Tutorial and Exercises with WordList in WordSmith Tools: Level I

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

For more information on WordSmith, visit www.lexically.net and/or read WordList's help files, which you can find under V:\WordSmith4 Help Files on the Writing Center computers or online (pages 132-167). From the 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. WordList is a tool that analyzes texts or groups of texts (corpora) to extract statistics about the words they contain. The results of this analysis are presented in word lists (alphabetical lists or lists ordered by frequency of occurrence) and tables. Such information can be used for multiple purposes: studying the lexical characteristics of text types and genres, tracking changes in lexical usage over time, identifying plagiarism and extracting terminology. This tutorial will pay special attention to this last application and will show how WordList can be used as a term extractor.

The exercises in this tutorial will show you how to:

- Choose texts to analyze,
- Generate a word list,
- View the results,
- Generate an index,
- Generate word-cluster lists.

Getting Ready

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*).
2. From the CERTT WebCT site, download to this sub-directory the file *Wind Power Manual EN* (**CORPORA AND OTHER RESOURCES > TRILINGUAL RESOURCES [EN-FR-SP] > TECHNICAL TEXTS [EN-FR-SP] > WIND POWER [EN-FR-SP]> WIND POWER MANUAL EN**).
3. Extract the files from the compressed folder to the sub-directory you just created. (For instructions on how to do this, read *Extracting files from a compressed folder*.
4. Start WordSmith Tools (**START > PROGRAMS > TRADUCTION > WORDSMITH 4 > WORDSMITH.EXE**).

5. 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.
6. The software opens and displays its main window, which gives you access to its tools and options. Each tool can be started by clicking its button (**CONCORD, KEYWORDS,** or **WORDLIST**). Program options can be found in the **UTILITIES** menu.

Choosing the texts

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

1. Select one or more files to work with:
   a. In the *WordSmith Tools* main window, go to the **FILE** menu and click on **CHOOSE TEXTS.**
   b. A *Choose texts* window appears, divided in two sections. On the right-hand side you can see the files that will be analyzed and on the left, you have your computer drive and file structure.
   c. WordSmith comes with a demo file (a chapter of *A Tale of Two Cities*). To remove it, select it and press the **DEL/SUPPR** key on the keyboard, or right click and select **DELETE** in the contextual menu that appears.
   d. In the drop-down list at the top left of the *Choose texts* window, select the drive where you saved the documents that you want to analyze (U:\ drive or *My Documents*).
   e. In the drop-down list immediately to the right of the drop-down list of drives, 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 display only plain text files, select (*.txt), for webpages (*.htm, *html) and for .xml documents (*.xml).

   f. On the left-hand 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).
   g. Once you located the files, select them all. (To do so, select the first one, and hold down the **SHIFT** key on the keyboard and while you select the last file).
   h. Click on the long, narrow vertical button with the two arrows to the right between the two panes, *Files available* and *Files selected*. The files appear on the right block.

NOTE: If you will work often with these files, you can save the selection by clicking the SAVE FAVOURITES button ( ) and saving the list to your U: drive (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 from your U: drive next time you use WordSmith, click the GET FAVOURITES button ( ).

   i. Once you have selected all the files you will need, close the window by validating the selection (click the green checkmark).

Generating a WordList

1. In WordSmith main window, click the **WORDLIST** button. The *WordList* window appears in front of the *Getting Started* dialogue box.

2. In the *Getting Started* dialogue box, click the **MAKE A WORD LIST NOW** button. The word list appears in the *WordList* window.

3. Observe the information that it is presented to you. At the bottom of the window you can see that it is displayed on 5 tabs.

   a. *Frequency* tab

      i. The information in this tab is divided into 5 columns

WORD	List of the words that appear in the analyzed texts ordered by decreasing frequency of occurrence.
FREQ.	Number of times each word occurs in the set of texts.
%	Percentage of the set of texts made up by the each word.
TEXTS	Number of texts that contain that word.
%	Percentage value of the number of texts that contain that word.

      ii. What are the most common words? Are there many terms among the highest-ranking words?

      iii. Do all of the entries in the list correspond to one and only one word or meaning? Does every word that appears in the texts correspond to one and only one entry in the list? What does this tell you about how WordList works?

      iv. Can you think of any strategy to filter the terms?

      v. Is the information found under the *Texts* column relevant for a terminologist? Why or why not?

   b. *Alphabetical* tab

      i. This tab shows the same information as the *Frequency* tab except that the words are shown in alphabetical order (A-Z).

      ii. What is the advantage of studying the word list in this order? What types of words appear together?

   c. *Statistics* tab

      i. This tab shows a series of data on the total of the texts and on each file. The names of the rows are self-explanatory.

NOTE: Each occurrence of word in a file is called a *token*, while each different word is referred to as a *type*. For example, if we analyzed a file that contained only the three words "bridge, bridge, bridge", the file would contain 3 tokens and 1 type.

   d. *Filenames* tab

      i. This tab lists the names of the files analyzed.

   e. *Notes* tab

      i. This tab allows you to make notes about the wordlist.

      ii. When can it be useful to analyze a text with a word list? Would a human do a better job at extracting terms? How might a frequency word list help or hinder the terminologist?

4. If you want to use this word list in the future, you can save it by opening the **FILE** menu in the *WordList* 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.

NOTE: When you choose the SAVE option, WordSmith saves the word list in its own proprietary format (.lst), which means you will only be able to read the file with WordSmith. You can also export the word list to other formats such as plain text, xml text or Excel spreadsheet. To do so, select *Save As* ( ) from the WordList FILE menu.

Viewing the results

In the previous section we generated a "raw" word list. The list contained all the words that appeared in the texts, and they were shown according to the default parameters. If we are looking for terminological units, we might be interested in eliminating grammatical words from the list in order to let more terms emerge, or in re-sorting the list according to other criteria. In this section we will learn how to do so.

1. Clean the word list of grammatical words. Grammatical words (articles, prepositions, auxiliary verbs, conjunctions…), due to their linking role in language, are very frequent in texts and always appear at the top of any word list.
    a. Make sure you are on the *Frequency* tab in the *Word* **LIST** window. If you are not, you can go back to it by clicking on the *Frequency* tab.
    b. The first word of the list appears selected in blue. In almost any list, this word will be the article *the*, as this is the most frequent word in English. To eliminate it from your list, just press the **DEL/SUPPR** key on your keyboard. The word *the* turns grey and has been stroked through.
    c. 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 *b* as many times as needed. If at any time you delete a word by mistake or you change your mind about deleting a word, you can re-insert it in the word list by selecting it and pressing the **INS** key on your keyboard.
    d. Once you have gone through your entire list (if you are analyzing a short text) or through the most frequent words 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 remove the deleted words and reorganize the word list based on the frequency of the remaining words.

NOTE: This process can be automated by means of a stoplist. To learn more about this process, see the WordSmith WordList Tutorial, Level II.

2. Study a word to decide whether it is worth keeping. Some words are easily identified as grammatical words that do not constitute a term. 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:
    a. Open the **COMPUTE** menu and select the **CONCORDANCE** option ( ). The software automatically launches a simple search for this word in its Concord tool. For more information on how to use Concord, see the WordSmith Concord Tutorial, Level I.
3. Re-sort the word list alphabetically, alphabetically by word ending, by word length or by consistency (presence across texts).
    a. Make sure you are on the *Alphabetical* tab in the *WordList* window. If you are not, you can go back to it by clicking on the *Alphabetical* tab.

b. Invert the alphabetical order, usually A-Z, into Z-A.

   i. Either click the **WORD** button on top of the words column, or open the **EDIT** menu and select the option **RESORT** ( ).

c. Resort the word list alphabetically by word ending.

   i. Open the **EDIT** menu select the **OTHER SORTS** option and then the **REVERSE WORD** ( ) option.

d. Resort the word list by word length.

   i. Open the **EDIT** menu select the **OTHER SORTS** option and then the **WORD LENGTH** ( ) option.

e. Resort the word list by word consistency. This option showcases the words of the word list that appear across the most texts. This is useful to separate words that are relevant to the field from words that are specific to a sub-area of specialization.

   i. Click on the **TEXTS** button on top of the fourth column.

Generating an index

An index records the position of each token and type of a text. It greatly resembles a word list in appearance. However, it differs in potential as it allows the user to compute word clusters, among other information.

1. Identify where you want to save your index.

   a. In the main *WordSmith Tools* window (the window where you access the three analysis tools Concord, KeyWords and WordList), open the **SETTINGS** menu and click on the **ADJUST SETTINGS**… option.

   b. Click on the *Index* tab.

   c. Under *Index File*, enter the path to the location where you want to store the index. You can either type it or browse to it by clicking on the **BROWSE** button ( ) that appears to the right. If you are working at the Writing Centre, remember to save it in your U: drive.

   d. Validate your changes by clicking on the **OK** button at the top right of the window.

2. Generate an index.

   a. In the main *WordSmith Tools* window, click the **WORDLIST** button**.**

   b. A *Getting Started...* dialogue box appears. Click the **MAKE/ADD TO INDEX** button.

   c. A new dialogue box appears, in which you can modify where the index will be stored. To do so, follow the same steps as in 1.c.

   d. In the same dialogue box, you can also choose between (a) deleting an existing index and creating a new one, (b) backing-up an existing index and adding words to it or (c) adding words to an existing index without backing it up. (The first time you create an index, WordSmith may not ask you to make this choice, however.)

   e. Select your preferred location and saving option and click the **OK** button. The main *WordSmith Tools* window will show the status of the process and an information box will appear notifying you that the indexes have been correctly saved.

   f. Open the index.

      i. In the *WordList* window, open the **FILE** menu and select the **OPEN** option.

ii. The software will automatically look for the folder where the indexes are stored. Should it not be able to find it automatically, browse through your computer and locate the sub-directory you indicated in 1.c or 2.c.

iii. Open the index directory and double-click on the file with the extension *.tokens*.

iv. A window that resembles a word list appears. Note that we know that we are working with an index because the name of the window is *Index: main_index***.**

Generating word-cluster lists

Terms can be single word or multi-word units. Word-cluster lists identify groups of words that tend to appear together and that therefore may be likely to be terms. However, not all groups of words that appear together are terms, and it is up to the user and his or her judgement to filter the lists.

1. Open your text or corpus index.
   a. If you have not yet generated an index, read the previous section, *Generating an Index*.
   b. If you have already generated an index, open the WordList tool from the *WordSmith Tools* window by clicking on the **WORDLIST** button.
      i. In the **WORDLIST** window, open the **FILE** menu and select the **OPEN** option.
      ii. The software will automatically look for the folder where the indexes are stored. Should it not be able to find it automatically, browse through your computer and locate the sub-directory you indicated in 1.c or 2.c.
      iii. Open the directory index and double-click on the file with the extension *.tokens*.
2. Once the *Index* window is open, open the **COMPUTE** menu and select the **CLUSTERS** option ().
   a. A dialogue box appears where you can configure the settings of your word clusters:
      i. Select whether to generate clusters for **all** the words in the list or only for a **selection.**
         - The **ALL** option will take longer and it will cover the totality of the words that appear in the list.
         - The **SELECTION** option can be very useful if you have previously identified words that are likely to be part of multi-word term units. To generate word-clusters from selected words you must first select them by clicking on them. If you want to select more than one word, click on the first one and then hold down the **CTRL** key on your keyboard while you click on the rest of the words you want to compute word-clusters for.
      ii. Set the **size** of your word clusters (how many words they will include). The minimum size is 2 words and the maximum, 8.
      iii. Set the **minimum frequency of** the word cluster. This will be the number of times that a word cluster must appear in the text to qualify to appear in the word-cluster list. The threshold chosen will depend on the size of the text/corpus analyzed. If the corpus is very large, the minimum frequency will be high; conversely, in a small corpus the minimum frequency will be lower.
      iv. Set the **maximum frequency percentage**. This will exclude from the computation clusters that begin with words that make up a percentage of the corpus that exceeds the maximum set as a limit. This option is designed to

eliminate clusters beginning with grammatical words such as *the*, *a*, and *is*, which are very frequent and would generate many noisy clusters.

NOTE: To decide where to set this threshold, you can consult the frequency of these words in your wordlist.

v. Set where the software must **stop**. This criterion tells the software to ignore clusters that include punctuation, a paragraph break or a sentence break, as terminological units do not occur across such structures.

vi. Generate the word-cluster list by clicking the **OK** button.

vii. A new window appears, looking exactly like a word list, except that instead of single words it contains word-clusters.

NOTE: You can clean and resort this list as if it were a word list. For more information on how to do so, read the section *Viewing the Results*.

3. If you want to use this word-cluster list in the future, you can save it 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.

NOTE: When you choose the SAVE option, WordSmith saves the word list in its own proprietary format (.lst), which means you will only be able to read the file with WordSmith. You can also export the word list to other formats such as plain text, xml text or Excel spreadsheet. To do so, select SAVE AS ( ) option from the FILE menu in the WordList window.

Questions for reflection

1. After having used WordList, what are your impressions of its interface and the searching options it offers?
2. What options could be more useful to you and why?
3. Compared to other corpus analysis tools or term extractors, what are WordList's advantages and disadvantages?
4. What are the limitations that present the greatest obstacles to identifying terms? Can you think of solutions that you would like WordList to offer?

NOTE: You may be able to find answers to some of your frustrations in the more advanced options of WordList or in WordSmith's KeyWords and Concord Tools. To keep exploring WordSmith, read the WordList Level II and Keywords and Concord tutorials.