ΣΩΜΑ ΕΛΛΗΝΙΚΩΝ ΚΕΙΜΕΝΩΝ: ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ

Διονύσης Γούτσος ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Abstract

The paper is the first public presentation of the research programme "Basic Corpus of Greek Texts", under the co-operation of the University of Athens and the University of Cyprus, aiming at building a new, extensive and representative corpus of Greek. In particular, the Corpus of Greek Texts (CGT) is envisaged as collecting a substantial amount of data (30 million words) in a short time span (1-2 years) as a basis for linguistic research and a resource for teaching applications. The scope and representativeness of the genres included, as well as free accessibility to it, will make CGT one of the most necessary tools for the study of Greek. The paper presents the research area, the aims and needs of the programme, the identity and structure of the CGT, as well as the methodological issues and linguistic implications and applications related with the compilation of the corpus.

Λέξεις - κλειδιά

υπολογιστική γλωσσολογία, ηλεκτρονικά σώματα κειμένων (ΗΣΚ)

1. Εισαγωγή

Η ανακοίνωση αυτή αποτελεί την πρώτη δημόσια παρουσίαση του ερευνητικού Προγράμματος «Βασικό Σώμα Ελληνικών Κειμένων», που χρηματοδοτείται από το Πανεπιστήμιο Κύπρου και εκπονείται σε συνεργασία του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών με το Πανεπιστήμιο Κύπρου. Σε αυτό συμμετέχουν ερευνητές από τα δύο πανεπιστήμια και συνεργάζονται ακαδημαϊκοί από πανεπιστήμια του εξωτερικού, καθώς και μεταπτυχιακοί και ερευνητικοί αλλά και εμπορικοί συνεργάτες.[1] Το Πρόγραμμα αποβλέπει στη σύσταση του Σώματος Ελληνικών Κειμένων (ΣΕΚ), ενός νέου ηλεκτρονικού σώματος κειμένων (ΗΣΚ) που στοχεύει να αποτελέσει σημείο αναφοράς για τη μελέτη της ελληνικής γλώσσας. Στην ανακοίνωση θα αναφερθούμε στην ερευνητική περιοχή, τους στόχους, την ταυτότητα και τη δομή του ΣΕΚ, καθώς και τις διαδικασίες που συνδέουν το σχεδιασμό με την υλοποίηση και την περαιτέρω ανάπτυξή του ΗΣΚ.

Το πρόγραμμα για τη σύσταση του ΣΕΚ αποτελεί βασική εφαρμογή στα Ελληνικά των πιο πρόσφατων εξελίξεων στο χώρο της υπολογιστικής γλωσσολογίας και ειδικότερα της χρήσης των υπολογιστών στη συλλογή και επεξεργασία σωμάτων κειμένων (computer corpus linguistics). Η ικανότητα των ηλεκτρονικών υπολογιστών να αποθηκεύουν έναν τεράστιο όγκο γλωσσικών πληροφοριών και να επεξεργάζονται γλωσσικά δεδομένα, επιτρέποντας την ταχύτατη αναζήτηση, ανάκληση, ταξινόμηση, αυτόματο υπολογισμό στατιστικών στοιχείων κ.λπ. αποτελεί πραγματική επανάσταση στη γλωσσολογική έρευνα (Μπαμπινιώτης 1999: 168, Leech 1992, Sinclair 1991). Η χρήση των υπολογιστών αποτελεί ένα «καίριο βοήθημα στην έρευνα, όχι μόνο της γλώσσας, αλλά και της γραμματικοσυντακτικής της δομής» (Μπαμπινιώτης 1999: 170, πρβλ. Aarts 1991) και της επικοινωνιακής δυναμικής της γλώσσας

(έ.ά.: 219). Επιπλέον, όπως έχει υποστηριχθεί, δεν προσφέρει απλώς μια νέα μεθοδολογία για τη μελέτη της γλώσσας αλλά ένα νέο ερευνητικό χώρο και μια νέα φιλοσοφική προσέγγιση στο αντικείμενο (Leech 1992: 106).

Σύμφωνα με τον Sinclair (1996), ως σώμα κειμένων θεωρείται κάθε συλλογή τμημάτων μιας συγκεκριμένης γλώσσας, τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα της γλώσσας αυτής. Η χρήση σωμάτων κειμένων διαθέτει μια μακρόχρονη παράδοση στο χώρο της γλωσσολογίας, αν και έως τις τελευταίες δεκαετίες, είχε υπερκερασθεί από άλλες μεθοδολογικές προσεγγίσεις που έδιναν έμφαση στη μη εμπειρική, διαισθητική συλλογή δεδομένων. Η ουσιαστική αναβίωση και ευρεία επικράτηση της χρήσης τους οφείλεται στη ραγδαία ανάπτυξη των δυνατοτήτων της τεχνολογίας που επέτρεψε τη δημιουργία ΗΣΚ, δηλαδή σωμάτων κειμένων κατάλληλων για ηλεκτρονική χρήση και ειδικά κωδικοποιημένων για τυποποιημένες και ομοιογενείς εργασίες ανάκτησης γλωσσικής πληροφορίας (έ.ά).

Όπως επισημαίνει ο Kennedy, το κύριο αίτημα της σύγχρονης έρευνας είναι «ένα συστηματικό και περιεκτικό πρόγραμμα έρευνας της δομής και της χρήσης των διαφόρων γλωσσών, που θα περιλαμβάνει την εύκολη πρόσβαση στα αποτελέσματα της έρευνας» (1998: 291). Για το σκοπό αυτό είναι απαραίτητη η δημιουργία εκτεταμένων και σύγχρονων ΗΣΚ, που θα προσφέρουν πρόσβαση σε ένα αντιπροσωπευτικό δείγμα κάθε γλώσσας. Πρέπει να σημειωθεί, ωστόσο, ότι, ενώ η επεξεργασία ευρωπαϊκών γλωσσών, όπως η αγγλική, γαλλική και γερμανική, έχει προχωρήσει σημαντικά, όπως θα διαπιστώσουμε πιο κάτω, για τα Ελληνικά παραμένει επίκαιρο το όραμα του Θησαυρού ολόκληρης της Ελληνικής Γλώσσας (Μπαμπινιώτης 1999: 170). Το ερευνητικό πρόγραμμα για τη δημιουργία του ΣΕΚ αποσκοπεί στο να θέσει τις βάσεις για την επίτευξη αυτού του στόχου.

2. Στόχοι του ερευνητικού προγράμματος

Σκοπός του προγράμματος είναι η συγκρότηση του Σώματος Ελληνικών κειμένων, μιας συλλογής σύγχρονων ελληνικών κειμένων, αποθηκευμένων σε ηλεκτρονική μορφή, αντιπροσωπευτικής των σημαντικότερων κειμενικών ειδών της γλώσσας, που θα αποβλέπει στη γλωσσολογική ανάλυση και σε εκπαιδευτικές εφαρμογές. Τα κύρια χαρακτηριστικά του ΣΕΚ είναι τα ακόλουθα:

- θα αποτελεί μια σαφώς καθορισμένη συλλογή κειμένων από μια ποικιλία κειμενικών ειδών πρωταρχικής σημασίας σε περιβάλλοντα επικοινωνίας στα Ελληνικά και σημαντικών για τη διδασκαλία της Ελληνικής ως μητρικής/ ξένης γλώσσας
- θα περιέχει ένα σημαντικό ποσοστό προφορικών δεδομένων, έτσι ώστε να αποτελέσει τη μεγαλύτερη υπάρχουσα συλλογή προφορικών κειμένων της Ελληνικής
- θα περιέχει ένα σημαντικό ποσοστό δεδομένων από τον κυπριακό χώρο, αναδεικνύοντας για πρώτη φορά τη γεωγραφική ποικιλία της Ελληνικής
- θα σχεδιαστεί ως βάση για μεγαλύτερα (π.χ. γενικά, πολύγλωσσα κ.λπ.) μελλοντικά σώματα κειμένων

- θα προσφέρεται σε μελετητές και σπουδαστές της γλώσσας μέσω εφαρμογών φιλικών προς το χρήστη.

Συνοπτικά, το πρόγραμμα στοχεύει στη συλλογή μιας σημαντικής ποσότητας δεδομένων (30 εκατομμύρια λέξεις) σε σύντομο χρονικό διάστημα (1-2 έτη) που θα αποτελέσει τη βάση της γλωσσολογικής έρευνας και πηγή πληροφοριών για διδακτικές εφαρμογές που σχετίζονται με την Ελληνική. Το εύρος και η αντιπροσωπευτικότητα των κειμενικών ειδών που περιλαμβάνονται, καθώς και η ελεύθερη πρόσβαση στο ευρύτερο κοινό θα καταστήσουν το ΣΕΚ ένα σημαντικό εργαλείο για τη μελέτη της ελληνικής γλώσσας.

Το ΣΕΚ θα καλύψει την ανάγκη για μια ικανού μεγέθους και αντιπροσωπευτική συλλογή ελληνικών κειμένων που θα μπορεί να συγκρίνεται για ερευνητικούς σκοπούς με τις αντίστοιχες συλλογές για τις ευρωπαϊκές γλώσσες. Οι υπάρχουσες συλλογές Ελληνικής είναι μικρού μεγέθους και εξειδικευμένες ή μη αντιπροσωπευτικές, καθώς δεν διαθέτουν αρκετά προφορικά δεδομένα (βλ. Goutsos, King και Hatzidaki 1994, για μια συνολική εκτίμηση). Επιπλέον, έχει σημειωθεί λίγη πρόοδος από τις αρχές της δεκαετίες του ΄90, στην οποία ανάγονται οι περισσότερες συλλογές κειμένων (έ.ά.) που δεν ξεπερνούν συνολικά τα 10 εκατομμύρια λέξεις. Ακόμη πιο σημαντικό είναι ότι, ακόμη και στις περιπτώσεις όπου υπάρχουν εκτεταμένες συλλογές κειμένων (π.χ. ΕΘΕΓ), δεν έχει γίνει σαφές το περιεχόμενο των συλλογών και ο βαθμός αντιπροσωπευτικότητάς τους αλλά και – κυρίως – οι συλλογές δεν είναι άμεσα προσβάσιμες από το ερευνητικό κοινό, τους δασκάλους και τους μαθητές της γλώσσας, μέσω δημόσιας, ελεύθερης πρόσβασης.

Το Βασικό Σώμα Ελληνικών Κειμένων σχεδιάζεται για να καλύψει το χάσμα μεταξύ της Ελληνικής και άλλων ευρωπαϊκών γλωσσών, για τις οποίες υπάρχουν ήδη μεγάλου μεγέθους, αντιπροσωπευτικές συλλογές κειμένων. Ειδικότερα, όσον αφορά το μέγεθος των συλλογών και τη συμμετοχή των προφορικών δεδομένων σε αυτές, είναι ενδεικτικά τα ακόλουθα στοιχεία:

Αγγλικά:	BNC Corpus: 100 εκατ. λέξεις (10 εκατ. προφορικά δεδομένα)
	Bank of English Corpus: περίπου 329 εκατ. λέξεις
	(60 εκατ. προφορικά δεδομένα)
	Cancode corpus: 5 εκατ. λέξεις προφορικά δεδομένα
Γαλλικά:	Ottawa-Hull: 3,5 εκατ. λέξεις προφορικά δεδομένα
	ELRA Parole Corpus: 20 εκατ. λέξεις
	TLF: 150 εκατ. λέξεις γραπτά δεδομένα
Γερμανικά:	Mannheim Corpus: 8 εκατ. λέξεις
	Muenster Textbank: 94 εκατ. λέξεις
Ιταλικά:	Pisa Corpus: 10 εκατ. λέξεις
Ισπανικά:	Corpus Oral De Referencia Del Espanol: 1,1 εκατ. προφορικά δεδομένα
	Mark Davies Modern Newspapers: 35 εκατ. λέξεις
Πορτογαλικά:	Mark Davies Modern Newspapers 26 εκατ. λέξεις
Ολλανδικά:	INL 1995 27 εκατ.
	INL 1996 38 εκατ. λέξεις
Σουηδικά:	10 εκατ. προφ. δεδομένα συνολικά (βλ. Goutsos, Hatzidaki, King 1994)

Πρέπει να τονιστεί εδώ ότι οι περισσότερες από τις πιο πάνω συλλογές όχι μόνο είναι προσιτές (δωρεάν ή με συνδρομή) στο κοινό, αλλά και διαθέτουν μια πλήρη εικόνα των δεδομένων που

αποτελούν τη συλλογή, προφορικών και γραπτών, και των κειμενικών ειδών από τα οποία προέρχονται.

Το Σώμα Ελληνικών Κειμένων θα καλύψει το χάσμα των δεδομένων για τα Ελληνικά με αυτά που υπάρχουν για τις μεγαλύτερες ευρωπαϊκές γλώσσες και θα συμβάλει έτσι σημαντικά στη διαγλωσσική έρευνα. Αυτό θα επιτευχθεί τόσο με το μέγεθος όσο και με τη σύσταση της συλλογής (βλ. παρακάτω, πρβλ. Renouf 1987 για το σχεδιασμό, εντοπισμό και απόκτηση κειμένων). Ο συνολικός αριθμός λέξεων που προτείνεται (30 εκατομμύρια) είναι ικανοποιητικός, τουλάχιστον σε πρώτη φάση και για σημαντικές εφαρμογές. Ας σημειωθεί, για παράδειγμα, ότι η πρώτη έκδοση του πρωτοποριακού Λεξικού Cobuild, που βασίστηκε εξ ολοκλήρου σε ΗΣΚ, άντλησε από μια βάση δεδομένων μόλις 20 εκατομμυρίων λέξεων (Sinclair 1987). Αποτελεί εκτίμησή μας ότι το ΣΕΚ θα καλύψει της ανάγκες της γλωσσολογικής έρευνας στα ελληνικά για τουλάχιστον την επόμενη δεκαετία. Επιπλέον, το ΣΕΚ αναμένεται να αποτελέσει κεντρική πηγή ανάπτυξης αυθεντικού διδακτικού υλικού για την Ελληνική, μια περιοχή στην οποία υπάρχει έλλειψη σε σχέση με τις άλλες ευρωπαϊκές γλώσσες (Georgakopoulou και Goutsos 1998).

3. Ταυτότητα και περιγραφή του ΣΕΚ

Με βάση τις ανάγκες που διαπιστώθηκαν και τους στόχους που τέθηκαν παραπάνω, το ΣΕΚ θα είναι ένα ηλεκτρονικό σώμα κειμένων

- Γενικό
- Μονόγλωσσο
- Συγχρονικό
- Μεικτό

Συγκεκριμένα, το ΣΕΚ συλλέγει ολόκληρα κείμενα από ένα πλήθος κειμενικών ειδών και όχι μια εξειδικευμένη γλωσσική περιοχή. Είναι μονόγλωσσο, στη φάση αυτή δηλαδή δεν περιέχει κείμενα σε άλλη γλώσσα εκτός της ελληνικής αλλά και μεταφρασμένα κείμενα στην ελληνική. Περιέχει κείμενα από μια συγχρονία της ελληνικής και συγκεκριμένα μετά το 1990 και έως σήμερα. Τέλος, περιέχει τόσο προφορικά όσο και γραπτά κείμενα.

Η γενική σύνθεση του ΣΕΚ παρουσιάζεται στον ακόλουθο πίνακα:

Στόχος αριθμού λέξεων:	30 εκατομμύρια λέξεις	
		ΠΗΓΗ
Προφορικά δεδομένα:	3 εκατ. λέξεις (10 %)	
Αυθόρμητη συνομιλία:	0,5 εκατ.	Μεταγραφή
Δημόσιες συνεντεύξεις:	1,5 εκατ.	Διαδίκτυο
Ραδιοφωνική, τηλεοπτική συνομιλία:	1 εκατ.	Μεταγραφή
Γραπτά δεδομένα:	27 εκατ. λέξεις (90 %)	
Βιβλία: λογοτεχνία:	5 εκατ.	Εκδότες
Βιβλία: ενημερωτικά:	5 εκατ.	Εκδότες
Ακαδημαϊκή γραφή:	5 εκατ.	Πανεπιστήμια
Τύπος: νέα:	5 εκατ.	Διαδίκτυο
Τύπος: απόψεις:	5 εκατ.	Διαδίκτυο
Επίσημα έγγραφα:	2 εκατ.	Διαδίκτυο

Τονίζεται ότι στο σύνολο των δεδομένων προβλέπονται 3 εκατομμύρια λέξεις από γραπτές και προφορικές πηγές του κυπριακού χώρου (1 εκατομμύριο προφορικά και 2 εκατομμύρια λέξεις γραπτά δεδομένα).

Επίσης, πρέπει να σημειωθεί εδώ ότι ο παραπάνω πίνακας είναι απλά ενδεικτικός για τη σύνθεση του ΣΕΚ. Η κωδικοποίηση των δεδομένων ακολουθεί μια λεπτομερέστερη ταξινόμηση με βάση διάφορες διαστάσεις των κειμένων (κειμενικός τρόπος, γένος, είδος, μέσο κ.λπ.) με αποτέλεσμα να διακρίνονται επιμέρους κατηγορίες. Το Παράρτημα παρουσιάζει αναλυτικά τις κατηγορίες αυτές και τους ενδεικτικούς στόχους για τον αριθμό λέξεων που πρέπει να συμπληρωθεί με βάση μελέτη που εκπονήθηκε από τους ερευνητικούς συνεργάτες του προγράμματος. Όπως θα τονιστεί και στη συνέχεια, οι κατηγορίες αυτές είναι ενδεικτικές και επιτρέπουν την αναταξινόμηση των δεδομένων και την ελεύθερη επιλογή υπο-συνόλων από αυτά.

4. Αντιπροσωπευτικότητα του ΣΕΚ

Ίσως το σημαντικότερο μεθοδολογικό ζήτημα που εγείρει η δημιουργία ενός ΗΣΚ συνδέεται με τη λεγόμενη αντιπροσωπευτικότητά του, το βαθμό δηλαδή στον οποίο επιτυγχάνει να θεωρείται «αντιπροσωπευτικό δείγμα» της γλώσσας, την οποία χρησιμοποιούν τα κείμενα που περιέχει, σύμφωνα με τον ορισμό του Sinclair (1996), στον οποίο αναφερθήκαμε πιο πάνω.[2] Το τι σημαίνει ακριβώς «αντιπροσωπευτικό δείγμα» αποτελεί σημείο προς συζήτηση και οι ερευνητές έχουν τοποθετηθεί με διαφορετικό τρόπο στο ζήτημα των γλωσσολογικών κριτηρίων με τα οποία επιλέγονται και διατάσσονται τα κείμενα κάθε ΗΣΚ. Σύμφωνα με τον Barnbrook (1996: 24), ένα τέτοιο δείγμα πρέπει να διαθέτει παρόμοια χαρακτηριστικά με τον γλωσσικό πληθυσμό που στοχεύει να αντιπροσωπεύσει στην ανάλυση μιας γλώσσας.

Στην πράξη, η δειγματοληψία, ιδίως σε περιπτώσεις ΗΣΚ που αντιπροσωπεύουν το γενικό πληθυσμό μιας γλώσσας όπως το ΣΕΚ, μπορεί να πάρει διαφορετικές μορφές Είναι χαρακτηριστικό ότι ΗΣΚ όπως το αγγλικό BNC (βλ. πιο πάνω) στηρίζεται σε μια αυστηρή ταξινόμηση κειμενικών ειδών, που χρησιμοποιεί στατιστική δειγματοληψία για το κομμάτι των προφορικών δεδομένων, ενώ, αντίθετα, το Bank of English έχει εξελιχθεί σε ΗΣΚ ελέγχου (monitor corpus), μια τεράστια βάση δεδομένων που συνεχώς ανανεώνεται δίνοντας πρωτίστως έμφαση στο μέγεθος έτσι ώστε να καλύπτει τις απαιτήσεις ενός ΗΣΚ γενικής γλώσσας μέσω της ευρείας συλλογής ενός πλήθους κειμένων (βλ. Barnbrook 1996: 25). Πρόκειται για κεντρικά παραδείγματα των δύο διαφορετικών τρόπων διαστρωμάτωσης που επικρατούν στη διεθνή πρακτική και αναφέρονται στη διαστρωμάτωση με βάση δημογραφικά στοιχεία και στη διαστρωμάτωση με βάση την κειμενική τυπολογία (Biber, Conrad και Reppen 1998: 248). Στην περίπτωση του ΣΕΚ η διαστρωμάτωση γίνεται με βάση ποικίλα κειμενικά κριτήρια (όπως το κειμενικό είδος, τη θεματολογία, το γνωστικό αντικείμενο, το μέσο κ.λπ.), στοχεύοντας στην αναγνώριση ενός εύρους καταστασιακών ιδιωμάτων (Γεωργακοπούλου και Γούτσος 1999: 56) που αναγνωρίζονται διαισθητικά από τη γλωσσική κοινότητα.

Σύμφωνα με τη συζήτηση που γίνεται στο Kučera (2002) σε σχέση με το Εθνικό ΗΣΚ Τσεχικών, μπορούμε να θεωρήσουμε ότι η αντιπροσωπευτικότητα αναφέρεται σε τρεις

διαστάσεις κάθε ΗΣΚ, το μέγεθος, την αυθεντικότητα και τις αναλογίες, τη σχετική «ισορροπία» δηλαδή μεταξύ των κειμενικών ειδών που το απαρτίζουν. Το ΣΕΚ διαθέτει για λόγους σχεδιασμού ένα συγκεκριμένο μέγεθος (30 εκατομμύρια), που ασφαλώς δεν μπορεί να εξασφαλίσει πλήρη αντιπροσωπευτικότητα της γλώσσας σε αυτή τη φάση ανάπτυξής του. Ωστόσο, για λειτουργικούς σκοπούς και σε σχέση με τη μέχρι τώρα διεθνή εμπειρία, παρόμοιου μεγέθους ΗΣΚ έχουν εξυπηρετήσει με επάρκεια μεγάλης κλίμακας γλωσσολογικούς στόχους (π.χ. λεξικογραφικούς, όπως στην περίπτωση της πρώτης έκδοσης του Λεξικού Cobuild, που αναφέρθηκε πιο πάνω).

Επιπλέον, το ΣΕΚ καλύπτει ικανοποιητικά τις άλλες δύο διαστάσεις και διαθέτει, επιπλέον, προσεκτικό σχεδιασμό σε άλλες παραμέτρους. Πιο συγκεκριμένα, η ταυτότητα του ΣΕΚ διασφαλίζει ότι μόνο κείμενα που ικανοποιούν συγκεκριμένα κριτήρια και μόνο ολόκληρα κείμενα (όπου αυτό είναι δυνατόν) περιλαμβάνονται σε αυτό. Τα κείμενα αυτά προέρχονται από περιβάλλοντα επικοινωνίας κεντρικής σημασίας για τα Ελληνικά (βλ. Παράρτημα) και δημιουργημένα με φυσικό τρόπο (όχι κάτω από πειραματικές συνθήκες) έτσι ώστε να μπορούν να χαρακτηρίζονται αυθεντικά.

Κατά δεύτερο λόγο, το ΣΕΚ επιχειρεί να καλύψει μια πρωτόγνωρη ποικιλία κειμενικών ειδών, με ιδιαίτερη έμφαση στα προφορικά δεδομένα και τα κείμενα από την κυπριακή γεωγραφική ποικιλία, συμβάλλοντας έτσι σε μια πιο ολοκληρωμένη εικόνα της γλώσσας.[3] Στην ουσία, η αντιπροσωπευτικότητα εξαρτάται άμεσα και καίρια από το στόχο του ΗΣΚ, που στην περίπτωσή μας, αφορά τη γενική εικόνα της ελληνικής γλώσσας και τις ευρύτερες δυνατές εφαρμογές. Με αυτή την έννοια το εύρος των κειμενικών ειδών που συμπεριλαμβάνονται βρίσκεται σε άμεση συνάρτηση με το πλήθος των ερευνητικών στόχων που υπηρετούνται. Επιπλέον, η αναλογία των επιμέρους κειμενικών ειδών στο τμήμα του γραπτού λόγου έγινε με βάση μελέτης γλωσσικής πρόσληψης και, ειδικότερα, αναγνωσιμότητας, σύμφωνα με τα σχετικά στοιχεία του Εθνικού Κέντρου Βιβλίου.[4] Πρέπει επίσης να τονιστεί ότι βασικό μέλημα στο σχεδιασμό του ΣΕΚ υπήρξε η λεπτομερής και συστηματική κωδικοποίηση στοιχείων για κάθε κείμενο που περιλαμβάνεται, έτσι ώστε να υπάρχει άμεση πρόσβαση στα πρωτογενή στοιχεία που συνθέτουν την ταυτότητα κάθε κειμένου, επιτρέποντας τον έλεγχο της κειμενικής τυπολογίας που χρησιμοποιήθηκε. (Η δυνατότητα αυτή διαφοροποιεί σημαντικά το ΣΕΚ από τα υπάρχοντα ΗΣΚ της ελληνικής, των οποίων η σύνθεση και δομή δεν μπορεί να ελεγχθεί επαρκώς).

Τέλος, ένα από τα πιο σημαντικά χαρακτηριστικά του ΣΕΚ είναι η δυνατότητά του να χρησιμοποιηθεί ως αρχείο γλωσσικών δεδομένων, στην οποία συμβάλλει η ευελιξία με την οποία θα μπορούν να επιλεγούν και να συνδυαστούν επιμέρους υπο-σώματα. Με αυτή την έννοια, οι αριθμητικοί στόχοι που έχουν τεθεί για κάθε επιμέρους κατηγορία (βλ. Παράρτημα) μπορούν να θεωρηθούν ως μια προσωρινή και τυχαία επιλογή, η οποία μπορεί να αντικατασταθεί από μια άλλη επιλογή του χρήστη, ανάλογα με τις ανάγκες της συγκεκριμένης έρευνας. Έτσι, κείμενα που συγκεντρώνονται και υπερβαίνουν τους αριθμητικούς στόχους κάθε κατηγορίας αποκλείονται από την παρούσα σύνθεσή του, φυλάσσονται όμως για μια μεταγενέστερη φάση ανάπτυξης του ΣΕΚ. Γι′ αυτό το λόγο, κάθε κείμενο που πληροί τα συγκεκριμένα κριτήρια του ΣΕΚ είναι χρήσιμο και δεν αποκλείεται εκ των προτέρων.

5. Από το σχεδιασμό στην υλοποίηση

Η σειρά των διαδικασιών που σχετίζονται με την υλοποίηση του προγράμματος συνοψίζεται στον ακόλουθο πίνακα:

- Διερεύνηση πηγών/ ανάπτυξη
- Συλλογή δεδομένων
- Μεταγραφή (για προφορικά δεδομένα)
- Καθαρισμός και αποθήκευση
- Τυποποίηση
- Κωδικοποίηση
- Σχολιασμός (σε μελλοντική φάση)

Ειδικότερα, ένα μεγάλο μέρος του ερευνητικού προγράμματος αφιερώνεται στη διερεύνηση των πηγών και την ευρύτερη ανάπτυξη των γλωσσικών πόρων που συνδέονται με το ΣΕΚ. Έτσι εξερευνώνται ιστότοποι στο διαδίκτυο που μπορούν να προσφέρουν πολύτιμα δεδομένα και αναζητούνται διάφορες, ιδιωτικές και δημόσιες, πηγές που μπορούν να προσφέρουν κείμενα στο πρόγραμμα. Η συλλογή των δεδομένων γίνεται με μαγνητοφώνηση ή βιντεοσκόπηση, για τα προφορικά κείμενα, ηλεκτρονική σάρωση, δακτυλογράφηση ή χρήση διαδικτύου, για τα γραπτά δεδομένα. Στην πρώτη περίπτωση, ακολουθεί αναγκαστικά μεταγραφή, που στη συγκεκριμένη φάση είναι ορθογραφική με ευρύ χαρακτηρισμό φωνητικών στοιχείων και χαρακτηριστικών συνεχούς λόγου (επικάλυψη, διακοπή, παύση, επιμήκυνση κ.λπ.).[5]

Στη συνέχεια, γίνεται καθαρισμός των αρχείων από περιττά, μη λεκτικά στοιχεία (π.χ. εικόνες, γραμμές, κενά κ.λπ.) και αποθήκευση με τη μορφή ASCII. Ακολουθεί η τυποποίηση, που συνίσταται σε βασικό χαρακτηρισμό παραγράφων, ενοτήτων, τίτλων, ομιλητών κ.λπ., όπου αυτό είναι απαραίτητο. Με άλλα λόγια, περιλαμβάνονται στο αρχείο βασικές πληροφορίες για την κειμενική δομή. Τέλος, κωδικοποιούνται σε ανεξάρτητη βάση δεδομένων τα στοιχεία ταυτότητας του κειμένου που περιλαμβάνουν συγγραφέα, ημερομηνία παραγωγής, τίτλο, πρώτες λέξεις, αριθμό λέξεων κ.λπ., καθώς και λεπτομερή στοιχεία ταξινόμησης. Η ταξινόμηση στο ΣΕΚ είναι πολλαπλή και αναφέρεται στα ακόλουθα στοιχεία:

- Τρόπος: προφορικός-γραπτός λόγος
- Μέσο: ραδιόφωνο, τηλεόραση, ζωντανό, βιβλίο, τηλέφωνο, εφημερίδα, περιοδικό, ηλεκτρονικό, άλλο
- Γένος: πληροφορίας-μη πληροφορίας
- Είδος: ακαδημαϊκός λόγος, ενημερωτικά κείμενα, νόμοι-διοίκηση, ιδιωτικά κείμενα, λογοτεχνία, ειδήσεις, άρθρα γνώμης, συνέντευξη, δημόσια ομιλία, αυθόρμητη συνομιλία, διάφορα
- Υπο-είδος: 01-99
- Γεωγραφική ποικιλία: κοινή-κυπριακή
- Λέξεις-κλειδιά

Αυτό που πρέπει να τονιστεί ιδιαίτερα σε σχέση με τα στοιχεία ταξινόμησης είναι η ευελιξία, η οποία αναφέρθηκε και πιο πάνω. Με άλλα λόγια, η ταξινόμηση δεν είναι δεσμευτική για την

ανάκληση μιας υπο-ομάδας κειμένων· αντίθετα, επιτρέπει την πολλαπλή και διαφοροποιημένη σύνθεση υπο-σωμάτων ανάλογα με τις ανάγκες της έρευνας και τις μεθοδολογικές προτεραιότητες του ερευνητή. Έτσι, για παράδειγμα, ερευνητές που δεν θεωρούν χρήσιμη ή δεν συμφωνούν με την ταξινόμηση του συστήματος σε κειμενικό γένος (κείμενα πληροφορίας-μη πληροφορίας) μπορούν να παρακάμψουν την ταξινόμηση αυτή και να χρησιμοποιήσουν άλλα κριτήρια. Το ίδιο ισχύει και στην περίπτωση του συνεχούς προφορικού-γραπτού λόγου, στο οποίο το σημείο τομής μπορεί να τεθεί διαφορετικά από κάθε χρήστη του ΣΕΚ. Η λεπτομερής και πολλαπλή ταξινόμηση επιτρέπει ακριβώς την ευελιξία στην επιλογή των δεδομένων, διασφαλίζοντας ταυτόχρονα τη λεπτομερή ταυτοποίηση κάθε κειμένου που περιλαμβάνεται στο ΗΣΚ.

Η υλοποίηση του προγράμματος προβλέπει τέσσερις φάσεις. Στο πρώτο έτος (2003), προβλέπεται η συλλογή προφορικών και γραπτών δεδομένων, η μεταγραφή μέρους των προφορικών δεδομένων, η πραγματοποίηση της ιστοσελίδας και ο προκαταρκτικός σχεδιασμός εφαρμογών. Για το δεύτερο έτος (2004) έχουν προγραμματιστεί η ολοκλήρωση της συλλογής δεδομένων, ο σχεδιασμός της περαιτέρω ανάπτυξης, η συντήρηση και επέκταση της ιστοσελίδας και η δοκιμαστική ανάπτυξη των εφαρμογών. Η τρίτη φάση θα περιλαμβάνει την ανάπτυξη και επέκταση του ΣΕΚ μέσω εφαρμογών που θα δοκιμαστούν σε ένα περιορισμένο κύκλο ενδιαφερόμενων ερευνητών. Στην τέταρτη φάση, το ΣΕΚ θα είναι διαθέσιμο στο ευρύτερο κοινό μέσω κατάλληλων εφαρμογών. Προς το παρόν, έχει υλοποιηθεί στο μεγαλύτερο βαθμό ο σχεδιασμός του πρώτου έτους και έχουν καλυφθεί οι στόχοι που τέθηκαν. Τα δεδομένα που έχουν συλλεγεί ανέρχονται σε 12 εκατομμύρια λέξεις περίπου. Επίσης, η ιστοσελίδα του προγράμματος βρίσκεται στο διαδίκτυο και παρέχει βασικές πληροφορίες στη διεύθυνση: www.ucy.ac.cy/ sek.

6. Ανάπτυξη και εφαρμογές

Είναι σαφές ότι το ΣΕΚ θα αποτελέσει ένα βασικό εργαλείο έρευνας με πολλαπλές εφαρμογές, κατά τα πρότυπα των ΗΣΚ σε διεθνές επίπεδο. Σύμφωνα με το ερευνητικό πρόγραμμα, τέσσερις κυρίως περιοχές αναμένονται να καλυφθούν:

α) Γλωσσολογική έρευνα: το ΣΕΚ θα προσφέρει πολύτιμα δεδομένα στην έρευνα με τη χρήση ηλεκτρονικών υπολογιστών για το λεξιλόγιο και τη σύνταξη της Ελληνικής, την περιγραφή του προφορικού και γραπτού λόγου, καθώς και υφολογικών φαινομένων, τη μελέτη της προφορικής γλώσσας και ιδιωματικών παραλλαγών, την κοινωνιο-γλωσσική και διαλεκτολογική έρευνα (πρβλ. Chafe, Du Bois και Thompson 1991: 64-66). Παραδείγματα πιθανών εφαρμογών, που έχουν χρησιμοποιήσει μέρος του ΣΕΚ ή κάποια προκαταρκτική μορφή του, δίνονται αναλυτικά στα Goutsos, Hatzidaki και King (1994), Γούτσος, King και Χατζηδάκη (1995), Georgakopoulou και Goutsos (1998) και Goutsos (1999).

β) Εκπαιδευτικές εφαρμογές: το ΣΕΚ θα δώσει τις δυνατότητες για την ανάπτυξη υπολογιστικών εργαλείων για το σχεδιασμό υλικού ή για χρήση στην αίθουσα διδασκαλίας (βλ. Wichmann, Fligelstone, McEnery και Knowles 1997, Χατζηδάκη, υπό δημοσίευση). Σε αυτά θα περιλαμβάνεται αδιαμεσολάβητη πρόσβαση των μαθητών στη συλλογή κειμένων για

αυτοδιδασκαλία, εργαλεία διαδικτύου για εξ αποστάσεως εκπαίδευση, καθώς και ειδικά σχεδιασμένες ασκήσεις, στη μορφή εγχειριδίου ή CD με βάση αυθεντικό γλωσσικό υλικό.

γ) Διασύνδεση με άλλα προγράμματα: το ΣΕΚ έχει σχεδιαστεί να συμβάλει στην ανάπτυξη:

1) των μεταφραστικών σπουδών, σε σύνδεση με συλλογές κειμένων σε άλλες γλώσσες καθώς και με πολύγλωσσα και παράλληλα σώματα κειμένων,
2) της ελληνικής λεξικογραφίας, συνδεόμενο με προγράμματα όπως το Λεξικό Μπαμπινιώτη (1998),
3) των ιστορικών γλωσσικών σπουδών, σε σύνδεση με ΗΣΚ που περιλαμβάνουν άλλες ποικιλίες, συγχρονικά και διαχρονικά (π.χ. με το Perseus Project ή το Θησαυρό της Κυπριακής).

δ) Υπολογιστικές εφαρμογές: το ΣΕΚ μπορεί να χρησιμοποιηθεί προσφέροντας υλικό για την ανάπτυξη τεχνικών εργαλείων γλωσσικής ανάλυσης. Παρότι σε αυτό το στάδιο η αποθήκευση των κειμένων γίνεται σε μη χαρακτηρισμένη μορφή, η ανάπτυξη του προγράμματος στο μέλλον περιλαμβάνει γραμματικό χαρακτηρισμό (tagging) και σχολιασμό (annotation) του γλωσσικού υλικού που περιλαμβάνεται στο ΣΕΚ.

Με την ολοκλήρωση των διάφορων φάσεων του προγράμματος, αναμένουμε ότι θα έχουν υλοποιηθεί οι βασικοί στόχοι του σχεδιασμού με έμφαση στη δημόσια πρόσβαση στο ΣΕΚ και στις συγκεκριμένες εφαρμογές. Είμαστε βέβαιοι ότι η υλοποίηση του προγράμματος θα αλλάξει ριζικά την εικόνα που έχουμε σήμερα για την ελληνική στις διάφορες κειμενικές πραγματώσεις της, πλησιάζοντας σε μια πιο ακριβή, ολοκληρωμένη και αυθεντική περιγραφή της γλώσσας.

Παράρτημα

Αναλυτική ταξινόμηση κειμένων στο ΣΕΚ

ΤΡΟΠΟΣ	ΕΙΔΟΣ	ΥΠΟ-ΕΙΔΟΣ	ΜΕΣΟ	ΣΤΟΧΟΣ
Προφορικά	Ειδήσεις	Επίκαιρα νέα	Ραδιόφωνο	50.000
	Ειδήσεις	Επίκαιρα νέα	Τηλεόραση	200.000
	Ειδήσεις	Ψυχαγωγικά νέα	Ραδιόφωνο	50.000
	Ειδήσεις	Ψυχαγωγικά νέα	Τηλεόραση	200.000
	Συνέντευξη	1 με 1	Ραδιόφωνο	50.000
	Συνέντευξη	1 με 1	Τηλεόραση	100.000
	Συνέντευξη	1 με 1	Ζωντανό	50.000
	Συνέντευξη	1 με 2-3	Τηλεόραση	50.000
	Συνέντευξη	1 με πολλούς	Τηλεόραση	150.000
	Συνέντευξη	1 με πολλούς	Ζωντανό	100.000
	Ομιλίες	Ακαδημαϊκές	Ζωντανό	250.000
	Ομιλίες	Μη ακαδημαϊκές	Ζωντανό	1.250.000
	Συνομιλία	1 με 1	Ζωντανό	200.000
	Συνομιλία	1 με 2-3	Ζωντανό	200.000
	Συνομιλία	1 με 1	Τηλέφωνο	90.000
	Συνομιλία	Άλλο	Τηλέφωνο	10.000
Γραπτά	Λογοτεχνία	Μυθιστόρημα	Βιβλίο	2.052.000
	Λογοτεχνία	Διήγημα	Βιβλίο	1.539.000
	Λογοτεχνία	Αυτοβιογραφία	Βιβλίο	564.300

	Λογοτεχνία	Ποίηση	Βιβλίο	410.400
	Λογοτεχνία	Θεατρικά έργα	Βιβλίο	359.100
	Λογοτεχνία	Παραμύθι	Άλλο	102.600
	Λογοτεχνία	Στίχοι τραγουδιών	Άλλο	51.300
	Λογοτεχνία	Ανέκδοτα	Άλλο	51.300
	Διάφορα	Άλλο	Άλλο	270.000
	Ειδήσεις	Κοινωνικά/ Πολιτικά	Εφημερίδα	1.166.400
	Ειδήσεις	Οικονομικά	Εφημερίδα	1.166.400
	Ειδήσεις	Ελεύθ. Χρόνου	Εφημερίδα	583.200
	Άρθρα Γνώμης	Κοινωνικά/ Πολιτικά	Εφημερίδα	1.270.080
	Άρθρα Γνώμης	Οικονομικά	Εφημερίδα	1.270.080
	Άρθρα Γνώμης	Ελεύθ. Χρόνου	Εφημερίδα	635.400
	Πληροφορίες	Διάφορα	Εφημερίδα	64.800
	Ακαδημαϊκά	Ανθρωπιστικά	Βιβλίο	1.382.400
	Ακαδημαϊκά	Κοινωνικά/ Οικονομικά	Βιβλίο	1.382.400
	Ακαδημαϊκά	Θετικές Επιστήμες	Βιβλίο	691.200
	Ακαδημαϊκά	Ανθρωπιστικά	Περιοδικό	259.200
	Ακαδημαϊκά	Κοινωνικά/ Οικονομικά	Περιοδικό	259.200
	Ακαδημαϊκά	Θετικές Επιστήμες	Περιοδικό	129.600
	Ακαδημαϊκά	Ανθρωπιστικά	Ηλεκτρονικό	86.400
	Ακαδημαϊκά	Κοινωνικά/ Οικονομικά	Ηλεκτρονικό	86.400
	Ακαδημαϊκά	Θετικές Επιστήμες	Ηλεκτρονικό	43.200
	Ενημερωτικά	Ανθρωπιστικά	Βιβλίο	777.600
	Ενημερωτικά	Κοινωνικά/ Οικονομικά	Βιβλίο	648.000
	Ενημερωτικά	Θετικές-Τεχνολογία	Βιβλίο	388.800
	Ενημερωτικά	Σπίτι-Χόμπυ-Διατροφή	Βιβλίο	259.200
	Ενημερωτικά	Αθλητισμός	Βιβλίο	129.600
	Ενημερωτικά	Πολιτισμός	Βιβλίο	129.600
	Ενημερωτικά	Άλλα	Βιβλίο	259.200
	Ενημερωτικά	Ανθρωπιστικά	Ηλεκτρονικό	259.200
	Ενημερωτικά	Κοινωνικά/ Οικονομικά	Ηλεκτρονικό	216.000
	Ενημερωτικά	Θετικές-Τεχνολογία	Ηλεκτρονικό	129.600
	Ενημερωτικά	Σπίτι-Χόμπυ-Διατροφή	Ηλεκτρονικό	86.400
	Ενημερωτικά	Αθλητισμός	Ηλεκτρονικό	43.200
	Ενημερωτικά	Πολιτισμός	Ηλεκτρονικό	43.200
	Ενημερωτικά	Άλλα	Ηλεκτρονικό	86.400
	Ενημερωτικά	Κοινωνικά/ Πολιτικά	Περιοδικό	777.600
	Ενημερωτικά	Οικονομικά	Περιοδικό	777.600
	Ενημερωτικά	Θετικές-Τεχνολογία	Περιοδικό	518.400
	Ενημερωτικά	Σπίτι-Χόμπυ-Διατροφή	Περιοδικό	518.400
	Ενημερωτικά	Αθλητισμός	Περιοδικό	518.400
	Ενημερωτικά	Πολιτισμός	Περιοδικό	518.400
	Ενημερωτικά	Γυναικεία	Περιοδικό	518.400
	Ενημερωτικά	Νεανικά	Περιοδικό	518.400
	Ενημερωτικά	Οικολογικά	Περιοδικό	259.200
	Ενημερωτικά	Άλλο	Περιοδικό	259.200
	Νόμοι-Διοίκηση	Νομοθεσία		756.000
	Νόμοι-Διοίκηση	Διοίκηση		756.000
	Ιδιωτικά	Επιστολές	Άλλο	32.400
	Ιδιωτικά	Ηλεκτρ. Ταχυδρομείο	Ηλεκτρονικό	81.000
	Ιδιωτικά	Εφήμερα	Άλλο	32.400

	Ιδιωτικά	Ηλεκτρ. Συνομιλία	Ηλεκτρονικό	81.000
	Ιδιωτικά	Ημερολόγιο	Άλλο	32.400
	Ιδιωτικά	Άλλο	Άλλο	64.800
	Διαδικαστικά			108.000
	Διάφορα			216.000

Σημειώσεις

[1] Οι συνεργάτες του Προγράμματος «Βασικό Σώμα Ελληνικών Κειμένων» είναι:

Επιστημονικοί υπεύθυνοι:
- Διονύσης Γούτσος (Πανεπιστήμιο Αθηνών)
- Παύλος Παύλου (Πανεπιστήμιο Κύπρου)

Συμμετέχοντες:
- Γεώργιος Μπαμπινιώτης (Πανεπιστήμιο Αθηνών, επόπτης)
- Αλεξάνδρα Γεωργακοπούλου (Department of Byzantine and Modern Greek Studies, King's College London)
- Philip King (School of English και EISU, Πανεπιστήμιο Birmingham)
- Σταματία Κουτσουλέλου-Μίχου
- Γιώργος Μικρός
- Καίτη Μπακάκου-Ορφανού
- Ελένη Παναρέτου-Τσίρη (Πανεπιστήμιο Αθηνών)

Εμπορικοί συνεργάτες:
- Εκδοτικός οίκος «Ελληνικά Γράμματα», Αθήνα

Μεταπτυχιακοί συνεργάτες:
- Σωτηρούλα Γιασεμή,
- Ρένα Σοφοκλέους

(Πανεπιστήμιο Κύπρου)

Επιστημονικοί συνεργάτες:
- Μαριάνθη Πατρώνα,
- Ελένη Γαρυφαλλάκη, Ματίνα Σπηλιοπούλου, Μαριάννα Χρήστου

Η ανακοίνωση αυτή γίνεται από τον επιστημονικό υπεύθυνο σχεδιασμού και υλοποίησης, ενώ όλοι οι συνεργάτες θα παρουσιάσουν άλλες πτυχές του προγράμματος.

[2] Η συζήτηση που ακολουθεί οφείλει πολλά στα προσωπικά σχόλια της Ράνιας Χατζηδάκη.

[3] Ειδικότερα, η συμβολή των προφορικών κειμενικών ειδών στο σύνολο του ΗΣΚ (10 %), παρότι φαινομενικά μικρή, αποτελεί τη διεθνή σταθερά, με βάση τις δυνατότητες και τις χρονικές απαιτήσεις για καταγραφή, απομαγνητοφώνηση και μεταγραφή προφορικών δεδομένων (Goutsos, King & Hatzidaki 1994).

[4] Η πτυχή αυτή του σχεδιασμού θα καλυφθεί σε ξεχωριστή ανακοίνωση από τον Γιώργο Μικρό.

[5] Πιο εξειδικευμένα στοιχεία επιτονισμού και φωνητικής δεν καταγράφονται στο παρόν στάδιο, αλλά υπάρχει η δυνατότητα για λεπτομερέστερη καταγραφή στο μέλλον, εφόσον φυλάσσεται ακουστικό αρχείο των δεδομένων.

Βιβλιογραφία

Aarts, Jan. 1991. "Intuition-based and observation-based grammars". *English Corpus Linguistics*, επιμ. Karen Aijmer και Bengt Altenberg, 44-62. London: Longman.

Barnbrook, Geoff. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Biber, Douglas, Conrad, Susan και Reppen, Randi. 1998. *Corpus Linguistics. Investigating Language, Structure and Use*. Cambridge: Cambridge University Press.

Chafe, Wallace L., Du Bois, John W. και Thompson, Sandra A. 1991. "Towards a new corpus of spoken American English". *English Corpus Linguistics*, επιμ. Karen Aijmer και Bengt Altenberg, 64-82. London: Longman.

Georgakopoulou, Alexandra και Goutsos, Dionysis. 1998. "Conjunctions versus discourse markers in Greek: The interaction of frequency, positions and functions in context". *Linguistics* 36 (5). 887-917.

Γεωργακοπούλου, Αλεξάνδρα και Γούτσος Διονύσης. 1999. *Κείμενο και επικοινωνία*. Αθήνα: Ελληνικά Γράμματα.

Goutsos, Dionysis. 1999. "Translation in bilingual lexicography. Editing a new English-Greek Dictionary". *Babel* 45 (2). 107-126.

Goutsos, Dionysis, Hatzidaki, Rania και King, Philip 1994. "A corpus-based approach to Modern Greek language research and teaching". *Themes in Greek Linguistics: Papers from the First International Conference on Greek Linguistics. Reading, September 1993*. επιμ. Irene Philippaki-Warburton, Katerina Nicolaidis και Maria Sifianou, 507-513. Amsterdam/Philadelphia: John Benjamins.

Goutsos, Dionysis, King, Philip και Hatzidaki, Rania. 1994. "Towards a Corpus of Spoken Modern Greek". *Literary and Linguistic Computing* 9 (3). 215-223.

Γούτσος, Διονύσης, King, Philip και Χατζηδάκη, Ράνια. 1995. "Η χρήση των corpus στη λεξικογραφία και περιγραφή της Νέας Ελληνικής". *Μελέτες για την Ελληνική Γλώσσα. Πρακτικά της 15ης ετήσιας συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης, 11-14 Μαΐου 1994*, 843-854. Θεσσαλονίκη: Αφοί Κυριακίδη.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kučera, Karel. 2002. "The Czech National Corpus: Principles, design and results". *Literary and Linguistic Computing* 17 (2). 245-257.

Leech, Geoffrey 1992. "Corpora and theories of linguistic performance". *Directions in Corpus Linguistics*, επιμ. Jan Svartvik, 105-122. Berlin/New York: Mouton de Gruyter.

Μπαμπινιώτης, Γιώργος. 1998. *Λεξικό της Νέας Ελληνικής Γλώσσας*. Αθήνα: Κέντρο Λεξικογραφίας.

Μπαμπινιώτης, Γιώργος. 1999. *Η Γλώσσα ως Αξία. Το Παράδειγμα της Ελληνικής*. Αθήνα: Gutenberg.

Renouf, Antoinette. 1987. "Corpus development". *Looking Up*. επιμ. John Sinclair, 1-40. London and Glasgow: Collins ELT.

Sinclair, John. (επιμ.) 1987. *Looking Up*. London and Glasgow: Collins ELT.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. "Preliminary recommendations on corpus typology". Έγγραφο EAGLES (στο http: / / www.ilc.pi.cnr.it/ EAGLES/ corpustyp/ corpustyp.html).

Wichmann, Anne, Fligelstone, Steven, McEnery, Tony και Knowles, Gerry (επιμ.). 1997. *Teaching and Language Corpora*. London: Longman.

Χατζηδάκη, Ουρανία. (υπό δημοσίευση). "Τα σώματα κειμένων και το διαδίκτυο ως πηγές για την προώθηση της Νέας Ελληνικής ως ξένης γλώσσας". *Πρακτικά του Ευρωπαϊκού Συμποσίου Γλωσσικός Πλουραλισμός και Πολιτική Ξένων Γλωσσών στην ΕΕ (19-22 Σεπτεμβρίου 2001).*