# Review of software applications for deriving collocations

**Nikolaos K Anagnostou and George R S Weir**
**Department of Computer and Information Sciences**
**University of Strathclyde**
**Glasgow G1 1XH**

### 1. Introduction

The field of collocation extraction has enjoyed considerable growth and vitality from the 1990s onwards. Our research uncovered a multitude of software programs that can derive collocations from textual data, but also introduced the following question: Which one is the most fitting for the task of extracting collocations from a corpus?

This paper attempts to answer the previous question. We start by stating the criteria on which we based our judgement of the software applications included in our review. We then move on to give a brief description of each package, in terms of its functions, merits and demerits. We conclude by stating which of the packages was deemed, according to our opinion, the most appropriate for our purposes and provide a summary table of the results of the software review.

### 2. Criteria used for the review

To identify the most relevant tool, we compiled a list of criteria to help us in the selection and conducted a mini software review. We examined four software packages and evaluated them based on the following five criteria:

1. **Capacity to extract collocations without keywords**: Many programs implement what Evert (2004, p. 27) calls a "directional" view of collocations. In this approach, a keyword has to be initially specified in order for the program to identify its collocates. Our aim is different; our focus is not on the collocates of a specific word, but on the collocations of a corpus as a whole; extracting as many as possible is essential. As a result, this criterion has the greatest significance and its inclusion in the list was mandatory.
2. **Measures of association**: The option of choosing between several association measures allows for flexibility when extracting collocations and running comparative evaluation tests. Also, the measure that was found to be the most robust and accurate for collocation identification was Log-likelihood. For these reasons, the inclusion of Log-likelihood, accompanied by the support of multiple association measures, constitutes a meaningful selection criterion.
3. **Capacity to handle XML files**: Today, most corpora are annotated and for this reason come in XML instead of plain text format, since the former allows for metadata encoding. The corpus used in this project, the BNC, is also in XML format. It is apparent that the capacity of a program to manipulate corpora in such a format is an essential criterion.
4. **Capacity to extract multiword collocations**: Extracting collocations with more than two words is one of the targets of our research and consequently a very significant factor in the selection process.
5. **Capacity to search/handle multiple files at the same time**: Most corpora are sample-based, i.e. they consist of text samples of a certain size. In order to do large scale collocation extraction, a program has to be able to manipulate many or preferably all the text samples from which a corpus consists of. To account for this need, we included this factor in the reviewing process.
6. **Presence of a Graphical User Interface (GUI)**: It is a truism to say that the ease of use of a piece of software is directly related to whether it has a GUI or not. It allows quick familiarisation with a program's capabilities and working fast and effectively. Thus, it is an important selection criterion.

## 3. Software review

The programs included in the review are: WordSmith Tools 4 (WST 4), Collocate, Xaira and the Ngram Statistics Package (NSP). The first two are commercial solutions; Xaira and the NSP are open source and freeware. WST 4, Collocate and Xaira run on Windows platforms, and the NSP has been designed for Unix and Linux platforms, but is known to run on Windows as well. Some of these software packages are fully fledged concordancers[1], while others specialise solely on collocation extraction.

For demonstration and selection purposes, we performed two tasks (when the application had the capacity to do so): collocation extraction, both general and keyword-based; and concordancing. Unless explicitly stated otherwise, the corpus used for both tasks was the BNC Baby. Figures 1-8 provide screen shots of the packages we included in the software review.

It also has to be noted that the applications included here are by no means the only ones available for collocation extraction. Other examples of programs capable of extracting collocations are SENTA (by Gaël Dias), Kolokacje (by Aleksander Buczyński), the Multilingual Corpus Toolkit (by Scott Piao) and the Utilities for Cooccurrence Statistics (by Stefan Evert). The only reason for their exclusion was simply lack of time.

### *3.1 WordSmith Tools 4*

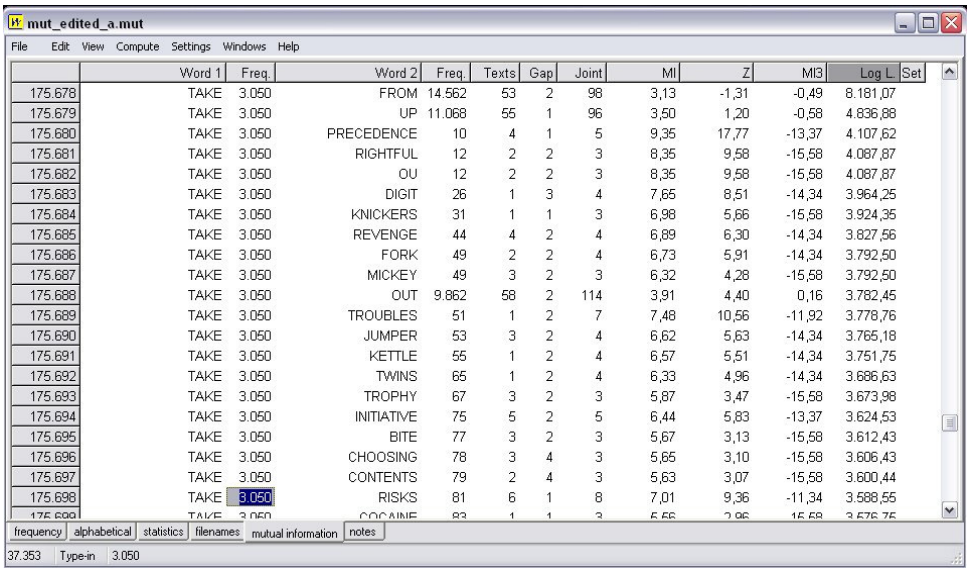
Developed by Mike Scott, this software package is published by the Oxford University Press. It provides a wide range of functions relevant to corpus linguistics in the form of an all-in-one suite. The package costs approximately £50 (US$92 or €75) for a single user licence and £260 (US$460, €376) for a 10-user licence.

WST 4 is without doubt the most versatile of the applications included in this review. Its functions are grouped in three main categories: Concord, Keywords and Wordlist. As the names of the categories suggest, the program can create concordances, perform keyword analyses and compile word-frequency lists. Another notable feature of the program is the WebGetter tool, which allows for on-the-fly creation of corpora from the Internet, based on a number of parameters, including languages to be considered. The package is accompanied by a thorough and informative manual.

WST 4 is capable of extracting collocations, both general and keyword-based, and implements four association measures to compute them, namely the MI, Z-score, MI3 (i.e. MI cubed) and Log-likelihood. It can also handle multiple files and supports XML, but shows an affinity for plain text formats. Some of its demerits include an unintuitive GUI, which takes a bit of time to get used to, along with lack of support for multiword collocation extraction and regular expressions, which can be very useful for extracting collocations based on syntactic patterns.

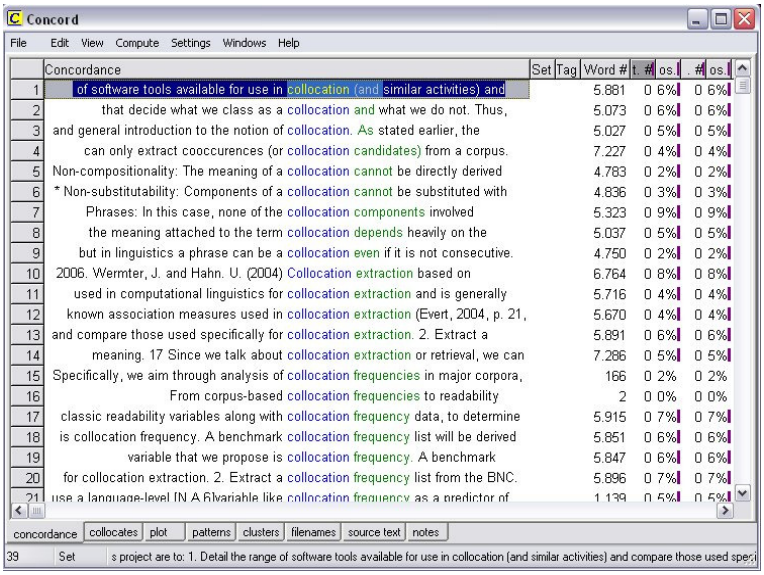
No evaluation versions of WST 4 are presently available. The software can be purchased directly from the OUP website (http://www.oup.co.uk/episbn/0-19-459400-9).

[1] A software package capable of producing concordances, i.e. all the occurrences of a keyword in its context (KWIC) in a corpus. For readers interested in concordancing software, a core tool for corpus manipulation, Wiechmann and Fuhs (2006) provide an excellent and quite detailed review of ten of the most popular concordancers. We are indebted to the authors for providing a pre-draft version of their parer.




| | Word 1 | Freq. | Word 2 | Freq. | Texts | Gap | Joint | MI | Z | MI3 | Log L | Set |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 175.678 | TAKE | 3.050 | FROM | 14.562 | 53 | 2 | 98 | 3,13 | -1,31 | -0,49 | 8.181,07 | |
| 175.679 | TAKE | 3.050 | UP | 11.068 | 55 | 1 | 96 | 3,50 | 1,20 | -0,58 | 4.836,88 | |
| 175.680 | TAKE | 3.050 | PRECEDENCE | 10 | 4 | 1 | 5 | 9,35 | 17,77 | -13,37 | 4.107,62 | |
| 175.681 | TAKE | 3.050 | RIGHTFUL | 12 | 2 | 2 | 3 | 8,35 | 9,58 | -15,58 | 4.087,87 | |
| 175.682 | TAKE | 3.050 | OU | 12 | 2 | 2 | 3 | 8,35 | 9,58 | -15,58 | 4.087,87 | |
| 175.683 | TAKE | 3.050 | DIGIT | 26 | 1 | 3 | 4 | 7,65 | 8,51 | -14,34 | 3.964,25 | |
| 175.684 | TAKE | 3.050 | KNICKERS | 31 | 1 | 1 | 3 | 6,98 | 5,66 | -15,58 | 3.924,35 | |
| 175.685 | TAKE | 3.050 | REVENGE | 44 | 4 | 2 | 4 | 6,89 | 6,30 | -14,34 | 3.827,56 | |
| 175.686 | TAKE | 3.050 | FORK | 49 | 2 | 2 | 4 | 6,73 | 5,91 | -14,34 | 3.792,50 | |
| 175.687 | TAKE | 3.050 | MICKEY | 49 | 3 | 2 | 3 | 6,32 | 4,28 | -15,58 | 3.792,50 | |
| 175.688 | TAKE | 3.050 | OUT | 9.862 | 58 | 2 | 114 | 3,91 | 4,40 | 0,16 | 3.782,45 | |
| 175.689 | TAKE | 3.050 | TROUBLES | 51 | 1 | 2 | 7 | 7,48 | 10,56 | -11,92 | 3.778,76 | |
| 175.690 | TAKE | 3.050 | JUMPER | 53 | 3 | 2 | 4 | 6,62 | 5,63 | -14,34 | 3.765,18 | |
| 175.691 | TAKE | 3.050 | KETTLE | 55 | 1 | 2 | 4 | 6,57 | 5,51 | -14,34 | 3.751,75 | |
| 175.692 | TAKE | 3.050 | TWINS | 65 | 1 | 2 | 4 | 6,33 | 4,96 | -14,34 | 3.686,63 | |
| 175.693 | TAKE | 3.050 | TROPHY | 67 | 3 | 2 | 3 | 5,87 | 3,47 | -15,58 | 3.673,98 | |
| 175.694 | TAKE | 3.050 | INITIATIVE | 75 | 5 | 2 | 5 | 6,44 | 5,83 | -13,37 | 3.624,53 | |
| 175.695 | TAKE | 3.050 | BITE | 77 | 3 | 2 | 3 | 5,67 | 3,13 | -15,58 | 3.612,43 | |
| 175.696 | TAKE | 3.050 | CHOOSING | 78 | 3 | 4 | 3 | 5,65 | 3,10 | -15,58 | 3.606,43 | |
| 175.697 | TAKE | 3.050 | CONTENTS | 79 | 2 | 4 | 3 | 5,63 | 3,07 | -15,58 | 3.600,44 | |
| 175.698 | TAKE | 3.050 | RISKS | 81 | 6 | 1 | 8 | 7,01 | 9,36 | -11,34 | 3.588,55 | |

**Figure 1. Using WST 4 to extract a list of two-word collocations. The figure depicts some of the collocations of "take". Results ranked by the Log-likelihood measure.**




| | Concordance | Set | Tag | Word # | t. # | os. | . # | os. |
|---|---|---|---|---|---|---|---|---|
| 1 | of software tools available for use in collocation (and similar activities) and | | | 5.881 | 0 | 6% | 0 | 6% |
| 2 | that decide what we class as a collocation and what we do not. Thus, | | | 5.073 | 0 | 6% | 0 | 6% |
| 3 | and general introduction to the notion of collocation. As stated earlier, the | | | 5.027 | 0 | 5% | 0 | 5% |
| 4 | can only extract cooccurences (or collocation candidates) from a corpus. | | | 7.227 | 0 | 4% | 0 | 4% |
| 5 | Non-compositionality: The meaning of a collocation cannot be directly derived | | | 4.783 | 0 | 2% | 0 | 2% |
| 6 | * Non-substitutability: Components of a collocation cannot be substituted with | | | 4.836 | 0 | 3% | 0 | 3% |
| 7 | Phrases: In this case, none of the collocation components involved | | | 5.323 | 0 | 9% | 0 | 9% |
| 8 | the meaning attached to the term collocation depends heavily on the | | | 5.037 | 0 | 5% | 0 | 5% |
| 9 | but in linguistics a phrase can be a collocation even if it is not consecutive. | | | 4.750 | 0 | 2% | 0 | 2% |
| 10 | 2006. Wermter, J. and Hahn. U. (2004) Collocation extraction based on | | | 6.764 | 0 | 8% | 0 | 8% |
| 11 | used in computational linguistics for collocation extraction and is generally | | | 5.716 | 0 | 4% | 0 | 4% |
| 12 | known association measures used in collocation extraction (Evert, 2004, p. 21, | | | 5.670 | 0 | 4% | 0 | 4% |
| 13 | and compare those used specifically for collocation extraction. 2. Extract a | | | 5.891 | 0 | 6% | 0 | 6% |
| 14 | meaning. 17 Since we talk about collocation extraction or retrieval, we can | | | 7.286 | 0 | 5% | 0 | 5% |
| 15 | Specifically, we aim through analysis of collocation frequencies in major corpora, | | | 166 | 0 | 2% | 0 | 2% |
| 16 | From corpus-based collocation frequencies to readability | | | 2 | 0 | 0% | 0 | 0% |
| 17 | classic readability variables along with collocation frequency data, to determine | | | 5.915 | 0 | 7% | 0 | 7% |
| 18 | is collocation frequency. A benchmark collocation frequency list will be derived | | | 5.851 | 0 | 6% | 0 | 6% |
| 19 | variable that we propose is collocation frequency. A benchmark | | | 5.847 | 0 | 6% | 0 | 6% |
| 20 | for collocation extraction. 2. Extract a collocation frequency list from the BNC. | | | 5.896 | 0 | 7% | 0 | 7% |
| 21 | use a language-level [N A 6]variable like collocation frequency as a predictor of | | | 1.139 | 0 | 5% | 0 | 5% |

**Figure 2. Using WST 4 to find the concordance of "collocation"[2]. The results are sorted alphabetically, based on the first word on the right of the search word.**

### *3.2 Collocate*

Developed by Michael Barlow, this package is published by Athelstan. It specialises in collocation extraction and costs approximately £24 (US$45, €35) for a single user licence or £185 (US$350, €275) for a 15-user licence.

Collocate is a tool focused on collocation extraction and in this field it excels. Its two main functions are: Extract and Full Extract. They used for deriving keyword-based and general collocations respectively. The package provides several options for collocation investigation including word/phrase search, regular expressions and word/tag search. Is also allows for n-gram extraction ($n_{max}$ = 6) and implements three association measures, T-score, MI and Log-likelihood, to determine their collocation strength.

[2] For this task we used the paper "From corpus-based collocation frequencies to readability measure" Anagnostou and Weir (the present volume).

The application has an easy to use GUI, with all the main functions accessible for the menu toolbar. It can manipulate multiple files and has good support for XML. On the downside, since Collocate specialises in collocation extraction, we believe that it should implement more association measures. Nevertheless, it is a well-rounded, capable and user-friendly collocation extraction tool.

Collocate can be ordered from the web and a demo version is also available (http://athel.com/product_info.php?products_id=29&osCsid=47a7029dbe235029ef686d7db90df9d2).

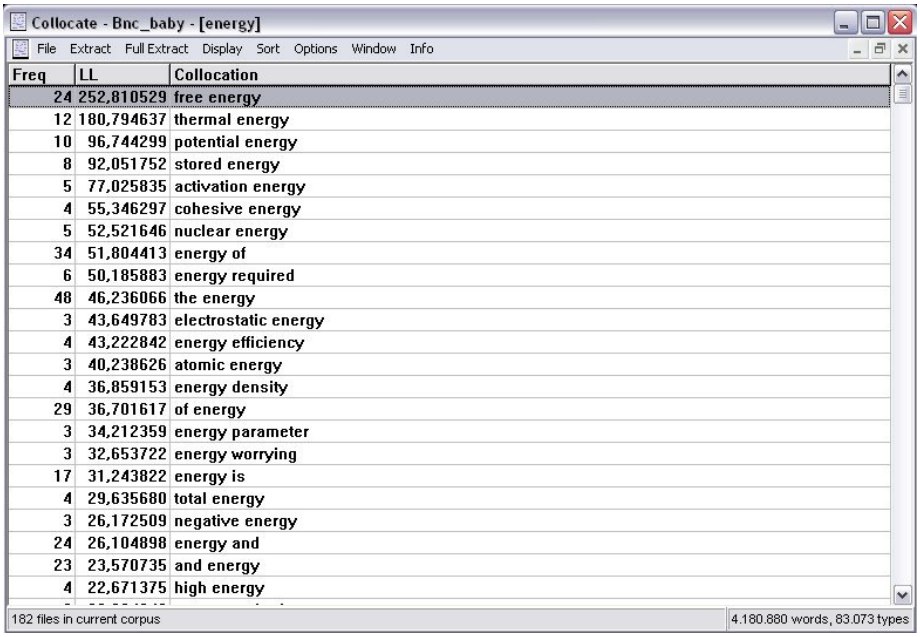



**Figure 3. Using Collocate to extract two-word collocations of "energy". Results ranked by the Log-likelihood measure.**

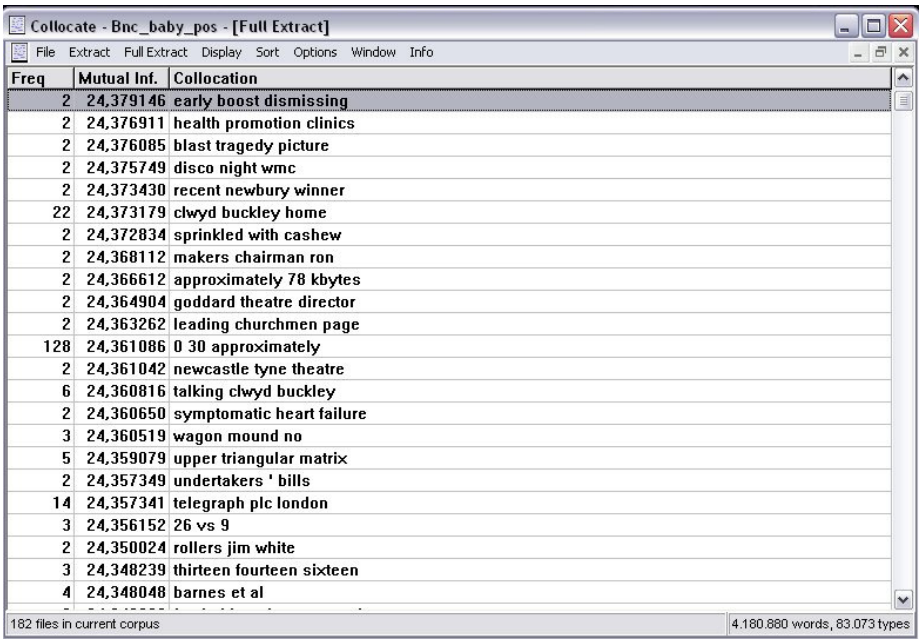



**Figure 4. Using Collocate for multiword collocation extraction.**

### *3.3 Xaira*

Xaira was developed by Lou Burnard and Tony Dodd, and is distributed by the Research Technologies Service at Oxford University Computing Services. The package is available for free, along with its source code, and it is bundled with the BNC.

Xaira is as a general purpose XML search engine. To search a corpus using Xaira, the corpus must first be indexed. These indexes are created by a separate program called the Xaira indexer. The program can generate concordances and extract collocations, albeit only keyword-based ones. It implements only one association measure, the Z-score. Xaira's true strength lies in the variety of ways it provides for searching a corpus. A user can search for words, patterns (i.e. regular expressions), specific XML tags or even run queries based on XQL, a language Xaira utilises to internally represent queries.

Xaira has a quite well organised GUI, but the user needs to spend some time reading the help file, to get accustomed to the program's idiosyncrasies and the syntax of the various search tools. As expected, the package has excellent support for XML. It handles multiple files indirectly, through the initial indexing process. In a nutshell, Xaira is a powerful tool for corpus interrogation but not for collocation extraction.

Xaira can be downloaded from Sourceforge, at the following URL:
http://sourceforge.net/project/showfiles.php?group_id=130289

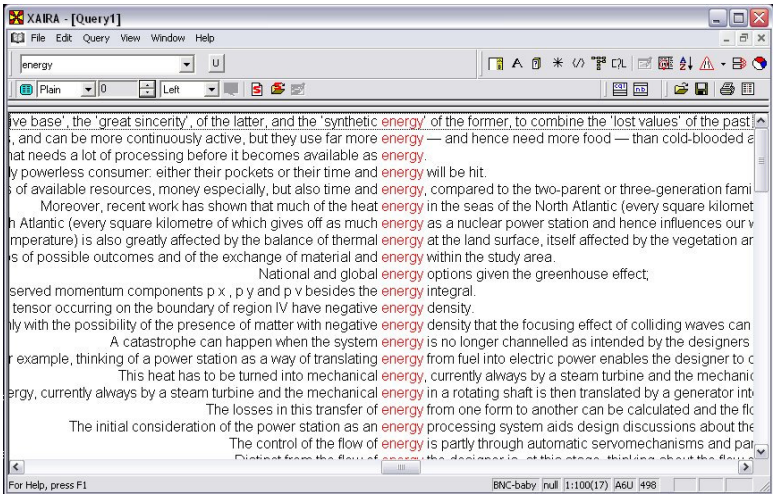



**Figure 5. Using Xaira to find the concordance of "energy".**

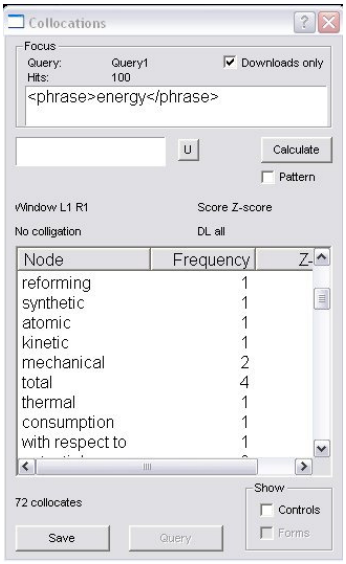



**Figure 6. Using Xaira to find the collocates of "energy". The words in the "Node" column are collocates of "energy", in a (-1,+1) collocational window.**

### *3.4 Ngram Statistics Package*

This software package is a collaborative effort, with the main developers being Ted Pedersen and Satanjeev Banerjee[3]. The package is open-source, programmed in Perl and can be described as a suite of programs for n-gram analysis of text files.

The NSP consists of two main programs: count.pl and statistics.pl. The first program takes as input text files and produces a list of n-grams occurring in these files. The second program takes as input the aforementioned n-gram list and runs an association measure selected by the user, in order to determine which of the n-grams can be considered as collocations. Out of all the programs included in this review, the NSP implements the largest collection of association measures. Examples include Log-likelihood, MI (true and pointwise), Poisson Stirling, $X^2$, T-score and more[4]. The package is a tool specialising in multiword collocation extraction and thus has no support for keyword-based extraction or any of the additional characteristics found in the previously reviewed packages. Finally, the NSP provides strong support for regular expressions.

.

The NSP's functions are accessed through a command line interface, with no GUI in place. Consequently, for users not accustomed to a command prompt, it is the most difficult to use from the packages reviewed here. It can process multiple files simultaneously, with the restriction of all of them being in the same directory. Also, the package cannot understand XML and the n-gram lists it produces are saved in a hazy format. All in all, in the confines of this software review, the NSP is the most capable tool for multiword collocation extraction, albeit a bit unwieldy.

The NSP is available for download at the following URL: http://ngram.sourceforge.net/

[3] See http://www.d.umn.edu/~tpederse/nsp.html for the full team.
[4] See final table for the full range of association measures implemented by the NSP.

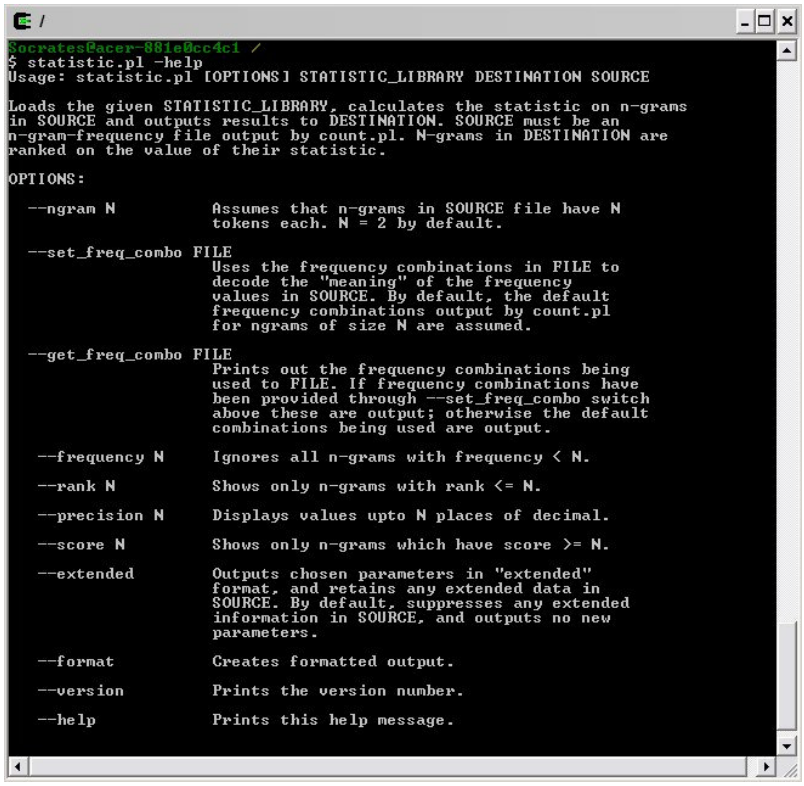



**Figure 7. List of some of the available command line arguments for the statistic.pl utility of the NSP[5].**

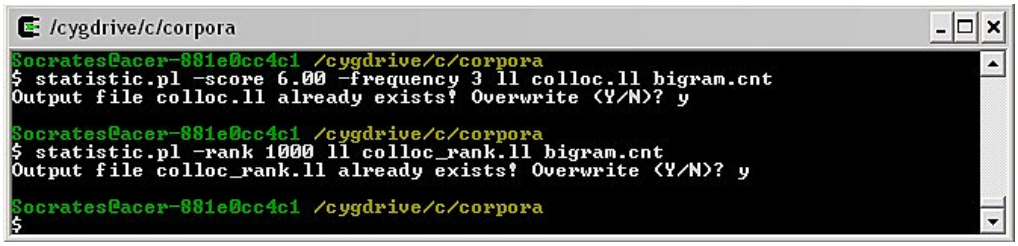



**Figure 8. Examples of usage for the statistic.pl utility. The first command creates a list of bigrams ranked by Log-likelihood ratios. It includes only those with scores of 6.00 or better among bigrams that occur more 3 or more times. The second command creates a list of the top 1000 bigrams as ranked by the Log-likelihood ratio. The file bigram.cnt is the input (created with the count.pl utility) and the files colloc.ll, colloc_rank.ll are the outputs for each command respectively.**

### 3. Results of the software review

Table 1 summarises the results of this software review. Out of the four packages included here, three were capable of extracting collocations without keywords, which stands as our most important selection criterion. Each of these has its advantages and disadvantages. WordSmith Tools is the best-rounded tool, offering a wealth of functions for corpus investigation, but cannot extract multiword collocations. The Ngram Statistics Package is capable of doing this and also has the largest collection of association measures, but lacks support for XML and produces cluttered results. Collocate is the only package that meets all of our selection criteria, but has a small collection of association measures. Based on these results, we decided that, at our present stage of research, the application most appropriate for our purposes is Collocate. We do have to state though that should a work-around of NSP's drawbacks be found, we are ready to reconsider.

[5] We ran the utility using Cygwin emulation (a Linux-like environment in Windows).

**Table 1. Summary of the software review.**

| | | Criteria | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Capacity to extract collocations without keywords** | **Measures of association** | **Capacity to handle xml files** | **Capacity to extract multiword collocations** | **Capacity to handle multiple files at the same time** | **Presence of a Graphical User Interface (GUI)** |
| **Software Packages** | **Wordsmith Tools 4** | ✓ | MI, **Log-likelihood**, MI3, Z-score | ✓ | × | ✓ | ✓ |
| | **Collocate** | ✓ | MI, **Log-likelihood**, T-score | ✓ | ✓[6] | ✓ | ✓ |
| | **Xaira** | × | Z-score | ✓ | × | ✓ | ✓ |
| | **Ngram Statistics Package** | ✓ | **Log-likelihood**, MI (true and pointwise), Poisson Stirling, $X^2$, T-score, Phi, Dice, Jaccard, Fisher's exact tests (left, right, two-tailed), Odds ratio | × | ✓ | ✓ | × |

(✓ = Yes, × = No)

[6] For collocations longer than two words, Collocate uses the MI association measure or the cost criterion (Kita et al., 1994).

## 4. Conclusion

This paper described the software review we conducted, in order to find the most appropriate package for collocation extraction. We tried, within the available time limits, to unveil the potential of four tools identified as possible candidates for the task above. To do this, we defined a list of six criteria, specifically tailored to the needs of our research. Consequently, we disregarded other factors commonly included in software reviews like speed of execution, resource management issues, sorting and display capabilities etc., considering them to be of lesser importance. The applications were judged along these criteria and a brief description of each program's functions, strong and weak points was given. The present review showed that the package that stands out most strongly is Collocate, closely followed by the NSP.

## References

**Anagnostou, N. and Weir, G. R. S. (2006).** 'From corpus-based collocation frequencies to readability measure', this volume.

**Evert, S. (2004).** The Statistics of Word Cooccurrences (Word Pairs and Collocations). Ph. D. dissertation, Universitat Stuttgart.

**Pedersen, T. and Banerjee, S. (2003)**. The Design, Implementation and Use of the Ngram Statistics Package. In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City.

**Wiechmann, D. and Fuhs, S. (2006)** Concordance Software. Corpus Linguistics and Linguistics Theory 2-1, 109-130