Corpus Analysis of Key Words

PAUL RAYSON

Introduction

If you look up “key word” in a dictionary you will find that it has various senses, but the relevant sense here is a word that has importance or significance especially to describe the contents of a document. In the context of corpus analysis, the accepted definition of a “key word” extends this sense to a *statistically* significant word characterizing a document, text, or corpus. The procedure to extract key words from a corpus is one of the most regularly used tools in a corpus linguist’s toolbox alongside frequency profiling, concordancing, n-grams (clusters or lexical bundles) and collocation analysis. Numerous studies in linguistics and other disciplines have exploited the key-words procedure for the analysis of textual data, ranging from a discourse analysis of refugees and asylum seekers in the UK press, studies of health communication, lexical simplification in translations, profiling of learner language, to e-learning materials development. The key-words procedure is inherently comparative. In order to discover the statistically significant key words in a corpus, we compare it to a (usually larger) reference corpus and extract the words that occur significantly more or less frequently compared to what we expect based on the reference document(s).

There are several further meanings of “key word” in the language- and text-processing literature that we turn to briefly here. In the area of information retrieval (IR), the science of indexing and searching for information in document collections or on the Web, a “key word” is a manually assigned word or phrase given as a label to a book or article by an author or human abstracter. Techniques have been proposed in IR to extract such key words automatically from text (notably Sparck Jones, 1971) and some of these measures are now finding their way across to the community of corpus linguistics (Oakes, 2008). Similarly, one can assign key words to photographs or images to assist in their retrieval (Zhou & Huang, 2002). Early work in semantics was carried out by Firth (1935) on “focus and pivotal words” and this laid the groundwork for much that was to follow. Cultural key words are “words which capture important social and political facts about a community” (Hunston, 2002, p. 117). Williams (1983) produced a set of around 120 words which were important in the culture but were selected subjectively. Stubbs (1996, p. 172) made his selection of words based on characteristic collocations to show the associations and connotations that they have. Wierzbicka (1997, p. 16) in trying to understand cultures through their key words had no “objective discovery procedure” for them. Finally, Stubbs (2010) compared and contrasted the concepts of cultural key words with the statistical key words that we shall focus on here.

Having now examined the core and related definitions for key words, in the next section we will further clarify what significant key words are and define the process by which they are calculated. Next, we will focus on problems and limitations of the technique and possible solutions to them, before concluding by considering a variety of applications for this technique by way of further literature for the reader to explore.

This is the submitted version of the following book chapter which has been published in final form as:
Rayson, P. (2019) Corpus analysis of key words. In Chapelle, Carol A. (Ed) *The Concise Encyclopedia of Applied Linguistics*. Wiley, p. 320-326.
https://www.wiley.com/en-gb/The+Concise+Encyclopedia+of+Applied+Linguistics-p-9781119147367

Method: How Are Key Words Calculated?

The procedure to calculate key words can be applied mechanically, is conceptually quite straightforward, and is comprised of three stages. The first stage is to compute a word-frequency list for each of the two texts that we wish to compare. One of these texts would usually be the larger reference corpus mentioned above. For each text, the word-frequency list records the different word forms (types) and how many times they occur (tokens). We also count the total number of running words in each text. The second stage is to compare the two resulting frequency lists. Some complexity arises in the application, choice of formula, and the assumptions made in such calculations and we will further describe these issues below. Conceptually the comparison is again quite straightforward. For each word in the two texts, we apply a metric or formula that compares its relative frequencies (i.e., percentages of occurrence) in the word-frequency lists. The value of the statistic or "keyness" is proportional to the difference in relative frequencies. In other words, the larger the difference in relative frequencies, the larger the value of the statistic or "keyness." The third and final stage of the process is to sort the words in terms of their keyness. This means that, all other things being equal, the most interesting key words with the largest keyness values appear at the top of the list. The least interesting words, whose relative frequency is similar in the two texts, are listed toward the bottom of the key-words list. We can further distinguish between positive and negative key words. Positive key words are those which are "overused" in the first text: their relative frequency is higher in this text compared to the second or reference text. Negative key words are said to be "underused" in the first text relative to the second since their relative frequency is lower in this text compared to the reference corpus. In most cases, it is the positive key words that are most interesting since they tell us about what occurs more often in the first text.

In terms of the second stage and the application of the keyness statistic, we first need to construct a "contingency" table for each word in the frequency lists. Then we apply our chosen statistic to calculate the keyness value. The most widely used significance method is log-likelihood with chi-squared now being less frequently used. Other keyness statistics have been proposed; see Baron, Rayson, and Archer (2009) for a detailed survey and discussion of the criticisms of these statistics. Chi-squared was first used in a corpus analysis context by Hofland and Johansson (1982) to compare word frequencies in corpora of 1 million words of American English (the Brown Corpus) with 1 million words of British English (the LOB Corpus). The chi-squared values can be looked up in statistical tables (of the chi-squared distribution) in order to identify those that are statistically significant at different levels of confidence, for example 5%, 1%, and 0.1%. Each level corresponds to a cutoff and a probability (or "*p*") value. For example, the 5% level (or *p* value of 0.05) corresponds to a keyness critical value of 3.84. Hence any words with a chi-squared value greater than or equal to this value are considered significant. At the 1% level, *p* value of 0.01, the critical value is 6.63. We can specify smaller *p* values and correspondingly higher critical values in order to be more certain of our results. This mathematical process is known as hypothesis testing and the default or "null hypothesis" in this test is that there is no difference between the actual frequencies we observe in the two corpora. A high enough keyness value allows us to reject this null hypothesis for a given word, although it should be noted that, strictly speaking, this does not logically entail support for the alternative hypothesis, that there is a difference between the actual frequencies.

Log-likelihood (LL) was first brought to the attention of the corpus community by Dunning (1993) for collocation analysis. LL has been shown to be more reliable than chi-squared in a number of different arrangements of corpus comparison (Rayson, Berridge, & Francis, 2004) such as varying the relative sizes of the corpora and across the range of word frequencies. Hence, we will present it here. This simpler version of the formula comes from Read and Cressie (1988, p. 3) who show that chi-squared and log-likelihood come from the same family of statistics.

Table 1 Contingency table for keyness calculation

	Corpus 1	*Corpus 2*	*Total*
Frequency of a word	a	b	$a + b$
Frequency of other words	$c - a$	$d - b$	$c + d - a - b$
Total	c	d	$c + d$

Table 1 shows the contingency table that we need to complete for each word in the frequency lists. The values "*a*" and "*b*" are the "observed" or actual frequencies of the words in the two corpora. The total size of each corpus is shown by "*c*" and "*d*." We first calculate "expected" values for each word using the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

The expected values are simply averages for each word adjusted for the corpus size. In this formula, "*N*" corresponds to the total number of words and "*O*" corresponds to the observed value. From Table 1, $N_1 = c$, and $N_2 = d$. So, we calculate $E_1 = c \times (a + b) / (c + d)$ and $E_2 = d \times (a + b) / (c + d)$. The final log-likelihood value is then calculated using the following formula:

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

The formula represents the distance of the word frequency in each corpus from the previously calculated expected or average values. In terms of Table 1, $LL = 2 \times ((a \times \ln(a / E_1)) + (b \times \ln(b / E_2)))$. This result can be calculated using a simple spreadsheet, although we have to be careful about calculating ln(0), so in practice we ignore that half of the calculation if "*a*" or "*b*" are zero.

A very important caveat is required at this point. Statistical goodness-of-fit procedures such as chi-squared, the log-likelihood metric and others assume that samples are random with independent observations, but of course we know that language corpora do not abide by these assumptions (Stubbs, 1995; Kilgarriff, 2005; Evert, 2006; Baroni & Evert, 2007). Gries (2005) also highlights the fact that corpus linguists rarely apply a correction for multiple "post-hoc" testing as takes place in the key-words procedure and points out that this correction is more common in other disciplines. Additionally, null hypothesis significance testing is increasingly being criticized in other research areas. The usual sidestep taken by linguists who are aware of these criticisms is to employ the metrics to calculate the keyness values and use them only to place the key words in rank order rather than determine significance for each word. If we put less reliance on the significance-testing component of the procedure, it can be used informally to guide our analyses (Rayson & Garside, 2000).

Scott defined, refined, and applied this key-words process with his WordSmith software in a number of papers (1997, 2000a, 2000b, 2001a, 2001b, 2002). As we have seen, the procedure is inherently comparative and usually involves a general reference corpus. It can also be used to compare a text with other relevant texts (Rayson & Garside, 2000; Seale, Ziebland, & Charteris-Black, 2006). There are at least three important considerations to bear in mind when choosing the reference text: representativeness, homogeneity, and comparability. Representativeness (Biber, 1993) is an important attribute for a reference corpus. To be representative, a corpus should contain samples of all major types of text (Leech, 1993) in some way proportional to their use in "everyday language" (Clear, 1992). This allows us to discover features in the study text with significantly different usage to that found in "general" language. Where we are comparing corpora of roughly equal sizes, as in the Hofland and Johansson (1982) study to compare British and American English, homogeneity is important since we may otherwise find that the results reflect certain sections within one of the corpora that are unlike the remainder of the two corpora being compared. Comparability refers to the sampling methods employed in the collection of each of the corpora and ideally the same sampling methods should be employed in each case. For example, the LOB Corpus was created to be comparable to the Brown Corpus and its compilers used the same design and collection procedures. This renders the results directly comparable and avoids any unintended surprises such as key words arising because the corpora are sampled from a different time period, genre, or domain.

Problems, Limitations, and Extensions of the Technique

In addition to the caveats already expressed above, there have been some criticisms of the key-words approach. Berber Sardinha (1999) highlighted one drawback in that there are normally far too many key words for the researcher to analyze. Baker (2004a) noted three points in relation to the technique. First, "a key word analysis will focus only on lexical differences, not lexical similarities" (Baker, 2004a, p. 349). Comparing, as Baker did, two corpora of gay and lesbian erotic narratives to general corpora such as Brown or LOB would produce different key words than when comparing the erotic narratives to each other. This reinforces the careful choice of a reference corpus as of prime importance. Second, a word may be key because it occurs very frequently in one part of a corpus. Hence, examining the range or dispersion of a key word is important. Third, "key words only focus on lexical differences, rather than semantic, grammatical, or functional differences" (Baker, 2004a, p. 354). It is possible to find cases where a word is key when it appears with a number of different meanings in the text and, in contrast, cases where a word does not get marked as key because counting all of its senses together hides the fact that one of the senses is key when counted separately. Gries (2006, p. 116) also stated one obvious limitation of key-word variability studies as "they have little or nothing of interest to offer a linguist who is primarily interested in grammatical or other phenomena." Recent research has pointed out the limitations of using one frequency count to represent the within group variation of occurrences in a corpus (Brezina and

Meyerhoff, 2014), problems of the assumption of independence (Lijffijt et al., 2016) and has proposed the use of a new effect size metric to complement existing significance measures (Hardie, 2014).

In terms of revisions to the technique, Scott (1997) right from the start extended the procedure to find "key key words." These are words that are shown to be key in a number of files within a corpus. Mahlberg (2007) showed that the keyness procedure applied to lists of clusters (recurrent phrases or n-grams) can produce useful results for studying local textual functions in literary stylistics. In order to address the criticisms of the keyword procedure as discussed above, Rayson (2008) proposed an extension to the process to include key parts of speech and key semantic domains as implemented in the Wmatrix software. This exploited automatic corpus-annotation tools that assign a grammatical label and semantic field tag to every word or phrase in a corpus. The frequencies of these tags were counted to produce tag-frequency lists and then the keyness calculation was applied to those lists in addition to the word-level lists. As a result, at the semantic level, words are grouped together into semantic fields which do not emerge from the word-level analysis, thus allowing a richer and deeper set of key items to emerge. The practical problem of too many words to examine is also partially solved because of the smaller number of semantic groups that need to be consulted.

Example Applications

By way of a small example, the key-words procedure was applied to the full text of "Alice's Adventures in Wonderland" (one of the most frequently downloaded texts from the Internet Archive and Project Gutenberg; see online resources). The total text is around 27,000 words long and it was compared to the British National Corpus written sampler (1 million words). As Scott notes, many of the top key words in any comparison are personal names and this can be seen in the results here. The top 20 key words are (in order of keyness): *Alice*, *she*, *said*, *turtle*, *hatter*, *gryphon*, *I*, *it*, *mock*, *you*, *herself*, *dormouse*, *Queen*, *rabbit*, *her*, *house*, *March*, *caterpillar*, *very*, *duchess*. Using a *p* value of 0.01, there are 1,189 words above the critical log-likelihood value of 6.63, giving plenty of words to examine in further analysis.

The key-words procedure has been used to address a wide range of research questions in linguistic analysis. Baker (2004b) used key words to examine discourses of homosexuality in the UK House of Lords debates on gay-male law reform, and later for a large-scale comparison of British and American English (2017). Toolan (2006) highlighted the growing trend of use of the key-word procedure for literary analysis as a way of discovering foregrounding, structuring, and reader-guiding in a text. There are many further examples on the Web pages of WordSmith and Wmatrix tools listed in the suggested readings below. Scott and Tribble (2006) also showed how key words and other related corpus analysis techniques can be used in language education and teaching.

SEE ALSO: Corpora: English-Language

References

Baker, P. (2004a). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics, 32*(4), 346–59.

Baker, P. (2004b). "Unnatural acts": Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics, 8*(1), 88–106.

Baker, P. (2017). *British and American English: Divided by a common language?* Cambridge, England: Cambridge University Press.

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies, 20*(1), 41–67.

Baroni, M., & Evert, S. (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics* (pp. 904–11). Prague, Czech Republic: Association for Computational Linguistics.

Berber Sardinha, T. (1999). Using keywords in text analysis: Practical aspects. DIRECT Working Papers 42, São Paulo and Liverpool.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–57.

Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, *19*(1), 1–28.

Clear, J. (1992). Corpus sampling. In G. Leitner (Ed.), *New directions in English language corpora*, (pp. 21–31). Berlin, Germany: De Gruyter.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik, 54*(2), 177–90.

Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society, 34*(1), 36–72.

Gries, St. Th. (2005). Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory, 1*(2), 277–94.

Gries, St. Th. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora, 1*(2), 109–51.

Hardie, A. (2014). *Log Ratio – an informal introduction*. CASS blog: http://cass.lancs.ac.uk/?p=1133 Accessed 26th July 2018.

Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. Bergen, Norway: The Norwegian Computing Centre for the Humanities.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, England: Cambridge University Press.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory, 1–2,* 263–76.

Leech, G. (1993). 100 million words of English: A description of the background, nature and prospects of the British National Corpus project. *English Today, 9*(1), 9–15.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing*, *31*(2): 374–397.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora, 2*(1), 1–31.

Oakes, M. P. (2008). Measures from information retrieval to find the words which are characteristic of a corpus. In B. Lewandowska-Tomaszczyj (Ed.), *Corpus linguistics, computer tools, and applications—state of the art: PALC 2007* (pp. 127–38). Frankfurt, Germany: Peter Lang.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519–49.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Le poids des mots: Proceedings of the 7th international conference on statistical analysis of textual data (JADT 2004), volume II, Louvain-la-Neuve, Belgium, March 10–12, 2004* (pp. 926–36). Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In A. Kilgarriff & T. Berber Sardinha (Eds.), *Proceedings of the workshop on comparing corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), 1–8 October 2000* (pp. 1–6). Hong Kong, People's Republic of China: Association for Computational Linguistics.

Read, T., & Cressie, N. (1988). *Goodness of fit statistics for discrete multivariate data*. New York, NY: Springer.

Scott, M. (1997). PC analysis of key words—and key key words. *System, 25*(2), 233–45.

Scott, M. (2000a). Reverberations of an echo. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC'99: Practical applications in language corpora: Papers from the international conference at the University of Lodz, 15–18 April 1999* (pp. 49–65). Frankfurt, Germany: Peter Lang.

Scott, M. (2000b). Focusing on the text and its key words. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the 3rd international conference on teaching and language corpora* (pp. 104–21). Frankfurt, Germany: Peter Lang.

Scott, M. (2001a). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 47–67). Amsterdam, Netherlands: John Benjamins.

Scott, M. (2001b). Mapping key words to *problem* and *solution*. In M. Scott & G. Thompson (Eds.), *Patterns of text: In honour of Michael Hoey* (pp. 109–27). Amsterdam, Netherlands: John Benjamins.

Scott, M. (2002). Picturing the key words of a very large corpus and their lexical upshots or getting at the Guardian's view of the world. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the 4th international conference on teaching and language corpora, Graz 19–24 July, 2000* (pp. 43–50). Amsterdam, Netherlands: Rodopi.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam, Netherlands: John Benjamins.

Seale, C., Ziebland, S., & Charteris-Black, J. (2006). Gender, cancer experience and Internet use: A comparative keyword analysis of interviews and online cancer support groups. *Social Science & Medicine, 62*(10), 2577–90.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London, England: Butterworths.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language, 2*(1), 23–55.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford, England: Blackwell.

Stubbs, M. (2010). Three concepts of keywords. In M. Bondi & M. Scott (Eds.), *Keyness in texts: Corpus linguistic investigations* (pp. 21–42). Amsterdam, Netherlands: John Benjamins.

Toolan, M. (2006). Top keyword abridgements of short stories: A corpus linguistic resource? *Journal of Literary Semantics, 35*(2), 181–94.

Wierzbicka, A. (1997). *Understanding cultures through their key words*. Oxford, England: Oxford University Press.

Williams, R. (1983). *Keywords: A vocabulary of culture and society* (2nd ed.). London, England: Fontana Press.

Zhou, X., & Huang, T. (2002). Unifying keywords and visual contents in image retrieval. *IEEE MultiMedia, 9*(2), 23–33.

Suggested Readings

Adamson, S., & Durant, A. (2007). Four ways of looking at a keyword: Introduction. *Critical Quarterly, 49*(1), 1–5.

Archer, D. (Ed.). (2009). *What's in a word-list? Investigating word frequency and keyword extraction*. Oxford, England: Ashgate.

Bennett, T., Grossberg, L., & Morris, M. (Eds.). (2005). *New keywords: A revised vocabulary of culture and society*. Oxford, England: Blackwell.

Online Resources

Internet Archive (2009). *Project Gutenberg, Alice's adventures in Wonderland*. Retrieved July 25, 2018 from https://archive.org/details/alicesadventures00011gut

Rayson, P. (2018). *Wmatrix corpus analysis and comparison tool*. Retrieved July 25, 2018 from http://ucrel.lancs.ac.uk/wmatrix/

Scott, M. (2018). *Research using WordSmith Tools*. Retrieved July 25, 2018 from http://www.lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm