Comparing corpora with WordSmith Tools: How large must the reference corpus be?

Tony BERBER-SARDINHA
LAEL, Catholic University of São Paulo
Rua Monte Alegre 984
05014-001 São Paulo SP, Brazil
tony4@uol.com.br

Abstract
WordSmith Tools (Scott, 1998) offers a program for comparing corpora, known as KeyWords. KeyWords compares a word list extracted from what has been called ‘the study corpus’ (the corpus which the researcher is interested in describing) with a word list made from a reference corpus. The only requirement for a word list to be accepted as reference corpus by the software is that must be larger than the study corpus. One of the most pressing questions with respect to using KeyWords seems to be what would be the ideal size of a reference corpus. The aim of this paper is thus to propose answers to this question. Five English corpora were compared to reference corpora of various sizes (varying from two to 100 times larger than the study corpus). The results indicate that a reference corpus that is five times as large as the study corpus yielded a larger number of keywords than a smaller reference corpus. Corpora larger than five times the size of the study corpus yielded similar amounts of keywords. The implication is that a larger reference corpus is not always better than a smaller one, for WordSmith Tools Keywords analysis, while a reference corpus that is less than five times the size of the study corpus may not be reliable. There seems to be no need for using extremely large reference corpora, given that the number of keywords yielded do not seem to change by using corpora larger than five times the size of the study corpus.

Introduction
WordSmith Tools (Scott, 1998) offers a program for comparing corpora, known as KeyWords. This tool has been used in several studies as a means for describing various lexicogrammatical characteristics of different genres (Barbara and Scott, 1999; Batista, 1998; Berber Sardinha, 1995, 1999a, b; Berber Sardinha and Shimazumi, 1998; Bonamin, 1999; Collins and Scott, 1996; Conde, 1999; Dutra, 1999; Freitas, 1997; Fuzetti, 1999; Granger and Tribble, 1998; Lima-Lopes, 1999; Lopes, 2000; Ramos, 1997; Santos, 1999; Scott, 1997; Silva, 1999; Tribble, 1998). The keywords identified by the program are not necessarily the ‘most important words’ in the corpus (Scott, 1997), or those that correspond to readers’ intuitions as to what the topics of the texts are. It is generally thought that a set of WordSmith Tools keywords indicate ‘aboutness’ (Phillips, 1989).

KeyWords compares a word list extracted from what has been called ‘the study corpus’ (the corpus which the researcher is interested in describing) with a word list made from a reference corpus. The result is a list of keywords, or words whose frequencies are statistically higher in the study corpus than in the reference corpus. The software also identifies words whose frequencies are statistically lower in the study corpus, which are called ‘negative keywords’, in contrast to positive keywords, which have higher frequencies in the study corpus. Negative keywords, though, will not be discussed in the present paper. Hence, whenever keyword is mentioned in this paper, it will mean ‘positive keyword’.

The only requirement for a word list to be accepted as reference corpus by the software is that must be larger than the study corpus. Thus, the composition and length of KeyWord lists can vary according to at least six parameters:

• The composition of the study corpus.
• The composition of the reference corpus.
• The size of the study corpus.
• The size of the reference corpus.
• The statistical test used in the comparison of frequencies (log-likelihood and chi-square are available).
• The level of significance (p) used as the ‘keyness’ benchmark (the cut-off point).

Since WordSmith Tools is Windows software, it has appealed to a large audience of applied linguists willing to do corpus-based research, to whom this platform is generally the only one that they know how to use. To them, one of the most pressing questions with respect to using KeyWords seems to be what would be the ideal size of a reference corpus. The aim of this paper is thus to propose answers to this question.

1 Using KeyWords

A KeyWord list is a portion of the study corpus word list. KeyWords compares the frequencies for each type in the study and reference corpora. The program calculates the log-likelihood (G²) or Chi-Square (X²) of each word form based on its distribution in both corpora, an example of which is given in the table below.

| Word form | Remaining word forms | Total |
|-----------|----------------------|-------|
| Study corpus | 10 (10%) | 90 (90%) | 100 (100%) |
| Reference corpus | 10 (1%) | 1000 (99%) | 1010 (100%) |

For a distribution such as the above, both the log-likelihood and chi-square statistics would probably flag the word form in question as a keyword, since its frequencies in the two corpora are so different (10% versus 1%). The way KeyWords processes word lists is not unique, and has been applied by researchers using other software (De Cock, Granger, Leech, and McEnery, 1998; Granger and Rayson, 1998; Milton, 1998).

After processing the word lists, the keyword lists appear in WordSmith Tools as illustrated below.

1 See Dunning (1992) for the formulae.
The reference corpora were compiled out of texts published in 'The Guardian'. The reason for choosing it is that newspaper text is the most typical kind of reference corpus used by applied linguists, mainly because it is easy to get. Therefore, the results obtained here would be relevant to the typical user of KeyWords. The reason for specifically choosing the Guardian is that Mike Scott, the author of WordSmith Tools, makes it available on his website a word list of 95 million tokens of The Guardian text on his website. This has become a popular choice for several WordSmith Tools users investigating English keywords. Once again, it was hoped that by using The Guardian, the investigation would mirror a typical choice of WordSmith users. For the present study, a portion of the Guardian word list was used, namely from texts published in 1994, taken randomly.

The size of the reference corpora varied according to the size of the study corpora. For each study corpus, 18 reference corpora were created. Each one was n times larger than the study corpus, with n being 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. For instance, the letters corpus had 11,761 tokens, and so for n=2 the size of the reference corpus was 23,522 tokens (11,761 * 2); for n=3, the reference corpus size was 35,283 (11,761 x 3), for n=4 47,044, and so on, up to n=100, whose size was 1,176,100 words.

The KeyWords settings used for the comparisons were as follows:

| Setting            | Value     |
|--------------------|-----------|
| Procedure          | loglikelihood |
| Max p. value       | 0.01      |
| Max wanted         | 16000*    |
| Min frequency      | 2         |

Table 1: KeyWords settings

The table below shows the size of all of the reference corpora used in the study:

| Size of reference corpora | N=2 | n=3 | n=4 | n=5 | N=6 | N=7 |
|---------------------------|-----|-----|-----|-----|-----|-----|
| Letters                   | Tokens 23,522 35,283 47,044 | 58,805 | 70,566 | 82,327 |
|                           | Types 5,543 7,409 8,863 | 10,161 | 11,163 | 12,249 |
| Editorials                | Tokens 109,252 163,878 218,504 | 273,130 | 327,756 | 382,382 |
|                           | Types 14,973 18,378 21,746 | 24,118 | 26,537 | 28,382 |
| Reviews                   | Tokens 71,482 107,223 142,964 | 218,964 | 273,130 | 327,756 |
|                           | Types 11,000 14,331 17,758 | 19,490 | 21,559 | 23,402 |
| Mystery                   | Tokens 96,596 144,894 193,192 | 241,490 | 289,788 | 338,086 |
|                           | Types 13,880 17,636 20,285 | 22,861 | 24,925 | 26,928 |
| Sci-Fi                    | Tokens 24,162 36,243 48,324 | 60,405 | 72,486 | 84,567 |
|                           | Types 5,644 7,550 9,032 | 10,325 | 11,318 | 12,422 |

| Size of reference corpora | n=8 | n=9 | n=10 | n=20 | n=30 | n=40 |
|---------------------------|-----|-----|------|------|------|------|
| Letters                   | Tokens 94,088 105,849 117,610 | 235,220 | 352,830 | 470,440 |
|                           | Types 13,095 13,896 14,879 | 22,650 | 27,763 | 31,471 |
| Editorials                | Tokens 437,008 491,634 546,260 | 1092,520 | 1,638,780 | 2,185,040 |
|                           | Types 30,292 31,825 33,672 | 47,305 | 57,325 | 65,237 |
| Reviews                   | Tokens 285,928 321,669 357,410 | 714,820 | 1,072,230 | 1,429,640 |
|                           | Types 24,940 26,524 27,812 | 38,610 | 47,081 | 53,695 |
| Mystery                   | Tokens 386,384 434,682 482,980 | 965,960 | 1,448,940 | 1,931,920 |
|                           | Types 28,563 30,084 31,669 | 44,755 | 53,867 | 61,531 |
| Sci-Fi                    | Tokens 96,648 108,729 120,810 | 241,620 | 362,430 | 483,240 |
|                           | Types 13,305 14,209 15,156 | 22,918 | 28,144 | 32,010 |
Table 2: Size of reference corpora

| n=  | Letters Tokens | Letters Types | Editorials Tokens | Editorials Types | Reviews Tokens | Reviews Types | Mystery Tokens | Mystery Types | Sci-Fi Tokens | Sci-Fi Types |
|-----|----------------|---------------|-------------------|------------------|---------------|---------------|----------------|---------------|---------------|-------------|
| 50  | 588,050        | 35,083        | 2,731,300         | 71,680           | 1,787,050     | 59,690        | 2,414,900      | 68,117        | 604,050       | 35,460      |
| 60  | 705,660        | 38,560        | 3,277,500         | 77,397           | 2,144,460     | 64,753        | 2,897,880      | 73,623        | 724,860       | 38,959      |
| 70  | 823,270        | 42,421        | 3,823,820         | 82,743           | 2,501,870     | 69,242        | 3,380,860      | 78,508        | 845,670       | 42,822      |
| 80  | 940,880        | 44,607        | 4,370,080         | 87,902           | 2,859,280     | 73,167        | 3,863,840      | 83,076        | 966,480       | 45,101      |
| 90  | 1,058,490      | 47,067        | 4,916,340         | 92,884           | 3,216,690     | 76,945        | 4,346,820      | 87,578        | 1,087,290     | 47,474      |
| 100 | 1,176,100      | 48,902        | 5,462,600         | 97,121           | 3,574,100     | 92,157        | 4,829,800      | 97,121        | 1,208,100     | 49,617      |

Table 3: Keyword totals (% = pct. of the total number of types in the study corpus).

| n=  | Letters Keywds. | Letters % | Editorials Keywds. | Editorials % | Reviews Keywds. | Reviews % | Mystery Keywds. | Mystery % | Sci-Fi Keywds. | Sci-Fi % |
|-----|-----------------|-----------|--------------------|--------------|----------------|-----------|----------------|-----------|----------------|---------|
| 2   | 279             | 11.6      | 433                | 5.0          | 401            | 5.2       | 583            | 9.3       | 137            | 4.6     |
| 3   | 347             | 14.4      | 686                | 8.0          | 582            | 7.5       | 748            | 11.9      | 203            | 6.8     |
| 4   | 354             | 14.7      | 637                | 7.4          | 496            | 6.4       | 728            | 11.6      | 196            | 6.6     |
| 5   | 481             | 19.9      | 963                | 11.2         | 889            | 11.5      | 1027           | 16.4      | 363            | 12.2    |
| 6   | 480             | 19.9      | 910                | 10.6         | 872            | 11.3      | 1035           | 16.5      | 361            | 12.1    |
| 7   | 450             | 18.6      | 892                | 10.4         | 829            | 10.7      | 1018           | 16.2      | 355            | 11.9    |
| 8   | 457             | 18.9      | 887                | 10.3         | 846            | 10.9      | 1037           | 16.5      | 350            | 11.7    |
| 9   | 457             | 18.9      | 880                | 10.3         | 822            | 10.6      | 1031           | 16.4      | 332            | 11.1    |
| 10  | 462             | 19.1      | 896                | 10.4         | 837            | 10.8      | 1050           | 16.7      | 330            | 11.1    |
| 20  | 506             | 21.0      | 967                | 11.3         | 935            | 12.1      | 1119           | 17.8      | 353            | 11.8    |
| 30  | 497             | 20.6      | 960                | 11.2         | 919            | 11.9      | 1116           | 17.8      | 364            | 12.2    |
| 40  | 507             | 21.0      | 953                | 11.1         | 926            | 12.0      | 1135           | 18.1      | 367            | 12.3    |
| 50  | 490             | 20.3      | 936                | 10.9         | 914            | 11.8      | 1123           | 17.9      | 373            | 12.5    |
| 60  | 492             | 20.4      | 942                | 11.0         | 933            | 12.0      | 1141           | 18.2      | 378            | 12.7    |
| 70  | 492             | 20.4      | 928                | 10.8         | 914            | 11.8      | 1140           | 18.1      | 368            | 12.3    |
| 80  | 485             | 20.1      | 948                | 11.0         | 929            | 12.0      | 1145           | 18.2      | 374            | 12.5    |
| 90  | 485             | 20.1      | 943                | 11.0         | 922            | 11.9      | 1130           | 18.0      | 383            | 12.8    |
| 100 | 475             | 19.7      | 952                | 11.1         | 939            | 12.1      | 1143           | 18.2      | 382            | 12.8    |

The results indicate that the number of keywords increases as the size of the reference corpus increases, but this increase is not linear. For instance, the keywords for n=2 in the letters corpus was 279, for n=3 it was 347, and for n=100 the total keywords was 475. Had the growth been linear, for n=3 there would be 418 keywords, and for n=100 13,950. Obviously, a total of 13,950 keywords could never have been obtained since the maximum possible number of
keywords in the letters corpus is 2,415, which is the total number of types. The same is true of all the other corpora.

This suggests that there must be a point at which the growth in number of keywords diminishes. This can be shown by plotting the number of keywords for each size of n across all the study corpora, as in the graph below.

Plot 1: Distribution of keywords

The plot shows that for all study corpora the keyword totals rose from n=2 to n=3, then fell or stabilized at n=4, rose again at n=5 and from then on basically reached a plateau. For instance, for the letters corpus, the keyword totals for n=2, n=3, n=4, n=5, and n=6 were respectively 11.6, 14.4, 14.7, 19.9, and 19.9. Hence, there was indeed a considerable rise from n=2 to n=3 (11.6 to 14.4), followed by a slight rise at n=4 (14.7), then a major increase at n=5 (19.9), and there was no change from n=5 to n=6 (19.9 to 19.9).

In order to check where the major changes occurred, an ANOVA was run on the keyword totals across the various n sizes. The results are shown in the table below.

| Source | df | SS      | F       | p     |
|--------|----|---------|---------|-------|
| Size of n | 21 | 1540.8087 | 267.98 | <0.0001 |
| Error | 68 | 18.6184 |         |       |
| Total | 89 | 1559.4271 |         |       |

Table 4: Results of ANOVA for keyword totals across reference corpora

The value of $F(21,68)=267.98$ is significant at $p<0.0001$, which indicates that size of the reference corpora had a significant effect on the keyword totals. This does not show us the differences in keyword totals among n sizes.

In order to know at which n sizes the keyword totals are statistically different, the REGWF (Ryan-Einot-Gabriel-Welsch) Multiple F Test was run in SAS. The results appear in the table below, in decreasing order of the average percentage of keyword totals across the five study corpora.

| Groupings | Avg. % keywords | Size of n |
|-----------|-----------------|-----------|
| A         | 14.8840         | 40        |
| A         | 14.8480         | 60        |
| A         | 14.7900         | 20        |
| A         | 14.7780         | 100       |
| A         | 14.7780         | 80        |
| A         | 14.7600         | 90        |
| A         | 14.7220         | 30        |
| A B       | 14.6940         | 70        |
| A B C     | 14.6780         | 50        |
| A B C D   | 14.2280         | 5         |
| A B C D   | 14.0660         | 6         |
| B C D     | 13.6860         | 8         |
| C D       | 13.6340         | 10        |
| D         | 13.5660         | 7         |
| D         | 13.4640         | 9         |
| E         | 9.100           | 3         |
| E         | 9.3280          | 4         |
| F         | 7.1300          | 2         |

Table 5: Results of REGWF test

The REGWF test presents the results in terms of groupings, identified by letters. Keyword totals in the same grouping are not statistically different. Hence, sizes of n equal to 40, 60, 20, 100, 80, 90, 30, 70, 50, 5, and 6 formed grouping A, which has on average 14.066% to 14.884% keyword totals. Likewise, n sizes equal to 70, 50, 5, 6, and 8 were in grouping B, with averages ranging from 13.68% to 14.694%.

Note that this is overlap among groupings, and so groupings A, B, C and D are in fact joined. This grouping comprises n sizes ranging from 5 to 100. The remaining groupings are non-overlapping: grouping E was formed by n sizes 3 and 4, and grouping F by n=2.

Therefore, there are two basic divisions in the previous table, namely at n sizes equal to 2, 3,
and 5. These correspond to the major peaks and plateaus visible in the plot.

The results suggest, then, that the critical value for a reference corpus seems to be five. In other words, the answer to the question 'what is the ideal size of a reference corpus' is five. A reference corpus that is five times as large as the study corpus yields a larger number of keywords than a smaller reference corpus. This means that the results of a keyword analysis based on a reference corpus that is less than five times the size of the study corpus could be very different from a study done on a corpus, say, just three times larger than the study corpus, in so far as the number of keywords go. Several potentially revealing keywords could be left out of the analysis if the reference corpus is not as large as five times or more.

Conclusion

The aim of this study was to estimate the ideal size of a reference corpus to be used in WordSmith Tools KeyWords procedure. KeyWords provides facilities for comparing a study corpus to a reference corpus, which, by default, must be larger than study corpus.

The results indicated that a reference corpus that is five times larger than the study corpus yields a similar amount of keywords than reference corpora that are up to 100 times larger than the study corpus. This was taken to mean that a reference corpus does not need to be more than five times larger than the study corpus.

In sum, a larger reference corpus is not always better than a smaller one, for WordSmith Tools Keywords analysis. There seems to be no need for using extremely large reference corpora, given that the number of keywords yielded do not seem to change by using corpora larger than five times the size of the study corpus. This may be important for WordSmith Tools users, who may be short of disk space and memory on their PCs to process large reference corpora. A suggestion that might come out of this finding is that researchers should not spend time and resources building, collecting or searching for larger and larger reference corpora. Resources would be better spent in the compilation of reference corpora that are more suitable in terms of their contents viz à viz the study corpus.

This study did not tackle several important questions. One of them is whether the keywords that were identified represent the main concepts or topics found the texts. A qualitative study would be needed to answer this, as an independent test of validity of the status of the keywords. Another question is the effect of the size of the study corpus. It is not known how study corpora of the same size behave in terms of the total keywords that they yield when compared to reference corpora of the same size.

Another question is the composition of the keyword lists obtained. This study restricted itself to quantitative aspects of keyword list variation, but it would be important that changes be assessed qualitatively as well. In particular, it would be pertinent to know which keywords were added or dropped as the levels of n changed. Finally, the fact that Brown corpus texts are short fragments and not whole texts may have upset the results, since the number of keywords seems to vary considerably as a function of the size of the texts (Mike Scott, personal communication). Shorter texts provide less room for repetition, which in turn influences word frequencies.

Acknowledgements

My thanks go to Mike Scott and the three anonymous reviewers for their comments.

References

Leila Barbara and Mike Scott (1999). Homing on a genre: invitations for bids. In “Writing Business: Genres, media and discourse”, In F. Bargiela-Chiappini & C. Nickerson, ed., Longman, New York, USA, pp. 227-254.

Maria Eugênia Batista (1998) E-Mails na troca de informação numa multinacional: o gênero e as escolhas léxico-gramaticais. Unpublished MA Thesis, LAEL, Catholic University of Sao Paulo, Brazil.

Tony Berber-Sardinha (1995). The OJ Simpson trial: Connectivity and consistency. Paper presented at the BAAL Annual Meeting, Southampton, England, 14 September 1995.

Tony Berber-Sardinha (1999a) Using KeyWords in text analysis: Practical aspects. DIRECT Papers, 42. LAEL, Catholic University of Sao Paulo, Brazil / AELSU, University of Liverpool, England. (Available online at www.direct.f2s.com)

3 This could be done in WordSmith itself through the 'consistency list' function.
Tony Berber-Sardinha (1999b) *Word sets, keywords, and text contents: an investigation of text topic on the computer*. Delta, 15, pp. 141-149. (Available online at www.scielo.br)

Tony Berber-Sardinha and Marilisa Shimazumi (1998) *Using corpus linguistics to describe the APU (Assessment of Performance Unit) archive of schoolchildren's writing*. Unpublished manuscript. (Available online at www.tonyberber.f2s.com)

Márícia Bonamin (1999) *Análise organizacional e léxico-gramatical de duas seções de revistas de informática, em inglês*. Unpublished MA Thesis, LAEL, Catholic University of São Paulo, Brazil. (Available online at www.lael.f2s.com/online.htm)

Heloísa Collins and Mike Scott (1996) *Lexical landscaping*. DIRECT Papers, 32. CEPRIL, Catholic University of São Paulo, Brazil, and AELSU, Liverpool University, England.

Helena Conde (1999) *Aspectos culturais da escrita de alunos de uma escola americana em São Paulo - Uma perspectiva baseada em corpus*. MA Project, LAEL, Catholic University of São Paulo, Brazil.

Sylvie De Cock, Sylvianne Granger, Geoffrey Leech and Tony McEnery (1998) *An automated approach to the phrasicon of EFL learners*. In “Learner English on Computer”, S. Granger, ed., Longman, New York, USA, pp. 67-79.

Ted Dunning (1992) *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19, pp. 61-74.

Patrícia Dutra (1999) *Análise léxico-gramatical baseada em corpus da música pop contemporânea*. MA Project, LAEL, Catholic University of São Paulo, Brazil.

Alice de Freitas (1997). *América mágica, Grã-Bretanha real e Brasil tropical: um estudo lexical de panfletos de hotéis*. Unpublished Doctoral Thesis, LAEL, Catholic University of São Paulo, Brazil. (Available online at www.lael.f2s.com/online.htm)

Heloísa Fuzetti (1999) *A interação oral entre crianças numa escola americana - Uma abordagem baseada em corpus*. MA Project, LAEL, Catholic University of São Paulo, Brazil.

Sylvianne Granger and Paul Rayson (1998) *Automatic profiling of learner texts*. In “Learner English on Computer”, S. Granger, ed., Longman, New York, USA, pp. 119-131.

Sylvianne Granger and Chris Tribble (1998) *Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning*. In “Learner English on Computer”, S. Granger ed., Longman, New York, USA, pp. 199-209.

Rodrigo Lima-Lopes (1999) *Padrões colocacionais dos participantes em cartas de negócios em língua inglesa*. Manuscript, LAEL, Catholic University of São Paulo, Brazil.

Maria Cecília Lopes (2000) *Homepages institucionais em português e suas versões para o inglês: Uma análise baseada em corpus de aspectos lexicais e discursivos*. Unpublished MA Thesis, São Paulo, Brazil, LAEL, Catholic University of São Paulo, Brazil.

John Milton (1998) *Exploiting LI and interlanguage corpora in the design of an electronic language learning and production environment*. In “Learner English on Computer”, S. Granger, ed., Longman, New York, USA, pp. 186-199.

Martin Phillips (1989) *Lexical Structure of Text*. Birmingham: ELR, University of Birmingham, 80 p.

Rosinda Guerra Ramos (1997) *Projeto de imagem através de escolhas linguísticas: Um estudo no contexto empresarial*. Unpublished Doctoral Thesis, LAEL, Catholic University of São Paulo, Brazil.

Valérie Branco Pinto dos Santos (1999) *Padrões interpessoais no gênero de cartas de negociação*. Unpublished MA Thesis, LAEL, Catholic University of São Paulo, Brazil. (Available online at www.lael.f2s.com/online.htm)

Mike Scott (1997) *PC Analysis of key words and key key words*. System, 25, pp. 233-245.

Mike Scott (1998) *Wordsmith Tools Version 3*. Oxford University Press, Oxford, England.

Maria Fernanda da Silva (1999) *Análise lexical de folhetos de propagandas de escolas de línguas e as representações de ensino*. Unpublished MA Thesis, LAEL, Catholic University of São Paulo, Brazil. (Available online at http://www.lael.f2s.com/online.htm)

Chris Tribble (1998) *Genres, keywords, teaching-towards a pedagogic account of the language of Project Proposals*. Paper presented at TALC98, Oxford, England.