

8.2. ΛΟΓΙΣΜΙΚΑ ΑΝΑΛΥΣΗΣ ΣΩΜΑΤΩΝ ΚΕΙΜΕΝΩΝ

Ο Πίνακας 8.2 που ακολουθεί περιλαμβάνει στοιχεία για τα κυριότερα λογισμικά που επιτρέπουν την ανάλυση σωμάτων κειμένων στον προσωπικό υπολογιστή και έχουν ελεγχθεί ότι λειτουργούν για ελληνικά δεδομένα με σχετικά απλό τρόπο.¹

Πίνακας 8.1: Λογισμικά ανάλυσης σωμάτων κειμένων

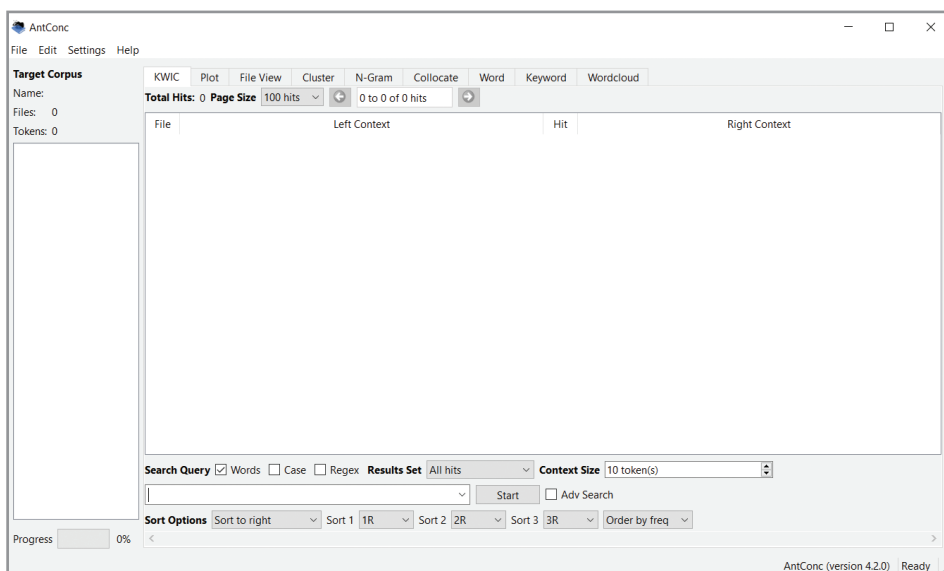
Λογισμικό	Τελευταία έκδοση	Ιστοσελίδα
AntConc	4.2.0 (Μάρτιος 2023)	https://www.laurenceanthony.net/software/antconc
CorpKit		https://interrogator.github.io/corpkite
LancsBox	5	https://lancsbox.lancs.ac.uk
Simple Concordance Program	5.0.4 (Μάιος 2023)	http://www.textworld.com/scp
Simple Corpus Tool	3	https://martinweisser.org/ling_soft.html#SCT
TXM	0.8.2 (Ιούνιος 2022)	https://txm.gitpages.huma-num.fr/textometrie/files/software/TXM/0.8.2
Voyant Tools	2.6.6 (2023)	https://voyant-tools.org
WordCruncher	7.1.107 (2023)	https://wordcruncher.com
WordStatix	2.0.0 (2020)	https://sites.google.com/site/wordstatix

Το δημοφιλέστερο και πιο εύχρηστο πρόγραμμα ανάλυσης σωμάτων κειμένων είναι το AntConc, η πρώτη έκδοση του οποίου αναπτύχθηκε από τον Laurence Anthony το 2002.² Από τα υπόλοιπα προγράμματα ιδιαίτερη μνεία πρέπει να γίνει για το Voyant Tools, που ανέπτυξαν το 2003 οι Stéfan Sinclair και Geoffrey Rockwell και είναι το μόνο από τη λίστα του Πίνακα 2 που λειτουργεί στο διαδίκτυο αλλά και με εγκατάσταση σε προσωπικό υπολογιστή. Το λογισμικό διαθέτει σαφώς διατυπωμένες αρχές σχεδιασμού (βλ. στην ιστοσελίδα: <https://voyant-tools.org/docs/#!/guide/about>), οι οποίες είναι κεντρικής σημασίας για κάθε έργο στις ψηφιακές ανθρωπιστικές επιστήμες, ενώ συνδέεται με ένα σημαντικό βιβλίο (Rockwell & Sinclair 2016), που αναπτύσσει τον προβληματισμό της κειμενικής ανάλυσης μέσω υπολογιστών ως μέρος μιας ευρύτερης ερμηνευτικής παράδοσης στις ανθρωπιστικές επιστήμες.

Για να γίνουν κατανοητά τα βασικά μεθοδολογικά εργαλεία της ανάλυσης σω-

1. Περισσότερα λογισμικά κειμενικής ανάλυσης σε σώματα κειμένων μπορεί κανείς να βρει στον κατάλογο της ιστοσελίδας: <https://www.clarin.eu/resource-families/corpus-query-tools>

2. Βασικές οδηγίες για το πρόγραμμα AntConc στα αγγλικά βρίσκονται στην ιστοσελίδα: <https://www.laurenceanthony.net/software/antconc/releases/AntConc420/help.pdf>



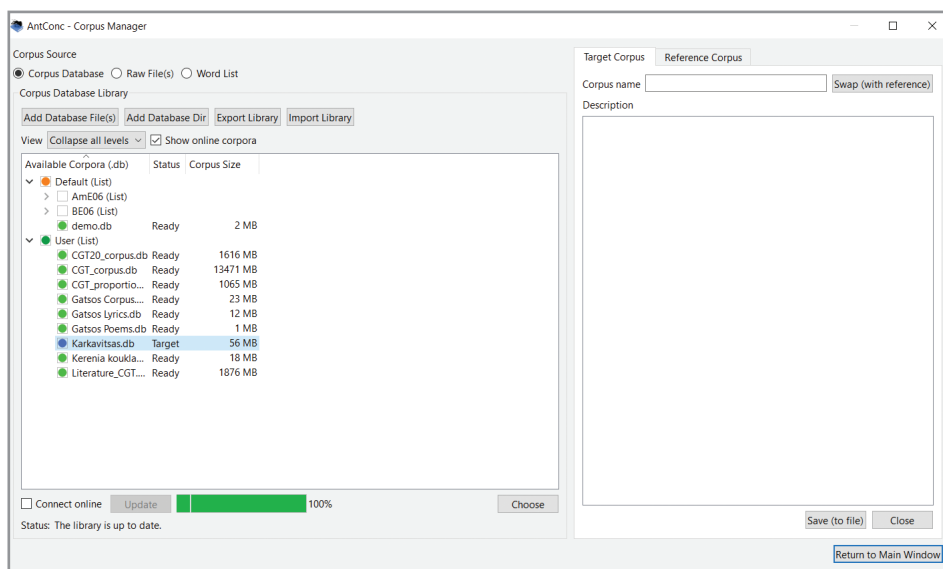
Εικόνα 8.5: Οθόνη έναρξης του προγράμματος AntConc

μάτων κειμένων θα παρουσιάσουμε ενδεικτικά τα εργαλεία που προσφέρει το πρόγραμμα AntConc στην τελευταία του έκδοση. (Για μια θεωρητική παρουσίαση των μεθοδολογικών εργαλείων βλ. Γούτσος & Φραγκάκη 2015: 61 κ.εξ.). Με την έναρξη του προγράμματος εμφανίζεται η βασική οθόνη που φαίνεται στην Εικόνα 8.5.

Στην οθόνη αυτή παρουσιάζονται οι κύριες επιλογές των μεθοδολογικών εργαλείων: συμφραστικοί πίνακες στην καρτέλα KWIC, διασπορά λεξιλογίου στην καρτέλα Plot, εμφάνιση του κειμένου στην καρτέλα File View, λεξικά συμπλέγματα στην καρτέλα Cluster, ν-γράμματα στην καρτέλα N-Gram, συνάψεις στην καρτέλα Collocate, κατάλογος συχνότητας στην καρτέλα Word, λέξεις-κλειδιά στην καρτέλα Keyword και συννεφώλεξο στην καρτέλα Wordcloud.

Για να λειτουργήσει το πρόγραμμα θα πρέπει να φορτωθούν τα αρχεία του σώματος κειμένων το οποίο ενδιαφερόμαστε να αναλύσουμε μέσω της επιλογής File, που προσφέρει δύο επιλογές, είτε να φορτώσουμε προσωρινά ένα ή περισσότερα αρχεία (Open File(s) as 'Quick Corpus'), είτε να επιλέξουμε ένα σώμα κειμένων που έχουμε ήδη δημιουργήσει (Open Corpus Manager), όπως φαίνεται στην Εικόνα 8.6.

Στην Εικόνα 8.6 φαίνονται σώματα κειμένων στα αγγλικά που περιλαμβάνονται στο λογισμικό (AmE06 και BE06) και άλλα σώματα κειμένων που αποθηκεύσαμε εδώ για τους σκοπούς της παρουσίασης. Για να δημιουργήσει κανείς ένα σώμα κειμένων, που θα είναι αποθηκευμένο στο πρόγραμμα μπορεί να χρησιμοποιήσει την επιλογή Raw File(s) (στην καρτέλα Target Corpus), που τον καθοδηγεί να δώσει όνομα στο σώμα κειμένων και να προσθέσει αρχεία ή και ολόκληρο φάκελο αρχείων που θα συμπεριληφθούν στο σώμα κειμένων. Με τη επιλογή Choose επιλέγεται το αποθη-



Εικόνα 8.6: Επιλογή σώματος κειμένων στο Corpus Manager του AntConc

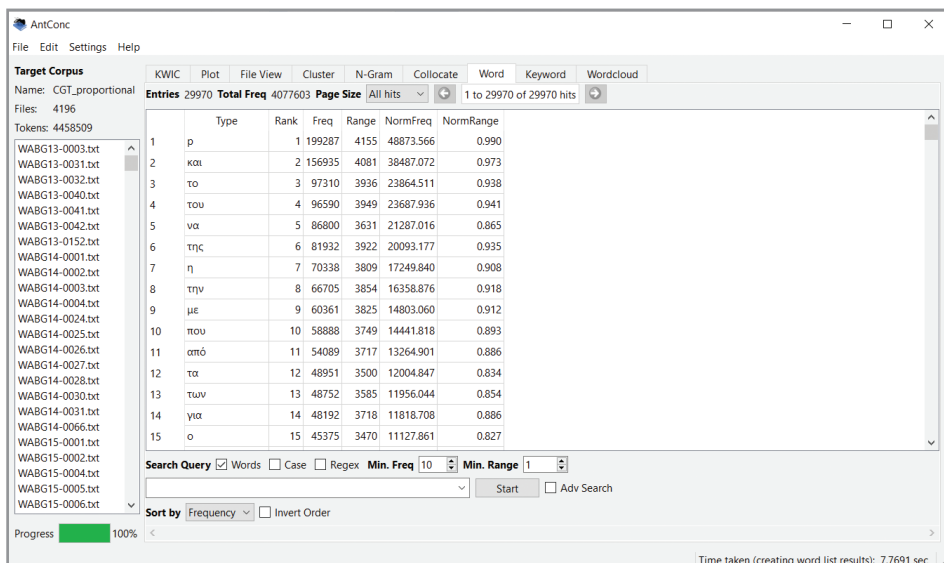
κειμένο σώμα κειμένων που επιθυμούμε να χρησιμοποιήσουμε και μπορούμε έτσι να επιστρέψουμε στην κύρια οθόνη (Εικόνα 8.5).

8.3. ΚΑΤΑΛΟΓΟΙ ΣΥΧΝΟΤΗΤΑΣ

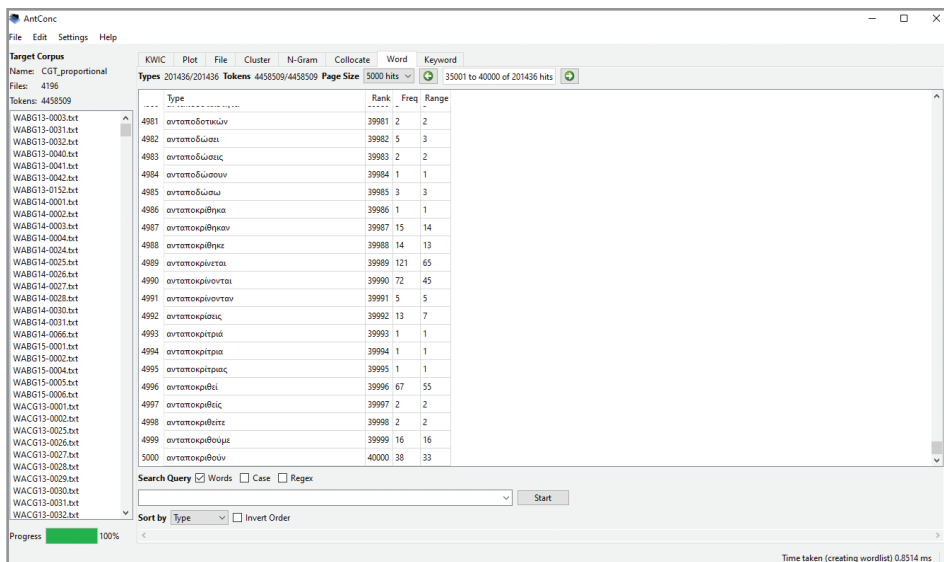
Αφετηρία κάθε κειμενικής ανάλυσης αποτελεί η αναζήτηση της ποικιλίας και της κατανομής του λεξιλογίου ενός σώματος κειμένων μέσω ενός καταλόγου συχνότητας (βλ. Γούτσος & Φραγκάκη 2015: 61-67). Στο AntConc οι κατάλογοι συχνότητας δημιουργούνται από την επιλογή Word στην αρχική οθόνη και έχουν τη μορφή που βλέπουμε στην Εικόνα 8.7.

Ο κατάλογος συχνότητας στην Εικόνα 8.7 προέρχεται από την αναζήτηση σε ένα αναλογικό υποσώμα του ΣΕΚ 4.196 αρχείων με 4.458.509 δείγματα λέξεων, όπως μπορούμε να δούμε στα στοιχεία επάνω αριστερά. Στο σύνολο αυτών των δειγμάτων λέξεων ο κατάλογος συχνότητας περιλαμβάνει 29.970 μοναδικές λέξεις που επαναλαμβάνονται από 1 έως 156.935 φορές στην περίπτωση του *και*. (Στον συγκεκριμένο κατάλογο πρώτο στοιχείο εμφανίζεται το *p*, που αντιστοιχεί στην επισημείωση για την αλλαγή παραγράφου). Όλες οι λέξεις του καταλόγου εμφανίζονται αν επιλέξουμε All hits στο μενού που αντιστοιχεί στο Page Size.³

3. Αυτό ισχύει για όλα τα εργαλεία του AntConc και είναι σημαντικό όχι μόνο για να μπορούμε να



Εικόνα 8.7: Κατάλογος συχνότητας στο AntConc με σειρά συχνότητας



Εικόνα 8.8: Κατάλογος συχνότητας στο AntConc με αλφαβητική σειρά

δούμε όλα τα αποτελέσματα με κύλιση προς τα κάτω αλλά και γιατί στην εξαγωγή των δεδομένων (από το File > Save Current Tab Results) πρέπει να έχει επιλεγεί All Hits για να αποθηκευτούν όλα τα στοιχεία.

Έχουμε τη δυνατότητα να ταξινομήσουμε τα αποτελέσματα του καταλόγου είτε με σειρά συχνότητας, επιλέγοντας Frequency στο μενού Sort by κάτω από τα αποτελέσματα, είτε με αλφαβητική σειρά, επιλέγοντας Type, όπως στην Εικόνα 8.8, ή με αντίστροφη αλφαβητική σειρά χαρακτήρων, δηλαδή με βάση την κατάληξη της λέξης και όχι την αρχή της, όπως βλέπουμε στον Πίνακα 3, που προέρχεται από τον ίδιο κατάλογο συχνότητας, επιλέγοντας TypeEnd είτε, τέλος, με βάση τη διασπορά κάθε στοιχείου στα αρχεία, δηλαδή σε πόσα διαφορετικά κείμενα του σώματος κειμένων απαντά η λέξη, επιλέγοντας Range. Στον Πίνακα 8.2 εμφανίζεται ο τύπος της λέξης, η σειρά της στον κατάλογο (Rank), η συχνότητά της (Freq), η κάλυψη, δηλαδή σε πόσα κείμενα απαντά (Range) και η στατιστικά κανονικοποιημένη συχνότητα (NormFreq) και κατανομή σε κείμενα (NormRange) σύμφωνα με τις επιλογές στο Global Settings (Statistics).

Πίνακας 8.2: Απόσπασμα καταλόγου συχνότητας στο αναλογικό ΣΕΚ με αντίστροφη αλφαβητική σειρά

Type	Rank	Freq	Range	NormFreq	NormRange
σταδιακά	944	195	138	47.822	0.033
καρδιαγγειακά	945	14	10	3.433	0.002
ενεργειακά	946	25	12	6.131	0.003
επιφανειακά	947	18	17	4.414	0.004
οικογενειακά	948	20	18	4.905	0.004
περιφερειακά	949	44	29	10.791	0.007
ενδοθηλιακά	950	32	1	7.848	0.000
αντιπηκτικά	951	13	1	3.188	0.000
πανεπιστημιακά	952	73	19	17.903	0.005
γορτυνιακά	953	56	1	13.734	0.000
επικοινωνιακά	954	11	9	2.698	0.002
τηλεπικοινωνιακά	955	10	4	2.452	0.001
αναπτυξιακά	956	23	17	5.641	0.004
ολυμπιακά	957	22	8	5.395	0.002
εργαστηριακά	958	10	8	2.452	0.002
χρηματιστηριακά	959	15	15	3.679	0.004
ευκαιριακά	960	13	12	3.188	0.003
οριακά	961	18	18	4.414	0.004
δημητριακά	962	39	26	9.564	0.006
εργασιακά	963	25	23	6.131	0.005

Μια τυπική κατανομή του λεξιλογίου κάθε σώματος κειμένων, αλλά και κάθε μεμονωμένου κειμένου, είναι αυτή που φαίνεται στον Πίνακα 8.3, που παρουσιάζει διαφορετικές ζώνες συχνότητας από τον ίδιο κατάλογο.

Πίνακας 8.3: Ζώνες συχνότητας 1-10, 101-120 και 501-520 στο αναλογικό ΣΕΚ

Σειρά	Λέξη	Συχν.	Σειρά	Λέξη	Συχν.	Σειρά	Λέξη	Συχν.
1	και	156935	101	θέμα	2583	501	επιχειρήσεις	686
2	το	97310	102	σύμφωνα	2577	502	λειτουργίας	686
3	του	96590	103	ενός	2561	503	δημοκρατίας	685
4	να	86800	104	θέση	2556	504	προγράμματα	681
5	της	81932	105	εδώ	2500	505	αγώνα	680
6	η	70338	106	αφού	2499	506	κέντρο	680
7	την	66705	107	κάνει	2413	507	παπανδρέου	680
8	με	60361	108	είχαν	2342	508	περισσότερες	679
9	που	58888	109	πρώτη	2336	509	νέων	675
10	από	54089	110	γίνεται	2322	510	συμμετοχή	674
11	τα	48951	111	τρόπο	2278	511	ημέρες	673
12	των	48752	112	ούτε	2237	512	παραγωγή	672
13	για	48192	113	άλλη	2214	513	νομίζω	669
14	ο	45375	114	άλλο	2190	514	περισσότερα	668
15	είναι	39821	115	ίδια	2146	515	σχήμα	668
16	σε	38600	116	τέλος	2140	516	σύνολο	668
17	οι	35029	117	ώστε	2135	518	μεγαλύτερο	664
18	τη	31929	118	περιοχή	2122	518	δικαίωμα	655
19	στο	31669	119	μεγάλη	2120	519	αναφέρεται	654
20	θα	31562	120	λίγο	2099	520	ομάδας	645

Στις πρώτες θέσεις κάθε καταλόγου συχνότητας εμφανίζονται αποκλειστικά γραμματικοί τύποι (βλ. Γούτσος 2020: 167-169), λειτουργικές λέξεις χωρίς αναφορικό περιεχόμενο, όπως *και*, *το*, *του*, *να*, *της* κ.λπ. Σε πιο χαμηλές θέσεις αρχίζουν να κάνουν την εμφάνισή τους και λεξικοί τύποι, όπως λ.χ. γενικά ουσιαστικά (*θέση*, *τρόπο*, *περιοχή*), ενώ πιο ειδικοί λεξικοί τύποι εμφανίζονται σε χαμηλότερες θέσεις στον κατάλογο συχνότητας. Στο τέλος κάθε καταλόγου συχνότητας εμφανίζονται τα *άπαξ λεγόμενα* (*hapaxes*), τύποι δηλαδή με μία μόνο εμφάνιση στο σώμα κειμένων (ή το κείμενο). Στη συνέχεια παρουσιάζονται οι 50 πρώτοι τύποι στον κατάλογο συχνότητας του αναλογικού ΣΕΚ με αντίστροφη σειρά (από τους λιγότερο στους περισσότερο συχνούς):

άβγαλτες, άβγαλτη, άβολες, άβουλες, άβουλη, άβουλο, άγνωστους, άγα-
μες, άγαμη, άγαμο, άγαμου, άγαν, άγαρμμη, άγαρμποι, άγγελε, άγγελό,
άγγιξέ, άγγιξα, άγγισμα, άγγλου, άγγλων, άγει, άγias, άγιοί, άγκιστρο,

Πίνακας 8.4: Πιο συχνόι γραμματικοί τύποι στα ποιήματα και τους στίχους του Νίκου Γκάτσου

Σειρά συχνότητας	Ποιήματα		Στίχοι	
	Λέξη	Συχνότητα	Λέξη	Συχνότητα
1	και	96	το	1134
2	τα	75	και	1076
3	να	74	μου	962
4	το	70	να	949
5	του	59	κι	881
6	της	57	τα	643
7	με	56	με	559
8	κι	51	του	488
9	θα	38	σου	482
10	στα	38	θα	432

Παρότι το δείγμα είναι σχετικά μικρό, έχει ενδιαφέρον ότι στους στίχους εμφανίζονται προσωπικές αντωνυμίες α' και β' προσώπου (*μου, σου*), όπως και το ότι ο τύπος *το* είναι πιο συχνός από το *και*. Οπωσδήποτε, μεγάλο ενδιαφέρον έχουν αντίστοιχα οι 10 πιο συχνόι λεξικοί τύποι σε αυτά τα σώματα κειμένων, όπως παρουσιάζονται στον Πίνακα 8.5.

Πίνακας 8.5: Πιο συχνόι λεξικοί τύποι στα ποιήματα και τους στίχους του Νίκου Γκάτσου

Σειρά συχνότητας	Ποιήματα		Στίχοι	
	Λέξη	Συχνότητα	Λέξη	Συχνότητα
1	καρδιά	150	μάτια	12
2	αγάπη	145	είναι	11
3	ζωή	104	καρδιά	7
4	παιδιά	101	φωτιά	6
5	κόσμου	77	παιδιά	5
6	μάννα	77	τραγούδι	5
7	νύχτα	74	φεγγαριού	5
8	πάει	74	χιόνι	5
9	πόρτα	72	αίμα	4
10	δρόμο	67	αυλή	4

Πιο συστηματικές συγκρίσεις μεταξύ καταλόγων συχνότητας γίνονται με τη χρήση του εργαλείου των λέξεων-κλειδιών, που θα παρουσιάσουμε σε επόμενη ενότητα (8.6).