ISSN 1442-438X
*CALL-EJ Online*, *7*(2), 55-58

## [Software Review]

# Freely available Software for Comprehensive Corpus-Based Analysis

**Katherine Moran**
Georgia State University, U.S.A.

## Introduction

Corpus-based linguistic analyses are highly dependent on the scope and type of corpus used. A benefit of using a corpus is to see how language patterns function within the context of a particular area of language use (Stubbs 2002). Using a small, specific corpus of complete texts, of newspaper articles written on politics, or a collection of student writing, for example, can be very useful in finding common lexical patterns otherwise not readily apparent. However, the results of such a study do not give much information in terms of general language patterns outside the context of political news writing or student writing. Comparing a small text analysis with language usage in a large, general corpus can reveal more information about both general language use as well as a different perspective on the use of language in the genre-specific corpus.

Finding effective analysis software for personal use can often be challenging. In this review, I will look at three programs available for free on the Internet that can be used for various types of corpus-based analyses. These programs serve very different purposes, but when used in conjunction with one another, they can allow for a more complete analysis and understanding of language function.

The first program I will discuss is called TextStat. This program is available for download and allows the user to compile data into a corpus and perform various analyses. The next program is called WebCorp and uses the Internet as its corpus. This program allows the user to set very specific search parameters to comb the web for as specific or general a query as desired. The third program is WASPS and is also available online. The program is based on the British National Corpus and uses 100 million words from this corpus to show word patterns and grammatical features based on user input.

By integrating all three programs, researchers can gain a better understanding of how language characteristics operate in various aspects of use.

## Description

### TextStat2

Matthias Huning created TextStat2, or simply TextStat as it was for several iterations, in 2000, in the department of Dutch Linguistics at the University of Berlin. The program interface is can be set to English, Dutch, or German and the software is capable of dealing with several other languages. TextStat is downloadable from www.niederlandistik.fu-

berlin.de/textstat/ and runs in Windows. The program balances very useful features without being highly complex.

The primary feature of the program is the corpus-building component. In corpus mode, accessible by clicking the button labeled corpus, the program is ready to open an established corpus or allow the user to create a new one. The corpus-building feature is accessible by either clicking the new corpus icon or by selecting a new corpus from the corpus pull-down menu at the top of the screen. Either of these actions opens a save window, prompting the user to name the corpus and choose the location where the corpus will be saved. The new corpus is now created and saved, but at this point, it has no files in it. A message will now pop up prompting the user to add files to the corpus and describing the types of files the program accepts. There are very clear directions written on the interface as the user accesses various features. Files can be added to the corpus locally or from the Internet. By creating a corpus, the user has control over the type of texts analyzed and can create a small, genre, or register specific corpus to search for frequent features relative to a particular type of text.

After compiling and selecting a corpus, the user can perform various analysis tasks by selecting either word form, to get frequency counts or concordance to get a Key Word In Context (KWIC) list. I suggest beginning with word form for several reasons. First, starting with a list of word frequencies can instantly reveal common features shared by texts in the corpus by showing which words are being used the most and which words are not. Looking at single words in a frequency list can be less daunting than trying to analyze lists of concordance lines. The context of a concordance line can also occlude the concise clarity of word frequency.

There are several options for creating word frequency lists in TextStat2. The entire text can be sorted alphabetically, by frequency starting with the most frequent or by frequency starting with the least frequent, labeled retrograde in the program. The user can set minimum and maximum occurrences for words to appear on the list or search for the frequency of a single word or word-form.

Clicking on a word in the frequency list brings up the KWIC index for all occurrences of the word. These concordance lines show forty characters to the left and right of the node word by default, but the settings can be adjusted in the concordance section of the program. Clicking on the node word in a concordance line gives the expanded context with the concordance line highlighted and a link to the file from which the word came. The concordance feature can be accessed separately as well, allowing for selection and variation in the search terms. In this section more advanced searching options are available including wildcard searches and searches involving two-node words.

The data obtained from TextStat2 will be highly dependent on the context and type of the corpus used. The patterns and frequencies found will be corpus-syncratic showing trends that pertain to the particulars of the texts involved. To understand more fully how some of the patterns or features of a particular selection of texts realize within the larger scope of the language, a comparison search in a large corpus comprised of texts representative of many discourses, register, and genre styles can be very helpful.

**WebCorp**

The Research and Development Unit for English Studies at the University of Central London created WebCorp in 1999. As the name suggests, WebCorp is a web-based

program that uses the Internet as a mammoth corpus. The program relies on commercial search engines such as Google or Altavista (there are several from which the user can choose, though Google is the default setting) to comb the web for the entered query. The reliance on "outside" search engines causes the program to process requests rather slowly, though the creators are in the process of updating the program with their search engine which will likely speed up the processing time (Antoinette Renouf, personal communication, March 2005). Concordance lines, the keyword in context, and collocations, words with a statistical likelihood of appearing with or near the search word, are the primary functions of this corpus analysis tool, though the program can also generate word frequency lists for specific URLs. The program can be accessed on the Internet at www.webcorp.org.uk.

The concordancing feature of this program allows for both simple and advanced searches. By choosing the advanced search option the user can control a host of factors from the number of concordance lines displayed to filtering specific words from the search. The option to display collocates is also on the advanced search page. The page is laid out and labeled, though full explanations are available for all the features of the advanced search by clicking a labeled tab at the top of the page.

The benefit of WebCorp is the virtually limitless data available. The information on the web is uncensored and not compiled by a researcher. It is a result of the language used by the general population and crosses most genres and discourse types. The Internet was never intended as a corpus, and though arguably the Internet has created a sub-culture of its own from which patterns may emerge, by its sheer size it represents an authentic collection of naturally occurring language. In this way, WebCorp represents a corpus in direct opposition to a self-constructed corpus one can create with TextStat2. A corpus created using TextStat2 is a collection of carefully selected texts and likely to be intentionally limited in scope and type.

The results obtained from the analysis of such a purposefully constructed corpus can be used as search terms in WebCorp for comparison purposes. Comparing the results of a small corpus investigation with a corpus reflective of more general language use can reveal the typical usage of words or phrases that had notable frequencies within the self-constructed corpus. While WebCorp is a valuable resource for its size alone, sometimes word pattern frequencies are not completely transparent. The volume of data can make it difficult to sort through the results and have a quick picture of how the word typically functions in the language.

**WASPS**

WASPS, created in 1999 at the University of Brighton, is an Internet access program designed to show the patterns and categories of word use. The program is based on 100 million words of the British National Corpus ( BNC). Getting into the program can be a bit confusing. Registration of a username and password is required and once signed in, the user must click on the link to WASPBENCH to enter the program.

From here the program is very clear. The user enters a node word and selects the descriptor noun, verb, or adjective (these are the only searches available at this time) for the entered word. Controls for the amount of output displayed and the number and length of concordance lines (called examples from the BNC) can also be set on this page. The results are displayed on a horizontally split screen with the node word and its frequency

in the BNC at the top of the page. They are categorized by the node word's co-occurrence with various grammatical words, beginning with prepositions. Consider, for example, the verb search. WASPS shows 4106 occurrences of the verb in the BNC. Of those, 1621 occur with the preposition for in the pattern search. Listed beneath each preposition is the frequency for each content word that follows in the pattern. Under search for, the word occurs 46 times and by clicking on the highlighted frequency to the right of the word all 46 occurrences are displayed in concordance lines on the bottom half of the screen. The results also display information such as the modifiers, subjects, and objects used with the node word.

Using WASPS can give a quick picture of the function of a search word in a way that is not possible with WebCorp. However, WASPS is based on a closed and constructed, albeit, large, corpus. The program serves to give the grammatical patterns of words but does not reflect the unrestricted scope of WebCorp. Because of the different purposes served by WebCorp and WASPS, using them together can give a fuller picture of word function in the language. The search terms chosen for WebCorp, because of their significance in the small corpus constructed in TextStat2, could also be used in WASPS to show the typical patterning of these words. The results from WASPS can be cross-compared to both WebCorp and TextStat2 results to reveal patterns in genre or atypical uses based on context.

## Conclusion

A caveat for corpus linguistic research is necessary concerning the context and scope of the language used for analysis. No corpus represents an entire language and it is dangerous to make generalizations based on the results of a search in even the largest corpus. However, the inherent dependency on context can be a benefit as well. Much can be learned from an investigation into a small, specific, carefully constructed corpus about the frequencies and patterns particular to that type of language use. Often, these patterns are not easily visible in a very large and varied corpus due to the volume of language represented. Comparing the results of a small investigation with language use and patterns in a large corpus can help overcome these limitations.

TextStat, WebCorp, and WASPS are three programs with vastly different capabilities and purposes. Each program can be accessed for free and unlimited use. By using them in conjunction with one another, a researcher can fairly easily obtain a better understanding of language as it functions in both limited and controlled environments and more general use.

## Reference

Stubbs, M. (2002). Words and Phrases: Corpus studies of lexical semantics. Oxford, Blackwell.