Σώματα κειμένων: Αρχές και μέθοδοι

Η ανάγκη όσων μελετούν μια γλώσσα να αντλούν πληροφορίες για αυτή από εμπειρικά δεδομένα είναι σύμφυτη με τον επιστημονικό χαρακτήρα της γλωσσολογίας. Όπως κάθε επιστήμη, έτσι και η γλωσσολογία βασίζεται για τις θεωρίες και τις υποθέσεις της σε άμεσα προσβάσιμα γλωσσικά δεδομένα, τα οποία προσφέρουν σε όλους/ες τη δυνατότητα ελέγχου (επιβεβαίωσης ή διάψευσης). Ωστόσο, στη διάρκεια του 20ού αιώνα επικράτησαν γλωσσολογικές αντιλήψεις, σύμφωνα με τις οποίες η διαίσθηση των φυσικών ομιλητών/τριών μιας γλώσσας ήταν επαρκής για να στηριχθεί μια γλωσσολογική υπόθεση. Οι αντιλήψεις αυτές άρχισαν να υποχωρούν όταν με την ευρεία χρήση των προσωπικών ηλεκτρονικών υπολογιστών δόθηκε η δυνατότητα για αποθήκευση, ταχύτατη ανάκληση και επεξεργασία τεράστιου όγκου γλωσσικών πληροφοριών. Έτσι, μια από τις πιο σύγχρονες γλωσσολογικές μεθόδους βασίζεται σε συλλογές μεγάλου αριθμού κειμένων, αποθηκευμένων σε ηλεκτρονική μορφή και επεξεργάσιμων με υπολογιστικά εργαλεία, τα οποία ονομάζονται *σώματα κειμένων*.

Τα σώματα κειμένων επιτρέπουν να εξαγάγουμε πληροφορίες για τη γλώσσα από δεδομένα που είναι:

- εμπειρικά: δεν βασίζονται σε εικασίες ή τη διαίσθηση του/της ερευνητή/τριας, αλλά προέρχονται από την «εξωτερική» πραγματικότητα
- αυθεντικά: δεν προέρχονται από πειραματικές συνθήκες ή άλλες τεχνητές συνθήκες, αλλά από την αυθόρμητη (προφορική και γραπτή) παραγωγή λόγου των ομιλητών/τριών μιας γλώσσας
- συστηματικά: έχουν συλλεγεί με βάση συγκεκριμένα κριτήρια και αρχές και όχι με τυχαίο ή ανεκδοτολογικό τρόπο
- κειμενικά: περιλαμβάνουν ολόκληρα κείμενα ή τμήματα κειμένων και δεν περιορίζονται σε μεμονωμένες λέξεις ή προτάσεις
- εκτεταμένα: διαθέτουν μεγάλο όγκο και δεν περιορίζονται σε λίγα παραδείγματα. Όπως λέει ο Sinclair, ένας από τους πρωτοπόρους στον χώρο, «η γλώσσα φαίνεται πολύ διαφορετική όταν κοιτάς ένα μεγάλο κομμάτι της ταυτόχρονα» (1991: 100).

Συνοψίζοντας τα χαρακτηριστικά αυτά, ο Sinclair (1996) ορίζει ως σώμα κειμένων κάθε «συλλογή τμημάτων μιας γλώσσας τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα της γλώσσας αυτής».

Ειδικότερα, θεωρεί ότι το ηλεκτρονικό σώμα κειμένων είναι «κατάλληλο για ηλεκτρονική χρήση, ειδικά κωδικοποιημένο για τυποποιημένες και ομοιογενείς εργασίες ανάκτησης γλωσσικών πληροφοριών». Παρόμοιος είναι ο ορισμός του σώματος κειμένων από τον Sampson (2001: 6) ως «εκτεταμένου δείγματος αυθεντικής χρήσης της υπό εξέταση γλώσσας, που συγκροτείται και χρησιμοποιείται ως πηγή στοιχείων για την παραγωγή ή εξέταση υποθέσεων για τη φύση της γλώσσας».

Μια βασική έννοια που εισάγει ο Sinclair στον παραπάνω ορισμό είναι εκείνη της *αντιπροσωπευτικότητας*: κάθε σώμα κειμένων πρέπει να δίνει μια αντιπροσωπευτική εικόνα της γλώσσας (ή της ποικιλίας της γλώσσας) που συλλέγεται. Είναι δύσκολο να καθορίσουμε με ακρίβεια τι σημαίνει αντιπροσωπευτικότητα, δηλαδή να έχουμε εκ των προτέρων ένα μέτρο για το πόσα και ποια κείμενα είναι απαραίτητο να συλλέξουμε. Το βασικό στην έννοια αυτή είναι η έμφαση που δίνει στην όσο το δυνατόν μεγαλύτερη ποικιλία κειμένων και κειμενικών ειδών, στην ισορροπία των αναλογιών μεταξύ τους και στην αυθεντικότητα των δεδομένων που συλλέγονται. Επιπλέον, από την παραδοχή ότι ένα σώμα κειμένων είναι αντιπροσωπευτικό προκύπτει ότι τα συμπεράσματα που εξάγονται από αυτό ισχύουν, τηρουμένων των αναλογιών, για όλη τη γλώσσα ή τη γλωσσική ποικιλία που περιέχει το σώμα κειμένων. Ταυτόχρονα, ο Sampson στον δικό του ορισμό παραπάνω δίνει έμφαση στο γεγονός ότι το αποθηκευμένο ηλεκτρονικά σώμα κειμένων δεν παράγει «γεγονότα» για τη γλώσσα, αλλά προσφέρει τη βάση για τη δημιουργία υποθέσεων από τον/την ερευνητή/τρια, που μπορεί να οδηγήσουν στην εξαγωγή επαληθευμένων συμπερασμάτων. Επομένως, η παρέμβαση του/της γλωσσολόγου είναι απαραίτητη, τόσο στη διαμόρφωση των ερευνητικών ερωτημάτων όσο και στην αξιολόγηση των ευρημάτων που προκύπτουν.

Από τις βασικές αυτές αρχές συνάγεται ότι το είδος των δεδομένων που εξάγονται από την ανάλυση των ηλεκτρονικών σωμάτων κειμένων εξαρτάται άμεσα από τη σύστασή τους. Από το πρώτο ηλεκτρονικό σώμα κειμένων, το λεγόμενο *Brown Corpus,* έως σήμερα έχει αναπτυχθεί πλήθος πολλών και διαφορετικών σωμάτων κειμένων σε πολλές γλώσσες του κόσμου. Πιο γνωστά σώματα κειμένων για τα Αγγλικά είναι το *British National Corpus* (BNC) και το *Bank of English*, που ξεπερνούν τα 100 εκατομμύρια λέξεις, προσπαθώντας να δώσουν μια ευρεία εικόνα του συνόλου της αγγλικής γλώσσας. Αυτά τα σώματα κειμένων αποτέλεσαν τη βάση γλωσσικής έρευνας και εφαρμογών (π.χ. βιβλία αναφοράς όπως λεξικά, γραμματικές

κλπ.), ενώ σήμερα ανάλογα σώματα κειμένων ξεπερνούν το 1 δισεκατομμύριο λέξεις (*Cambridge International Corpus, Oxford English Corpus*). Αντίστοιχη ήταν η ανάπτυξη των σωμάτων κειμένων και στις άλλες μείζονες γλώσσες, αν και με σχετική καθυστέρηση σε σχέση με τα Αγγλικά. Στο Πλαίσιο 1 παρουσιάζεται μια επιλογή σωμάτων κειμένων με τις ιστοσελίδες τους στο διαδίκτυο, όπου μπορούν να αναζητηθούν περισσότερες πληροφορίες.

Πλαίσιο 1: Ηλεκτρονικά σώματα κειμένων σε ευρωπαϊκές γλώσσες

Αγγλικά:	*Ιστοσελίδα*
Bank of English Corpus	http://www.titania.bham.ac.uk/
British National Corpus (BNC)	http://www.natcorp.ox.ac.uk/
Cambridge International Corpus	http://www.cup.cam.ac.uk/gr/elt/catalogue/subject/item2701617/Cambridge-International-Corpus
International Corpus of English (ICE)	http://ice-corpora.net/ice/
Oxford English Corpus	http://www.oxforddictionaries.com/page/oec
Γαλλικά:	
Corpus de Référence du Français parlé	http://sites.univ-provence.fr/delic/corpus/index.html
Artfl – Frantext	http://artfl-project.uchicago.edu/content/artfl-frantext
Γερμανικά:	
Mannheimer Corpora	http://www.ids-mannheim.de/kl/corpora.html
Ιταλικά:	
Banca dati dell'italiano parlato	http://languageserver.uni-graz.at/badip/
Ipercorpus	http://culturitalia.uibk.ac.at/kic-index.htm
Ισπανικά:	
Mark Davies' Corpus Del Español	http://www.corpusdelespanol.org
Corpus Oral De Referencia Del Español	http://www.lllf.uam.es/corpus/corpus.html
Corpus Diacronico (Real Academia)	http://corpus.rae.es/cordenet.html

Στα Ελληνικά, τα πρώτα ηλεκτρονικά σώματα κειμένων εμφανίζονται στη δεκαετία του '80 και περιλαμβάνουν παλαιότερα λογοτεχνικά έργα (π.χ. κρητική λογοτεχνία, απομνημονεύματα του Μακρυγιάννη), ενώ το 1994 οι Goutsos κ.ά. διαπιστώνουν ότι, αν και υπάρχουν αρκετά σχετικά ερευνητικά προγράμματα, τα σώματα κειμένων είτε δεν χρησιμοποιούνται καθόλου στη γλωσσική έρευνα είτε δεν έχουν αξιοποιηθεί πλήρως. Στη δεκαετία του 1990 εμφανίζονται τα δύο μεγαλύτερα ηλεκτρονικά σώματα κειμένων της Ελληνικής, ο Εθνικός Θησαυρός της Ελληνικής Γλώσσας (ΕΘΕΓ), που περιλαμβάνει 40 εκατομμύρια λέξεις, κυρίως από εφημερίδες, και το Σώμα Ελληνικών Κειμένων (ΣΕΚ), που περιέχει 30 εκατομμύρια λέξεις από μια πιο ισορροπημένη ποικιλία κειμενικών ειδών (βλ. Γούτσος 2003, Goutsos 2010). (Βλ. Πλαίσιο 2).

Πλαίσιο 2: Ηλεκτρονικά σώματα κειμένων της Ελληνικής

Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ)	40 εκατ. λέξεις, κυρίως από εφημερίδες	http://hnc.ilsp.gr/subcorpus.asp#
Σώμα Ελληνικών Κειμένων (ΣΕΚ)	30 εκατ. λέξεις, διάφορα κειμενικά είδη	http://www.sek.edu.gr
Πύλη για την Ελληνική Γλώσσα	σώματα κειμένων από εφημερίδες και σχολικά εγχειρίδια	http://www.greek-language.gr/greekLang/modern_greek/index.html
Πανεπιστήμιο Αθηνών	σώματα κειμένων για τη διδασκαλία και την εκμάθηση της Ελληνικής ως ξένης γλώσσας	http://greekcorpora.isll.uoa.gr/gr/Default.aspx

Εκτός από αυτά τα γενικά σώματα κειμένων, τα οποία περιλαμβάνουν κείμενα και κειμενικά είδη από το σύνολο της γλώσσας, επιδιώκοντας να αποτελέσουν *σώματα κειμένων αναφοράς* (reference corpora) για όλη τη γλώσσα, έχει αναπτυχθεί ένα πλήθος μικρότερων, εξειδικευμένων ηλεκτρονικών σωμάτων κειμένων για ειδικές γλωσσικές ποικιλίες ή ειδικούς σκοπούς. Έτσι, κάποια σώματα κειμένων συλλέγουν δεδομένα από ορισμένη ηλικιακή ομάδα (π.χ. το *CHILDES Corpus* από την παιδική γλώσσα ή το *COLT* από την προφορική γλώσσα των εφήβων), από ένα πεδίο του λόγου (λ.χ. το *Michigan Corpus of Academic Spoken English*), από τη διαχρονία της γλώσσας (π.χ. *Helsinki Corpus*) κ.ά. Τέλος, τα κείμενα που αποθηκεύονται στα σώματα κειμένων είναι συνήθως κωδικοποιημένα για ορισμένα χαρακτηριστικά ή μεταδεδομένα, όπως το έτος συγγραφής, στοιχεία των συγγραφέων ή των ομιλητών/τριών, κειμενικό είδος και υποείδος κλπ. Ορισμένα σώματα κειμένων είναι επιπλέον και *χαρακτηρισμένα* (annotated), περιέχουν δηλαδή πληροφορίες για τον χωρισμό παραγράφων, τα διάφορα κειμενικά μέρη (τίτλοι κλπ.), την αλλαγή ομιλητών κ.ά. ή και *επισημειωμένα* (tagged), δηλαδή περιέχουν πληροφορίες για τη γραμματική κατηγορία στην οποία ανήκουν όλες οι λέξεις των κειμένων.

Βασικά μεθοδολογικά εργαλεία για την ανάλυση και την επεξεργασία των δεδομένων που περιλαμβάνονται στα ηλεκτρονικά σώματα κειμένων είναι οι *κατάλογοι συχνότητας* λέξεων, οι *συμφραστικοί πίνακες* (concordances) και τα *συμφραστικά πλαίσια*. Το Πλαίσιο 3 παρουσιάζει μια επιλογή από υπολογιστικά προγράμματα που διατίθενται για τους σκοπούς της ανάλυσης.

Πλαίσιο 3: Εργαλεία ανάλυσης σωμάτων κειμένων

Monoconc	ειδικευμένο λογισμικό	http://www.monoconc.com/
Wordsmith Tools	ειδικευμένο λογισμικό	http://www.lexically.net/wordsmith/
Paraconc	πολύγλωσσα σώματα κειμένων	http://www.paraconc.com/
Antconc	ελεύθερο λογισμικό	http://www.antlab.sci.waseda.ac.jp/antconc_index.html
ConcApp	ελεύθερο λογισμικό	http://www.edict.com.hk/PUB/concapp/
Sketch Engine	ειδικευμένο για τη λεξικογραφία	http://www.sketchengine.co.uk/
WebCorp	για αναζήτηση στο διαδίκτυο	http://www.webcorp.org.uk/

Στον Πίνακα 1 παρουσιάζεται ένας κατάλογος με τις πιο συχνές λέξεις σε 28 εκατομμύρια λέξεις του ΣΕΚ:

Σειρά	**Λέξη**	**Συχνότητα**	**Ποσοστό %**
1	και	963.280	3,41
2	του	625.351	2,49
3	το	588.223	2,20
4	να	565.740	2,03
5	της	529.597	1,95
6	η	455.343	1,86
7	την	436.917	1,57
8	που	382.458	1,42
9	με	375.649	1,40
10	από	335.973	1,26
11	των	317.900	1,18
12	για	312.804	1,17
13	ο	293.353	1,10
14	τα	285.829	1,09
15	είναι	248.907	0,93
16	σε	235.302	0,87
17	οι	220.451	0,80
18	θα	207.441	0,78
19	τους	198.279	0,75
20	τη	196.015	0,72
21	στο	195.193	0,71
22	δεν	192.624	0,71
23	στην	190.424	0,71
24	τον	175.947	0,65
25	ότι	162.005	0,61

Πίνακας 1: Κατάλογος συχνότερων λέξεων στο ΣΕΚ (28 εκατ.)

Πληροφορίες από καταλόγους συχνότητας όπως αυτός του Πίνακα 1 μπορούν να αξιοποιηθούν με διάφορους τρόπους στη γλωσσική έρευνα, καθώς προσφέρουν μια

εκτίμηση των περισσότερο και λιγότερο σημαντικών στοιχείων για τη γλώσσα και της σχέσης μεταξύ τους.

Στον Πίνακα 2 περιλαμβάνεται ένας συμφραστικός πίνακας με είκοσι επιλεγμένες σειρές του τύπου *ξεχνάμε*. Όπως μπορούμε να δούμε, ο συμφραστικός πίνακας περιλαμβάνει την υπό εξέταση λέξη ή *κομβική λέξη* στο κέντρο, ενώ αριστερά και δεξιά δίνονται οι λέξεις με τις οποίες συνεμφανίζεται, τα στενά της δηλαδή συμφραζόμενα. Στον συγκεκριμένο πίνακα οι σειρές έχουν διαταχθεί αλφαβητικά με βάση την πρώτη λέξη στα αριστερά της κομβικής λέξης, ενώ περιλαμβάνεται στα δεξιά και μια στήλη με κωδικοποιημένο το κειμενικό είδος από το οποίο προέρχεται κάθε σειρά.

1	στο τραυματικό γεγονός. Γιατί	**ξεχνάμε**	τα όνειρά μας; Όπως κ	ΛΟΓΟΤ
2	υνδυασμός // Άλλωστε εμείς δεν	**ξεχνάμε**	ότι τα Γιάννινα ήταν	ΛΟΓΟΤ
3	ακρινή Κόρντοβα, αυτήν που δεν	**ξεχνάμε**	– εκείνη που αγαπήσαμ	ΕΝΗΜΕΡ
4	από 30 χρόνια στο Μόναχο. Δεν	**ξεχνάμε**	την τραγωδία και γι'α	ΕΝΗΜΕΡ
5	ο, τα λόγια και οι εκδηλώσεις;	**Ξεχνάμε**	ποια ήταν η κυβέρνηση	ΛΟΓΟΤ
6	φαινόμενο" Φαίνεται ότι εύκολα	**ξεχνάμε**	ότι ο ίδιος αέρας κυκ	ΑΚΑΔΗΜ
7	(σαρδάμ). Είμαστε αφηρημένες ή	**ξεχνάμε**	τις υποχρεώσεις μας.	ΕΝΗΜΕΡ
8	σία. "Κρατάμε την πρόκριση και	**ξεχνάμε**	την εμφάνιση", δήλωσε	ΑΡΘΡ.ΓΝ
9	σμού και ξεχνάμε – θέλουμε και	**ξεχνάμε**	– πως τη συγκεκριμένη	ΕΙΔΗΣ
10	η πόλη μας, η Αθήνα. Γιατί μη	**ξεχνάμε**	ποτέ ότι η Αθήνα είνα	ΠΡΟΦΟΡ
12	σία, συντροφιά, όλα. Και ας μη	**ξεχνάμε**	πάντα ότι η ανάγνωση	ΝΟΜΙΚ
13	ι της μειονότητας. Αλλά, ας μη	**ξεχνάμε**	ότι η Συνθήκη της Λωζ	ΑΡΘΡ.ΓΝ
14	ν εξυπηρετείται η περιοχή. Μην	**ξεχνάμε**	ότι είναι μια περιοχή	ΛΟΓΟΤ
15	ννησης, θανάτου, κ.λ.π. Ας μην	**ξεχνάμε**	όμως, τουλάχιστον τα	ΑΚΑΔΗΜ
16	μοκρατικών ενεργειών. (Ας μην	**ξεχνάμε**	ότι στις αρχές του 20	ΕΝΗΜΕΡ
17	ο οποίο σημαίνει λόφος. Ας μην	**ξεχνάμε**	ότι και το χωριό πάνω	ΑΚΑΔΗΜ
18	ης Αμερικής. Έχουμε αρχίσει να	**ξεχνάμε**	τον παραδοσιακό τρόπο	ΕΝΗΜΕΡ
19	υμπιάδας. Δεν πρέπει, όμως, να	**ξεχνάμε**	και τη στενή διαπλοκή	ΑΡΘΡ.ΓΝ
20	ομίζουμε. Δεν πρέπει πάντως να	**ξεχνάμε**	ότι η εικόνα του σκλη	ΑΚΑΔΗΜ

Πίνακας 2: Συμφραστικός πίνακας του *ξεχνάμε* στο ΣΕΚ (επιλογή)

Οι συμφραστικοί πίνακες προσφέρουν τις περισσότερες γλωσσικές πληροφορίες για μια λέξη (ή φράση, εφόσον αναζητήσουμε μια ολόκληρη φράση), καθώς επιτρέπουν να εντοπίσουμε τις λέξεις με τις οποίες συνεμφανίζεται, δηλαδή τις *συνάψεις* της.[1] Επιτρέπουν επίσης να διαπιστώσουμε τις άμεσες συνταγματικές σχέσεις της λέξης (ή φράσης), δηλαδή τα ευρύτερα δομικά σύνολα στα οποία χρησιμοποιείται. Επιπλέον, μπορούμε να διακρίνουμε τα περισσότερο σημαντικά δομικά σχήματα στη χρήση

[1] Για μεγαλύτερη ανάλυση και θεωρητική τοποθέτηση των σχετικών όρων, βλ. Γούτσος (2006).

μιας λέξης (ή φράσης) από χρήσεις που εμφανίζονται ελάχιστες φορές και γι' αυτόν τον λόγο κρίνονται περιθωριακές.

Γούτσος, Δ. (2006). Ανάπτυξη λεξιλογίου-Από το βασικό στο προχωρημένο επίπεδο. Στο Γούτσος, Δ., Σηφιανού, Μ. & Α. Γεωργακοπούλου *Η ελληνική ως ξένη γλώσσα: Από τις λέξεις στα κείμενα*. Αθήνα: Πατάκης, 13-92.