

ACCELA Corpus Linguistics Training with Oxford WordSmith Tools version 4.0

Session 2: Jul 27/04

Objectives:

- To become familiar with the **WordSmith Corpus Tools Controller** and user Interface
- To learn how to prepare original documents for working with WordSmith Corpus Tools
- To become familiar with WordSmith filing system and choosing texts

Part one: become familiar with the WordSmith Corpus Tools Controller

1.1 Getting started

When you first run WordSmith (by running `c:\wsmith4\wshell.exe`) you'll see this in the top left corner of your PC.

This is the main screen of the Oxford WordSmith Tools Controller.



It has four main menu options, a saying (which keeps on changing and which you can edit), three buttons for the main Tools, and a series of tabs. At the moment we see the "Main" tab showing that 3 texts have been chosen for Concord.

I. become familiar with the WordSmith Corpus Tools graphic user interface (GUI)

2.2 Controller



This program controls the Tools. It is the one which shows and alters current defaults, handles the choosing of text files, and calls up the different Tools.

It will appear at the top left corner of your screen.

You can minimise it, if you feel the screen is getting [cluttered](#).

For a step-by-step view with screenshots, click [here to visit the WordSmith website](#).

2.3 Concord



Concord is a program which makes a [concordance](#) using [DOS](#), [Text Only](#), [ASCII](#) or [ANSI](#) text files.

To use it you will specify a search word, which Concord will seek in all the text files you have

chosen. It will then present a concordance display, and give you access to information about

2.4 KeyWords



The purpose of this program is to locate and identify key words in a given text. To do so, it compares the words in the text with a reference set of words usually taken from a large corpus of text. Any word which is found to be outstanding in its frequency in the text is considered "key". The key words are presented in order of outstandingness.

The distribution of the key words can be [plotted](#).

Listings can be [saved](#) for later use, edited, printed, copied to your word-processor, or saved as text files.

This program needs access to 2 or more word lists, which must be created first, using the [Word List](#) program.

See also: KeyWords Help Contents Page, [The buttons](#)

2.5 WordList



This program generates word lists based on one or more [ANSI](#) or [ASCII](#) text files. Word lists are shown both in alphabetical and frequency order. They can be [saved](#) for later use, edited, printed, copied to your word-processor, or saved as text files.

Part Two: Preparing original documents for working with WordSmith Corpus Tools

Since WordSmith tools work ONLY we plain text files, then you need to convert all your files in your corpus, before you actually start using WordSmith tools. The recommended conversion options are:

plain text: Refers to textual data in ASCII format. Plain text is the most portable format because it is supported by nearly every application on every machine. It is quite limited, however, because it cannot contain any formatting commands.

UNICODE: A text format standard which uses 2 "bytes" per character. This allows for over 65,000 different characters and symbols to be displayed and makes it possible to show Chinese, Japanese, Cherokee and a whole lot of other languages.

I. Converting multiple Microsoft word files into plain text files (.txt) Using the WordSmith Text Converter utility

1. Launch WordSmith tools

2. In the controller window, go to the Utilities menu and click on the **Text Converter** option
3. Choose **Files** (the top left tab). Decide whether you want the program to process sub-folders of the one you choose. There is no limit to the number of files Text Converter can process in one operation.
4. Click on the **Conversion tab**, and:
5. Decide whether you want to make copies of the text files, or to over-write the originals.
Obviously you must be confident of the changes to choose to over-write; copying however may mean a problem of storage space.

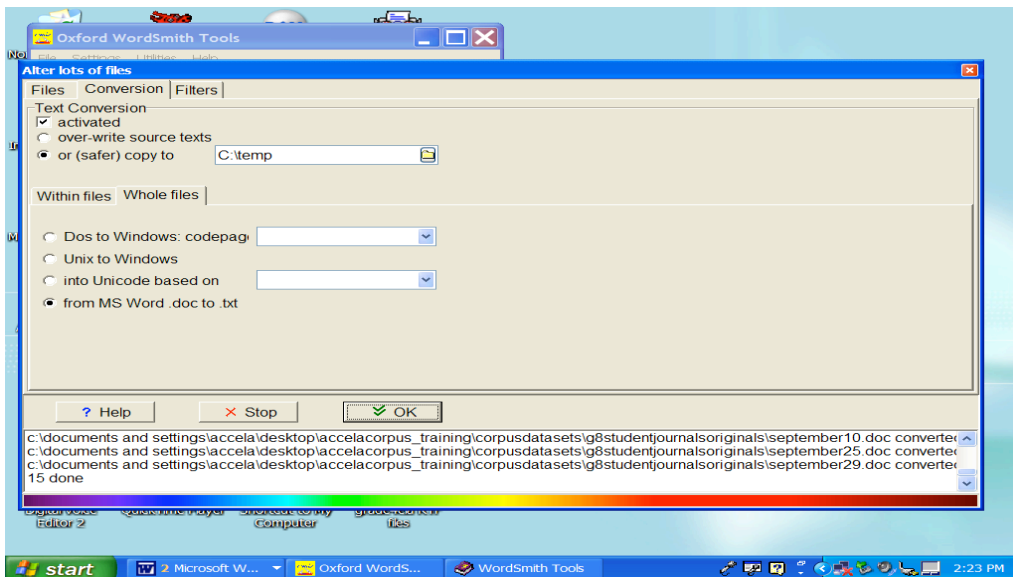
6. Specify what to convert, that is the search-words and what you want them to be replaced with. For a quick conversion you can simply type in a word you want to change and its replacement (e.g. Just one change so that responsible becomes responsible) or you can choose your own pre-prepared Conversion File.

7. If you might want some files not to be converted, or simply don't want any conversions but instead to place files in appropriate sub-folders, choose the Filters tab.

If you choose Over-write Source texts, Text Converter will work more quickly and use less disk space, but of course you should be quite sure your conversion file codes are right before starting!

Note that some space on your hard disk will be used even if you plan to over-write. The conversion process does its work, then if all is well the original file is deleted, and the new version copied. There has to be enough room in the destination folder for the largest of your new files; it is much quicker for it to be on the same drive as the source texts. If it isn't, your permission will be asked to use the same drive.

Press OK to start; you will see a list of results below.



Results showing 15 files converted from MS Word to txt format.

If you want to stop Text Converter at any time, click on the Cancel button or press Escape.

Hands on practice

From the Corpus CD files, double click on the hands on folder, and repeat the above steps to convert the spring1998g4math to the spring2004g4math files

II. Converting PDF files into plain text files (.txt)

1. From Corpus CD files, double click on the **hands on folder**, and open the **spring1998g4ela.pdf** document
2. Within Acrobat reader, go to file and click on the **Save as** option
3. On the Save As window, click on the **Save as type** arrow menu (under Object name)
4. Click on the **Text (Accessible) (*.txt)** option
5. **Rename** the file adhering to the following guidelines:
 - i. Use a meaningful (yet brief and easy to decipher) name
 - ii. Use lowercase letters and numbers only
 - iii. Delete spaces between words
 - iv. Make sure the file is a **txt file**

Part Three: become familiar with WordSmith filing system and choosing texts

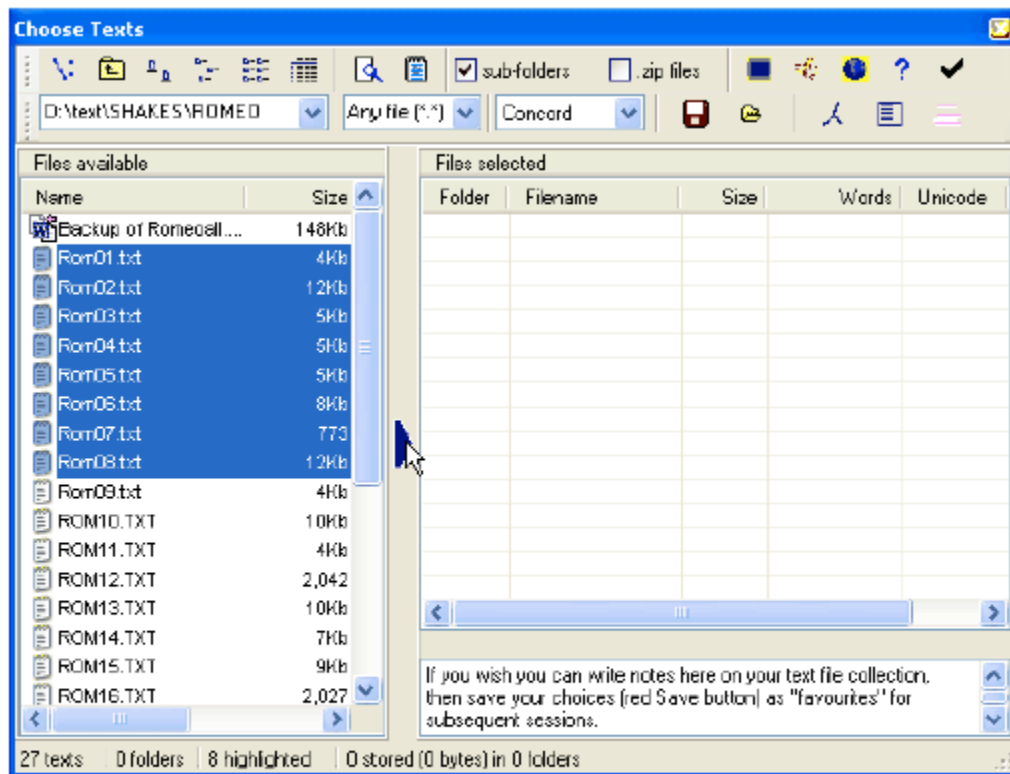
This function is accessed from the File menu in the Controller and the Settings menu or New menu item () in the various Tools.

1.2 Choosing texts

To choose text files, click the File menu in the main [Controller](#).





When you click *Choose Texts*, you will see something like this:



At the left is a fairly standard text file explorer, at the right an area for *Files selected*. Click the big blue button (where the cursor is), or drag some text files from left to right. You should see something like this:

Drives and Folders

Double-click on a folder to enter it. You can re-visit a folder if its name is in the folder window history list, and easily go back with the standard Windows "back" button . Or click on the  button to choose a new drive or folder.

Sub-Folders

If checked, when you select a whole driveful or a whole folderful of texts at the left, you will select it plus any files in any sub-folders of that drive or folder.

View

Allows you to browse within the currently selected file so as to check whether to include it. Any accented characters (e.g. æ, é) or currency symbols such as £, ¥, ¢, and [tags](#) will appear according to current [Text Characteristics settings](#). You can change these while [viewing](#) the file.

View in Notepad

Lets you see the text contents in the standard Windows simple word-processor for text files, *Notepad*.



Get from Internet

Allows you to access [WebGetter](#) so as to download text from the Internet.

Zip files

If you double-click on a [zip file](#) you can enter that as if it were a folder and see the contents. You can view these too.

Favourites

Two buttons on the right ( and ) allow you to save or get a previous file selection, saving you the trouble of making and remembering a complex set of choices.

Test for Unicode

This button tests any files selected. In the screenshot above, no tests have been done so the display shows ? for each file. If the text file is in [Unicode](#), the display shows U, if plain ASCII or [Ansi](#) text, if it's a Word .doc file, D.

Clear

As its name suggests, this allows you to change your mind and start afresh. If any selected filenames are highlighted, only these will be cleared.

OK

This puts the current file selection into store. All files of the type you've specified in any sub-folders will also get selected if the "Sub-folders too" checkbox is checked. You can check on which ones have been selected under *All Current Settings*.

3.3 getting started with WordList

For a step-by-step view with screenshots click [here to visit the WordSmith website](#).

I suggest you start by trying the Wordlist program. In the main Oxford WordSmith Tools window (the one with Oxford WordSmith Tools [Controller](#) in its title bar), choose the Tools option, and once that's opened up, you'll see Wordlist. Click and WordList will open up, on the right hand side of your screen.

Start

You will see a dialogue box which lets you [choose your texts](#) or change your choice, and make a new word list.

If you have [never used WordSmith](#) before you will find a text has been selected for you automatically to help you get started.

There are other settings which can be altered via the menu, but usually you can just go straight ahead and make a new word list, individually or as a [Batch](#).

You'll find that WordList starts processing your file(s) and a [progress](#) window in the main Controller shows a bar indicating how it's getting on. After WordList has finished making the list, you will see three windows showing the words from your text file in alphabetical order and in frequency order, and statistics.

Don't forget to [save the results](#) (press F2 or ) if you want to keep the word list for another time.

See also: WordList Help Contents.

Sources:

Source: WordSmith Tools Help (c) Mike Scott

<http://www.pcwebopedia.com/>