Analysing keyword lists

Paul Rayson[1] and Amanda Potts[2]

Abstract Frequency lists are useful in their own right for assisting a linguist, lexicographer, language teacher, or learner analyse or exploit a corpus. When employed comparatively through the keywords approach, significant changes in the relative ordering of words can flag points of interest. This conceptually simple approach of comparing one frequency list against another has been very widely exploited in corpus linguistics to help answer a vast number of research questions. In this chapter, we describe the method step-by-step to produce a keywords list, and then highlight two representative studies to illustrate the usefulness of the method. In our critical assessment of the keywords method, we highlight issues related to corpus design and comparability, the application of statistics, and clusters and n-grams to improve the method. We also describe important software tools and other resources, as well as providing further reading.

1 Introduction

As we have seen from Chap. 4, frequency lists are an essential part of the corpus linguistics methodology. They allow us to see what words appear (and do not appear) in a text, and give an indication of their prominence if we sort the list in frequency order. Beyond the world of the corpus linguistic researcher, frequency lists can be used directly or indirectly to support language learners by providing a way to focus on the more frequent words in a text, or suggesting priorities for language teachers when preparing lesson materials. Lexicographers use frequency lists indirectly when constructing traditional printed dictionaries. Frequency lists have also been allied with other kinds of grammatical information (such as major word class) and translational glosses (both for the words themselves and example sentences) and turned directly into frequency dictionaries for learners and teachers in a number of languages (e.g. Juilland et al., 1970; Tono et al., 2013).

When used in simple form, frequency lists can sometimes be misleading, and care needs to be taken when making generalisations from them. For example, if frequency counts are derived from a large representative corpus such as the British National Corpus (BNC), we may reasonably claim that high frequency words in the corpus may also have a similarly high usage in the language. However, a word may have a high frequency count in the BNC not because it is widely used in all sections of the corpus, but because it has a very high frequency in only certain parts of the corpus and not in others, e.g. conversational speech rather than newspaper articles. Hence, as described in Chap. 4, and more particularly in Chap. 5, we need to pay careful attention to dispersion or range measures, which can help us estimate, for instance, how widely represented a word is across the various texts, domains, or genres within a written corpus, or across speakers within a spoken corpus.

Using computer software to automate the creation of frequency lists from texts saves the researcher significant amounts of time, but the results can be overwhelming in terms of the amount of information to analyse. One option to reduce the wealth of information is to compare a frequency list from one corpus with another in order to highlight differences in word rank or frequency, since significant changes to the relative ordering of words can flag points of interest (Sinclair, 1991: 31).

[1] Paul Rayson, Lancaster University, UK. E-mail: p.rayson@lancaster.ac.uk
[2] Amanda Potts, Cardiff University, UK, E-mail: pottsa@cardiff.ac.uk

Corpus linguistics is inherently comparative, so a method has evolved to support the comparison of corpora, which helps us study, for example, the differences between tabloid and broadsheet newspapers (Baker, Gabrielatos & McEnery, 2013), vocabulary variation on the basis of age and gender (Murphy, 2010), or grammatical, lexical, and semantic change over 100 years of British and American English (Baker, 2017). The resulting widely-used method of keyword analysis is our focus in this chapter.

Hofland and Johansson (1982) were early pioneers of this approach when they carried out a large (for the time) comparison of one million words of American English (represented by the Brown corpus) with one million words of British English (in the LOB corpus). Their study employed a difference coefficient defined by Yule, which varies between +1 and –1 to calculate the difference between the relative frequencies of a word in the two corpora. In addition, Pearson's statistical goodness-of-fit test, called the chi-squared test, was applied, enabling Hofland and Johansson to mark statistically significant differences at the 5%, 1% and 0.1% confidence levels (cf. Chap. 20).

Another major milestone in the development of the keywords approach was the inclusion of the method by Mike Scott in his WordSmith Tools software. A number of authors had used significance tests to determine the importance of differences of specific words or linguistic features between corpora, but Scott's approach (1997) allowed for a systematic comparison of full word frequency lists. Scott demonstrated that the keyword results enable a researcher to understand the 'aboutness' of a text or corpus.

2 Fundamentals

The keywords method is conceptually simple, relying on the comparison of (normally) two word frequency lists. The complexity of the method lies in the choice of statistics and frequency cut-offs to appropriately filter the results (for further discussion of this, see Section 4). As part of the corpus linguist's toolbox, the keywords method is most appropriate as a starting point to assist in the filtering of items for further investigations, rather than an end in and of itself.

As a first step, two frequency sorted word lists are prepared, one from the corpus being studied (the 'target') and one from a reference corpus. Each word list contains a list of tokens and associated frequencies. The reference dataset in corpus linguistics studies is usually a general corpus, representative of some language or variety of language. However, depending on the research question and aims, a suitable comparison set may also be used, e.g. from a corpus representing a different variety, time, or genre. Rather than using two different corpora entirely, some researchers use various subcorpora from the same (reference) corpus for the 'target' and 'reference' sets, and this approach is further exemplified in the two representative studies summarised below. Next, the frequency of each word in the target corpus is compared to its frequency in the reference dataset in order to calculate a keyness value. Finally, the word list for the target corpus is reordered in terms of the keyness values of the words. The resulting sorted list contains two kinds of keyword: *positive* (those which are unusually frequent in the target corpus relative to the reference corpus) and *negative* (words which are unusually infrequent in the target corpus). It is also common to describe these two groups of words as *overused* (for positive) and *underused* (for negative), particularly in the learner corpus literature (cf. Chap. 13).

In order to perform the keyness calculation for each word in the list, corpus tools set up a 2 by 2 contingency table, as shown in Table 1 (see also Chap. 7). The value 'c' is the total number of words in the target corpus and 'd' is the total number of words in the reference corpus. Numbers 'a' and 'b' are termed the 'observed' (O) values (i.e. the actual frequencies of a given word in each corpus).

	Target Corpus	**Reference Corpus**	**Total**
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

Table 1: Contingency table for each word in the list

Two main significance test statistics have been used in the corpus linguistics literature: chi-squared (as employed by Hofland and Johansson, 1982) and log-likelihood (LL) (described by Rayson and Garside, 2000 for use in keywords calculations and by Dunning, 1993 for calculating collocations). Rayson et al. (2004a) compared the reliability of the two statistics when used for keyness calculations under varying conditions (corpus size, frequency of words) and showed that the log-likelihood test is preferred over the chi-squared test since it is a more accurate statistic for heavily skewed comparisons (e.g. a high ratio of target corpus to reference corpus sizes or low frequency words), so here we present only the log-likelihood formula. First, we need to calculate the expected values (E) corresponding to each observed value (O) in Table 1 and then insert these values into the second equation below.

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

Here, N1 = c, and N2 = d. Hence, for E1 = c*(a+b)/(c+d) and E2 = d*(a+b)/(c+d). Note that the calculation of expected values takes account of the sizes of the corpora, so the raw frequency figures should be used in the table.

$$-2ln\lambda = 2\sum_i O_i ln\frac{O_i}{E_i}$$

The log-likelihood score in this case is LL = 2*((a*ln(a/E1)) + (b*ln(b/E2))).[3] [4] Once this calculation has been performed for each word and the resulting word list has been sorted on the LL value, we

[3] Online calculators and downloadable spreadsheets are available at http://corpora.lancs.ac.uk/sigtest/ and http://ucrel.lancs.ac.uk/llwizard.html

[4] It should be noted that this formula represents the 2-cell calculation (Rayson and Garside, 2000) which can be used since the contribution from the other two cells is fairly constant and does not affect the ranking order. Other tools, e.g. AntConc, and statistical calculators also support the 4-

can see the words that are most indicative (or characteristic) of the target corpus relative to the reference corpus at the top of the list. Words which occur in the two corpora with roughly similar relative frequencies appear lower down the sorted list. At this point in the method, different researchers have applied different cut-offs relating to position in the list—significance value or p-value—and there is little agreement about a preferred approach. We will return to this discussion in our critical assessment in Section 3.

Representative study 1

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,* 2577–2590.

This work set out to achieve two aims: 1) to discuss comparative keyword analysis as a possible substitute for the sort of qualitative thematic analysis that a large number of previous illness studies found basis in; and 2) to apply this method in analysing gender differences in the online discourse of people with breast and prostate cancer.

Previous studies of people with cancer found gender differences which align broadly with findings in more wide-reaching sociolinguistic analyses, namely that women's style is more expressive compared to men's style, which is more instrumental (Boneva and Kraut, 2002). Seale et al. (2006) expanded considerably on previous studies by incorporating tools from corpus linguistics, notably keyness analysis.

The corpora utilised contained two types of data: research interviews and Internet-based support groups. Qualitative interviews were adopted for secondary analysis from the Database of Individual Patient Experiences project; these were conducted in the UK, 2000-2001, with 97 people with cancer (45 women with breast cancer; 52 men with prostate cancer), totalling 727,100 words. Posts were inspected to extract only those written by people with cancer (as opposed to family members, carers, or those experiencing symptoms that may be associated with cancer), which resulted in a final corpus comprising 12,757 posts and 1,629,370 words.

Often, stylistic, grammatical, or syntactical features of the target corpus are highlighted through keyness comparison with a general reference corpus. However, the comparative keyword analysis reported here did not involve a general reference corpus (which may conflate or obscure gender differences in the specific data sets. Instead, the breast cancer texts were compared to the prostate cancer texts to facilitate analysis of meanings made by their female and male authors, respectively.

Keywords were calculated using WordSmith Tools (Scott, 2004). Measures of 'keyness' (expressed in positive and negative log-likelihood values with the prostate cancer corpus serving as the target corpus) were provided. WordSmith Tools also provides their corresponding p-values, $p<0.00000001$ for all items, rendering this an ineffective method of describing or differentiating results. The 'top 300' results from both the breast and prostate cancer corpora were analysed, with some exclusions. Concordances and clusters around keywords were analysed by hand, and keywords were manually

cell calculation incorporating contributions from frequencies of the other words into the Log-Likelihood value.

placed into semantic categories. In the opinion of the researchers, "[t]his enabled important and meaningful comparative aspects of these large bodies of text to be identified" in a "more economical and potentially replicable manner than conventional qualitative thematic analysis based on coding and retrieval" (ibid.). This answered the first broad purpose of the paper, as introduced above.

Analysis of the interviews and web fora led to findings broadly aligned with previous research on gender differences in the experience of serious illness. Qualitative analysis of the interview corpus also supported findings from previous studies: women were more likely to claim to seek social support on the Internet, whereas men said that they use the Internet to look for information. Quantitative analysis using keyness helped to identify further areas of interesting difference. Compared to women with breast cancer, men with prostate cancer had a much greater number of keywords pertaining to TREATMENT (i.e. *catheter, brachytherapy, hormone, Zoladex, treatment*), TESTS AND DIAGNOSIS (*biopsy, MRI, screening*), SYMPTOMS AND SIDE EFFECTS (*incontinence, impotence*), and DISEASE AND ITS PROGRESSION (*PSA* [prostate specific antigen], *staging, cancer, aggressive*). By contrast, women with breast cancer had a greater number of keywords under categories such as SUPPORT (i.e. *help, supportive*), FEELINGS (*scared, hope, depressed*), PEOPLE (*I, you, husband, ladies*), CLOTHING AND APPEARANCE (*wear, clothes*), and SUPERLATIVES (*lovely, definitely, wonderful*).

Grouping keywords into a range of semantic categories helps to generalise and distil meanings. Viewed as a whole, findings in Seale et al. suggested "that men's experience of their disease appears to be more localised on particular areas of the body, while women's experience is more holistic" (2006, p. 2588).

Seale et al. (2006) conceded that the study had a number of limitations. The first limitation was to do with sampling: as individual posts had not been linked to poster identity, there was the possibility that overuse of certain keywords by specific individuals has resulted in patterns being interpreted as common across the corpus/sample as a whole. As with other keyness studies, the researchers acknowledged that this work focussed on difference rather than similarity, which has the effect of reifying gender differences. Finally, more complex syntactical or semantic patterns (such as tag questions or cognitive metaphors) were not highlighted and discussed with this methodology or in this study.

Other limitations *not* acknowledged by the researchers were also present in this study. The use of WordSmith's 'keyness measure' was used to rank results, with the 'top 300' skimmed from each corpus for further analysis. However, with all p-values nearing zero and no thresholds of log-likelihood included, it is unclear whether the most salient semantic categories were, in fact, included and populated. Finally, while we acknowledge that the creation of *ad hoc* semantic categories is useful for thematic content analysis, particularly when the researchers are very well-versed in the content of their corpora, we wonder about the accompanying detriment that this brings, particularly in relation to replicability. A number of other studies make use of further computational methods to undertake semantic annotation and categorisation. Below, we summarise one such work.

Representative study 2

Culpeper, J. (2009). Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29-59.

Culpeper (2009) moved beyond analysis of keywords in isolation to consider both key parts-of-speech and semantic categories in a corpus stylistic analysis of character-talk in *Romeo and Juliet.*

In this study, the speech of the six characters with the highest frequency of speech in *Romeo and Juliet* was isolated to create subcorpora varying in length from 1,293 to 5,031 words. Culpeper posited that studying these subcorpora would expose the differing speech styles of characters (which, in turn, contribute to reader perception of characterisation). In this case, the very small sizes of data under analysis also allowed for full consideration of all results.

Earlier literary studies have described (contextualised) frequency of features as indicators of an author's or character's style. Culpeper's aim was to "produce key items that reflect the distinctive styles of each character compared with the other characters in the same play, rather than ... stylistic features relating to differences of genre ... or aspects of the fictional world" (2009, p. 35). Therefore, in this study, the subcorpus of any individual character's speech was compared to a reference corpus of all other character's speech (exclusive).

This study also made use of WordSmith Tools (Scott, 2006) to calculate keywords. It found that keyword analysis did provide evidence for results that might be predictable (for instance, that "Romeo is all about love" (Culpeper, 2009, p. 53)), but also exposed features that were less easily observable (Juliet's subjunctive keywords, which may be linked to an anxious style), without relying on intuitions about which parts of the text or which features to focus on (ibid., p. 53). The study then went on to experiment with analysing key parts-of-speech and key semantic domains using a combination of software.

Key parts-of-speech and key semantic categories are calculated by applying the keywords procedure to part-of-speech and semantic tag frequency lists in the same way as described in section 2 for word frequency lists. Keyword analysis can be illuminating, but can also be misleading, because semantic similarity is not explicitly taken into account during analysis. While many (or even most) researchers who use keyword analysis end up grouping or discussing words in semantic categories, this can be a somewhat flawed method, as these categories are subjective, and words which are too infrequent to appear as key on their own will be discounted. In small corpora (such as the *Romeo and Juliet* corpus), many content words of interest would necessarily be low-frequency. The incorporation of a computational annotation system allows for systematic, rigorous annotation followed by statistical analysis.

In this study, semantic domains were annotated and analysed using UCREL's (University Centre for Computer Corpus Research on Language, based at Lancaster University) Semantic Analysis System (USAS) in Wmatrix. The input is part-of-speech tagged text produced by CLAWS, which is then run through SEMTAG, which uses lexicons to assigns semantic tag(s) to each lexical item or multiword unit (for details, see Wilson & Rayson 1993, 1996). The accuracy rate for SEMTAG is said to be approximately 91% for modern, general language (Rayson et al., 2004b), though the author acknowledged that semantic shift led to some 'incorrect' classifications. Some of these were corrected using a historical lexicon, but Culpeper cautiously checked and interpreted the results to be more certain of their meaning and importance.

In the analysis section of the paper, a number of key semantic categories and constituent items (words and set phrases) were presented for Romeo, Nurse, and Mercutio. An indicative selection of Romeo's key semantic categories appears in Table 2.

Semantic category, including the tag code and frequency	**Items within the category (and their raw frequencies) up to a maximum of ten types if they are available**
RELATIONSHIP: INTIMATE/SEXUAL (S3.2) (48)	***love*** (34), *kiss* (5), *lovers* (3), *kisses* (2), *paramour* (1), *wantons* (1), *chastity* (1), *in love* (1)
LIKING (E2+) (38)	***love*** (15), ***dear*** (13), *loving* (3), *precious* (2), *like* (1), *doting* (1), *amorous* (1), *loves* (1)
COLOUR AND COLOUR PATTERNS (O4.3) (33)	*light* (6), *bright* (4), *pale* (3), *dark* (3), *green* (2), *stained* (2), *black* (2), *golden* (1), *white* (1), *crimson* (1)

Table 2: Romeo's 'top three' semantic categories, as rank-ordered for positive keyness (i.e. relatively unusual overuse). Keywords—identified independently earlier in the paper—are emboldened. Adapted from Culpeper (2009, p. 48).

The first two semantic categories (RELATIONSHIP: INTIMATE/SEXUAL and LIKING) are clearly linked semantically and metonymically. The keyness of these themes was predictable, but did provide empirical and statistical evidence for topicality of Romeo's speech and a sense of his role as a lover. The third category in Table 2 is illustrative of key semantic domains which were much less predictable. Romeo describes literal light, but also makes use of conventional metaphors, such as light/dark for happiness/unhappiness. Exposure of metaphorical usage is an interesting and useful feature of the semantic tagger.

Interestingly, very few items within key semantic domains were also identified as keywords in this study. In Table 2 above, only three of the 26 listed items were also keywords; in the full table in Culpeper (2009, p. 48), only five out of 77 total items in key semantic categories were independently identified keywords. This result highlighted a major advantage of key semantic domain analysis: lexical items which may not turn up in keyword analysis due to low frequency combine together to highlight larger fields of key meaning.

As far as indicators of aboutness and style, all methods have merit. Key part-of-speech categories, keywords, and key semantic categories are usually dominated by a small number of very frequent items, allowing for overlap. Culpeper found that items identified as keywords dominate 66.6% of the semantic categories, meaning that a keyword analysis would reveal most conclusions, but also leave out a not-insignificant amount of findings. The areas not overlapping (particularly the 33.4% between key semantic domains and keywords) are very salient but also very difficult to predict. So, "[w]hen keywords are dominated by ideational keywords, capturing the 'aboutness' of the text, the part-of-speech and particularly the semantic keyness analyses have much more of a contribution to make, moving the analysis beyond what is revealed in the keywords" [ibid.].

3 Critical assessment and future directions

In this chapter so far, we have described the origins and motivations for the development of the keywords method in corpus linguistics, and shown two representative studies using the technique. In the next few sub-sections, we will undertake a critical assessment describing some of the pitfalls and misconceptions about the use of the technique, along with a summary of criticisms, concluding with future directions for the approach.

3.1 Corpus preparation

Given that the input to the keywords method is word frequency lists, then the specific details of the preparation of the lists are important—but often overlooked—in the corpus linguistics literature. What counts as a word in such lists (on the basis of tokenisation, punctuation, capitalisation, standardisation, etc.) can potentially make a large difference to the results. Usually, corpus software tools tokenise words by identifying boundaries with white space characters and removing any punctuation characters from the start and end of words. However, different texts and authors might use hyphenation differently; for instance, "todo", "to-do" and "to do", must be carefully cross-matched or standardised, or else the frequencies in contingency tables will not compare word types consistently. Most corpus tools avoid the capitalisation issue completely by forcing all characters to upper- or lowercase, but this can cause issues with different meanings, e.g. "Polish" versus "polish". Wmatrix makes an attempt to preserve capitalisation if words are tagged as proper nouns but this, in turn, relies on the accuracy of the POS tagger software.

Corpus methods are increasingly being applied to historical corpora, as we have seen in Culpeper (2009), described in Representative study 2. Many authors rely on modernised or standardised editions to avoid spelling variation issues. Baron et al. (2009) carried out a detailed study to assess the degree to which keyword results are affected by spelling variants in original editions. First, they estimated the extent of spelling variation in various large Early Modern English corpora and found that, on average, in texts from 1500, the percentage of variant types is over 70% and variant tokens is around 40%. In terms of preparing frequency lists, this means, for example, that rather than counting all occurrences of the word "would", corpus software needs to take account of frequencies for other potential variants: "wolde", "woolde", "wuld", "wulde", "wud", "wald", "vvould", "vvold", and so on. The amount of spelling variation drops down to less than 10% of types and tokens in corpus texts from around 1700 onwards. In terms of impact on the keyness method, spelling variation is a significant problem, since the rank ordering of words will be affected by the distribution of variant frequencies. Baron et al. (2009) estimated the difference with rank correlation coefficients on keyword lists calculated before and after standardisation and found that Kendall's Tau scores can drop as low as 0.6 (where a score of 1 indicates that the two lists are the same, 0 indicates that the rankings are independent and –1 indicates that one ranking is the reverse of the other; cf. Chap. 17). A similar effect will be observed when applying the keywords approach to computer-mediated communication (CMC) varieties e.g. online social media, emails, and SMS, so care must be taken with data preparation.

3.2 Focus on differences

One of the central drawbacks to keyness analysis is the innate focus on difference (and obfuscation of similarity). Baker (2004) undertook a comparative study on online erotica, and explained that while *large* is a keyword in gay male erotic texts compared to lesbian erotic narratives from the same erotica website, other semantically related words (i.e. *huge*) may have occurred with comparable frequency in both corpora. This may lead analysts to (erroneously) over-generalise the keyness of 'size' in the gay male corpus, overlooking the central tenet of keyword analysis, which is allowance for findings and discussion at the *lexical* level. Baker (2004) proposed one way to circumvent this focus on differences: to carry out comparisons on more than two sets of data. This is helpful when undertaking keyword analysis on two target corpora (rather than one target corpus compared to a reference corpus). By calculating keywords in two target corpora against one another and then, for instance, against a larger reference corpus, differences and similarities may be highlighted in the emerging results.

Another issue in keyword analysis highlighted by Baker (2004) is that the 'strongest' words tend to reveal obvious patterns. While this does provide confirmatory evidence in new studies that the technique is working as expected, this bias can contribute to an unmanageable number of unsurprising keywords being thrown up for analysis. Possible proposed work-arounds have already been demonstrated in some of the studies discussed above: researchers may apply cut-off points related to relative dispersion across texts (see also Chap. 5), frequency in the entire corpus, or maximum p-values, or even switch the focus to use dispersion instead of frequency for keyness calculations (Egbert and Biber, 2019). No robust guidelines as to 'appropriate' cut-offs for any of these measures have been recommended in the literature, which can be seen as a need for further development of the method. Similar issues with method settings and parameters can be observed in the area of collocation research (see Chap. 7).

3.3 Applications of statistics

There have been a number of criticisms of the keywords approach in relation to the application and interpretation of the significance test statistics used in the procedure. The method described in Section 2 can be seen as a goodness-of-fit test, where the null hypothesis is that there is no difference between the observed frequencies of a word in the two corpora. If the resulting metric (log-likelihood in our case) exceeds a certain critical value, then the null hypothesis can be rejected. After choosing a degree of confidence, we can use chi-squared statistical tables to find the critical value, e.g. for the 5% level ($p < 0.05$) the critical value is 3.84, and for 1% ($p < 0.01$), it is 6.63 (cf. Chap. 20). For a comparison of two corpora, we use values with 1 degree of freedom, i.e. one less than the number of corpora. However, if the value calculated from the contingency table does not exceed the critical value, this only indicates that there is not enough evidence to reject the null hypothesis and we cannot conclude that the null hypothesis is true (i.e. which would indicate that there is no significant difference).

It was Dunning (1993) who first brought the attention of the community to the log-likelihood test, proposing it for collocation analysis rather than keywords. Dunning cautioned that we should not rely on the assumption of a normal distribution when carrying out statistical text analysis and recommended log-likelihood as parametric analysis based on the binomial or multinomial distributions instead. There is some disagreement in the literature here, with some authors stating that chi-squared assumes a multinomial distribution, making no special distributional assumptions of normality. Cressie and Read (1984) showed that Pearson's X^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are two statistics in a continuum defined by the power-divergence family of statistics and refer to the long running discussion (since 1900) of the statistics and their appropriateness for contingency table analysis. Kilgarriff (1996) considered the Brown versus LOB corpus comparison by Hofland and Johansson (1982) and highlighted that too many common words were marked as significant using the chi-squared test. In order to better discriminate interesting from non-interesting results, he suggested making use of the Mann-Whitney test instead, as this makes use of frequency ranks rather than frequency directly. However, with a joint LOB/Brown frequency above 30 where the test could be applied, 60% of the word types were still marked as significant. Results using Mann-Whitney also suffer towards the low end of the frequency spectrum, especially when words have a frequency of zero in one of the two corpora. This is because a large number of words occur with the same frequency (indeed, usually half of the types in a corpus occur with a frequency of one), so they cannot be satisfactorily ranked. For tables with small expected frequencies, many researchers have used Yates' corrected chi-squared statistic (Y^2), and some prefer Fisher's exact test; for more discussion see Baron et al. (2009).

More recent papers have also investigated similar issues of statistical validity and appropriateness of the keywords procedure as currently envisaged for comparing corpora with specific designs. Brezina and Meyerhoff (2014, p. 1) showed that using a keywords approach to compare whole corpora "emphasises inter-group differences and ignores within group variation" in sociolinguistic studies. The problem is not the significance test itself, but rather the aggregation of frequency counts for a target linguistic variable, e.g. a word across speaker groupings. They recommend the Mann-Whitney U test instead, to take account of separate speaker frequency counts and variation within datasets. As Kilgarriff (2005) reminded us, language is not random, and the assumption of independence of words inherent in the chi-squared and log-likelihood tests "may lead to spurious conclusions when assessing the significance of differences in frequency counts between corpora" (Lijffijt et al., 2016: 395), particularly for poorly dispersed words. Paquot and Bestgen (2009) and Lijffijt et al. (2016) recommended representing the data differently in order to make the assumption about independence at the level of texts rather than the level of words. Lijffijt et al. (2016) recommended other tests that are appropriate for large corpora, such as Welch's t-test, the Wilcoxon rank-sum test and their own bootstrap test (see also Chap. 24). In response to Kilgarriff (2005), Gries (2005) pointed out the importance of multiple corrections for *post-hoc* testing (e.g. Bonferroni, or the more recent Šidák correction), since, after applying those, the expected proportion of significant results are observed. Gries (2005) also directed readers to other methods such as effect sizes, Bayesian statistics (later picked up by Wilson, 2013) and confidence intervals, and highlighted that null hypothesis significance testing has been criticised in other scientific disciplines for many decades. Concerns over the reproducibility and replicability of scientific results have led the editors of the *Basic and Applied Social Psychology* journal to ban p-values (null hypothesis significance testing) and the American Statistical Association produced a policy statement to discuss the issues (Wasserstein and Lazar, 2016).

Many misconceptions about statistical hypothesis testing methods are observable in the corpus linguistics literature and beyond; for further details, see Vasishth and Nicenboim (2016). One specific example that we can demonstrate here illustrates the usefulness of including effect size measures alongside significance statistics to allow for comparability across different sample sizes. As with significance metrics, there are a number of different effect size formulae that could be used. Effect size measures show the relative difference in sizes between word frequencies in two corpora, rather than factoring in how much evidence we have in the corpus samples. This means that, unlike log-likelihood measures, they are not affected by sample size. Consider three hypothetical experiments for the frequencies of the words 'blah', 'ping' and 'hoot' in four corpora, as show in Table 3. Here, we are using log-likelihood (LL) as our significance measure and Log Ratio (LR) as the effect size measure (Hardie, 2014). In experiment 1, LL tells us that there is enough evidence to reject the null hypothesis at $p < 0.0001$ (critical value 15.13) and the effect size shows the doubling of the frequency of the word in corpus 1 relative to corpus 2. Compare this with experiment 2, where the word frequencies and corpus sizes are all ten times larger than in experiment 1. As a result, the LL value is ten times larger, indicating more evidence for the difference, but the LR is still the same, given that the ratio of 1,000 to 500 is the same as the ratio of 100 to 50. In experiment 3, we retain the same sized corpora as in experiment 2, but the frequencies of the word are closer together, and they illustrate that a smaller relative frequency difference is still shown to be significant at the same p-value as in experiment 1. Importantly, we should note the lack of comparability of the LL score between experiments 1 and 2 (as well as between 1 and 3) because they employ differently sized corpora. In contrast, effect size scores can be compared across all three experiments without the same concerns.

Experiment	Word frequencies and corpus sizes	Significance and effect size results
1	Corpus 1 and 2 contain 10,000 words each. Frequency of 'blah' in corpus 1 = 100 Frequency of 'blah' in corpus 2 = 50	Significance (LL) = 16.99 Effect size (LR) = 1.00
2	Corpus 3 and 4 contain 100,000 words each. Frequency of 'ping' in corpus 3 = 1,000 Frequency of 'ping' in corpus 4 = 500	Significance (LL) = 169.90 Effect size (LR) = 1.00
3	Corpus 3 and 4 contain 100,000 words each. Frequency of 'hoot' in corpus 3 = 1,000 Frequency of 'hoot' in corpus 4 = 824	Significance (LL) = 17.01 Effect size (LR) = 0.28

Table 3: Three hypothetical keywords experiments

3.4 Clusters and n-grams

Both Baker (2004) and Rayson (2008) have pointed out a serious limitation of the keywords procedure, which is that it can really only be used to highlight lexical differences and not semantic differences. This means that a word which has one significant meaning might not be correctly signalled as key when its various senses are counted together, thus masking something of interest. To some extent, the procedure implemented in Wmatrix and employed by Culpeper (2009) as described in Section 3.2 will address this issue, because words are semantically tagged and disambiguated before the keyness procedure is applied.

Researchers may also be interested in (semantic) meaning beyond the single word. The USAS semantic tagger can be used to identify semantically meaningful multiword expressions (MWEs) since these chunks need to be analysed as belonging to one semantic category or are syntactic units e.g. phrasal verbs, compounds, non-compositional idiomatic expressions. Wmatrix then treats these MWEs as single elements in word lists, allowing key MWEs to emerge alongside keywords. Consider an example MWE 'send up'. If this were not identified in advance as a semantically meaningful chunk meaning 'to ridicule or parody', then separate word counts for 'send' and 'up' would be observed and merged with the other occurrences of those words in the corpus, potentially incorrectly inflating their frequencies.

Without the benefit of a semantic tagger, Mahlberg (2008) combines, for the first time, the keywords procedure with clusters or n-grams, i.e. repeated sequences of words counted in corpora. Once the clusters have been counted, then key clusters can be calculated using the same procedure as for keywords. Mahlberg then groups key clusters by function to draw conclusions about local textual functions in a corpus of Charles Dickens' writing, which formed the basis of a corpus stylistic investigation. This key clusters (or key n-grams) approach can be seen as an extension of the keywords approach. The simple keywords approach is, in fact, a comparison of n-grams of length 1. It has proved to be a very fruitful line of investigation with a number of other studies employing this method. Paquot (2013, 2014, 2017) used key clusters to identify French learners' lexical preferences and potential transfer effects from their native language. Additionally, others have used the key n-gram approach to support native language identification (Kyle et al, 2013) and automatically assessing essay quality (Crossley et al, 2013).

3.5 Future directions

Many current studies using the keywords method are on English corpora. As this method is readily available in software such as WordSmith and AntConc—which work well in most languages—more thought should be given to how well keywords work in languages other than English, especially where much more complex inflectional and derivational morphology occurs, e.g. Finnish. For these languages, it might be the case that comparing surface forms of words from the corpus works less well than comparing lemmas, because the frequencies are too widely dispersed across different word forms (in a similar way to historical spelling variants) to be comparable.

In terms of future directions for keyness analysis, we recommend that more care is taken in the application of the technique. Rather than blindly applying a simple method to compare two relative frequencies, more thought is required to consider the criticisms and shortcomings that have been expressed in the preceding sections. Any metadata subdivisions present within a target corpus or reference corpus should be better explored via comparison so that they are not hidden; the corpora should be carefully designed and constructed with the aim of answering specific research questions and facilitating comparability; issues such as tokenisation, lemmatisation, capitalisation, identification of n-grams and multi-word expressions, and spelling variation should be considered; and differences as well as similarities should be taken into account when undertaking the analysis of the keyword results. As a corpus community, we need to agree on better guidelines and expectations for filtering results in terms of minimum frequencies and significance and effect size values rather than relying on *ad hoc* solutions without proper justifications.

In the future, we recommend investigating the use of statistical power calculations in corpus linguistics. Power calculations can be used alongside significance testing and effect size calculations and are increasingly employed in other disciplines, e.g. psychology. Statistical power allows us to calculate the likelihood that an experiment will detect an effect (or difference in frequency in our case of comparing corpora) when there is an effect to be detected. We can use higher statistical power to reduce the probability of a Type-2 error, i.e. concluding that there is no difference in frequency of a word between two corpora, when there is in fact a difference. This might mean setting the effect size in advance and then calculating (*a-priori*) how big our corpora need to be, or at least being able to (*post-hoc*) calculate and compare the power of our corpus comparison experiments. This might help us answer the perennial question, 'How big should my corpus be?' and help researchers determine comparability and the relative sizes of sub-corpora defined by metadata such as socio-linguistic variables. Finally, related to the experimental design and interpretation of results, issues of corpus comparability, homogeneity and representativeness are highly important to consider alongside reliability of the statistical procedure (Rayson and Garside, 2000). It should not be forgotten that interpretation of the results of any automatic procedure is the responsibility of the linguist, and the results of the keywords method are a starting point to help guide us rather than an end point of the research.

4 Tools and resources

4.1 Tools

Keyness was arguably one of the later additions to the quiver of corpus linguistic tools; many papers published between 2001-2008 (including the two representative studies summarised in this chapter) discussed the infancy of its adoption. Now, however, this is considered one of the standard five methods of the field, alongside frequency, concordance, n-gram and collocation. As a result, nearly all concordancers and corpus linguistic tools will offer some assistance in the calculation of keyness. Distinguishing features, then, are:

1. the incorporation of more sophisticated **taggers**, allowing for calculation of key lemmas, parts-of-speech (POS), or semantic domains;
2. the inclusion of built-in **reference corpora**, often general corpora or subsections thereof, allowing for immediate calculation against a known 'benchmark' without the necessity of sourcing or collecting a comparable corpus;
3. the selection of **measures of keyness** available (see Section 3.3 for discussion).

We have provided an overview of popular tools and these features in Table 4.

Tool	*Keywords (*lemma-tised)*	*Key clusters*	*Key POS*	*Key semantic domains*	*Upload own corpora*	*Built-in reference corpora*	*Measures of keyness available*
AntConc v3.4.4	✓	✓			✓		Chi-square, LL
CQPweb v3.2.25	✓*	✓	✓	~		✓	LL, Log Ratio (unfiltered, LL filter, Confidence Interval filter)
SketchEngine	✓*	✓	✓	~	✓	✓	Simple maths[5]
Wmatrix v3	✓		✓	✓	✓	✓	LL, Log Ratio
WordSmith Tools v5	✓	✓			✓		Chi-square (Yates correction), LL

Table 4: Overview of available tools. A tick mark indicates full usability of a given feature; a tilde indicates partial capacity (i.e. beta development or use restricted to special access).

If a user has both a reference and target corpus and is simply interested in straightforward calculation of keywords, we can recommend both AntConc and WordSmith as good beginner-level tools for this method. CQPweb has the greatest variety of measures available; with robust tagging systems and a range of reference corpora, it also allows for calculation of keyness across features and genres. However, key semantic domains are difficult to access, and inability to upload target corpora may inhibit use for many users interested in exploring their own data. SketchEngine is extraordinarily powerful, with part-of-speech tagging and lemmatisation on a huge number of languages. However, semantic tagging is still under development, and some may disagree with the application of Simple Maths. The most powerful tool for semantic processing is inarguably Wmatrix. The main interface also offers easy access to key POS and keywords, and a small number of reference corpora are accessible. We recommend Wmatrix for keyness analysis, although with a caveat about size restrictions since it is currently suitable for corpora up to around five million words. The keywords method can also be implemented directly in programming languages such as Python, R and Perl (see Chapt. 9).

4.2 Resources (Word lists)

Generation of keywords in a target corpus necessitates some point of comparison, usually either a second target corpus, a reference corpus, or a word list from a large, general corpus. Selection of a reference corpus will impact the results, and some care should be taken to select an appropriate 'benchmark' to highlight differences aligned with a given research question. Many research questions necessitate the collection of specialised reference corpora, or entail comparison of

[5] For details, see https://www.sketchengine.eu/documentation/simple-maths/

subcorpora. Those wishing to answer more general questions (e.g. 'aboutness') may choose to make use of a general reference corpus. Word lists of many of the largest general corpora are readily available online; we have provided a sample of some English-language resources in Table 5, although care should be taken when selecting these to ensure that tokenisation decisions are well documented and comparable.

Source	*Description*	*Link*
BNC	A number of word lists from the British National corpus, including subcorpora divisions (e.g., written or spoken)	http://ucrel.lancs.ac.uk/bncfreq/flists.html
Brown family	Word lists from the 1961, 1991, and 2006 American and British Brown family corpora: Brown, LOB, Frown, FLOB, AmE06, and BE06	http://ucrel.lancs.ac.uk/wmatrix/freqlists/
COCA	A range of word and phrase lists from the Corpus of Contemporary American English	http://corpus.byu.edu/resources.asp
Kelly	Multilingual word lists of the most frequent 9,000 words in nine languages	https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/kelly.html
Moby	A range of word and phrase lists, including: five languages; root words, synonyms, and related words; the complete works of Shakespeare; with some lists part-of-speech or IPA coded	http://icon.shef.ac.uk/Moby/
Wiktionary	A huge range of word lists from a range of sources and domains (including, i.e., Project Gutenberg), from a large number of languages	https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists

Table 5: Selection of available word lists, with descriptions and weblinks.

5 Further Reading

Bondi, M. and Scott, M. (Eds.). (2010). Keyness in texts. Amsterdam: John Benjamins.

This is quite a comprehensive guide for scholars with particular interest in keywords and phrases. The collection is divided into three sections: 1) Exploring keyness; 2) Keyness in specialised discourse; and 3) Critical and educational perspectives. Section one deals with a number of the issues that we have touched upon here in greater detail, with leading scholars such as Stubbs and Scott outlining the main concepts and problems in keyword analysis. Sections two and three function as interesting collections of case studies on corpora drawn from engineering, politics, media, and textbooks, from a range of time periods and places.

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

This edited collection has a number of chapters of particular relevance for scholars interested in keyness. Mike Scott explores reference corpus selection and discusses the eventual impact on findings. Tony McEnery and Paul Baker have chapters using keyness to critically examine the discourses in media and politics, respectively. Those interested in Culpeper's (2009) paper above may like to read a wider study on Shakespeare's comedies and tragedies, by Archer, Culpeper, and Rayson. Finally, Archer makes an argument for wider incorporation of frequency and keyword extraction techniques in the closing chapter.

References

Baker, P. (2004). Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis. Journal of English Linguistics, 32(4).

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press. Cambridge: Cambridge University Press.

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

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. Anglistik, 20(1), 41-67.

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.

Cressie, N. & Read, T. (1984). Multinomial Goodness-of-Fit Tests. Journal of the Royal Statistical Society. Series B (Methodological) 46.3: 440-464.

Crossley, S. A., Defore, C., Kyle, K., Dai, J., & McNamara, D. S. (2013). Paragraph specific n-gram approaches to automatically assessing essay quality. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 216–219). Heidelberg, Berlin, Germany: Springer.

Culpeper, J. (2009). Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29-59.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19.1, 61-74.

Egbert, J. and Biber, D. (2019) Incorporating text dispersion into keyword analysis. *Corpora*, 14(1), 77-104.

Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. Corpus Linguistics and Linguistic Theory, 1(2), 277-294.

Hardie, A. (2014). Log Ratio – an informal introduction. CASS blog: http://cass.lancs.ac.uk/?p=1133 Accessed 17th July 2017.

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

Juilland, A., Brodin, D., and Davidovitch, C. (1970). Frequency dictionary of French words. Paris: Mouton & Co.

Kilgarriff, A. (1996) Why chi-square doesn't work, and an improved LOB-Brown comparison. Proceedings of the ALLC-ACH Conference. Bergen, Norway: 169-172.

Kilgarriff, A. (2005) Language is never ever ever random. Corpus Linguistics and Linguistic Theory 1 (2): 263-276.

Kyle, K., Crossley, S., Daim J., and McNamara, D. (2013) Native Language Identification: A Key N-gram Category Approach. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 242–250, Atlanta, Georgia, June 13 2013.

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

Mahlberg, M. (2008). Clusters, key clusters and local textual functions in Dickens. *Corpora, 2*(1), 1–31.

Murphy, B. (2010). *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk.* Amsterdam: John Benjamins.

Paquot, M. (2013). Lexical bundles and transfer effects. International Journal of Corpus Linguistics 18, 3, 391-417.

Paquot, M. (2014). Cross-linguistic influence and formulaic language: recurrent word sequences in French learner writing. In Leah Roberts, & Ineke Vedder, Jan Hulstijn, EUROSLA Yearbook. Amsterdam: Benjamins, 2014, pp. 216-237.

Paquot, M. (2017). L1 Frequency in Foreign Language Acquisition: Recurrent Word Combinations in French and Spanish EFL Learner Writing. *Second Language Research*. 33 (1): 13-32.

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction. In Jucker, A., Schreier, D., and Hundt, M. (eds), Corpora: Pragmatics and Discourse. Amsterdam: Rodopi, pp. 247–69.

Rayson, P. (2003). *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. Ph.D. thesis, Lancaster University.

Rayson, P. (2008). From key words to key semantic domains. International Journal of Corpus Linguistics. 13:4 pp. 519-549.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In 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, Hong Kong, pp. 1 - 6.

Rayson P., Berridge D. and Francis B. (2004a). Extending the Cochran rule for the comparison of word frequencies between corpora. In Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004, Presses universitaires de Louvain, pp. 926 - 936.

Rayson, P., Archer, D., Piao, S. L., and McEnery, T. (2004b). The UCREL semantic analysis system. In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal. Paris: European Language Resources Association, 7–12.

Scott, M. (1997) PC analysis of key words – and key key words. System, 25(2), 233-245.

Scott, M. (2004). WordSmith Tools. Version 4.0. Oxford: Oxford University Press.

Scott, M. (2008). WordSmith Tools Help Manual. Version 5.0. Liverpool: Lexical Analysis Software.

Scott, M. and Tribble, C. (2006) Textual Patterns: keyword and corpus analysis in language education. Amsterdam: John Benjamins.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tono, Y., Yamazaki, M. and Maekawa, K. (2013) A Frequency Dictionary of Japanese. Routledge.

Vasishth, S., and Nicenboim, B. (2016) Statistical Methods for Linguistic Research: Foundational Ideas – Part I. Language and Linguistics Compass, 10: 349–369. doi: 10.1111/lnc3.12201.

Wasserstein, R. L. & Lazar, N. A. (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician, 70:2, 129-133.
Wilson, A. (2013) Embracing Bayes factors for key item analysis in corpus linguistics. In: New approaches to the study of linguistic variability. Language Competence and Language Awareness in Europe. Peter Lang, Frankfurt, pp. 3-11.