1. INTRODUCTION AND OVERVIEW. WordSmith Tools 5 (henceforth WST 5) is, according to its self-description, “software for finding patterns in text.”

This review is written from a South African perspective where WST 5, to the best of our knowledge, is mainly used by linguists, lexicographers, and terminologists of Afrikaans and English, as well as for the nine official Bantu languages of South Africa. Certain problems encountered using WST 5 mentioned in this review should not necessarily be debited to WST 5. They can be attributed to factors such as computers not meeting minimum compatibility requirements, lack of experience with such software, or limited bandwidth. This review attempts to highlight the most prominent features of the latest version of WST 5 (5.0.0.334) for those who are not familiar with the program in terms of our hands-on experience.

The corpus linguistics era dawned in South Africa, and Africa as a whole, in the nineties, and with it came the need for text manipulation. Linguists are mainly interested in studying the complex morphological and syntactic structures of especially the Bantu languages, terminologists perform term extraction, and lexicographers require a tool to compile lemma lists, accurate definitions, and authentic examples for dictionaries.

Generally speaking, the computer skills of users in South Africa vary from basic to fairly advanced. This is in line with WST’s own description of its intended target users, namely “students and teachers and researchers interested in language, history, politics, medicine etc.”

It is important to note that WST is not a fixed product or static program with formal release sub-version numbers every few years, but according to its self-description “evolves gradually and is re-issued very often (about once a week)”; the suggestion is that the user should regularly download the latest version. Version 5.0.0.334 is the latest available version of WST 5 (March 5, 2011). Upgrading to the latest version can be done automatically from the main menu.

An immediate plus point of WST 5 is that a fully functional demo version can be downloaded which enables use of the program without delay. The user can thus immediately start generating, for example, word lists and concordance lines. The only limitation is a restriction on the amount of data output, such as a maximum of 100 words in an al-

---


2. See De Schryver & Prinsloo (2000a and 2000b) for corpus applications using an earlier version of WST.

alphabetical word list, or 10 lines in a keyword in context query, followed by a “past demo limit” notification. This demo version can then be updated to the full version by purchasing a registration code from Lexical Software Analysis⁴ or Oxford University Press⁵ (UK) at the very reasonable price of £50 (approximately US$70–80 or €65), with a 50% discount if updating from a previous version.

It is advisable to familiarize oneself at an early stage with the full capabilities, restrictions, and default settings by reading through the settings and help functions. The following section provides an overview.

2. WORDSMITH SETTINGS AND TOOLS. The opening screen features the main menu and the three fundamental functions of WST 5, namely Concord, KeyWords, and WordList, along with the main categories of settings: text characteristics, tags, concord, keywords, and wordlist.

![Figure 1: The main functions of WordSmith Tools](image)

The first-time user is advised to check these settings before embarking on large-scale text manipulation. It has also been remarked that the many settings can be confusing, and users might take a while to comprehend the changes in output resulting from changes in the settings. Users should consult the very detailed get-started guides at http://www.lexically.net/wordsmith/support/get_started_guides.html. A very simple “get started” video tutorial would be a useful addition to the software, as would a separate tutorial for more advanced users. New users should be encouraged to “play” with the software and its settings before undertaking their actual research data runs. Some users reported that they learned the program through trial and error but would prefer step-by-step guidance in addition to the existing guides.

---

⁴ http://www.lexically.net/wordsmith/purchase/index.htm

⁵ http://ukcatalogue.oup.com/product/9780194597005.do
Figure 2 shows a user-friendly screen for performing file selection, a necessary action for using the Concord, KeyWords, or WordList functions.

3 WORDLIST. WordList generates word lists based on one or more input text files. Supported formats include HTML or any other plain-text format that uses encodings such as ASCII or Unicode. A complete list of supported formats and encodings is provided in the Help files as well as on the website Help section. Lists can be compared within WordList, or a consistency analysis can be performed for the purpose of stylistic comparison. Word lists generated in this way are required as input to generate keywords and key-key words (these will be discussed in a separate section).

The main outputs of the WordList function are a frequency list, an alphabetical list, and statistics of the manipulation (see figure 3). WST 5 gives a frequency list as a default, with tabs at the bottom of the window for the other functions, including tabs for filenames and notes.

WST 5 provides additional information, such as the number of texts and the percentage of occurrence in texts in the word list.

Users report being quite satisfied with the processing speed of WST 5, even when used
on standard desktops or laptop computers. Two instances of slow processing were reported: one occurred while running WST 5 on a corpus containing 750,000,000 tokens (running words) in more than 1,300 different text files, and the other running WST 5 under Parallels on a MacBook Pro. Processing speed depends on the specification and performance of the computer. Performance relies heavily on RAM. Computers with sufficient and higher-speed RAM perform best, based on our experience using a lower-end computer specification of Intel Celeron 2.16Ghz FSB 666Mhz, 1Gb DDR2, 333Mhz RAM in contrast with a higher-end Intel Core 2 Duo 3Ghz FSB 1333.3Mhz, 4Gb DDR3, 1666Mhz RAM.

A word of caution: the user should make sure that the word lists are saved under different names. We found that saving the newer lists results in replacement of older ones without warning.

Frequency and alphabetical lists are invaluable, for instance, to the lexicographer as the starting point of a lemma list for a new dictionary. They could also be useful in the revision of an existing list, for example, in terms of inclusion versus omission of lemmas in the dictionary, which is probably the most common usage of the WordList function in South Africa.

The search function of WST 5 presents a window in which searches can be performed on Word, Frequency, Percentage, % in corpus, Number of texts, % of texts, and so on. What is apparently missing in WST 5 is a “match entire cell contents” option. For example, a search for 100 in the frequency column of WST 5 brings up 353,100 (exact matches are highlighted here); on the next click, the result might be 100,924 or 100,817. In the case of words, the default setting is full word, but the asterisk can be used for parts of words: *voer* renders forms such as invoerders, kragtoevoer, and voertaal. A useful addition to WST would be a search direction (up/down) option. In some cases it is also necessary to clear the result box before a new search is successful; for example, clearing the previous frequency value 50 before the new value 200 could be searched, as shown in figure 4.

![Figure 4: Search window in WST 5](image)

Lemmatization can be performed within WST 5, and frequency counts are adjusted automatically, as with forgives lemmatized to forgive in figure 5.
4. CONCORD. The Concord function makes a concordance using input files. It renders keywords in context, which simply means it runs through the text(s) and returns a search word the user pre-entered, in numerous contexts and with co-texts to the left and right, as in figure 6.

A very useful function for our applications proved to be the ability to re-sort lines based on whether words preceded or followed the keyword – a choice of 1–5 words to the left (L1-L5) or right (R1-R5) (see figure 7). Three sorting options can run simultaneously. The input box for the search word function allows for Boolean operators: the search form *sebenz* returns thousands of derivations of -sebenz- containing prefixes and or suffixes in a Zulu corpus.
Figure 7: Input box for search words in WST 5

Figure 8 gives a brief impression of the outputs of other major functions of Concord, e.g. Collocates, Plot, Patterns, and Clusters:

**Collocates in concord, e.g., local occurs 7 times with government:**

**Patterns in concord, e.g. frequent patterns with break:**
Plot in concord, i.e., positions of the word in each file:

<table>
<thead>
<tr>
<th>File</th>
<th>Word</th>
<th>Concord</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>File 1</td>
<td>bank</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>File 2</td>
<td>loans</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>File 3</td>
<td>assets</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>File 4</td>
<td>to break</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>File 5</td>
<td>will break</td>
<td>20</td>
<td>3</td>
</tr>
</tbody>
</table>

Clusters in concord, e.g., containing break:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Word</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Break in the</td>
<td>break</td>
<td>20</td>
</tr>
<tr>
<td>To break through</td>
<td>break</td>
<td>10</td>
</tr>
<tr>
<td>To break</td>
<td>break</td>
<td>10</td>
</tr>
<tr>
<td>Will break in</td>
<td>break</td>
<td>10</td>
</tr>
<tr>
<td>The break of</td>
<td>break</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 8: Selected functions of Concord in WST 5

A problem encountered with the generation of concordance lines in WST 5 is that all of the data is processed before the actual generation and display of the output commences. This means that generating concordance lines from a very large corpus is impossible in a live demonstration or classroom situation. This problem did not exist in WST (1996), in which generation and display started immediately. A setting that allows one to preview and use at least some of the concordance lines while it is processing would be a welcome addition to WST.

5. KEYWORDS. For our purposes, key words are defined as words that appear substantially more frequently than expected (positive keys) or substantially less than expected (negative keys) in a given corpus than in a comparable general corpus. For example, in a word list culled from a corpus containing the annual reports of five commercial banks, the top ranking positive keys (see figure 9) include typical banking terms such as bank, loans, assets,

---

6 The reason was technical and had to do with a fundamental change Microsoft imposed when moving to 32-bit processing (Windows 95) (Mike Scott: personal email correspondence, 30 April 2011).
and credit. A word list generated from a general corpus of English words would be expected to include a lower percentage of these financial terms than does this partial corpus.

Terminologists find this function helpful in extracting terms for the compilation of Language for Specific Purposes (LSP) dictionaries.

More advanced functions of WST 5, described below, include Key-keys, ConcGrams, File Viewer, Corpus Corruption Detector, and Text Converter (from PDF and DOC formats).

6. KEY-KEY WORDS. Key-keys are keywords shared across multiple lists of keywords, or in WST’s own terminology, “key-words which are most frequent over a number of files.”

The process to generate key-key words is similar to the one for normal keyword lists. In the bank example presented above, we would end up with 5 keyword lists from 5 wordlists for the creation of a keyword database. Words such as account, activities, assets, average, and balance eventually came up as key-keys in Figure 10.

---

**Figure 9:** Positive keys in annual reports of five commercial banks

**Figure 10:** Key-keys in annual reports of five commercial banks

---

7. **WSConcGram.** Our attempt to define ConcGrams in a simpler way is to say its purpose is to find all the different combinations of two or more words—say, words A and B in combinations such as AB, BA, ACB, BCA where C stands for one or more other words.⁸

Consider, for instance, WST’s own example of combinations in for hands and pockets in figure 11.

![ConcGrams in WST 5 for hands and pockets](image)

**Figure 11:** ConcGrams in WST 5 for hands and pockets

8. **INSTALLATION ON A REMOVABLE DRIVE.** The ability to install this application on a removable drive could be useful to certain types of WST users, such as students who are more likely to roam from computer to computer. WST 5 seems to be self-sufficient in terms of computer software setup. It does not require the installation of software such as Java Virtual Machine or the .NET framework to operate. The only software requirement is that it is run under the Windows operating system.

9. **A LEARNING CURVE FOR BEGINNERS.** In addition to the recommendation to study WST’s settings prior to the large-scale processing mentioned above, there is also a slight learning curve to overcome with WST 5. Some interface interactions are different from the standard Windows applications. An example is the search function, which is traditionally located under the Edit menu for most Windows applications (such as the Office Suite); in WST 5, it is found in the View menu. These are minor issues, but could unnecessarily have a negative impact on new users’ first impressions of WST 5.

An instance of unexpected behavior is illustrated in figure 12, in which a keyword list was created that contained more than the default 500-maximum keyword count. A pop-up message informs the user that this limit was reached, and that the operation is complete. The problem is that the main Window stays in focus, and the popup-message is obscured. This creates the impression that the application froze at 39% completion. Other minor unexpected behaviors or interface design decisions may also take some getting used to.

It is also important to note that the user must decide and be aware of setting changes that only take effect for a certain work session, as opposed to settings which become the default value. The user should not assume that a specific change in settings will apply to current and future sessions, when in fact it may only be valid for the current session. The best solution is to decide upon the most useful default values, and then only change the settings required for a specific session without losing the default ones.

In figure 13, ticking or clearing the tick in the Save button box accounts for the crucial difference: to change default settings, tick the Save button. Users often tend to overlook this.

A final point in discussing the Wordsmith learning curve is that although the Help files and online Help functions in general clearly describe WST’s functionalities, a few
improvements could benefit users. For instance, using corpus texts that are in the public domain, and offering step-by-step, example-driven guidance, would be especially useful for the more complicated functions such as key-keys and ConcGrams. Users could then follow such step-by-step guidance using the same wordlists and concordance lines, and compare their results with those shown in the examples. In this way, users could feel more confident in working with WST 5.

10. CONCLUSION. WordSmith Tools 5 is an excellent program for text manipulation and offers a huge variety of functions. WST 5 is easy to use and has been substantially improved since its pioneering release in 1996. Its popularity is constantly growing, and for many linguists, lexicographers, and terminologists, WST is the only sufficient and efficient solution to their text manipulation needs.

**Pros:** The ability to manipulate huge amounts of data stored in thousands of files. Easy to use, no advanced computational knowledge required, and operates on computers with relatively slow processors. Very reasonably priced. Full version can be downloaded.

**Cons:** A learning curve to overcome. Interface interactions unique to the program, and not quite the same as expected from standard Windows OS applications.

**Primary function:** The creation of frequency lists and alphabetical wordlists, as well as keywords in context (concordance lines).

**Platforms:** WordSmith Tools version 5 is intended for use with Windows 2000 or later, including Windows 7 (32- or 64-bit versions), or one of the previously mentioned Windows versions on Virtual Machine instances run from Linux or Apple Mac Operating Systems.

**From:** Lexical Software Analysis or Oxford University Press (UK)


**Application size:** Approximately 30 MB

---

9 Note the capital S in the final two versions (http://www.lexically.net/publications/citing_wordsmith.htm).
REFERENCES:

Akita Bank: http://www.bank-ir.ne.jp/akita/pdf/akita05.pdf
ICICI Bank: http://www.icicibank.com/aboutus/annual.html

D.J. Prinsloo
danie.prinsloo@up.ac.zs

Daniel Prinsloo
dprinsloo@gmail.com