



Tutorial and Exercises with KeyWords in WordSmith Tools: Level I

WordSmith Tools, developed by Mike Scott, is a corpus analysis tool that includes three text analysis tools: a monolingual concordancer (**CONCORD**) and wordlist extractors (**WORDLIST** and **KEYWORDS**). This tutorial will focus on the basic features of **KEYWORDS**. In CERTT you can also find a basic and an intermediate tutorial on [WORDLIST](#), as well as on [CONCORD](#).

For more information on WordSmith, visit www.lexically.net and/or read WordList's help files, which you can find on the Writing Center computers under V:\WordSmith4 Help Files or [on-line](#) (pages 132–167). From the <http://www.lexically.net> site, you can also download a demo version of WordSmith Tools to install at home; this version gives you access to all of the functions of the tool, but limits the number of results displayed to 25.

WordSmith Tools can process .html, .xml and .txt files. KeyWords is a program that identifies the "key" words in a text, i.e., the words that are unusually frequent. This type of information can be used to study a genre, a language for special purposes, a writer's idiolect, etc.

In order to establish what frequencies are common and what are especially high, the program needs a reference corpus that will establish the norm against which word frequencies will be compared. It is important to know that KeyWords compares word lists and not raw texts.

The exercises in this tutorial will cover how to:

- [generate a key word list](#),
- [display the results](#),
- [manipulate the results](#) and
- [choose the right reference corpus](#).

Getting Ready

Before starting to work with KeyWords, you will need to prepare a series of files. First of all you will have to download the KeyWords Resource Package; then you will learn how to select the texts you want to work with in WordSmith and finally you will generate the word lists that KeyWords will work with.

Downloading the Resource Package and Starting WordSmith

1. Create a sub-directory on the U: drive (aka *My Documents*). (For instructions on how to do this, see the document [Creating a sub-directory](#) (**TUTORIALS AND EXERCISES IN ENGLISH > WINDOWS XP**)).



2. Download to this sub-directory the compressed folder *KeyWords Resource Package* from the CERTT site (**FILES TO DOWNLOAD > KEYWORDS RESOURCE PACKAGE**).
3. Extract the files from this compressed folder to the sub-directory you just created. (For instructions on how to do this, see the document [Extracting files from a compressed folder](#) (**TUTORIALS AND EXERCISES IN ENGLISH > WINDOWS XP**)).
4. Open the files in the KeyWords Resource Package and read them.
5. Start WordSmith Tools (**START > PROGRAMS > TRADUCTION > WORDSMITH 4 > WORDSMITH.EXE**).
6. When WordSmith asks whether you want to only see the basic functions, click on **NO** so that you have access to all the functions of the software.
7. The software opens its main window, which gives you access to its tools and options. Each tool can be started by its respective button (**CONCORD, KEYWORDS, WORDLIST**) and the program options can be found in the **UTILITIES** menu.

Choosing the Texts

The first step to start working with WordSmith Tools is to select the texts that you want to analyze. You can use texts of your own or if you do not have any at hand you can use the files you downloaded in the previous section.


NOTE: If this is your first time using **KeyWords**, we highly recommend that you use the sample texts.

1. In the main WordSmith Tools window, go to the **FILE** menu and click on **CHOOSE TEXTS**
2. A *Choose texts* window appears divided in two sections. On the right you can see the files that will be analyzed and on the left, you have your computer root tree.
3. WordSmith comes with a demo file (a chapter of *A Tale of Two Cities*). To remove it, select it and press **DEL/SUPPR.** or right click and select **DELETE** in the contextual menu.
4. In the drop-down list that appears at the top left-hand corner of the dialogue box, select the drive where you saved the documents that you want to analyze (a sub-directory of the U:\ drive or *My Documents*).
5. In the dropdown list immediately to the right of the drive drop-down list, you can select the format of the files you want to have displayed.


NOTE: By default the software shows all files in all formats (*.*). If you want to only display plain text files, select (*.txt); for webpages (*.htm, *.html) and for .xml documents (*.xml).

6. On the left side, under *Files Available*, browse the drive you selected to find the files you want to work with. (To open a sub-directory, double-click on it).
7. Once you have located the files, select the one you want to work with: *Environment_EN.txt*.

NOTE: Do not open all of the files at the same time. Start by selecting only the file named *Environment_EN.txt*.

8. Click on the long vertical button with the blue arrow between the two blocks, *Files available* and *Files selected*, or drag and drop the file to the *Files selected* block on the right-hand side. The file appears in the *Files selected* block.
9. Once you selected the file you will need, close the window by validating the selection (click on the green tick ).

Generating a WordList



1. In WordSmith main window, click on the **WORDLIST** button. The **WORDLIST** window appears together with the *Getting started* dialogue box.
2. In the *Getting Started* dialogue box, click on the **MAKE A WORD LIST NOW** button. The word list appears in the **WORDLIST** window.
3. Save this word list by opening the menu **FILE** in the word list window and selecting the *Save* option (). You will then have to select a sub-directory to store the file and give it a name. Remember to save it on the U: (*My Documents*) if you are at the Writing Centre.
4. You now have created a word list for the text *Environment_EN.txt*.

Generating WordLists for comparison

1. In order to carry out the tutorial we still need two other wordlists. Repeat all the above steps in the [Choosing the Texts](#) and [Generating a Wordlist](#) sections to generate a wordlist of all the texts you will find in the folder *Reference Corpus 1* and do it all over again for the texts in the folder *Reference Corpus 2*. Remember to save each word list.

WARNING: Make sure to remove the *Environment_EN.txt* file from the *Files Selected* on the right. Otherwise your new wordlist will contain this text also. To remove the file you only need to select it and press on the key **DEL./SUPPR.** on your keyboard.

NOTE: To add multiple files to the *Files selected* list when you create these lists, in the *Files available* list, select the first file, then press the **SHIFT** key and hold it down while you select the last file.

If you think you will work often with these files, you can save the selection by clicking on the **SAVE FAVOURITES** button () and saving the list on the U: drive or *My Documents*. (Remember that all files saved in any other drive at the Writing Centre will be deleted once you log off your computer.) To retrieve the selection next time you use WordSmith, click on **GET FAVOURITES** (.

2. At the end of this section you should have created and saved three word lists.

Evaluating the WordList

1. Look carefully at the word list you generated from the *Environment_EN.txt* text.
 - a. What words have surfaced in the word list?

- b. If we set aside articles, prepositions and other grammatical words, what words (or terms) are the most frequent in the text?
- c. Do you consider that these are the most distinctive words in the text?
- d. If you had to choose the key words in the text, would you have selected these?

Generating a Keyword List

1. On the main *WordSmith Tools* window, open the KeyWords tool by clicking on the **KEYWORDS** button.
2. The *KeyWords* window opens. Open the **FILE** menu of this window and select the **NEW...** option (🟢).
3. A dialogue box appears and requests you to load two wordlists. Click on the **BROWSE** button (📁) of the first text field and locate on your computer (probably in a sub-directory of the U: drive) the word list you generated from the text *Environment_EN.txt*. Once you have located it, double-click on it. The location of this file will now appear in the field.
4. Click on the **BROWSE** button (📁) of the field below and locate on your computer (probably in a sub-directory of the U: drive) the word list you generated for the texts in the folder *Reference Corpus 1*. Once you have located it, double-click on it. The location of this file will now appear in the field.

WARNING: Make sure that you enter the word list of the text you want to analyze in the first field and the reference corpus in the one below. Otherwise, the list generated by the program will be useless.

5. Once both word lists are located, generate the keyword list by clicking on the button **MAKE A KEYWORD LIST NOW**.
6. The program automatically generates a keyword list by comparing the frequency of the words in each list and extracting the words that are proportionately more frequent in the text to be analyzed than in the reference corpus. You can see this list appear in the **KEYWORDS** window.
 - a. Observe this keyword list. Is it different that the word list you generated for the *Environment_EN.txt* text?
 - b. Is the list just as long or is it shorter?
 - c. How do the words in this list differ from the words in the initial word list?
 - d. Do you agree in that they are key words of this text?

Displaying the Results

In the previous section we generated a keyword list. Let's have a closer look at the information that is displayed in the *KeyWords* window.

1. Look at the bottom of the *KeyWords* window and note that it is divided into seven tabs: *KWs*, *plot*, *links*, *clusters*, *filenames*, *notes*, *source text*. To shift from tab to tab you have only to click on its name.

2. Click on the **KWS** tab and observe its content. The information on it is divided into seven columns.

Column	Description
Keyword	list of keywords
Freq.	number of occurrences of each keyword in the source text(s) in which these key words are key
%	percentage value of the frequency of the keyword in the source text
R.C. Freq.	number of occurrences of each keyword in the reference corpus (R.C)
R.C. %	percentage value of the frequency of the keyword in the reference corpus
Keyness	statistical calculation that factors in the frequency of a word in each wordlist and limits it with the probability value (p)
P	value used in statistics to indicate the probability of obtaining a wrong results; a high p value will imply high chances of that word not being a key word

3. Click on the **PLOT** tab and observe its contents. The information on it is divided into six columns.

Column	Description
Keyword	list of keywords
Dispersion	statistical calculation to assess whether a keyword appears evenly throughout the whole text or is highly concentrated in very specific sections. It ranges from 0 to 1 where 0 is a very uneven use of the keyword while 1 is a perfectly even presence of the keyword along the whole text.
Keyness	statistical calculation that factors in the frequency of a word in each wordlist and limits it with the probability value (p)
Links	Links are co-occurrences of key-words within sets of 5 words
Hits	number of occurrences of each keyword in the source text(s) in which these key words are key
Plot	graphical representation of where the occurrences of each word appear in the text

The information displayed on this tab can be very useful when studying a genre or a type of text.

4. Click on the **LINKS** tab and observe its contents. This *Links* window shows the number of links followed by a column headed "in" and a percentage. This percentage represents the number of links divided by the total number of occurrences of the word in question (the "in" column number).
5. Click on the **CLUSTERS** tab and observe its contents. This tab shows keywords that appear close to each other or together in the text. If the two keywords are

separated by brackets with one or more dots inside, this means that the words do not appear side by side.

6. Click on the **FILENAMES** tab and observe its contents. This tab shows the name of the file and its location on your computer.
7. Click on the **NOTES** tab and observe its contents. In this tab you can write notes about the keyword list.
8. Click on the **SOURCE TEXT** tab and observe its contents. This tab displays the full source text.

Manipulating the results

In the previous section we generated a raw keyword list. The list contained all the words that appeared to be unusually frequent in the analyzed text when compared to a reference corpus. This section presents how to eliminate from the results those words that do not interest us for our research, how to generate a concordance from a keyword and the different options available to sort the keywords list.

Cleaning up the KeyWords list

Clean the keyword list of any words that you do not deem key. Statistical calculations do give impressively accurate results but they are not error-proof. You may want to eliminate certain words from the list.

1. Make sure you are on the *KWs* tab in the **KEYWORDS** window. If you are not, you can go back to it by just clicking on the **KWS** tab.
2. To eliminate an entry you need only to select it by clicking on it, and then press the **DEL/SUPPR** key on your keyboard. The selected word turns grey and has been stroked through.
3. Browse down the list with the arrow keys on your keyboard and delete all the words you do not want to appear in your word list. To do so, repeat step 2 as many times as needed.

NOTE: If at any time you delete a word by mistake or you change your mind on a deleted word, you can re-insert it to the word list by selecting it and pressing the **INS** key on your keyboard.

4. Once you have gone through your entire list (if you are analyzing a short text) or through the most key keywords of the list (if you are working with a corpus or a very long text), and are sure you want to remove the deleted words **permanently**, go to the **EDIT** menu and select the **ZAP** option (🚧). Zapping will cut out the deleted words and reorganize the word list based on the frequency of the remaining words.

WARNING: Zapping will eliminate the previously deleted words from your keywords list permanently. There is no undo option.

NOTE: This process can also be automated by means of a stop-list. To learn more about it, read the [WordSmith WordList Tutorial, Level II](#).

Generating a Concordance

Study a word to decide whether it is worth keeping. Some words are easily identified as keywords while others fall into a rather grey zone. The context in which they appear can help us decide on their nature. To verify a word's context:

1. Make sure you are on the *KWs* tab in the **KEYWORDS** window. If you are not, you can go back to it by just clicking on the **KWS** tab.
2. Select the word you want to generate a concordance with by clicking on it.
3. Open the menu **COMPUTE** and select the option **CONCORDANCE** (C). The software automatically launches a simple search for this word in its *Concord* tool. For more information on how to use *Concord*, read the tutorial [WordSmith Concord Level I](#).

Sorting the Results

1. By clicking on the button on top of each column the program re-sorts the results in increasing or decreasing order of the value in that column. For example, if we click on the **KEYWORD** button, the program will re-sort the results alphabetically from A-Z or from Z-A alternatively. If we click on **KEYNESS**, the program will re-sort the results from highest to lowest Keyness or vice-versa.
2. We can also re-sort the keyword list from the **EDIT** menu. The options here are to re-sort the results alphabetically, alphabetically by word ending or by word length.
 - a. Make sure you are on the **KWS** tab in the **KEYWORDS** window. If you are not, you can go back to it just by clicking on the **KWS** tab.
 - i. Invert the alphabetical order, usually A-Z, into Z-A by opening the **EDIT** menu and selecting the option **RESORT** (R).
 - b. Resort the word list alphabetically by word ending, by opening the **EDIT** menu and selecting the option **OTHER SORTS** and in the new menu, selecting the option **REVERSE WORD** (RW).
 - c. Resort the word list by word length, by opening the **EDIT** menu, selecting the option **OTHER SORTS** and, in the new menu, selecting the option **WORD LENGTH** (WL).

Choosing the Right Reference Corpus

KeyWords calculates each word's keyness in function of a reference corpus. Therefore, what this corpus contains has a direct effect on the keyword list that is presented to us.

The KeyWords Resource Package you downloaded at the beginning of the exercise contains a text on the environmental effects of wind turbines and two reference corpora. The *Reference Corpus 1*, which we have been using until now, contains a series of texts on income taxes and capital gains taxes, while the *Reference Corpus 2* consists of a series of texts on wind power. In this second corpus, words relating to wind turbines and wind power should be much more frequent than in the first one, and therefore the keyword list should differ greatly.

1. Generate a keyword list with the *Environment_EN.txt* text word list and the *Reference Corpus 2* word list. To do so repeat the steps in the section [Generating a Keyword List](#) substituting the *Reference Corpus 1* word list with the *Reference Corpus 2* word list.
 - a. How do the results differ? Has the program retrieved the same keywords?
 - b. Do keywords that appear in both keyword lists have the same keyness value? And p value?

Wrapping up

1. To make a copy of your files as a backup or to transfer them to another computer outside the Writing Centre:
 - c. In **MY COMPUTER** or from the **START** menu, find the sub-directory you created to store the files for this exercise.
 - d. Make a compressed folder that contains this sub-directory. (For instructions, see the file [Create a compressed folder.](#))
 - e. Copy this compressed folder to a USB key or diskette, or if it is less than 2 MB, send a copy as an attachment to your e-mail.

Questions for reflection

1. After having used KeyWords, what uses can you think of for this tool?
2. How do KeyWords results differ from those obtained with WordList?
3. Compared to other corpus analysis tools or term extractors, what are some advantages and disadvantages of KeyWords?
4. Based on what you have seen in the *Choosing the Right Corpus* section, how can the results of KeyWords be refined?