TEXTHAMMER, VER. 1.5. USER MANUAL

INTRODUCTION

The TextHammer software package is currently being developed by Mikhail Mikhailov and Juho Härme at the University of Tampere. It is used for searching the different corpora stored on the mustikka.uta.fi server.

The TextHammer package is being developed so that it is possible to access and search both monolingual and parallel corpora via a web interface. The corpora are stored on a server in Postgresql databases. The application consists of PHP scripts which run SQL queries on databases containing the corpus data and display search results in a web browser. The primary function of the software is to carry out searches of various kinds. It is not designed to perform filtering, sorting, rearranging or reordering, or numerous sophisticated statistical tests. These operations are readily available in spreadsheet and database applications (e.g. R, SPSS, Microsoft Excel), and it is easier therefore to load the search results from TextHammer into the relevant software in order to perform categorization, reorganizing or quantitative analysis.

TextHammer has both search utilities and maintenance tools. In this manual, we will only describe the search utilities and give some hints on their use.

MAIN MENU

After logging into the system, TextHammer's main menu is displayed on the user's computer screen together with the current news of the project.

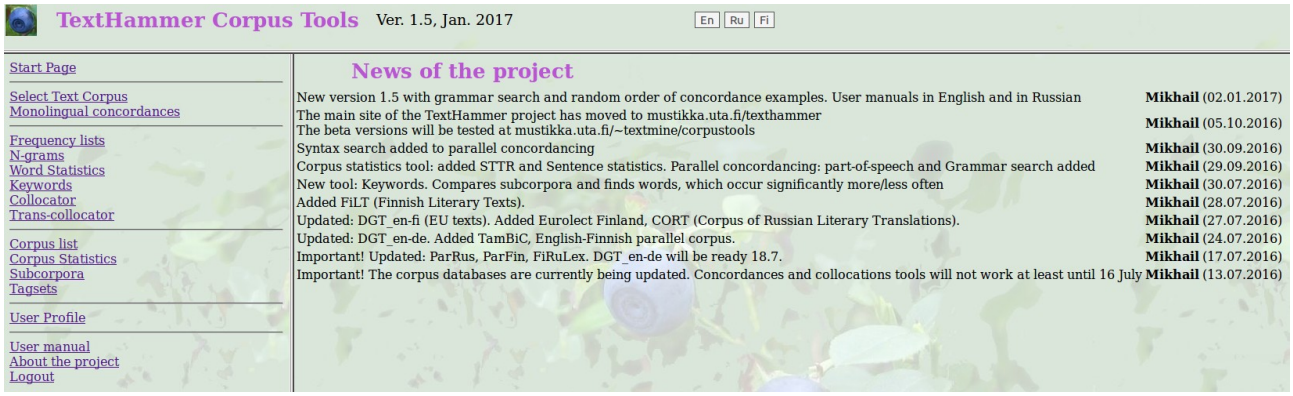


Figure 1 TextHammer: main menu

The current version of the package consists of the following functions:

- **Start page.** The main menu.

- **Select Text Corpus**. Begin by choosing from the list of available corpora the one you wish to work with. Click on **Select.** Access to different corpora is granted by the database administrator. This means that some corpora are available to all users, while others can only be used by a limited number of users or by their developers. To use corpora that are not currently in your list, contact the administrator of the server. The user can search only one corpus at a time.

- **Monolingual concordances, Parallel concordances.** These perform searches, the results of which are presented in single-language or two-language usage examples, aka concordances.

- **Frequency Lists**. This creates various kinds of frequency lists for the whole corpus or for its subcorpora.

- **N-grams.** This creates frequency lists of re-occurring multiword units from the texts of the corpus: bigrams (two words), trigrams (three words), etc.

- **Word Statistics**. This calculates more elaborate frequency statistics for specific word forms or lemmas, within different texts and subcorpora.

- **Keywords**. This tool compares two subcorpora and compiles lists of words that occur in one subcorpus significantly more often than in another.

- **Collocator**. This searches for the collocates (words that occur in close contexts with the search item) of word forms or lemmas.

- **Trans-collocator.** This searches for collocates across languages, i.e. those words which occur frequently in translations of the segments containing the search item.

- **Corpus list**. This gives a list of all the texts in the active corpus with the most important metadata (author, title, year of publication, publisher, etc).

- **Corpus statistics**. This provides general statistics on the corpus, subcorpora and separate texts, e.g. word count, number of sentences, etc.

- **Subcorpora**. Here the user can define any subcorpora inside the active corpus in order to perform searches on selected texts.

- **Tagsets.** Collection of links to annotation manuals for the parsers used for tagging the corpora.

- **User Profile**. Here the user can update the personal information, change language of the interface (English, Russian, Finnish) and his/her password.

The main menu remains visible on the left of the browser window, while the interface for the tool in use is displayed in the main part of the window, to the right of the browser menu. This makes navigation between different tools easier. The user can enter search parameters via the web form and start the search by clicking on the start button. The search results are then displayed in the same window and can be either copy/pasted or downloaded as a delimited text file.

The most important functions of the TextHammer program are described below.

SUBCORPORA

Although the **Subcorpora** tool is at the end of the menu and does not output any data in itself, it is one of the central elements of the program. Often the user will not need data from all the texts in the corpus, but only data from texts of a certain genre, texts by particular authors or just one specific text. The Subcorpora tool allows the user to create such groups of texts for conducting searches on different parts of the corpus. For creating subcorpora, the user can perform searches within list of texts by criteria: author, title, language, etc (see Figure 2). The user can also load, edit and save under different name an existing subcorpus. And if one wishes to work with a single text, it is possible to create a subcorpus consisting of that text only.

When working with a parallel corpus, one does not have to (and cannot) include texts in all languages into a single corpus. The subcorpora for parallel corpora are created having in mind the language, on which the searches will be performed. Often, several subcorpora can be created for the same texts to search in different directions (e.g. English → Finnish and Finnish → English).

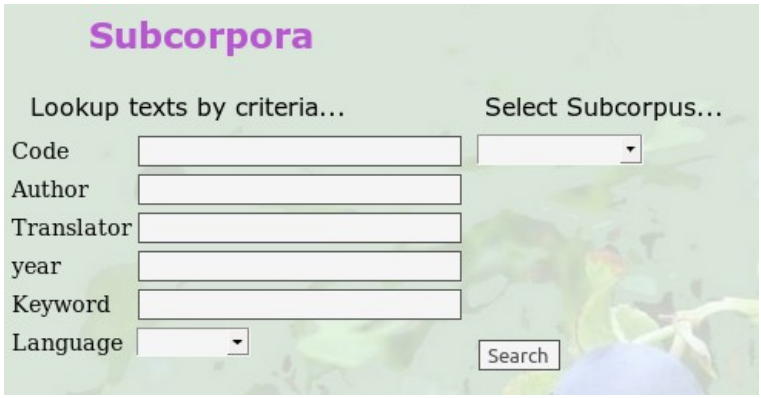


Figure 2 The online query form of the Subcorpora tool

After performing a search, the list of texts answering the specified criteria is displayed in the browser screen (Fig. 3). The user can tick the texts to be included in a new subcorpus (without necessarily including all the search results) and save it under a suitable name. The subcorpus will then be available for use with other TextHammer tools. The subcorpus will also be available to other users of the corpus, and so it should be given a suitable description. Note that a subcorpus which has been created in this way is a virtual data set: no texts are physically copied and if the user deletes the subcorpus, no data is actually deleted.

Figure 3 The interface for creating a subcorpus

MONOLINGUAL AND PARALLEL CONCORDANCES

The program has two concordancer tools: one for searching in monolingual corpora and another for parallel corpora. With parallel corpora, both monolingual and parallel concordancing are possible depending on the research task. The interfaces of both tools are similar, the main difference being that the parallel concordancer works with parallel corpora and outputs bitexts (corresponding text segments). Here, only the parallel concordancer will be described.

The concordance search query is defined and submitted via the web form on Figure 4. The TextHammer program generates only bilingual concordances, even if there are more than two languages in the corpus. The search is performed on the texts in **Language 1**. **Language 2** is the language for which corresponding segments will be displayed.

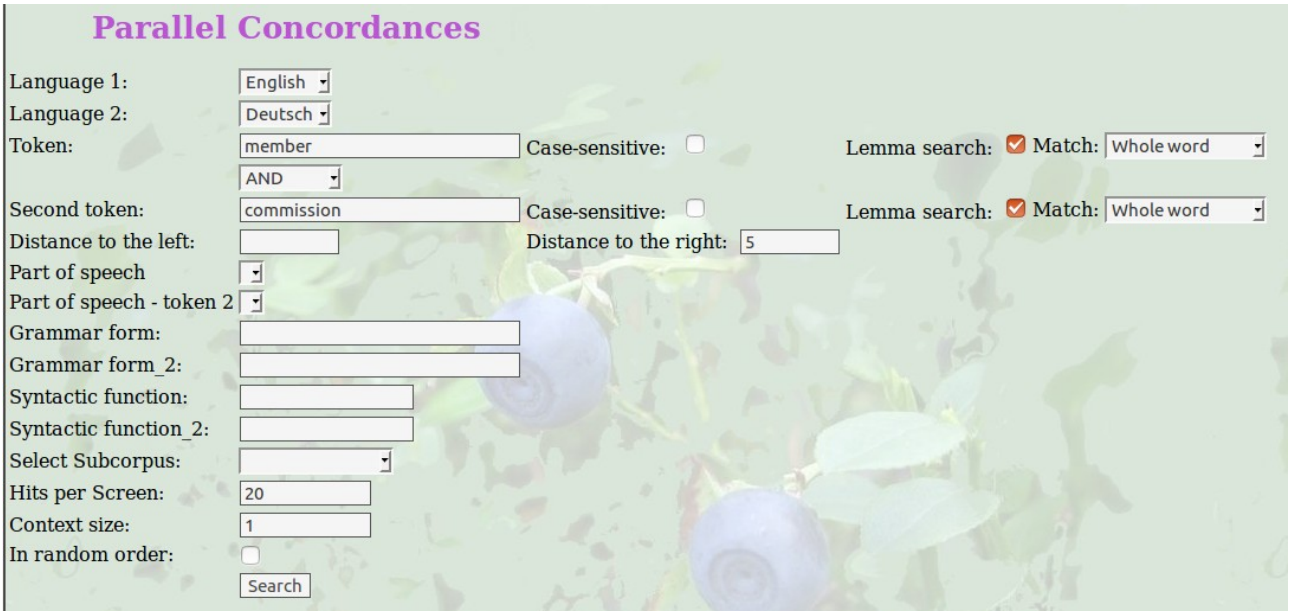


Figure 4 Concordance query interface

The **Token**, i.e. the search item, can be a single word form or a lemma. The search engine can look for exact matches, or for the start, end or any part of a word form or lemma. The search can be performed for one or two tokens and the user can specify the relations between them: both items present (AND, which is the default), one of the items present (OR), or the second item not present (NOT).

The **Distance to the left** and **Distance to the right** parameters specify the length of context in which the second token (if used) is expected to be found (or not found); the default values are 1 to the right and 1 to the left.

If the texts are morphologically tagged, **parts of speech** and **grammatical forms** can be set as additional search criteria. The grammar tags depend on the parser used for grammatical analysis. The links for the relevant tagsets' descriptions are available from the **Tagsets** in the main menu. If the part of speech or grammatical form is specified and the search string is left blank, the program will generate a concordance for the part of speech/grammatical form (e.g. all nouns in the Dative).

If **In random order** check box is checked, the concordance search will be performed in random order. This feature is useful when a large number of examples is expected and the user needs examples from different texts.

If the user needs the complete concordance, it is recommended to set **Hits per screen** to a value exceeding expected number of examples and the option **In random order** should be switched off.

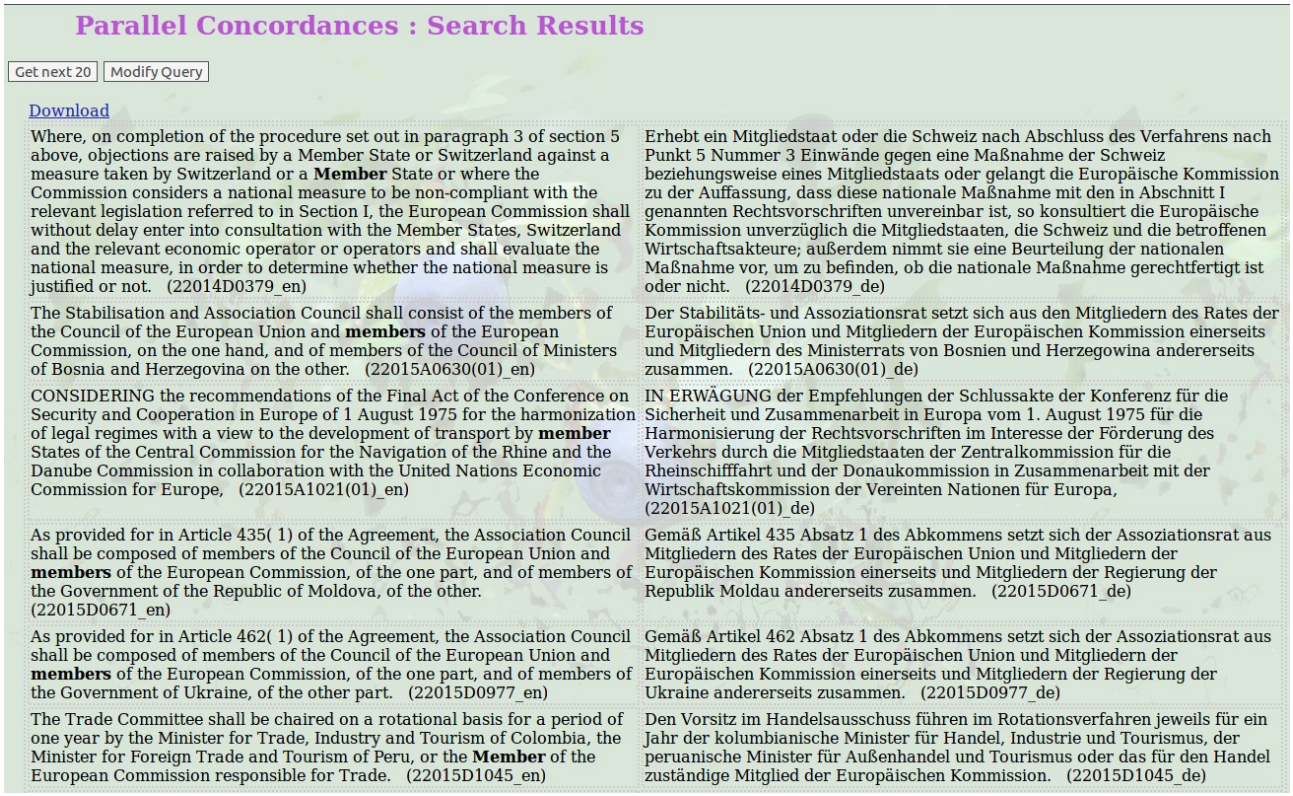


Figure 5 Search results in the window of the TextHammer concordancer, showing the expression *member + commission* in the DGT_en-de corpus.

Searches can be performed on all the texts in the corpus or on a subcorpus (see **subcorpora** above) which can be selected from a drop-down list. The user can also specify the number of examples to be found, and the size of the surrounding context.

The search results can be downloaded to the researcher's workstation by clicking on the hyperlink **Download**. They are saved as delimited text files and can be loaded directly into spreadsheet software (see **Introduction** above).

WORD FREQUENCIES

This tool generates frequency lists for the corpus or for any subcorpora. The program can generate lists of word forms, or lists of lemmas if the corpus is lemmatized. If the corpus is grammatically annotated, the user can create a frequency list for all the grammatical tags. The user should be aware that the lemma lists and grammatical frequency lists may also contain errors if the annotated texts have not been manually checked.

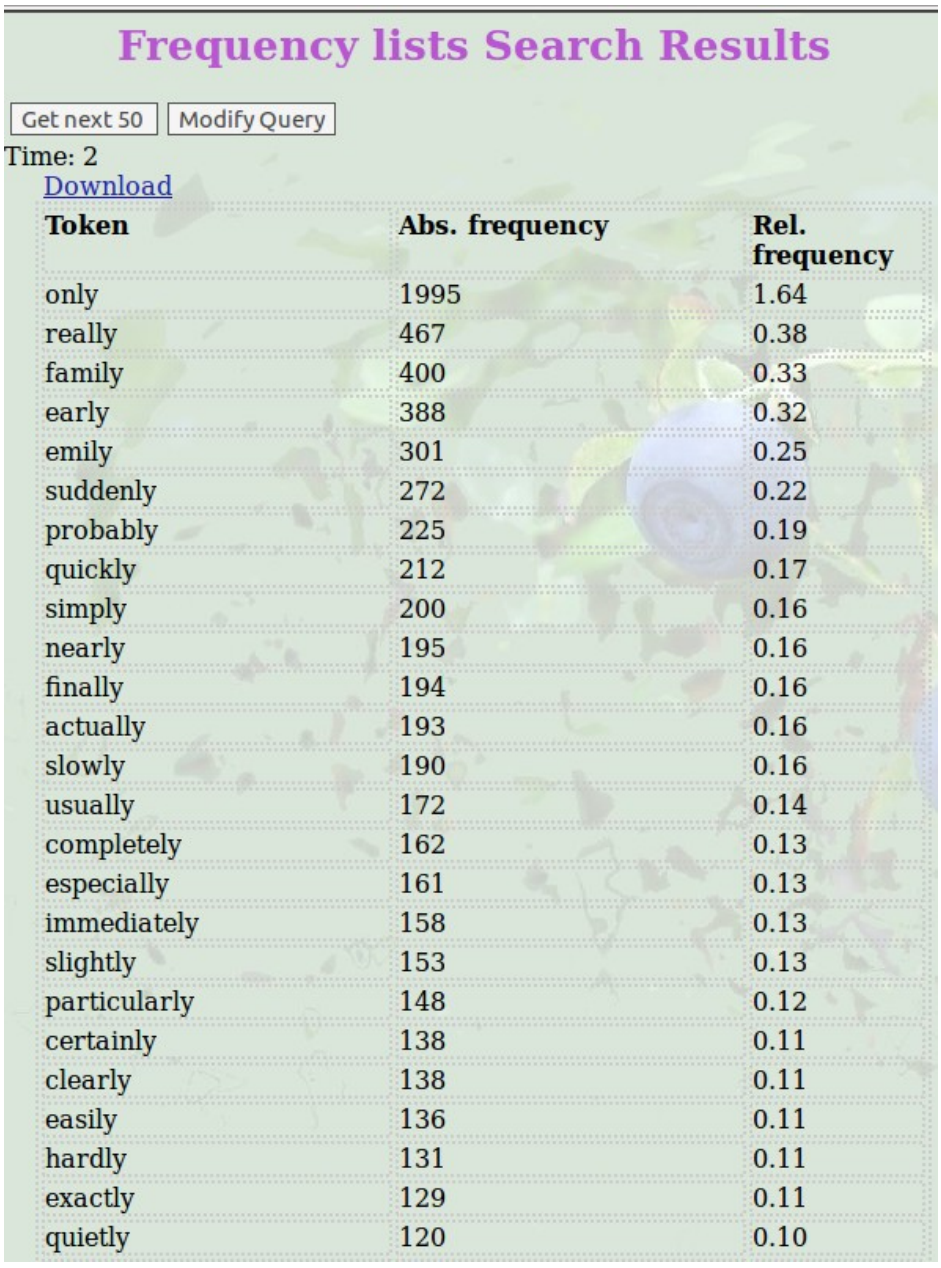
Figure 6 Web form for defining the parameters for the new word listCreating a complete frequency list can take a long time, especially if the corpus is large. If information on only one word or group of similar words (e.g. with the same stem) is required, the user can enter a search substring like that in Figure 6. This makes the search much faster.

In many cases, lists are needed for separate texts or groups of texts. To obtain these, the user chooses the relevant subcorpus from the list of subcorpora. If there is no relevant subcorpus available, the user must first create the subcorpus (see. the section Subcorpora).

When working online with the search results, the user can also display the results gradually, in groups. A display list of 10 items fits conveniently into the browser window and the search results are displayed faster (because outputting a very long list on-screen might take a long time). To get the next portion of elements or to return to the previous portion, press Next/Previous X words.

The program calculates both absolute and relative frequencies (per 1000 running words). The lists can also be ordered in descending order of frequency or in ascending alphabetical order. These options are useful if the user wishes just to have a look at the list. If further work on the search results (sorting, filtering, etc) is planned, it is better to download the list and transfer it into a spreadsheet for further processing.

Figure 7 shows the frequencies for words ending with the string *-ly*. The search was carried out on a lemmatized English word list. The items of interest are the adverbs with the suffix *-ly*. When irrelevant items (*only, family, Emily,* etc.) have been removed, the list can give the researcher a good idea of what can be found in the corpus.



Frequency lists Search Results

Get next 50 | Modify Query

Time: 2

Download

Token	Abs. frequency	Rel. frequency
only	1995	1.64
really	467	0.38
family	400	0.33
early	388	0.32
emily	301	0.25
suddenly	272	0.22
probably	225	0.19
quickly	212	0.17
simply	200	0.16
nearly	195	0.16
finally	194	0.16
actually	193	0.16
slowly	190	0.16
usually	172	0.14
completely	162	0.13
especially	161	0.13
immediately	158	0.13
slightly	153	0.13
particularly	148	0.12
certainly	138	0.11
clearly	138	0.11
easily	136	0.11
hardly	131	0.11
exactly	129	0.11
quietly	120	0.10

Figure 7 Word Frequencies: search resuts

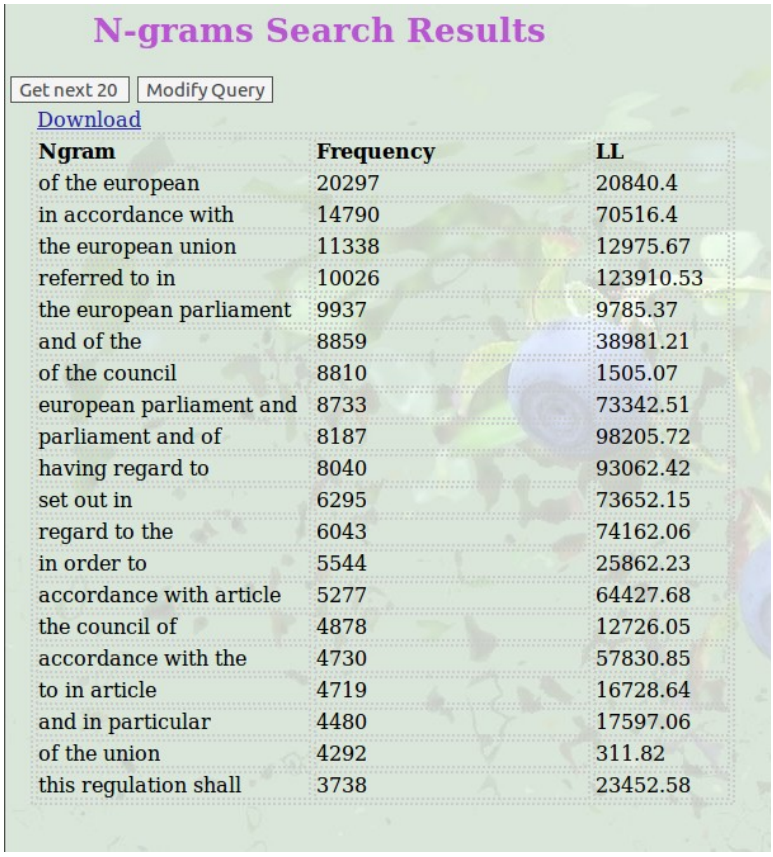
The list can be downloaded to the user's computer by clicking the hyperlink **Download**. The list is thereby saved as a delimited text file (csv) with the table columns separated by tab characters, and the rows by paragraph marks. Such files can easily be opened, of course, by any text editor or word processor, but spreadsheet software (Microsoft Excel, LibreOffice Calc, etc) is much more effective if further processing is to be carried out. (It is important to remember that the spreadsheet program may be confused by the conventions for representing the frequencies when the csv file is imported. In British and American usage, decimals are signaled by a decimal point, while in many European countries a comma is used. Thus, if the country in the regional settings of the operational system of the workstation uses a comma, the system will not recognize sequences like '1.3' or '15.2' as decimal numbers, but as dates (i.e. March 1st and February 15th). To overcome this problem, the user should define the column as a number column in the file import dialogue box.)

N-GRAMS

The N-grams tool finds in the corpus multiword units which co-occur above a certain frequency limit. The p-value is calculated to evaluate the collocation strength of the elements. The tool helps to find terms, proper names, idioms and cliches in the corpus. Please note that processing of large corpora can take a lot of time.

Figure 8 N-grams tool: query interface

To make the search faster, the researcher can set a higher lower frequency limit and/or work with subcorpora and not with the whole corpus.



N-grams Search Results

Get next 20 | Modify Query

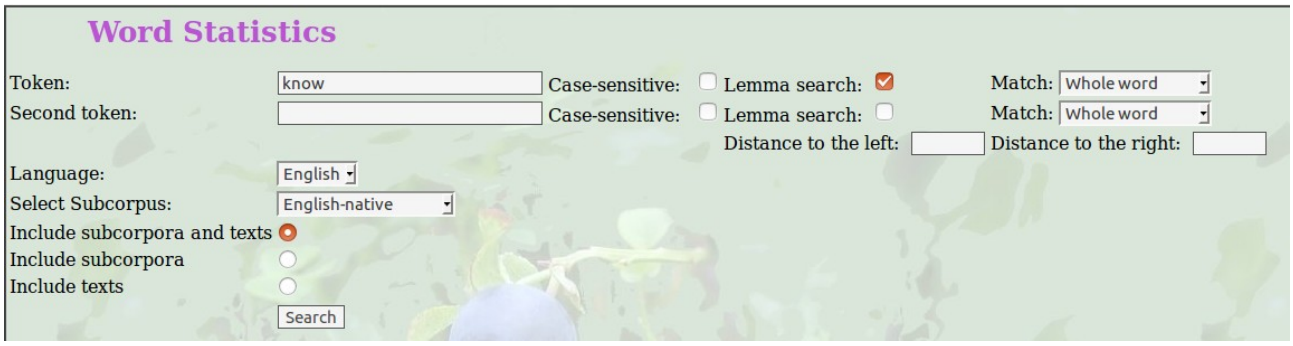
Download

Ngram	Frequency	LL
of the european	20297	20840.4
in accordance with	14790	70516.4
the european union	11338	12975.67
referred to in	10026	123910.53
the european parliament	9937	9785.37
and of the	8859	38981.21
of the council	8810	1505.07
european parliament and	8733	73342.51
parliament and of	8187	98205.72
having regard to	8040	93062.42
set out in	6295	73652.15
regard to the	6043	74162.06
in order to	5544	25862.23
accordance with article	5277	64427.68
the council of	4878	12726.05
accordance with the	4730	57830.85
to in article	4719	16728.64
and in particular	4480	17597.06
of the union	4292	311.82
this regulation shall	3738	23452.58

Figure 9. N-grams tool: search results

WORD STATISTICS

The **Word Frequencies** tool generates frequency lists for the whole corpus or for a specified subcorpus. To study the distribution of a word across a number of texts and/or subcorpora, the user would have to run the **Word Frequencies** program many times and might easily forget to check some of the subcorpora. The **Word Statistics** tool was developed therefore to make it possible to calculate the frequencies of words in different texts quickly.



Word Statistics

Token: know Case-sensitive: Lemma search: Match: Whole word

Second token: Case-sensitive: Lemma search: Match: Whole word

Distance to the left: Distance to the right:

Language: English

Select Subcorpus: English-native

Include subcorpora and texts

Include subcorpora

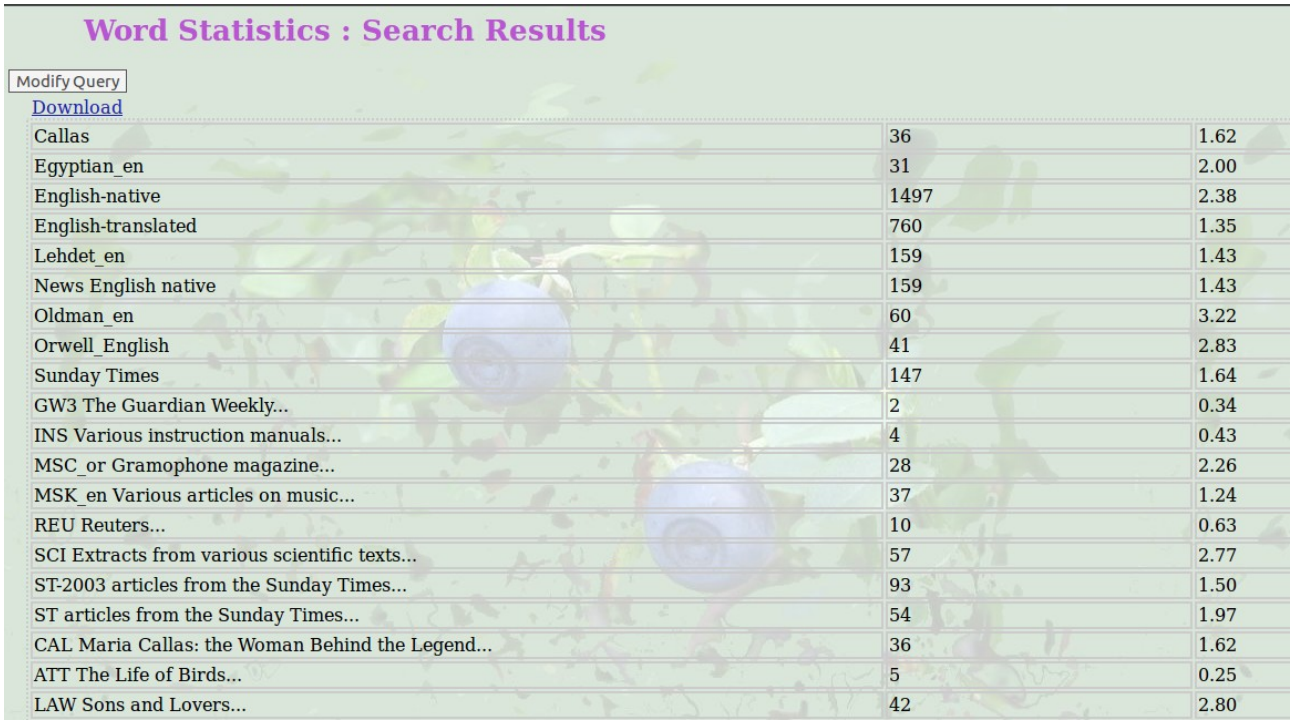
Include texts

Search

Figure 10 Word Statistics tool: query interface

The query interface for the Word Statistics tool follows the same principles as those for **Concordances**, **Word Frequencies** and **Collocations**: searches can be performed for one or two items (both word forms and lemmas), different kinds of matching can be used, and search can be limited to a subcorpus, if necessary. The search results can include frequencies for different subcorpora and/or separate texts.

The results of the search are displayed in table form, and they can also be downloaded to the user's computer. The tool can be very useful for studying the dispersion of words across various subcorpora or different texts, and for detecting significant differences between frequencies in different texts.



Callas	36	1.62
Egyptian_en	31	2.00
English-native	1497	2.38
English-translated	760	1.35
Lehdet_en	159	1.43
News English native	159	1.43
Oldman_en	60	3.22
Orwell_English	41	2.83
Sunday Times	147	1.64
GW3 The Guardian Weekly...	2	0.34
INS Various instruction manuals...	4	0.43
MSC_or Gramophone magazine...	28	2.26
MSK_en Various articles on music...	37	1.24
REU Reuters...	10	0.63
SCI Extracts from various scientific texts...	57	2.77
ST-2003 articles from the Sunday Times...	93	1.50
ST articles from the Sunday Times...	54	1.97
CAL Maria Callas: the Woman Behind the Legend...	36	1.62
ATT The Life of Birds...	5	0.25
LAW Sons and Lovers...	42	2.80

Figure 11 Word Statistics tool: search results

KEYWORDS

This tool performs lexical comparison of subcorpora. It creates frequency lists (textforms or lemmas) and finds the elements, which occur in one list significantly more often than in another. The log-likelihood index is used for the purpose. The tool works the same way as the Keywords tool in WordSmith Tools program package. The main difference is that in TextHammer one can compare lemmatized word lists, which is very important for languages with rich morphology.

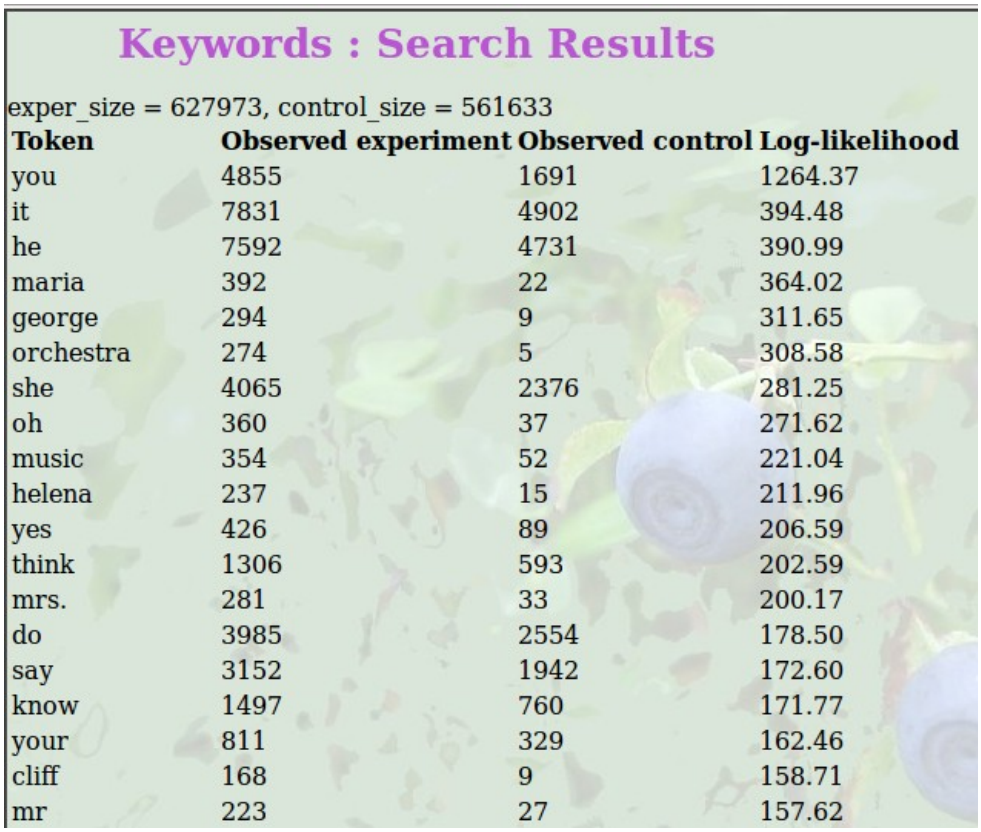
To perform keyword search:

- Using the Subcorpora tool, create the relevant subcorpora (they can also consist of single texts if the task is to compare single texts)
- In the Keywords menu select the language and the two subcorpora to compare.
- If the subcorpora are very large, it may be practical to increase the minimal frequency and the mininal log-likelihood value.
- The tool compares lemmatised word lists by default. Untick the Lemma search, if needed.

Figure 12. Keywords tool: menu.

The tool compiles two frequency lists, which can take a long time, if subcorpora are large.

Then it outputs to the screen the results. It shows frequencies of words in both subcorpora and the log-likelihood index value, which is negative for the words more frequently used in the second subcorpus. The results are sorted by LL in descending order.



Keywords : Search Results

exper_size = 627973, control_size = 561633

Token	Observed experiment	Observed control	Log-likelihood
you	4855	1691	1264.37
it	7831	4902	394.48
he	7592	4731	390.99
maria	392	22	364.02
george	294	9	311.65
orchestra	274	5	308.58
she	4065	2376	281.25
oh	360	37	271.62
music	354	52	221.04
helena	237	15	211.96
yes	426	89	206.59
think	1306	593	202.59
mrs.	281	33	200.17
do	3985	2554	178.50
say	3152	1942	172.60
know	1497	760	171.77
your	811	329	162.46
cliff	168	9	158.71
mr	223	27	157.62

Figure 13 Keywords tool: search results.

COLLOCATOR

The Collocator tool searches for words occurring in the immediate context of the search item. The program can look for word forms or lemmas, and it can also lemmatize the collocates. The user can also define the span of the context to be included (**Distance to the left/to the right**) and the minimal total frequency of the collocate. This is the sum of all the occurrences of each word occurring with the search item in the specified word span. If the value is set to 1, the program will find all the words adjacent to the search word. The program calculates the log-likelihood coefficient (LL) for the collocate candidates, which shows the strength of the collocation, and removes those with very low LL values. As with the Concordancer, searches can be performed on the whole corpus or on a subcorpus.

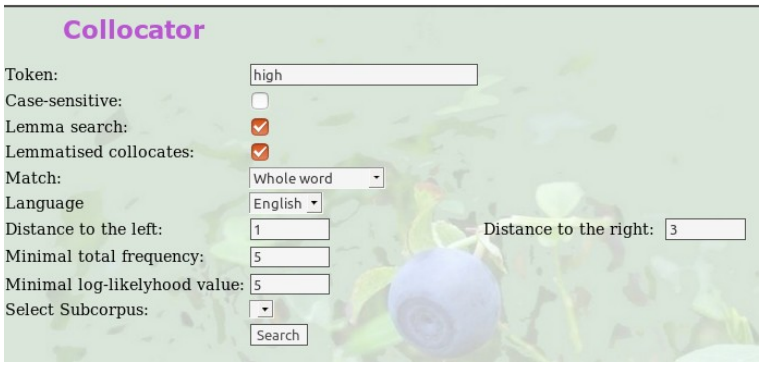


Figure 14. Collocations: the query interface

The resulting list of collocates is displayed in descending order of LL. The column headings refer to the distance of the collocate from the search word: L1 = first word to the left, L2 = second word to the left, etc; R1 = first word to the right, R2 = second word to the right, etc.

The results of a search for the collocates of the word *high* in the DGT corpus (Figure 13) reveal some of the most frequently co-occurring word combinations in the corpus: *high representative*, *high level*, *very high* etc. If necessary, these phrases can be studied further with the help of the **Concordances** tool.

Collocator : Search Results

Download

Modify Query

Search word: **high**, (sub)corpus size: 8412651

Word	L1	R1	R2	R3	Sum	LL
representative	0	279	0	0	279	524.53
level	8	213	31	9	261	330.68
voltage	0	130	0	4	134	269.29
permeability	0	64	0	0	64	208
very	74	0	0	4	78	131.23
quality	1	90	0	1	92	127.89
the	763	1	39	309	1112	120.19
bus	0	0	56	1	57	97.85
of	151	8	582	77	818	93.86
tenacity	0	24	0	0	24	91.88
speed	0	61	4	2	67	87.65
chromatography	0	0	0	28	28	70.72
sea	0	44	0	0	44	63.63
too	28	0	0	0	28	62.13
relatively	30	0	0	1	31	55.25
performance	0	39	12	4	55	52.82
extremely	21	0	2	0	23	52.06
value	0	63	12	1	76	51.25
density	0	15	9	6	30	49.28
degree	0	35	0	0	35	45.92

Figure 15. Collocations: search results for the word *high*.

The search results can be downloaded to the user's workstation by clicking on the hyperlink **Download**. This saves them in the form of a delimited text file and they can then be loaded into spreadsheet software.

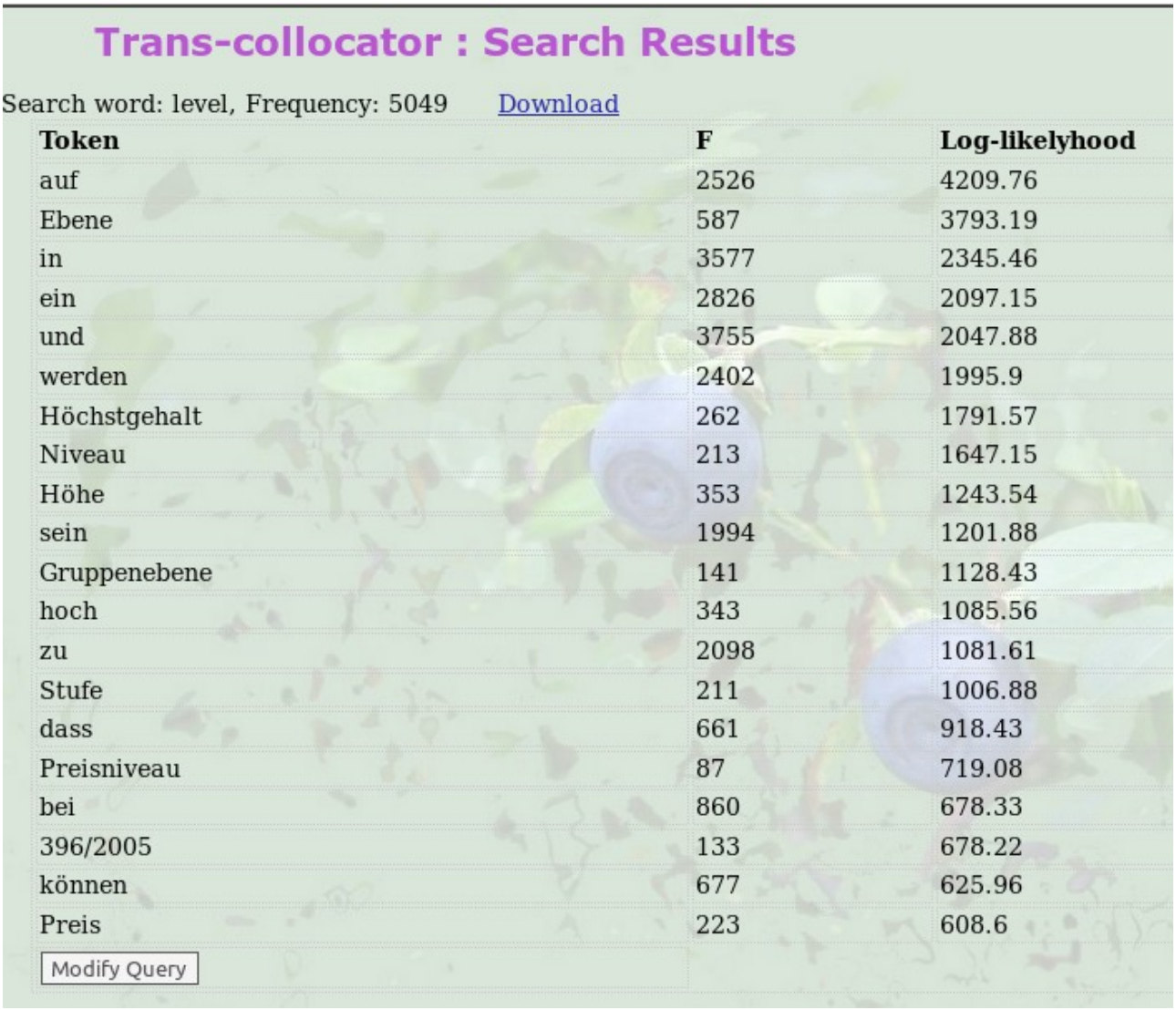
TRANS-COLLOCATOR

This is an experimental tool that searches for those items which occur frequently in the translations of the segments containing the search item. These might be translation equivalents or strong collocates of the search item. The shorter the segments in the parallel texts, the better the tool works.

The query interface for this tool (Figure 16) is quite different from that of the Collocator query. The user cannot define the span of the surrounding context: the tool operates with whole segments only. Nor is it necessary to define the desired frequencies: the application automatically uses the frequency of the search item as its starting point.

Figure 17 shows search results for the German trans-collocates of the English word *level* in the DGT-Acquis corpus. Among the words which are of little or no interest, the tool has found three possible German equivalents: *Ebene*, *Niveau*, *Gruppenebene*, *Preisniveau* as well as thematically connected words: *Höchstgehalt* ('maximum content'), *Höhe* ('height'), *Preis* ('price'). It should be added, that the tool only finds certain elements answering search criteria and arranges them according to certain statistic parameter. It is the researcher, who analyses and interprets the results.

Figure 16. Trans-collocator: the query interface.



Trans-collocator : Search Results

Search word: level, Frequency: 5049 Download

Token	F	Log-likelyhood
auf	2526	4209.76
Ebene	587	3793.19
in	3577	2345.46
ein	2826	2097.15
und	3755	2047.88
werden	2402	1995.9
Höchstgehalt	262	1791.57
Niveau	213	1647.15
Höhe	353	1243.54
sein	1994	1201.88
Gruppenebene	141	1128.43
hoch	343	1085.56
zu	2098	1081.61
Stufe	211	1006.88
dass	661	918.43
Preisniveau	87	719.08
bei	860	678.33
396/2005	133	678.22
können	677	625.96
Preis	223	608.6

Modify Query

Figure 17 Trans-collocator: search results for the English word *strong* in DGT-Acquis

***

The other tools included in the TextHammer package are all quite straightforward and do not require any special description or explanation; these display different kinds of corpus statistics. We wish to stress, however, that TextHammer is still being developed: the existing tools are constantly being improved and more new functionalities added.

Questions on using software, reports on bugs, and suggestions on program functionalities can be sent by e-mail to the address mikhail.mikhailov@staff.uta.fi.