Gabrielatos, C. (2018) Keyness Analysis: nature, metrics and techniques. In Taylor, C. & Marchi, A. (eds.) (2018) Corpus Approaches to Discourse: A critical review. Oxford: Routled...

Costas Gabrielatos
Edge Hill University

This is the **preprint** of the chapter published in: Taylor, C. & Marchi, A. (eds.) (2018) *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge. 225-258.

Please note that this is the submitted version (i.e. before peer review) and there are differences with the published version.

Keyness Analysis: nature, metrics and techniques

Costas Gabrielatos

1. Introduction

The notion of *keyness*, as it is understood in corpus linguistics,[1] was developed in the mid-to-late 1990s, and the procedure of keyness analysis was first incorporated in Wordsmith Tools (Scott, 1996). Scott (1997) defined a "key word" as "a word which occurs with unusual frequency in a given text […] by comparison with a reference corpus of some kind" (ibid.: 236). The focus of Scott (1997) was establishing words in a corpus, which, when grouped together in "culturally significant ways", would "provide a representation of socially important concepts" (ibid.: 233). It seems, then, that from its very introduction keyness analysis was used to examine issues that are at the heart of current corpus approaches to discourse studies. The notion of keyness is closely related to the notion of *aboutness*, that is, the understanding of the main concepts, topics or attitudes discussed in a text or corpus (Phillips, 1989: 7-10, 26, 53-54). Phillips (1989: 7) argues that "aboutness stems from the reader's appreciation of the large-scale organisation of text". The notion of aboutness informs work on keyness (e.g. Scott, 2001: 110) and may have influenced its development, in that a keyness analysis is a way to establish aboutness, for example, through the examination of (groups of) key words.[2] However, in Phillips (1989) aboutness was not established on the basis of frequency differences between (sub-)corpora, but on the examination of lexical patterns within a (sub-)corpus:

> [A] distributional analysis of textual substance, invoking no knowledge of the semantic content, the syntactic organisation or the lexical meaning of the text, would reveal the existence of global patternings in the lexis of the text. […] What the text is about may be specified by providing a semantic interpretation for the formally identified macrostructure. […] By relating the phenomenon of collocation to that of text macrostructure a beginning can be made on the explication of aboutness" (Phillips, 1989: 11, 16).

Despite the above difference, the two techniques share a core characteristic: the analysis does not take into account the meaning of the linguistic forms in focus; rather, considerations of meaning are only introduced in the and interpretation of results (Phillips, 1989: 21).

During the same period (mid-to-late 1990s), the notions of keyness and aboutness (although not described using these terms), were also extensively investigated by Kilgarriff (1996a, 1996b, 1997) within the framework of research on corpus similarity. Kilgarriff (1997: 233) posited that "any difference in the linguistic character of two corpora will leave its trace in differences between their word frequency lists", and that, in such an approach, "the individual meanings of texts are taken out of focus, to be replaced by the character of the whole" (ibid.: 232). The former statement can very well be seen as a justification for carrying out a keyness analysis, whereas the latter statement can be seen as describing the aboutness of a corpus. Of course, a keyness analysis on word-forms in two raw corpora (as is usually the case in

[1] See Stubbs (2010) for a discussion of different conceptions of the term 'keyword' and, indirectly, 'keyness'.
[2] For alternative approaches to establishing topics, see Gabrielatos, et al. (2012) and Jaworska & Nanda (2016).

corpus-based discourse studies) is “a fairly blunt instrument” (Gabrielatos & Baker, 2008: 28), as it does not cater for a host of linguistic features, most notably homography, multi-word units, polysemy, part of speech, and syntactic relations. However, even in this case, the results can be expected to be useful, as, for example, the different senses of a word-form can be expected to have different sets of collocates, at least some of which can be expected to be key. This can be shown through Kilgarriff’s (1997) example of the word-form *bank*. Let us assume that the corpora compared have similar frequencies of this word form as a noun, but different frequencies of its two senses (related to money and rivers). Even if the word-form itself is not key, the difference in content is expected to be revealed “because the one corpus will use *money, account* and *Barclays* more, the other, *river* and *grassy*” (Kilgarriff, 1997: 233).

At this point, we need to take into account that corpus linguistics research had been carrying out frequency comparisons between corpora long before the notion of keyness was introduced. For example, Aarts (1971/2004) used a sub-corpus from the Survey of English Usage to compare the frequency of different types of noun phrase (e.g. containing a pronoun or noun) in different syntactic positions (e.g. Subject or Object). Closer to the nature of keyness analysis as it is currently understood in corpus linguistics, Krogvig & Johansson (1985) compared the frequencies of the modal verbs *will*, *would*, *shall* and *should* in two general corpora of American and British English (Brown and LOB, respectively). In a study that can be seen as the first to use a corpus-based approach to discourse studies, and the first such study to employ keyness analysis (although without using this term), Leech & Fallon (1992) compared the frequencies of all the word-forms in the Brown and LOB corpora to study “social, institutional, linguistic, and other factors which distinguish one culture from another” (1992: 31).

The last two studies above also exemplify two broad approaches to frequency comparisons, which will be termed *focused* and *exploratory*, respectively (see also Gries, 2010a: 285). In Krogvig & Johansson (1985), the comparison focused on the frequency of particular language items in the two corpora, whereas in Leech & Fallon (1992), the frequencies of all words in the two corpora were compared. Focused frequency comparisons are carried out when the researchers have already decided on the linguistic item(s) that their study will examine, and have already formulated hypotheses or research questions, which the results of the pairwise frequency comparisons are expected to help address. In a focused approach, there is no limit to the selection of the unit of analysis, as such studies usually employ random samples of manageable sizes, which can be manually annotated for particular lexical groups, grammatical constructions, lexicogrammatical patterns, or semantic/pragmatic meanings. In this way, a study can establish whether, for example, a particular modal sense or grammatical construction is much more frequent in one of the two compared corpora. Exploratory frequency comparisons are not motivated by particular hypotheses, and any research questions that motivate them are expected to be quite general (e.g. What topics are mentioned more frequently in the two corpora?). Rather, in an exploratory approach, frequency comparisons are used as “a way in to texts”; as a technique for identifying linguistic items (usually words) that can indicate aboutness, and “repay further study” (Archer, 2009: 4-5), or generate hypotheses (Gries, 2010a: 285). Exploratory studies use automated techniques for both the frequency comparisons and the corpus tagging/annotation (if required). Once the unit of analysis is selected (e.g. word-forms, n-grams), the frequencies of all such units are compared (see also section 2 below). It would seem, then, that keyness analysis, particularly as it is usually used in corpus-based discourse studies, is an exploratory approach. However, exploratory and focused approaches are not entirely discrete, but can be combined, as shown in the two examples below.

- Example 1: The research starts with an exploratory approach, by deriving a list of key items ranked according to the value of the keyness metric used in the study. At this point, the researcher may switch to a targeted approach and select particular types of items for concordance analysis, according to explicit criteria, such as their normalised or raw frequency, part of speech, core sense, or relation to a particular topic.
- Example 2: The research starts with a targeted approach, by specifying items to be included in, or excluded from, the analysis (e.g. only nouns, or words below/above a frequency threshold). Members of the resulting key item list are then selected according to explicit criteria.

In light of the above, a keyness analysis is essentially a comparison of frequencies. As it is currently practised, it usually aims to identify large differences between the frequency of word-forms in two corpora (usually referred to as the *study* and *reference* corpus) – although there is increasing interest in using keyness analysis to establish similarity (Taylor, 2013), or absence (Partington, 2014), which can be seen as an extreme case of frequency difference (see also sections 3.2, 4.1 and 5). Unfortunately, the influence of practices in other quantitative disciplines, and contradicting definitions of keyness, led to the adoption of inappropriate metrics, which, in turn, led to a number of misconceptions relating to the nature of keyness and keyness analysis, the linguistic units that can be the focus of a keyness analysis, the metrics used to measure keyness, and the attributes of the corpora to be compared.

Of course, a study employing keyness analysis does not stop at the identification of key items; rather, this is only the first stage, as a manual analysis is required to establish the use of the items in context (e.g. Baker, 2006, Baker et al., 2008, 2013; Duguid, 2010; Partington et al., 2013). However, the accurate and principled (if not objective) identification of key items is crucial, as their selection will greatly influence the conclusions of such a study. That is, even when the manual analysis is thorough and context-informed, if the selection of key items is flawed, so are the results and conclusions. As the identification of key items, and the selection of those to be included in the manual analysis, is multifaceted and, currently, influenced by a number of misconceptions, it merits a detailed examination. Therefore, due to space limitations, the discussion related to the stage of manual analysis of key items falls beyond the scope of this chapter. The remainder of this chapter will first discuss the nature of *keyness* and *keyness analysis*, the definitions of which will then inform the discussion of the possible linguistic units that can be the focus of a keyness analysis, and the selection of appropriate metrics for establishing keyness. This section will also offer a brief historical overview of the notion of keyness and, more generally, the use of frequency comparisons in corpus linguistics. The chapter will then move on to considerations of principled (if not objective) techniques for selecting the key items to be included in the manual analysis, and issues relating to the selection of the corpora to be compared. The chapter will conclude with an example study. All the sections will incorporate discussion of relevant misconceptions.

2. Definitions and related issues

This section will focus on the definition of the terms *keyness*, *keyness analysis*, and *key item*, and will distinguish between the nature of keyness and the ways that keyness is measured. The definitions will be discussed extensively, as their nature informs the discussion of all other aspects, in particular, the selection of appropriate metrics for keyness, and the corpora to be compared.

It needs to be clarified from the outset that using ‘keyword’ as a default term for referring to the linguistic unit of focus in a keyness analysis is restricted and restricting – as is the term ‘keyword analysis’. Frequency comparisons can involve a host of other types of linguistic units, particularly if the corpus or sample has been lemmatised, or annotated for grammatical, syntactic, or semantic categories. For example, exploratory keyness studies have been carried out on lemmas (Utka, 2004), n-grams (Andersen, 2016), multi-word units (MWUs) (Gerbig, 2010), part of speech (POS) tags (Culpeper, 2009), lexicogrammatical patterns (Miki, 2011), and semantic fields (Rayson, 2008). In focused studies using random samples of manageable sizes there are no restrictions regarding the unit of frequency comparisons, as the feasibility of manual annotation allows focus on any type of linguistic unit (form or meaning) or level (e.g. semantic, pragmatic, discoursal). Perhaps the area of corpus studies in which this has been best demonstrated is research on learner language, particularly in studies employing corpus-based “contrastive interlanguage analysis” (Granger, 1996), one of the main foci of which is identifying over/underuse of particular linguistic items in learner language. Any over/underuse in learner language is established through frequency comparisons of the target linguistic features in L2 and L1 corpora/samples (e.g. Granger et al., 2011). Therefore, it would be more appropriate to use the term ‘keyword’ only when the frequency of word-forms is compared, and adopt the inclusive term *key item* proposed by Wilson (2013: 3). What also emerges from the discussion so far is that the type of keyness analysis typically employed in corpus-based discourse studies, that is, one involving the automated comparison of the frequency of word-forms in two raw corpora, is only one option among many, and it would be restricting to treat it as the default approach.

Definitions of the terms *keyness* or *keyword* have tended to conflate their nature with the proposed metric for measuring the level of keyness. Very early on, in the help manual of the first version of WordSmith Tools, keywords were defined as “words whose frequency is unusually high in comparison with some norm” (Scott, 1996: 53). It is straightforward to derive from this definition that a keyword is identified by way of a frequency comparison. It should clearly follow, then, that an appropriate metric for keyness would reflect the size of the frequency difference, and that the larger the difference, the more ‘key’ a word would be. However, elaborations on the definition also tied the nature of keywords to a different type of metric. For example, Scott (1998: 71) adds that “a word is said to be "key" if [...] its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a *p* value specified by the user”. In other words, the proposed metric for keyness was not the size of a frequency difference itself, but its statistical significance, or, simply put, the extent to which we can trust an observed frequency difference, irrespective of its size (see sections 2.2. and 2.3 for details). In adopting a statistical significance score as the indication of keyness, WordSmith Tools conformed to contemporary widespread practice in disciplines employing quantitative analyses (Ellis, 2010: viii; Ziliak & McCloskey, 2008: xv-xviii, 1-2). In fact, it is not unlikely that the wording of the definition of keywords was influenced by (or reflected) the choice of the particular statistical significance metric in Wordsmith Tools, log likelihood (G^2, also frequently indicated as LL). Dunning (1993) developed the log likelihood test in order to accurately identify the statistical significance of rare events. The focus on rare events seems to be reflected in the wording of early definitions: “unusually high [frequency]” (Scott, 1996: 53), “unusual frequency” (Scott, 1997: 236). However, this is not to say that, at the time (i.e. the mid-1990s), there was consensus among corpus linguists regarding the use of G^2

(or any other test of statistical significance) as a metric for frequency differences. Kilgarriff's work on corpus similarity, based on frequency comparisons, focused on critically examining different types of metrics (e.g. Kilgarriff, 1996a, 1996b, 1997; Kilgarriff & Rose, 1998) – a clear indication that, at the time, the issue of selecting/devising an the appropriate metric for frequency comparisons was anything but settled within corpus linguistics. This is also suggested by the variety of metrics used in corpus studies before 1996. For example, Aarts (1971/2004) used the Chi-Squared test (X^2), which returns the statistical significance of a frequency difference, whereas Krogvig & Johansson (1985) used the difference coefficient (Hofland & Johansson, 1982), a metric that reflects the size of a frequency difference, whereas Leech & Fallon (1992) combined the difference coefficient with the Chi-Squared (X^2) value – that is, they took into account both the size and statistical significance of frequency differences (see section 3 for a discussion of metrics). Soon after 1996, however, due to the availability of an affordable corpus tool that enabled corpus linguists to easily carry out automated frequency comparisons, and given that corpus linguistics researchers tend to rely on, and trust, corpus tools (Gries, 2010b: 124-125), the G^2 score (or the associated *p*-value)[3] was adopted as the metric for keyness by almost all corpus-based studies, as well as corpus tools (Gries, 2015: 55). Evidence for the widespread adoption of statistical significance scores as a metric for keyness comes from Pojanapunya & Watson Todd (2016: 3-10), who reviewed thirty studies employing keyness analysis, published between 2002 and 2013. Out of the twenty studies that specified the metric of keyness, all used a statistical significance metric (13 used G^2, 7 used X^2). It can also be expected that the studies that did not specify a keyness metric also used a statistical significance metric, as, at the time of the above studies, it was the default/only keyness metric available in almost all corpus tools. It is also interesting to note that, at the time when corpus linguistics was about to adopt a statistical significance metric to measure frequency differences, researchers in other fields (e.g. STEM, psychology) were vocally challenging its use as the main/only metric in their studies (e.g. Thompson, 1998). This is an important consideration in view of the very recent, and rather sudden, shift in corpus linguistics towards the use of effect size metrics for keyness, and the inclusion of a large number of other statistical metrics in corpus tools, not all of which measure effect size, or are appropriate for all types of keyness analysis. The next section discusses the issue of metrics and looks at the metrics currently offered in corpus tools.

3. Identifying key items: Appropriate metrics

A core distinction made in any current introductory book on statistics is between *effect size* and *statistical significance*. The effect size "indicates the magnitude of an observed finding" (Rosenfeld & Penrod, 2011: 342), that is, it shows "whether the difference or relationship we have found is strong or weak" (Mujis, 2010: 70, see also Ellis, 2010: 3-5). Statistical significance indicates "the high probability that the difference between two means or other finding based on a random sample is not the result of sampling error but reflects the characteristics of the population from which the sample was drawn" (Sirking, 2006: 306). Simply put, statistical significance does not reveal the size of a frequency difference, but, indirectly, the level of confidence we can have that the difference we have observed

[3] As readers may be familiar with different statistical significance tests (which may return different values for the same significance level), and as the values of every null-hypothesis significance test correspond to a *p*-value, the discussion of statistical significance will refer to *p*-values; however, the corresponding scores of the most commonly used significance test, log likelihood (G^2), will also be indicated. For reviews of different statistical significance tests, see Gries (2006, 2010a, 2010b, 2015), Hoffmann et al. (2008: 149-158), Kilgarriff (1996a, 1996b, 1997, 2005), Kilgarriff & Rose (1998), Paquot & Bestgen (2009), Rayson et al. (2004).

(however large or small) is dependable (e.g. Andrew, Pederson & McEvoy, 2011: 60; Sirking, 2006: 304).

3.1 Comparing effect-size and statistical significance

Statistical significance tests examine the *null hypothesis* (H_0); in the case of frequency comparisons, the null hypothesis would be that there is no real frequency difference, irrespective of the size of the observed difference. The values returned by significance tests correspond to particular *p*-values. Wilson (2013: 4) explains that "the *p*-value tells us the probability of obtaining an equal or more extreme result, given the null hypothesis […] If the *p*-value is very small, then one conventionally infers that either (a) a very rare event has occurred or (b) the null hypothesis is unlikely to be true", i.e. that it is unlikely that there is no frequency difference. Please note that the relationship between *p*-values and the level of statistical significance they indicate is an inverse one: the lower the *p*-value, the higher the statistical significance; whereas the relationship between the value returned by the statistical significance test and the statistical significance level it indicates is direct: the higher the value the higher the significance level.

Wilson (2013: 4) also stresses that the *p*-value should not be understood "as being the actual probability that an observed difference in proportional frequencies between two texts or corpora has occurred by chance" (see also Ellis, 2010: 17). For example, if p=0.01, this should *not* be interpreted as meaning that the frequency difference we have observed has a 1% probability of having occurred by chance, or, conversely, that we can be 99% confident that the observed frequency difference is real. Rather, it should be interpreted as meaning that there is a 1% chance that we would get the same or a larger frequency difference when, in reality, no such difference exists.

In view of the nature of keyness, and the nature of effect-size and statistical significance metrics, it is clear that keyness needs to be established via an effect-size metric (see also Gabrielatos & Marchi, 2011; Gries, 2010a: 284-285; Kilgarriff, 2001). It must also be stressed that effect-size and statistical significance metrics are not alternative measures of keyness, despite the size of the difference being, indirectly, taken into account in statistical significance tests. Simply put, the two metrics measure different aspects of a frequency difference. Killgarriff (2005: 264) observed that "there are number of papers in empirical linguistics literature where researchers […] used the confidence with which H_0 could be rejected as a measure of salience, whereas in fact they were merely testing whether they had enough data to reject H_0 with confidence". In addition to being an inappropriate method for measuring frequency differences, statistical significance tests exhibit a number of important limitations. All these issues are discussed below in detail.

Focused studies involving the manual examination of frequency differences of particular sets of words (Gabrielatos 2007; Gabrielatos & McEnery, 2005) revealed large discrepancies in the ranking between, on the one hand, values of frequency difference and, on the other, values of statistical significance. Using an exploratory approach, Gabrielatos & Marchi (2011) carried out frequency comparisons between specialised corpora of different sizes, and compared the ranking of scores derived from an effect size metric (the percent difference between the two normalised frequencies, %DIFF)[4] and a statistical significance one (log likelihood, LL), with a cut-off *p*-value of 0.01 (G^2=6.63). They used two large corpora, *SiBol*

[4] See Section 3.2 for details on this metric.

1993 (96 mil. words) and *SiBol 2005* (156 mil. words), each comprising all articles published in British broadsheets in 1993 and 2005, respectively, and two small corpora, comprising different sections from the *Guardian* in 2005: media section (1 mil. words) and home news section (6 mil. words). Gabrielatos & Marchi (2012) added three more comparisons, using a small specialized corpus (Hutton enquiry, 1 mil. words) and two general corpora, one small (FLOB, 1 mil. words) and one large (BNC, 100 mil. words). If the two types of metric were alternatives, then they would return the same ranking of keywords -- for example, the fiftieth keyword according to effect size would also be the fiftieth keyword according to statistical significance. In other words, the two rankings would fully correlate. Also, even if the two rankings did not fully correlate, the extent to which they did would provide useful indications regarding their similarity in assigning keyness scores. The correlations of the ranking returned by the effect-size and statistical significance metrics was measured using Spearman's Rank Correlation (r_s), a metric used when values "are measured on a ranked scale" (Ellis, 2010: 11). A value of '1' indicates full positive correlation (i.e. the two metrics produce identical rankings); a value of '-1' indicates full inverse correlation (i.e. the two metrics produce exactly opposite rankings); a value of '0' indicates no correlation (ibid.). The extent of the correlation can also be shown via the graphical plotting of each keyword's rank according to each metric. For example, a full positive correlation (r_s=1) would produce a straight line, ascending from left to right (Figure 1). The correlation analysis of the two rankings (by effect-size and statistical significance) revealed extremely weak correlations in all five comparisons, with r_s scores ranging from 0.010 to 0.122 (i.e. extremely close to no correlation). This is also clearly indicated in Figures 2-6 (which also show the r_s value). For example, in the comparison between Hutton and BNC, the word 'pound' ranked at position 12 according to LL, but position 10744 according to %DIFF. That is, it would be seen as a very strong candidate for analysis if statistical significance was used as a metric, but it would not be considered on the strength of the actual frequency difference (as shown by the effect-size metric). On the contrary, the rankings according to %DIFF and another effect-size metric (*Ratio*,[5] Kilgarriff, 2001) were identical for all keywords.

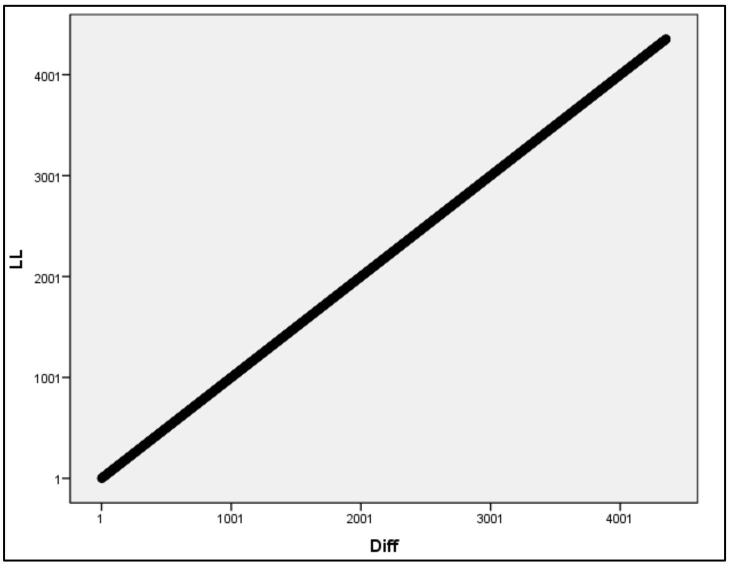

Figure 1. Putative full positive correlation between LL and %DIFF rankings (r_s=1)

[5] See section 3.2 for details.

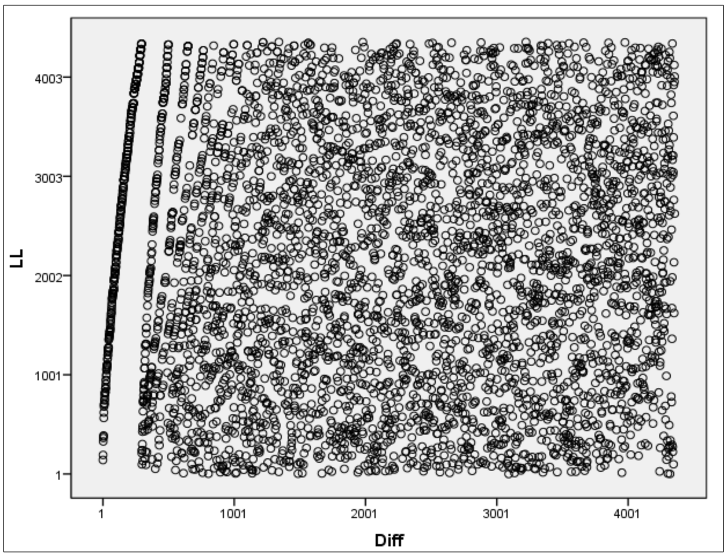

Figure 2. Correlation of rankings in the comparison between SiBol 1993 and SiBol 2005 (r_s= 0.010)

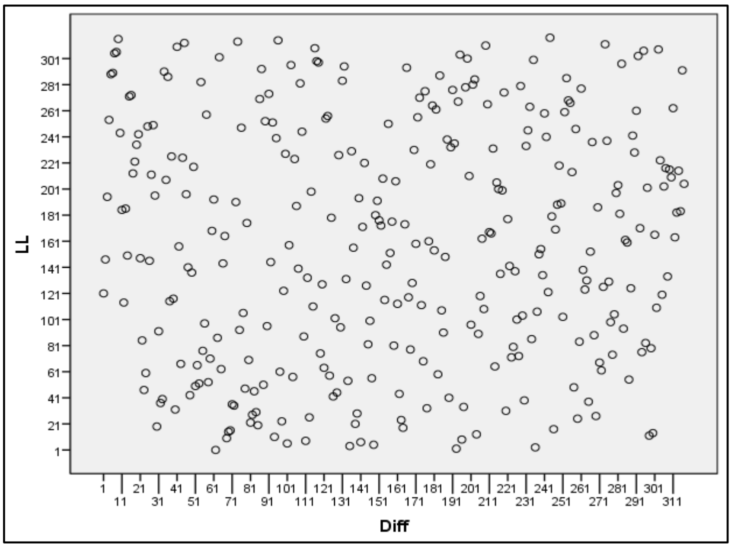

Figure 3. Correlation of rankings in the comparison between the two Guardian sections (r_s= 0.040)

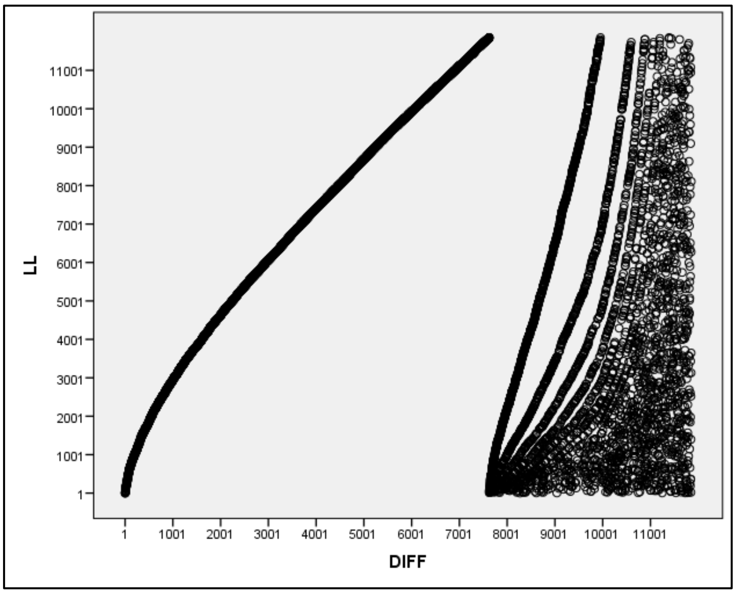

Figure 4. Correlation of rankings in the comparison between Hutton and BNC (r_s= 0.094)

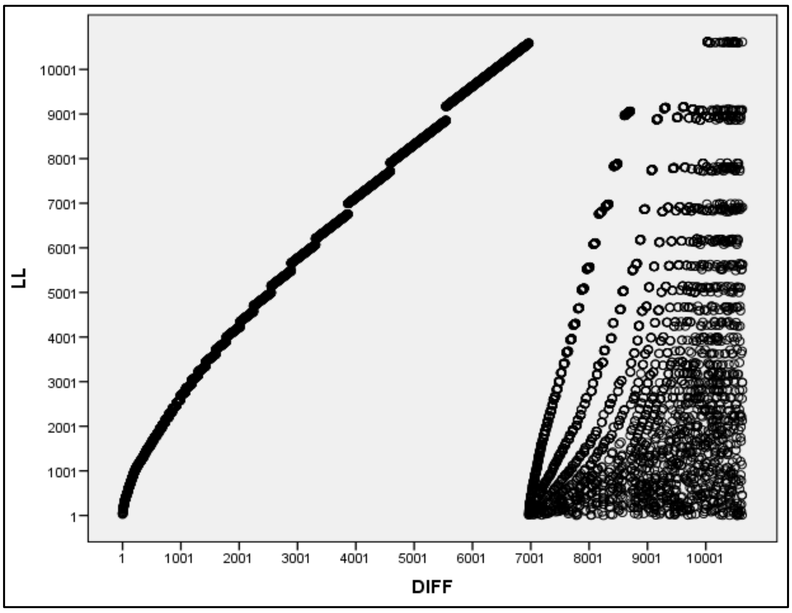

Figure 5. Correlation of rankings in the comparison between Hutton and FLOB (r_s= 0.122)

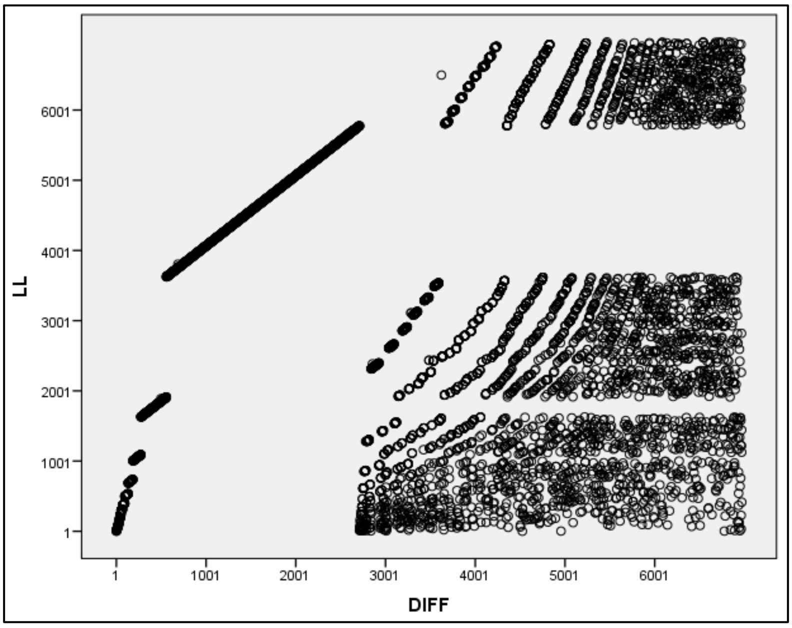

Figure 6. Correlation of rankings in the comparison between FLOB and BNC (r_s= 0.055)

However, Gabrielatos & Marchi (2012) also considered the possibility that the extremely low correlations between rankings might mask very small ranking differences among the top-N keywords. For example, a word might rank in position 10 according to one metric and 20 according to the other – which would mean that both words would be selected for analysis even if a small sub-set were chosen. To investigate that, they compared the overlap in the top 100 keywords returned by both metrics in all comparisons. Again, the analysis showed very little overlap (Table 1).

Table 1. Overlap in top-100 keywords returned by the two metrics

Compared corpora	Shared in top-100
SiBol 1993 vs. SiBol 2005	3
Guardian 2005: Media vs. Home	0
Hutton vs. BNC	2
Hutton vs. FLOB	8
FLOB vs. BNC	22

The above clearly indicate that the statistical significance score does not accurately reflect the size of a frequency difference. More precisely, Gabrielatos & Marchi (2011, 2012) observed the following cases:

- A very large frequency difference may have very low statistical significance.
- A very small frequency difference (even one so small that it could be deemed to show similarity rather than difference) may have very high statistical significance.
- Two very similar frequency differences may have very different levels of statistical significance.
- Two very different frequency differences may have very similar levels of statistical significance.

They concluded that statistical significance values are an unreliable and misleading measure of keyness, as they would exclude true keywords from the analysis and/or would result in treating a low-level keyword as a high-level one. Simply put, the two types of metric are not alternatives, as they measure different aspects of a frequency difference: its size and the level of confidence we can have this size is dependable.

The observations in Gabrielatos & Marchi (2011, 2012) can also be explained when we consider another aspect of statistical significance metrics. Statistical significance scores are sensitive to the size of the sample, in that, the larger the sample, the higher the statistical significance of all effect-sizes, however small they might be (Ellis, 2010: 5; Rosenfeld & Penrod, 2011: 84). In fact, “if we increase the sample […] we would ultimately reach the point where all null hypotheses would be rejected” (Owen & Jones, 1977: 359, cited in Kilgarriff, 1997: 237). In a keyness analysis, this sensitivity is related not only to the size of the compared corpora, but also to the corpus frequency of an item. That is, given a frequency difference, the higher the raw frequencies of an item in the two corpora and/or the larger the two corpora, the higher the statistical significance value will be. This is why function words (e.g. articles, prepositions), which are very frequent, tend to be among the top items when the results of a keyness analysis are ranked according to statistical significance (Kilgarriff, 1996a: 34), even when the actual frequency differences are very small. This also explains why items with similar effect-size scores can have very different statistical significance scores (Gabrielatos & Marchi, 2011, 2012). The corollary of this shortcoming is that statistical significance scores are not comparable across different keyness analyses. That is, an item may show the same effect-size in two different comparisons, but, because of its different corpus frequencies and/or different corpus sizes, the same effect size may have different levels of statistical significance in the different comparisons. For example, let us examine a number of hypothetical pairwise comparisons between three corpora: A (10 mil. words), B (100 mil. words) and C (1000 mil. words). Let us also assume that an item has different raw frequencies in these corpora, but in all of the pairwise comparisons this item shows the same very small frequency difference, say only 10%. If we calculate the statistical significance of the same frequency difference in the different comparisons, we derive the results shown in Table 2.[6]

Table 2. Different G^2 values for the same frequency difference

RF in A	RF in B	Size of A	Size of B	G^2
1100	10000	10,000,000	100,000,000	8.78
RF in A	RF in C	Size of A	Size of C	G^2
1100	100000	10,000,000	1,000,000,000	9.58
RF in B	RF in C	Size of B	Size of C	G^2
11000	100000	100,000,000	1,000,000,000	87.76

When examining the results in Table 2, it would be a mistake to conclude that our hypothetical item is much more key in the comparison between corpus B and corpus C, as in all three cases the frequency difference is the same. What we can only conclude regarding the latter comparison is that we can be much more confident that there is a very small frequency difference (Gabrielatos & Marchi, 2011). Conversely, a lower statistical significance does not necessarily indicate a smaller frequency difference, as it may be the result of the low frequency of an item and/or small corpus sizes. For example, if the frequency of our hypothetical item in corpora A and B was half of that shown in Table 2 (i.e. it was 550 and 5000, respectively), the frequency difference would be the same as before (10%), but it would have a much smaller statistical significance (G^2=4.39) – in fact, as this p-value is above 0.01 (the highest threshold usually used in exploratory keyness analyses), the item would not even be considered to be key. The above examples show that a high statistical significance does not necessarily indicate a large frequency difference, and a low statistical significance does not necessarily indicate a low frequency difference. The latter limitation also entails that statistical significance metrics cannot help pinpoint frequency similarities between corpora, whereas effect-size metrics can.

The sensitivity of statistical significance values to the size of one or both of the compared corpora results to the following characteristic: the larger the corpora compared, the higher the statistical significance value of all differences, however small (Rosendfeld & Penrod, 2011: 84;) and, therefore, the higher the number of frequency differences that will be statistically significant. This characteristic led to two related misconceptions: a) that there is an ideal range of corpus sizes, which returns an optimum number of keywords, and b) that the reference corpus must be larger than the study corpus (e.g. Berber-Sardinha, 2000). Of course, the smaller the corpora, the smaller the number of frequency differences that can be expected to cross the threshold of statistical significance. However, the objective of a keyness analysis is not to maximise, or minimize, the number of CKIs, but to derive as true a picture as possible of the differences and similarities of item frequencies between two corpora, using an effect-size metric -- corpus size is not as important as the representativeness of each corpus, and the principled selection of corpora to be compared.

Kilgarrif (1996b, 2005) argues against the use of null-hypothesis testing in corpus linguistics for two reasons. The first is that “language is never random, so the null hypothesis is never true” (Kilgarriff, 2005: 273). The second reason is related to the sensitivity of statistical significance values to corpus sizes (ibid.):

[6] RF= raw frequency.

> [H]ypothesis testing has been used to reach conclusions, where the difficulty in reaching the conclusion is caused by sparsity of data. But language data, in this age of information glut, is available in vast quantities. A better strategy will generally be to use more data. Then the difference between the motivated and the arbitrary will be evident without the use of compromised hypothesis testing.

However, the above should not be understood to imply that statistical significance metrics are useless in keyness analysis – quite the contrary, provided that we understand the nature and extent of the contribution of statistical significance to establishing keyness. In fact, Kilgarriff's (2005: 273) second argument can be seen to point towards the utility of using statistical significance testing when the corpora are small (e.g. data collection is difficult/costly, or the focus of the corpus is restricted). Also, Kilgarriff (2001: 239) states that G^2 "gives an accurate measure of how surprising an event is even where it has occurred only once" and that "early indications are that, at least for low and medium frequency words […] it corresponds reasonably well to human judgements of distinctiveness". In light of the above, statistical significance testing seems useful in cases of small corpora and/or items with low raw frequency -- when even large frequency differences may be unreliable. In such cases, statistical significance scores can indicate whether an observed large difference in the frequency of an item is also dependable enough to merit incorporating the item in the subsequent manual analysis. Therefore, a key item combines two characteristics: a) it exhibits a sizeable frequency difference between the compared corpora, and b) the difference is also statistically significant (Gabrielatos & Marchi, 2011, Gries, 2010b: 130). The nature of keyness, when focussing on differences, can be simply represented in a Venn diagram (Figure 7, Gabrielatos, 2014). The two circles can move closer or further apart depending on the set thresholds for effect-size and statistical significance: the stricter the thresholds, the smaller the shared area (i.e. the shorter the list of key items).

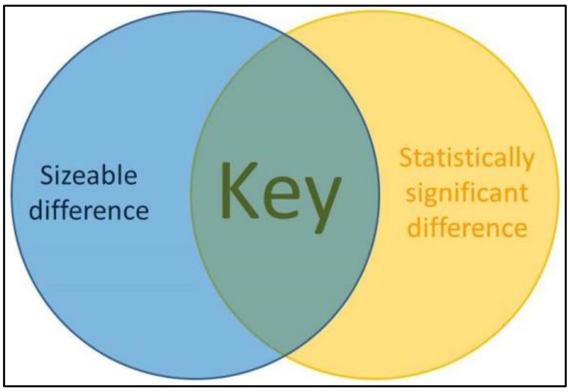

Figure 7. Keyness at a glance

3.2 Effect-size metrics

This section will examine the effect-size metrics available in the most widely used corpus tools: AntConc (Anthony, 2017),[7] WordSmith Tools (Scott, 2016), Sketch Engine (Kilgarriff et al., 2014). To these we should add the Excel document developed by Paul Rayson, which allows for both manual entry of raw frequencies and corpus sizes (useful for targeted keyness studies), as well as copy-pasting frequency lists derived in other corpus tools (useful for

[7] Please note that this relates to a version under development (3.5.0); the current version only offers a statistical significance metric.

exploratory studies).[8] At this point, we need to mention that the term ‘effect size’ may be a misnomer as far as keyness analysis is concerned. The choice of the term ‘effect’ seems to have been motivated by the use of such metrics in studies that aim to measure some kind of cause-effect relationship (e.g. the effect of a medical treatment or a teaching technique) or a correlation/association, that is, “the relationship between two variables” (Everitt, 2002: 20) (e.g. between the use of a particular linguistic item and sociolinguistic factors, such as age and gender).[9] This is clearly indicated by the wording of definitions in statistics textbooks, with the following definition in Ellis (2010: 3-4) being typical:

> An effect can be the result of a treatment revealed in a comparison between groups (e.g., treated and untreated groups) or it can describe the degree of association between two related variables (e.g., treatment dosage and health). An effect size refers to the magnitude of the result as it occurs, or would be found, in the population.

However, in a keyness analysis, at least as used in corpus-based discourse studies, no effect is measured; that is, the frequency of an item in one corpus is not expected to influence the frequency of, or interact with, the same item in another corpus. Therefore, measures of association (e.g. Dice Coefficient)[10] do not seem appropriate for a keyness analysis, unless, of course, what is compared is not the frequencies of items, but their ranking according to frequency in each corpus (e.g. Forsyth & Lam, 2009). Also, some effect size metrics focus on the difference of means in the compared datasets (e.g. Cohen’s *d*, Phi Coefficient). Again, this is irrelevant in a keyness analysis in corpus-based discourse studies, as what is compared is not means of groups of frequencies, but two distinct frequencies. Of course, such metrics are appropriate for other types of frequency comparisons. For example, in research on learner language, it is usually required to compare means of the frequency of particular items or types of errors in the output of learners grouped according to their proficiency level (e.g. Gablasova et al., 2017). Finally, some metrics that are presented as measuring effect-size in some corpus tools either measure statistical significance or are not purely effect-size metrics, as they include the value of a statistical significance metric in their calculation (e.g. BIC, Cramer’s V, Phi Coefficient, t-test); that is, they are “hybrid” metrics (Hoffmann et al., 2008: 151; see also Ellis, 2010: 10; Everitt, 2002: 285-286; Kilgarriff, 1996a: 35). In this light, the above types of metrics do not seem to be appropriate for keyness analysis.

This section will conclude with a discussion of five appropriate effect-size metrics (in order of simplicity) used in one or more of the corpus tools mentioned above, examining their calculation, the interpretation of values, and any particular characteristics or limitations. The calculation of all metrics below takes into account one or more of the following: the size of the compared corpora (C1, C2), the raw frequencies of an item in the two corpora (RFC1, RFC2), or the normalised frequencies of the item (NFC1, NFC2).

Ratio (Kilgarrif, 2009)

$$\boxed{\text{Ratio} = \frac{\text{NFC1}}{\text{NFC2}}}$$

[8] http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx (latest version, 4 July 2016). Paul Rayson also maintains a webpage offering a statistical significance calculator, as well as information on a large number of metrics: http://ucrel.lancs.ac.uk/llwizard.html
[9] For more examples, and a detailed outline, see Ellis (2010: 4, 7-15)
[10] See Rychly (2008) for a discussion on Dice and LogDice.

This is the simplest of the effect-size metrics, only involving the normalised frequencies of an item in the compared corpora. A value of ‘1’ indicates that the item has equal normalised frequency in the two corpora; higher/lower values indicate higher/lower NF in C1 – for example, a value of ‘4’ indicates that the item is four times more frequent in C1 than C2. However, it must be noted that the values are directional; i.e. they depend on which corpus is used as the study corpus. To use the example above, if C1 is the study corpus, then the value is ‘4’, whereas, if C2 is the study corpus, then the value will be ‘0.25’. Therefore, researchers using this metric need to be able to understand that the two scores (4 and 0.25) indicate the *same* size of difference, examined from two different perspectives. From the perspective of C1, the item has four times higher frequency, whereas from the perspective of C2 the item has four times lower frequency.

Odds Ratio (OR) (Everitt, 2002: 271; Pojanapunya & Watson Todd, 2016: 15)

$$\text{OR} = \frac{\text{RFC1} / (\text{C1} - \text{RFC1})}{\text{RFC2} / (\text{C2} - \text{RFC2})}$$

This metric takes into account raw frequencies, but also the sizes of the compared corpora. As in the case of Ratio, its values are directional.

Log Ratio (Hardie, 2014)

$$\text{Log Ratio} = \log \frac{\text{NFC1}}{\text{NFC2}}$$

This metric is the binary logarithm of the ratio of normalised frequencies and, therefore, produces different values. Equal normalised frequencies are indicated by a value of ‘0’, whereas ‘1’ indicates that NFC1 is twice NFC2. Overall, an increase of one in the Log Ratio value indicates a doubling of the frequency difference. For example, a value of ‘2’ indicates that NFC1 is four times NFC2. An advantage of Log Ratio is that, although it is a directional metric, the directionality does not manifest itself in different values (as in the case of the other directional metrics), but in the same value being positive or negative. Let us use the example given in the discussion of Ratio. If RFC1 is four times RFC2, the Ratio value would be ‘4’ if C1 was the study corpus, but ‘0.25’ if C2 was the study corpus – whereas, in the case of Log Ratio, the values would be ‘2’ and ‘-2’ respectively.

%DIFF (Gabrielatos & Marchi, 2011)

$$\%\text{DIFF} = \frac{(\text{NFC1}-\text{NFC2}) * 100}{\text{NFC2}}$$

This metric takes into account the normalised frequencies of an item in the compared corpora. Equal normalised frequencies are indicated by a value of ‘0’. Positive values show higher frequency and negative values indicate lower frequency. A value of ‘100’ indicates twice the frequency, and every increase of ‘100’ adds one to the difference – for example a value of ‘500’ indicates six times higher frequency. It is also a directional metric; in the above example, when C1 is the study corpus, the value is ‘300’, whereas when C2 is the study corpus, the value is ‘-75’. In the latter case, the interpretation is that the item has 75% lower frequency in C2 compared to C1, or, in different terms, that the frequency of the item in C2 is one-quarter of its frequency in C1. The limitation of this metric is that, although its score has no upper limit, there is a lower limit: negative scores stop at '-100'.

Difference Coefficient (Hofland & Johansson, 1982)

$$\text{Diff Coefficient} = \frac{\text{NFC1} - \text{NFC2}}{\text{NFC1} + \text{NFC2}}$$

As above, this metric takes account of normalised frequencies. The scores range from '1' to '-1', and are interpreted as follows: '1' indicates that the item only exists in C1 (i.e. it has zero frequency in C2); '0' indicates that the item has the same normalised frequency in the two corpora; '-1' indicates that the item only exists in C2 (i.e. it has zero frequency in C1). That is, although the metric is directional, its values do not create problems of comparison, due to the plus/minus sign. However, the interpretation of values is less straightforward. For example, if (as in the example above) NFC1 is four times NFC2, the value is '0.6'.

Regarding the directionality of values, when effect-size values are reported, it must be made clear which corpus was treated as the study corpus (i.e. which corpus was first in the comparison). Overall, however, it is simpler if values are reported in terms of the higher frequency. What is important is that all the above metrics return the same ranking of CKIs. Therefore, the selection of one over another hinges on their availability in corpus tools, and the extent to which particular researchers find their values easy to interpret.

A limitation of all but one (Difference Coefficient) of the above metrics is that, when an item has zero frequency in C2, the calculation cannot be performed (due to division by zero). Three techniques to dealing with this limitation have been proposed. One technique is to remove items with zero frequency from the comparison. However, excluding such instances may well remove very useful differences. If, as is the case in a keyness analysis, we think it is interesting that a corpus has more occurrences of an item compared to another corpus, then it is even more interesting that one corpus has no occurrences when the other corpus has some. This is because these items can be seen as characterising not only of the corpus with non-zero occurrences, but also the corpus with zero occurrences – particularly if these items are related by meaning or use. The importance of zero (and very low) frequencies in a corpus increases with a) the frequency of the item in the other corpus, and b) the size of the corpus lacking the item. Simply put, the difference between nothing and something is potentially salient, and the larger the frequency/corpus, the more salient the absence. The second technique, usually termed 'add 1' (Kilgarriff, 2009: 2), is to add a small number (no more than '1') to the frequency of every item in each corpus. However, this technique has two flaws. First, it increases the size of the corpora by the number of types in each (or a fraction, if a number smaller than '1' is added). Second, it increases frequencies unevenly: the smaller the frequency of an item, the higher the proportion of increase in frequency resulting from the addition of a fixed number. For example, if we add '1' to three items with frequencies of 100, 10, and 1, then the frequency of these items would increases by 1%, 10%, and 100%, respectively. The resulting increase in corpus sizes, and the non-proportionate increase in the frequency of individual items can be expected to skew the results of a keyness analysis. The third technique is to replace zero frequencies with an infinitesimally small number (0.000000000000000001 – one quadrillionth), which, for practical purposes, is an adequate proxy for zero (Gabrielatos & Marchi, 2011). The limitation of this technique is that it results in extremely high values in the effect size metrics with no upper limits (e.g. %DIFF, Log Ratio). However, this can also be seen as a strength, as it flags up instances of absence.

The next section will discuss the decisions that must be taken after the effect-size and statistical significance scores have been calculated.

4. Selecting key items for analysis

Unless the corpora compared are very similar, it is unlikely that a study employing an exploratory keyness approach can carry out a manual analysis of all key items. For example, all thirty of the studies using keyness analysis reviewed in Pojanapunya & Watson Todd (2016: 3-10) focused on a sub-set of key items. It follows then that the technique used to select key items for manual analysis is of paramount importance, as it will greatly influence (if not determine) the results of a study. Clear indications of the main techniques preferred are provided in the findings of Pojanapunya & Watson Todd's review (2016: 3-10):

a) More than half (16) of the studies selected the top N words. The number of keywords examined ranged from 10 to 1000, with the average being about 100.
b) About one in four (7) specified a threshold of statistical significance, usually a very high one (with *p*-values ranging from 0.05 to 10^{-14}),[11] which was expected to reduce the number of keywords returned from the automated comparison.
c) A small number of studies (2) combined a corpus frequency threshold with a statistical significance threshold.
d) One in six (5) selected keywords that were deemed to be related to particular topics.

Of course, as the studies above used statistical significance as a measure of keyness (and ranked key items by the statistical significance score), the top-N items were those with the highest statistical significance. Similarly, the studies that set a very high threshold of corpus frequency also derived items with the highest statistical significance (as statistical significance scores increase as corpus frequency increases). Therefore, there is little difference between approaches (a)-(c), which were employed in the vast majority (25/30) of the studies examined in Pojanapunya & Watson Todd (2016).

As was mentioned in section 2, the level of keyness of an item needs to be established via the combination of two metrics, which complement each other. The effect size score will enable the ranking of the items returned from the automated frequency comparison according to the size of the frequency difference. The statistical significance score will provide information regarding the level of confidence we can have that the observed frequency difference is dependable -- or, to look at the issue from a different perspective, whether the item is frequent enough and/or the corpora are large enough for the observed differences to be dependable. As statistical significance has been the standard metric for keyness so far, there is a small body of relevant work in corpus linguistics to draw on for guidelines regarding establishing an appropriate cut-off *p*-value. However, there is very little work on establishing effect size threshold values for keyness analysis. In other words, the inclusion of an item in the list returned by the automated frequency comparison does not necessarily entail that the item is key. In this light, it seems wise to initially view the items returned via the keyness function of a corpus tool as *candidate key items* (CKIs).[12] The remainder of this section will first discuss the issue of threshold values for item frequency and statistical significance, and

[11] Negative exponents of the number 10 indicate the result of dividing the number 1 by a number starting with 1 and followed by as many zeroes as specified by the exponent. So, 10^{-14} is 1/100000000000000, or 0.00000000000001.
[12] The term is influenced by the use of "candidate collocates" in Sketch Engine (Kilgarriff et al., 2014).

then propose an objective and versatile technique, based on effect-size values, for facilitating the selection of key items in exploratory keyness studies.

4.1 Frequency thresholds

As was shown in Pojanapunya & Watson Todd (2016: 3-10), the majority of keyness studies tend to set frequency thresholds, usually removing low-frequency items from the comparison, either directly or indirectly (via setting high statistical significance thresholds). However, this may have unintended consequences. For example, if C1 contains some items in very low frequency while C2 contains these items in a (relatively) much higher frequency, then these items can be expected to register high effect-size values. Applying a low-frequency threshold may remove potentially important items, which may index very pronounced differences (e.g. topics, attitudes). In the same vein, removing items with zero frequency in one of the compared corpora will prevent the examination of absence. The same applies to setting high-frequency thresholds, which are usually used to filter out function words. However, function words can point towards particular attitudinal differences between the two corpora (e.g. Duguid, 2008; McEnery, 2006). In his work on swearing, McEnery (2006: 147) found that syntactic co-ordinators, in particular the word ‘and’ demonstrated “the important function of linking objects of offence to form networks of offence”. If frequency thresholds are to be set, then they should be specified in terms of normalised frequencies (e.g. per million words; pmw), not raw frequencies. This is because, in corpora of uneven sizes, the same raw frequency may correspond to very uneven normalised frequencies. For instance, a raw frequency of 5 in a corpus of 10 million words, translates into a normalised frequency of 0.5 pmw, whereas in a corpus of 100,000 words, it translates into 50 pmw. However, even if thresholds are set according to normalised frequencies, they may remove important key items from consideration. As McEnery (2006: 148) puts it, “it is a brave, or rather foolish, analyst who assumes that, in any given data set, the words are so unlikely to be key that they can be safely ignored from the very start”. Therefore, it seems wise to avoid setting frequency thresholds, but generate a list of CKIs which include all items (i.e. all types in both corpora). Researchers can then make principled decisions regarding which items are to be examined, taking into account the effect-size and statistical significance of CKIs (see sections 4.2, 4.3 and 5), as well as the particular foci of the study.

4.2 Statistical significance thresholds

Before examining the utility of using statistical significance thresholds, that is, setting a *p*-value that is deemed low enough for results to be seen as statistically significant, we must consider that such thresholds are arbitrary (Hoffmann et al., 2008: 88). That is, researchers in a particular discipline may agree that a particular *p*-value indicates a level of confidence that they feel is high enough for results to be accepted as dependable. As a result, the threshold for statistical significance varies between disciplines. For example, in most of the social sciences the usual threshold is p=0.05, i.e. a *p*-value of/below 0.05 is regarded as indicating statistical significance (Wilson, 2013: 8), whereas in corpus linguistics the threshold is usually p=0.01 at the most. However, as keyness analyses (particularly of large corpora) tend to return too many CKIs for researchers to examine manually, the usual practice (as indicated in Pojanapunya & Watson Todd, 2016) is to set a much lower *p*-value (e.g. 0.000000001), partly in order to reduce the CKIs, and partly because of the misconception of the *p*-value as a measure of keyness – that is, setting a very low *p*-value threshold is expected to return only the items with the highest keyness. In light of the discussion so far, we need to examine two interrelated issues: a) the *p*-value that can be seen as low enough for the corresponding frequency difference to be deemed dependable, and b) the wisdom of setting extremely low *p*-value thresholds to reduce the number of CKIs returned by the automated frequency comparison.

It was clarified in section 2, that the *p*-value does not directly indicate the probability that an observed frequency difference is due to chance. However, this is not to say that this probability cannot be calculated; rather, a different statistical measure is needed. Wilson (2013: 5-6) proposes using the approximate Bayes Factor (BIC), the value of which provides an estimate of "the degree of evidence against the null hypothesis" (H_0). For the purposes of keyness analysis (i.e. pairwise frequency comparisons), BIC is calculated using a) the log-likelihood (LL) value of the frequency difference (provided by all current corpus tools) and b) the combined size of the compared corpora (N), as follows: $BIC \approx LL – \log(N)$.[13] The resulting value is interpreted as indicating the amount of evidence against H_0, as shown in Table 3 below (Raftery, 1999: 420; Wilson, 2013: 6).[14] However, it must be clarified that BIC is not an effect-size metric; it can be used instead of, or in addition to, other statistical significance tests, but it cannot replace an effect-size metric.

Table 3. BIC values and their interpretation

BIC	Degree of evidence against H_0
<0	No evidence – favours H_0
0-2	Not worth more than a bare mention
2-6	Positive evidence against H_0
6-10	Strong evidence against H_0
>10	Very strong evidence against H_0

An example frequency comparison carried out by Wilson (2013: 6-8) between two small corpora (approximately 10000 and 150000 words) yielded the following correspondence between *p*-values and BIC values, and the evidence level they indicate (Table 4).[15] As Wilson (2013: 8) points out, the BIC values in Table 4 suggest that the usual threshold of p=0.01 (G^2=6.63) provides considerably less than positive evidence. The values also put in perspective the threshold of p=0.0001 (G^2=15.13) proposed by Rayson, Beridge & Francis (2004), which, in light of Wilson's (2013: 8) results seems to provide at least positive evidence.

Table 4. Correspondence between p-values and degrees of evidence

BIC	Degree of evidence against H_0	*p*-value	G^2
2-6	Positive evidence against H_0	0.00018	13.98
6-10	Strong evidence against H_0	0.000014	18.81
>10	Very strong evidence against H_0	0.0000024	22.22

However, Wilson (2013: 7) clarifies that, as BIC takes into account the sizes of the compared corpora, "there will not always be a direct correspondence" between G^2 and BIC values. Given the sensitivity of G^2 values to corpus sizes, it would seem advisable to set statistical significance thresholds in terms of BIC values (Wilson, 2013: 8). However, currently, no corpus tool includes BIC as a metric, although it is included in Paul Rayson's Excel sheet. Until BIC is included in corpus tools, the following two approaches are possible. The

[13] The symbol '≈' indicates that the value is approximate.
[14] I am grateful to Andrew Wilson (Lancaster University) for clarifications on, and references for, BIC.
[15] Please note that *p*-values are rounded up.

simplest approach is to treat the correspondences in Table 4 as general guidelines for selecting a *p*-value threshold. A more reliable approach would be to set the corpus tool threshold to the highest acceptable *p*-value in corpus linguistics (i.e. p=0.01) and then copy-paste the corpus tool output to Paul Rayson's Excel sheet, and decide on which CKIs will be considered key based on the BIC values (see Section 5 for examples).

In light of the above, would it be reasonable to argue that the lower the *p*-value the better? The short answer is, no: this would privilege items with very high corpus frequency, which may not show very high frequency differences (effect sizes), and may well filter out key items with very high effect sizes, only because these items do not have very high corpus frequencies. Another limitation of such an approach is that, if very large effect sizes were filtered out (i.e. they were not present in the list returned by the automated frequency comparison), the researcher would not even be aware of their existence. As a result, this practice would be expected to remove useful key items, and reduce the scope for identifying groups or key items, which could help the analysis to more accurately identify patterns of use, and corresponding semantic preferences and discourse prosodies (see Baker, 2004; Leech & Fallon, 1992: 31). More precisely, given the *p*-value indicating the threshold for very strong evidence in Wilson's (2013) study (table 4 above), it would seem that a *p*-value threshold below 0.0000001 (i.e. above a G^2 score of 28.38)[16] would be inadvisable, as it could remove very large frequency differences from consideration, particularly if the items do not have extremely high corpus frequencies, or the corpora are not particularly large. Hoffmann et al. (2008: 88) suggest an alternative approach: "instead of using pre-defined thresholds, you […] can simply decide whether you are willing to take the risk indicated by the *p*-value". This approach allows researchers to have a clear view of CKIs and decide on the items to be included in the manual analysis after examining the range of frequency differences (via effect-size values), and the corresponding range of statistical significance levels, or, better still, levels of evidence against H_0 (via BIC scores). This approach is particularly useful when small corpora are compared, and even very high frequency differences can be expected to have low statistical significance. In such cases, the researcher can accept lower significance values than the ones shown in Table 3, and mitigate the corresponding discussion accordingly.

However, it must be clarified that the above approach is suitable only when differences are sought. If the study aims to identify similarities, then statistical significance thresholds should not be used. This is because, as was discussed in section 3.1, statistical significance scores indicate the level of confidence against H_0 – in other words, they show how confident we can be that there is a difference. As a result, statistical significance thresholds remove items with similar frequencies; that is, they remove the very items that the study seeks to identify. However, corpus tools always use statistical significance thresholds. Therefore, before carrying out a keyness analysis aiming to identify similarities, the maximum *p*-value must be set at '1', that is, the output of the frequency comparison will contain the effect-size and statistical significance values of all types in the compared corpora.

4.3 Effect-size thresholds

As the range of effect size values is expected to vary according to the level of difference or similarity between the two corpora (Gabrielatos & Marchi, 2011), such a threshold can be expected to be comparison-specific. Even with a high threshold of statistical significance,

[16] This p-value is derived by rounding down the value of p=0.0000024 in Table 4.

frequency comparisons are expected to return a wide range of effect sizes, some of which will be too small, at least compared to items higher up the list, and may even be small enough to effectively signal similarity. For example, a difference of 100% is comparatively very high if the majority of differences are below 50%, but comparatively very low if the majority is above 100%. The approach of selecting the top-N CKIs has two important limitations. First, it does not consider the proportion of key items that N represents; for example, the top 100 represent 10% of key items if the total is 1000, but 50% if the total is 200. Second, it does not consider whether there are items below rank position 100 which have only marginally lower scores than the 100th item. For example, it does not make sense to include the 100th item with a difference of 100%, but exclude the 101st item with a difference of 99.5%. Therefore, neither selecting the top-N CKIs nor setting a universal threshold would seem advisable.

The approach proposed here is adapted from Gabrielatos (2010: 52-54, 205-221) and Gries (2010a: 285-288). The CKIs are clustered according to their respective effect-size scores. The clustering method suggested is *hierarchical cluster analysis*: a family of statistical techniques used in assigning objects (in this case, CKIs) to groups according to their degree of similarity/dissimilarity in relation to one or more variables (in this case, the effect-size score) (Everitt, 1993: 1, 6-7; Gan *et al.*, 2007: 3-5, Romesburg, 1984: 2). More precisely, the "agglomerative" method is suggested, which initially treats each CKI as a separate cluster, and then combines CKIs to clusters according to the (dis)similarity of their effect-size scores (Everitt, 1993: 55-57; Gan *et al.*, 2007: 9). The degree of (dis)similarity is measured using the "Euclidian distance", which computes the square root of the sum of the squares of the pairwise differences in the effect-size scores (Gan *et al*., 2007: 326). The distance between clusters, or already established clusters and CKIs not yet assigned to a cluster, is calculated using "average group linkage": the average of the distances between all the scores in each cluster (Sneath & Sokal, 1973: 222). This determines the allocation of CKIs to clusters, as well as the conflation of existing clusters into more inclusive ones. This method has been shown to consistently produce clear and useful classifications (Adamson & Bawden, 1981: 208).

In order to accommodate the usual restriction in the number of CKIs that can be examined manually, an additional setting is to determine the number of clusters. The number of clusters will vary according to a) the number of CKIs and b) the number of key items that can be examined manually in the particular study (MEKIs). As a rule of thumb, the number of clusters should be the number of CKIs divided by the number of MEKIs. For example, if a keyness analysis returns 1000 CKIs, but only about 50 can be examined manually, then twenty clusters should be specified. Of course, as will be seen in section 5, CKIs are not necessarily grouped neatly in clusters of equal sizes. However, this calculation allows researchers to start from the cluster with the highest effect-size scores (if the focus is differences) or the lowest ones (if the focus is similarity) and, if the cluster does not contain enough CKIs, to move to the adjacent one. Of course, another option is to determine the same number of clusters for both CKI lists. Whatever the number of clusters, this approach results in a continuum of clustered CKIs ranked from highest to lowest frequency difference. What needs to be stressed is that, as CKIs are clustered according the close similarity of their effect-size scores, once one item in a cluster has been selected for manual analysis, all other items in the cluster must also be selected. The next section brings together all aspects discussed so far, and exemplifies the suggested procedures through a case study.

5. Selecting key items: a case study

5.1 Aims, data and methodology

This section will present a case study of keyness analysis, which a) examines both differences and similarities; b) takes into account both effect-size and statistical significance, and c) demonstrates different alternatives to the principled selection of KIs for further manual analysis. As was clarified in Section 1, the case study does not aim to carry out a manual analysis of CKIs; rather it is used as a springboard for a discussion of the methodological options and issues focused on in previous sections.

The corpora to be compared are the 2017 UK election manifestos of the Conservative (CM2017, 29954 words) and Labour (LM2017, 23691 words) parties. The largest frequency differences are expected to index aspects characterising each manifesto (as compared to the other), whereas the smallest differences are expected to index similarities. It will be shown that even with such small corpora, and fairly strict thresholds of statistical significance, there is still a good number of KIs that can be usefully included in the manual analysis. The texts were downloaded from Paul Rayson's *Wmatrix* webpage[17] (Rayson, 2003, 2009). The texts were converted from the original PDFs and automatically cleaned by Paul Rayson, but further manual cleaning was deemed necessary in order to remove (all of) the following: page numbers, chapter/section numbers; headers and footers, and characters indicating bullet points *(•)* and quotation marks *(&bquo;)*.

Two corpus tools were combined: WordSmith 7 (Scott, 2016), and Paul Rayson's Excel document. WordSmith 7 was used to derive frequency lists and lists of CKIs (from which only the raw frequencies of CKIs were retained). All other calculations were carried out using Paul Rayson's Excel document, as it offers more effect-size metrics and, more importantly, both G^2 and BIC scores. In Wordsmith 7, only 'positive' key items were derived, as both corpora alternated being the study and reference corpus. Also, in order to avoid removing items from consideration, the following settings were selected in WordSmith: The minimum word frequency was set to '1'. The maximum *p*-value was set to '1'; that is, initially, statistical significance was ignored. This allowed to derive an effect-size score for all types in the corpora, and be able to establish similarities as well as differences. This resulted in 2316 CKIs in CM2017 and 2657 CKIs in LM2017. In Excel, effect-size was measured by %DIFF with zeros replaced by 0.000000000000000001, and statistical significance was established via G^2 and BIC. The cluster analysis was carried out using SPSS 22.[18] Procedures of KI selection differed according to whether the focus was differences or similarities. In all instances, CKIs were ranked or clustered according to their effect-size score.

5.2 Identifying differences

<u>Differences: Alternative 1</u>

This approach filters out all differences with BIC<2, that is, only differences that show at least positive evidence against H_0 are retained. In the particular comparisons, a BIC value of '2' corresponded to G^2 scores of about 13 (p<0.001), which is similar to the G^2 score (13.98)

[17] http://ucrel.lancs.ac.uk/wmatrix/ukmanifestos2017.

[18] Please note that, in SPSS, 'average group linkage' is referred to as 'between-groups linkage'.

corresponding to BIC=2 in Wilson (2013: 8).[19] Due to the small size of the corpora, this leaves a very manageable number of KIs for both comparisons (particularly as their corpus frequencies are very low): 31 for CM2017 and 34 in LM2017 (Tables 5 and 6, respectively). Frequencies are normalised per thousand words (ptw);[20] CKIs are ranked according to effect-size.

A first observation is that, in both comparisons, some CKIs have zero frequencies in the other corpus (5 in CM2017, 14 in LM2017), despite the small item frequencies and corpus sizes. This supports the inclusion of zero-frequency items in keyness comparisons, as their exclusion would prevent pinpointing potentially useful absences. For example 'universities' and 'United Kingdom' do not appear at all in LM2017, whereas 'equality' and 'LGBT' are not mentioned at all in CM2017. Another interesting observation is that 'Labour' and 'Conservative' are CKIs in LM2017, but is not in CM2017.

Table 5. Differences: CKIs in CM2017 (BIC≥2)

CKIs in CM2017	RF CM2017	RF LM2017	NF (ptw) CM2017	NF (ptw) LM2017	%DIFF	G^2	*BIC*
UNITED	63	0	2.10	0	2.10E+17	*73.42*	*62.53*
KINGDOM	45	0	1.50	0	1.50E+17	*52.45*	*41.56*
UNIVERSITIES	16	0	0.53	0	5.34E+16	*18.65*	*7.76*
SHALL	15	0	0.50	0	5.01E+16	*17.48*	*6.59*
SHALE	12	0	0.40	0	4.01E+16	*13.99*	*3.10*
STABLE	20	1	0.67	0.04	1481.83	*16.90*	*6.01*
DATA	33	2	1.10	0.08	1205.01	*26.40*	*15.51*
BELIEVE	37	3	1.24	0.13	875.46	*26.71*	*15.82*
GENERATIONS	20	2	0.67	0.08	690.91	*13.17*	*2.28*
GO	20	2	0.67	0.08	690.91	*13.17*	*2.28*
ONLINE	26	3	0.87	0.13	585.46	*15.92*	*5.02*
IF	57	7	1.90	0.30	544.03	*33.69*	*22.80*
INSTITUTIONS	24	3	0.80	0.13	532.73	*14.04*	*3.15*
LEADERSHIP	24	3	0.80	0.13	532.73	*14.04*	*3.15*
TECHNICAL	24	3	0.80	0.13	532.73	*14.04*	*3.15*
OPPORTUNITY	24	3	0.80	0.13	532.73	*14.04*	*3.15*
TECHNOLOGY	30	4	1.00	0.17	493.18	*16.87*	*5.98*
DIGITAL	59	9	1.97	0.38	418.49	*30.32*	*19.43*
GREAT	39	6	1.30	0.25	414.09	*19.92*	*9.03*
STRONG	51	9	1.70	0.38	348.18	*23.42*	*12.53*
BETTER	45	9	1.50	0.38	295.46	*18.50*	*7.61*
WANT	40	8	1.34	0.34	295.46	*16.44*	*5.55*
HELP	79	17	2.64	0.72	267.54	*30.21*	*19.32*
UNION	47	11	1.57	0.46	237.94	*16.41*	*5.52*
WORLD	106	27	3.54	1.14	210.51	*33.46*	*22.57*
DO	66	17	2.20	0.72	207.06	*20.54*	*9.65*

[19] In CM2017, BIC=2.28 corresponded to G^2=13.17; in LM2017, BIC=2.19 corresponded to G^2=13.08.
[20] The usual normalisation per million words is not appropriate, as it does not make sense to normalise to a corpus size larger than the ones examined.

CONTINUE	82	22	2.74	0.93	194.79	*24.20*	*13.31*
BEST	48	13	1.60	0.55	192.03	*13.99*	*3.10*
SO	102	40	3.41	1.69	101.68	*15.41*	*4.52*
CAN	99	40	3.31	1.69	95.75	*13.92*	*3.03*
WE	949	419	31.68	17.69	79.14	*105.26*	*94.37*

Table 6. Differences: CKIs in LM2017 (BIC≥2)

CKIs in LM2017	**RF LM2017**	**RF CM2017**	**NF (ptw) LM2017**	**NF (ptw) CM2017**	**%DIFF**	***G*²**	***BIC***
LABOUR'S	21	0	0.89	0	8.86E+16	*34.33*	*23.44*
EQUALITY	19	0	0.80	0	8.02E+16	*31.06*	*20.17*
UNIONS	15	0	0.63	0	6.33E+16	*24.52*	*13.63*
LGBT	12	0	0.51	0	5.07E+16	*19.62*	*8.72*
REINSTATE	11	0	0.46	0	4.64E+16	*17.98*	*7.09*
SCRAP	10	0	0.42	0	4.22E+16	*16.35*	*5.46*
PRIVATISATION	9	0	0.38	0	3.80E+16	*14.71*	*3.82*
BANKS	9	0	0.38	0	3.80E+16	*14.71*	*3.82*
RENTERS	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
WOMEN'S	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
FAILURE	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
ENFORCE	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
EXTENDING	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
CENTRES	8	0	0.34	0	3.38E+16	*13.08*	*2.19*
LABOUR	319	3	13.47	0.10	13344.38	*490.90*	*480.01*
CUTS	24	2	1.01	0.07	1417.23	*27.46*	*16.57*
OFFICERS	12	1	0.51	0.03	1417.23	*13.73*	*2.84*
OWNERSHIP	20	2	0.84	0.07	1164.36	*21.62*	*10.73*
CRISIS	19	2	0.80	0.07	1101.14	*20.18*	*9.29*
GUARANTEE	18	3	0.76	0.10	658.62	*15.69*	*4.80*
REGIONAL	17	3	0.72	0.10	616.47	*14.38*	*3.49*
ARRANGEMENTS	16	3	0.68	0.10	574.33	*13.08*	*2.19*
VITAL	16	3	0.68	0.10	574.33	*13.08*	*2.19*
STAFF	22	5	0.93	0.17	456.32	*15.91*	*5.02*
RIGHTS	66	16	2.79	0.53	421.55	*45.59*	*34.70*
WOULD	22	6	0.93	0.20	363.60	*13.86*	*2.97*
WORKERS	62	17	2.62	0.57	361.12	*38.88*	*27.99*
STANDARDS	40	12	1.69	0.40	321.45	*23.19*	*12.30*
UNDER	35	12	1.48	0.40	268.77	*17.79*	*6.90*
BACK	34	12	1.44	0.40	258.24	*16.76*	*5.87*
CONSERVATIVES	50	19	2.11	0.63	232.73	*22.66*	*11.77*
JOBS	34	14	1.44	0.47	207.06	*13.94*	*3.05*
ALL	100	56	4.22	1.87	125.78	*25.04*	*14.15*
ON	215	168	9.08	5.61	61.81	*22.06*	*11.17*

Differences: Alternative 2

This approach is appropriate for keyness comparisons returning a large number of CKIs, or, irrespective of the number of CKIs, for studies preferring to base selection decisions on a fine-grained grouping of CKIs, rather than on a simple ranking. It is also suggested for studies preferring to start with a larger pool of CKIs, from which to select or remove particular types of items. For example, a study may focus only on nouns or verbs, as it aims to identify key social actors or processes (van Leeuwen, 1996). In the latter case, researchers can select a lower statistical significance threshold, irrespective of the BIC value. In the present case study, if a threshold of $p \leq 0.01$ ($G^2 = \geq 6.63$) is selected, about three times the number of CKIs is returned (92 for CM2017 and 107 for LM2017). Let us assume that a fine-grained grouping of these CKIs is required, with about ten CKIs per group. Using the simple formula presented in Section 5.1, these CKIs will need to be grouped in ten clusters (Tables 7 and 8 – numbers before CKIs indicate their ranking position). The way to interpret the clusters is as follows (other filtering criteria notwithstanding): a) CKIs in higher clusters are more key than CKIs in lower clusters, b) all CKIs sharing a cluster should be treated as equally key. A first observation is that CKIs do not cluster neatly in sets of ten; this is because the clustering takes into account the distance between the effect-size scores of consecutive CKIs. A second observation is that the two expanded sets of CKIs derived after lowering the statistical significance threshold contain all CKIs derived with the higher ones.

Table 7. Differences: CKIs in CM2017 ($G^2 \leq 0.01$) grouped in ten clusters

Cluster	Difference: CKIs in CM2017
1	1:UNITED
2	2:KINGDOM
3	3:UNIVERSITIES
4	4:SHALL
5	5:SHALE
6	6:YOUNGER; 7:AHEAD; 8:YOUR
7	9:EASIER; 10:MERITOCRACY
8	11:DESIGN; 12:MIGHT ; 13:ELDERLY; 14:COMPETITIVE; 15:DEEP; 16:ACTIVE; 17:ATTRACT; 18:PUPILS
9	19:EXCEPTIONAL; 20:THINGS; 21:LEADERS; 22:WRONG; 23:GLOBE; 24:EDINBURGH 25:REGULATORS; 26:EXPLORE; 27:COMBAT; 28:WORRY; 29:GOVERN
10	30:STABLE; 31:DATA; 32:PROSPEROUS; 33:DIFFICULT; 34:FRAMEWORK; 35:BELIEVE; 36:MUCH; 37:GENERATIONS; 38:GO; 39:INFORMATION; 40:ONLINE; 41:IF; 42:INSTITUTIONS; 43:LEADERSHIP; 44:TECHNICAL; 45:OPPORTUNITY; 46:TECHNOLOGY; 47:OLD; 48:SIGNIFICANT; 49:POOR; 50:DIGITAL; 51:GREAT; 52:REMAIN; 53:WORLD'S; 54:STRONG; 55:PARTNERSHIP; 56:THERESA; 57:BETTER 58:WANT; 59:MARKETS; 60:STRONGER; 61:HELP; 62:INTERESTS; 63:PROSPERITY 64:NATION; 65:UNION; 66:GREATER; 67:NOW; 68:WORLD; 69:DO; 70:TOGETHER 71:LEAVE; 72:SCHOOL; 73:CONTINUE; 74:BEST; 75:EUROPEAN; 76:RIGHT; 77:SHOULD; 78:ABOUT; 79:USE; 80:AROUND; 81:TAKE; 82:BRITISH; 83:SO; 84:THOSE 85:CAN; 86:MAKE; 87:WE; 88:THIS; 89:IT; 90:BRITAIN; 91:PEOPLE; 92:IN

Table 8. Differences: CKIs in LM2017 grouped in ten clusters

Clusters	Difference: CKIs LM2017
1	1:LABOUR'S
2	2:EQUALITY
3	3:UNIONS
4	4:LGBT

5	5:REINSTATE
6	6:SCRAP
7	7:PRIVATISATION; 8:BANKS
8	9:RENTERS; 10:WOMEN'S; 11:FAILURE; 12:ENFORCE; 13:EXTENDING; 14:CENTRES 15:NEGOTIATING; 16:PROBATION; 17:ADULT
9	18:PROCUREMENT; 19:INSECURE; 20:WAGES; 21:HIV; 22:TOURISM; 23:PRIORITISE 24:REINTRODUCE; 25:PROFIT; 26:YOUTH; 27:TRANSITION; 28:REVERSE 29:RESOLUTION; 30:NEGLECT; 31:ABOLISH; 32:PROFITS; 33:MATERNITY 34:OPERATIVE; 35:UNLIKE; 36:LIBRARIES; 37:RECOGNITION; 38:LATE 39:CONTROLS; 40:HANDS; 41:BALANCE; 42:MUSIC; 43:DELIVERS; 44:JUDICIAL 45:OPTIONS; 46:FARES
10	47:LABOUR; 48:CUTS; 49:OFFICERS; 50:UN; 51:FAILED; 52:OWNERSHIP; 53:EQUAL 54:ECONOMIES; 55:CRISIS; 56:WAR; 57:FORMS; 58:PEACE; 59:ALLOWANCE; 60:TARGETS; 61:FEES; 62:GUARANTEE; 63:REGIONAL; 64:LEGISLATION; 65:TRADING; 66:ARRANGEMENTS; 67:VITAL; 68:STAFF; 69:LED; 70:RANGE; 71:PLANS; 72:RIGHTS; 73:HOURS; 74:TOWARDS; 75:WOULD; 76:FULLY; 77:OWNED; 78:WORKERS; 79:DISABILITIES; 80:STANDARDS; 81:DISCRIMINATION; 82:FOOD; 83:UNDER; 84:BACK; 85:CLIMATE; 86:CONSULT; 87:CUT; 88:CONSERVATIVES; 89:PRIVATE; 90:JOBS; 91:ENVIRONMENTAL; 92:TRANSPORT; 93:INVEST; 94:WOMEN; 95:EMPLOYMENT; 96:SECTOR; 97:HOMES; 98:END; 99:MANY; 100:ALL; 101:FUNDING 102:PROTECT; 103:REVIEW; 104:BEEN; 105:COMMUNITIES; 106:INTO; 107:ON

5.3 Identifying similarities

Assuming that about a hundred CKIs for each corpus could be manually examined, the whole set of CKIs (2315 in CM2017 and 2656 in LM2017) was grouped into 232 and 266 clusters respectively, using the simple formula presented in Section 4.3 (Tables 9 and 10 – cluster '1' contains CKIs with the lowest %DIFF score). The smaller the frequency differences, the more a CKI can be deemed to index similarity. A first observation is that there is very little overlap between the CKIs in Tables 9 and 10. This is because each set contains CKIs with smallest frequency differences from the perspective of each corpus. Therefore, a study focusing on similarity would need to combine the two lists. Looking at CM2017 (Table 9), 92 CKIs show the smallest %DIFF scores, and are grouped in 74 clusters – a very fine-grained classification, as quite a large number of clusters was specified. If this was deemed unsatisfactory, a smaller number of clusters could have been specified. The %DIFF scores of the CKIs range from -.040% to 15.59% in Table 9, and between 0.68% to 18.53% in Table 10. BIC scores are between -6.19 and -10.89 in Table 9, and between -7.79 and -10.89 in Table 10 -- all indicating that H_0 (no difference) is strongly supported. If more CKIs can be examined, then CKIs in subsequent clusters can be added. If fewer items are needed, items in lower clusters can be removed, or, alternatively, a lower effect-size threshold can be set (e.g. %DIFF=5%).

Table 9. Similarities: CKIs with lowest %DIFF in CM2017

Cluster	Similarity: CKIs CM2017
1	1:COMPANIES
2	2:BUILD
3	3:HOUSING
4	4:FOR
5	5:AND
6	6:BRITAIN'S
7	7:TAKING
8	8:FAIRER
9	9:RECORD

10	10:NORTHERN
11	11:FROM
12	12:SUPPORT
13	13:WORKING
14	14:DEAL
15	15:TERM
16	16:BEFORE
17	17:TACKLE
18	18:PARENTS
19	19:SHARE
20	20:POLICIES
21	21:DISABILITY
22	22:RETAIN
23	23:AGREEMENT
24	24:GOVERNMENTS
25	25:GENDER
26	26:REFORMING
27	27:LAUNCH
28	28:PROMISE
29	29:REQUIRED
30	30:MEETING
31	31:RESPOND
32	32:MEMBERSHIP
33	33:FISCAL
34	34:PAYMENTS
35	35:FORM
36	36:IMPLEMENTATION
37	37:KIND
38	38:FOUND
39	39:INFLATION
40	40:TARIFF
41	41:CASES
42	42:STREET
43	43:VOTE
44	44:THING
45	45:TOP
46	46:USERS
47	47:THIRD
48	48:VETERANS
49	49:STARTING
50	50:DOUBLE
51	51:SEA
52	52:SCALE
53	53:DISABLED
54	54:COUNTER
55	55:SPECIFIC; 56:DECENT; 57:LAW; 58:INCREASE
56	59:OUR
57	60:SUSTAINABLE
58	61:GIVE
59	62:BETWEEN; 63:ADDRESS
60	64:TO
61	65:NEEDS

62	66:THE
63	67:FUTURE
64	68:CHANGES; 69:RESPONSIBILITY
65	70:CREATE
66	71:POWERS; 72:MAKING
67	73:BUSINESSES
68	74:COMMITMENT; 75:DEBT; 76:CENTRE; 77:CORPORATE; 78:LOOK
69	79:ENGLAND
70	80:HAVE
71	81:FUND; 82:KEY; 83:PLANNING; 84:STUDENTS; 85:RECEIVE
72	86:PERSONAL; 87:MARKET
73	88:DOMESTIC; 89:PROVIDING; 90:COUNCILS; 91:WHOLE
74	92:ACTION

Table 10. Similarities: CKIs with lowest %DIFF in LM2017

Cluster	**Similarity: CKIs LM2017**
1	1:WHICH
2	2:WITHIN
3	3:GIVING
4	4:CURRENT
5	5:HOLD
6	6:BANKING
7	7:BROADBAND
8	8:COVERAGE
9	9:DUE
10	10:PAYING
11	11:DIVERSE
12	12:GOVERNANCE
13	13:ROYAL
14	14:DIRECTLY
15	15:SECOND
16	16:EMPLOYED
17	17:SPEND
18	18:RECENT
19	19:NON
20	20:FUEL
21	21:TURN
22	22:HEALTHY
23	23:CAPACITY
24	24:AVERAGE
25	25:PRICES
26	26:CRIME
27	27:SYSTEM
28	28:OF
29	29:RURAL
30	30:SUCH

31	31:LEGISLATE
32	32:IRELAND
33	33:PENSIONERS
34	34:IMMEDIATE
35	35:COMPANY
36	36:DEVOLUTION
37	37:TIMES
38	38:PRINCIPLE
39	39:MEDICAL
40	40:UK
41	41:LOCAL
42	42:YEARS
43	43:POLICE
44	44:US
45	45:ECONOMY
46	46:NHS
47	47:GAP
48	48:DEVOLVED
49	49:ARE
50	50:GOVERNMENT
51	51:WILL
52	52:TOO
53	53:LIVING
54	54:PROGRAMME
55	55:CONSIDER
56	56:RUN
57	57:CURRICULUM
58	58:REPEAL
59	59:INTEREST
60	60:APPROPRIATE
61	61:TEN
62	62:WEALTH
63	63:TAKEN
64	64:FOCUS
65	65:A
66	66:ENERGY
67	67:WHEN
68	68:ACT
69	69:PROTECTIONS
70	70:PROPERLY
71	71:PREVENT
72	72:OFFICE
73	73:LEVELS
74	74:AT
75	75:HAS
76	76:STATE
77	77:CURRENTLY
78	78:UK'S
79	79:HIGH
80	80:DEVELOPMENT
81	81:TWO
82	82:LONDON
83	83:FOUR

84	84:FREE
85	85:FIRST
86	86:OUT
87	87:AN
88	88:HEALTH
89	89:OR
90	90:LEAST
91	91:PROMOTE
92	92:FACE
93	93:ENVIRONMENT
94	94:ESTABLISH
95	95:BOTH
96	96:FULL
97	97:EXISTING
98	98:ONE
99	99:ROLE
100	100:WITH

6. Conclusion

Keyness analysis can be used to identify not only difference, but also similarity and absence. Establishing keyness involves a multi-faceted process. Although keyness needs to be primarily established via an effect-size metric (reflecting the size of a frequency difference), this needs to be supplemented by a statistical significance metric (reflecting the level of confidence that an observed frequency difference is dependable). However, not all effect-size metrics used in statistics are appropriate for keyness analysis, particularly as the technique is used in discourse studies. Also, although statistical significance is a useful additional metric, its utility is limited to reflecting the level of reliability of a given frequency difference. High statistical significance does not necessarily indicate a large difference, nor does low statistical significance necessarily point at similarity. As *p*-values are sensitive to item frequency and corpus sizes, the same *p*-value can be expected to have different importance in different comparisons. A more useful way of establishing the level of confidence in a frequency difference is via the BIC score, which also allows for comparisons of statistical significance between studies. It is, therefore, recommended that all corpus tools include BIC among the statistical significance metrics they make available.

It was also shown that the reference corpus does not need to be larger than the study corpus. If the corpora are too small for an observed frequency difference to be dependable, this will be reflected in the BIC score. If the comparison does not yield enough dependable frequency differences, then the researchers can either select a lower statistical significance threshold, or accept that their study requires larger corpora. Neither does the reference corpus be a general corpus – as was shown in the case study. In fact, the terms ‘study’ and ‘reference’ corpus might be misleading. There is nothing intrinsic in a corpus that renders it a good selection for a ‘study’ or ‘reference’ role. Any two corpora can be compared, as long as their characteristics (nature, content, time-period) help address the particular research questions or hypotheses.

Finally, keyness is not a straightforward attribute. However objectively effect-size and statistical significance are calculated, the identification of an item as *key* depends on a multitude of decisions regarding frequency, effect-size, and statistical significance thresholds, and selections related to the word class or meaning of an item. It is, therefore, crucial that these decisions are both principled and explicitly stated. More precisely, studies need to report on, and justify, any thresholds, inclusion/exclusion of particular types of CKIs, and the proportion of CKIs selected for analysis.

References

Adamson, G.W. & Bawden, D. (1981) Comparison of hierarchical cluster analysis techniques for automatic classification of chemical structures. *Journal of Chemical Information and Computer Sciences*, 21, 204-209.

Aarts, F.G.A.M. (1971) On the distribution of noun-phrase types in English clause structure. *Lingua*, 26, 281-93. Reprinted in Sampson, G. & McCarthy, D. (eds.) (2004) *Corpus Linguistics: Readings in a widening discipline*. London: Continuum. 35-57.

Andersen, G. (2016) Using the corpus-driven method to chart discourse-pragmatic change. In Pichler, H. (ed.) *Discourse-Pragmatic Variation and Change in English: New methods and insights*. Cambridge: Cambridge University Press. 21-40.

Andrew, D.P.S., Pederson, P.M & McEvoy, C.D. (2011) *Research Methods in Design in Sport Management*.

Anthony, L. (2017) AntConc (Version 3.5.0 Dev) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net

Archer, D. (2009). Does frequency really matter? In Archer, D. (ed.), What's in a word list? (pp. 1-16). Farnham/Burlington: Ashgate.

Baker, P. (2004) Querying keywords: questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics*. 32(4), 346-359.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., Gabrielatos C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273-305.

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

Berber-Sardinha, T. (2000) Comparing corpora with WordSmith tools: How large must the reference corpus be? *Proceedings of the Workshop on Comparing Corpora* Vol. 9, Hong Kong 7 October 2000. 7-13

Culpeper, J. (2009) Keyness: 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.

Duguid, A. (2008). Men at work: How those at Number 10 construct their working identity. In Garzone,G & Sarangi, S. (eds.) *Discourse, Ideology and Specialized Communication*. Bern: Peter Lang. 453-484.

Duguid, A. (2010) Newspaper discourse informalisation: A diachronic comparison from keywords. *Corpora*, 5(2), 109-138.

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

Ellis, P.D. (2010). *The Essential Guide to Effect Sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Everitt, B.S. (1993) *Cluster Analysis* (3rd ed.). London: Edward Arnold.

Everitt, B.S. (2002) *The Cambridge Dictionary of Statistics* (2nd ed.). Cambridge: Cambridge University Press. [e-copy in EHU library]

Forsyth, R. & Lam, P. (2009) Keyness as correlation: Notes on extending the notion of keyness from categorical to ordinal association. In Mahlberg, M., González-Díaz, V. & Smith, C. (eds.) *Proceedings of the Corpus Linguistics Conference* (C L2009), University of Liverpool, UK, 20-23 July 2009.
Gablasova, D., Brezina, V. & McEnery, T. (2017) Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*. DOI: 10.1111/lang.12226
Gabrielatos, C. (2007). *If*-conditionals as modal colligations: A corpus-based investigation. In Davies, M., Rayson, P., Hunston, S. & Danielsson, P. (eds.) *Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007*. Birmingham: University of Birmingham.
Gabrielatos, C. (2010) A corpus-based examination of English *if*-conditionals through the lens of modality: Nature and types. PhD Thesis. Lancaster University.
Gabrielatos, C. (2014). Corpus approaches to discourse studies: The basics. Discourse & Communication Studies Research Team, Örebro University, 12 September 2014. https://www.academia.edu/8406977
Gabrielatos, C. & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics*, 36(1), 5-38
Gabrielatos, C. & Marchi, A. (2011). Keyness: Matching metrics to definitions. *Corpus Linguistics in the South* 1, University of Portsmouth, 5 November 2011.
Gabrielatos, C. & Marchi, A. (2012). Keywords: Appropriate metrics and practical issues. *CADS International Conference*, Bologna, Italy, 13-15 September 2012.
Gabrielatos, C. & McEnery, T. (2005) Epistemic modality in MA dissertations. In Fuertes Olivera, P.A. (ed.) *Lengua y Sociedad: Investigaciones recientes en lingüística aplicada. Lingüística y Filología no. 61.* Valladolid: Universidad de Valladolid. 311-331
Gabrielatos, C., McEnery, T., Diggle, P. & Baker, P. (2012) The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, 37(2), 151-175.
Gan, G., Ma, C. & Wu, J. (2007). *Data Clustering: Theory, algorithms and applications*. Philadelphia: ASA-SIAM.
Gerbig, A. (2010) Key words and key phrases in a corpus of travel writing. In Marina Bondi, Mike Scott (ed.s) *Keyness in Texts*. John Benjamins. 147-168.
Granger, S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer K., Altenberg B., Johansson M. (eds.) *Languages in Contrast: Text-based cross-linguistic studies*. Lund: Lund University Press. 37-51
Granger, S., Gilquin, G., Meunier, F. (2011) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses Universitaires de Louvain.
Gries, S.Th. (2006) Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2). 109-151.
Gries, S.Th. (2010a). Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269-291. Frankfurt am Main: Peter Lang.
Gries, S.Th. (2010b) Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods. In Tony Harris & María Moreno Jaén (eds.), *Corpus linguistics in language teaching*, 121-146. Frankfurt a.M.: Peter Lang.

Gries, S.Th. (2015) Quantitative designs and statistical techniques. In Douglas Biber & Randi Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics*, 50-71. Cambridge: Cambridge University Press.

Hardie, A. (2014) Log Ratio – an informal introduction. Post on the website of the ESRC Centre for Corpus Approaches to Social Science (CASS). http://cass.lancs.ac.uk/?p=1133

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund-Prytz, Y. (2008) *Corpus Linguistics with BNCweb - a practical guide*. Frankfurt: Peter Lang.

Hofland, K. & Johansson, S. (1982). *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Jaworska, S. & Nanda, A. (2016) Doing well by talking good? A topic modelling-assisted discourse study of corporate social responsibility. *Applied Linguistics*, https://doi.org/10.1093/applin/amw014

Kilgarriff, A. (1996a). Which words are particularly characteristic of a text? A survey of statistical approaches. In Evett, L.J. & Rose, T.G. (eds.) *Language Engineering for Document Analysis and Recognition (*LEDAR), AISB96 Workshop proceedings, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK. 33 – 40.

Kilgarriff, A. (1996b) Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. *ALLC-ACH Conference*, June 1996, Bergen, Norway. 169-172.

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong, pp. 231 – 245.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133.

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

Kilgarriff, A. (2009). Simple maths for keywords. In Mahlberg, M., GonzálezDíaz, V. & Smith, C. (eds.) *Proceedings of the Corpus Linguistics Conference*, *CL2009*. Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/CL2009/171_FullPaper.doc

Kilgarriff, A. & Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing*, Granada, Spain. 46-52.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014) The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.

Krogvig, I. & Johansson, S. (1985). *Shall, will, should*, and *would* in British and American English. *ICAME News*, 5, 32-56.

Leech, G. & Fallon, R. (1992). Computer corpora - What do they tell us about culture? *ICAME Journal*, 16, 29-50.

McEnery, A.M. (2006) *Swearing in English: Bad language, purity and power from 1586 to the present*. London: Routledge.

Miki, N. (2011) Key colligation analysis: Discovering stylistic differences in significant lexico-grammatical units. proceedings of the Corpus Linguistics 2011 Conference, ICC Birmingham, 20-22 July 2011. 1-23. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-208.pdf

Mujis, D. (2010) *Doing Quantitative Research in Education with SPSS* (2nd ed.) Sage.

Paquot, M. & Bestgen, Y. (2009) Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Jucker A., Schreier D. & Hundt M.

(eds.) *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. Amsterdam: Rodopi. 247-269.

Partington, A. (2014) Mind the gaps: The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, 19(1), 118–146.

Partington, A., Duguid A & Taylor, C. (2013) *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies*, Amsterdam & Philadelphia: John Benjamins.

Phillips, M. (1989) *Lexical Structure of Text*. Discourse Analysis Monograph no. 12. English Language Research, University of Birmingham.

Pojanapunya, P. & Watson Todd, R. (2016) Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Lingustic Theory*. DOI: 10.1515/cllt-2015-0030

Raftery, A.E. (1999) Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods & Research*, 27(3), 411-427.

Rayson, P. (2003) Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis, Lancaster University.

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

Rayson, P. (2009) Wmatrix: A web-based corpus processing environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Rayson P., Berridge D. & Francis B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In 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), Vol* 2. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain. 926-936.

Romesburg, H.C. (1984). *Cluster Analysis for Researchers*. Belmont, CA: Wadsworth.

Rosenfeld, B. & Penrod, S.D. (2011) *Research Methods in Forensic Psychology*. Wiley.

Rychlý, P. (2008) A lexicographer-friendly association score. In Petr Sojka & Aleš Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, pp. 6–9. Masaryk University, Brno 2008

Scott, M. (1996). *WordSmith Tools Manual*. Oxford: Oxford University Press.

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

Scott, M. (1998). *WordSmith Tools Manual, Version 3.0*. Oxford: Oxford University Press.

Scott, M. (2001) Mapping key words to *problem* and *solution*. In Scott, M. & Thompson, G. (eds.) *Patterns of Text: In honour of Michael Hoey*. John Benjamins. 109-128.

Scott, M. (2016) *WordSmith Tools version 7*. Stroud: Lexical Analysis Software.

Sirking, R.M. (2006). *Statistics for Social Sciences* (3rd ed.) London: Sage.

Sneath, P.H.A & Sokal, R.R. (1973). *Numerical Taxonomy.* San Francisco: Freeman.

Stubbs, M. (2010) Three concepts of keywords. In Bondi, M. & Scott, M. (eds.) *Keyness in Texts: Corpus Linguistic Investigations*. Amsterdam: Benjamins. 21-42. Amsterdam: John Benjiamins.

Taylor, C. (2013) . Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8(1), 81-113.

Thompson, B. (1998) Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, 5(2), 33-38.

Utka, A. (2004) Analysis of George Orwell's novel 1984 by statistical methods of corpus linguistics. *Sankirta: A Yearly Internet Journal of Lithuanian Corpus Linguistics*. http://donelaitis.vdu.lt/publikacijos/adrtmain.htm .

van Leeuwen, T. (1996) The representation of social actors. In Caldas-Coulthard, C.R. & Coulthard, M. (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge. 32-71.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger, & A. Koll-Stobbe (Eds.), *New approaches to the study of linguistic variability*. (pp. 3-11). (Language Competence and Language Awareness in Europe; Vol. 4). Frankfurt: Peter Lang.

Ziliak, S.T. & McCloskey, D.N. (2008) *The Cult of Statistical Significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.