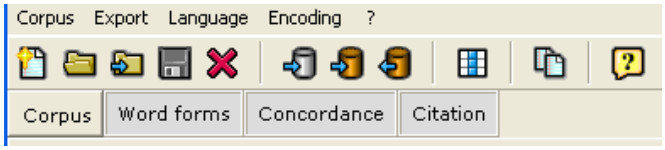
The Humanities Resource Center

TextSTAT Quickstart Guide to text analysis with TextSTAT

TextSTAT is a concordance program which was designed to be user friendly and provide simple Internet functionality. Texts can be combined to form corpora (which can also be stored as such). The program analyses these text corpora and displays **word frequency lists**, **concordances,** and **keywords in context** according to search terms. With TextSTAT you can search large amounts of text. You learn how often a certain word occurs or in what contexts it is used. Word combinations can also be examined.

Creating Your Own Corpora

When you open TextSTAT, you will see a window with a menu bar and several tabs. In the foreground is the tab sheet 'Corpus'. You can now add files and, in this way, put together a corpus. Put your mouse over the menu icons to learn what each one does.



Save Corpus / Open Corpus

You can save the opened files so that you can use them again as a corpus at a later stage (via the appropriate button and/or menu entry). You can decide the name of the file that is then created. We recommend storing the corpora in a separate folder.

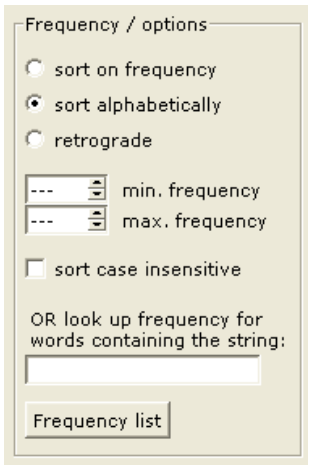

Add a file from the Internet to your corpus.


Add a text file from your computer (note that textSTAT cannot work with Microsoft Word files. These files would have to be saved first as .txt files)


Remove a file from your corpus.

Word Forms

After compiling a corpus from one or several files or after loading an existing corpus, you can obtain frequency information on the word forms contained in the corpus by clicking on the 'Word Forms' tab. Click on the 'Frequency list button' to generate a default word frequency list. Note that this does not convert any of the words to all lowercase, so the same word may appear twice in the list with the first letter of the word either uppercase or lowercase

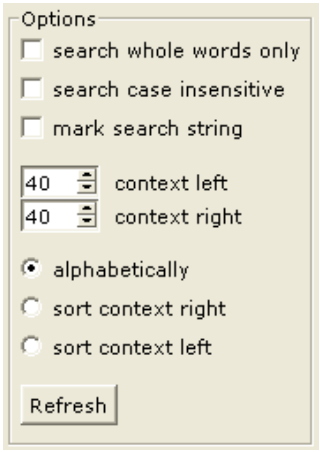


The options menu on the right hand side of the screen allows you to sort your word list in different ways. To convert all uppercase letters to lowercase, check off the **sort case insensitive** checkbox.

Retrograde sorts the words starting with the last letter of each word. You can also limit the frequency range to be displayed. Here you should take into account that '0' means no restrictions (therefore: if min.=0 and max.=0, all word forms will be displayed). After the display options have been changed, you will have to 'Update list'. If you double-click on a word form, then it will be searched for in the corpus and a concordance will be created.

Search / Concordance

The Search/Concordance tab shows a word form or a keyword in context. The terms found can be sorted according to different criteria, and the length of the context to be displayed can be determined. The search term is displayed in upper case by default. This marking can be deactivated.



When you enter a search string, it will be assumed by default that a word has been entered. This setting: search for 'whole words only' can be deactivated. A new search and/or a change in the display options can be activated with the button 'Search/Update'.When searching, you can use regular expressions (see below). If you double-click on a line of text, this will be searched for in the corpus and the citation (a text passage with more context) will be displayed.

Citation

The Citation tab will display a text passage in which the sought string will be shown in more context. Moreover, the name of the file from which the passage is taken, will also be displayed. The position (in characters) of the passage in the original file will be given in brackets.

A double-click on the file name opens the original file with the program that is linked with the file extension. In the case of websites, you are connected with the Internet and see the original file displayed in the browser.

Regular Expressions

When defining the search term (in 'Search/Concordance'), you can use so-called 'regular expressions'. While these are not particularly user friendly, they are extremely powerful in executing very precise search queries.

Important special characters used in regular expressions:	**Examples:**
'.' (the dot) stands for any character you like **'\w'** stands for any alphanumeric character **'\W'** stands for any non-alphanumeric character (e.g. space, punctuation marks) **'+'** the preceding character is repeated once or any number of times **'*'** the preceding character is repeated any number of times, including zero **'*?', '+?'** make sure that '*' and '+' are not 'greedy' (see examples) **'|'** stands for or **'[ ]'** square brackets define a set of characters which are searched for alternatively.	**b\wr** finds 'but', 'bit', 'bet' and 'bat' **b\w+r** finds 'but', 'bit', 'bet', 'bat', 'boat' and 'built' **w[ao]nder** finds 'wander' and 'wonder' **(this|that)** finds 'this' or 'that' **so.+e** finds the string 'sold me her house' in the text: 'My sister sold me her house' **so.+?e** finds the string 'sold me' in the text: 'My sister sold me her house' **s.+r** finds the string 'sister sold me her' in the text: 'My sister sold me her house' **s\w+r** finds the string 'sister' in the text: 'My sister sold me her house'