the ESP, São Paulo, vol. 22, nº 1 87-99

# COMPARING CORPORA WITH WORDSMITH KEYWORDS *
## Comparação de corpora com WordSmith KeyWords

Tony BERBER SARDINHA (LAEL, PUC-SP)

**Abstract**
*In this article, I review some of the features available for language analysis in the KeyWords tool of WordSmith (Scott, 1996). In general, KeyWords has proved a reliable tool for comparing language samples, be them texts or corpora. Its results have helped identify not only differences across texts but also within texts. I present some arguments related to the use of chi-square in comparing word frequencies, and propose two techniques for extracting a representative subset of key words for analysis.*

**Keywords:** *Corpus Linguistics; KeyWords; corpora; WordSmith Tools.*

**Resumo**
*O artigo apresenta uma resenha do programa KeyWords, para análise lingüística, que faz parte do pacote WordSmith Tools (Scott, 1996). KeyWords tem se apresentado como uma ferramenta confiável na comparação de amostras de linguagem, sejam elas textos individuais ou corpora. O programa identifica 'palavras-chave', isto é, aquelas que possuem freqüência estatisticamente superior àquela encontrada em um corpus de referência. O artigo ainda discute questões relacionadas com a estatística qui-quadrado, empregada pelo programa, e propõe duas técnicas para extrair subconjuntos representativos de listas de palavras-chave para análise.*

**Palavras-chave:** *Lingüística de Corpus; KeyWords; corpora; WordSmith Tools.*

---

* Earlier versions of this paper appeared in Liverpool Working Papers in Applied Linguistics (LWPAL) 2.1:81-90, 1996, under the title 'Applications of WordSmith Key Words', and in DIRECT Papers 42, entitled 'Using key words in text analysis: practical aspects'.

## 1. Introduction

Nowadays there is a greater amount of electronic texts available than ever before. As a result, more research has been carried out on texts and corpora using computer programs. WordSmith tools is a recently-published suite of programs which offers many innovations for those interested in computing and studying word frequencies and word patterns. One of its strengths is the range of features which each of the individual tools offers (see Berber Sardinha, 1996a for a review). It is aimed at those people who do research in text using either a single text or a corpus. One of its innovative tools is KeyWords, a program which carries out comparisons between word lists. In this article, I intend to introduce some of the basic features of KeyWords and report on some studies in which they have been used.

## 2. Comparing Corpora

Lately there has been interest in the issue of contrasting corpora by comparing the frequencies of the words in them. Kilgariff (1996a&b) explains that one of the interests in comparing corpora is that in some contexts (e.g. lexicography) one needs to decide whether to use one corpus or another. In order to decide it is crucial that researchers know what the similarities and differences are between the two corpora. Further, sometimes it is necessary to predict whether the results obtained in previous research by using corpus 'x' can be generalised to results using corpus 'y'.

## 3. Understanding Key Words

KeyWords has not been designed to address these exact issues, but it can be used to help answer related questions. For example, 'how is text 1 different from or similar to text 2?' Or, 'what are the possible topics being discussed in text collection A as opposed to text collection B?' These issues have been discussed on-line on the CORPORA distribution list and the results of the discussion can be found on the Internet at http://www.liv.ac.uk/~tony1/corpus.html.

What is meant by key word is something different from 'important word' because in the program keyness is defined by frequency. Thus, a word will be key if its frequency is either unusually high or unusually low in comparison to a reference corpus.

A key word analysis normally involves at least two files. Typically, one will be the target text or texts (the one under consideration), and the other the reference text or texts, but one can simply compare two individual texts. The reader is reminded that by file is meant a WordList file, not a raw text. Mike Scott has put a huge word list of newspaper stories on-line, containing about 95 million words, which will be excellent as a reference corpus, at http://www.liv.ac.uk/~ms2928/homepage.html. It is a single-word word list, though, which will not work if you are planning to look at clusters. A key words analysis need not involve only two files though. The programme can also handle multiple comparisons, that is, many target files against a single reference file. This type of comparison is done by selecting 'batch processing'. Optionally, you may choose a stop list in case you want to weed out the commonest words such as 'the' and 'of'.

## 4. Statistics

The results of the comparison are shown on the screen in a table containing the words which are 'key' together with their frequencies in the two files plus some additional statistical information (chi-square and p value, if the right conditions are met). There has been a debate about the use of the chi-square statistic in comparing word frequencies (e.g. Kilgariff, unknown). It has been argued that one of the problems with the chi-square statistic is that what one is testing by using it is whether two samples (that is, two texts, two sets of texts, two corpora, etc) have been randomly drawn from the same population. If the chi-square value is high enough, one can reject the hypothesis that the samples have been drawn from the same population. In other words, one can assume that sample 1 and sample 2 are different with regard to the use of a word or certain words. However, one cannot assume that words have been drawn at random, because words are chosen depending

on a number of reasons (syntax, topic, usage, etc). Further, as Kilgariff (1996a) has shown, the comparison of most high frequency words by chi-square tends to result in significant values. As Owen and Jones (cited in Kilgariff, 1996b:8) have argued, the chi-square statistic can only tell us whether 'the sample size is too small to reject the null hypothesis'. Nevertheless, even critics such as Kilgariff have been using adapted versions of the chi-square statistic for comparison of language samples. Crucially, researchers who have investigated how to identify topical units in texts (e.g. Thomas and Wilson, 1996) have resorted to chi-square; therefore, there is strong reason to believe that chi-square is not ill-suited for the task of comparing corpora in the way KeyWords does.

## 5. Key Word Analyses

Key words have been used in a number of investigations. For instance, Berber-Sardinha (1995b) used key words to explore a model of intertextual lexical cohesion (cf. Hoey 1991 & 1995). He extracted key words from a corpus of newspaper texts and then computed collocations of these key words. The analysis of recurrent key words published in different months helped reveal texts which referred to each other meaningfully. Also, Shimazumi and Berber Sardinha (1996) used key words to compare frequency lists from adult and schoolchildren text and found that the set of schoolchildren key words indicated important stylistic and developmental characteristics of the texts they had written. These characteristics are interpreted as showing the process of acquisition of literacy by schoolchildren. Key words have been used to study text internal differences as well. For example, Berber-Sardinha (1995d) extracted key words from business reports. Based on their position in the texts, he found that key words tended to group together into two distinct meaning sets, 'company' and 'non-company' words. These sets indicate a broad division in the business reports, namely between topics related to the company itself and topics related to the company's employees, investments, etc.

Using key word plots (see section 6) the author found that company key words were typically being used near (though not in the

same collocational environment) non-company words. Similarly, Berber-Sardinha (1995a & 1996b) compared the placement of key words within sections of business reports. He found that the most striking difference is not between individual sections of the report, but between introductory and non-introductory sections.

In recent versions of WordSmith tools, the sorting facility has been expanded to include an option which sorts key words by first appearance. This can be very helpful in identifying internal topic boundaries in the texts, since it becomes possible to see at which points in the text new key words were introduced which in turn might reveal the beginnings of new topics. Other sorting options include sorting by range and keyness. The former can provide the opportunity to distinguish between 'local' and 'global' topics, while the latter can perhaps be used to differentiate between 'major' and 'minor' topics. It must be stressed, though, that descriptive categories such as 'local', 'major' and so on are not inherent in key word analysis. Their applicability will depend on the interpretation of one's individual data.

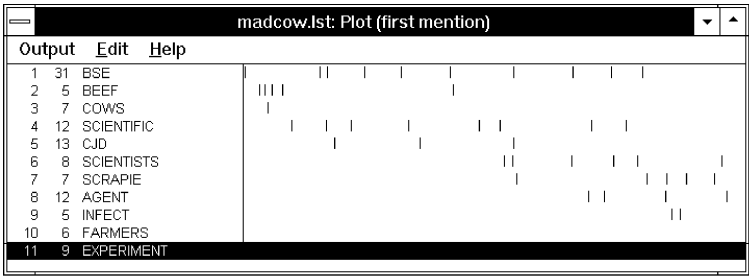


**FIGURE 1 – Partial KeyWords plot of a report about 'mad cow disease'**

## 6. Plots

The key words list can be complemented by the key words plot – a diagram that shows the distribution of the key words within the text (so it is a good idea to keep one text per file otherwise the plot will be

useless). An example of a key words plot appears in FIGURE 1 The screen shot shows distribution of key words of a text about mad cow disease (The Independent, 10 December 1995). The plot offers interesting initial insights about the placement of topics in the text. For instance, 'beef' occurs only at the very beginning, which suggests that the text is not exactly about beef itself, but about problems associated with eating beef.

Other key words such as 'BSE' and 'CJD' corroborate this interpretation. A further contrast is perhaps signalled by the different positions occupied by 'scientists' and 'farmers'; the former are found across the middle of the text, whereas the latter appear towards the end. Note that the frequencies of key words in the text as shown down the second column of numbers do not match the number of markers on the plot because the markers represent how many user-defined portions of text contained at least one key word. In this example, each portion corresponded to 9.9% of the text, which means that more than one key word can occur in that portion.

Unlike the table (the default display of keywords), whose capacity is virtually unlimited, the plot can handle at most 200 key words, but this is surely enough for most applications. If it is not, then this is an indirect sign that the user must rethink and select more selective settings so that fewer key words are obtained. The online help mentions 40 key words as a reasonable limit, but this is simply a rule of thumb; if you do not treat each key word individually but rather as a member of a set, then certainly an output having much more than 40 key words can be interpreted without much trouble.

## 7. Selecting a Sample of Key Words

The key word lists produced by the program normally hold more key words than it is possible for the researcher to analyse. As a result, most researchers select a portion of the total key words to interpret. There is no consensus as to what would be a suitable sample size, and as a result people tend to use intuitive figures such as 100 or 200. The question that arises is what would be the ideal size of the portion, that

is, what would be the smallest *representative* number of key words from a given set. In this section I want to propose two basic ways which can be used to select a representative subset of key words.

The first method is simply choosing the *majority* of key words. This can be easily implemented by getting half of your key word types (i.e. individual key words) plus one. To do this, it is advisable that the user override the default number of key words returned by the program, which is normally 500, otherwise you would have a subset of a subset of the key words. Nevertheless, this is not a steadfast rule and you could in principle apply the majority method to the 500 key word list but you would presumably have to argue harder. Once you have the starting set of key words, simply divide the total number of individual key words by two and add one; thus, if you have 500 key words, a majority subset would be 251[1].

Another version of the majority method is to count tokens instead of types. Here you would have to add the frequency of all the key words in your starting set for your target text or corpora (i.e. not the reference corpus) and get 50% of them plus one. Notice that the key words program does not give you the total number of tokens. You would have to obtain that figure by other means. A spreadsheet such as MS-Excel can do this fairly easily. In the key word listing, first sort the key words by descending frequency in the target corpus (the left-hand side columns) so that the most frequent key words appear at the top of the list. Then select the columns that give the key words and their frequencies in the target corpus. Click on 'copy' and then paste that selection into MS-Excel. Remove any extra headings that might have come with the selection by choosing 'delete rows'. Now go to the bottom of the spreadsheet underneath the column with the frequencies and call up the 'sum' function. If everything goes well, you will now have a total figure for tokens. Use MS-Excel (or a calculator) to divide that total by two and that will be your majority number of key word tokens. The next task is to choose the subset of individual key words whose added frequencies correspond to the majority figure. In other words, you want

[1] In the cases where the division by two does not result in an integer (e.g. 25.5 for a 51-word list), it is best to round it off to the nearest higher integer (e.g. 26 for 25.5).

to select those key words whose frequencies account for more than half of the key word tokens. In theory, there would be two ways to select the key words: by counting either from the top of the list (thus starting with the most frequent key words) or from the bottom of the list. However, the most natural way is to choose words from the top of the list, so that your list includes those key words that account for most of your target corpus tokens. To do this, you need to obtain the cumulative frequencies for the key words in the spreadsheet. This is not as simple as getting the overall total, though. When you have obtained the cumulative totals, spot the place in the cumulative totals that match the majority figure. The key words up to that point will be your subset.

The second method is obtaining a *significant* subset of key words. By significant is meant a statistically significant subset. The chi-square test can be used to identify a significant portion of the key words. Before running the test, think of your key words as falling into one of two categories: 'Chosen' and 'Not chosen'. The 'chosen' key words are those that will make up the significant subset, and the 'not chosen' ones will be ignored. The chi-square test works by comparing observed frequencies to expected frequencies and assessing whether the difference between observed and expected is higher than a criterion. Therefore, you need to compare the observed and expected frequencies for chosen and not chosen key words. Table 1shows the smallest values for 'chosen' that reach significance at $p<.05$ for a selected number of key word sample sizes; higher values would yield even better significance values. Importantly, in these calculations we assume that the chosen key words outnumber the not chosen ones and thus we assign the higher number in the contingency table to 'chosen'. As with the majority method, there are two figures that can be used for a significant sample: types or tokens. Thus, if you are considering types, and your total for key word types is 1000, according to Table 1a significant sample would be 531, which is significant at $p=.0499$. This is equal to 53.1% of the starting key word set, which is more than what you will get if you follow the majority method, which would be 501. If you are considering tokens, the majority sample would be the same size, 531, since this does not make a difference to the calculation of the chi-square test. The figures in Table 1 show

that the smaller the starting set the higher the significant sample. So, for instance, if your original key word list has 100 words, you will need 60% of the key words to make up a significant sample, but if you have 20,000 total key words, you will need just above 50% (more precisely 50.695%) for a significant sample.

| Total key words | Smallest significant sample of chosen key words | % | p |
|---|---|---|---|
| 1,000,000 | 500,980 | 50.098 | 0.0499 |
| 20,000 | 10,139 | 50.695 | 0.0493 |
| 1,000 | 531 | 53.1 | 0.0499 |
| 500 | 272 | 54.4 | 0.0491 |
| 100 | 60 | 60 | 0.0455 |
| 10 | 9 | 90 | 0.0114 |

**Table 1 – Smallest significant sample sizes of selected total key word values**

To obtain the exact figure for other total key word sample sizes not shown in Table 1 you can run the chi-square test in MS-Excel. Supposing that you obtained a total of 400 key words and wanted to extract a significant sample of types, in the spreadsheet, enter the totals in a layout similar to Table 2

| | Chosen | Not chosen | Total |
|---|---|---|---|
| Observed | 220 | 180 | 400 |
| Expected | 200 | 200 | 400 |

**Table 2 – Table layout for calculating significant sample sizes**

Assuming that the value for 220 is in cell B5, enter the following formula in a separate cell outside the table: *=TESTE.QUI(B5:C5,B6:C6)*[2]. This will return 0.0455002705 which is the significance value, meaning that a subset of 220 key words out of a total of 400 key words is a significant sample. Lowering this value to 210, for instance, would yield p= 0.3173108131, which means that a 210-word subset is not a significant sample.

The significant criterion is more appealing than the majority method since in principle it embodies the notion of objectivity because of the use of statistical tests in order to estimate the sample size. At the same time it is more controversial since there are no widely agreed methods for estimating samples of words. The researcher wanting to select a subset of key words based on these guidelines should be aware of the potentially controversial nature of this method.

## 8. Extending Key Words Analysis

A typical key word analysis involves the extraction of key words from a set of different texts. Once we have separate key word lists for the individual texts we generally want to answer the question 'what is the most recurrent of these key words?' The KeyWords program can give an extra bit of help in these situations because it incorporates a facility that computes in how many files each key word was key. This is accomplished by means of the 'key key words' option which picks out those key words which occurred at least twice and then lists in what percentage of your batch of files they were key in. Thus, a word which was key in at least two texts will be a key key word of those texts. This concept has found an interesting application in the comparison of the testimony of major witnesses in the OJ Simpson trial (Berber Sardinha, 1995c). First, key words were obtained for the various kinds of examination, for example *direct, cross, redirect, recross, etc.* Then key key words were extracted, namely those which were key in most examinations. Finally the defence witness's key key words were

[2] This is how the command reads in the Brazilian Portuguese version of MS-Excel 97.

compared with those for the prosecution witness. The results indicated consistent choices of key words over the length of each witness's testimony. Thus, one of the ways key key words may be interpreted is as being markers of consistency of one's style or stance. Similarly, Scott (1997) has used key key words to identify sets of recurrent topics among thousands of newspaper reports. He used the 'associates' and 'clumps' facilities of key words to identify groups of texts which had key words in common. One of his analyses revealed some of the topics which seem to be commonly associated with 'the British' in the press.

The plot is helpful not only because it works as a visual aid to the distribution of key words but also because it provides additional information about the co-occurrence of key words. Beside each key word it displays a number that indicates how many times a given key word appeared in the collocation span of another. This will give you a rough measure of the interrelatedness of the key words. However, the figures by themselves do not mean much without knowing which words actually co-occurred. This is obtained by clicking on the individual key words one is interested in, which brings up a small window which in turn presents the 'Links' for that word, namely the other key words that appeared in its collocation horizon.

One of the difficulties of using key key words is that there is no way of knowing which of them were key in the same texts. The 'Associates' option can partly remedy this situation because it shows which words were key in the same texts as each key key word. Remember that the words in the 'Associates' listing are not key key words since they will not have occurred of necessity at least twice but they will have been key words anyway in the same texts as the particular key key word you have chosen.

The key words tool can help in the investigation of word patterns. With the publication of Cobuild's new book on verb patterns (Francis & Hunston, 1996) there is a likelihood that lexical patterns will become part of foreign language teaching methodology and therefore it is a matter of time before more and more teachers and learners will start looking for patterns themselves in their own texts. The WordSmith package will be of great help in these situations because of its 'clusters'

facilities. The 'associates' and 'clumps' features in key words can provide a different way of looking at patterns from that in Francis & Hunston (1996), where lexical patterns mean co-occurrence at a narrow distance, while 'associates' and 'clumps' address co-occurrence within the same text or group of texts. Hence, the KeyWords sense of co-occurrence is in many ways similar to the old meaning of 'collocation', advocated by Firth, Sinclair, Halliday (Scott 1997), which is different from the contemporary meaning of collocation as words which co-occur within a four- or five-word span.

## 9. Final comments

The KeyWords facility of WordSmith is an extremely helpful tool to investigate differences and similarities both across and within texts. It is hoped that this short article has helped illustrate some of its possible applications in research.

Recebido em: 07/2000. Aceito em: 10/2000.

## References

BERBER SARDINHA, A. P. 1995a Annual business reports sections: key words. *DIRECT Papers*. *Working Paper* **25**. CEPRIL, PUC-SP, Brazil, and AELSU, Liverpool University, England.

_____ 1995b Intertextual lexical cohesion in newspaper reports. Paper presented at the 40th Annual Conference of the International Linguistic Association, Georgetown University, Washington, DC, USA, March 10, 1995.

_____ 1995c The OJ Simpson trial: connectivity and consistency. Paper presented at the BAAL Annual Meeting, Southampton, UK, 14 September 1995.

_____ 1995d Segmentation and choice in written business English. Paper presented at the 7th International Systemic Functional Workshop, Universidad Menéndez Pelayo, Valencia, Spain, 26-29 July 1995.

_____ 1996a Review of WordSmith tools. *Computers & Texts* **12:**19-21.

_____ 1996b Sections as linguistic units: Key words. 5th Annual Postgraduate Conference, 9 March 1996, University of Manchester, Manchester, UK.

FRANCIS, G., & S. HUNSTON. 1996 *Verbs. Grammar Patterns, 1*. HarperCollins, Cobuild.

HOEY, M. 1991 *Patterns of Lexis in text. Describing the English Language*. OUP.

_____ 1995 The lexical nature of intertextuality: A preliminary study. IN: WARVIK, B., S.-K. TANSKANEN, & R. HILTUNEN (ed.) *Organization in Discourse*. Proceedings from the Turku Conference. Anglicana Turkuensia, 14, 73-94. Turku: Abo Akademi.

KILGARIFF, A. 1996a Using word frequency lists to measure corpus homogeneity and similarity between corpora. Proceedings, COLING workshop on very large corpora.

KILGARIFF, A. 1996b Which words are particularly characteristic of a text? A survey of statistical approaches. Language Engineering for Document Analysis and Recognition. Brighton, England. AISB Workshop series.

_____ (unknown). Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. ITRI, University of Brighton, UK.

OWEN, F., & R. JONES 1977 *Statistics*. Polytech Publishers.

SCOTT, M. 1996 *WordSmith Tools*. Oxford University Press.

SCOTT, M. 1997 PC Analysis of key words - and key key words. *System* **25:** 233-245.

SHIMAZUMI, M., & A.P. BERBER SARDINHA 1996 Approaching the Assessment of Performance Unit (APU) archive of schoolchildren's writing from the point of view of corpus linguistics. Paper presented at the TALC96 Conference, Lancaster University, UK, 11 August 1996.

THOMAS, J., & A. WILSON 1996 Methodologies for studying a corpus of doctor-patient interaction. IN: THOMAS, J. & M. SHORT (ed.) *Using Corpora for Language Research*. Longman. 92-109.

*Tony Berber Sardinha holds a PhD in English from the University of Liverpool. He is currently Assistant Professor at the Catholic University of São Paulo, where he carries out research in Corpus Linguistics.*