

WordSmith Tools, version 4 - A short user guide

More information about the program, including step-by-step guide with screenshots can be found at <http://www.lexically.net/wordsmith/>

The help files in the program are unusually helpful and easy to use.

WordSmith Tools has many advanced functions that you should explore further when you have learned the basics. There is often more than one way to do something – try the different ways until you find what suits you best.

Before you start

To use the program you need (at least) one text. The program only reads plain text, so files called .txt, .htm(l), .sgm(l), and .xml are ok. *Some sample texts are available for the course.*

The program can deal with texts in many languages and you should set its preferences to the language(s) you work with. *UK English by default.* You can also change the settings to exclude markup or only search parts of a file (see the WsT help files for more information).

Open the program by clicking on the WordSmith Tools icon (or use the Start menu).

Concordance

The Concordance tool will allow you to search for a word/phrase and display it as concordance lines. The tool has many other functions, not only allowing you to sort, delete and mark lines but also to find multi-word units, see a graphic display of where in a text the search word is found and generate a list of collocations.

1. Click on the ‘Concord’ button in the WordSmith Tools main screen. *A new window ‘Concord’ opens*
2. In the ‘File’ menu, select ‘New’ (or use Ctrl+ N). *The ‘Getting Started’ window opens.*
3. Select texts:
 - a. Click on the ‘Texts’ tab. *You will see what texts (if any) are already selected.*
 - b. If the selected text is what you want, move on to Define Search Word (4) below.
 - c. If the selected text(s) are not what you want, click ‘Change Selection’. *The ‘Choose Texts’ box opens.*
 - d. Browse the files in the left hand window. Select the ones you want to use by highlighting them and clicking on the blue arrow. *The text appears in the table to the right.* To remove texts from the selection, click on them in the table and then on the delete button on your keyboard.
 - e. Close the ‘Choose Texts’ window to return to the ‘Getting started’ box.
4. Define search word
 - a. Click on the ‘Search Word’ tab.
 - b. Type your search string into the upper box. *You can find instructions on how to formulate your search below the search box.*

- c. Click 'OK'. *The concordance window opens and your hits are listed. You may be asked if you want to sort the lines before display. This is optional.*

Examining the concordance lines

More context

*Either 'drag' on the dividing lines between the rows or columns (best for expanding one line)
or use the 'View' menu and select 'Grow'
or click F8
or double-click on the concordance line to see the whole source text.*

Sorting

1. Use the Edit menu -> 'Resort' *or click F6. The Concordance Sort window opens.*
2. Select your Main Sort, and optionally Sort 2 and Sort 3. *You can choose whether to have ascending or descending alphabetical order. NB: The sort is only effective if the 'Activated' box is ticked.*

Exploring further

Use the tabs at the bottom of the Concordance screen (see the WordSmith manual for more information about these options).

WordList

The WordList tool will allow you to create word lists ordered alphabetically or by frequency. It will also generate statistics about the text(s), such as number of words, type-token ratio and word length. You can choose to group items into lemmas via the word list.

1. In the WordSmith Tools main screen, click on the 'WordList' button. *A new window 'WordList' opens*
2. In the 'File' menu, select 'New' (or use Ctrl+ N). *The 'Getting Started' window opens.*
3. Click on the 'Main' tab. *If you have any texts selected, these are listed.*
4. To select texts, click 'Choose texts now'. To change the current selection, click on 'Change Selection'. *The 'Choose texts' window opens.* Browse the files in the left hand window. Select the ones you want to use by high-lighting them and clicking on the blue arrow. *The text appears in the table to the right.* To remove texts from the selection, click on them in the table and then on the delete button on your keyboard. Close the 'Choose Texts' window to return to the 'Getting started' box.
5. Click 'Make a word list now' to create the word list.
6. Use the tab along the bottom of the screen to choose a different list (frequency, alphabetical or general statistics)
7. To save the list, use the 'File' menu -> 'Save' option and define where to store the file.

Key Words

The KeyWords tool will compare two (or more) word lists and identify the words that are more frequent (in relation to text size). You can compare lists from two texts or one list from a text to a list from a (reference) corpus.

0. Before you can start you need (at least) two wordlists: one from the text/corpus you are working with and one from a reference corpus (the 'reference corpus' list can be a list generated from another text, from a number of texts, or from a corpus).
1. Click on the 'KeyWords' button in the WordSmith Tools main screen. *A new window 'KeyWords' opens*
2. In the 'File' menu, select 'New' (or use Ctrl+ N). *The 'Getting Started' window opens.*
3. Select the 'Key words' tab. Click on the 'open file' icon to the right of the upper empty box and select the file(s) you want to work with (=the list(s) which you want to compare to the reference list). Then select your reference corpus list in the lower empty box and click
4. Click on 'Make a keyword list now' to create the keyword list.
5. Sort the list by clicking on the headings above each column.
6. Use the tabs along the bottom of the KeyWords screen to display a plot (graphically displaying where in the file the words occur), links or clusters (key words appearing close to other key words) or to see the whole text (see the WordSmith manual for more information about these options).

Suggested exercises

Aim: Explore the (basic) functions in WordSmith Tools.

Pride and Prejudice is on many reading lists. Use the following exercises to explore the novel using WordSmith Tools. Some suggestions of how to solve the tasks are given in the 'Tips and hints' section below. The tasks are grouped under Concordance, WordList and KeyWords but can be approached from different angles. Feel free to explore alternative approaches and to switch between the tools.

Concordance

1. Search in P&P for the words *pride* and *proud*. Look at the concordance lines and see if any patterns strike you at first glance.
2. How many instances are there? Which word is more frequent – *pride* or *proud*?
3. Sort the concordances on the words preceding the search word. What patterns can you see now? Sort on the words following. Any new patterns?
4. Where in the text do you find *pride/proud*? Are the words more frequent in the beginning or end of the novel?
5. What words are used with *pride/proud*? Where are they found (before or after the word? Immediately preceding/following?)
6. Is *pride/proud* used in 'bundles' with the same items before or after? Can you find any with more than three words?
7. How often is *pride/proud* used as a positive or negative feature? Which is more frequent? Alternatively - whom/what is referred to by *pride/proud*?
8. *And* is often used with *pride/proud* but not only immediately preceding or following the search term. Can you find all instances? What kind of words are co-ordinated with *pride/proud*? Does the co-ordination emphasise any particular aspect of *pride/proud*?
9. Look only at instances that are co-ordinated or used with an attribute (for example *pride and vanity* and *brotherly pride*). Is the author using this feature in any particular way?

Tips and hints

1. To search for two words, use a slash (/) between the words in the search box and you can search for both at the same time.
2. The search should find 73 instances of the two words (bottom left of screen), 51 of which are *pride*. Just glancing at the list may be enough to see that *pride* is more frequent but you can also use the sort function. (Press F6 and select *centre* in the *Main sort* tab. The first instance of *proud* is number 52 in the list.) If you find more or less than 73 instances, check the *settings*. (You get different results depending on whether you, for example, exclude the header bit – see the *Tags* tab.)
3. Use the 'Sort' function (the F6 key) and set the main sort to L1 to sort by the word immediately preceding *pride/proud* (you can use the Sort 2 and Sort 3 options as well if you like). You will see that there are many instances of *his pride* for example. Among the words following (Main sort on R1) are *and* and *of*.

4. Use the 'Plot' tab to see a graphic representation of how the occurrences are distributed across the novel.
5. Use the 'Collocates' function (tab at bottom of screen). The most frequent collocates are *and*, *his*, *of*. By examining the table, you can see that *and* follows the search word more often than precedes it, while *his* and *of* are more frequent before the *pride/proud*. You can also see that there are instances of *pride* found close to *pride*. These are in sentences like: *But his pride, his abominable pride, his shameless avowal of what he had done...*
6. Use the 'Cluster' tab. Select Compute -> Clusters and change the settings to explore different options (change minimum frequency to 2 or 3). There are some with five words, but they only occur twice each.
7. You may find it useful to mark the lines that have a particular meaning and then sort on this mark-up. Do this by clicking on the line you are analysing and just type a letter (P for positive, N for negative etc). Then include 'Set' in your sorting.
8. You can search for *pride/proud* and *and* by using the 'Advanced search' tab on the search screen. Define the span (how many words on either side) to include (for example L5, R5).
9. Go through the concordance lines (you may find it useful to sort them in different ways to find the instances more quickly). When you find one that is irrelevant, click on it and then press the delete key. The line is marked. To delete all marked instances, press Ctrl+Z or use the Edit->Zap menu. You can un-mark an un-zapped line by clicking on it and pressing the Insert key.

WordLists

1. Create a word list based on *Pride and Prejudice* and save it.
2. What is the most frequent word? The most frequent words are function words/words from closed word classes. What is the most frequent content word? How often does that occur?
3. How many times does *Darcy* occur? Is that more or less than *Elizabeth*?
4. How many words are there in the novel. How many different ones?
5. It has been suggested that about half the words in any text occur only once. Is that true for this text? What does this mean?
6. Can you find any 'odd' items in the list? (try looking for words beginning with X) Can you find an explanation? (How/where are they used?) Solution?
7. Create other word lists and compare these to the *Pride and Prejudice* one. For example from one or all of the other Austen novels, other Austen texts (letters, unfinished work), other contemporary writing, a modern novel, a modern or contemporary corpus.

Tips and hints

1. Use the WordList function and choose 'Save'.
2. Look at the Frequency list (tab at bottom of screen). The most frequent word is *the*.

3. Scroll down the alphabetical list until you find the relevant instances, or start typing the word (the program jumps to the place in the list where that letter combination is). You could also make a concordance search for the words.
4. The 'Statistics' tab will provide information about the text – size, number of running words (tokens), number of different words (types) and much more.
5. In this text less than half the words occur only once. This could suggest that the author is less flexible, the text easier to read since there are fewer new words or something else. Impossible to say though without further investigation. Also remember that the list is not lemmatised.
6. One word beginning with X is Xmetal. A search for xmetal (in WordList window, select Compute ->Concordance) shows that it is mentioned in the text header (presentation of the text preceding the text itself). You can change the settings in WordSmith to ignore this part of the text (in the WsT start window, choose Settings ->Adjust settings -> Tags. Add relevant item in the box by 'Document header ends'). You need to look in the text to know where the header ends (double-click on any concordance line and scroll to the top). Here: Add 1796-1812 in the box by 'Document header ends'.

KeyWords

0. If you have not already done so, create and save one word list from *Pride and Prejudice* and one from the other five Austen novels (emma, mansfield, nothanger, persuasion, sense). We'll refer to these as 'pride' and 'austen5'.
1. What words do you think are used proportionately more in *Pride and Prejudice* than in Austen's other novels? Which are used proportionately less? Guess and compare with your keyword list.
2. Where in the text are the key words found? Can you relate this to themes in the novel? Is Brighton talked about throughout? When does Fitzwilliam appear? Mr Collins? Charlotte?
3. What does 'keyness' mean? What are negative keywords? What use are they?
4. Explore the links and clusters. Use the Concord function to see the instances in context.
5. Generate keywords for other texts. What difference does it make if you use the same reference corpus (austen5), create a new corpus of five Austen novels (including *Pride and Prejudice*, excluding the novel you are comparing to), or use all six novels. What is best?

Tips and hints

1. Create a keyword list, using austen5 as your reference corpus. The items at the top of the list are all proper nouns. The items at the bottom (used proportionately less in *Pride* than in *Austen5*) include proper nouns but also some other, probably unexpected items such as *quite, it, very*.
2. Use the 'Plot' tab to display a graphic representation of the distribution (you can change the display via the 'View' option. Make the 'plot' area as wide as possible for best display). Try sorting the list in different ways (click on the list heading). Look at Fitzwilliam and compare it to Rosings (he turns up during the visit to Rosings), see that

Brighton is used intensively for a short(ish) time, Charlotte features most in the second quarter of the work, Bingly is not used at the same time as Brighton, etc.

3. 'Keyness' is a measure that compares the *relative* frequencies of a word in two texts. A word that is frequent on one text and rare in the other gets a high keyness value.
4. Both links and clusters show how key words are used close to each other. 'Links' finds other keywords within a certain distance (which you can define). Clusters are combinations of keywords occurring in a set pattern, forming phrases with gaps (where non-key words are found). Good information in the WsT help files/manual.
5. If you include the text you are looking at in your reference corpus as well, the frequency difference between the text and corpus will be less and you will get a lower keyness value. If you use the fist austen5 'corpus' (all novels but Pride) and compare it to – for example – Persuasion, the error will be even greater since not only are you including your text in the reference corpus but you are also missing one novel (Pride). It could possibly be ok as an experiment where you want to just try out the functions and compare the results (such as in this workshop) but is not suitable for 'real' work.