Documentation for the use of the eAQUA function 'explorative search'

André Bünte

Abstract

The aim of this article is to provide a concise and comprehendible technical documentation of the eAQUA tool "explorative search" for students and scholars of classical and ancient studies. So in plain terms it shall be described what kind of information the user obtains, how this information is generated and which conclusions might be drawn from it. This pattern has been implanted in the composition of this technical documentation, which consists of four parts. First the functionality is on focus followed by the description of the results and thirdly by the definition of these results. To round it off the fourth part will show the analysis of these results and give possibilities to interpret them for a subsequent integration into the further work.

Keywords

Explorative search – text mining – word net – documentation

The following scheme depicts the tripartite structure of the documentation together with an overview of all functions of the explorative search.

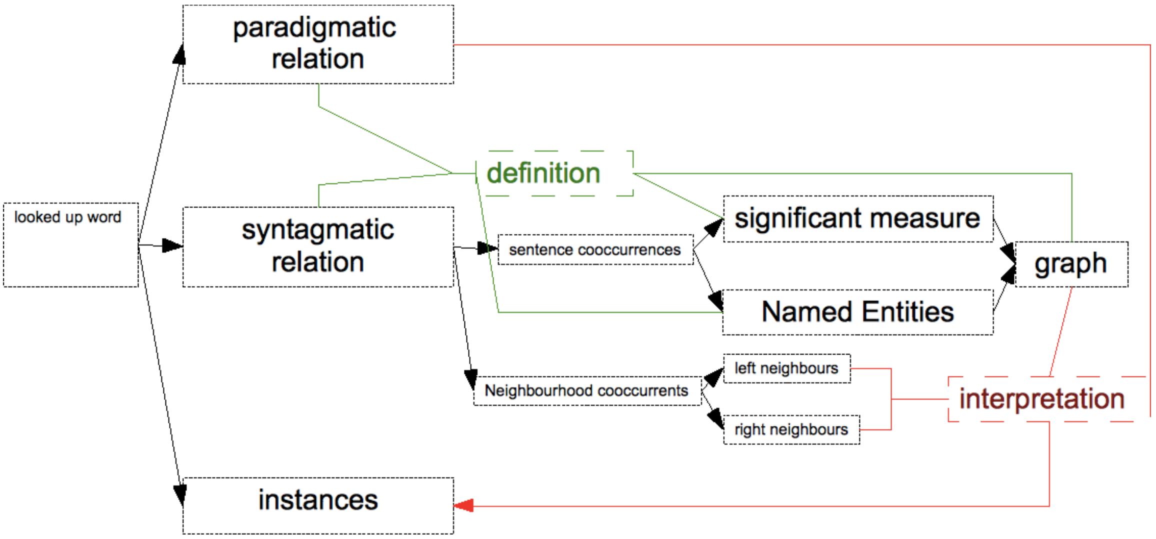


Fig. 1: Concept of the documentation

1. Functionality

At the very beginning the users shall enter a word in which they are interested. To be able to optimally observe the possibilities of the method it is advisable to chose a word that bears already some familiarity.

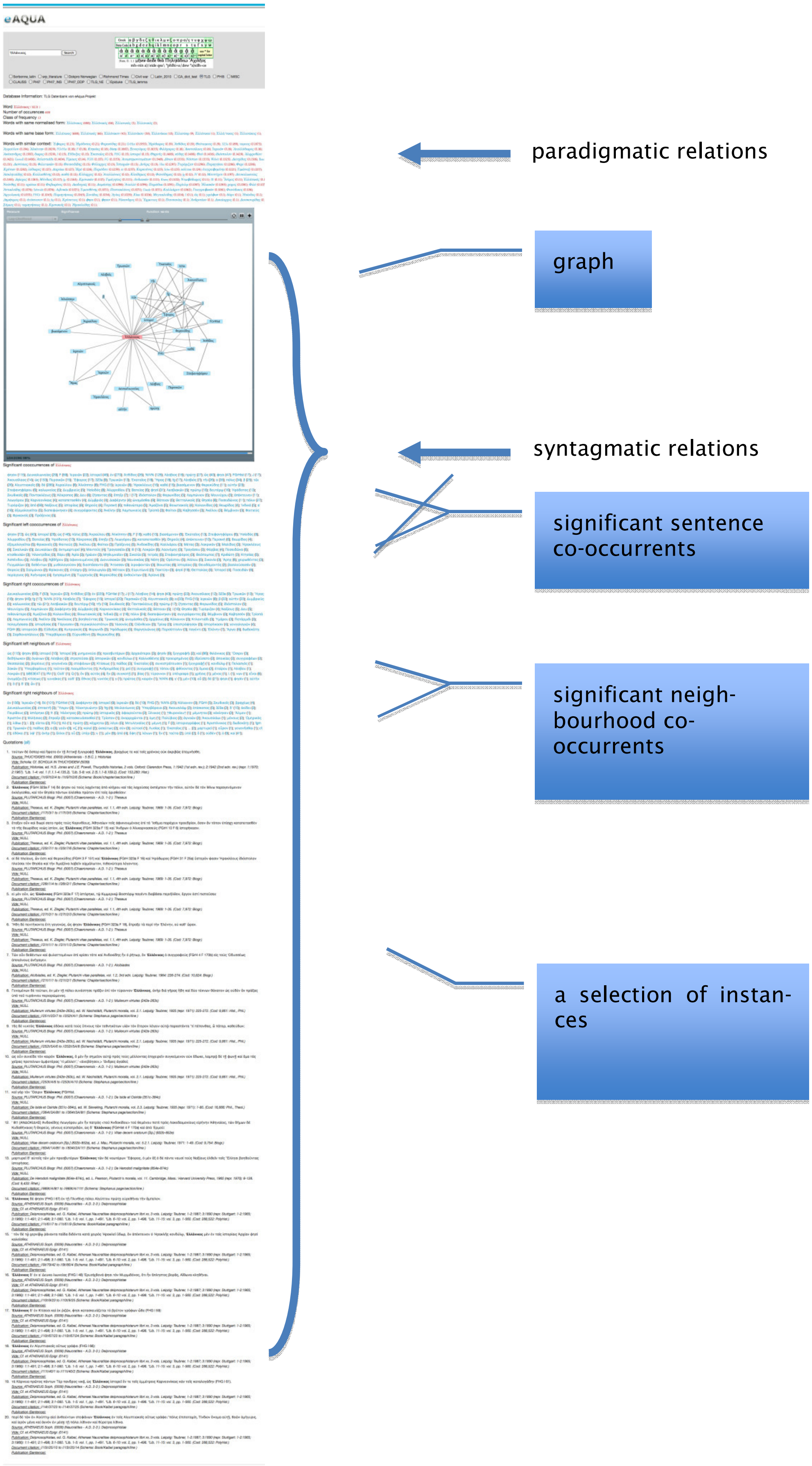
Example: Ἐλλάνικος

It is possible to enter the word in Unicode with diacritics or without. The software will combine all possibilities in which the word appears inside the corpus no matter if the word is written upper case or lower case, contains diacritics or not. Furthermore a Greek word might be looked up when it has been typed in BetaCode. The latter being a Latin transcription of the Greek alphabet. The picture on the right hand showing the keyboard layout aims to aid transcribing. Some examples of different methods entering the word would be

The next obligatory step is selecting the corpus in which the word should be analysed. There are several corpora available. For the example Ἐλλάνικος shall be analysed in the Greek literary texts, so "TLG" would be appropriate.

Example: TLG

Then, clicking on the "Search"-Button will initiate the analysis provided that the word is found in the corpus. Possible reasons for a non-detection will be discussed further below. Next the screen gets enlarged providing the results of the analysis. In Fig. 2 the whole page is depicted and the results are denoted orderly.


eAQUA
Search
Database Information: TLG Datenbank von eAqua Projekt
Word
Number of occurences
Class of frequency
Words with same normalised form:
Words with same base form:
Words with similar context:

paradigmatic relations

graph

syntagmatic relations

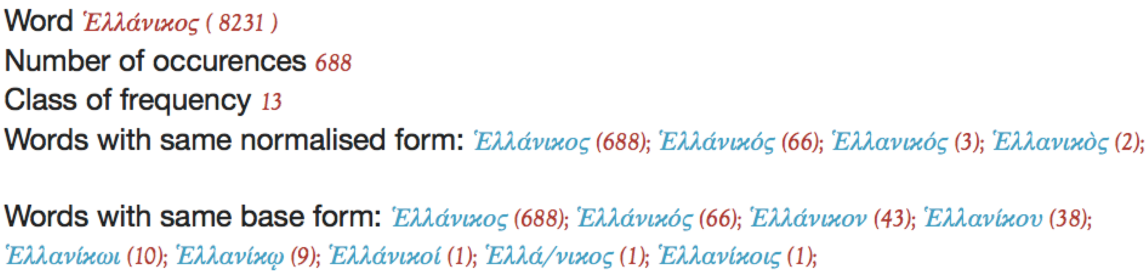
significant sentence co-occurrents

significant neighbourhood co-occurrents

a selection of instances

Fig. 2: page with the results

2. Description of the results



Word Ἑλλάνικος (8231)
Number of occurences 688
Class of frequency 13
Words with same normalised form: Ἑλλάνικος (688); Ἑλλάνικός (66); Ἑλλανικός (3); Ἑλλανικὸς (2);

Words with same base form: Ἑλλάνικος (688); Ἑλλάνικός (66); Ἑλλάνικον (43); Ἑλλανίκου (38); Ἑλλανίκωι (10); Ἑλλανίκῳ (9); Ἑλλάνικοί (1); Ἑλλά/νικος (1); Ἑλλανίκοις (1);

Fig. 3: basic statistical information of the word Ἑλλάνικος

Fig. 3 shows that the results of the explorative search can be grouped into four categories:

I) statistical information concerning the looked up word,
II) paradigmatic context of the looked up word,
III) syntagmatic context of the looked up word,
IV) a selection of instances including the looked up word that might be expanded.

Generally two types of values appear in brackets behind a word in the output: the first type consists of integral numbers mostly indicating a frequency in the corpus, the second concerns values between 0 and 1 that indicates a measure of similarity.

I) The statistical information

The values are always to be seen in relation to the selected corpus. "Word" contains the word itself with a bracketed integral number, in this case giving a biunique value for the word as an internal reference.

Example: Ἑλλάνικος (8231)

"Number of occurrences" states how often a word appears in the whole corpus.

Example: 688

"Class of frequency" states the power of relation of the most frequent word of the corpus with the looked up word.

Example: 13

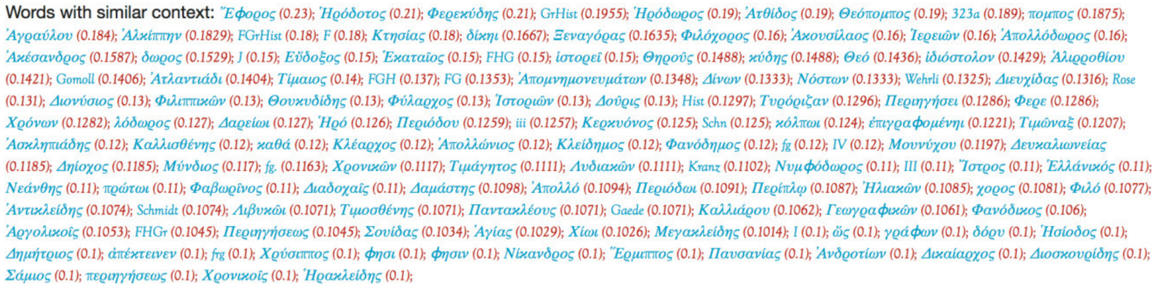
The most frequent word inside the TLG corpus 'καὶ' appears 2^{13} (which equals 8192) times more often than the looked up "Ἑλλάνικος', the latter having a frequency of 688, the former a frequency of 4.022.447.

"Words with same normalised form" states all other appearing words which are found to be similar with the looked up word but have a different notation e.g. they contain capital letters or have a diacritic in a different position. These words are treated generally as independent forms.

"Words with same base form" states all other appearing words that have been found to share the same base form with the looked up word, so that the different grammatical cases are covered. They are treated generally as independent forms.

II) The paradigmatic context

The paradigmatic context is represented by "Words with similar context".



Words with similar context: Ἔφορος (0.23); Ἡρόδοτος (0.21); Φερεκύδης (0.21); GrHist (0.1955); Ἡρόδωρος (0.19); Ἀτθίδος (0.19); Θεόπομπος (0.19); 323a (0.189); πομπος (0.1875);
Ἀγραύλου (0.184); Ἀλκίππην (0.1829); FGrHist (0.18); F (0.18); Κτησίας (0.18); δίκηι (0.1667); Ξεναγόρας (0.1635); Φιλόχορος (0.16); Ἀκουσίλαος (0.16); Ἱερειῶν (0.16); Ἀπολλόδωρος (0.16);
Ἀκέσανδρος (0.1587); δωρος (0.1529); J (0.15); Εὔδοξος (0.15); Ἑκαταῖος (0.15); FHG (0.15); ἱστορεῖ (0.15); Θηροῦς (0.1488); κύδης (0.1488); Θεό (0.1436); ἰδιόστολον (0.1429); Ἁλιρροθίου
(0.1421); Gomoll (0.1406); Ἀτλαντιάδι (0.1404); Τίμαιος (0.14); FGH (0.137); FG (0.1353); Ἀπομνημονευμάτων (0.1348); Δίνων (0.1333); Νόστων (0.1333); Wehrli (0.1325); Διευχίδας (0.1316); Rose
(0.131); Διονύσιος (0.13); Φιλιππικῶν (0.13); Θουκυδίδης (0.13); Φύλαρχος (0.13); Ἱστοριῶν (0.13); Δοῦρις (0.13); Hist (0.1297); Τυρόριζαν (0.1296); Περιηγήσει (0.1286); Φερε (0.1286);
Χρόνων (0.1282); λόδωρος (0.127); Δαρείωι (0.127); Ἡρό (0.126); Περιόδου (0.1259); iii (0.1257); Κερκυνόνος (0.125); Schn (0.125); κόλπωι (0.124); ἐπιγραφομένηι (0.1221); Τιμῶναξ (0.1207);
Ἀσκληπιάδης (0.12); Καλλισθένης (0.12); καθά (0.12); Κλέαρχος (0.12); Ἀπολλώνιος (0.12); Κλείδημος (0.12); Φανόδημος (0.12); fg (0.12); IV (0.12); Μουνύχου (0.1197); Δευκαλιωνείας
(0.1185); Δηίοχος (0.1185); Μύνδιος (0.117); fg. (0.1163); Χρονικῶν (0.1117); Τιμάγητος (0.1111); Λυδιακῶν (0.1111); Kranz (0.1102); Νυμφόδωρος (0.11); III (0.11); Ἴστρος (0.11); Ἑλλάνικός (0.11);
Νεάνθης (0.11); πρώτωι (0.11); Φαβωρῖνος (0.11); Διαδοχαῖς (0.11); Δαμάστης (0.1098); Ἀπολλό (0.1094); Περιόδωι (0.1091); Περίπλῳ (0.1087); Ἠλιακῶν (0.1085); χορος (0.1081); Φιλό (0.1077);
Ἀντικλείδης (0.1074); Schmidt (0.1074); Λιβυκῶι (0.1071); Τιμοσθένης (0.1071); Παντακλέους (0.1071); Gaede (0.1071); Καλλιάρου (0.1062); Γεωγραφικῶν (0.1061); Φανόδικος (0.106);
Ἀργολικοῖς (0.1053); FHGr (0.1045); Περιηγήσεως (0.1045); Σουίδας (0.1034); Ἀγίας (0.1029); Χίωι (0.1026); Μεγακλείδης (0.1014); I (0.1); ὥς (0.1); γράφων (0.1); δόρυ (0.1); Ἡσίοδος (0.1);
Δημήτριος (0.1); ἀπέκτεινεν (0.1); frg (0.1); Χρύσιππος (0.1); φησι (0.1); φησιν (0.1); Νίκανδρος (0.1); Ἕρμιππος (0.1); Παυσανίας (0.1); Ἀνδροτίων (0.1); Δικαίαρχος (0.1); Διοσκουρίδης (0.1);
Σάμιος (0.1); περιηγήσεως (0.1); Χρονικοῖς (0.1); Ἡρακλείδης (0.1);

Fig.4: List of Words with similar context like Ἑλλάνικος

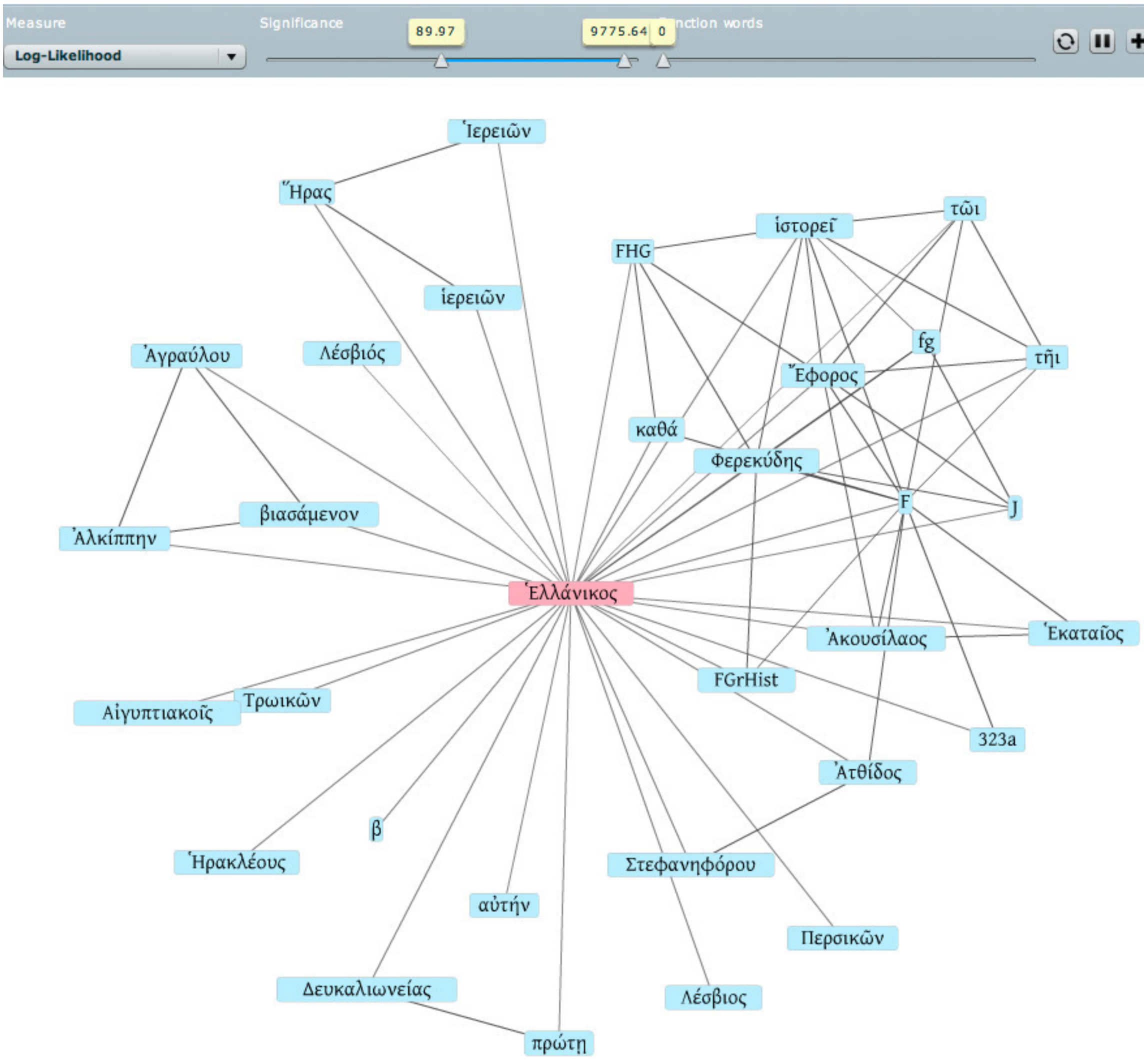
The indicated value of similarity is computed in relation to the looked up word. All words seeming to bear some analogy to the looked up word are given, because their co-occurrence profile is in some way similar. The results are sorted by similarity, starting with the most similar form.

Example: Ἔφορος (0.23)

III) The syntagmatic context

The visualisation as a word net and the lists of significant co-occurrences and significant neighbours show the syntagmatic context triply. The position in relation to the looked-up word forks additionally the significant coocurrences as well as the significant neighbours.

a) The word net



Measure: Log-Likelihood

Significance: 89.97 – 9775.64

Function words: 0

Fig. 5 Graph depicting a word net of the example Ἑλλάνικος with the standard adjustments

The word net consists of significant co-occurrences of the looked up word. The significance is indicated by proximity using a graph-layout, which is driven by a force-based algorithm.

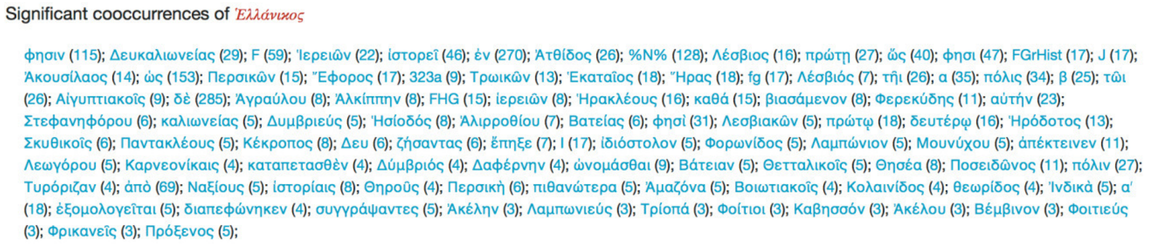
The mouse-over gesture changes the colour of the affected item as well as its co-occurrents and uncovers direct connections inside the word net easily.

Using the click-and-hold gesture it's possible to move any item around. If the dynamic mode is activated the graph will automatically be re-laid out. The button is located in the upper right corner.

The users select the significance measure in the drop down menu in the upper left corner. The 'Significance' slider next to it is a boundary marker for the significance measure. The word net only displays the values between the two sliders. With the slider on the right side the users choose a limit for the so-called 'Function words'. If it's set to 'zero' all function words are to be displayed, if it's set to 1000, the 1000 most frequent words of the corpus are to be dropped out. The setting 'zero' for example, displays all function words, whereas the setting '1000' leaves the 1000 most frequent words of the corpus out.

b) Significant co-occurrences

The hundred most significant co-occurrents are displayed here. Their significance is computed using the log-likelihood measure. The most significant co-occurrent is displayed first, the least significant last. The bracketed figures indicate the number of co-occurrences with the looked up word.



Significant cooccurrences of Ἑλλάνικος

φησιν (115); Δευκαλιωνείας (29); F (59); Ἱερειῶν (22); ἱστορεῖ (46); ἐν (270); Ἀτθίδος (26); %N% (128); Λέσβιος (16); πρώτη (27); ὥς (40); φησι (47); FGrHist (17); J (17); Ἀκουσίλαος (14); ὡς (153); Περσικῶν (15); Ἔφορος (17); 323a (9); Τρωικῶν (13); Ἑκαταῖος (18); Ἥρας (18); fg (17); Λέσβιός (7); τῆι (26); α (35); πόλις (34); β (25); τῶι (26); Αἰγυπτιακοῖς (9); δὲ (285); Ἀγραύλου (8); Ἀλκίππην (8); FHG (15); ἱερειῶν (8); Ἡρακλέους (16); καθά (15); βιασάμενον (8); Φερεκύδης (11); αὐτήν (23); Στεφανηφόρου (6); καλιωνείας (5); Δυμβριεύς (5); Ἡσίοδός (8); Ἀλιρροθίου (7); Βατείας (6); φησὶ (31); Λεσβιακῶν (5); πρώτῳ (18); δευτέρῳ (16); Ἡρόδοτος (13); Σκυθικοῖς (6); Παντακλέους (5); Κέκροπος (8); Δευ (6); ζήσαντας (6); ἔπηξε (7); I (17); ἰδιόστολον (5); Φορωνίδος (5); Λαμπώνιον (5); Μουνύχου (5); ἀπέκτεινεν (11); Λεωγόρου (5); Καρνεονίκαις (4); καταπετασθὲν (4); Δύμβριός (4); Δαφέρνην (4); ὠνομάσθαι (9); Βάτειαν (5); Θετταλικοῖς (5); Θησέα (8); Ποσειδῶνος (11); πόλιν (27); Τυρόριζαν (4); ἀπὸ (69); Ναξίους (5); ἱστορίαις (8); Θηροῦς (4); Περσικὴ (6); πιθανώτερα (5); Ἀμαζόνα (5); Βοιωτιακοῖς (4); Κολαινίδος (4); θεωρίδος (4); Ἰνδικὰ (5); α' (18); ἐξομολογεῖται (5); διαπεφώνηκεν (4); συγγράψαντες (5); Ἀκέλην (3); Λαμπωνιεύς (3); Τρίοπά (3); Φοίτιοι (3); Καβησσόν (3); Ἀκέλου (3); Βέμβινον (3); Φοιτιεύς (3); Φρικανεῖς (3); Πρόξενος (5);

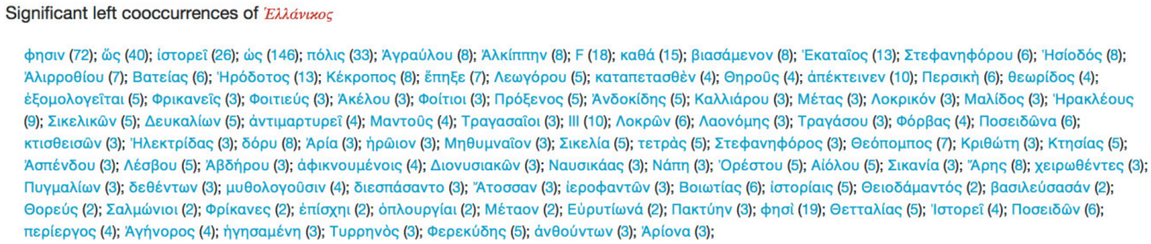
Fig. 6: List of the hundred most significant co-occurrents of Ἑλλάνικος

Example: φησιν (115)

There are 115 sentences in the corpus containing 'φησιν' along with 'Ἑλλάνικος'.

b 1) Significant left co-occurrences

The list shows the hundred most significant co-occurrents of the looked up word that appear in the sentence left to it. Their significance has been computed using the log-likelihood measure. The most significant co-occurrent is displayed first, the least significant last. The bracketed figures indicate the number of co-occurrences with the looked up word.



Significant left cooccurrences of Ἑλλάνικος

φησιν (72); ὥς (40); ἱστορεῖ (26); ὡς (146); πόλις (33); Ἀγραύλου (8); Ἀλκίππην (8); F (18); καθά (15); βιασάμενον (8); Ἑκαταῖος (13); Στεφανηφόρου (6); Ἡσίοδός (8); Ἀλιρροθίου (7); Βατείας (6); Ἡρόδοτος (13); Κέκροπος (8); ἔπηξε (7); Λεωγόρου (5); καταπετασθὲν (4); Θηροῦς (4); ἀπέκτεινεν (10); Περσικὴ (6); θεωρίδος (4); ἐξομολογεῖται (5); Φρικανεῖς (3); Φοιτιεύς (3); Ἀκέλου (3); Φοίτιοι (3); Πρόξενος (5); Ἀνδοκίδης (5); Καλλιάρου (3); Μέτας (3); Λοκρικόν (3); Μαλίδος (3); Ἡρακλέους (9); Σικελικῶν (5); Δευκαλίων (5); ἀντιμαρτυρεῖ (4); Μαντοῦς (4); Τραγασαῖοι (3); III (10); Λοκρῶν (6); Λαονόμης (3); Τραγάσου (3); Φόρβας (4); Ποσειδῶνα (6); κτισθεισῶν (3); Ἠλεκτρίδας (3); δόρυ (8); Ἀρία (3); ἡρῷον (3); Μηθυμναῖον (3); Σικελία (5); τετρὰς (5); Στεφανηφόρος (3); Θεόπομπος (7); Κριθώτη (3); Κτησίας (5); Ἀσπένδου (3); Λέσβου (5); Ἀβδήρου (3); ἀφικνουμένοις (4); Διονυσιακῶν (3); Ναυσικάας (3); Νάπη (3); Ὀρέστου (5); Αἰόλου (5); Σικανία (3); Ἄρης (8); χειρωθέντες (3); Πυγμαλίων (3); δεθέντων (3); μυθολογοῦσιν (4); διεσπάσαντο (3); Ἄτοσσαν (3); ἱεροφαντῶν (3); Βοιωτίας (6); ἱστορίαις (5); Θειοδάμαντός (2); βασιλεύσασάν (2); Θορεύς (2); Σαλμώνιοι (2); Φρίκανες (2); ἐπίσχηι (2); ὁπλουργίαι (2); Μέταον (2); Εὐρυτίωνά (2); Πακτύην (3); φησὶ (19); Θετταλίας (5); Ἱστορεῖ (4); Ποσειδῶν (6); περίεργος (4); Ἀγήνορος (4); ἡγησαμένη (3); Τυρρηνὸς (3); Φερεκύδης (5); ἀνθούντων (3); Ἀρίονα (3);

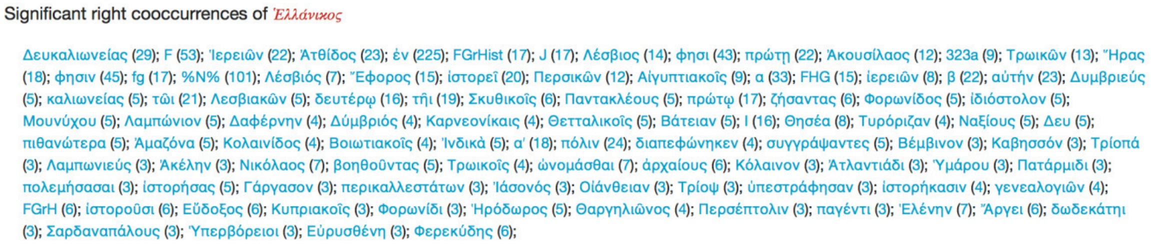
Fig. 7 List of the hundred most significant co-occurrents of Ἑλλάνικος appearing in the sentence before it

Example: φησιν (72)

There are 72 sentences in the corpus that have 'φησιν'somewhere left of 'Ἑλλάνικος'.

b 2) Significant right co-occurrences

This list shows the hundred most significant co-occurrents of the looked up word that appear in the sentence right to it. Their significance has been computed using the log-likelihood measure. The most significant co-occurrent is displayed first, the least significant last. The bracketed figures indicate the number of co-occurrences with the looked up word.



Significant right cooccurrences of Ἑλλάνικος

Δευκαλιωνείας (29); F (53); Ἱερειῶν (22); Ἀτθίδος (23); ἐν (225); FGrHist (17); J (17); Λέσβιος (14); φησι (43); πρώτη (22); Ἀκουσίλαος (12); 323a (9); Τρωικῶν (13); Ἥρας (18); φησιν (45); fg (17); %N% (101); Λέσβιός (7); Ἔφορος (15); ἱστορεῖ (20); Περσικῶν (12); Αἰγυπτιακοῖς (9); α (33); FHG (15); ἱερειῶν (8); β (22); αὐτήν (23); Δυμβριεύς (5); καλιωνείας (5); τῶι (21); Λεσβιακῶν (5); δευτέρῳ (16); τῆι (19); Σκυθικοῖς (6); Παντακλέους (5); πρώτῳ (17); ζήσαντας (6); Φορωνίδος (5); ἰδιόστολον (5); Μουνύχου (5); Λαμπώνιον (5); Δαφέρνην (4); Δύμβριός (4); Καρνεονίκαις (4); Θετταλικοῖς (5); Βάτειαν (5); I (16); Θησέα (8); Τυρόριζαν (4); Ναξίους (5); Δευ (5); πιθανώτερα (5); Ἀμαζόνα (5); Κολαινίδος (4); Βοιωτιακοῖς (4); Ἰνδικὰ (5); α' (18); πόλιν (24); διαπεφώνηκεν (4); συγγράψαντες (5); Βέμβινον (3); Καβησσόν (3); Τρίοπά (3); Λαμπωνιεύς (3); Ἀκέλην (3); Νικόλαος (7); βοηθοῦντας (5); Τρωικοῖς (4); ὠνομάσθαι (7); ἀρχαίους (6); Κόλαινον (3); Ἀτλαντιάδι (3); Ὑμάρου (3); Πατάρμιδι (3); πολεμήσασαι (3); ἱστορήσας (5); Γάργασον (3); περικαλλεστάτων (3); Ἰάσονός (3); Οἰάνθειαν (3); Τρίοψ (3); ὑπεστράφησαν (3); ἱστορήκασιν (4); γενεαλογιῶν (4); FGrH (6); ἱστοροῦσι (6); Εὔδοξος (6); Κυπριακοῖς (3); Φορωνίδι (3); Ἡρόδωρος (5); Θαργηλιῶνος (4); Περσέπτολιν (3); παγέντι (3); Ἑλένην (7); Ἄργει (6); δωδεκάτηι (3); Σαρδαναπάλους (3); Ὑπερβόρειοι (3); Εὐρυσθένη (3); Φερεκύδης (6);

Fig. 8 List of the hundred most significant co-occurrents of Ἑλλάνικος appearing in the sentence after it

Example: φησιν (45)

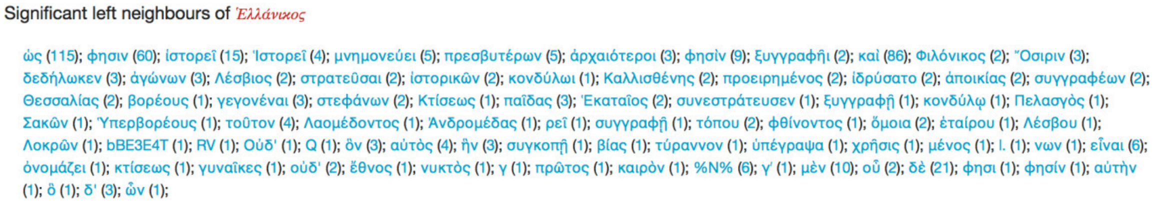
There are 45 sentences in the corpus that have 'φησιν' somewhere right of 'Ἑλλάνικος'.

c) Neighbourhood co-occurrences

The neighbourhood co-occurrences fork into significant left and right neighbours.

c 1) Significant left neighbours

The list shows the hundred most significant left neighbours of the looked up word. The significance has been computed using the log-likelihood measure. The most significant left neighbours displayed first, the least significant last. The bracketed figures indicate the number of left neighbourhood co-occurrences of the looked up word.



Significant left neighbours of Ἑλλάνικος

ὡς (115); φησιν (60); ἱστορεῖ (15); Ἱστορεῖ (4); μνημονεύει (5); πρεσβυτέρων (5); ἀρχαιότεροι (3); φησὶν (9); ξυγγραφῆι (2); καὶ (86); Φιλόνικος (2); Ὄσιριν (3); δεδήλωκεν (3); ἀγώνων (3); Λέσβιος (2); στρατεῦσαι (2); ἱστορικῶν (2); κονδύλωι (1); Καλλισθένης (2); προειρημένος (2); ἱδρύσατο (2); ἀποικίας (2); συγγραφέων (2); Θεσσαλίας (2); βορέους (1); γεγονέναι (3); στεφάνων (2); Κτίσεως (1); παῖδας (3); Ἑκαταῖος (2); συνεστράτευσεν (1); ξυγγραφῆ (1); κονδύλῳ (1); Πελασγὸς (1); Σακῶν (1); Ὑπερβορέους (1); τοῦτον (4); Λαομέδοντος (1); Ἀνδρομέδας (1); ρεῖ (1); συγγραφῇ (1); τόπου (2); φθίνοντος (1); ὅμοια (2); ἑταίρου (1); Λέσβου (1); Λοκρῶν (1); bBE3E4T (1); RV (1); Οὐδ' (1); Q (1); ὃν (3); αὐτὸς (4); ἦν (3); συγκοπῇ (1); βίας (1); τύραννον (1); ὑπέγραψα (1); χρῆσις (1); μένος (1); l. (1); νων (1); εἶναι (6); ὀνομάζει (1); κτίσεως (1); γυναῖκες (1); οὐδ' (2); ἔθνος (1); νυκτὸς (1); γ (1); πρῶτος (1); καιρὸν (1); %N% (6); γ' (1); μὲν (10); οὗ (2); δὲ (21); φησι (1); φησίν (1); αὐτὴν (1); ὃ (1); δ' (3); ὢν (1);

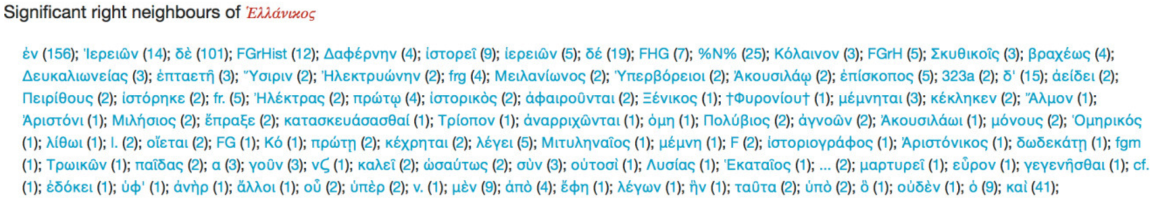
Fig. 9 List of the most significant words that appear in the sentence close before □λλάνικος

Example: φησιν (60)

There are 60 sentences in the corpus that have 'φησιν' as the left neighbour of 'Ἑλλάνικος'.

c 2) Significant right neighbours

The hundred most significant right neighbours of the looked up word are listed here. The significance has been computed using the log-likelihood measure. The most significant right neighbour appears in the first position, the least significant in the last. The bracketed figures indicate the number of right neighbourhood co-occurrences of the looked up word.



Significant right neighbours of Ἑλλάνικος

ἐν (156); Ἱερειῶν (14); δὲ (101); FGrHist (12); Δαφέρνην (4); ἱστορεῖ (9); ἱερειῶν (5); δέ (19); FHG (7); %N% (25); Κόλαινον (3); FGrH (5); Σκυθικοῖς (3); βραχέως (4); Δευκαλιωνείας (3); ἑπταετῆ (3); Ὕσιριν (2); Ἠλεκτρυώνην (2); frg (4); Μειλανίωνος (2); Ὑπερβόρειοι (2); Ἀκουσιλάῳ (2); ἐπίσκοπος (5); 323a (2); δ' (15); ἀείδει (2); Πειρίθους (2); ἱστόρηκε (2); fr. (5); Ἠλέκτρας (2); πρώτῳ (4); ἱστορικὸς (2); ἀφαιροῦνται (2); Ξένικος (1); †Φυρονίου† (1); μέμνηται (3); κέκληκεν (2); Ἄλμον (1); Ἀριστόνι (1); Μιλήσιος (2); ἔπραξε (2); κατασκευάσασθαί (1); Τρίοπον (1); ἀναρριχῶνται (1); ὁμη (1); Πολύβιος (2); ἀγνοῶν (2); Ἀκουσιλάωι (1); μόνους (2); Ὁμηρικός (1); λίθωι (1); l. (2); οἴεται (2); FG (1); Κό (1); πρώτη (2); κέχρηται (2); λέγει (5); Μιτυληναῖος (1); μέμνη (1); F (2); ἱστοριογράφος (1); Ἀριστόνικος (1); δωδεκάτῃ (1); fgm (1); Τρωικῶν (1); παῖδας (2); α (3); γοῦν (3); νϛ (1); καλεῖ (2); ὡσαύτως (2); σὺν (3); οὑτοσὶ (1); Λυσίας (1); Ἑκαταῖος (1); ... (2); μαρτυρεῖ (1); εὗρον (1); γεγενῆσθαι (1); cf. (1); ἐδόκει (1); ὑφ' (1); ἀνὴρ (1); ἄλλοι (1); οὗ (2); ὑπὲρ (2); v. (1); μὲν (9); ἀπὸ (4); ἔφη (1); λέγων (1); ἦν (1); ταῦτα (2); ὑπὸ (2); ὃ (1); οὐδὲν (1); ὁ (9); καὶ (41);

Fig. 10 List of the most significant words that appear in the sentence short after Ἑλλάνικος

Example: ἐν (156)

There are 156 sentences in the corpus that have 'ἐν' as the right neighbour of 'Ἑλλάνικος'.

3. Definition of the results

In order to make the results understandable successively will be defined in the following paragraph a) basic definitions, b) syntagmatic relations, c) paradigmatic relations, d) the significance measure and e) the word net.

a) Basic definitions

It's obligatory to understand the general and elementary structural relations between two linguistic tokens like phonemes, morphemes or words when using methods derived from the natural language procession, where mainly statistics of texts are computed and evaluated. This understanding derives from the linguistic structuralism.

Definition: The local context of a token is the set of tokens with whom together its appearing in one sentence.

Local context relates to the concept of sentence and is therefore limited to the linguistic level of sentences.

Example: Ἑλλάνικος ἐν Αἰγυπτιακοῖς οὕτως γράφει (FHG I 66):

If 'Ἑλλάνικος' is the looked up word then its local context consists of -'ἐν'; -'Αἰγυπτιακοῖς'; -'οὕτως'; -'γράφει'; -'FHG'; -'I'; -'66'.

b) Syntagmatic relationship

Definition: Two tokens are in syntagmatic relation, when they occur together. That means at least one local context exists that contains both tokens.

Example: Ἑλλάνικος ἐν Αἰγυπτιακοῖς οὕτως γράφει (FHG I 66):

The two tokens 'Ἑλλάνικος' and 'γράφει' are in a syntagmatic relation, because they appear together in at least one sentence. The joint appearance of two tokens in a local context is also called co-occurrence.

Definition: Two words are in a statistically syntagmatic relation, if they are in a syntagmatic relation that is statistically significant, quasi their joint appearance is not casually pertaining to a yet to be defined significance measure.

Example: Ἑλλάνικος ἐν Αἰγυπτιακοῖς οὕτως γράφει (FHG I 66):

The token 'Ἑλλάνικος' appears 688 times in the TLG corpus. The token 'ἐν' appears 783.892 times in the TLG corpus. These tokens appear jointly in one sentence 270 times. Therefore the high frequent token 'ἐν' appears in 39 % of all sentences containing 'Ἑλλάνικος'. This is statistically significant. In contrast to this the token 'Ἑλλάνικος' appears only in 0.03% of all sentences containing 'ἐν', therefore not being statistically significant. One can conclude that for the token 'Ἑλλάνικος' the co-occurrent 'ἐν' is significant while for the token 'ἐν' the co-occurrent 'Ἑλλάνικος' lacks statistical significance.

c) Paradigmatic relationship

Definition: The global context of a token can be defined as the set of all those tokens that have a syntagmatic statistical relationship to it.

Example: Ἑλλάνικος

The global context of Ἑλλάνικος contains all its significant co-occurrents (the hundred most significant are given):

φησιν (115); Δευκαλιωνείας (29); F (59); Ἱερειῶν (22); ἱστορεῖ (46); ἐν (270); Ἀτθίδος (26); %N% (128); Λέσβιος (16); πρώτῃ (27); ὥς (40); φησι (47); FGrHist (17); J (17); Ἀκουσίλαος (14); ὡς (153); Περσικῶν (15); Ἔφορος (17); 323a (9); Τρωικῶν (13); Ἑκαταῖος (18); Ἥρας (18); fg (17); Λέσβιός (7); τῆι (26); α (35); πόλις (34); β (25); τῶι (26); Αἰγυπτιακοῖς (9); δὲ (285); Ἀγραύλου (8); Ἀλκίππην (8); FHG (15); ἱερειῶν (8); Ἡρακλέους (16); καθά (15); βιασάμενον (8); Φερεκύδης (11); αὐτήν (23); Στεφανηφόρου (6); καλιωνείας (5); Δυμβριεύς (5); Ἡσίοδός (8); Ἀλιρροθίου (7); Βατείας (6); φησὶ (31); Λεσβιακῶν (5); πρώτῳ (18); δευτέρῳ (16); Ἡρόδοτος (13); Σκυθικοῖς (6); Παντακλέους (5); Κέκροπος (8); Δευ (6); ζήσαντας (6); ἔπηξε (7); I (17); ἰδιόστολον (5); Φορωνίδος (5); Λαμπώνιον (5); Μουνύχου (5); ἀπέκτεινεν (11); Λεωγόρου (5); Καρνεονίκαις (4); καταπετασθὲν (4); Δύμβριός (4); Δαφέρνην (4); ὠνομάσθαι (9); Βάτειαν (5); Θετταλικοῖς (5); Θησέα (8); Ποσειδῶνος (11); πόλιν (27); Τυρόριζαν (4); ἀπὸ (69); Ναξίους (5); ἱστορίαις (8); Θηροῦς (4); Περσικὴ (6); πιθανώτερα (5); Ἀμαζόνα (5); Βοιωτιακοῖς (4); Κολαινίδος (4); θεωρίδος (4); Ἰνδικὰ (5); α′ (18); ἐξομολογεῖται (5); διαπεφώνηκεν (4); συγγράψαντες (5); Ἀκέλην (3); Λαμπωνιεύς (3); Τρίοπά (3); Φοίτιοι (3); Καβησσόν (3); Ἀκέλου (3); Βέμβινον (3); Φοιτιεύς (3); Φρικανεῖς (3); Πρόξενος (5).

The global context of Ἔφορος contains the following tokens:

F (75); τῆι (52); φησιν (77); %N% (146); ἱστορεῖ (33); Θεόπομπος (27); Κυμαῖος (17); δ' (136); Καλλισθένης (19); ἐν (209); ὥς (37); ἄλλοι (36); Κύμης (13); FGrH (16); Ἑλλάνικος (17); ὡς (150); ἱστοροῦσι (15); Ὀρθαγόραν (8); ἱστόρηκεν (11); FHG (17); τῶι (27); φησὶν (41); δὲ (300); Ἀθηναίων (24); πόλις (33); J (12); Τίμαιος (14); εἰκοστῷ (11); Ἡσίοδός (9); Ἀκουσίλαος (9); sc. (16); εἴρηκεν (18); Ξενοφῶν (14); I (20); Κροίσου (10); ζήσαντας (7); Νεστανίαν (5); Σαυροματῶν (7); II (18); fg (12); περὶ (85); Ἀναξιμένης (9); Μέροπος (7); Κυμαῖον (6); Θουκυδίδης (13); FGrHist (9); Κλαζομενίων (6); Καρίας (10); φησι (27); πεμφθέντος (7); Ἑκαταῖος (11); Ἱστοριῶν (8); Καλάθουσαν (4); Ἀκραιφνίους (4); Φάλαννον (4); Ἑστιαῖός (4); Ἡρακλειδῶν (8); ἀνέγραψε (7); Σάμιος (8); Μαιωτῶν (5); Ἰσοκράτους (8); κτιζόντων (4); Κλείταρχος (6); Ναξίους (5); Λοκρῶν (8); μάντιν (8); Δαμάστης (5); ἔθνος (15); Δοῦρις (7); Ἀκαρνανίαν (6); ἱστοριῶν (8); δεδηλώκασι (4); Τιμοφάνους (4); Εὐρυσθένη (4); Ἀβαρνίδος (4); ἀρχαίους (7); λογιώτατοι (5); III (12); Ἀριστοτέλης (15); Κῦρον (8); κτίσμα (10); ἧς (23); μεταβαλλομένου (5); Ἡρόδοτος (10); νῆσον (11); Ζάλευκον (4); Δείνων (5); κατῴκισαν (5); οὐδετέρως (8); ἀντεξέπλευσαν (3); Βούδαρον (3); Φλογίδαν (3); ξενολογίᾳ (3); Νεστάνιος (3); Τυχίαν (3); Φωκαίδι (3); Καρπίδας (3); Κηφισσόδωρος (3); συνυποδεῖξαί (3); Βυβάστιον (3).

There are 21 tokens co-occurring both with ''Ἔφορος' and with 'Ἑλλάνικος'.

φησιν; F; ἱστορεῖ; ἐν; %N%; ὥς; φησι; FGrHist; J; Ἀκουσίλαος; ὡς; Ἑκαταῖος; fg; τῆι; πόλις; τῶι; δὲ; FHG; Ἡσίοδός; Ἡρόδοτος; Ναξίους.

The amount of correspondences suffices to define both contexts as similar as they are beyond a set threshold.

d) Significance measures

There are different measures to reckon the significance of a co-occurrence. These measures are called association measures as they interpret co-occurrence frequency data. For each acknowledged pair of words (co-occurrence) in the corpus the software has been computed an association score. Hence all measures consist of a different mathematic formula, some measures are heuristic and some are based on statistical hypothesis tests. The results or association scores computed by the measures cannot be compared directly.

The simplest measure is the word frequency. It is motivated by the principle of contiguity and focuses on the joint frequency of A and B and so, is a heuristic combination of the observed joint and marginal frequencies. The value of the computed significances is questionable since co-occurrences between two frequent but casually co-occurring words are also frequent and appear with this measure as significant. The principle of contiguity is meaningful if the words are equally distributed. Instead they are "Zipf-distributed".[1]

Mutual information is an association measure, whose basic concept is the comparison of the probability of observing word A and B together against probability of observing them independently. In case of low frequencies this measure is known to be prone of overestimating which leads to the (erroneous) display of rarely occurring co-occurrents.[2]

The Dice coefficient or "mutual expectation" is shown as 'sig_dice' in the menu. It is related to the Jaccard coefficient, which is defined as the size of the intersection divided by the size of the union of the representative sets of terms of both texts. It's an association measure estimating a maximum likelihood for the coefficient of association strength. It has a large sampling error rate especially in low-frequency data. On the other hand it is particularly sensitive to strong directional associations meaning that almost all occurrences of a word A co-occur with a word B as e.g. when dealing with idioms.[3] So the word net would display mainly co-occurrences consisting of occurrents that occur almost always together.

[1] Zipf, G.K., The Psycho-Biology of Language. An Introduction to Dynamic Philology. Cambridge, Mass. 1935.

[2] Church & Hanks (1990), Evert (2005).

[3] Evert (2005).

Log-likelihood has been used to compute the significance of the co-occurrents in the lists as its results are seen to correspond much closer to a true significance than the other measures. As stated above it is generally assumed that the occurrence of a word is independent of the occurrence of another word (statistical independency). Every language follows certain patterns, like the use of idioms that yields fixed co-occurrences. In case of low frequent co-occurrents, the assumption of statistical independence will give no reliable results, as the difference between the observed amount of co-occurrences and the estimated amount of the co-occurrences could be substantial enough to infer that the co-occurrence is significant while it is not. Therefore a statistical hypothesis test is performed comparing the observed co-occurrences and sample distributions in order to check, whether the observation is an unlikely outcome of the sample. A problem is the definition of 'unlikely'. The Log-likelihood ratio test takes into consideration that preferences (e.g. idioms) in the use of a language interfere with the assumed statistical independency that declines in relation to the distance between the co-occurrent and the looked up term and so defines the observations and expectations by two independent binomial distributions. This is an approximation to the Zipf-distribution. In contrast to the chi-squared test it has no explicit ranking of contingency tables but derives the ranking from the sampling distribution.[4]

The Local mutual information measure corresponds to the contribution of a co-occurrence to the total average mutual information of the corpus. It scales the mutual information measure with the co-occurrence frequency as a rough indicator of the amount of evidence provided by the co-occurrent.[5]

The chi-squared test is displayed as 'X^2' in the menu. Like the Log-likelihood ratio test it is an asymptotic hypothesis test that can deal with rather extreme outcomes. Both tests differ in the way they compute the sample distribution, as the test statistic of the chi-squared test is based on the comparison of the observed frequencies with the expected frequencies under the point null hypothesis of independence. It does not take the Zipf-distribution of words into consideration and fails on outliers.[6]

To sum it up, there is no 'right' significance measure. It's up to the users and their experience with the tool to know what they want to see and what not.

e) The word net

The graph depicting the word net is a visualisation of the syntagmatic relations. It can be defined as a graphic of the significant co-occurrences of a word. The co-occurrents are displayed as nodes being connected through edges representing the significance. In this way a netlike structure is formed as the co-occurrents are connected with each other depending on their occurring. In order to increase the usability of the network, like avoiding the overlay of two nodes, the graph is driven by a force based layout manager. Under the assumed basic condition that any node pushes off the others, the significance of co-occurrence of the two related nodes tightens their connectivity. The result is a word net where non-significant nodes are as far away as possible from each other and at the same time as close as possible to their significant co-occurrents.

4. Interpretation of the results

In this paragraph ways are shown how the users can interpret their results in connection with the instances. So, for example, questions like 'How can I include paradigmatic and syntagmatic relations along with left and right neighbours and the word net in my research?' will be answered.

a) Paradigmatic relations

[4] Dunning (1993); Evert (2005).

[5] Evert (2005).

[6] Evert (2005).

De Saussure called paradigmatic relationships between linguistic elements ‘associative’ relationships[7], because they represent the relationship between individual elements in specific environments with such elements in the memory that can potentially replace them. Paradigmatic relationships are based on the criteria of selection and distribution of linguistic elements, and are, for example, the basis for establishing the phoneme inventory of a language through the construction of minimal pairs, the replacement of sounds in an otherwise constant environment, which leads to a difference in meaning. Elements that have a paradigmatic relationship can potentially occur in the same context but are mutually exclusive in an actual concrete context because they stand in opposition to one another. Therefore it’s often impossible to discover them by just looking at the instances. So, the instances containing two linguistic elements that are determined to have a paradigmatic relationship have to be compared in order to see the evidence.

Example: Based on the example of 3 c), the paradigmatic relationship between 'Ἑλλάνικος' and ''Ἔφορος' in the TLG-E corpus, three instances for each form are shown that contain one of the 22 significant co-occurrents, which they have in common.

Ἑλλάνικος

sentence 1: Ὁμηρίδαι γένος **ἐν** Χίωι, ὅπερ Ἀκουσίλαος **ἐν** <γ> (2 **F** 2), **Ἑλλάνικος ἐν τῆι** Ἀτλαντιάδι ἀπὸ τοῦ ποιητοῦ φησὶν ὠνομάσθαι. (FGrHist 4 F 20 line 1-3)

sentence 2: ὅθεν ἡ **πόλις** Θῶνις ὠνόμασται, ὡς **ἱστορεῖ Ἑλλάνικος.** (FGrHist 4 F 20 line 5)

sentence 3: ἢ ὅτι ἔπηξε τὸ δόρυ ἐκεῖ ὁ Ἄρης **ἐν** τῇ πρὸς Ποσειδῶνα ὑπὲρ Ἁλιρροθίου δίκῃ, ὅτε ἀπέκτεινεν αὐτὸν βιασάμενον Ἀλκίππην τὴν αὐτοῦ καὶ Ἀγραύλου τῆς Κέκροπος θυγατρός, **ὥς φησιν Ἑλλάνικος** <**ἐν**> α΄ (FGrHist 4 **F** 38 et 323a **F** 1).

Ἔφορος

sentence 1: **πόλις** ἐστὶ τῆς Τρῳάδος Κεβρὴν, Κυμαίων ἀποικία, **ὥς φησιν Ἔφορος ἐν** α΄. (Harpocration p. 172 line 13-14)

sentence 2: **Ἔφορος** μὲν οὖν **φησιν** (FGrH 70 F 219), **ὡς** ἁλισκομένης τῆς νεὼς ἑαυτὸν ἀνέλοι, Τιμωνίδης **δέ**, πραττομέναις ἐξ ἀρχῆς ταῖς πράξεσι ταύταις μετὰ Δίωνος παραγενόμενος καὶ γράφων πρὸς Σπεύσιππον τὸν φιλόσοφον, **ἱστορεῖ** (FGrH 561 **F** 2) ζῶντα ληφθῆναι τῆς τριήρους εἰς τὴν γῆν ἐκπεσούσης τὸν Φίλιστον: (Plutarchus Dion 35 4-5)

sentence 3: ἀπὸ τοῦ πεμφθέντος ὑπὸ Κροίσου ἐπὶ ξενολογίαν μετὰ χρημάτων, **ὥς φησιν Ἔφορος**, εἶτα μεταβαλλομένου πρὸς Κῦρον. (Suda epsilon 3718a Adler line 1-3)

Exceptions: Under certain circumstances two normally not combinable lingustic elements like those having a paradigmatic relationship may occur, nonetheless, together in one sentence. In case of two proper names like Hellanicus and Ephorus this is undoubtly possible.

Example: περίεργος δ' ἂν εἴην ἐγὼ τοὺς ἐμοῦ μᾶλλον ἐπισταμέ νους διδάσκων ὅσα μὲν Ἑλλάνικος Ἀκουσιλάῳ περὶ τῶν γενεαλογιῶν διαπεφώνηκεν, ὅσα δὲ διορθοῦται τὸν Ἡσίοδον Ἀκουσίλαος, ἢ τίνα τρόπον Ἔφορος μὲν Ἑλλάνικον ἐν τοῖς πλείστοις ψευδόμενον ἐπιδείκνυσιν, Ἔφορον δὲ Τίμαιος καὶ Τίμαιον οἱ μετ' ἐκεῖνον γεγονότες, Ἡρόδοτον δὲ πάντες. (Flavius Josephus Contra Apionem 1.16-17)

Interpretation: 'Ἑλλάνικος' as well as ''Ἔφορος' are proper names. Additionally, they seem to have a paradigmatic relationship. The kind of this relationship can be made understandable, when their common significant co-occurrents are further investigated. Even in those six examples above one can easily see in a rather similar manner "XY φησιν..." ("XY says...") or "ὥς φησιν XY ἐν..." ("as XY says in...") that authors are mentioned by their ancient colleagues. As the works of both authors Hellanicus as well as Ephorus are lost to us, these citations are the only remains left nowadays. So, the way they are cited is a very significant criterion for them. Their common significant co-occurrents 'ὥς', 'φησιν' and 'ἐν' reveal this feature. Other com-

[7] de Saussure, F. (1916).

mon significant co-occurrents like 'ἱστορεῖ' or 'πόλις' show they were both historians who wrote at some point about cities. In our exception case they were cited in one sentence, because that witness was narrating about the relationship between the historians.

In linguistics, paradigmatic relationships can be determined by several types:

- synonymy - the relationship of the same or nearly the same meaning indicated by partial or total synonymy,
- hyponymy - the relationship of an inclusive meaning,
- opposition - the relationship of opposite meanings, which might be further diversified into antonymy, directional opposition, complementarity, heteronymy, incompatibility and conversity,
- semantic field - small semantic fields might be constituted by antonyms, directional oppositions and complementary expressions, whilst large semantic fields might by constituted by taxonomies and mereologies.

In the example Ἔφορος and Ἑλλάνικος have a similarity of denotation as both are 'cited historians' though they are not synonym. Therefore, they belong to the same (large) semantic field as they are inside a taxonomy, the hyponyms of 'cited historian'.

b) Syntagmatic relations

The syntagmatic relation of two terms suggests their common occurrence within the same syntagma and makes a relation of meaning obvious. In the method that is discussed here, the syntagmatic relationship is determined by the significance of the co-occurrences. The more significant a co-occurrence is, the more obvious is the syntagmatic relationship. Based on the general quantity of occurrences of each co-occurrent and the common occurrences, as well as their tf-idf weight[8], the explorative search method makes a selection of syntagmatic relations. Thus, the significance can be determined by selecting the significance measure, which is only possible in the graphical word net representation. The method shows therefore only the most important relationships of meaning of a term. This allows the quick and rough overview of the relationship of meanings of any word by means of his most significant co-occurrents.

b 1) The word net

Based on the distance between the nodes, the graph represents the significance of co-occurrences. The closer the terms are arranged, the more significant is their relationship. Occasionally, the word net yields another important feature displaying the ambiguity of the looked up term. This is possible when the meanings are so ambiguous that on one hand the co-occurrents belonging to a certain meaning are very narrowly arranged and on the other hand the distance to the co-occurrents that belong to other meanings are far enough away, so they can be clearly distinguished. A meaning would appear then as a nucleus or a cluster on the periphery of the word net.

b 2) Significant co-occurrents

The significance of the listed terms has been computed using the log-likelihood measure. According to the results the most significant term is on the first position followed subsequently by the next 99 significant terms. This group of terms constitutes a selection of contexts for the looked up word. The kind of context can vary a lot, depending largely on the corpus. It is a fundamental rule, to look in the given instances for the sentences containing the co-occurrent in focus, in order to read a meaningful interpretation of the result. Through this further details of the co-occurrence can be revealed. Why the co-occurrent is so significant, should always be a mandatory question.

[8] Tf-idf (term frequency–inverse document frequency) is a statistical measure that is used in order to estimate the importance of a term within a document or a corpus, cf. Salton (1989).

The lists of left and right significant co-occurrents dissociate further the co-occurrents according to their position in the sentence in relation to the looked up term. As the word order in classical languages is looser than in modern languages, it seems less meaningful. However, many idioms exist that have a fixed word order and can be perfectly analysed using these lists - it depends on the problem and matters of the individual researcher. Inversely, idioms can be recognised, if a fixed word order is obvious through examining these lists.

b 3) Left and right neighbours

The co-occurrents listed here are ordered by their significance. Words frequently occurring in an immediate proximity often have a special relationship. This relation highly depends on the part of speech taken by the looked up word. In case of nouns, this could be articles, possessive pronouns or, as it is common in ancient Greek, an adjective being inserted between the article and the noun, yielding a direct reference to the latter. Also, other nouns regularly having a direct reference to the looked up noun, can appear here, as it is very common in case of names, for example. There are many further instances where the immediate proximity of two terms is crucial for their relationship, as in case of measurements the figure and indication of measurements. Because of the loose word order, it is again mandatory to clarify the indications given by the neighbour- listings with the instances.

André Bünte
Lehrstuhl für Alte Geschichte, Historisches Seminar, Universität Leipzig
abunte@uni-leipzig.de

Literature

Church, K. W. / Hanks, P. (1990) Word association norms, mutual information and lexicography. In: *Computational Linguistics* 16(1): 22–29.

Evert, S. (2005) *The Statistics of Word Co-occurrences. Word Pairs and Collocations.* Dissertation, Universität Stuttgart.

Dunning, T. E. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* 19(1): 61–74.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: *Philosophical Magazine* 5 50(302): 157–175.

de Saussure, F. (1916) *Cours de linguistique générale*. ed. Bally, C. / Sechehaye, A. Paris.

Salton, G. (1989) *Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer*. Menlo Park: Addison-Wesley Publishing Company.