation 2001: 13).

[1] Για τη συχνότητα ως παράγοντα βασικότητας ενός σημασιολογικού πεδίου, βλ. Barsalou (1992: 35) «Τόσο και η συχνότητα εμφάνισης όσο και η εννοιολογική αναγκαιότητα φαίνονται να συμβάλλουν στον πυρήνα των σημασιολογικών πεδίων (frames)»

2 Μεθοδολογία

2.1 Χαρακτηριστικά της προτεινόμενης προσέγγισης

Η προτεινόμενη προσέγγιση έχει ως αφετηρία της τις παραπάνω θεωρητικές αρχές και στοχεύει στην ανάπτυξη ενός ΒΛ για τη ΝΕ που θα στηρίζεται σε περισσότερο ασφαλή δεδομένα από τη διαισθητική εμπειρία των συντακτών της. Τα χαρακτηριστικά της έγκεινται στα εξής σημεία:

α) το υλικό που αποτελεί τη βάση για την ανάπτυξη καταλόγων ΒΛ προέρχεται από την αξιοποίηση των δυνατοτήτων που παρέχουν τα ΗΣΚ.

β) η συγκέντρωση του υλικού έγινε για μια συγκεκριμένη θεματική περιοχή, τις ΑΓΟΡΕΣ (καταναλωτικά αγαθά – καταστήματα – υπηρεσίες), όπως αυτή περιγράφεται στα Αναλυτικά Προγράμματα[2] για τη Διδασκαλία της ΝΕ ως ΞΓ, με σκοπό να αναλυθεί σε όλες τις περαιτέρω λειτουργίες και επικοινωνιακές καταστάσεις στις οποίες εμπλέκεται ο χρήστης της,

γ) ο σχεδιασμός και οι απαιτήσεις του υλικού είχαν ως λογικό επακόλουθο τη διασπορά του σε μια ποικιλία κειμενικών ειδών (προφορικού και γραπτού λόγου, αν και για τις ανάγκες της συγκεκριμένης παρουσίασης θα στηριχτούμε αποκλειστικά σε υλικό από γραπτό λόγο[3]) σχετιζόμενων με όλο το εύρος των δραστηριοτήτων που μπορούν να ενταχθούν στη θεματική περιοχή που επιλέχθηκε (διαφημίσεις, άρθρα από εφημερίδες, αγγελίες, ενημερωτικά φυλλάδια, οδηγίες χρήσης κλπ.) Με τον τρόπο αυτό προσπαθήσαμε να οριοθετήσουμε την έννοια της «βασικότητας» στο λεξιλόγιο της συγκεκριμένης θεματικής περιοχής, δημιουργώντας ένα μεθοδολογικό εργαλείο που θα μπορεί να εφαρμοστεί στη συνέχεια για τον λεξιλογικό καθορισμό και των λοιπών θεματικών κύκλων των Αναλυτικών Προγραμμάτων της ΝΕ ως ΞΓ.

[2] Η συγκεκριμένη θεματική ενότητα απαντά αυτόνομα στο Αναλυτικό Πρόγραμμα για το «Ενδιάμεσο Επίπεδο για τα Νέα Ελληνικά» του Κέντρου Ελληνικής Γλώσσας (2001), ενώ εντάσσεται μέσα το ευρύτερο πλαίσιο για την Καθημερινή Ζωή στο Αναλυτικό Πρόγραμμα για το «Εισαγωγικό 1 και το Βασικό 2 Επίπεδο» του Πανεπιστημίου Αθηνών (1998). Και οι δύο προσεγγίσεις, ωστόσο, δεν διαφέρουν μεθοδολογικά ούτε ως προς την περιγραφή που δίνουν για το συγκεκριμένο θεματικό πεδίο, ούτε ως προς τα λεξιλογικά στοιχεία που περιλαμβάνουν σε αυτό.

[3] Για τις παρατηρήσεις από τα δεδομένα του προφορικού λόγου, βλ. Ιακώβου, Μαρκόπουλος & Μικρός (2003)

2.2 Ηλεκτρονικά Σώματα Κειμένων: Θεωρία και εφαρμογές στην διδασκαλία της γλώσσας

2.2.1 Τα ΗΣΚ στη γλωσσική έρευνα

Η χρήση *ηλεκτρονικών* ή *μηχανικά αναγνώσιμων σωμάτων κειμένων* (ΗΣΚ) στην έρευνα για τη γλώσσα συνιστά μια ξεχωριστή μεθοδολογία, η οποία βασίζει τις αποδείξεις της και αντλεί τα επιχειρήματά της από μια καθαρά εμπειρική περιγραφή της γλωσσικής μαρτυρίας, όπως αυτή εμφανίζεται σε συλλογές κειμένων του γραπτού ή / και του προφορικού λόγου (corpora).

Η συγκεκριμένη μεθοδολογία γλωσσικής περιγραφής ενσωματώνει την ποσοτικοποίηση κατανομής των γλωσσικών στοιχείων ως μέρος της ερευνητικής δραστηριότητας. Όπως χαρακτηριστικά σημειώνει ο Leech (1992:107), η έρευνα εστιάζει στη γλωσσική πλήρωση (performance) παρά στη γλωσσική ικανότητα (competence) και στην παρατήρηση μιας δυνάμει χρήσης της γλώσσας, η οποία μας οδηγεί στη θεωρητική υπόθεση και όχι το αντίστροφο.

Κατ' αυτή την έννοια μια γλωσσική περιγραφή που βασίζεται στην εξέταση ΗΣΚ διαφέρει από γλωσσικές προσεγγίσεις που εξαρτούν τις αποδείξεις τους από τη διαίσθηση και την ενδοσκόπηση και αντανακλά τη διάκριση μεταξύ της εμπειρικής και της ορθολογιστικής μεθοδολογίας. Η διάκριση αυτή δεν σημαίνει απαραίτητα την προτίμηση της μιας προσέγγισης και τον αποκλεισμό της άλλης. Η χρήση ΗΣΚ ως πηγών μαρτυρίας για τη γλώσσα δεν είναι απαραίτητα ασύμβατη με τις γλωσσολογικές θεωρίες. Οποιεσδήποτε δηλώσεις γίνονται για τη γλώσσα θα πρέπει να στηρίζονται σε εμπειρικές αποδείξεις από τη χρήση της. Οι αποδείξεις αυτές θα πρέπει να βασίζονται στο γλωσσικό αίσθημα των φυσικών ομιλητών της γλώσσας ή σε σώματα κειμένων. Η διαφορά, κατά τον Kennedy (1998:8), έγκειται στον πλούτο των αποδείξεων και στην εμπιστοσύνη που μπορούμε να έχουμε στη γενίκευση αυτών των αποδείξεων, όσον αφορά στην εγκυρότητα και στην αξιοπιστία τους.

Ως εκ τούτου, τα όρια ανάμεσα στις δύο προσεγγίσεις της γλωσσικής περιγραφής δεν είναι αυστηρά και, σύμφωνα με τον Fillmore (1992:35), για να πετύχει κανείς στο εγχείρημα αυτό θα πρέπει να κάνει χρήση και των δύο πηγών μαρτυρίας: τα ΗΣΚ προσφέρουν γλωσσικό πλούτο που κανένας ερευνητής δεν μπορεί να φθάσει ωθούμενος μόνο από ενδοσκόπηση και βασιζόμενος στους συλλογισμούς του· και ταυτόχρονα κάθε φυσικός ομιλητής έχει στέρεη γνώση της

γλώσσας του που καμία μαρτυρία κειμενικών δεδομένων δεν μπορεί από μόνη της να υποστηρίξει ή να αντικρούσει.

2.2.2 Πλεονεκτήματα από τη χρήση ΗΣΚ

Τα πλεονεκτήματα που παρουσιάζει μια έρευνα βασισμένη σε ΗΣΚ είναι άμεσα συνδεδεμένα με τις δυνατότητες που μας παρέχει σήμερα η τεχνολογία των ηλεκτρονικών υπολογιστών. Τα σύγχρονα υπολογιστικά μέσα μαζικής καταγραφής και αποθήκευσης δεδομένων από τη μια και οι υψηλές ταχύτητες επεξεργασίας τους από την άλλη καθιστούν στην ουσία κάθε φυσικό ομιλητή και κάτοχο ενός προσωπικού υπολογιστή συλλέκτη γλωσσικού υλικού. Η τεράστια αυτή ευκολία μεταφερόμενη στο ερευνητικό επίπεδο και υποστηριζόμενη από ειδικά προγράμματα επεξεργασίας και ανάλυσης γλωσσικών δεδομένων καθιστούν τον ηλεκτρονικό υπολογιστή έναν πολύτιμο βοηθό στη σπουδή της δομής και της χρήσης της γλώσσας.

Οι τεχνολογικές δυνατότητες των Η/Υ όσον αφορά στην διαχείριση σωμάτων κειμένων αναφαίνονται στις διαδικασίες αναζήτησης, ανάκτησης, ταξινόμησης και υπολογισμού γλωσσικών δεδομένων, είτε αυτά αποτελούν κείμενο σε ηλεκτρονική, μηχανικά-αναγνώσιμη μορφή είτε ψηφιοποιημένη φωνή από προφορικά δεδομένα. Συγκεκριμένα, ο υπολογιστής είναι σε θέση να ψάξει μέσα στα ΗΣΚ για μια συγκεκριμένη λέξη ή ακολουθία λέξεων ή ακόμα και για ένα συγκεκριμένο μέρος του λόγου. Επιπλέον, μπορεί να ανακτήσει και να εμφανίσει όλα τα παραδείγματα μιας λέξης με τα συμφραζόμενά της και να υπολογίσει τον αριθμό εμφανίσεων της λέξης αυτής με σκοπό τον καθορισμό της συχνότητας εμφάνισής της μέσα σ' ένα σώμα κειμένων.

Συνοψίζοντας τις δυνατότητες αυτές σε συνδυασμό με όσα ελέχθησαν παραπάνω θα λέγαμε ότι μια γλωσσική έρευνα που βασίζει την περιγραφή της σε ΗΣΚ έχει τα ακόλουθα πλεονεκτήματα έναντι άλλων προσεγγίσεων:

- Ασχολείται με παραδείγματα από την καθημερινή μαρτυρία στη χρήση της φυσικής γλώσσας σε αντίθεση με παραδείγματα 'κατασκευασμένης' γλώσσας, προϊόντα ενδοσκόπησης και γλωσσικού διαλογισμού
- Ένα ΗΣΚ περιέχει περισσότερες εμφανίσεις, περισσότερα παραδείγματα απ' όσα μπορεί να σκεφθεί ένας ερευνητής
- Ένα ΗΣΚ αποτελείται από επαληθεύσιμα δεδομένα και κατά συνέπεια επιτρέπει τη διεξαγωγή μιας εμπειρικής έρευνας

- Ένα ΗΣΚ μας αποκαλύπτει τη συχνότητα εμφάνισης και την κατανομή μιας λέξης ή ακολουθίας λέξεων και των συμφραζομένων της
- Ένα ΗΣΚ επιτρέπει στους ερευνητές την αναγνώριση και ανάλυση *τύπων συνάφειας* (association patterns)[4]

2.2.3 Τα ΗΣΚ και η διδασκαλία της γλώσσας

Η χρήση ΗΣΚ στη διδασκαλία και εκμάθηση της γλώσσας κρίνεται ιδιαίτερα σημαντική και τυγχάνει τον τελευταίο καιρό μιας διαρκώς αυξανόμενης αποδοχής. Βέβαια, όπως επισημαίνει ο Tribble (2001), ο προβληματισμός για μια άμεση χρήση σωμάτων κειμένου στο μάθημα της γλώσσας έχει τεθεί από τις αρχές της δεκαετίας του ’90 (Widdowson 1991, Sinclair 1991, 1997). Η χρήση παραδειγμάτων από ΗΣΚ κατά τους McEnery και Wilson (1996:104) εκθέτει τους μαθητές στο πρώιμο στάδιο της μαθησιακής διαδικασίας σε είδη προτάσεων και λεξιλόγιο με το οποίο έρχονται σε επαφή κατά την ανάγνωση αυθεντικών κειμένων ή κατά τη χρήση της γλώσσας σε αυθεντικές επικοινωνιακές περιστάσεις. Ιδιαίτερα στη διδασκαλία της ξένης γλώσσας οι παιδαγωγοί αναγκάζονται να κάνουν κάποιες επιλογές, οι οποίες σχετίζονται με τα ενδιαφέροντα και την εκπαίδευσή τους και εξαρτώνται από παράγοντες όπως τα αναλυτικά προγράμματα και το γλωσσικό επίπεδο των μαθητών.

Μια από τις μεθοδολογικές προκλήσεις για τον δάσκαλο ή τη δασκάλα της ξένης γλώσσας αποτελεί η απόφαση για το εύρος του διδασκόμενου λεξιλογίου. Μια τέτοια απόφαση έχει άμεσο αντίκτυπο στο λεξικό περιεχόμενο των μαθησιακών δραστηριοτήτων και στην αξιολόγηση της γλωσσικής επίδοσης. Λόγω της προόδου που έχει επιτευχθεί στην εφαρμογή των ΗΣΚ στην έρευνα του βασικού λεξιλογίου, κατέστη δυνατή η διεξαγωγή ποσοτικών αναλύσεων στη σπουδή της λέξης με άμεση αναφορά στη διδασκαλία της ξένης γλώσσας (Nation 1990, Meunier 1998). Λίστες συχνοτήτων και λεξικές αναλύσεις γραπτών και προφορικών σωμάτων κειμένων έθεσαν σαφείς και ακριβείς ‘λεξιλογικούς’ στόχους για την εκπαίδευση στη ξένη γλώσσα.

Στην παρούσα ανακοίνωση εξετάζεται αυτή ακριβώς η προσέγγιση στην έρευνα για το βασικό λεξιλόγιο διδασκαλίας της ΝΕ ως ξένης γλώσσας. Για το σκοπό

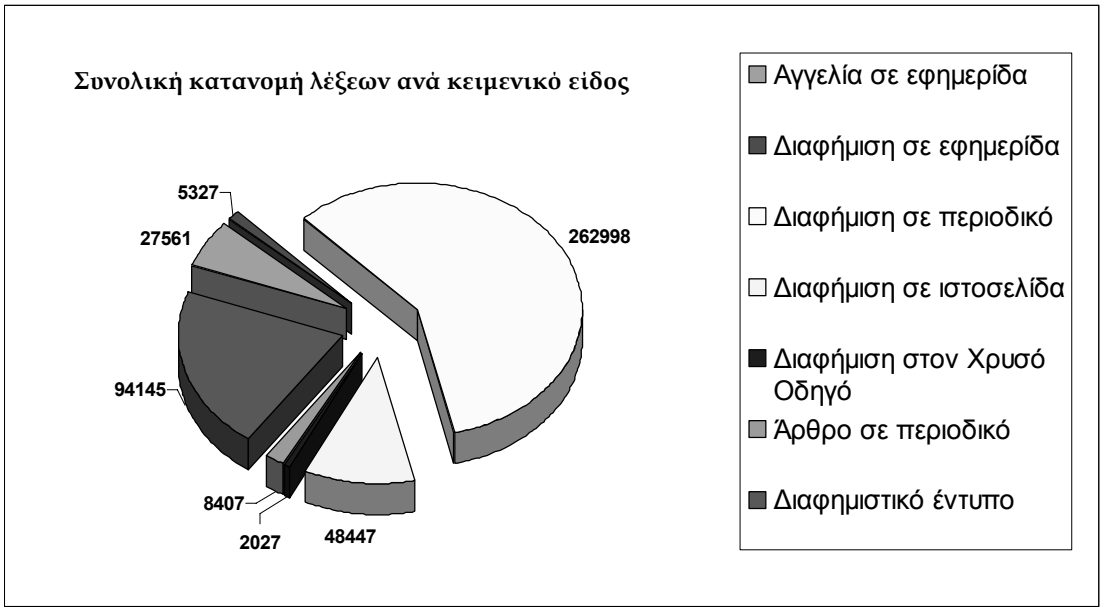
[4] Ο Biber (Biber *et al.*, 1998) ορίζει τους τύπους συσχέτισης ως «...τους συστηματικούς τρόπους με τους οποίους αναλύονται γλωσσικά στοιχεία σε συσχετισμό με άλλα γλωσσικά και μη στοιχεία». Ως γλωσσικά στοιχεία νοούνται οι διάφοροι λεξικοί και γραμματικοί τύποι μιας γλώσσας σε αντιπαράθεση με μη γλωσσικά στοιχεία όπως η συχνότητα εμφάνισης και η κατανομή τους σε διαφορετικά κειμενικά είδη, σε διαφορετικές διαλέκτους και χρονικές περιόδους.

αυτό συλλέχθηκε ένα ΗΣΚ, το οποίο για τις ανάγκες της παρουσίασης περιορίστηκε κατ' αρχάς σε δεδομένα του γραπτού λόγου με άξονα τη θεματική περιοχή ΑΓΟΡΕΣ[5].

2.2.4 Περιγραφή του ΗΣΚ

Το σώμα κειμένων που καταρτίστηκε για τις ανάγκες της παρούσας ανακοίνωσης απαριθμεί ένα σύνολο από 434.634 λέξεις (tokens) και 47.039 διαφορετικούς τύπους λέξεων (types)[6], που κατανέμονται σε 2.065 διαφορετικά αρχεία σε ηλεκτρονική μορφή. Η συλλογή του υλικού βασίστηκε σε τυχαία δειγματοληψία με μέσο εύρος αρχείου τις 217 λέξεις. Τη συλλογή ακολούθησε η καταγραφή και ταξινόμηση του υλικού σε πίνακες με στοιχεία για το όνομα κάθε αρχείου, τον αριθμό λέξεων και τύπων που περιέχει, το κειμενικό είδος και το θεματικό περιεχόμενο, το οποίο προέκυψε από την ανάλυση της βασικής θεματικής περιοχής σε επιμέρους λειτουργικές κατηγορίες.

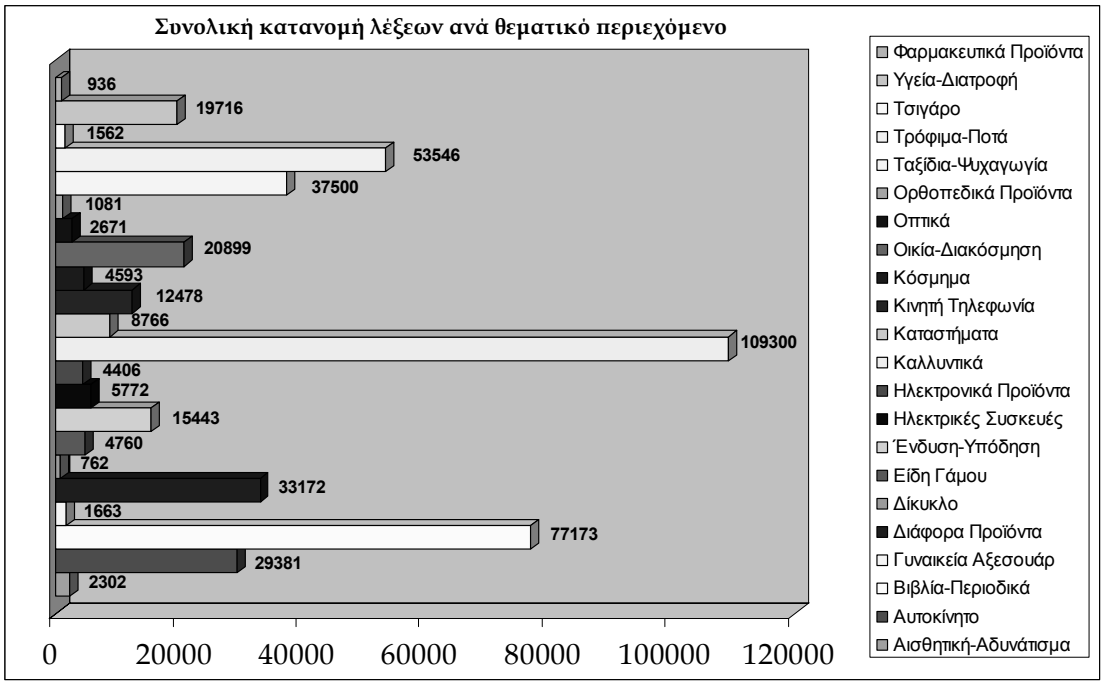
Ως προς το κειμενικό είδος οι πηγές που χρησιμοποιήθηκαν αναλύονται στο γράφημα που ακολουθεί μαζί με τη συνολική κατανομή των λέξεων ανά τύπο κειμένου:



[5] Στο σημείο αυτό θεωρούμε χρέος μας να αναφερθούμε και να ευχαριστήσουμε τις φοιτήτριες και τους φοιτητές του Διατμηματικού Προγράμματος της ΝΕ ως Ξένης Γλώσσας του ακαδημαϊκού έτους 2001-02 για την πολύτιμη βοήθεια τους στη συλλογή και την ηλεκτρονική καταγραφή του γλωσσικού υλικού.

[6] Η διαφορά μεταξύ tokens και types έγκειται στην εμφάνιση των λέξεων. Όταν για παράδειγμα λέμε ότι ένα κείμενο έχει 1000 λέξεις, πολλές από τις λέξεις αυτές επαναλαμβάνονται και οι επανεμφανίσεις τους προσμετρώνται στο τελικό σύνολο των λέξεων / tokens του κειμένου. Από τις 1000 αυτές λέξεις, 400 αποτελούν μοναδικές εμφανίσεις διαφορετικών τύπων / types του κειμένου (Nation 2001).

Όσον αφορά τα θεματικά υποπεδία της υπερκείμενης θεματικής ενότητας «Αγορές» το ΗΣΚ που δημιουργήθηκε εμφανίζει την ακόλουθη κατανομή:



2.3 Στατιστική επεξεργασία του λεξιλογίου

Η εξαγωγή των σημαντικών λέξεων του θεματοποιημένου ΗΣΚ που θα ακολουθήσουμε εδώ θα στηριχθεί στην αξιοποίηση ειδικευμένων στατιστικών δεικτών οι οποίοι θα μας βοηθήσουν να εντοπίσουμε τις λέξεις εκείνες που παρουσιάζουν ιδιάζουσα κατανομή και επομένως είναι υποψήφιες για την ένταξή τους στο βασικό λεξιλόγιο.

Η μεθοδολογία που χρησιμοποιήσαμε στηρίζεται στη σύγκριση της Λίστας Συχνότητας Λεξιλογίου (ΛΣΛ) του θεματοποιημένου ΗΣΚ με την ΛΣΛ του ΗΣΚ γενικής γλώσσας το οποίο στην παρούσα περίπτωση είναι ο Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ)[7]. Για τον λόγο αυτό κατασκευάστηκαν δύο ΛΣΛ. Η πρώτη στηρίχθηκε στο θεματοποιημένο ΗΣΚ και περιείχε 47.039 τύπους (μοναδικές εμφανίσεις λέξεων) ενώ η δεύτερη αξιοποίησε μια παλαιότερη μορφή του ΕΘΕΓ που αριθμούσε περίπου 10 εκ. λέξεις και περιείχε 252.084 τύπους. Η διαδικασία

[7] Ο ΕΘΕΓ αποτελεί ένα ΗΣΚ το οποίο αριθμεί στην παρούσα φάση του 29 εκ. λέξεις και καλύπτει τη γραπτή παραγωγή της σύγχρονης Νέας Ελληνικής. Αναπτύχθηκε από το Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ) και διατίθεται μέσω του Διαδικτύου στην ακόλουθη διεύθυνση: http://corpus.ilsp.gr

εντοπισμού μιας υποψήφιας λέξης για βασικό λεξιλόγιο είναι ίδια με αυτήν που χρησιμοποιείται για τον εντοπισμό των όρων ενός κειμένου (Μικρός 2004). Μια λέξη ξεχωρίζει όταν σε ένα θεματοποιημένο ΗΣΚ εμφανίζεται σε υψηλές συχνότητες και σε ένα ΗΣΚ γενικής γλώσσας εμφανίζεται σε μικρές (ή και καθόλου). Η όλη διαδικασία μπορεί να εξηγηθεί σχηματικά στον παρακάτω πίνακα:

Πίνακας 1: Σύγκριση της συχνότητας εμφάνισης μιας λέξης σε δύο διαφορετικά ΗΣΚ

Λέξη: πληροφορίες	**Θεματοποιημένο ΗΣΚ**	**Γενικό ΗΣΚ**
Συχνότητα εμφάνισης	1.061	2.372
% (επί του συνόλου των λέξεων του ΗΣΚ)	0,2	0,02
Συχνότητα μη - εμφάνισης	408.188	9.864.424
% (επί του συνόλου των λέξεων του ΗΣΚ)	99,8	99,98
Σύνολο (εμφανίσεις)	409.249	9.866.796
Σύνολο (ποσοστό)	100	100

Η συγκεκριμένη λέξη αν και εμφανίζεται τουλάχιστον διπλάσιες φορές στο Γενικό ΗΣΚ αποτελεί μια υποψήφια λέξη-κλειδί καθώς το ποσοστό εμφάνισής της στο ειδικό ΗΣΚ είναι δεκαπλάσιο από το ποσοστό εμφάνισης στο ΗΣΚ γενικής γλώσσας. Στην πιο γενική του μορφή ο Πίνακας 1 μπορεί να επαναγραφεί ως εξής:

Πίνακας 2: Γενική μορφή του πίνακα σύγκρισης συχνοτήτων εμφάνισης μιας λέξης σε δύο ΗΣΚ

	Θεματοποιη-μένο ΗΣΚ	**Γενικό ΗΣΚ**	**Σύνολο**
Συχνότητα λέξης (Σ)	α	β	α+β
Συχνότητα υπόλοιπων λέξεων (-Σ)	γ-α	δ-β	γ+δ-α-β
Σύνολο	γ	δ	γ+δ

Η στατιστική σύγκριση της εμφάνισης μιας λέξης στα δύο ΗΣΚ μπορεί να γίνει με μια σειρά από μεθόδους οι οποίες είναι ευρύτερα γνωστές ως μετρήσεις λεξικής συνάφειας - ΜΛΣ (lexical association measures). Η μαθηματική λογική που υφέρπει των περισσότερων ΜΛΣ στηρίζεται στην μηδενική υπόθεση ότι οι πραγματώσεις του λεξικού ζεύγους $\alpha \sim \beta$ είναι ανεξάρτητες μεταξύ τους. Η συγκεκριμένη μηδενική υπόθεση εξετάζεται στη συνέχεια με τη βοήθεια ποικίλων στατιστικών μεθόδων για να ελεγχθεί κατά πόσο οι παρατηρούμενες συχνότητες του συγκεκριμένου λεξικού

ζεύγους (α ~ β) αποκλίνουν με στατιστικά σημαντικό τρόπο από τις αναμενόμενες συχνότητες, αυτές δηλ. που θα περιμέναμε να συναντήσουμε αν ίσχυε η μηδενική υπόθεση.

Από τα διάφορα ΜΛΣ που έχουν κατά καιρούς προταθεί επιλέξαμε το Log Likelihood (LL) για λόγους που αναφέρονται αναλυτικά στο Ιακώβου, Μαρκόπουλο και Μικρό (2004).

Ο υπολογισμός του LL βάσει του γενικού πίνακα 2 μπορεί να γίνει ως εξής:

$$LL= 2(\alpha\log(\alpha) + \beta\log(\beta) + \gamma\log(\gamma) + \delta\log(\delta) - (\alpha + \beta)\log(\alpha + \beta) - (\alpha + \gamma)\log(\alpha + \gamma) - (\beta + \delta)\log(\beta + \delta) - (\gamma + \delta)\log(\gamma + \delta) + (\alpha + \beta + \gamma + \delta)\log(\alpha + \beta + \gamma + \delta))$$ [8]

Τα αποτελέσματα από την εφαρμογή του LL στα δεδομένα μας οργανώθηκαν με τον τρόπο που εμφανίζονται παρακάτω (Πίνακας 3) :

Πίνακας 3: Εξαγωγή βασικού λεξιλογίου σε Θεματοποιημένο ΗΣΚ βάσει του LL

Α/Α	*Λέξη*	*Συχνότητα Θ- ΗΣΚ*	*% Θ- ΗΣΚ*	*Συχνότητα ΗΣΚ Γενικής Γλώσσας.*	*% ΗΣΚ Γενικής Γλώσσας*	*LL*	*p*
1	ευρώ	1.162	0,27	9		7.258,00	<0,001
2	χρυσή	980	0,23	207		5.128,30	<0,001
3	σας	2.061	0,47	5.672	0,06	4.582,70	<0,001
4	ευκαιρία	1.016	0,23	1.459	0,01	3.211,30	<0,001
5	τιμή	961	0,22	1.442	0,01	2.978,30	<0,001
6	επιδερμίδα	475	0,11	17		2.862,60	<0,001
7	καλλυντικά	488	0,11	60		2.717,90	<0,001
8	αγγελία	448	0,1	16		2.700,10	<0,001
9	πληροφορίες	1.061	0,24	2.372	0,02	2.680,80	<0,001
10	τής	604	0,14	333		2.635,60	<0,001
11	διαφήμιση	536	0,12	404		2.145,70	<0,001
12	έπιπλα	400	0,09	102		2.035,80	<0,001
13	δέρμα	387	0,09	169		1.783,00	<0,001
14	μαλλιά	430	0,1	373		1.646,80	<0,001

Από τους 47.039 λεξικούς τύπους που αριθμούσε συνολικά το θεματοποιημένο ΗΣΚ εντοπίστηκαν με τη χρήση του LL 3.784 λεξικοί τύποι (8% του συνολικού λεξιλογίου του θεματικού ΗΣΚ) οι οποίοι αποτέλεσαν το βασικό λεξιλόγιο[9] της ενότητας «Αγορές» με το οποίο ασχοληθήκαμε στην παρούσα έρευνα. Επιπλέον, για κάθε όρο υπολογίσαμε τον αριθμό των κειμένων στα οποία αυτός εμφανίζεται και καταλήξαμε σε έναν δείκτη κειμενικής διασποράς ο οποίος μας αποκάλυψε πόσο διευρυμένη ή

[8] Ο υπολογισμός του LL μπορεί να γίνει αυτόματα μέσω Διαδικτύου στην ακόλουθη διεύθυνση: http://lingo.lancs.ac.uk/llwizard.html

[9] Αν χρειαστεί να περιοριστεί ο αριθμός των εξαγομένων λέξεων μπορεί να γίνει χρήση του τεστ χ^2 για να καθοριστεί το υποσύνολο εκείνο του λεξιλογίου που θα αποτελεί το «βασικότερο του βασικού λεξιλογίου». Για τη σχετική μεθοδολογία βλ. Berber (1999).

στενή είναι η χρήση μιας λέξης. Η συγκεκριμένη μέτρηση αξιοποιήθηκε ιδιαίτερα στην εκτίμηση της χρήσης συγκεκριμένων γραμματικών δομών, όπως η χρήση των κλιτικών για την οποία θα γίνει ιδιαίτερη αναφορά παρακάτω.

3 Αποτελέσματα

Ως γενικό συμπέρασμα η χρήση του στατιστικού δείκτη LL παράγει αξιόπιστο βασικό λεξιλόγιο το οποίο κρίνεται κατάλληλο για ενσωμάτωση σε διδακτικές δραστηριότητες. Η μεθοδολογία που περιγράφηκε, τέλος, μπορεί να χρησιμοποιηθεί και πέρα από τη συγκεκριμένη θεματική περιοχή που επιλέχθηκε πιλοτικά και να εφαρμοστεί και σε άλλες θεματικές περιοχές δίνοντας τις κατευθυντήριες γραμμές για μια θεματοποιημένη προσέγγιση του βασικού λεξιλογίου των πρώτων επιπέδων ελληνομάθειας. Η ποσοτική και ποιοτική εξέταση του βασικού λεξιλογίου που προέκυψε από την παρούσα έρευνα οδήγησε σε επιπλέον συμπεράσματα που αφορούν στην διδακτική του αξιοποίηση και συνοψίζονται ως εξής:

α) οι πρώτες θέσεις καταλαμβάνονται από γραμματικές λέξεις (grammatical words) σε αντιδιαστολή προς τις λέξεις πλήρεις περιεχομένου (content words), καθώς μόλις στην 33η θέση απαντά λέξη αυτής της κατηγορίας: *το ευρώ*. Ωστόσο, και οι υπόλοιπες λέξεις υψηλής συχνότητας - *πληροφορίες* (36η θέση), *ευκαιρία* (37η θέση), *χρυσή* (ως σύναψη με την προηγούμενη: 39η θέση), *τιμή* (40η θέση)- φαίνεται ότι βρίσκονται τόσο υψηλά στην ΛΣΛ ακριβώς γιατί λειτουργούν για τη συγκεκριμένη θεματική περιοχή όπως οι γραμματικές λέξεις λειτουργούν για κάθε κομμάτι λεξιλογίου. Στην περίπτωση αυτή είναι το ίδιο σημαντικές για την οριοθέτηση βασικών λειτουργιών του θεματικού κύκλου ΑΓΟΡΕΣ (αξία αγορών, επίτευξη χαμηλού κόστους, πληροφόρηση…). Σε υψηλές, επομένως, θέσεις από άποψη συχνότητας αναμένονται λεξιλογικά στοιχεία χαρακτηρισμένα τόσο από το θεματικό πεδίο όσο και από τα κειμενικά είδη που το απαρτίζουν (πρβ. τον αριθμό των διαφημίσεων και των αγγελιών επί του συνολικού αριθμού των κειμένων)

β) Αντίθετα, με χαμηλό βαθμό συχνότητας απαντούν λεξιλογικά στοιχεία, τα οποία μπορούν να θεωρηθούν με μια ευρύτερη έννοια του όρου η «μεταγλώσσα» του συγκεκριμένου θεματικού πεδίου. Ειδικότερα, παρατηρούμε ότι λέξεις που αποδίδουν τις υπερκείμενες κατηγορίες όπως «εμπορεύματα» (αντί για τα προς εμπορεία είδη, όπως «καλλυντικά», «αυτοκίνητο», «καφές», «κινητό»), «συναλλαγή» (αντί για τις σχετικές πράξεις και ενέργειες, όπως περιγράφονται μέσα από ρήματα «αγοράζω», «προσφέρω»), «αγορά» (ως αφηρημένη ενέργεια στον ενικό αριθμό αντί του

πληθυντικού «αγορές» που παραπέμπει ταυτόχρονα σε συγκεκριμένο αντικείμενο αναφοράς) έχουν πολύ χαμηλό βαθμό συχνότητας. Η παρατήρηση αυτή φαίνεται να ενισχύει την άποψη ότι η υψηλή συχνότητα εμφάνισης μιας λέξης παραπέμπει στον πυρήνα ενός σημασιολογικού πεδίου παρέχοντας σημαντικές πληροφορίες για τη βασικότητα των στοιχείων που τον συνιστούν σε σχέση προς τις εννοιολογικές ανάγκες που καλύπτουν.

γ) υψηλό είναι και το ποσοστό που εμφανίζουν τα ξένα λεξιλογικά στοιχεία με τα οποία οι μαθητές μπορεί να είναι ιδιαίτερα εξοικειωμένοι από την προηγούμενη γλωσσική τους εμπειρία (για τη συγκεκριμένη θεματική ενότητα καλύπτουν ένα ποσοστό της τάξης του 27,63% επί των συνολικών λεξιλογικών τύπων), στον βαθμό που αυτά μπορούν να λειτουργήσουν με ποικίλους τρόπους στη μετάβαση των μαθητών στο λεξιλογικό σύστημα της γλώσσας-στόχου.

δ) οι επικοινωνιακές ανάγκες της συγκεκριμένης θεματικής περιοχής επηρεάζουν και τη συχνότητα εμφάνισης τόσο των δομών όσο και των μορφολογικών χαρακτηριστικών με τα οποία απαντούν οι διάφορες γλωσσικές της πραγματώσεις. Ειδικότερα, παρατηρούμε ότι στο υλικό που συγκεντρώθηκε οι ρηματικοί τύποι απαντούν σε μεγάλο ποσοστό με εκφορές δευτέρου και τρίτου προσώπου, δεδομένης της νοηματικής ιδιομορφίας της θεματικής περιοχής που επιλέχθηκε (διαφημίσεις ως κίνητρο αγορών, περιγραφή των σχετικών προϊόντων, αμεσότητα συναλλαγής). Είναι χαρακτηριστικό ότι ακόμα και για τα κλιτικά, ο τύπος «σας», στη δευτεροπρόσωπη εκφορά του, παρουσιάζει τη μεγαλύτερη διασπορά σε σχέση με τα λοιπά λεξιλογικά στοιχεία (12,41%), όπως και τον υψηλότερο βαθμό σημαντικότητας σε σχέση με τις λοιπές γραμματικές λέξεις (4.582,7), γεγονός που καθιστά το σημασιολογικό του περιεχόμενο (αυτό του αποδέκτη του μηνύματος) κυρίαρχο για τη συγκεκριμένη θεματική περιοχή. Μια τέτοιου τύπου προσέγγιση φαίνεται να ανατρέπει την παραδοσιακή εστίαση στη διδασκαλία του πρώτου προσώπου των ρηματικών τύπων ως κατ'εξοχήν φορέα σημασιολογικών και μορφολογικών χαρακτηριστικών, καθώς οι πρωτοπρόσωπες εκφορές καλύπτουν μόλις το 2,39% των συνολικών ρηματικών εμφανίσεων. Είναι μάλιστα αξιοσημείωτο ότι στη μορφή αυτή απαντά μόνο το απόλυτα ταυτιζόμενο με το συγκεκριμένο θεματικό πεδίο ρήμα «αγοράζω» με 11 μόνο εμφανίσεις στην 4149^{η} θέση. Το ίδιο ισχύει και για τον αριθμό με τον οποίο εμφανίζονται οι ρηματικοί τύποι. Η ιδιομορφία της θεματικής περιοχής προκρίνει υψηλές συχνότητες στον πληθυντικό αριθμό των δύο πρώτων προσώπων με 91,21% και 82,18% αντίστοιχα, γεγονός που δημιουργεί υψηλές απαιτήσεις από πλευράς

κατανόησης και παραγωγής ως προς τη λειτουργικότητα και των σχηματισμό των συγκεκριμένων ρηματικών καταλήξεων. Ανάλογες είναι και οι παρατηρήσεις μας σχετικά με τον χρόνο και τον τρόπο εκφοράς των ρηματικών τύπων, καθώς η θεματική αυτή ενότητα μπορεί να θεωρηθεί πολύ καλή ευκαιρία τόσο για τη διδασκαλία χαρακτηριστικών τύπων του Ενεστώτα (απαντούν στο υλικό σε ποσοστό 77,9% έναντι 22,1% των εκφορών σε Αόριστο) όσο και της Συνοπτικής Ρηματικής άποψης σε ποικίλες δομές (αποτελεί για παράδειγμα το 57,3% των δομών σε Προστακτική).

Επομένως, τα στοιχεία αυτά, χαρακτηριστικά μιας θεματοποιημένης προσέγγισης του βασικού λεξιλογίου, μπορούν να δώσουν στη γλωσσική διδασκαλία τις εξής κατευθύνσεις:

α) δυνατότητα σπονδυλωτής διδασκαλίας θεματικών ενοτήτων ανάλογων προς τα ενδιαφέροντα και τις ανάγκες των μαθητών, οι οποίοι θα μπορούν με τον τρόπο αυτό να διαθέτουν έναν κατάλογο με τα απαραίτητα βασικά λεξιλογικά στοιχεία για την πραγμάτωση των επικοινωνιακών λειτουργιών σε κάθε θεματικό κύκλο.

β) εξοικείωση των μαθητών με αυθεντικό υλικό ανά θεματική ενότητα

γ) εστίαση στις γραμματικές δομές που καλύπτουν τις επικοινωνιακές ανάγκες της κάθε θεματικής ενότητας. Η υψηλή συχνότητα[10] εμφάνισης συγκεκριμένων δομών για κάθε θεματικό κύκλο αποτελεί χρήσιμο οδηγό στην προσπάθεια μιας περισσότερο επικοινωνιακής προσέγγισης στη διδασκαλία της ΝΕ ως ξένης γλώσσας.

[10] Για τη σχέση συχνότητας στη διδακτική προσέγγιση της γραμματικής, βλ. Biber, D. & R. Reppen (2002).

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Barsalou, L. (1992) "Frames, Concepts and Conceptual Fields" στο Leher, A. & E. F. Kittay (eds) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*, Lawrence Elbaum Associates, Pb.

Berber, S. T. (1999) "Using KeyWords in text analysis: Practical aspects" *DIRECT Papers*, 42, 1-8.

Biber, D. & R. Reppen (2002) "What does frequency have to do with grammar teaching?" *Studies in Second Language Acquisition*, 24, 199-208.

Biber, D., S. Conrad & R. Reppen (1998) *Corpus Linguistics*: *Investigating Language Structure and Use.* Cambridge, CUP.

Carter, R. (1996[2]) *Vocabulary: Applied Linguistics Perspectives*, London: Routledge.

Daille, B. (1995) "Combined approach for terminology extraction: lexical statistics and linguistic filtering" Technical Report, 5, UCREL, Lancaster University.

Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence" *Computational Linguistics,* 19, 61-74.

Fillmore, C. (1992) "Corpus Linguistics" or " Computer-aided armchair Linguistics". Στο Svartvik, J. (eds) *Directions in Corpus Linguistics.* Berlin, Mouton de Gruyter.

Hofland, K. & S. Johansson (eds) (1982) *Word frequencies in British and American English.* The Norvegian Computing Center for the Humanities, Bergen, Norway.

Ιακώβου, Μ. Μαρκόπουλος, Γ. & Μικρός, Γ. (2004) «Θεματοποιημένο Βασικό Λεξιλόγιο μέσω ΗΣΚ: Πρακτική εφαρμογή στη διδασκαλία της ΝΕ ως ΞΓ». Στο *Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας.* Διαθέσιμο: http://www.philology.uoc.gr/conferences/6thICGL/ebook/a/iakovou&markopoulos&mikros.pdf (3 Ιουνίου 2004).

Kennedy, G. (1998) *An Introduction to Corpus Linguistics.* New York, Longman.

Kilgariff, A. (2001) "Comparing corpora" *International Journal of Corpus Linguistics*, 6, 97-133.

Leech, G. & R. Fallon (1992) "Computer corpora – What do they tell us about culture?" *ICAME Journal*, 16, 29-50.

Leech, G. (1992) "Corpora and theories of linguistic performance". Στο Svartvik, J. (eds) *Directions in Corpus Linguistics,* Berlin, Mouton de Gruyter.

McEnery, T. & Wilson, A. (1996) *Corpus Linguistics.* Edinburgh, Edinburgh University Press.

Meunier, F. (1998) "Computer Tools for the Analysis of Learner Corpora". Στο Granger, S. (eds) *Learner English on Computer,* New York, Longman.

Μικρός, Γ. (2004) «Ηλεκτρονικά Σώματα Κειμένων και Ορολογία». Στο Κατσογιάννου, Μ. & Ε. Ευθυμίου (εκδ.), *Ελληνική ορολογία: έρευνα και εφαρμογές,* Αθήνα, Καστανιώτης.

Nation, I.S.P (2001) *Learning Vocabulary in Another Language,* Cambridge Applied Linguistics.

Nation, P. (1990) *Teaching and Learning Vocabulary.* Heinle & Heinle.

Owen, F. & R. Jones (1977) *Statistics,* Polytech Publishers.

Rayson, P., G. Leech & M. Hodges (1997) "Social differentiation in the use of English vocabulary: some analysis of the conversational component of the British National Corpus" *International Journal of Corpus Linguistics,* 2, 133-152.

Schmitt, N. (2000) *Vocabulary in Language Teaching*, Cambridge University Press.

Sinclair, J.M. (1991) "Shared knowledge". Στο Alatis, J.E. (eds) *Linguistics and language pedagogy: the state of the art,* Washington, D.C., Georgetown University Press.

Sinclair, J.M. (1997) "Corpus Evidence in Language Description". Στο Wichmann, A., S. Fligelstone, T. McEnery, & G. Knowles, (eds) *Teaching and language corpora,* New York, Longman.

Tribble, C. (2001) "Corpora and teaching: adjusting the gaze" *ICAME 2001 Future Challenges in Corpus Linguistics*, Louvain-la-Neuve, Belgium, 16-20 May 2001.

Wichmann, A., et al. (1997) *Teaching and Language Corpora,* Longman: London & New York.

Widdowson, H.G. (1991) "The description and prescription of language". Στο Alatis, J.E. (eds) *Linguistics and language pedagogy: the state of the art,* Washington, D.C., Georgetown University Press.