

WordSmith Tools 8.0 - Distinguishing features

This document is intended to be read in conjunction with the 'Choosing a CAQDAS Package Working Paper' which provides a more general commentary of common CAQDAS functionality. This document does not provide an exhaustive account of all the features and functions provided by WordSmith Tools 8.0, rather is designed to highlight some of its distinguishing features. The Comment section at the end details our opinions on certain aspects of functionality and usability. With thanks to Mike Scott for ensuring the accuracy of this document.

Background and philosophy

<https://lexically.net/wordsmith/>

WordSmith Tools is a program suite developed for the lexical analysis of large corpora of texts by the British linguist Mike Scott (University of Liverpool). First released in 1996, it was based on MicroConcord (Mike Scott and Tim Johns, Oxford University Press, 1993). The suite consists of tools developed to create interactive concordances and has been used by Oxford University Press to prepare dictionaries, by language teachers and by researchers investigating linguistic patterns in many different languages worldwide. This document reviews version 8, distributed by Lexical Analysis Software Ltd.

Minimum System Specifications (recommended by developer)

WordSmith Tools version 8 is for Windows 7 or later (either 32 or 64-bit). A fairly modern (less than 4 years old) laptop or desktop PC is recommended. WordSmith will run on Mac OS using a virtual machine with Windows installed. 100 Mb disk-space and 1GB of RAM minimum.

Figure 1. The Controller window – starting point for Concord, WordList and KeyWords plus Utilities



SETTINGS icons and UTILITIES listed in main window left and centre –including **Language Chooser**–unlimited choice of languages –several can be chosen for current work; **Corpus Checker** –checks relevance of data; **Text Converter** –batch converts plain text files into Utf16 encoding OR processes changes; **Character profiler** – counts characters; **CharGrams** – see Figure 3.; **WSConcGram** – finds word pairings, triplets, quadruplets. (Red flags indicate active processes)

Structure of work in WordSmith

Three *main tools* form the central (separate but integrated) functions of the software - **Concord**, **WordList** and **KeyWords** – and they are supplemented by an extensive set **utility programs**. Main tools and utility programs have discrete interfaces and are launched from the **Controller** (see Figure 1.) ■ There is no overall single 'project file', rather work begins by opening selected texts within the relevant tool ■ The focus of work is on exploring and accessing the occurrence and/or structure of words, language and syntax, and there are many ways to count, compare and break down these elements within texts ■ The purpose and end-results of any analytic exercise can be diverse as each tool allows the flexibility to study a textual corpus of any size and how its vocabulary and language varies. This work is entirely content based ■ The user can work qualitatively at the text level and/or use quantitative computations to compare, evaluate or build a valid lexicon or dictionary from the study corpus ■ **Settings** for the *main*

tools are set up in the **Controller** with *What you get* (parameters) and *What you see* (display) tabs allowing the basis of operations to be determined ■ **Saving work** - the results of any operation within any *main tool* or *utility program* can be saved as one or more files and these are key to how work progresses ■ There is no set pattern of, or sequence of work so different users may use the tools in unique ways. For instance, translators might only use the more qualitative **Concord** tool and the **Aligner** Utility (see Figure 3) to compare versions of texts in different languages, whereas Lexicographers might undertake analytic tasks that involve only the more quantitative **WordList** and **KeyWord** tools. Although there are 3 “main tools”, the **Utility** tools can perform equally prominent tasks

Data Types and Formats in WordSmith

Files must be plain text (.txt) and ideally UTF-16 encoded to support almost universal languages and character sets ■ The **Text Converter** *utility program* converts texts into UTF-16 and can also process and modify vast amounts of data across folders or within files to facilitate the use of all WordSmith Tools.

Searching content using the Concord Tool

The **Concord Tool** (see Figure 2) generates **Concordances** and is a straightforward way to explore content for user-specified words and phrases or collections of them ■ The Concord tool produces interactive lists of places (**concordances**) where words / phrases appear using Key Word in Context (KWIC) functionality, as well as situational information such as position within sentence, paragraph and documents (see Figure 2) ■ The *View* menu allows a voiced representation of selected text ■ More quantitatively, tabs at the base of the table denote frequency, collocation, patterns, clusters, timeline, filename and graphic plots representing dispersion of a word within each file are interactively connected with text ■ Hits can be categorized into **Sets**, with options to reformat source data to include the Set information to make it searchable on that basis

Searching content using the WordList Tool

The **WordList tool** provides many ways to explore texts qualitatively and quantitatively ■ Generate lists of words to study vocabulary in one text, across a *study corpora* or a whole *reference corpus* ■ Compare the frequency of a word in different text files or across genres ■ Translations comparison of associated words or translation equivalents across different languages ■ Compare a smaller study corpora with the reference corpora ■ Visualise alternative representations via the Compute menu e.g. **Concordances**, **KeyWords**, **Wordclouds** and further content-based calculations such as **Lemmas**, **Entropy**, **Compounds**, **Dispersion** and extensive ways to view additional information and statistics.

Searching content using the Keywords Tool

The Keywords tool is used in conjunction with saved WordLists (see above) to extrapolate and compare significant words ■ Wordlist files generated from the *study corpora* (or subsets of it) are compared with the *reference corpora* ■ **KeyWords** can also be launched via an option in the WordList file menu to Compare two Word Lists. Similarly to WordList options, via the KeyWords/Compute menu a selected word/line can produce a Concordance and therefore interactive connection with relevant lines of text. Most Compute options available in the WordList program (above) are also available in KeyWords.

Relationship between main tools and utilities

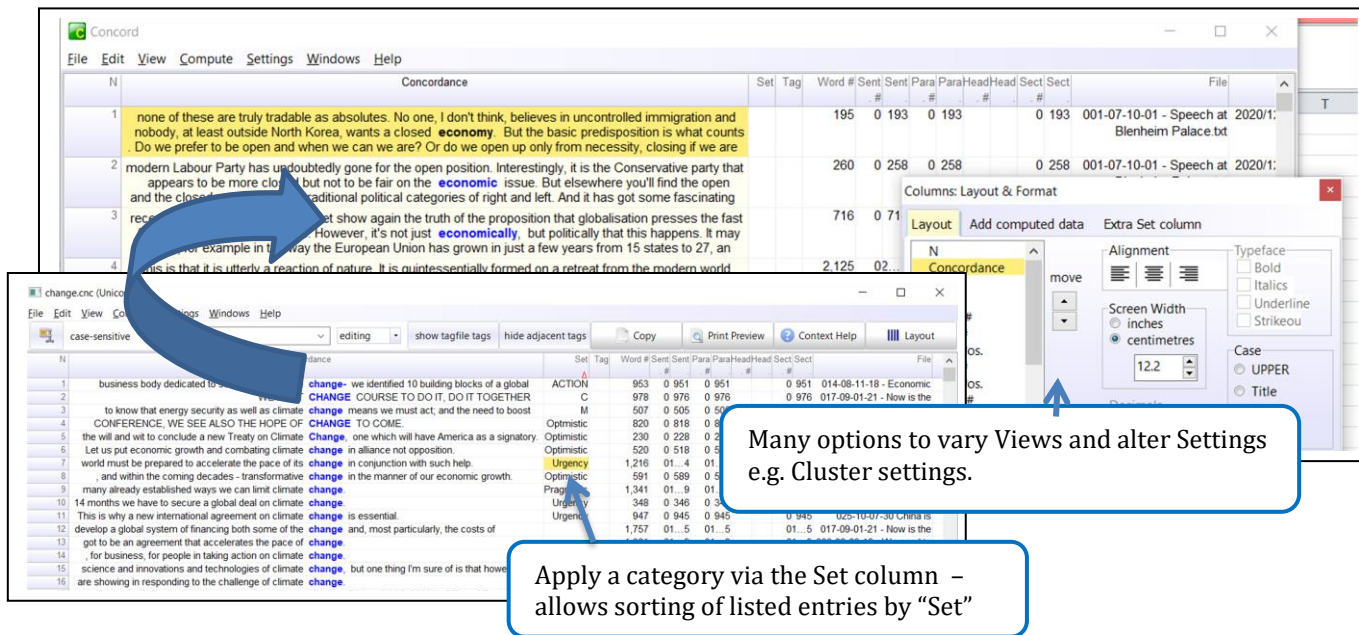
The separate main tools **Concord**, **WordList** and **Keywords** perform independent tasks on discrete datasets selected by the user but they are also interconnected. For example, WordList and KeyWords can connect to the qualitative dimension for each line of text or word selected at a Concordance tab or menu option ■ Separate analytic processes can build on one another as required, dependent on systematic saving or changing of files as work progresses ■ The **Keywords** tool relies on the earlier creation of **Wordlists** both for the *reference corpora* and for the *study corpora* ■ There are many different ways of varying settings to view, compute and export data ■ The **Utility** programs offer other extensive ways to break down content (see below).

The role and functioning of Utility Programs

Utility programs serve two overlapping purposes - set-up and analysis ■ **Utility programs for preparing texts and setting up for use.** The Version Checker ensures the most up-to-date version of WordSmith tools is installed via continual updating. The Data Converter converts files in batches into the optimal format for analysis (plain text with UTF16 encoding). The Minimal pairings program finds possible typos and pairs of words which are minimally different from each other (minimal pairs). The Language Chooser allows several languages to be simultaneously analysed and compared. The Corpus checker checks relevance of dataset (as above) ■ **Utility programs related to more analytic functions** The Aligner for translation purposes allows different language

versions of data to be merged according to chosen settings and presents sentences or paragraphs in each language sequentially to allow comparison (see Figure 3). The merged files can then be fed through e.g. Concord tool or other Main tools to produce specific searches. The Character profiler counts single characters, symbols or punctuation marks. The Chagrams tool enables the tracking and frequency of particular sequences of letters according to particular locations within words. The WSCongrams tool tracks the usage of common word pairings, triplets or quadruplets

Figure 2. The Concord tool – creating interactive lists of occurrences (concordances) with KWIC



Closeness to data – interactivity in WordSmith

There is good interactivity between the Main Tools. In a *qualitative* sense, views of Concordances (typically consisting of rows of text) provides the best interaction between ‘results’ and the whole context qualitative data (see Figure 2.) ■ **Concordance** views can be summoned from within quantitative lists in **WordList** or **KeyWords** programs and this is key to interactivity across all main tools and some Utilities ■ **Graphic Plot** indicators generated in the KeyWords tool are interactive with underlying text ■ Saved output files importantly still retain clickable interactivity with source text when reviewing them within the Wordsmith tools.

Categorising finds in WordSmith

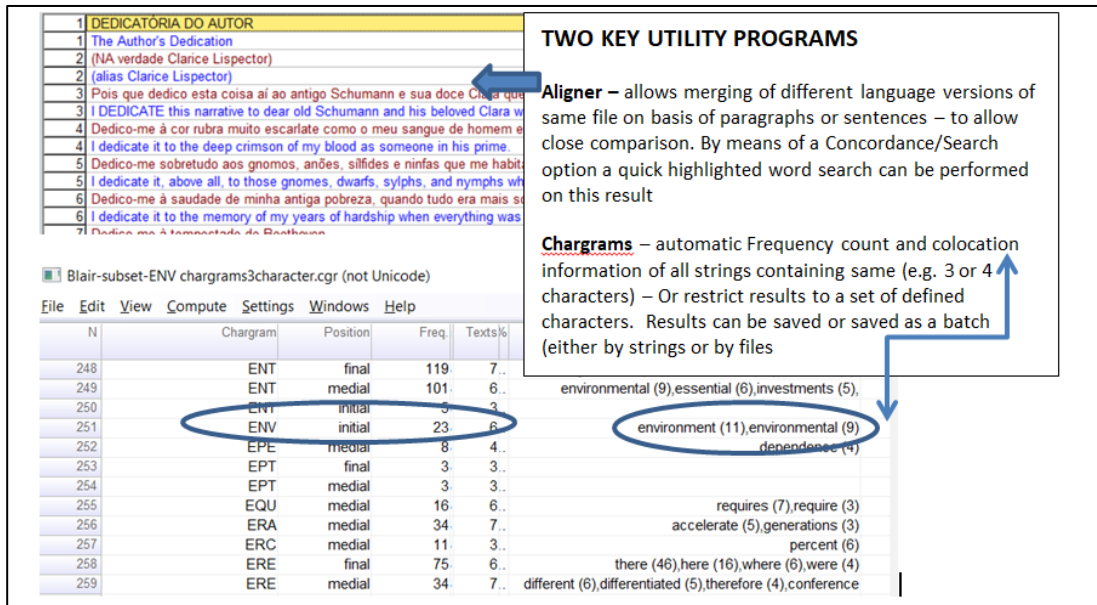
The **Set** and **Tag** options in the Concordance tool allows results (lines or columns) to be coded and results can subsequently be sorted by these categories. This feature could also be used to denote linguistic or syntactical characteristics ■ The **Text Converter** Utility allows source text to be selectively or globally modified to indicate sets/tags which can then act as precoding annotations to enable the Concord tool to search for occurrences.

Organisation of data in in WordSmith

Organisation of data for use by WordSmith Tools is focused at the data preparation stage since there is no one project file structure so there are no structural tools *per se* to organize data within one ■ However, the files and folders being interrogated can be saved as analysis proceeds, and therefore the naming of these selections can have an organizational purpose ■ In addition, the **Text Converter** utility offers ways to prepare and filter data in a variety of ways (e.g. globally or selectively) so that searches made in the main tools can take place on already filtered selections of files or identify required passages and structures based on flags or syntax or “tags” already inserted in the data ■ Finally, some main tools depend on defining whether a stage or work concerns the Reference Corpus or the Study Corpus so these sets of data can have an important organizational dimension. For example, a linguistic or lexical study of a whole body of literature may have to compare content from the specific *study corpus* with the larger body of appropriate *reference corpus* or *corpora*, and the **Corpus checker** utility allows the

relevance of a corpus of texts or a file to be assessed. For instance, a news story may refer incidentally to an issue you are interested in, but really be mainly about something else. This utility finds a random set and can save that for sharing with others so you can agree on criteria.

Figure 3. Utility programs include and Aligner and Chagrams



TWO KEY UTILITY PROGRAMS

Aligner – allows merging of different language versions of same file on basis of paragraphs or sentences – to allow close comparison. By means of a Concordance/Search option a quick highlighted word search can be performed on this result

Chagrams – automatic Frequency count and colocation information of all strings containing same (e.g. 3 or 4 characters) – Or restrict results to a set of defined characters. Results can be saved or saved as a batch (either by strings or by files)

N	Chagram	Position	Freq	Texts%
248	ENT	final	119	7..
249	ENT	medial	101	6..
250	ENV	initial	5	3..
251	ENV	initial	23	6..
252	EPE	medial	8	4..
253	EPT	final	3	3..
254	EPT	medial	3	3..
255	EQU	medial	16	6..
256	ERA	medial	34	7..
257	ERC	medial	11	3..
258	ERE	final	75	6..
259	ERE	medial	34	7..

Interrogating the dataset in WordSmith

WordSmith is all about interrogation - as described above the main tools and most of the Utilities are concerned wholly with interrogation and accessing and calculating 'content' in different ways. Significant statistical, collocational, positional and interactive plot data can be computed alongside the results produced by each main tool ■ Each new operation can be undertaken on a subset selection of data.

Output in WordSmith

The results of most interrogations can be saved as files that can be recalled and worked on further at any point ■ Files can be saved in multiple formats (including .txt, .xml, .xls, .rtf or concordance lists ■ Saved files provide standalone or interactive summaries, computations, word lists and concordances from which to write up analysis ■ Saved files can also provide the starting parameters for subsequent interrogations

Teamworking in WordSmith

WordSmith Tools is a single-user software, but since results are regularly saved for further work (within or outside the software) analysis is generally shareable. The range of file formats for saving work enable further work by team members who do not have WordSmith. Output in xml format enables viewing in a web browser ■ **Export** functions allow operations to be repeated on the same or another version of WordSmith.

Comment on WordSmith

WordSmith tools provides a huge array of detailed and sophisticated analysis possibilities originating from a Lexicology tradition and therefore is excellent for a multitude of linguistic analysis purposes. Many of the tools also facilitate the more quantitative aspects of Discourse, Conversation and Text analysis and the general content analysis of documents and texts across Social Science disciplines.

Wordlist and Keywords can quickly reveal vocabularies, content-based patterns, syntactical structures with excellent interactivity between each line/word listed to parallel Concordances (Figure 1). This provides an easy way to move from the quantitative dimension of texts to the qualitative. The excellent utility Aligner (Figure 3) might be particularly useful for researchers involved in Translation studies or Multi-national team projects where the comparison of idiomatic

phraseology is important. Such tools provide speedy proactive access to the physical content of a mass of textual data or documentation along topical lines.

Concordances could be useful to the more interpretive researchers needing an overview of dominant (or relevant) issues in a large body of supporting documentation, for example for a literature review or mixed-methods study. While WordSmith may have less overall utility to the more interpretive researcher because of limited data management or reflective tools for coding or memoing, WordSmith is helpful for exploring texts which is often part of interpretive processes, and for any study where there is a need to measure the physical content of large amounts of text.

Work consists of a series of discrete and complete operations which are quickly run using a Main program or a Utility on whatever selection of data is current. Though programs are discrete, there is good interactivity between tools as there are many menu options and tabs to move between different views and operations. The results generated within one program are then saved for output or further interactive examination in another.

*WordSmith tools is designed to process very large datasets. The developer compares 3 million to 30 million words in describing differing speeds of processing. The range of languages, useable character sets and size of datasets is almost unlimited and this is enabled by the use of plain text (UTF16). The **Corpus Checker** is a significant extra Utility device which provides a set of statistical measures concerning the relevance of a body of data to prevent unnecessary effort in data which may only have a superficial or spurious connection with subject matter. Considerable effort might be put to earlier 'processing' and 'preparing' data using its Utility tools. For the purpose of analysis it's also worth noting that unusually, the results of some tasks (e.g. using the Set tool in Concordances) can be used to automatically modify source data to enable further progressive searching functions..*

Saving and exporting work is designed for transferal to other users of WordSmith or into other software. Saving files can be dynamic and may be an important aspect of progression within WordSmith. However, since there are so many separate tools and processes and no overall mechanism to track processes or progression to another task, the user must develop a very systematic way of storing and labelling saved files which are the result of examining or even sometimes changing content - especially when varying text selection for each operation to ensure the correct set of data are used next time. The software is so very extensive in its parts that there is no one path to be taken through its tools. Rather it is a massive tool chest – from which any one device could produce just what is wanted for a particular purpose. This and its budget cost give it a universal utility.

Accessibility – in addition to its ability to handle and analyze almost any language, it also provides a voiced read-back tool when particular segments of data are selected.

*Software updates and support are excellent. This is a very advanced, low cost software where the look of the software remains the same but there are continual updates so that it stays compatible with operating system changes etc. The **Version Checker** in Controller prompts the user to download the most up-to-date version.*

The no-frills user interface is countered by excellent contextual help resources which are available at every stage of working, rather than via a more corporate web-based support infrastructure. Some parts of the software are easy to use. Other aspects are not so user friendly and the reason is often to do with very extensive detailed technical settings which can be varied for every operation. There are three excellent Get Started videos, but these really only scratch the surface of the enormous range of possibilities provided by WordSmith Tools.

Further Reading

Scott, M (2008) **Developing WordSmith**. International Journal of English Studies, Vol 8(1), pp95-106

Wilkinson, M (2011) **WordSmith Tools: The best corpus analysis program for translators?** Translation Journal. Vol 15. No.3

International Journal of Corpus Linguistics <https://benjamins.com/catalog/ijcl>

Corpora <https://www.eupublishing.com/loi/cor>

WordSmith resources <https://lexically.net/wordsmith/research/>