

Key words when text forms the unit of study

Sizing up the effects of different measures

Stephen Jeaco

Xi'an Jiaotong-Liverpool University

Throughout the social sciences, there has been growing pressure to present effect sizes when publishing empirical data (see American Psychological Association, 2001; Parsons & Nelson, 2004). While it seems indisputable that for the majority of quantitative research foci, effect size is an essential element of statistical analysis, this paper argues that specifically for key word analysis in corpus linguistics, the means of reporting effect size must depend on the level of the unit of study of each investigation (single text, collection or large corpus). After exploring some main criticisms of the log-likelihood measure, this paper unpacks the parameters of different measures for keyness and how they might address underlying concerns. It maintains that for the exploration of foregrounded/deviant/salient/marked features in text, the use of log-likelihood scores to rank the results is still fit for purpose and coupled with Bayes Factors is a solid approach for key word analyses.

Keywords: keyness, effect size, key word analysis, log-likelihood, ranking

1. Introduction

Over the last 25 years, there has been a pressing need to re-examine quantitative research methods in the social sciences to address the fact that effect size may have often been overlooked in earlier research. This need is evidenced in the change to the APA manual rolled out in its 5th edition (American Psychological Association, 2001). A simple definition for effect size is given by Grissom & Kim as follows: “an effect size usually quantifies the degree of difference between or among groups or the strength of association between variables such as a *group-membership* variable and an *outcome* variable” (Grissom & Kim, 2012: 1; emphasis in original). They go on to distinguish between statistical significance and effect size as follows:

Whereas a test of statistical significance is traditionally used to provide evidence (attained p level) that a null hypothesis is wrong, an effect size (ES) measures the degree to which such a null hypothesis is wrong (if it is wrong).

(Grissom & Kim, 2012: 5)

The mathematical basis of these two measures may not need to be thoroughly understood by researchers in order for them to be applied appropriately and effectively, but some care is needed to ensure that rules-of-thumb for analytical approaches and expectations of peer reviewers in different disciplines are developed appropriately. For wider applications and for research in fields of linguistics such as second language acquisition, a number of issues related to effect size are of importance (Plonsky & Oswald, 2014). Within corpus linguistics, there will also be many aspects where effect size measures can be easily and readily added to provide richer and deeper evidence. However, this paper explores one specific kind of corpus technique, known as Key Word (KW) Analysis (Scott, 1997; Scott & Tribble, 2006). Two fundamental reasons for setting out this discussion are the well-known fact that word frequencies have what is known as a Zipf distribution (see Croft et al., 2010; Oakes, 1998), and the practical experience of hundreds or thousands of researchers over the years who have generated KW lists sorted by keyness in descending order and found these data-driven results to be intuitively fitting and highly productive as starting points for further exploration. Sorting in descending order puts the largest values at the top of the list. This order is also known as reverse sorting and is used for the remainder of this article unless stated otherwise.

A fundamental concept explaining the background for KW Analysis is Zipf's (1935: 40–41) rank frequency distribution, which builds on his observation that “a few words occur with very high frequency while many words occur but rarely”. By ranking words in descending frequency along the x-axis and then plotting their frequencies in a text on the y-axis, Zipf noted that an extremely sharp decline can be observed. Figure 1 shows a rank frequency distribution chart using raw frequencies for the famous Wilkie Collins novel *The Moonstone*.¹

Looking at Figure 1, it can be observed that the most frequently occurring words (which typically for any English text include *the* and *of*) are at the top of a very steep drop. At the other end of the curve, what is striking is the length of the near-flat line, or rather the number of words with extremely low frequencies. This has later been supported by findings in corpus linguistics and computer science that hapex legomena (words occurring just once in a corpus) typically account for 50% of the

1. First published in 1868; plain text downloaded from www.gutenberg.org (file header and license removed prior to processing). Tokens: 237,772.

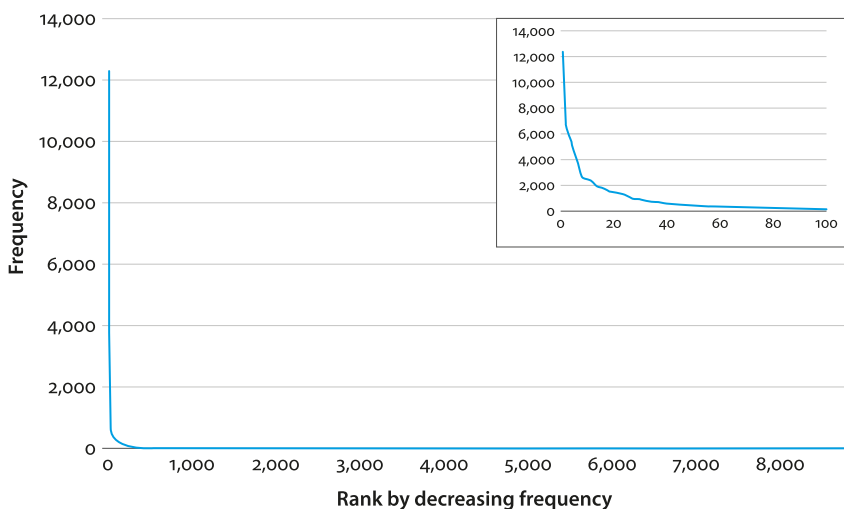


Figure 1. Rank frequency distribution chart for *The Moonstone*, with additional panel for the Top 100 items

types in a large text sample (Croft et al., 2010). Put simply, Zipf's approximation was that a word's rank multiplied by a word's frequency is a constant value (Zipf, 1935; Croft et al., 2010). An important point about Zipf's work was that it was a predictive model (Oakes, 1998) and could be used to estimate probability (Croft et al., 2010). Although the shape of such curves and the accuracy of Zipf's approximation will vary across different texts, the steepness at the beginning and the length of the curve can inform the present exploration of the purpose of KW Analysis. The shape of the curve must reflect to some extent a reader's expectations of typical word frequencies. It explains why a doubling of the frequency of a word from one part of the curve could be very different in terms of salience to the doubling of the frequency of a word from another part of the curve. The KW procedure is a means of approximating the importance of changes between frequencies found in the wordlist of a text (or collection of texts) of interest and what might be expected based on word frequencies in a reference corpus. The result for each item is often described in terms of its 'keyness'. A starting point for a definition of keyness might be the explanation of the purpose and means of calculation provided in the *WordSmith Tools* manual: "[k]ey words are those whose frequency is unusually high in comparison with some norm" (Scott, 2019b). The manual explains the computation of key words as follows:

To compute the "keyness" of an item, the program therefore computes

- its frequency in the small word-list
- the number of running words in the small word-list
- its frequency in the other corpus

- the number of running words in the comparison corpus and cross-tabulates these. (Scott, 2019a)

For cross-tabulation using a statistical test, these four parameters are put into a contingency table. Scott (1997) explained that keyness can be calculated using the chi-square statistic, and some earlier publications using keyness also used chi-square (e.g. Rayson et al., 1997), but in following years, the usual statistic became log-likelihood (LL) and the usual configuration and formula is shown in Table 1. With this well-known technique, “Corpus one” and “Corpus two” could be a study corpus and reference corpus, or two sub-corpora formed by splitting a larger corpus.

Table 1. Contingency table and formula for key words*

	Corpus one	Corpus two	Total
Freq. of word	a	b	a + b
Freq. of other words	c-a	d-b	c + d-a-b
Total	c	d	c + d

* $E1 = c * (a + b) / (c + d)$; $E2 = d * (a + b) / (c + d)$; $LL = 2 * ((a * \log(a / E1)) + (b * \log(b / E2)))$.
Table and formulae from Rayson & Garside (2000: 3); citing approach from Read & Cressie (1988).

As well as KWs that are based on frequencies across the whole corpus, Scott (1997) also proposed the notion of ‘Key Key Words’ (KKW). To calculate KKWs, the software generates batches of KW lists (one for each text), and then calculates the proportion of texts in which each candidate KKW is key. For an overview of KWs and KKWs, and practical applications of these methods, see Scott & Tribble (2006). More recently, Egbert & Biber (2019) have also proposed text dispersion keyness to measure keyness across texts in a corpus. For many researchers, the KW tool is one of the most important features of *WordSmith Tools* (Scott, 2016), and now text dispersion keyness is also available (Scott, 2019a). To get a sense of the importance of KWs in corpus linguistics, the reader is encouraged to consider the multitude of citations in the literature for *WordSmith Tools* itself. The measure of keyness based on LL has been the default software setting in other popular corpus tools including *AntConc* (Anthony, 2019), *CLiC* (Mahlberg et al., 2016) and *WMatrix* (Rayson, 2008). KKW functionality is included in software such as *WordSmith Tools* (Scott, 2016) and *The Prime Machine* (Jeaco, 2017). However, in other popular software, the term ‘Key Words’ or (‘keywords’) is used, but the statistical measure (or default statistical measure) for ranking is different. Software such as *Sketch Engine* (Kilgariff et al., 2004), *CQPWeb* (Hardie, 2012), *LangsBox* (Brezina et al., 2015) and *Lextutor* (Cobb, 2000) would fall into this category. The gap between these approaches perhaps also reflects a gap between

researchers' different purposes. Although KW features in *WordSmith Tools* and the other LL based software tools have been used for a wide variety of studies over the last fifteen years or so, there have been some recent reservations related to LL as a means of ranking results. Gabrielatos & Marchi (2012) present a thought-provoking examination of the definition and methods used in KW analysis. Their paper has sent ripples through the corpus linguistics community, and their views and findings have been echoed prominently elsewhere (Gabrielatos, 2018; Hardie, 2014a, 2014b). Gabrielatos & Marchi (2012) introduce an alternative measure of keyness called %DIFF which is a simple ratio calculation giving the percentage difference between the normalized frequency in the study corpus compared to the reference corpus. They make a number of claims leading towards the suggestion that %DIFF would be a more suitable method for ranking keyness. A summary of Gabrielatos & Marchi's (2012) claims is as follows:

- i. The usual definition of 'key word' is not consistent with the metric for 'keyness'.
- ii. The statistical significance of a frequency difference is not a good metric for keyness.
- iii. "The current threshold for statistical significance ($p \leq 0.01$, $LL \geq 6.63$) is arbitrary" (Gabrielatos & Marchi, 2012: 30).
- iv. A measure of effect size is needed to ensure that statistically significant results are not just a result of having a large sample.
- v. The best available measure of effect size would be %DIFF which is noted as "[t]he % difference of the frequency of a word in the study corpus when compared to that in the reference corpus" (Gabrielatos & Marchi, 2012: 10).

These claims bring together a number of important threads including: definitions of keyness and KW; the use of the LL measure; and the selection of reference corpora. The aim of this paper is to explore each claim and to consider whether there are alternative ways to make LL based keyness measures more robust.

2. Analysis

Sections 2.1 and 2.2 will address the claims of Gabrielatos & Marchi (2012) related to definitions and the difference between statistical significance and linguistic importance. The issue of appropriateness will be considered in Section 2.3. Section 2.4 presents some analysis of the different values that contribute to LL and effect size measures and some issues with non-occurrence in the reference corpus. Finally, the analysis will return to rank frequency distribution charts to demonstrate differences between the range of matches in the reference corpus for LL and %DIFF -ranked candidate KWs of a single text.

2.1 Defining keyness

Gabrielatos & Marchi (2012) criticize that the usual definition of keyness confuses statistical significance with effect size; or perhaps that the LL measurement and the definition of keyness given in many studies which use it blur the line between what keyness is and how it can be measured. A similar blurring in the definition of another linguistic feature in corpus linguistics – that of collocation – has been reflected on and discussed by Hoey (2005). Just as he reflects on the need to consider the difference between statistical and psychological definitions for collocations, the definitions often given for KWs are statistical definitions. The psychological importance of keyness and the clues as to why this phenomenon should exist are summarized by Scott as follows:

keyness is a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail. What the text “boils down to” is its keyness, once we have steamed off the verbiage, the adornment, the blah blah blah.

(Scott & Tribble, 2006: 55–56)

2.2 Two measures of keyness: LL and %DIFF

As Wilson (2013:1) points out, despite “some aura of novelty”, when keyness is measured using LL, it is “nothing more than an ordinary null hypothesis significance test applied to the frequencies of words or other items in two texts or corpora”. The LL measure has been applied to collocation and is one of the common statistics available in modern concordancers. Dunning (1993) proposed the use of LL, as a way of balancing the bias towards low frequency items which exists in many of the other collocation measures. When he proposed applying LL as a collocation measure, the values used in the formula were the combined frequency, the separate frequencies and the total corpus size. However, once LL began to be applied to keyness calculations, questions began to be raised about appropriate cut-off points for minimum statistical significance and also about the risks of over reporting relationships when it was applied to very large corpora. Both of these issues are addressed by Rayson et al. (2004) who demonstrated why in corpus applications the LL scores should be considered significant when greater than 15.13. Keyness in *WordSmith Tools*, *AntConc* (Anthony, 2019), *CLiC* (Mahlberg et al., 2016) and *The Prime Machine* (Jeaco, 2017) is based on the application of this kind of statistic, with the highest LL or chi-squared results being interpreted as being relatively more important, and results sorted in descending order using this measure by default. Nevertheless, the need for caution when interpreting keyness scores and relative rankings between items has been emphasized (Scott, 2019c).

Gabrielatos & Marchi (2012), however, argued that as in other fields of research, effect size has been overlooked, and they propose the use of the percentage difference in frequency (or %DIFF). They suggest that %DIFF should be used, with a miniscule number added to any item not appearing in the reference corpus. Since at the time of their presentation the procedure of selecting all statistically significant items and then ranking them using their effect size measure was not available in concordancers, they proposed a manual method for calculating this in a spreadsheet. They presented %DIFF as given in Equation 1.

Equation 1. Formula for %DIFF (Gabrielatos & Marchi, 2012: 12)

$$((\text{NF in SC} - \text{NF in RC}) / \text{NF in RC}) * 100$$

NF = Normalised frequency

SC = study corpus

RC = reference corpus

%DIFF is closely related to the KW metrics used in *Lextutor* (Cobb, 2000) or *Sketch Engine* (Kilgariff et al., 2004; Lexical Computing Ltd., 2014). Both %DIFF and LL approaches use the *relative* frequencies of words. However, since only normalized frequency is used in %DIFF, it could be argued that it reduces the data, leaving out other important information (see Section 2.4). For most analyses where individual texts form the unit of study, effect size should arguably also consider the relative frequency with respect to the whole sample. The “(a + b)/(c + d)” part of the equation is missing in %DIFF; that is to say the %DIFF equation does not measure the frequency of the phenomenon across the *combined* corpus. An important reminder given in the fourth claim presented earlier is that large sample sizes lead to high LL scores. The main reason given for this is that very large corpora can mean that the normal *p* value cut-off points are too low. Gabrielatos & Marchi (2012) propose setting a threshold “relative to the resulting range of %DIFF values”. However, having relative thresholds determined subjectively means that different studies will set their own cut-off thresholds and cross-study comparisons will be extremely difficult. Another cautionary note regarding the interpretation of LL scores for keyness across studies is that cut-off points are usually set by each individual user (Baker, 2004). The issue of balancing for effect size is also raised by Wilson (2013), where he proposes using an approximation of Bayes Factors. He recommends the use of Bayes Factors in keyness and other calculations in corpus linguistics to distinguish between very strong evidence and less strong evidence based on the overall size of the corpus. Wilson (2013: 4) contrasts Bayesian statistics with “frequentist” statistics and explains that an approximate Bayes Factor can be provided using a formula for Bayesian Information Criterion (BIC). Table 2 shows how these are calculated and to be interpreted.

Table 2. Approximate Bayes Factors and equation for BIC approximation *

Approximate Bayes Factor (BIC)	Degree of evidence against Ho ²
0–2	not worth more than a bare mention
2–6	positive evidence against Ho
6–10	strong evidence against Ho
>10	very strong evidence against Ho

* $BIC \approx LL - \ln(N)$. Formula from Raftery (1986) and Kass & Raftery (1995) given in Wilson (2013).

Using the BIC approximation it is possible to calculate the minimum size of a corpus required in order to satisfy the 15.13 level proposed by Rayson et al. (2004) above, while still reaching a BIC of 2. The natural log of 500,000 is 13.12, so subtracting this from a LL of 15.13 or more will generate a BIC value of at least 2. Thus, following Wilson’s (2013) proposal means that a BIC cut-off point of 2 on a corpus of half a million tokens or more will provide a level of stringency equivalent or greater than the 15.13 level and provides scalability to make comparisons between corpora of larger sizes possible. So, as the total combined corpus size increases, the LL score has to be higher in order to meet the same BIC. It should be noted that Gabrielatos (2018) recommends filtering out candidate keyword items based on BIC values below 2 and *WordSmith Tools* now includes BIC, too.

2.3 Determining appropriate measures for keyness

In order to demonstrate some of the dangers of blindly applying LL to data without considering the actual difference in frequency, Gabrielatos & Marchi (2012) compared the proportion of overlap between the rankings of LL and %DIFF for all KWs and the Top 100 KWs for several corpora. They claimed that low overlap would mean “one metric is inappropriate”. However, it would be fairer to say that low overlap would mean the metrics measure different phenomena. Different kinds of focus might lead researchers to prefer a %DIFF calculation over LL in certain circumstances. Indeed, several important considerations for the application of KW and KKW techniques are given by Scott & Tribble (2006) as they explore the results obtained when the scope of the wordlist and the reference corpus are adjusted. For example, when reporting results for one Shakespeare play using his other plays as a

2. Ho stands for the null hypothesis. In the context of collocation, the hypothesis would be that the items occur together more frequently than would be expected by chance. The null hypothesis is the opposite of this: that the items do not occur more frequently together than would be expected by chance. In other words, Ho is the hypothesis that the words do not form a collocation.

reference corpus, they explain how some KWs in the results “do not reflect importance and aboutness”, and might be considered to be indicators of style (Scott & Tribble, 2006: 60). They suggest that concordancing some of these “intruders” can provide useful insights. However, when moving from individual texts to collections or entire corpus databases, Scott & Tribble (2006) demonstrate that both KWs and KKWs can have a tendency to become more similar to a raw frequency wordlist, especially if the study corpus is more general in nature. Through the examples provided by Scott & Tribble (2006) in experimenting with different reference corpora and looking at KKWs on genre or whole database level, it is clear that the “aboutness” is not always easily obtained and careful consideration needs to be given to both these factors.

Although these potential pitfalls are introduced clearly in these examples and others in the help pages of *WordSmith Tools*, one of the most striking revelations of the comparison provided by Gabrielatos & Marchi (2012) for %DIFF versus LL was the extremely high LL values and very high rankings for the two most common words in English (*the* and *of*) despite only small differences in %DIFF. When conducting computer lab sessions teaching sophomore English majors how to perform KW analysis on small corpora, my own experience has been that if *the* appears in the results, either one of the case-sensitivity options in *AntConc* had been missed or the KW reference file had been incorrectly set up and so the results were due to a procedural error. For LL scores to be so high while the underlying differences in frequency are small, the results for these two words presented by Gabrielatos & Marchi seem to be of greater concern than some of the differences between LL and %DIFF for proper names, etc. Gabrielatos (personal communication, May 31, 2013) kindly provided the KW lists they had used to generate these results. He informed me that the data were taken from a comparison of the *SiBol93* and *SiBolo5* corpora, which are both substantial corpora based on newspapers with half a million texts or more in each. Partington (2010) provides information about the sources for these corpora, explaining that *SiBol93* contains every article published in 1993 in *The Times* and *Sunday Times*, the *Telegraph* and *Sunday Telegraph*, and the *Guardian*, while *SiBolo5* contains every article published in 2005 in these same UK broadsheets with the addition of *The Observer*. The *WordSmith Tools* KW files indicated that these corpora had been loaded from approximately 60 individual files, so the texts were not organised in such a way that individual news articles (or even individual newspaper issues) were contained in separate files. A conservative estimate of the number of actual news stories contained in each of these corpora, if a rough average from the *Guardian* corpus of 400 words per article is applied, would suggest *SiBol93* would have more than 240,000 texts and *SiBolo5* would have 390,000 texts. In their presentation, Gabrielatos & Marchi (2012: 24) showed the following results and conclusion:

- The LL = 32,366.01 (2nd) **but** %DIFF = 9.7% (4302nd)
- Of LL = 20,935.05 (5th) **but** %DIFF = 11.8%

What the high LL values indicate here is that we can be highly confident that there is a very small frequency difference

While they do highlight an important point, it could be argued that the suggestion that these are just small frequency differences is a little misleading. Table 3 shows the raw frequencies and LL keyness score for the Top 15 KWs in their data with the 1993 corpus as the study corpus and the 2005 corpus as reference. Table 4 shows the Top 15 KWs sorted by descending %DIFF value. As the KW list was loaded directly into *WordSmith Tools*, small differences in the keyness values from those given in the original presentation are likely to be a result of slightly different configurations. However, the rankings of the items seem to be identical to those Gabrielatos & Marchi (2012) presented. It should also be noted that since these results were drawn from KW lists which had been cut-off at the Top 500 KWs in descending order of LL score, some high scoring %DIFF candidates could be missing from this list. As recommended by Gabrielatos (2018), researchers working with full corpus data would be able to include all candidate KWs that meet the minimum threshold.

Table 3. LL ranked key words from the data of Gabrielatos & Marchi (2012)

Rank	Key word	Study f.	Ref. f.	LL
1	MR	206523	176385	30473.98
2	THE	6001857	8247131	29461.11
3	EC	15204	623	23281.55
4	CLINTON	19793	3264	20843.41
5	OF	2782374	3743488	19958.21
6	BOSNIA	13488	910	18890.35
7	1991	18233	4008	16561.94
8	RECESSION	12484	1101	16386.54
9	YELTSIN	9829	217	16162.65
10	CORRESPONDENT	14743	2521	15268.38
11	MAJOR	41747	24322	14473.85
12	MAASTRICHT	8669	300	13583.99
13	WHICH	316733	359203	13253.90
14	MILLION	84491	70938	13115.26
15	BOSNIAN	9159	623	12804.46

Table 4. %DIFF ranked key words from the data of Gabrielatos & Marchi (2012)

Rank	Key word	Study f.	Ref. f.	%DIFF
1	VANCE-OWEN	1254	0	1.30674E+15
2	KHASBULATOV	1008	0	1.05039E+15
3	PSBR	928	0	9.67028E+14
4	BT ₃	658	0	6.85673E+14
5	RUTSKOI	615	0	6.40865E+14
6	BRAER	592	0	6.16897E+14
7	1ST-HALF	576	0	6.00224E+14
8	FT-SE	1683	2	126359.787
9	AIDID	826	1	124030.463
10	VITEZ	819	1	122978.510
11	AIDEED	697	1	104644.470
12	POPIOLEK	537	1	80599.829
13	FIMBRA	512	1	76842.853
14	INV	506	1	75941.179
15	IPOUNDS	488	1	73236.157

Looking at the Top 15 ranked in descending order by LL, the presence of *the* and *of* is immediately striking. Adjusting for the differences in size, *the* occurs 1.5 million times more or one extra time in every one hundred words, and *of* occurs 0.78 million times more or one extra time in every two hundred words. Using the conservative estimates of the number of texts contained in these corpora, this would equate to an average of four and two more occurrences in each text, respectively. The difference is perhaps small given the high frequency of these two items in any English text, but the main issue is that they are not usually considered to be interesting or noticeable (though see Leech et al., 2009 on diachronic grammatical changes and the decline in *of*-phrases), and neither can really point to the aboutness of these corpora. However, the LL data shows indications of what was important in the news in 1993, with the *Maastricht* treaty, president names and the British Prime Minister's names, the *EC* as well as the troubles in Bosnia being evident. The other "intruders" such as *Mr*, *correspondent* and *which* might well be reasonable starting points for further investigation of changes in style. The suggestion that these differences are small is not really accurate; it would be more accurate to suggest that KW analysis based on a wordlist of each complete corpus is unlikely to provide very detailed information about changes in what is newsworthy over two periods, or as a resource for the exploration of what was happening in the world at those times.

Looking at the Top 15 results when ranked by descending %DIFF score, it is clear that the first 7 results have extremely high values. The dominance of these is caused by the handling of non-occurrence of these items in the reference corpus (see Section 2.4). Nevertheless, the remaining 8 items all have very high %DIFF scores because they only occurred once or twice in the reference corpus.

Using wordlists from the combined texts may bring out some names of major political leaders, major themes and some stylistic changes when results are ranked using LL, and they may bring out long lists of names and other entities when ranked using %DIFF. Yet neither of these lists really answers the question of what would have been noticeably different to a reader of newspapers during each year. In order to measure how changes took place in terms of what was newsworthy, it would arguably be better to work with each text in a separate file since the KWs for each individual text can correspond to what might be psychologically more salient. Rather than looking at differences in relative frequencies across large numbers of texts together, KKWs are based on a measurement of relative frequencies in each separate text (a unit of study better matching the way a reader typically reads a newspaper) compared with a larger collection (the reference corpus as an approximation of typical frequencies a reader might encounter). Therefore, measuring changes in the proportion of texts in which an item is key (using KKWs) ought to be a better means of investigation. In their presentation, Gabrielatos & Marchi (2012) used *adventists* and *ex-communist* as examples where there was a large %DIFF but relatively low LL; however, the raw frequencies for these were 94:6 and 134:26 respectively. Given the corpora had hundreds of thousands of texts it is just as unlikely that a reader would notice these large proportional changes, as it would be that they might notice the decreased use of *Mr*, *the*, or *of*. The important point that these examples illustrate is not so much that one measure is superior to another, but rather that organization of texts into files and the choice of reference corpus are very important when determining the best strategy for a specific research question.

2.4 Parameters used in different measures

Having argued that different measures will be appropriate for different kinds of study, the parameters of several effect size measures will be explored and compared, drawing particular attention to whether raw values are included and each measure's sensitivity to important contrasts between them. For the comparison of different effect size measures, this paper draws on the measures available in the very useful and popular online tool, the UCREL Log-likelihood and Effect Size Calculator (Rayson, n.d.). Table 5 shows each of the measures and indicates whether or not it is sensitive to the following aspects, which correspond in different ways to effect size in terms of different perspectives and different outlooks. For

clarity, all formulae have been reconfigured to show the four parameters which are entered into the UCREL Log-likelihood and Effect Size Calculator (or into the spreadsheet version which is also available from the same website); these are:

- a. the frequency of a word in the study corpus.
- b. the frequency of the same word in the reference corpus.
- c. the total running words of the study corpus.
- d. the total running words of the reference corpus.

Table 5. Parameters and underlying perspectives or outlooks for different measures

Parameters/perspectives		LL	LL+ Bayes Factors	%DIFF	Relative risk	Log ratio	Odds ratio
Raw frequencies	i. a vs. b	Y	Y				Y
	ii. a vs. c-a						Y
	iii. b vs. d-b						Y
Normalized frequencies	iv. a vs. c	Y	Y	Y	Y	Y	
	v. b vs d	Y	Y	Y	Y	Y	
Relative sizes of study and reference corpora	vi. a + b vs. c + d	Y	Y				
Comparability Parameters*	vii. c + d vs. X		Y				
Arbitrary near zero adjustment		N	N		N		N

* For aspect (vii), X denotes other studies with combined corpora of different sizes.

It is suggested that for parameters (i)–(vii), measures that do not reduce the data, and have more sensitivity to changes in magnitude of the underlying parameters should be favoured, with the letter Y used to mark such measures. The bottom row indicates whether an arbitrary value has to be added when an item does not occur in one of the corpora, as the measure is unable to cope with frequencies of zero. In this last row, those measures which do not require an adjustment should be favoured – these are marked with “N”. The only measure available at the time of writing which has not been analysed here is that of ELL (Effect Size for Log-likelihood Ratio), as this uses the LL result in its calculation, meaning the parameters are shared, and it will be discussed briefly later.

It can be seen in Table 5 that LL+Bayes Factors addresses the important contrasts raised in this paper. The critical values for confidence levels in the LL measure of keyness are simply a way of mapping the results of this parametric test which is not dependent on an assumption that the model has a normal distribution. The actual keyness value is not merely a measure of statistical significance; it represents the kind of measure that Bradley (1960) termed the “exact cumulative probability”.

As statisticians often remind us, the *p*-values typically obtained are not the probability of the specific outcome, but the probability of these or more extreme data. The same cautionary note regarding how to combine probabilities using chi-square holds as a note on how to rank results, namely: they “must be cumulative probabilities, not simply ‘significance levels’ within which the cumulative probability has fallen” (Bradley, 1960, p.373). It is the LL score for keyness which already considers all the factors relevant for studies where text is the unit of study. In addition, the LL score, viewed as a cumulative probability, helps us interpret just how marked the frequency in our study corpus is in terms of the frequency of our study corpus, the frequency of the word overall, and the size of the combined corpora overall.

When other measures (such as those in Table 5) are being put forward as being superior to LL, it is usually with the suggestion of retaining a statistical test like LL to determine which items should be included in a list and then that the results should be re-sorted using another measure. However, for research where the unit of study is the text, measures such as %DIFF which rely on adding a very small number can have serious problems in terms of ranking. Where the frequency in the reference corpus is zero, the size of the reference corpus becomes irrelevant, meaning that essentially for these items the %DIFF returns a magnified raw frequency list. Taking 15.13 as the cut-off for a LL score first and using the “Goal Seek” function in *Microsoft Excel*, it is simple to establish the minimum (integer) frequency for an item in a hypothetical study and reference corpora of varying sizes. Table 6 shows the results of these operations, along with the LL and Bayes Factor scores. In cases where the item does not occur in the reference corpus, the reduced formula for %DIFF is provided in Equation 2. Although empirical results would be preferable, Table 6 reveals important insights into how low minimum frequency thresholds for such items would be, while demonstrating the effectiveness of Bayes Factor.

Equation 2. Reduced formula for %DIFF for items not occurring in the reference corpus

$$\%DIFF = \frac{100 \times NF}{A}$$

NF = Normalised frequency in study corpus
A = Arbitrary near zero adjustment

Gabrielatos (2018: 237) claims that the prominence of items that are absent in the reference corpus when results are ranked using %DIFF is a “strength”, but for studies looking at companies or novels, the ranked list is likely to be dominated by company, product, character or place names, ranked simply by descending frequency. This can be illustrated by comparing the LL ranked KWs against those for

Table 6. Minimum frequencies in study corpora for items not occurring in reference corpora for hypothetical corpora of varying sizes, meeting the LL cut-off of 15.13

Study corpus	Reference corpus	Min. freq.	Normalised freq.	LL	Bayes Factor	BIC interpretation
10000	10000000	2	0.0002	27.64	11.52	Very strong evidence
10000	1000000	2	0.0002	18.46	4.64	Positive evidence
100000	10000000	2	0.00002	18.46	2.33	Positive evidence
10000	100000	4	0.0004	19.18	7.57	Strong evidence
100000	1000000	4	0.00004	19.18	5.27	Positive evidence
1000000	10000000	4	0.000004	19.18	2.97	Positive evidence
100000	100000	11	0.00011	15.25	3.04	Positive evidence
1000000	1000000	11	0.000011	15.25	0.74	Not worth more than a bare mention
10000000	10000000	11	0.0000011	15.25	< 0	–

%DIFF from two small corpora. Table 7 and Table 8 are for the 2019 Tesco PLC Annual Report using “BNC: Other Publications” as a reference corpus.³ Table 9 and Table 10 are from *The Moonstone* using the Fiction Collection 37 × 1 as a reference corpus.⁴ Due to space limitations, only a small number of top ranked results have been presented in these and subsequent tables and as mentioned earlier caution is needed when drawing conclusions based on the importance of top rankings. Extended tables for the Top 100 ranks can be found at <https://doi.org/10.1075/ijcl.18053.jea.additional>.


The LL tables were generated using *The Prime Machine*, where an indication of the difference in relative frequency is made through the use of arrows and a multiplier. For items not found in the reference corpus, the indicator is a special symbol (☀ ↑). As can be seen, such items do not crowd the LL ranked tables as much as they do the %DIFF tables. This is even clearer in the extended tables.

The other issue regarding the adjustment for items not found in the reference corpus is the smaller the tiny value, the larger these magnified raw frequencies appear. Equation 2 demonstrates that the need to choose an arbitrary adjustment

3. Tesco annual report from www.tescopl.com/media/476423/tesco_ar_2019.pdf; tables and charts removed and plain text extracted using *Adobe Acrobat Pro*; correction of word breaks made manually using *Microsoft Word* spelling prompts. Tokens: 96,012. The BNC: Other Publications sub-corpus is based on categories from Lee (2001) containing 20,504,136 tokens (includes punctuation, other symbols and numbers). Numbers and symbols were removed from the key word lists.

4. The Fiction Collection 37 × 1 corpus had been created using Project Gutenberg texts, using one novel from each of the 37 authors as listed in Mahlberg (2013) containing 7,377,506 tokens (includes punctuation, other symbols and numbers).
















Table 7. Top 15 key words ranked using LL for the 2019 Tesco PLC Annual Report compared against the “BNC: Other publications” sub-corpus

Rank	Word	Study f.	Ref. f.	Arrows	LL	Bayes
1	group	953	7171	≥ 10X ↑	4425.47	Very strong evidence
2	tesco	398	155	≥100X ↑	3618.73	Very strong evidence
3	financial	513	3247	≥ 10X ↑	2542.22	Very strong evidence
4	audit	259	243	≥100X ↑	2087.79	Very strong evidence
5	our	864	22729	≥ 5X ↑	2078.58	Very strong evidence
6	committee	382	3255	≥ 10X ↑	1687.94	Very strong evidence
7	ifrs	140	0	 ↑	1503.20	Very strong evidence
8	directors	220	779	≥ 10X ↑	1316.14	Very strong evidence
9	executive	243	1335	≥ 10X ↑	1265.87	Very strong evidence
10	booker	136	63	≥100X ↑	1212.38	Very strong evidence
11	value	301	3469	≥ 10X ↑	1165.31	Very strong evidence
12	remuneration	129	56	≥100X ↑	1158.76	Very strong evidence
13	assets	184	625	≥ 10X ↑	1113.96	Very strong evidence
14	colleagues	195	855	≥ 10X ↑	1093.84	Very strong evidence
15	impairment	110	26	≥100X ↑	1048.62	Very strong evidence

will lead to %DIFF values varying by several degrees of magnitude, depending on whether 1×10^{-18} is used (the default on the webpage at the time of writing) or another larger near zero value such as 1×10^{-12} , 1×10^{-8} or 0.001. Ranking results of a pre-selected set of items to map, for example, change over time or geographic area in the use of specific words will not result in such problems, but for the typical situation for research where the unit of study is the text, this is a serious disadvantage.

With regard to the ELL measure, which is available in the UCREL Log-likelihood and Effect Size Calculator, there still remain some problems if the observed frequency in the reference corpus is zero. The reference on the website for this measure is Johnston et al. (2006), and it is explained that the LL result is normalised by dividing by a number based on the minimum value of the expected frequency. KW candidates will almost always have an expected frequency for the study corpus that is lower than the reference corpus (especially, when research is based on using a small corpus of a specific style or register and comparing it to a larger corpus which is supposed to represent language use more broadly). In these cases, the ELL will normalise the result using the expected frequency for the study corpus (denoted by E1 in Table 1) and this will lead to high results for any items where the frequency in the reference corpus is zero. For this specific application of the KW technique, more will need to be considered regarding the appropriate-

Table 8. Top 15 key words ranked using %DIFF for the 2019 Tesco PLC Annual Report compared against the BNC: Other publications sub-corpus

Rank	Word	Study f.	Ref. f.	Arrows	%DIFF
1	ifrs	140	0	 ↑	1.46E+17
2	psp	57	0	 ↑	5.94E+16
3	www	30	0	 ↑	3.12E+16
4	APMs	18	0	 ↑	1.87E+16
5	tescoplc	18	0	 ↑	1.87E+16
6	brexit	17	0	 ↑	1.77E+16
7	website	16	0	 ↑	1.67E+16
8	payables	16	0	 ↑	1.67E+16
9	clubcard	15	0	 ↑	1.56E+16
10	shareview	11	0	 ↑	1.15E+16
11	remeasurements	11	0	 ↑	1.15E+16
12	equiniti	11	0	 ↑	1.15E+16
13	golsby	10	0	 ↑	1.04E+16
14	homeplus	10	0	 ↑	1.04E+16
15	ecl	10	0	 ↑	1.04E+16

ness of ELL, but it does seem that it will also push rare events back towards the top of the list.

2.5 Rank frequency distributions of Candidate KWs

In Sections 2.1 to 2.4, the claims of Gabrielatos & Marchi (2012) have been used as a way of unpacking some very important issues for researchers to consider when selecting appropriate metrics for keyness. It has been demonstrated that careful decisions regarding the choice of metrics are important, especially in cases of smaller corpora and where the unit of study is text. In this final section of analysis, we will return to Zipf's rank frequency distribution and see how this applies to candidate KWs. As was argued in the introduction to this paper, the power of KW analysis lies in its ability to use frequencies of words in a reference corpus to predict which words in a study corpus are likely to be most prominent for a human encountering these texts. In this sense, contrary to a point argued by Gabrielatos (2018), a similarity between the ranking of KW candidates and the size of their frequency difference may not be desirable for many kinds of KW studies. Essentially, the question about the importance of a candidate KW may be less "How many times bigger is its frequency?" and more "How noticeable is its change in frequency likely to be for the
















Table 9. Top 15 key words ranked using LL for *The Moonstone* compared against the fiction collection 37×1 corpus

Rank	Word	Study f.	Ref. f.	Arrows	LL	Bayes
1	franklin	537	24	≥100x ↑	3526.41	Very strong evidence
2	rachel	483	8	≥100x ↑	3267.50	Very strong evidence
3	sergeant	506	68	≥100x ↑	3094.80	Very strong evidence
4	betteredge	350	0	☀ ↑	2426.62	Very strong evidence
5	verinder	297	0	☀ ↑	2059.16	Very strong evidence
6	mr.	161	12652	≥ 3x ↑	1933.30	Very strong evidence
7	diamond	338	92	≥100x ↑	1902.79	Very strong evidence
8	blake	272	10	≥100x ↑	1800.04	Very strong evidence
9	rosanna	249	0	☀ ↑	1726.37	Very strong evidence
10	godfrey	248	3	≥100x ↑	1687.10	Very strong evidence
11	bruff	230	0	☀ ↑	1594.64	Very strong evidence
12	cuff	258	48	≥100x ↑	1525.94	Very strong evidence
13	my	265	37259	≥ 2x ↑	1244.43	Very strong evidence
14	moonstone	171	1	≥100x ↑	1173.35	Very strong evidence
15	ablewhite	157	0	☀ ↑	1088.51	Very strong evidence

text overall?”. This point can be illustrated by plotting top candidate KWs ranked by different metrics (after applying the recommended BIC 2 cut-off) on a graph of the rank frequency distribution using their frequencies in the reference corpus. For this comparison, a single text from the “BNC: Academic” sub-corpus was selected. The text is in file *JoT* in the Social Sciences section and is a text taken from “Global geomorphology: an introduction to the study of landforms” by Michael Summerfield, published in 1991. The reference corpus used contained all the other texts in the “BNC: Academic Sub-corpus” (using the categories from Lee, 2001).⁵ Candidate KWs for this text were extracted using *The Prime Machine*. A total of 657 candidate KWs met the BIC 2 threshold. This figure is well below the maximum number of KWs permitted in the software, so all KWs meeting the BIC 2 threshold were returned. % DIFF scores were calculated in *Microsoft Excel* and tables were then produced by ranking the results according to the two metrics. The Top 15 items can be seen in Tables 11 and 12. Extended tables can also be found at <https://doi.org/10.1075/ijcl.18053.jea.additional>. The reference corpus frequencies were then used to plot the rank frequency distributions. Figure 2 shows all 657 candidates when ranked by LL and Figure 3 shows the same candidates, but with an enforced maximum fre-

5. The selected text contains 50,366 tokens, leaving 18,036,863 tokens in the rest of the sub-corpus. Token counts include punctuation, other symbols and numbers.

Table 10. Top 15 key words ranked using %DIFF for *The Moonstone* compared against the fiction collection 37×1 corpus

Rank	Word	Study f.	Ref. f.	Arrows	%DIFF
1	betteredge	350	0	 ↑	1.47E+17
2	verinder	297	0	 ↑	1.25E+17
3	rosanna	249	0	 ↑	1.05E+17
4	bruff	230	0	 ↑	9.67E+16
5	ablewhite	157	0	 ↑	6.6E+16
6	luker	128	0	 ↑	5.38E+16
7	spearman	115	0	 ↑	4.84E+16
8	frizinghall	91	0	 ↑	3.83E+16
9	seegrave	51	0	 ↑	2.14E+16
10	murthwaite	48	0	 ↑	2.02E+16
11	yolland	38	0	 ↑	1.6E+16
12	merridew	34	0	 ↑	1.43E+16
13	herncastle	28	0	 ↑	1.18E+16
14	cannot	21	0	 ↑	8.83E+15
15	cobb	19	0	 ↑	7.99E+15

quency of 400. Figures 4 and 5 show the corresponding rank frequency distributions for candidates sorted by %DIFF.

While Figures 2 and 3 show hits across a wide range of reference corpus frequencies for the entire set of candidates (essentially no strong trends), there is a clear upward trend for Figures 4 and 5. If the aim of KW Analysis is to identify words that are likely to stand out in texts, the intensity of hits in the lower-left portion of the graph in Figures 4 and 5 is not promising. Figure 6 shows the same data as Figure 4, but with additional plots for a ranking simply based on the ascending frequency in the reference corpus.

The similarity in the curves is very striking and appears to be a mirror image of a Zipf distribution. Indeed, using Spearman's Rank Correlation, the descending ranking by %DIFF and the ascending ranking by the reference frequencies have a statistically significant strong positive correlation ($\rho=+0.93$, $p<0.0005$, one tailed), while results for LL compared by the reference frequency are very weak indeed ($\rho=+0.07$, $p=0.064$, one tailed). Comparing the study corpus frequencies with the %DIFF scores at the top of Table 12 shows the pattern described in Section 2.4: items where the reference frequency is zero have arbitrarily enormous %DIFF values and are simply ordered by the raw study corpus frequency. Further

Table 11. Top 15 key words ranked using LL for *JoT* compared against the remainder of the “BNC: Academic” sub-corpus

Rank	Word	Study f.	Ref. f.	Arrows	LL	Bayes Factor
1	continental	227	291	≥100x ↑	1962.62	Very strong evidence
2	lithosphere	166	6	≥100x ↑	1901.35	Very strong evidence
3	crust	160	43	≥100x ↑	1673.36	Very strong evidence
4	weathering	173	122	≥100x ↑	1636.33	Very strong evidence
5	plate	202	434	≥100x ↑	1584.34	Very strong evidence
6	oceanic	144	44	≥100x ↑	1490.15	Very strong evidence
7	volcanic	141	87	≥100x ↑	1356.51	Very strong evidence
8	uplift	123	72	≥100x ↑	1190.95	Very strong evidence
9	subduction	102	8	≥100x ↑	1142.97	Very strong evidence
10	margin	116	250	≥100x ↑	909.24	Very strong evidence
11	mantle	88	77	≥100x ↑	807.95	Very strong evidence
12	arc	94	125	≥100x ↑	807.63	Very strong evidence
13	rock	122	562	≥ 10x ↑	797.29	Very strong evidence
14	fig.	168	2761	≥ 10x ↑	705.69	Very strong evidence
15	margins	82	140	≥100x ↑	673.27	Very strong evidence
















down beyond these, as confirmed by Figure 6, the influence of the raw reference corpus frequency is clearly evident. By throwing out the LL score in the ranking, the top results have become heavily biased towards rare events.

3. Implications

The analysis presented in this paper leads to the implication that reflection on the appropriateness of one measure or another needs to be keenly attuned to the purposes and aims of the research. Table 13 shows the author’s own reflections on how different kinds of research may differ in terms of aims and purposes of keyness analysis. The table lists five examples of research aims, together with the unit of study, an approach, potential corroboration and follow-up work. The intention is not to provide a set of rules, but to outline some examples of possible types of research and how KWs could be used.

Table 13 suggests that for a type E study it may be most meaningful to compare %DIFF scores for sets of pre-defined items. This is because when looking at a specific communicative function or a group of related language choices, it would

Table 12. Top 15 key words ranked using %DIFF for *JoT* compared against the remainder of the “BNC: Academic” sub-corpus

Rank	Word	Study f.	Reference f.	Arrows	%DIFF
1	lithospheric	42	0	 ↑	8.34E+16
2	orogen	26	0	 ↑	5.16E+16
3	orogens	26	0	 ↑	5.16E+16
4	tephra	22	0	 ↑	4.37E+16
5	subducted	17	0	 ↑	3.38E+16
6	underthrusting	14	0	 ↑	2.78E+16
7	terranes	13	0	 ↑	2.58E+16
8	sunda	12	0	 ↑	2.38E+16
9	isostasy	11	0	 ↑	2.18E+16
10	domal	10	0	 ↑	1.99E+16
11	pyroclasts	9	0	 ↑	1.79E+16
12	upwarps	9	0	 ↑	1.79E+16
13	rifted	8	0	 ↑	1.59E+16
14	strato	7	0	 ↑	1.39E+16
15	calderas	7	0	 ↑	1.39E+16

seem purposeful to describe in terms of magnitude how tendencies of use for one item in the group compare to another. In many of the other types of study described in the table, it is suggested that using different configurations of texts (single texts, groups of texts or whole corpus) can be very effective.

4. Conclusion

This paper has argued that LL based KW calculations can be used effectively for a range of different kinds of research, but often work best with texts and moderately large collections of texts rather than with very large corpora at the entire corpus level. It has also been acknowledged that when corpus studies are of a more sociolinguistic nature, and particularly when looking at the spread of pre-determined features across different discourse communities, using measures based on relative frequencies such as %DIFF could very well be a fruitful approach. Corpus methods can be applied to discourse analyses (Baker et al., 2008; Gabrielatos et al., 2010) where the unit of study might be specific words. For culturally charged language items, it seems very reasonable to argue that a two- or three fold increase is impor-

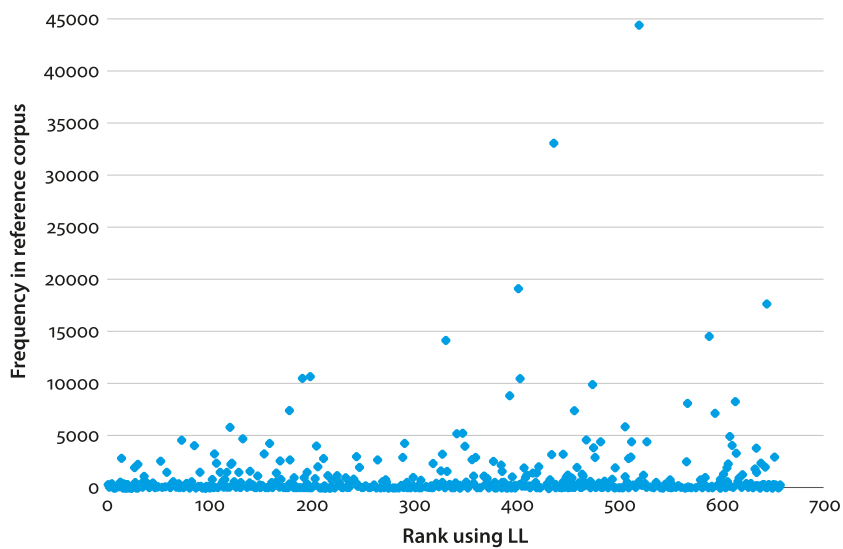


Figure 2. Rank frequency distribution chart based on the reference corpus frequencies for key words ranked using LL

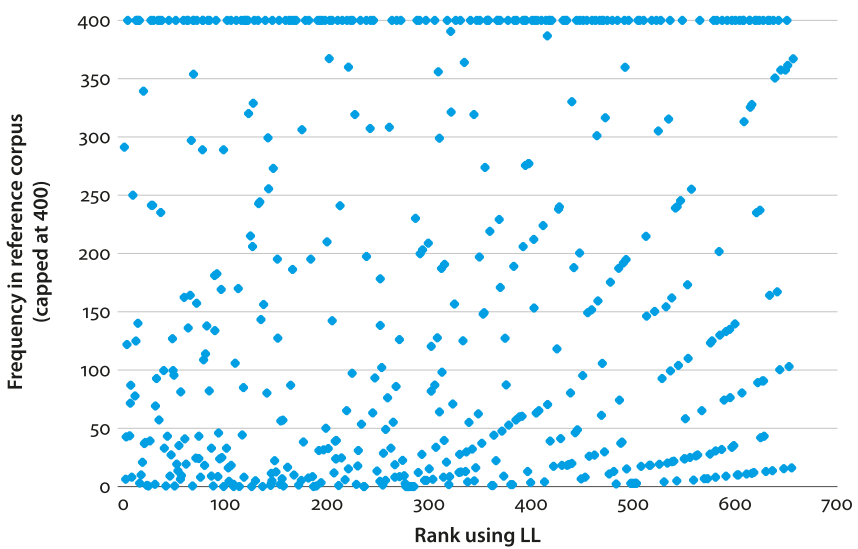


Figure 3. Rank frequency distribution chart based on the reference corpus frequencies for key words ranked using LL, with frequencies capped at 400

tant whether or not the items themselves are relatively rare in comparison with the sample sizes (i.e. the sizes of the corpora). This paper does not set out to challenge

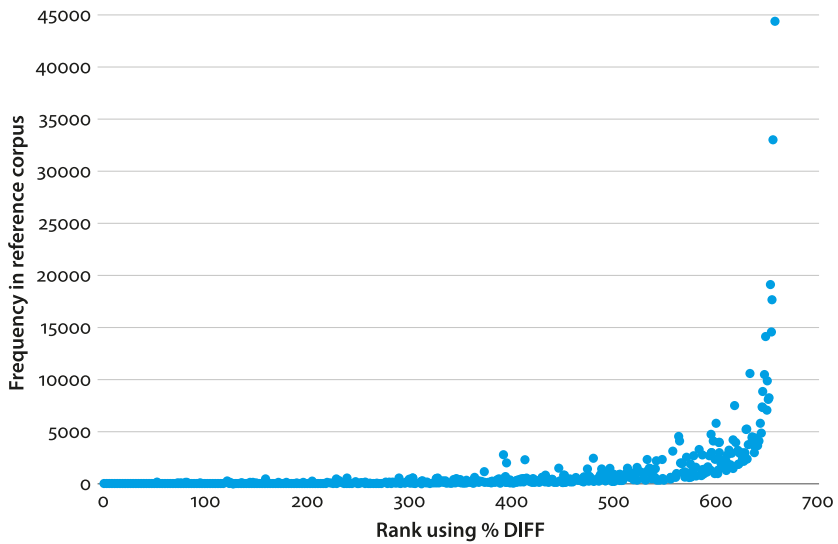


Figure 4. Rank frequency distribution chart based on the reference corpus frequencies for key words ranked using %DIFF

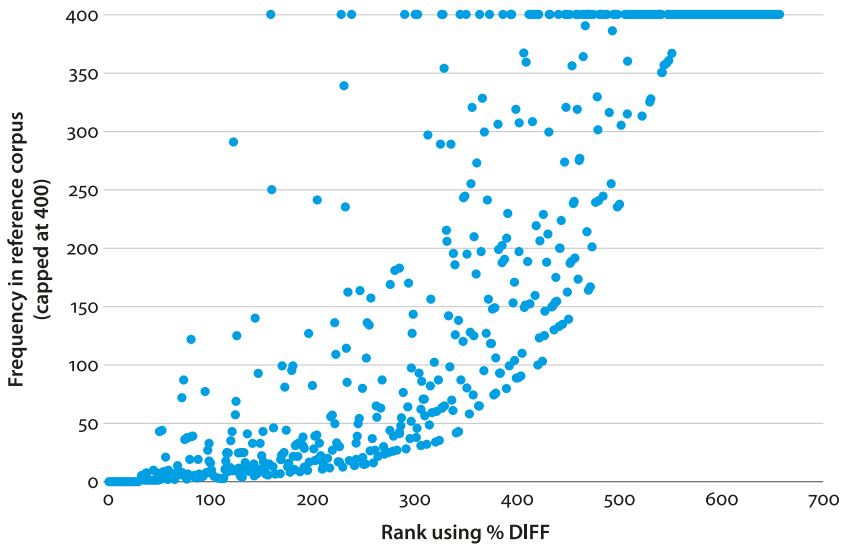


Figure 5. Rank frequency distribution chart based on the reference corpus frequencies for key words ranked using %DIFF, with frequencies capped at 400

studies working with the effect size measures for these kinds of research question. However, even for these it might also be helpful to find a means of also integrating

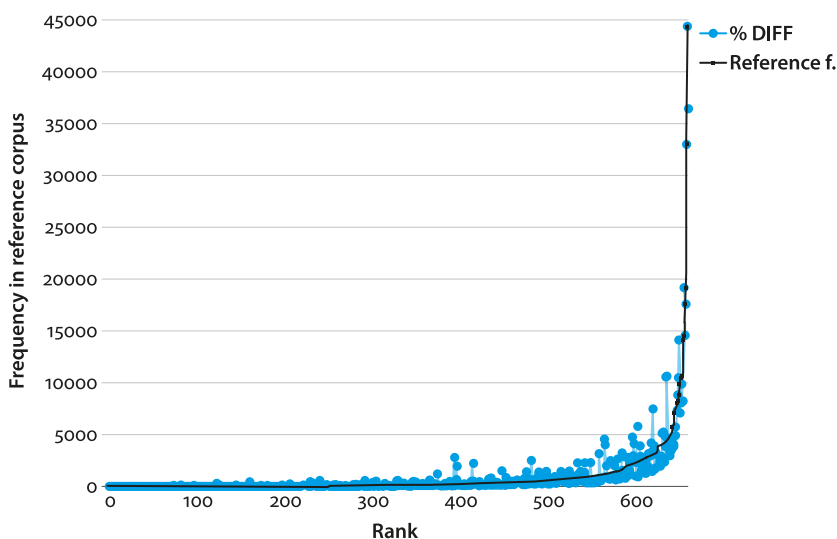


Figure 6. Rank frequency distribution chart based on the reference corpus frequencies for key words ranked using %DIFF (descending) and the reference corpus frequencies (ascending)

the $a + b$ vs. $c + d$ dimension. In brief, the %DIFF measure (and similar measures) could be good at revealing:

- i. Items in a pre-determined set of words or phrases that have larger differences in frequency;
- ii. Similarities between two corpora based on small %DIFF results (a topic beyond the scope of this paper);
- iii. Easy to understand figures for changes in relative frequency when items exist in both the study corpus and the reference corpus.

However, for studies where the unit of study is the text (individual texts, text samples, or collections of texts), it has been suggested that Bayes Factors should actually work very well, even when the reference corpus is very large. The LL measure is particularly good at revealing:

- i. What is thematically prominent in a text sample, a text or a collection of texts;
- ii. What might be features likely to be foregrounded/deviant/salient/marked (Leech & Short, 1981/2007) in a text sample, a text or a collection of texts;
- iii. What might be indicators of register or style.

Based on the findings of this paper, there are no specific suggestions for the UCREL Log-likelihood and Effect Size Calculator. It is reasonable for tools like

Table 13. Different research aims and uses of keyness values

A. Aim:	To determine the “aboutness” of each individual text
Research Type(s):	Corpus Stylistics studies on single novels or chapters in novels; Discourse Analysis on culturally prominent texts (e.g. political manifestos, Chairman’s Statements)
Unit of Study:	Individual texts
Suggested approach(es):	Measure KWs in each individual text, using a larger reference corpus from a similar genre/domain, or possibly using the other texts in the corpus as a reference corpus
Potential corroboration:	The reader of the text might agree that the text was about these things.
Follow up:	KW lists ranked by LL are likely to be revealing in themselves. A researcher might use some of these KWs as good starting points for further analysis.
B. Aim:	To identify words which are important in text of a particular kind.
Research Type(s):	Genre/Register analysis, using collections of texts from the target text varieties.
Unit of Study:	Corpus or sub-corpus of target genres/register
Suggested approach(es):	Measure KWs in separate genres or registers; Consider text dispersion keyness; Measure KKWs across individual texts within a genre or register
Potential corroboration:	A reader familiar with this genre or register would acknowledge that the KWs or KKWs are important to the field.
Follow up:	A researcher might use lists of results to show differences in importance of topic indicators or to find potential candidates which occur in two different sets to explore how use of these items differs.
C. Aim:	To identify ways in which certain groups of people are represented.
Research Type(s):	Critical Discourse Analysis
Unit of Study:	Corpus containing texts on specific topics or about specific groups
Suggested approach:	Measure KWs in the corpus against a reference corpus of similar texts about other groups.
Potential corroboration:	Readers/listeners may not be aware of how these language choices can manipulate their view of reality.
Follow up:	KW lists ranked by LL are likely to reveal some interesting differences; both lexical and grammatical items may lead to interesting discoveries about representation and argumentative structure.

Table 13. (continued)

D. Aim:	To determine major topics and themes in the discourse of specific social groups or differences in register and style (such as use of pronouns and contractions).
Research Type(s):	Sociolinguistic studies or Learner Corpus studies without predefined language items
Unit of Study:	Corpus
Suggested approach:	Measure KWs in one entire corpus against a comparable reference corpus.
Potential corroboration:	The reader of all the texts may not be aware of some of the differences because they are likely to be widely distributed across all the texts; a text chosen from this collection might “seem” to fit into the category; a text not exhibiting these major features might not “seem quite right” in terms of style.
Follow up:	A researcher would need to consider the results balanced against some understanding of how other expressions or non-linguistic features could be used by writers/speakers to perform similar communicative goals.
E. Aim:	To determine processes such as the adoption or adaptation of specific variants, specific forms and/or culturally loaded expressions
Research Type(s):	Sociolinguistic studies with pre-defined language items
Unit of Study:	Whole corpus
Suggested approach:	Measure differences for set of specific items in one entire corpus against a reference corpus; effect size measures such as %DIFF may be particularly useful.
Potential corroboration:	Small increases in the occurrence of culturally loaded expressions may provoke responses in certain kinds of reader/listener but not in others; with other features, the reader/listener may or may not be aware of the differences.

this to present a wide range of measures, and given the fact that the user enters values for one result at a time, responsibility rightly falls to the researchers to match appropriate measures to their specific study focus. Indeed, it is an excellent tool for small quantities of data and can be very helpful for double-checking results. Software developers of corpus tools will no doubt need to select a default setting for metrics like KW analysis, and this decision should be made according to the main kinds of research project envisaged for its users. *WordSmith Tools* now includes a log-ratio threshold setting that can be adjusted according to differ-

ent needs (Scott, 2019d). However, it is the researcher-users who need to be most keenly aware of how different measurements have advantages and disadvantages for different kinds of research.

Statistical methods and modelling must have developed further in recent years, but the interpretation of other measures is beyond the scope of this paper. The aim has been to re-consider the measures which are widely used, have been widely published in corpus linguistic research and are widely available in corpus software. Future evaluation of the applicability of other measures to the pertinent question for corpus linguists who are interested in texts rather than changes for a specific word will, however, need to consider the special requirements of being able to keep a balance between the different ways of looking at relationships between types and texts, text length and corpus, and the total sizes of combined study and reference corpora used in different studies. Researchers in the corpus linguistics community should recognise that different measures are currently available, and that it is not the case that one measure is the best for all situations. Providing transparent information about measures employed with a rationale related to each study's research questions should be a priority. Editors and reviewers of research articles should be wary of dismissing results on the basis of categorical judgements of the suitability (or otherwise) of KW metrics; rather they should evaluate the suitability of each choice according to the research purposes.

In conclusion, researchers using *WordSmith Tools*, *AntConc*, *WMatrix*, *CLiC*, *The Prime Machine* (or similar tools), should rest assured when taking candidate LL-ranked KWs and using these to explore concordance lines of words which are likely to be foregrounded/deviant/salient/marked. Looking down the list of key words ranked by LL (and with reference to the BIC scores) remains an excellent way to keep a proper balance between a word's raw frequency as a proportion of each corpus AND the size of the target corpus in comparison with the reference corpus AND the overall size of the combined study and reference corpora. This also holds for the application of the KW technique to other features such as groups of words from the same semantic field, n-grams or tags. When researchers select the top candidates from this ranked list (Top 10, Top 20, Top 50, etc.) including the keyness scores and the BIC scores in their tables of results will mean comparisons between other studies can be drawn more readily and risks of selection bias can be reduced.

Acknowledgements

The author is grateful for the detailed comments and suggestions from the anonymous reviewers. The author also gratefully acknowledges the help and support of Debby Gill in discussing and rearranging various equations.

References

- Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software>
- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). American Psychological Association.
- Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, 32(4), 346–359. <https://doi.org/10.1177/0075424204269894>
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, 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), 273–306. <https://doi.org/10.1177/0957926508088962>
- Bradley, J.V. (1960). *Distribution-free Statistical Tests*. Air Research and Development Command. <https://doi.org/10.21236/AD0249268>
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Cobb, T. (2000). *The Compleat Lexical Tutor* (Version 8.3) [Computer software]. Retrieved November, 2019, from <http://www.lextutor.ca>
- Croft, W.B., Metzler, D., & Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. Addison-Wesley.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.
- Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1), 77–104. <https://doi.org/10.3366/cor.2019.0162>
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.) *Corpus Approaches to Discourse: A Critical Review*. Routledge. <https://doi.org/10.4324/9781315179346-11>
- Gabrielatos, C., & Marchi, A. (2012). *Keyness: Appropriate metrics and practical issues* [Paper presentation]. CADS International Conference 2012, University of Bologna, Italy. https://www.researchgate.net/publication/261708842_Keyness_Appropriate_metrics_and_practical_issues
- Gabrielatos, C., Torgersen, E.N., Hoffmann, S., & Fox, S. (2010). A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics*, 38(4), 297–334. <https://doi.org/10.1177/0075424209352729>
- Grissom, R.J., & Kim, J.J. (2012). *Effect Sizes for Research: Univariate and Multivariate Applications*. Routledge. <https://doi.org/10.4324/9780203803233>

- Hardie, A. (2012). CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409. <https://doi.org/10.1075/ijcl.17.3.04har>
- Hardie, A. (2014a). Log Ratio – an informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)*. <http://cass.lancs.ac.uk/?p=1133>
- Hardie, A. (2014b). *Statistical identification of keywords, lockwords and collocations as a two-step procedure* [Paper presentation]. ICAME 35 Conference, University of Nottingham, Nottingham, UK.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Routledge.
- Jeaco, S. (2017). Concordancing lexical primings: The rationale and design of a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances*. John Benjamins. <https://doi.org/10.1075/scl.79.11jea>
- Johnston, J. E., Berry, K. J., & Mielke Jr, P. W. (2006). Measures of effect size for chi-squared and likelihood-ratio goodness-of-fit tests. *Perceptual and Motor Skills*, 103(2), 412–414. <https://doi.org/10.2466/pms.103.2.412-414>
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773. <https://doi.org/10.1080/01621459.1995.10476572>
- Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). *The Sketch Engine* [Paper presentation]. The 2003 International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.
- Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3), 37–72.
- Leech, G. N., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511642210>
- Leech, G. N., & Short, M. H. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose* (2nd ed.). Pearson Longman. (Original work published 1981)
- Lexical Computing Ltd. (2014). Statistics used in the Sketch Engine. <https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>
- Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. Routledge. <https://doi.org/10.4324/9780203076088>
- Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 11(3), 433–463. <https://doi.org/10.3366/cor.2016.0102>
- Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.
- Parsons, T. D., & Nelson, N. W. (2004). Paradigm shift in social science research: A significance testing and effect size estimation rapprochement? *PsycCRITIQUES*, 49(Suppl 3).
- Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, 5(2), 83–108. <https://doi.org/10.3366/cor.2010.0101>
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Raftery, A. E. (1986). A note on Bayes Factors for Log-Linear contingency table models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2), 249–250. <https://doi.org/10.1111/j.2517-6161.1986.tb01408.x>

- Rayson, P. (n.d.). UCREL Log-likelihood and effect size calculator. Retrieved November, 2019, from <http://ucrel.lancs.ac.uk/llwizard.html>
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549. <https://doi.org/10.1075/ijcl.13.4.06ray>
- Rayson, P., Berridge, D., & Francis, B. (2004). *Extending the Cochran rule for the comparison of word frequencies between corpora* [Paper presentation]. The 7th International Conference on Statistical Analysis of Textual Data, Louvain-la-Neuve, Belgium. https://eprints.lancs.ac.uk/id/eprint/12424/1/rbfo4_jadt.pdf
- Rayson, P., & Garside, R. (2000). *Comparing corpora using frequency profiling* [Paper presentation]. The Workshop on Comparing Corpora, Hong Kong University of Science and Technology, Hong Kong. https://eprints.lancs.ac.uk/id/eprint/11882/1/rg_acl2000.pdf
- Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–152. <https://doi.org/10.1075/ijcl.2.1.07ray>
- Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer. <https://doi.org/10.1007/978-1-4612-4578-0>
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2), 233–245. [https://doi.org/10.1016/S0346-251X\(97\)00011-0](https://doi.org/10.1016/S0346-251X(97)00011-0)
- Scott, M. (2016). *WordSmith Tools* (Version 7.0) [Computer software]. Stroud: Lexical Analysis Software.
- Scott, M. (2019a). WordSmith Tools online manual “KeyWords: Calculation”. Retrieved November, 2019, from https://lexically.net/downloads/version7/HTML/keywords_calculate_info.html
- Scott, M. (2019b). WordSmith Tools online manual “KeyWords”. Retrieved November, 2019, from <https://lexically.net/downloads/version7/HTML/keywords2.html>
- Scott, M. (2019c). WordSmith Tools online manual “KeyWords: Thinking about keyness”. Retrieved November, 2019, from https://lexically.net/downloads/version7/HTML/thinking_about_keyness.html
- Scott, M. (2019d). WordSmith Tools online manual “KeyWords: Keyness definition”. Retrieved November, 2019, from https://lexically.net/downloads/version7/HTML/keyness_definition.html
- Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. John Benjamins. <https://doi.org/10.1075/scl.22>
- 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. 3–12). Peter Lang.
- Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

Address for correspondence

Stephen Jeaco
Department of Applied Linguistics
HS431, Xi'an Jiaotong-Liverpool University
111 Ren'ai Lu, Suzhou Industrial Park
Suzhou, Jiangsu Province
P. R. China
steve.jeaco@xjtlu.edu.cn