Incorporating text dispersion into keyword analyses

Jesse Egbert[1] and Doug Biber[1]

Abstract

Keyword analysis has become an indispensable tool for discourse analysts, being applied to identify the words that are especially characteristic of the texts in a target discourse domain. But, surprisingly, the statistical computation of keyness makes no reference to those texts. Rather, once a corpus has been constructed, it is treated as a homogeneous whole for the computation of keyness. As a result, the keywords in such lists are relatively frequent in the corpus, but they are often not widely dispersed across the texts of that corpus and are thus not truly representative of the target discourse domain. The purpose of this study is to propose a new method for keyword analysis – text dispersion keyness – that is based on text dispersion, rather than corpus frequency. We compare the effectiveness of this measure to four other methods for computing keyness, carrying out a series of case studies to identify the keywords that are typical of online travel blogs. A variety of quantitative and qualitative analyses are carried out to compare these methods based on their content-generalisability and content-distinctiveness, demonstrating that text dispersion keyness is a superior measure for generating keyword lists.

Keywords: distinctiveness, generalisability, keyword analysis, lexical dispersion, word importance.

1. Introduction

Keyword analysis is one of the most widely used methods in corpus linguistics and corpus assisted discourse studies (CADS). The first writers (e.g., Firth, 1957; and Williams, 1983) to make reference to keywords ‘intuitively focused on words that they believed embodied important

[1] English Department, PO Box 6032, Northern Arizona University, Flagstaff, AZ 86011, USA.
Correspondence to: Jesse Egbert, *e-mail*: Jesse.Egbert@nau.edu

Corpora 2019 Vol. 14 (1): 77–104
DOI: 10.3366/cor.2019.0162
© Edinburgh University Press
www.euppublishing.com/cor

concepts that reflected societal or cultural concerns' (Baker, 2004: 346). Since then, other scholars have broadened the construct of keyness beyond words that are socially and culturally significant to any words that offer important insights into the 'aboutness' of a text or corpus. Baker (2004: 347) summarises this broad conceptual definition for keywords: 'An examination of the keywords that occur when two corpora are compared together should reveal the most significant lexical differences between them, in terms of aboutness and style'.

More than twenty years ago, Mike Scott proposed a simple yet important definition for keywords: words that occur 'with unusual frequency' in a target corpus when compared with a reference corpus (Scott, 1997: 236). This definition was operationalised as words that occur with significantly greater frequencies in the target corpus. Scott initially used chi-square and later the log-likelihood statistic, to calculate keyness (Scott and Tribble, 2006). In this paper, we refer to this approach as 'corpus frequency keyness'. This operational definition makes it possible to identify keywords automatically using a computer program – an approach that is simple to understand and easy to perform. As a result, this approach has been widely accepted by researchers in corpus linguistics and CADS.

Most researchers who use corpus frequency keyness are interested in identifying words that are strongly associated with the content of texts in a target discourse domain, such as erotic narratives (Baker, 2004), newspaper articles (Baker *et al.*, 2004; and Gabrielatos and Baker, 2008), student essays (Römer and Wulff, 2008), health consultations (Adolphs *et al.*, 2004), conversations (Xiao and McEnery, 2005), academic books and articles (Xiao and McEnery, 2005; and Paquot and Bestgen, 2009), and web documents (Kilgarriff, 2012). Yet, surprisingly, the statistical computation of corpus frequency keyness makes no reference to those texts, and offers little to no information on word use within them. Rather, once individual texts have been compiled into a corpus, the corpus itself is treated as the single unit for the computation of keyness. One underlying – but usually unrecognised – assumption of this approach is that the corpus is homogeneous, and thus words are evenly distributed across the corpus. In actual fact, though, corpus frequency keywords can be, and often are, frequent in a corpus, but are not widely dispersed across the texts of that corpus. As a result, such words are not truly typical of the discourse domain represented by the corpus.

The effectiveness of a keyword analysis can be evaluated with respect to two criteria: content-distinctiveness and content-generalisability. Content-distinctiveness refers to the strength of the relationship between a keyword and the content of the discourse domain represented by the target corpus, in contrast to all other discourse domains (represented by the reference corpus). Content-distinctive keywords should be interpretable and relevant. Thus, there are two sub-considerations here: (*1*) these keywords should be typical of the target discourse domain in contrast with words that are typical of other discourse domains; and (*2*) these words should reflect the content-'aboutness' of texts in the target discourse domain (rather than reflecting the grammatical characteristics associated with the register of that domain).

Content-generalisability is the degree to which a keyword represents the content of the discourse used across the full range of texts in the target corpus. This criterion requires that words be both (*a*) generalisable to many texts in the target domain, and (*b*) representative of the words that offer insight into the actual content-'aboutness' of those texts. Content-generalisability is important because it determines the extent to which a keyword is representative of, and meaningful for, the entire target corpus. The criterion of content-generalisability gives preference to content words over grammatical or function words. We acknowledge that some researchers in the past have been interested in key function words (see, for example, Baker, 2004; and Culpeper, 2009, 2014). However, the goal of most keyword studies is to describe the content of texts in a discourse domain (see Firth 1957; Williams 1983; Xiao and McEnery, 2005; Baker, 2004; Baker *et al.*, 2004; and Gabrielatos and Baker, 2008). While we agree that the study of grammatical patterns is also an important component of discourse analysis, we would argue that there are much more effective methods for carrying out such analyses than keyword analysis. Thus, our methods for evaluation here focus on content-generalisability and discount the representation of isolated function words.

In summary, for the purposes of discourse analysis, we believe that a corpus should be regarded as a representative collection of texts; and, by extension, the statistical computation of keyness should capture words that are characteristic of the content of those texts. The purpose of this study is to propose a new method for identifying keywords – text dispersion keyness – that is based on a word's dispersion across the texts of a corpus rather than its overall frequency in the corpus. We demonstrate that this approach results in the identification of keywords that achieve both greater content-distinctiveness and content-generalisability.

In the sections below, we first document several problems associated with the corpus frequency approach to keywords. We then review previous attempts to account for text dispersion in keyword analysis and introduce the rationale and methods for our new proposed method: text dispersion keyness. The remainder of the paper, then, focusses on an empirical comparison of different methods for measuring keywords, with a focus on the advantages of text dispersion keyness.

1.1 Corpus frequency keyword analysis

Corpus frequency keyword analysis treats the corpus as its unit of observation; thus, a study using this approach has a total of two units: a target corpus and a reference corpus. After counting the frequency for every word in these two corpora, the keyness for each word in the target corpus is calculated using one of several simple statistics, described below. Words that are used with statistically greater frequencies in the target corpus (in comparison to their frequency in the reference corpus) are considered to be 'key'. Keyness values are continuous scores, and so keywords are usually ranked according to the keyness statistic, with the assumption that words with higher keyness values are more strongly associated with the target corpus. Some keyness formulas also produce *p*-values that may be compared with a pre-established alpha level to determine whether a word is key.

Several statistical measures have been proposed for measuring keyness, including chi-square (Scott, 1997), log-likelihood (Rayson and Garside, 2000), log ratio,[2] simple frequency difference (Gabrielatos and Marchi, 2012) and simple maths (Kilgarriff, 2009). Although these measures differ in their mathematical computation, they are all examples of corpus frequency keyness in that they compare the frequency of a word between two corpora, without regard to the word's dispersion across texts. Some scholars have identified serious limitations with the corpus frequency approach (see, for example, Johnson and Ensslin, 2006; and Baker, 2004). However, these limitations are identified on theoretical grounds, and to date, we know of no previous empirical comparisons of the different methods.

Relatively few previous studies have addressed the content-distinctiveness and generalisability of corpus frequency keyword lists. Johnson and Ensslin (2006: 10) describe how extensive manual work is required to derive 'a "true" keyword list from the "rough" keyword list produced by the software'. In part, this process is required because current methods for deriving keywords are prone to assigning high keyness values to words that are not distinctive and generalisable to the content of the target discourse domain.

Researchers have dealt with the 'roughness' of corpus frequency keyword lists by sorting through dozens, hundreds or even thousands of keywords and manually selecting a relatively small number of words that are deemed to be content-distinctive (see, for example, Adolphs *et al.*, 2004). As a result, it is quite uncommon for researchers to actually report the complete keyword lists produced by corpus frequency software programs in published research. Instead, researchers typically include only the subset of the original wordlist that is deemed to be content-distinctive.

In a rare exception, Culpeper (2009) includes the full lists of corpus frequency keywords for each of six characters in Shakespeare's *Romeo and Juliet*. Among the first few keywords for each character are examples of words that do not satisfy the criterion of content-distinctiveness, as we have described it above. For example, the keyword list for *Mercutio* includes *a*, *of*, *the* and *an*, and the lists for *Romeo* and *Juliet* both include the word *that*. All of these words are general, high-frequency function words, making it difficult to associate them with the aboutness of discourse produced by a particular character. In other words, these words are not content-distinctive.

[2] See: http://cass.lancs.ac.uk/?p=1133; see also Baron *et al*. (2009).

Similarly, corpus frequency keyword lists often fail to achieve content-generalisability. Baker (2004) discusses the (non-)generalisability of corpus frequency keywords, noting that the corpus frequency keyword approach is prone to identifying words as key even though they only occur in a single text in the corpus. He illustrates this using the example of *wuz*, a word identified as a gay keyword using the corpus frequency approach because it occurred thirty-two times in a corpus of about 350 gay texts and zero times in a corpus of about 350 lesbian texts. However, a closer analysis revealed that all thirty-two occurrences of *wuz* were restricted to a single text.

In another study, Baker (2010: 105) includes small sub-samples of keywords for the Lancaster–Oslo/Bergen (LOB) corpus when compared with the Freiberg Lancaster–Oslo/Bergen (FLOB) corpus, and *vice versa*, in an effort to identify evidence of language change between 1961 (LOB) and 1991 (FLOB). However, based on our own analysis of these two corpora, we found that some of these keywords occur in only a very small number of texts. For the LOB keywords, these include *rhodesia* and *kenya*, which occur in 2.2 percent (eleven texts) and 1.6 percent (eight texts) of the corpus, respectively. For the FLOB keywords, these include the words *privatisation* and *fucking*, which occur in 3 percent (fifteen texts) and 1.8 percent (nine texts) of the corpus, respectively. It appears that these are cases of words that are distinctive but not generalisable.

The limitations of the corpus frequency approach are often masked by the qualitative methods used in these studies. That is, discourse analysts typically select words from the keyword list that are deemed meaningful or interesting to investigate further, while simply disregarding words that are non-distinctive and/or non-generalisable. But the need for such an approach indicates that the quantitative corpus methods are not as effective as they could be. If a keyword measure is performing well, we should be able to account for and explain the highest ranking keywords in the list. To put it another way, if high-ranking words in our keyword lists are difficult or impossible to explain, even with advanced knowledge about the target domain, we should see this as evidence that our keyword methods need to be improved. We propose that the limitations of corpus frequency keyness documented here, namely the lack of content-distinctiveness and content-generalisability, can be overcome by accounting for text dispersion.

1.2 Text dispersion in keyword analysis

Only a few previous corpus keyword studies have attempted to account for text dispersion. These studies have relied on one of two approaches: key keyword analysis and minimum range.[3] The first approach is referred to as

[3] A third possible method is simple maths (Kilgarriff, 2009). This measure is based on the ratio between frequencies in two corpora, with a constant (default = 1) added to both the numerator and the denominator to address the problem of low or zero frequencies. Simple

'key keyword analysis' (Scott, 1997). Key keywords are words that are key in a large proportion of the texts in a corpus (Baker *et al.*, 2013). To find key keywords, a separate frequency-based keyword analysis is performed to compare each text in the target corpus to the entire reference corpus. Key keywords are those that show up as key in a large number of texts from the target corpus. For example, Baker (2004) uses key keyword analysis to identify keywords in two corpora, one of gay texts and the other of lesbian texts. The top key keyword in the gay corpus was *his*, which was key in 334 out of the 350 texts, followed by *he*, which was key in 328 texts. In the lesbian corpus, *her* and *she* were the top keywords, occurring in 327 and 320 texts, respectively, out of a corpus of 350 texts.[4] In that study, Baker found that only about twenty words were key in more than twenty texts in corpora of 350 texts. As a result, he describes key keyword analysis as an overly conservative approach that results in a list that (*1*) contains very few keywords, and thus (*2*) 'confirms expectations, rather than reveal[s] hidden patterns' (Baker, 2004: 351).

In response to the limitations of key keyword analysis, Baker (2004) proposes a second approach to account for text dispersion. This method uses the traditional corpus frequency keyword approach but adds an additional step that eliminates words that do not meet a pre-determined text dispersion or range. Range is often measured as a proportion of the total texts in the corpus (Baker, 2004, 2010). For example, Millar and Budgell (2008) limited their keyword lists to words that occurred in a minimum of 30 percent of the texts in the corpus. However, Baker (2004) notes that among studies that use range to account for keyword dispersion there is no consensus on the ideal text frequency or percentage to use as a threshold.

None of these methods of accounting for dispersion has been empirically evaluated to determine whether they are successful in producing lists of keywords that are distinctive and generalisable. However, noting deficiencies in the key keyword and minimum dispersion approaches, Baker (2004: 351) notes that 'it would be useful to find a way that combines the strengths of key keywords with those of keywords but is neither too general or exaggerates the importance of a word based on the eccentricities of individual files'. In an effort to respond to calls such as these, and in the spirit of searching for the most effective operational definition for keyness, we propose a new method for keyword analysis, which is introduced and described in the next section.

maths can be based on ratios of raw frequencies or average reduced frequencies (ARF; see Savický and Hlaváčová, 2002). An anonymous reviewer noted that simple maths can also be computed in SketchEngine based on a ratio of text counts (i.e., the number of texts a word appears in). However, we are not aware of any published research that has used this method.

[4] For a more complete introduction to key keyword analysis, the reader is referred to McEnery *et al.* (2006: 311–18).

1.3 A new method for keyword analysis

We propose a different method for keyword analysis that is entirely dispersion-based. This new method uses the text, rather than the corpus, as the unit of observation. In addition, this method entirely disregards word frequency and instead generates keyword lists based solely on word dispersion across texts. Considering the long tradition of corpus frequency keyword analysis, some might find it unsettling to entirely disregard token frequency when measuring keyness. In an effort to address these concerns, we offer some details about how this new method came about.

Recently, we used keyword analysis for a large-scale corpus-based investigation of online registers of English, using the Corpus of Online Register of English (CORE) (see Biber and Egbert, 2018). In the early stages of that project, we attempted to use the corpus frequency keyword method, with no effort to account for dispersion, to generate keyword lists for the registers in our corpus. To do this, we treated all texts of a single register of interest as the target corpus with all other texts in the corpus as the reference corpus. Although some of the words in these lists met the criteria of being both distinctive and generalisable, we soon became aware of major limitations in the keyword lists that were produced. It was obvious that many of these issues were related to not accounting for text dispersion.

In response, following Baker (2004), we began experimenting with a variety of minimum dispersion cut-offs to determine the best threshold to use in our research. Early in this process, it was easy to see that dispersion is critically important to discovering distinctive and generalisable keywords as the quality of the keyword lists began to improve dramatically. However, we also learned that it is extremely difficult, if not impossible, to find a single dispersion threshold that works for corpora of all sizes. Because it is based on a random sample from the searchable web, the sub-corpora in CORE vary drastically in their sizes. For example, whereas the News Reports sub-corpus contains 10,399 texts, the Short Story corpus contains only 272 texts. We found that a dispersion requirement that relied on a simple percentage, while reasonable in theory, did not work well for the very large corpora. Even a very low threshold such as 5 percent was difficult to achieve with the News Reports, in which words were required to occur in 520 texts before they would be considered to be key. On the opposite extreme, we found that stricter minimum percents, such as the 30 percent cut-off used by Millar and Budgell (2008), strongly favoured high-frequency function words which were generalisable but often not distinctive.

During this time, we began to think broadly about the construct of keyness and other possible ways of operationalising it. We determined that in order to be content generalisable and distinctive, keywords should be used by many different writers/speakers. We were not interested in how many times authors repeated words, especially since the repeated words were often text-skewed and non-generalisable. These repeated words are often significant for a particular text, offering insights into its topic and content, but not for the entire corpus. As a result, we set out to develop a method for identifying distinctive and content generalisable keywords regardless of their frequency in a corpus.

We hypothesised that keyness could be measured without making any reference to word frequency by focussing entirely on the text dispersion of words. In part, this hypothesis was based on the fact that a word occurring in numerous texts will necessarily also have at least a moderate frequency. Instead of comparing the total frequency for a word in the target and reference corpora, this new method compares word use between the target and reference corpus in terms of the total number of texts where a word occurs at least once. As with many applications of the corpus frequency approach, we chose to use log-likelihood, or G^2, to compare these numbers statistically. The formula for G^2 is[5]:

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

Where O_i is the observed number of texts where the word occurs in the target and reference corpus and E_i is the expected number of texts where the word occurs in the target and reference corpus.

The expected values (E_i) are calculated using the following formula:

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

Where N_i is the total number of texts in the target and reference corpora.

We decided to use log-likelihood because the dispersion of words across texts, like word frequencies, tend to follow a Zipfian distribution, and the log-likelihood statistic estimates probabilities more accurately when counts are very low (see Dunning, 1993; and Kilgarriff, 2005). The words in the resulting list are ranked in order from highest to lowest log-likelihood value. We have given this new method the label of ‘text dispersion’ keyword analysis because of its focus on word dispersion across texts. Table 1 compares the characteristics of corpus frequency keyness and text dispersion keyness with regard to several parameters.

1.4 Study aims and outline

Up to this point, we have surveyed the current state of the art for methods of identifying keywords. We have also discussed the importance of dispersion in keyword analysis and introduced a new dispersion-based method for identifying keywords. With that background, our primary objective for this

[5] More information can be found at: http://ucrel.lancs.ac.uk/llwizard.html.

	Corpus frequency keyness	Text dispersion keyness
Definition	Words that are statistically more frequent in a target corpus when compared with a reference corpus	Words that occur in statistically more texts in a target corpus when compared with a reference corpus
Variable	Frequency in corpora	Dispersion across texts

Formula	Log-likelihood
	O = observed word frequency
	E = expected word frequency
		Log-likelihood
		O = observed word dispersion (number of texts)
		E = expected word dispersion (number of texts)

Requirements	– At least one text in target and reference corpora – Software (e.g., AntConc, WordSmith) or specialised program.	– Many texts in target and reference corpora – Specialised program

Table 1: Side-by-side comparison of corpus frequency keyword analysis and text dispersion keyword analysis.

study is to compare empirically the effectiveness of the following five keyword methods:

(*1*) Corpus frequency keyness;
(*2*) Corpus frequency keyness, with a minimum dispersion criteria of ten percent of the texts;
(*3*) Corpus frequency keyness, with a minimum dispersion criteria of thirty percent of the texts;
(*4*) Key keyword analysis; and,
(*5*) Text dispersion keyness.

Each of these keyness methods will be used to generate keyword lists for the discourse domain of online travel blogs. The results will then be compared quantitatively and qualitatively based on the previously discussed criteria of content-distinctiveness and content-generalisability.

Section 2 introduces the corpus used in this study, the keyword methods that are compared, and the quantitative and qualitative methods used to compare them. Section 3 describes the results of those comparisons. Finally, in Section 4 we summarise the results and evaluate the strengths and limitations of the various methods.

2. Methods

This section describes the methods used in this study. Section 2.1 reports on the design and composition of the target corpus and reference corpus we used. In Section 2.2, we introduce the five keyword methods compared in this study. In Section 2.3, we describe the quantitative and qualitative methods used to make those comparisons.

2.1 Corpora

The two corpora used for all of the analyses in this study are part of the Corpus of Online Registers of English (CORE). CORE was developed as part of a large-scale project focussed on exploring the registers that exist on the web and describing them linguistically (see Biber *et al.*, 2015; Egbert *et al.*, 2015; and Biber and Egbert, 2018). CORE consists of 48,571 documents sampled randomly from the searchable web. Each of these documents was coded for its situational characteristics by four independent coders recruited through Amazon's Mechanical Turk. These situational ratings were then used to classify each document into a register category (e.g., discussion forums, news reports and encyclopedia articles).

The target corpus used in the keyword analyses reported below contains all of the online travel blogs from CORE (371 texts; 330,918 words). Travel blogs are one of the narrative registers in CORE, along with news reports, personal blogs and sports blogs. Travel blogs are typically written by either a professional travel writer or a travel enthusiast to an audience of readers who are interested in a particular location or in exotic travel destinations in general. These blogs are usually pre-planned and edited by the author, and their purpose is to narrate travel experiences and describe and comment on travel destinations, logistics and tips. We chose this register because there has been very little corpus linguistic research focussed on it, and also because we expected that it would contain interesting and highly distinctive words.

The reference corpus used throughout this study comprises all of the documents in CORE except the travel blogs. This corpus contains 48,200 texts (about fifty-two million words) that fall into a wide range of register categories.[6] All of these documents share the characteristic of being published on the searchable web. This is the ideal reference corpus for learning about the distinguishing characteristics of travel blogs because it is large, varied and it representative of the web, the general domain that travel blogs come from. This last characteristic makes it possible to control for many variables associated with online language, allowing us to focus on identifying keywords that are associated with the specific purpose and topic of travel blogs, rather than on characteristics shared by other web registers.

[6] The reader is referred to Biber and Egbert (2018) for a complete description of these texts and the register categories they belong to.

2.2 Keyword methods

In this study, we evaluate and compare five keyword methods. These methods were chosen to represent the range of approaches to keyword analysis that have been proposed in previous literature.

The first method is corpus frequency keyword analysis – the traditional, frequency-based method which disregards dispersion (see Section 1.1). The second and third methods are both corpus frequency keyword analysis, combined with minimum text dispersion thresholds of 10 percent and 30 percent, respectively. The minimum range of 30 percent for the third method was included because it represents the highest minimum dispersion range that we have encountered in previous literature (Millar and Budgell, 2008). Including two minimum range thresholds is desirable because it allows us to measure the effect of changing the minimum range on keyword results. The first three methods were performed using computer programs written in Python, based on the log-likelihood formula.[7]

The fourth method is key keyword analysis. Key keyword analysis identifies and ranks keywords based on the percentage of the texts in the target corpus in which it was key. The key keyword analyses were carried out using WordSmith Tools version 7.0.

The fifth and final method is text dispersion keyword analysis – the dispersion-based method introduced in Section 1.3. A Python program was created to generate keyword lists using this method. As with Methods 1 to 3, this was based on the log-likelihood formula. The difference is that this formula used text dispersion as the variable rather than corpus frequency.

For the purposes of this study, the 100 highest ranked keywords from each method are analysed. Some previous researchers have used tests of significance (with an alpha criterion and *p*-values) in order to establish which words are key. However, different methods rely on different measures of significance, resulting in different numbers of 'significant' keywords, and thus lists of 'significant' keywords cannot be meaningfully compared across studies. We believe that the top 100 keywords is enough to learn about the strengths and limitations of the keyword methods, while still allowing us to conduct in-depth qualitative investigations.

2.3 Methods for evaluation and comparison

In order to compare the five methods described in the previous section, it was necessary to develop metrics that could be used to evaluate the effectiveness of keyness measures. Following Baker (2004: 347), we determined that effective keyness methods 'should reveal the most significant lexical differences between [two corpora], in terms of aboutness and style'.

[7] The formula can be found at: http://ucrel.lancs.ac.uk/llwizard.html.

As noted in Section 1, our primary interest is in identifying words that represent the ‘content-aboutness’ of a large proportion of the texts in a discourse domain. Based on this we developed two criteria for evaluating keyness measures: content-generalisability and content-distinctiveness. These two criteria are measured using a variety of quantitative methods, including relative frequency (types and tokens) and relative dispersion. Relative frequency is measured by comparing type and token frequencies in the target corpus with those of the reference corpus. Relative dispersion compares the percent of the texts in the target and reference corpora that contain keywords. Relative frequency is a strong indicator of content-distinctiveness. Relative dispersion is used to evaluate both content-generalisability and content-distinctiveness. We also quantitatively analyse lexical categories that are typically non-distinctive (‘function words’ and ‘frequent verbs’) and non-generalisable (‘proper nouns’ and ‘abbreviations’). We have found that function words and frequent verbs are typically difficult to clearly associate with the target corpus because they tend to be highly frequent and widely dispersed in all discourse domains. We have also found that it is rare for proper nouns and abbreviations to occur in a large proportion of the texts in a corpus. While we believe these simple methods are useful for evaluating the various keyword methods, future research could explore more sophisticated metrics for evaluating and comparing keyword lists.

In Section 3.2 we turn to a qualitative investigation of the distinctiveness and content-generalisability of the top 100 words produced by the five keyword methods. The presence of words from the non-distinctive (function words and frequent verbs) and non-generalisable (proper nouns and abbreviations) lists are compared between the text dispersion keyness method and the other four methods. We also discuss the overall lists in terms of their quality in relation to our goals of identifying words that are distinctive to the domain of travel blogs and generalisable to the content of the texts in that domain.

3. Results and discussion

After generating keyword lists using each of the methods described in Section 2.2, ordered from highest to lowest keyness value, we extracted the top 100 keywords. This resulted in five 100-word lists that serve as the basis for all of the analyses and comparisons in this section.

3.1 Quantitative results

In this section, we present quantitative results for the frequency (tokens and types) and dispersion of the top 100 keywords produced by each of the five keyness methods. These measures are reported for the Travel Blog (TB) corpus and CORE. In addition, the ratio of these measures (TB/CORE) is

Method	Top 100 keywords Tokens per 1,000 words		
	Travel blogs	CORE	Travel blogs/ CORE
Corpus frequency	138.70	101.89	1.36
Corpus frequency + min. of 10 texts	174.82	136.38	1.28
Corpus frequency + min. of 30 texts	346.40	315.52	1.10
Key keyword	189.58	130.31	1.45
Text dispersion	38.97	6.77	5.75

Table 2: Frequency of occurrence (per 1,000 words) for the top 100 keyword lists in the TB corpus and CORE.

Method	Top 100 keywords Types per 1,000 words		
	Travel blogs	CORE	Travel blogs/ CORE
Corpus frequency	25.24	11.89	2.12
Corpus frequency + min. of 10 texts	32.80	16.86	1.95
Corpus frequency + min. of 30 texts	55.48	38.39	1.45
Key keyword	23.12	9.16	2.52
Text dispersion	17.28	3.24	5.34

Table 3: Average percent of the types from the top 100 keyword lists that occur (per 1,000 words) in the TB corpus and CORE.

reported to provide insight into the relative frequency and relative dispersion of the keywords in the target corpus, when compared with the reference corpus. First, we present the rates of occurrence (i.e., tokens) per 1,000 words in Table 2, where Column 2 contains the combined rate of occurrence for the top 100 keyword tokens (per 1,000 words) in the target corpus, Column 3 contains the same count for the CORE, and Column 4 contains the ratio of these two token counts. Next, we present the average percent of keyword types from the top 100 keyword lists that occur, on average, per 1,000 words in Table 3,[8] which is organised in the same fashion as Table 2. Finally, we compare the text dispersion for each word in the target corpus and reference corpus. These results are displayed in the scatterplots in Figure 1, where the percentage of the texts that contain each word are plotted for the travel blog corpus (*y*-axis) and the CORE (*x*-axis). The dispersion results are also summarised in Table 4.

[8] To calculate these average percentages we counted the number of word types from each keyword list that occured in each text, normalised those counts (per 1,000 words) based on the text length, and calculated the mean across all texts in the corpus.

Method	Top 100 keywords Proportion of texts containing type		
	Travel blogs	CORE	Travel blogs/ CORE
Corpus frequency	0.26	0.15	1.74
Corpus frequency + min. of 10 texts	0.34	0.22	1.60
Corpus frequency + min. of 30 texts	0.58	0.49	1.19
Key keyword	0.24	0.12	2.06
Text dispersion	0.18	0.04	4.29

Table 4: Mean dispersion rates (percent of texts containing types) in the TB corpus and CORE.

As Table 2 shows, within the TB corpus, the top 100 keywords from the text dispersion keyness method had the lowest token count of the five methods. This shows that text dispersion keyness tends to identify keywords that are relatively less frequent than other keyness methods. In contrast, the keywords identified by the corpus frequency (CF) method were more than three times more frequent than the text dispersion keywords.

For all five methods, the average token counts for keywords were higher in the TB corpus than in CORE. This confirms that all of the methods are identifying words that are occurring more frequently in the target corpus. While we expect this for the three corpus frequency methods and for the key keyword method, this is not necessarily what we would expect for text dispersion keyness since this method entirely disregards frequency. There are important differences, however, in the ratio of TB and CORE frequencies. For the three CF methods and key keyword analysis, this ratio is between 1.10 and 1.45. In other words, for the first four methods the top 100 words are between 10 percent and 45 percent more frequent in the target corpus than in the reference corpus. Contrast this with the top 100 words from the text dispersion method which are, on average, 475 percent more frequent in Travel Blogs than in the rest of CORE.

The results for type frequencies are contained in Table 3. In contrast with the token frequency counts, which are rates of occurrence for the frequency of keyword tokens, the type results represent the percent of word types on average (out of the lists of 100 keyword types) that occur (at least once) per 1,000 words. In other words, these type results represent the percent of the types from the list of 100 keywords we would expect to see in the average 1,000 word run of text. As with the token frequency analysis, the top 100 keyword types from the text dispersion method were used the least frequently. In addition, the word types from the corpus frequency methods were used more frequently as the minimum text range increased.

The top 100 words identified by the text dispersion method were 434 percent more frequent in the TB corpus when compared with CORE. In contrast, the ratios from the other four methods ranged between 1.45 and 2.52, revealing that the method with the second highest target-reference

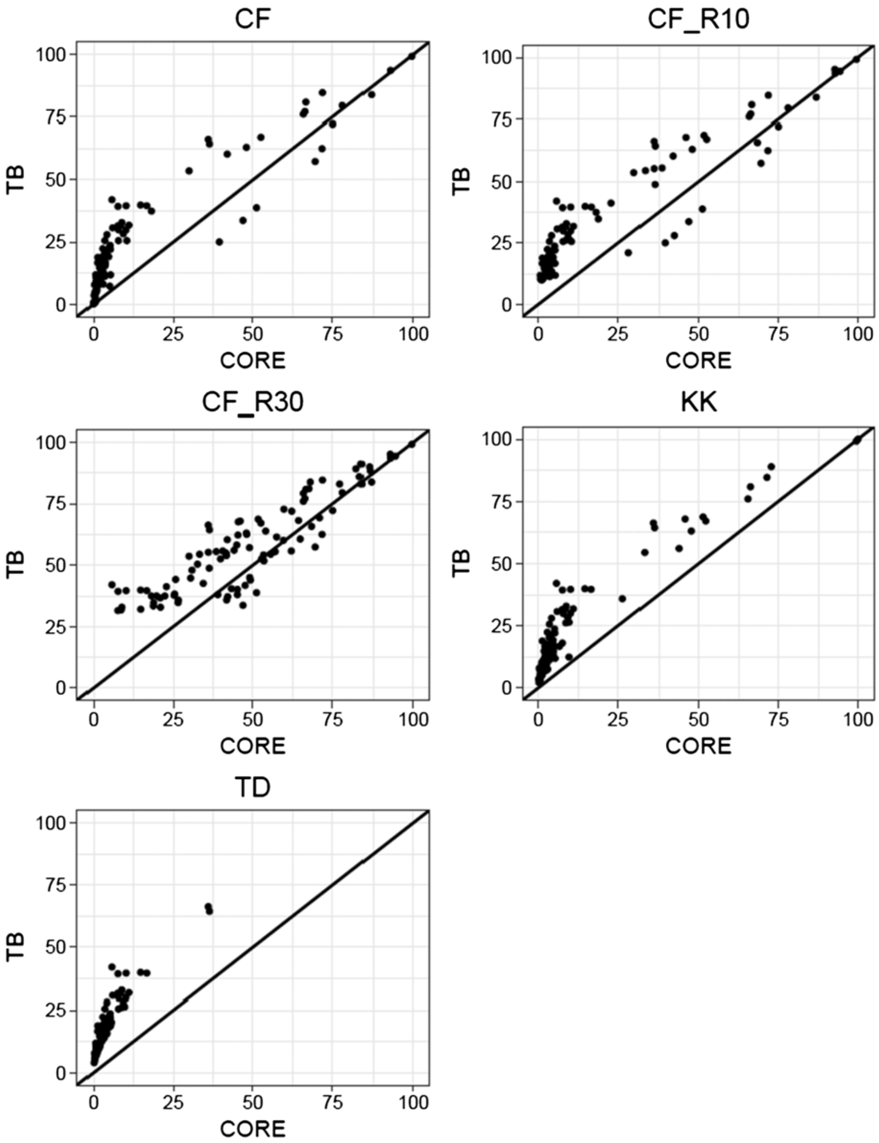

Figure 1: Scatterplots showing dispersion (in percent of texts) of keywords from the five methods in the TB and CORE corpora.

ratio was key keywords, in which the top 100 words were 152 percent more frequent in the TB corpus as in CORE.

Next we look at the results for relative word dispersion, which disregards word frequency entirely and instead measures the percent of the texts in the TB corpus and CORE that contain each of the keywords on the lists from the five methods. Figure 1 displays these results in five separate scatterplots. Each point represents one of the 100 keywords identified by the method labelled above the plot. The location of each point in the graph shows the percent of the texts in the CORE (*x*-axis) and travel blogs (*y*-axis) corpora where that word occurs. The 45 degree line is present as a point of reference. Points that fall near the line represent words that are approximately equal in their dispersion in TB and CORE. Points above the line represent words that are more widely dispersed in the TB corpus than CORE. Points below the line represent words that are more dispersed in CORE. In other words, this line could be thought of as a 'content-distinctiveness line', where the higher above the line that words fall, the more distinctive they are to the domain of travel blogs.

The results in Figure 1 reveal that some of the words identified as key by the three CF methods are actually more widely dispersed in CORE than in the TB corpus. This pattern is particularly apparent in the CF_R10 and CF_R30 lists. While the key keyword method differed from the three CF methods in that it only identified words that are more dispersed in the TB corpus, like the three CF methods many of the words labelled as key by the key keyword method are dispersed quite widely in both corpora. With the exception of the few words that fell below the line in the CF plot, the plots from the CF and KK methods are quite similar. In contrast, the TD method identified only words that (*1*) are more widely dispersed in the TB corpus, and (*2*) have relatively low dispersion in CORE. However, it should also be recognised that some of the words in the TD list are the same as those from the CF and KK methods, hence the similar patterns in the bottom left corner of those plots.

Table 4 summarises the dispersion results from Figure 1 in a different form. The numbers in Column 2 represent the average proportion of texts in the TB corpus that contain the keywords from each of the keyness methods. Similarly, Column 3 contains the same information for the texts in the CORE corpus. In other words, the results in these two columns are averages of the values represented by the points in the plots in Figure 1, where Column 2 contains averages of the values on the *y*-axis and Column 3 contains averages of the *x*-axis values. It can be seen in Column 4 that the words identified by the text dispersion method are 329 percent more widely dispersed in travel blogs than in CORE, whereas the ratio for the other four methods ranges between 19 percent and 106 percent higher dispersion in travel blogs. The results from this table confirm the findings discussed in the preceding paragraphs: text dispersion keyness is much more effective at capturing words that are more dispersed in the target corpus relative to the reference corpus.

Based solely on the frequencies for tokens and types, and the dispersion rates, of the top 100 words identified by these methods, the text dispersion method identifies words that are much more strongly related to the target travel-blog corpus than the reference corpus. One could argue that it is a problem that the frequency and dispersion values are much lower for the text dispersion words than the other methods. We believe this is actually a strength of the text dispersion method since often the most frequent/dispersed words are general high-frequency words that are not distinctive to the target corpus. It is not actually these rates that we are concerned with. Rather, the most important consideration is *relative* frequency and *relative* dispersion. A single example should suffice to illustrate this point. The key keyword method identified *and* as one of the top 100 keywords for the TB corpus. There is no question that *and* is among the most highly frequent and widely dispersed words in the TB corpus. However, *and* is also highly frequent and widely dispersed not only in CORE, but in almost every register of English. Thus, it is a poor keyword because it is not highly frequent and widely dispersed relative to the reference corpus. Based on the results reported here, text dispersion keyness is the only method that consistently identifies words that meet these criteria.

At this point we take a closer look at the content-distinctiveness and content-generalisability of the keyword lists produced by the five keyness methods. As mentioned previously, we believe that keywords should be both generalisable (representative of a large proportion of the corpus) and distinctive (strongly associated with the target domain). Although there is considerable overlap in these two constructs, we believe they each contribute important and distinctive attributes. A word can be generalisable but not distinctive. For example, high-frequency function words (e.g., *and*, *the* and *for*) are often dispersed across the entire target corpus. However, they are not distinctive in that they are not strongly associated with the target domain in contrast to other discourse domains. On the other hand, a word can be distinctive but not generalisable. A place name (e.g., *Paphos* and *Madrid*) is easy to interpret as a travel-related word, but if it only occurs in one or two texts then it is not useful as a keyword because it is not generalisable to the entire domain of travel writing.

The purpose of this section is to explore the quality of keyword lists through an analysis of particular word classes that are strong indicators of content-generalisability and content-distinctiveness (see Section 2.3). After an extensive review of many keyword lists, we have identified ‘abbreviations’ and ‘proper nouns’ as two types of words that are almost always questionable in their content-generalisability. We have also identified two types of words – ‘function words’ and ‘high frequency verbs’ – as categories that tend to be questionable in their content-distinctiveness. The list of high frequency verbs used in this study include all forms of (*1*) three primary verbs (*be*, *have* and *do*) and (*2*) the top ten most frequent lexical verbs in English (*say*, *get*, *go*, *know*, *think*, *see*, *make*, *come*, *take* and *want*) (see Biber *et al.*, 1999: 110). We should emphasise that it is certainly not always true that abbreviations and proper nouns are non-generalisable and function words and high frequency verbs are non-distinctive. However, it does seem to be true in the vast majority of cases. The frequencies for words in each of these four categories can be seen in Table 5.

Nearly one-third of the words on the CF list are questionable in their generalisability or content-distinctiveness. While the dispersion criteria of the CF_R10 and CF_R30 methods eliminate most of the non-generalisable words, it also seems to lead to the introduction of many more words that are questionable in their content-distinctiveness. This trend is so extreme that more than three-fourths of the words in the list from the CF_R30 method

	Generalisable?		**Distinctive?**		TOTAL
	Abbreviations	*Proper nouns*	**Function words**	**High frequency verbs**	
CF	*2*	*7*	**15**	**6**	***30***
CF_R10	*0*	*0*	**23**	**9**	***32***
CF_R30	*1*	*0*	**56**	**20**	***77***
KK	*1*	*7*	**11**	**4**	***23***
TD	*0*	*0*	**1**	**0**	***1***

Table 5: Frequencies for word categories with questionable content-generalisability and content-distinctiveness.

are non-distinctive. The KK method performs better than the CF methods with its list containing only eight non-generalisable words and fifteen non-distinctive words. The text dispersion method performed the best by far, including only one word from the non-distinctive list and no words from the non-generalisable list.

The quantitative results reported in this section only tell part of the story. Hence, in the next section we qualitatively investigate the actual words included in the top 100 keyword lists for each of these five methods in order to compare their quality in terms of content-generalisability and content-distinctiveness.

3.2 Qualitative results

In this section we focus on using a qualitative approach to further explore the quantitative patterns described in the previous section. As mentioned above, the numbers in Table 5 should be interpreted with caution since it is possible for abbreviations and proper nouns to be generalisable, and for function words and high frequency verbs to be distinctive. A more informative approach is to qualitatively investigate each of the words in the five keyword lists to evaluate the extent to which they are generalisable and distinctive. That is the purpose of the remainder of this section.

We begin this qualitative investigation by introducing the top 100 keywords for the text dispersion keyness method (see Figure 2). This list will be used as a point of reference in subsequent comparisons with the other four keyness methods since it emerged as the method that produces the highest quality keyword lists, based on the quantitative data reported in the previous section. We can see that there are no words in this list that come from lexical categories deemed to be difficult to generalise. Only one of the words comes from the lexical categories that are often not distinctive (shown in boldface). This means that, based on this dataset, 99 percent of the words identified as ‘keywords’ by the text dispersion method are distinctive and content generalisable. Interestingly, a closer examination reveals that the only function word that was included in the top 100 words from the text

adventure
afternoon
airport
amazing
around
arrived
attractions
beach
beaches
beautiful
beer
biking
boat
booked
bus
city
cliffs
day
delicious
destination
dinner
enjoyed
explore
exploring
famous
ferry
flight
flights
gardens
guide
headed
hike
hiking
hills
holiday
hostel
hostels
hotel
hotels
hour
island
islands
journey
locals
located
lovely
lunch
mountain
mountains
museum
nearby
night
north
park
photo
photos
places
restaurant
restaurants
ride
river
road
rocks
scenery
scenic
sea
shops
sights
south
spectacular
steep
streets
stunning
sun
sunny
sunset
swimming
tour
tourist
tourists
tours
town
trail
trails
travel
travellers
travelling
trees
trip
village
villages
visit
visited
visiting
visitors
walk
walked
walking
water
weather

Figure 2: Top 100 text dispersion (TD) keywords.

dispersion method was *around*, which is actually an exceptional case in that it is quite easy to interpret as a travel-related word (e.g., *walk around* and *travel around*).

The vast majority of the words in this list appear to be strongly associated with travel blogs (e.g., *trip*, *travel*, *tour*, *visit* and *tourists*). This list also contains words that refer to modes of transportation (e.g., *bus*, *walk*, *boat* and *flight*), geographical features (e.g., *beach*, *island*, *river*, *mountain* and *sea*), activities and attractions for tourists (e.g., *park*, *museum*, *hiking*, *attractions*, *restaurants*, *swimming* and *exploring*), language for describing travel locations (e.g., *amazing*, *beautiful*, *scenic*, *stunning*, *sunny* and *spectacular*), and words related to food and dining (e.g., *beer*, *delicious*, *dinner* and *lunch*), all of which are clearly associated with the language of travel blogs.

In the remainder of this section, we evaluate the top 100 words produced by each of the other four keyness methods. In each case, this is done through a comparison with the top 100 words from the text dispersion keyness method. These comparisons are made based on the content-distinctiveness and content-generalisability of the words identified in these lists. In Tables 6 to 9, the far right column contains words that appear in the text dispersion keyword list, but not in the list in question. The far left column in these tables contains words that occur in the list in question, but not in the text dispersion keyword list. The middle column contains words that appear in both keyword lists.

CF only		Both		TD only	
a	metres	adventure	mountain	afternoon	photos
along	morning	airport	mountains	biking	restaurant
asia	**not**	amazing	museum	booked	ride
back	**or**	**around**	night	cliffs	rocks
be	**our**	arrived	park	delicious	scenic
campsite	*paphos*	attractions	places	enjoyed	shops
canyon	path	beach	restaurants	explore	sights
castle	place	beaches	river	exploring	spectacular
climb	rain	beautiful	road	ferry	streets
col	refuge	beer	scenery	flights	stunning
contiki	ridge	boat	sea	gardens	sun
de	route	bus	south	guide	sunny
had	**said**	city	steep	hike	sunset
has	tent	day	tour	hills	swimming
he	*thai*	destination	tourist	hostels	trails
his	*thailand*	dinner	tourists	hour	travellers
hrp	**that**	famous	tours	journey	trees
i	**up**	flight	town	located	village
krakow	valley	headed	trail	lovely	villages
la	**was**	hiking	travel	nearby	visiting
lake	**we**	holiday	travelling	north	visitors
madrid	**will**	hostel	trip	photo	weather
		hotel	visit		
		hotels	visited		
		island	walk		
		islands	walked		
		locals	walking		
		lunch	water		

Table 6: Comparison between lists of top 100 corpus frequency (CF) keywords and top 100 text dispersion (TD) keywords.

In Table 6, we can see that there is a 56 percent overlap between the top 100 words from the CF method and the CF_R10 method. In other words, 44 of the top 100 words from the CF method did not occur in the text dispersion keyword list and *vice versa*. Of the forty-four words that were included *only* in the CF list (far left column), more than two-thirds were labelled as non-generalisable (italicised) or non-distinctive (boldface). The list of non-distinctive words that were labelled as key by the CF keyness method include pronouns (*he*, *his*, *I*, *our* and *we*), other function words (e.g., *a*, *along*, *not*, *or* and *that*), and general high-frequency verbs (e.g., *had*, *has*, *said*, *was* and *will*). Although some of these words (e.g., *along*, *back* and *up*) are easier to associate with the discourse domain of travel blogs than others (e.g., *that*, *not* and *or*), all of them are quite frequent in most other registers of English, making it difficult to interpret them as being distinctive to travel blogs when compared with other web registers.

The ‘CF only’ list also contains nine words that are questionable in terms of their generalisability to the entire domain of travel blogs (italicised in Table 6). These words include proper nouns (*contiki*, *krakow*, *madrid*, *thai*, *asia*, *thailand* and *paphos*) and abbreviated forms (*HRP* and *col*). As mentioned above, by classifying these words as difficult-to-generalise we are not saying that these words are unrelated to travel. Rather, we are simply

CF_R10 only		Both		TD only	
a	**may**	adventure	night	afternoon	photo
along	morning	airport	north	attractions	photos
back	**my**	amazing	park	biking	restaurant
be	**not**	**around**	places	booked	restaurants
coast	**off**	arrived	ride	cliffs	rocks
days	**on**	beach	river	delicious	scenic
de	**or**	beaches	road	enjoyed	shops
down	**our**	beautiful	scenery	explore	spectacular
evening	place	beer	sea	exploring	steep
great	rain	boat	sights	ferry	streets
had	route	bus	south	flights	stunning
has	**said**	city	tour	gardens	sun
he	**that**	day	tourist	guide	sunny
here	train	destination	tourists	hike	sunset
him	**up**	dinner	tours	hiking	swimming
his	**was**	famous	town	hills	trails
i	**we**	flight	trail	hostel	travellers
if	**went**	headed	travel	hostels	villages
is	**were**	holiday	travelling	located	visiting
la	**will**	hotel	trees	lovely	visitors
lake		hotels	trip	nearby	
		hour	village		
		island	visit		
		islands	visited		
		journey	walk		
		locals	walked		
		lunch	walking		
		mountain	water		
		mountains	weather		
		museum			

Table 7: Comparison between lists of top 100 corpus frequency keywords, with minimum range or 10 percent of the corpus (CF_R10), and top 100 text dispersion (TD) keywords.

questioning whether these words are used in enough texts by enough different authors to give us confidence that they are representative of the domain of travel blogs, and not a mere artefact of the topic of only one or a few texts.

There was a 59 percent overlap between the text dispersion keyword list and the corpus frequency list with a minimum text dispersion requirement of 10 percent. Of the forty-one words in the 'CF_R10 only' list there are no words with questionable content-generalisability, but thirty-two words fall into the non-distinctive category. It seems that, while the range criteria tend to eliminate words that are difficult to generalise, it also tends to introduce words that are not distinctive. Among these are high frequency verbs (e.g., *went*, *may* and *is*) and function words (e.g., *if*, *on* and *off*).

The top 100 keyword lists for the CF list with a higher minimum range requirement of 30 percent and the text dispersion method had a very small overlap of only 10 percent. Of the ninety words in the 'CF_R30 only' list, seventy-six were labelled as non-distinctive. As mentioned above, there seems to be a strong relationship between the minimum range criteria and the number of non-distinctive words in the list. In essence, the addition of

CF_R30 only			Both	TD only		
a	**he**	**take**	**around**	adventure	holiday	shops
after	**here**	**that**	beautiful	afternoon	hostel	sights
along	**his**	**the**	city	airport	hostels	south
and	home	**their**	day	amazing	hotel	spectacular
any	**how**	**them**	night	arrived	hotels	steep
are	**i**	**there**	road	attractions	hour	streets
area	**if**	**these**	town	beach	island	stunning
as	**into**	**they**	travel	beaches	islands	sun
at	**is**	**think**	trip	beer	journey	sunny
away	**know**	**this**	visit	biking	locals	sunset
back	little	**those**		boat	located	swimming
be	looking	**through**		booked	lovely	tour
because	**make**	time		bus	lunch	tourist
before	**me**	**took**		cliffs	mountain	tourists
being	**my**	**top**		delicious	mountains	tours
best	**next**	**up**		destination	museum	trail
bit	**no**	**us**		dinner	nearby	trails
days	**not**	**very**		enjoyed	north	travellers
do	**off**	**was**		explore	park	travelling
down	old	**way**		exploring	photo	trees
few	**on**	**we**		famous	photos	village
found	**or**	**went**		ferry	places	villages
from	**other**	**were**		flight	restaurant	visited
go	**our**	**what**		flights	restaurants	visiting
got	**out**	**where**		gardens	ride	visitors
great	place	**who**		guide	river	walk
h	**see**	**will**		headed	rocks	walked
had	small	**would**		hike	scenery	walking
has	**so**	years		hiking	scenic	water
have	**some**	**your**		hills	sea	weather

Table 8: Comparison between lists of top 100 corpus frequency keywords, with minimum range or 30 percent of the corpus (CF_R30), and top 100 text dispersion (TD) keywords.

dispersion criteria seems to compound some of the problems of the corpus frequency method. Clearly, it is not sufficient to enforce a simple dispersion range criteria if the goal is to increase both the content-generalisability and content-distinctiveness of keywords since words that are non-distinct tend to be widely dispersed in most corpora.

A comparison of the words that occurred on the key keyword list and those on the keyword list produced by the text dispersion method revealed a 52 percent overlap. Although this overlap is slightly lower than the overlap between the CF_R10 and TD methods, the KK method produced a higher quality keyword list than all of the CF methods, with only eight words with questionable content-generalisability and fifteen words with questionable content-distinctiveness. Seven of the eight non-generalisable words were proper nouns (e.g., *Australia*, *Bangkok*, *London* and *Santiago*). Most of the non-distinctive words were function words, including several personal pronouns (e.g., *I*, *my*, *our*, *us*, *we* and *you*). The results in Table 9 confirm the quantitative findings from Section 3.1, showing that the key keyword method is much more effective than the corpus frequency approach, regardless of the minimum range criteria. However, the KK method included many more

KK only		Both		TD only	
#	guided	adventure	museum	afternoon	photo
accommodation	**had**	airport	night	amazing	photos
and	**i**	**around**	north	arrived	ride
australia	**la**	beach	park	attractions	scenery
bangkok	lake	beautiful	places	beaches	scenic
bar	*london*	beer	restaurant	biking	shops
bay	**my**	boat	restaurants	booked	sights
bike	**our**	bus	river	cliffs	spectacular
bikes	plenty	city	road	delicious	steep
bridge	rain	day	rocks	destination	streets
camp	ridge	dinner	sea	enjoyed	stunning
campsite	rock	famous	south	explore	sun
canyon	route	flight	tour	exploring	sunny
castle	*santiago*	flights	tourist	ferry	sunset
cathedral	*spanish*	headed	tourists	gardens	swimming
climbing	*sydney*	hike	tours	guide	trails
crossing	*thailand*	hiking	town	hills	travellers
cruise	**the**	holiday	trail	hostels	travelling
de	train	hostel	travel	journey	village
evening	**us**	hotel	trees	locals	villages
festival	**was**	hotels	trip	located	visited
fish	**we**	hour	visit	lovely	visiting
got	**were**	island	walk	lunch	visitors
great	**you**	islands	walking	nearby	walked
		mountain	water		
		mountains	weather		

Table 9: Comparison between lists of top 100 key keywords (KK) and top 100 text dispersion (TD) keywords.

words with questionable content-distinctiveness and content-generalisability than the TD method.

Using the metrics of content-generalisability and content-distinctiveness, the text dispersion measure performs remarkably well, effectively excluding all of the non-generalisable words and all but one word with questionable content-distinctiveness. It seems that the quality of corpus frequency keyword lists declines as the minimum range requirement increases. This is probably related to the fact that very few words occur in a large proportion of the texts in any corpus, and these words are typically function words and high frequency words. Next to the text dispersion method the key keyword method performed best, with 77 percent of the words being classified as distinctive and content generalisable.

4. Conclusion

This study has the following three goals, to: (*1*) establish the importance of text dispersion in keyword analysis, (*2*) introduce text dispersion keyness, and (*3*) compare this new measure to four keyness measures that have been used in previous research. With regard to the first goal, we have attempted to show that disregarding text dispersion and only accounting for corpus frequency in keyword analysis can produce words that are not distinctive and/or not generalisable to the content of the entire corpus. Moreover, measuring keyness at the level of the text, rather than the level of the corpus, has many advantages. Texts are linguistically valid in that they are self-contained units of naturally occurring discourse (see, for example, Egbert and Schnur, 2018). A corpus, on the other hand, is a contrived unit that does not represent the way natural language is actually organised and produced. In addition to the logical arguments in favour of accounting for text dispersion in corpus research in general, we believe that the empirical results presented in the previous section offer strong evidence in support of measuring keyness using text dispersion rather than corpus frequency.

In order to address the third goal of the study, we quantitatively and qualitatively evaluated word lists produced by five keyness measures in terms of their content-distinctiveness and content-generalisability. All of the analyses we carried out showed that text dispersion keyness – the new method introduced in this study – outperformed the other four keyness methods. The frequency and dispersion rates of the top 100 text dispersion keywords were four-to-five times higher in the target corpus when compared with the reference corpus. In contrast, the relative frequency and dispersion rates for the other four methods was only one-to-two times higher in the target corpus than the reference corpus. Based on the results of this study, text dispersion keyness is also more likely to produce words that are generalisable and distinctive. Ninety-nine percent of the text dispersion keywords satisfied both of those criteria. In contrast, between 23 percent and 69 percent of the words produced by the other four methods were questionable in their content-generalisability or content-distinctiveness.

Somewhat surprisingly, the two corpus frequency measures that account for dispersion in the form of a minimum text range (CF_R10, CF_R30) performed quite poorly on all of the metrics accounted for in this study. In fact, the enforcement of a minimum dispersion rate actually systematically decreased (*1*) the target corpus to reference corpus ratios for frequency and dispersion, and (*2*) the number of distinctive and content-generalisable words. This suggests that there are fundamental problems with the corpus frequency approach that cannot be remedied with simple dispersion criteria. These problems seem to stem from the fact that the statistical procedure accounts only for frequency. This results in a list that is biased in favour of words that are statistically higher in frequency but not necessarily more meaningful and distinctive for the target corpus as a whole. Because these high-frequency words typically occur in many, if not most, of the texts in both corpora, the minimum dispersion does not actually seem to improve the list at all. In fact, as we saw in this study, it can actually result in a poorer keyword list.

Next to text dispersion keyness, the method that performed best was key keyword analysis. The list of words produced by the key keyword method had higher target corpus to reference corpus ratios for both word frequency (types and tokens) and dispersion than the three corpus frequency methods. Furthermore, the key keyword list contained fewer non-distinctive and non-generalisable words than the three corpus frequency methods. This shows that key keyword analysis is indeed a substantial improvement upon traditional keyness methods.

Based on this finding, the reader may wonder whether they should simply use key keyword analysis as opposed to adopting text dispersion keyness for their research. We believe that there are several reasons that discourse analysts and corpus researchers should use text dispersion keyness when performing keyword analysis on their corpora. First, it performs better at identifying high-quality keywords, while excluding words that lack the qualities of content-distinctiveness to the target discourse domain and generalisability to the content of the texts in the corpus. Second, text dispersion keyness is an elegant, parsimonious computational solution to the problem of identifying keywords. Text dispersion requires less data preparation and fewer methodological steps than key keyword analysis. Finally, text dispersion keyness produces keyword lists that are very easy to interpret. Simply put, text dispersion keyness identifies words that are used in a larger proportion of the texts in the target corpus than in the reference corpus. This allows even novice researchers and readers to understand the construct of keyness and its operationalisation.

The results of this study offer strong support for the claims we make in the introduction about the importance of texts in corpus analysis. We believe that, for many applications, text dispersion keyness is a better option than corpus frequency keyness and key keyword analysis. However, it is important to mention a few potential limitations of text dispersion keyness. First, as discussed throughout this paper, text dispersion keyness does not account for token frequency in any way. Thus, it would not be an option if a researcher's goal is to make specific claims about word frequencies in a corpus. Second, in order to perform text dispersion keyness it is necessary for corpora to be collected and organised in the form of texts. This seems to be the standard practice for most corpora, but there are cases where this may be difficult. Third, it is currently easier to produce corpus frequency keyword lists because this method is built into existing concordancing software programs (e.g., AntConc, WordSmith and MonoConc). Text dispersion keyness, on the other hand, is not built into these programs.

Finally, as we demonstrated throughout Section 3, text dispersion keyness is heavily biased in favour of including content words and excluding grammatical (i.e., function) words in keyword lists. In our view, this is ideal, rendering keyword analysis and grammatical analysis as separate enterprises, and focussing keyword analysis on the content of the target discourse domain, rather than grammatical characteristics associated with the register represented by that domain.

We thus believe that the identification of content generalisable words takes keyword analysis back to its roots, when researchers like Firth and Williams focussed on identifying 'words that they believed embodied important concepts that reflected societal or cultural concerns' (Baker, 2004:

346; see also Firth, 1957; and Williams, 1983). However, this approach would not be ideal for research that is focussed on identifying grammatical features through keyword analysis (e.g., Culpeper, 2014). For example, researchers who are attempting to use keyword analysis as a replacement for Multi-Dimensional analysis (e.g., Xiao and McEnery, 2005; and McEnery *et al.*, 2006) would not find text dispersion keyness to be particularly useful.

It is our hope that this study will raise awareness of the importance of text dispersion in corpus linguistics and discourse analysis. More importantly, we hope to see a trend in these fields in the direction of using the text – rather than the corpus – as the primary unit of analysis. While this study has focussed on keyword analysis, we believe that there are many benefits of focussing on texts instead of corpora in all areas of corpus linguistics and CADS.

References

Adolphs, S., B. Brown, R. Carter, P. Crawford and O. Sahota. 2004. 'Applying corpus linguistics in a health care context', Journal of Applied Linguistics 1 (1), pp. 9–28.

Baker, P. 2004. 'Querying keywords: questions in difference, frequency, and sense in keyword analysis', Journal of English Linguistics 32 (4), pp. 346–59.

Baker, P. 2010. 'Corpus methods in linguistics' in L. Litosseliti (ed.) Research Methods in Linguistics, pp. 95–113. New York: Continuum.

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyzanowsky, T. McEnery and R. Wodak. 2004. 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', Discourse and Society 19 (3), pp. 273–306.

Baker, P., A. Hardie and T. McEnery. 2013. A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press.

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

Biber, D. and J. Egbert. 2018. Register Variation Online. Cambridge: Cambridge University Press.

Biber, D., J. Egbert and M. Davies. 2015. 'Exploring the composition of the searchable web: a corpus-based taxonomy of web registers', Corpora 10 (1), pp. 11–45.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. Longman Grammar of Written and Spoken English. Harlow, England: Longman.

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), pp. 29–59.

Culpeper, J. 2014. 'Developing keyness and characterization' in D.L. Hoover, J. Culpeper and K. O'Halloran (eds) Digital Literary Studies: Corpus Approaches to Poetry, Prose and Drama, pp. 35–63. New York: Routledge.

Dunning, T. 1993. 'Accurate methods for the statistics of surprise and coincidence', Computational Linguistics 19 (1), pp. 61–74.

Egbert, J. and E. Schnur. 2018. 'The role of the text in corpus and discourse analysis: Missing the trees for the forest' in C. Taylor and A. Marchi (eds) Corpus Approaches to Discourse: A Critical Review. New York: Routledge.

Egbert, J., D. Biber and M. Davies. 2015. 'Developing a bottom-up, user-based method of web register classification', Journal of the Association for Information Science and Technology.

Firth, J.R. 1957. Papers in Linguistics 1934–1951. London: Oxford University Press.

Gabrielatos, C. and P. Baker. 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), pp. 5–38.

Gabrielatos, C. and A. Marchi. 2012. Keyness: Appropriate metrics and practical issues. Paper presented at Critical Approaches to Discourse Studies 2012, Bologna, 14 September 2012. Accessed 27 March 2017 at: http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf.

Johnson and Ensslin 2006. 'Language in the news: some reflections on keyword analysis using WordSmith Tools and the BNC', Leeds Working Papers in Linguistics and Phonetics 11, pp. 96–109. Accessed March 2017 at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.9338&rep=rep1&type=pdf.

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

Kilgarriff, A. 2009. 'Simple maths for keywords' in M. Mahlberg, V. González-Díaz and C. Smith (eds) Proceedings of Corpus Linguistics Conference CL2009. 20–23 July 2009. University of Liverpool, UK.

Kilgarriff, A. 2012. 'Getting to know your corpus' in P. Sojka, A. Horák, I. Kopeček and K. Pala (eds) Text, Speech and Dialogue, pp. 3–15. Berlin and Heidelberg: Springer.

McEnery, T., R. Xiao and Y. Tono. 2006. Corpus-based Language Studies: An Advanced Resource Book. Taylor and Francis.

Millar, N. and B. Budgell. 2008. 'The language of public health – a corpus-based analysis', Journal of Public Health 16 (5), pp. 369–74.

Paquot, M. and Bestgen, Y. 2009. 'Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction' in A. Jucker, D. Schreier and M. Hundt (eds) Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29), pp. 247–69. Amsterdam: Rodopi.

Rayson, P. and R. Garside. 2000. 'Comparing corpora using frequency profiling' in proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), pp. 1–6. 1–8 October 2000. Hong Kong.

Römer, U. and S. Wulff. 2008. 'Applying corpus methods to written academic texts: explorations of MICUSP', Journal of Writing Research 2 (2), pp. 99–127.

Savický, P. and J. Hlaváčová. 2002. 'Measures of word commonness', Journal of Quantitative Linguistics 9, pp. 215–31.

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

Scott, M. and C. Tribble. 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. (Volume 22.) Amsterdam: John Benjamins.

Williams, R. 1983. Keywords. London: Fontana.

Xiao, R. and T. McEnery. 2005. 'Two approaches to genre analysis: three genres in modern American English', Journal of English Linguistics 33 (1), pp. 62–82.