rEFLections
Vol 32, No 3, September - December 2025

Comparing Frequency and Dispersion Keywords: Effects of Variations in Target and Reference Corpora

PUNJAPORN POJANAPUNYA
School of Liberal Arts, King Mongkut's University of Technology Thonburi, Thailand
Author email: punjaporn.poj@mail.kmutt.ac.th

Article information

Article history:
Received: 3 Oct 2024
Accepted: 25 Sep 2025
Available online: 6 Oct 2025

Keywords:
Keyword analysis
Dispersion keyword analysis
Dispersion keyness
Frequency keyness
Small corpus

Abstract

Dispersion keyword analysis, which identifies words that occur in significantly more texts in the target corpus than in the reference corpus, has recently been introduced as a more effective method than traditional frequency keyword analysis. Previous research has used this method to identify keywords within a target corpus, usually consisting of hundreds of texts, and used a much larger corpus as a reference. However, questions remain regarding its applicability for cases involving fewer texts and comparisons between smaller specific corpora. This study compares the top 100 frequency keywords and dispersion keywords identified under several conditions, which varied in terms of the number of texts in the target corpus (24, 100, and 200 texts) and the types of reference corpora used. Both methods identified unique and shared keywords; however, frequency keywords are found more frequent and widely dispersed not only within the target corpus but also in the reference corpus compared to dispersion ones, which are notably more relevant to the target corpus. The selection between frequency and dispersion methods and the relevance of frequency and dispersion keywords in research with differing focuses are discussed.

INTRODUCTION

Keyword analysis, popularized by Scott (1997) and recognized as a robust method in corpus linguistics, has been widely applied in various disciplines, including the humanities and social sciences (Bancroft-Billings, 2020; Li & Lu, 2020), healthcare communication (Bailey, 2018; Ji & Li, 2024), and environmental science (Lam et al., 2019). A key reason for its broad adoption is that it provides a quantitative means of analyzing qualitative data, by identifying keywords that indicate either the aboutness or stylistic features of a text (Scott & Tribble, 2006). Aboutness keywords, which typically consist of open-class words, reflect the content or main topics of a corpus. They help reveal what a corpus is about, offering insights into its subject matter and highlighting meaningful characteristics within a specific discourse domain (Egbert & Biber, 2019; Egbert et al., 2020). Style keywords, on the other hand, usually consist of closed-class

This paper was specially selected to be published from the 5th Doing Research in Applied Linguistics (DRAL 5) International Conference that was run from 5 to 7 September 2024 in Bangkok, Thailand.

words and do not convey specific content; rather, they indicate stylistic choices. These keywords reveal register characteristics related to the interpersonal and structural aspects of a corpus (Culpeper, 2009). Such keywords can signal the author's stance, level of formality, or engagement with the audience.

Keyword analysis is an automated process that identifies keywords in a target corpus by comparing it with a reference corpus. The reference corpus may be a general corpus, which includes texts from various genres that represent general language use, or a specific corpus composed of texts from a particular genre. Stages for keyword analysis include 1) uploading a target corpus, which is purposefully collected by the researcher according to their research purpose, and a reference corpus into a corpus analysis program, 2) generating a list of all words in the target corpus, ranked by a keyness statistic from highest to lowest, which suggests the degree of importance of words in the target corpus compared to the reference corpus, and 3) choosing some high-ranked words, considered keywords for further interpretive process to address their research questions (see also Rayson, 2013). One of the key decisions during these stages is selecting an appropriate parameter, such as frequency or dispersion, to serve as the basis for comparing two corpora and thus affect the keywords identified. This choice is the main focus of the present study.

Frequency keyword analysis (F-KWA) identifies keywords based on frequency of words in a target and in a reference corpus. Keywords, in this context, are words that appear with significantly higher frequency or overused in a target corpus as compared to a reference corpus (Egbert et al., 2020; Gries, 2016 cited in Gries, 2021). It has been criticized for not considering the dispersion of words across various texts in the corpus, treating the corpus as a single unit of observation. By this method, words can be identified as key if they appear frequently within a few texts or even in a single text, even if they are not widely dispersed across the corpus, which may make them less typical. Therefore, these words may not truly reflect or represent overall characteristics within the discourse domain represented by the target corpus (Egbert & Biber, 2019; Egbert et al., 2020).

Dispersion keyword analysis (D-KWA) has been proposed by Egbert and Biber (2019) to address the limitation of the F-KWA. Dispersion refers to how dispersed a word is across the texts throughout a corpus (Egbert & Burch, 2023; Sönning, 2022a). D-KWA highlights the distribution of words across texts in a corpus (i.e., the number of texts in which words appear in both the target and reference corpora rather than their overall frequency). Instead of comparing frequencies, D-KWA compares the number of texts where the words appeared (the range) in the two corpora. So, dispersion keywords (D-KWs) are words which appear in significantly more texts, greater ranges in a target corpus than in a reference corpus. To enable the calculation of dispersion, the primary requirement is that both the target and reference corpora must be collected as several texts, treating each text as its observational unit. Keywords identified through this method are argued to be more effective as they capture broader themes rather than specific contexts (Egbert & Biber, 2019; Egbert, et al., 2020).

The primary distinction between the two approaches lies in the comparative parameter used for keyword identification: F-KWA compares frequency, while D-KWA compares the range, i.e., the number of texts within a corpus in which a word appears.

When comparing D-KWA with F-KWA, the F-KWA is widely recognized and has greatly influenced research in many disciplines. It is commonly conducted using various keyword analysis tools such as AntConc (Anthony, 2023), WordSmith Tools (Scott, 2024), Wmatrix (Rayson, 2009), Sketch Engine (Lexical Computing, n.d.), and LancsBox (Brezina et al., 2021). The application of D-KWA remains limited. Although its use has increased recently since Egbert and Biber introduced it in 2019, it is primarily investigated in methodology-focused studies that evaluate keyness approaches within corpus linguistics among specialists who have developed their own specialized programs for their research (such as, Python in Xu & Jhang, 2020).

Although previous research using D-KWA is less extensive than that using F-KWA, its growing recognition and application will make the approach more accessible to a wider range of analysts. Exploring the use of D-KWA in comparison to F-KWA will provide useful insights for analysts working with qualitative data, helping them make keyword analysis more beneficial for their research.

Methodological choices of frequency keyword analysis and dispersion keyword analysis

The concepts and methodological choices of frequency and dispersion keyword analysis are summarized in Table 1.

Table 1

Overview of F-KWA and D-KWA

No	Methodological choice	F-KWA	D-KWA
1	Approach to keywords	Frequency	Dispersion
2	Definition	Words that are statistically more frequent in a target corpus than in a reference corpus (Egbert et al., 2020)	Words which appear with significantly greater ranges in a target corpus than in a reference corpus (Egbert & Biber, 2019)
3	Target and benckmark corpus: text files	Multiple text files treated as a single unit of observation	Multiple texts, each treated as a separate unit of observation
4	Target and benckmark corpus: characteristics	Directed by research purposes (e.g., similar-genre corpus, general corpus)	
5	Keyness statistics	For example, log-likelihood (LL), odds ratio	For example, log-likelihood, key keywords, binomial regression
6	Thresholds for identifying keywords	Directed by research purposes	

Frequency keyword analysis

In F-KWA, the characteristics of the target and reference corpora (no. 4 in Table 1) largely depend on the research purposes. There are several keyness statistics to choose from (no. 5), including chi-square (Scott, 1997), log-likelihood (Rayson & Garside, 2000), log ratio, simple frequency difference (Gabrielatos & Marchi, 2012), odds ratio (Pojanapunya & Watson Todd, 2018), and simple maths (Kilgarriff, 2009), as cited in Egbert and Biber (2019), with log-likelihood (LL) being the most commonly used. After generating a list of words ranked by keyness statistic,

analysts usually set a threshold (no. 6) to identify keywords, such as consider 100 words ranked by keyness as keywords. These factors have an effect on the keyword outputs. In this study, apart from the characteristics of the target and reference corpora, keyness statistics, and thresholds (no. 5 and 6) were controlled when generating F-KWs and D-KWs for analysis. (Methodological choices are described in the methods section.)

Dispersion keyword analysis

The procedure for generating F-KWs and D-KWs is the same. Instead of comparing frequency of words, the D-KWA compares the number of texts that a word occurs in texts across each of the corpora. Similar to the F-KWA, the characteristics of a target corpus and a reference corpus, keyness statistics, and the thresholds (Choices 4-6 in Table 1) depend on the research purposes and the analyst's decisions, which are further detailed in the methods section.

Egbert and Biber (2019)'s dispersion approach prioritizes the existence of individual texts within a corpus, rather than treating the entire corpus as a single unit of observation. This approach more closely aligns with the natural occurrence of texts as "a text is a valid unit of language production, but a corpus is not" (Egbert et al., 2020, p. 30). They evaluated dispersion keyness by comparing it with traditional frequency keyness using a corpus of online travel blogs from the Corpus of Online Register of English (CORE) (371 texts, 330,918 tokens) and all documents in CORE except the travel blogs (48,200 files, 52 million words). They encourage discourse and corpus analysts to use D-KWA, which is more effective at identifying high-quality keywords. Their study shows that the top 100 F-KWs are more frequent and more dispersed across a corpus than the D-KWs. However, D-KWs perform better when evaluated in terms of relative frequency and relative dispersion rates, showing that D-KWs are more frequent and more widely dispersed in the target corpus compared to the reference corpus (Egbert et al., 2020). They argued that D-KWs are often more meaningful, interpretable, and relevant to the target corpus's discourse domain in terms of content-distinctiveness (strongly associated with the target domain) and content-generalisability (representative of a large proportion of texts in the corpus). Given these reasons for using D-KWA, they also noted that analysts should be aware that it is not suitable if the corpus is not collected as separate texts, and more importantly, if the research aims to make specific claims about word frequencies, as this approach does not initially account for word frequency.

It is worth noting that dispersion has long been considered a key criterion for keyword identification to ensure that keywords are both frequent and distributed across multiple texts. However, earlier studies typically applied dispersion as a post hoc filter after extracting frequency-based keywords and comparing corpora. For example, Clarke et al. (2022) used the F-KWA approach to generate keywords and then applied range as a criterion to identify the actual keywords by excluding those occurring in fewer than 5% of the texts in the target corpus.

Current state of dispersion keyword analysis and its applications

While the concept of dispersion is not new within corpus linguistics and keyword analysis, its application among discourse analysts and researchers outside corpus linguistics remains in its early stages.

Apart from Egbert and Biber's dispersion, other approaches to dispersion include, for example, key keywords—words that are identified as key in several texts within a corpus (Baker et al., 2013; Scott, 1997) and Binomial regression (Sönning, 2022a; Sönning, 2022b). A two-dimensional approach called Kullback-Leibler Divergence (Gries, 2021; Langenhorst et al., 2023), which integrates both the frequency and dispersion of a word over a corpus into keyness computations has also been introduced. These methods have been discussed among specialists, with several quantitative methods evaluated in methodology-oriented papers. However, application of these methods remains limited. This is partly due to the complexity and requirements of specialized programs, which are not often accessible to general users of keyword analysis. For example, Egbert and Biber (2019) developed a specifically designed Python program for dispersion. Although WordSmith Tools (Scott, 2024), Wmatrix (Rayson, 2009), Sketch Engine (Lexical Computing, n.d.), and LancsBox (Brezina et al., 2021), provide information about the dispersion of words, these tools typically present dispersion as descriptive data for keywords beyond frequency. However, none of these tools offered functions that use dispersion metrics as the primary parameter for identifying keywords during the period in which this research was conducted.

Previous methodology-focused papers typically use target corpora comprising several hundred to several thousand texts. For example, Egbert and Biber (2019) used a target corpus of 371 texts and a reference corpus of 48,200 texts. This raises questions about the effectiveness of the method when the corpora contain far fewer texts. The small number of texts in a corpus can influence statistical calculations and, consequently, the identification of keywords. For example, in a target corpus of 20 texts totaling 20,000 words, the frequency parameter for F-KWA can reach high values depending on the corpus size, providing a wide numerical range for comparison. In contrast, the range parameter for D-KWA is much narrower, limited to values between 0 and 20. As a result, comparisons within the larger range naturally show greater absolute differences than those within the smaller range. Given the small number of texts in the corpus, questions arise regarding whether F-KWA and D-KWA can still generate keywords that are meaningful for interpretation, and how their results may be similar or different in these conditions. Furthermore, comparisons conducted in previous research typically involve comparing specific corpora (a corpus representing a text collected from a specific genre) with general corpora (corpora consisting of texts from a variety of genres collected to represent general English). For example, Egbert & Biber (2019) used a 371 text component of CORE, Xu & Jhang (2020) compared frequency and dispersion keywords using a target corpus of English charter parties (CEC), consisting of 156 texts, and used the BNC Baby (3 million words) as the reference corpus.

The dispersion approach has been applied in a few application papers. Several papers which have cited the text dispersion keyness of Egbert and Biber did not apply their D-KWA directly. Instead, many of them used F-KWA and then measured dispersion of F-KWs as an additional step, such as by setting a minimum range for the number of texts in which the keywords should occur across a corpus (Baker, 2004; 2010 as cited in Egbert & Biber, 2019) (e.g., keywords should appear in 1% (Dayter & Messerli, 2022), 5% (Clarke et al., 2022), or 30% (Millar & Budgell, 2008) of texts in a corpus). However, there is no consensus regarding the ideal proportion to use as a threshold (Gries, 2021).

Scope of this study

This study focuses on the application of D-KWA by comparing it with F-KWA. Comparable conditions for producing F-KWs and D-KWs were designed. Since this study followed Egbert and Biber (2019)'s dispersion keyness, both F-KWs and D-KWs were produced based on the LL statistic, enabling comparable settings. LL is available on AntConc 4.2.4 (Anthony, 2023) for both F-KWA and D-KWA, ensuring accessibility to a wide range of users.

Purposes of the study

This research has investigated the quantitative characteristics of D-KWs generated under various conditions and compared them with F-KWs. These conditions varied in terms of 1) the number of texts (24, 100, and 200), and 2) comparison types (specific vs. specific and specific vs. general corpora). The findings of this study will provide input for researchers who are new to dispersion keyness and are uncertain about whether to use frequency or dispersion as a basis for their research, particularly regarding the applicability and effectiveness of dispersion keyness for small corpora with a few texts.

Research questions

This study aims to address the following questions:

1. Is the dispersion approach effective in the given context?
    a. A small corpus with a limited number of texts
    b. A specific corpus serving as a reference corpus
2. How relevant are frequency and dispersion keywords to the target corpus?
3. What are the differences and similarities between frequency keywords and dispersion keywords?

To answer questions 1 and 2, effectiveness and relevance are measured in terms of content distinctiveness and content generalizability, while question 3, which focuses on the differences and similarities of keywords from the two approaches, is measured in terms of rank difference, as well as shared and unique keywords.

METHODS

To investigate and compare traditional F-KWA and the more recently introduced D-KWA, this study set up conditions for generating lists of two types of keywords for analysis. The conditions vary in terms of 1) the number of text files in a target corpus (TC) and 2) types of comparisons for generating keywords.

Regarding types of comparisons, there are two major categories. First, a comparison between a specific corpus and a general corpus, which is the common type found in the current research using D-KWA. The second type is a comparison between a specific corpus and another specific

corpus. Since D-KWA has not been widely used yet, research compared specific corpora against each other is still limited.

Criteria for choosing sample specific corpora

Sample corpora are required to investigate keywords under various conditions. Main criteria for choosing the corpora are, firstly, the corpora should consist of multiple texts in a corpus, with a sufficient number to measure dispersion. Second, the corpora should narrate a clear story and quite obvious to recognize what words likely to be keywords. This helps determine the quality of keywords.

Conditions and variations

The conditions under investigation vary in terms of the number of texts and the types of corpus comparisons. Target corpora (TC) include a corpus of 24 chapters of The Wonderful Wizard of Oz (Oz) e-book, available on Project Gutenberg (https://www.gutenberg.org/), a corpus of 100 research article abstracts in applied linguistics (AL), and a corpus of 200 abstracts in science areas (SCI).

Reference corpora (RC) are general and specific corpora. General corpora (Gen) used in the study are AmE06 (a general corpus of American English) and BE06 (a general corpus of British English). Both corpora are one million word corpora consisting of 500 text files available for use in AntConc 4.2.4 (Anthony, 2023). Second, the same-genre specific corpora (GS) are a corpus of 17 chapters of Peter Pan e-book, available on Project Gutenberg, the AmE06-Fiction-General which is sub-corpus of fiction from AM06, and SCI.

The description and details of the corpora used in this study, including both the TC and RC, are presented in Table 2. Oz and AL served exclusively as the TC, while PP, Fic, AmE, and BE were used solely as the RC. The SCI corpus was used as either TC or RC, depending on the specific case.

Table 2

Corpora

No.	Corpus	Type	#types	#tokens	#files
1	The Wonderful Wizard of Oz (Oz)	TC	2,871	39,619	24
2	Research article abstracts in applied linguistics (AL)	TC	3,173	18,079	100
3	Research article abstracts in sciences (SCI)	TC/RC	7,199	45,021	200
4	Peter Pan	RC	4,796	47,878	17
5	AmE06-Fiction-General (Fic)	RC	8,335	59,568	29
6	AmE06	RC	44,434	1,017,879	500
7	BE06	RC	43,436	1,007,532	500

To allow analysis of keywords across conditions varying in the number of texts in a target corpus and types of a reference corpus (either general or specific), six comparisons which represent various settings were conducted (see Table 3).

Table 3

Details of conditions of comparisons

No.	TC	#texts in TC	RC	Types of RC	#texts in RC	Types of comparisons
1	Oz	24	Peter pan	GS	17	Specific vs. Specific
2			Fic	GS	29	Specific vs. Specific
3			AmE06	Gen	500	Specific vs. General
4	AL	100	SCI	GS	200	Specific vs. Specific
5			BE06	Gen	500	Specific vs. General
6	SCI	200	BE06	Gen	500	Specific vs. General

Generating keyword lists

F-KWs and D-KWs for the analysis were generated using AntConc 4.2.4 (Anthony, 2023). There are 'Likelihood Measure' options either based on frequency or dispersion on the Tool Settings menu. While F-KWA uses tokens, the D-KWA uses the number of text files for calculating LL. The top 100 words ranked by LL as the keyness statistic were selected as sample keywords from six conditions and compared.

Data analysis

The top 100 keywords from both approaches and all six conditions were analyzed and compared in terms of content distinctiveness, content generalizability, rank differences, and shared and unique keywords.

Content-distinctiveness and content-generalisability

According to Egbert and Biber (2019), content distinctiveness and generalizability indicate the effectiveness of keywords, specifically regarding how well these keywords reflect the content of a target corpus. Distinctiveness indicates the strength of the relationship between a keyword and the content of the discourse domain represented by the target corpus, in contrast to other discourse domain represented by the reference corpus. Generalizability refers to the degree to which a keyword represents the content across the full range of texts in the target corpus. Content distinctiveness and content generalizability are measured using frequency and frequency ratios, dispersion and dispersion ratios, the number of function words, the number of proper nouns, and the number of abbreviations.

Frequency and frequency ratio

Frequency and frequency ratio are indicators of content-distinctiveness. Of each list, normalised frequencies of the top 100 keywords generated by the F- and D-keyness approaches were calculated for both in the target corpus and in the reference corpus. Then, the means of normalised frequencies were calculated. The mean of normalized frequencies provides a rough explanation of how common or uncommon the top 100 keywords are across the methods. In addition, this value serves as a metric to compare the commonness and specificity of keywords identified by the two methods. A ratio of average frequency of the top 100 keywords in TC and RC was also calculated to give the relative appearance of these keywords in the TC compared to the RC.

Frequencies are normalized, taking into account the number of tokens per file and the order of magnitude of the corpora in comparisons. The values for normalization are presented below.

Oz = 39,619 tokens, 24 files (1,650 tokens/file, normalised per 2,000)
Peter Pan = 47,878 tokens, 17 files (2,800 tokens/file, normalise per 2,000)
AL = 18,079 tokens, 100 files (180/file, normalise per 200)
SCI = 45,021 tokens, 200 files (225/file, normalise per 200)
Fic = 59,568 tokens, 29 files (2,054/file, normalised per 2,000)
AmE06 = 1,017,879 tokens, 500 files (normalised per 1M)
BE06 = 1,007,532 tokens, 500 files (normalised per 1M)

Dispersion and dispersion ratio

Dispersion and dispersion ratio are used to assess both content-generalisability and content-distinctiveness. The number of texts in the corpus in which each of the top 100 words appears in TC and RC was calculated as a percentage. This value indicates whether the keywords generated from each method are words that are dispersed throughout the corpus or being specific to a small number of texts. To compare the dispersion of keywords in the target corpus and the reference corpus, the dispersion ratio—the ratio of the range of keywords in TC to RC—was calculated.

The number of function words, proper nouns, and abbreviations

The number of keywords that are typically non-distinctive or non-generalizable was identified and compared. Following Egbert and Biber (2019) and Xu and Jhang (2020), function words are considered non-distinctive since they tend to be highly frequent and widely dispersed across all discourse domains, while proper nouns and abbreviations were typically non-generalizable. The frequency of these words indicates the extent to which the keywords are relevant to the target corpus.

Rank difference

The difference in rank order of words when ranked based on F-keyness and D-keyness was measured. The analysis indicates whether the top F-KWs remain among the top D-KWs and vice versa. Table 4 shows examples for comparing ranks of F-KWs and D-KWs when they appear in the other list.

Table 4

Examples of calculation of rank difference

Frequency keywords			Dispersion keywords		
Rank F-KWs	**Rank D-KWs**	**Difference**	**Rank D-KWs**	**Rank F-KWs**	**Difference**
1	1	0	1	1	0
2	3	1	2	9	7
3	5	2	3	2	1
4	22	18	4	17	13
5	1557	1552	5	3	2

Frequency keywords			Dispersion keywords		
Rank F-KWs	**Rank D-KWs**	**Difference**	**Rank D-KWs**	**Rank F-KWs**	**Difference**
6	6	0	6	6	0
7	12	5	7	20	13
8	9	1	8	15	7
9	2	7	9	8	1
10	47	37	10	21	11
	Mean difference	162.3		Mean difference	5.5

The first three columns show the comparisons of the ranks of the F-KWs, while the remaining columns show the comparisons of the ranks based on the D-KWL. For F-KWs, the first column shows the rank of the top 10 F-KWs (1-10), and the second column displays the rank of each F-KWs in the D-KWL. For example, the 5th F-keyword is ranked 1557th in the D-KWL, showing a large rank difference. The third column shows the difference in rank for the top 10 F-KWs compared to their ranks in the D-KWL. The mean difference was then calculated to show an approximate difference, whether these top 10 remain top keywords when ranked by the other method. The same analysis was conducted for the D-KWs compared to their ranks in the F-KWL. In the analysis, the mean rank difference of D-KWs relative to F-KWs and the mean rank difference of F-KWs relative to D-KWs were calculated for the top 100 keywords in each case.

Unique and shared keywords

To compare how different (or similar) the keywords returned by both approaches are, keywords which appeared in both F-KWL and D-KWL (shared keywords) and those found in any of the list only (unique words) were counted and presented as a percentage.

RESULTS

This section presents the analysis of top 100 keywords from both approaches. Content distinctiveness and content generalisability are measured through frequency, dispersion, the number of function words, proper nouns, and abbreviations, while differences and similarities of keywords from these approaches are in terms of rank difference, and shared and unique keywords.

To illustrate the results of the quantitative data analysis, an example of one KWL derived from the frequency-based approach is presented in the appendix. This example is one of six corpus pairs, each analyzed using two approaches, resulting in 12 cases in total. It shows the top 100 F-KWs from the comparison of Oz (as the TC) and Peter Pan (as the RC). The calculated metrics for this KWL, including the mean of normalized frequency (as presented in Table 5), the average percentage of the range (as presented in Table 6), and the average rank difference (as presented in Figure 3), are displayed in the bottom row of the table.

Frequency and frequency ratio

The means of normalized frequencies along with frequency ratio of the top 100 keywords in the TC and the RC, showing the overall occurrence of these top 100 words in the target corpus and in the reference corpus by both methods, are presented in Table 5.

Table 5

Mean of normalised frequency of the top 100 keywords in the target and in the reference corpus

KWLs	Comparisons	Frequency keywords			Dispersion keywords		
		f(TC)	f(RC)	f(TC):f(RC)	f(TC)	f(RC)	f(TC):f(RC)
1	Oz vs. Peter Pan	5.46	2.16	2.53	1.70	0.02	85.00
2	Oz vs. Fic	6.02	2.53	2.38	1.74	0.05	34.80
3	Oz vs. AmE	7.53	3.15	2.39	1.81	0.05	36.20
4	AL vs. Sci	0.41	0.09	4.56	0.29	0.02	14.50
5	AL vs. BE06	0.47	0.11	4.27	0.24	0.01	24.00
6	Sci vs. BE06	0.30	0.08	3.75	0.11	0.00	0.11:0

Note: Division by zero is undefined. Therefore, the calculation for 0.11:0 was not performed.

Generally, the frequencies of the top 100 F-KWs are higher than those from D-KWs, both in the target corpus and the reference corpus.

When examining the frequency ratio of keywords in the TC as compared to the RC, main advantage of D-KWs over F-KWs is that D-KWs are very rare in the reference corpus. This suggests that the D-KWs are more unique and specific to the TC, performing better in distinguishing between the target and reference corpora.

Dispersion and dispersion ratio

Average percents of the number of occurrence (normalised occurrence per 100 texts) of keywords across texts in a target and in a reference corpus along with their ratio of range in the TC and the RC are presented in Table 6.

Table 6

Average percents of the dispersion of keywords and their ratio in a target and in a reference corpus

KWLs	Comparisons	Frequency keywords			Dispersion keywords		
		Range (TC)	Range (RC)	R(TC): R(RC)	Range (TC)	Range (RC)	R(TC): R(RC)
1	Oz vs. Peter Pan	52.54	20.00	2.63	37.75	1.53	24.67
2	Oz vs. Fic	58.33	24.10	2.42	41.42	3.28	12.63
3	Oz vs. AmE	63.17	24.30	2.60	43.71	3.18	13.75
4	AL vs. Sci	16.18	3.68	4.40	14.25	1.31	10.88
5	AL vs. BE06	17.72	8.51	2.08	12.02	2.31	5.20
6	Sci vs. BE06	12.07	7.44	1.62	7.30	1.15	6.35

In general, the top 100 F-KWs more disperse in a corpus than the D-KWs, for all corpus sizes and types of comparisons. The F-keyness can identify keywords that are distributed across a larger proportion of texts within a target corpus; however, they are also found in a certain

number of texts in the reference corpus. For example, Figure 1 shows the dispersion of the top 100 keywords from F- and D-KWA of Oz vs AmE06. We can see that F-KWs are much widely dispersed in the Oz corpus (TC) based on the raw number of texts where the keywords appeared. When considering dispersion ratio D-KWs show high dispersion in the target corpus relative to the reference corpus. On the other hand, D-KWs are more dispersed in the Oz corpus according to relative dispersion, the number of texts that keywords appeared in Oz relative to the reference corpus.

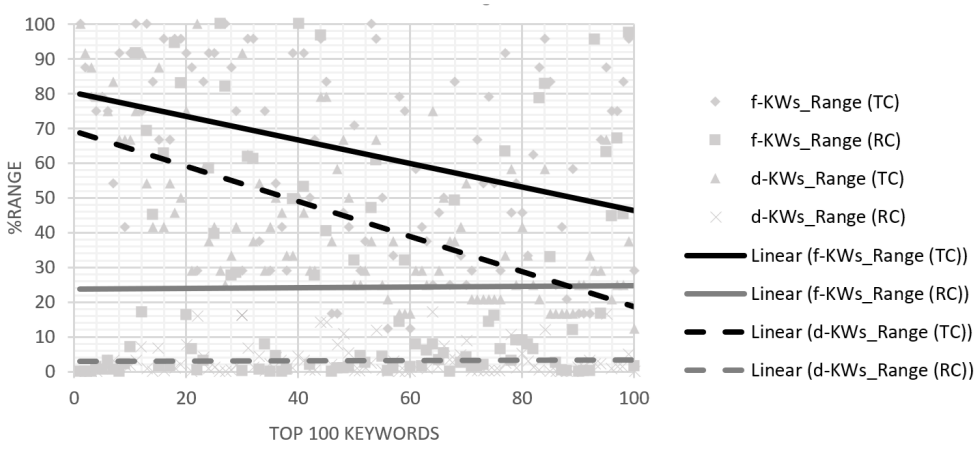

Figure 1 Dispersion of the top 100 keywords from F- and D-KWA of Oz vs AmE06

This means that both approaches produced keywords which are relevant to the target corpus either based on individual dispersion or relative dispersion, depending on whether the particular study required keywords which clearly disperse in the target corpus only or whether the occurrence of keywords in reference corpus is acceptable. To determine whether keywords well addressed their focus, examining more closely to keywords themselves is recommended.

The number of function words, proper nouns, and abbreviations

Given their commonality across corpora, function words were considered non-distinctive in terms of content. Proper nouns and abbreviations, with their limited dispersion, were deemed non-generalizable. The frequency of these words in all KWLs was counted to measure distinctiveness and generalizability.

Figure 2 shows that the top 100 F-KWs have a greater prevalence of function words (13 words on average) than D-KWs (3 words). The average number of proper nouns and abbreviations, however, does not show a clear difference.

When examining individual lists, the number of proper nouns in both F-KWL and D-KWL are high in all comparisons with Oz corpus as the target corpus and the number of abbreviaitons are high in the KWLs when abstract corpora are used as the target corpus. These differences are likely due to the nature of the target corpora, when proper nouns are found in novels and abbreviations are found in research.

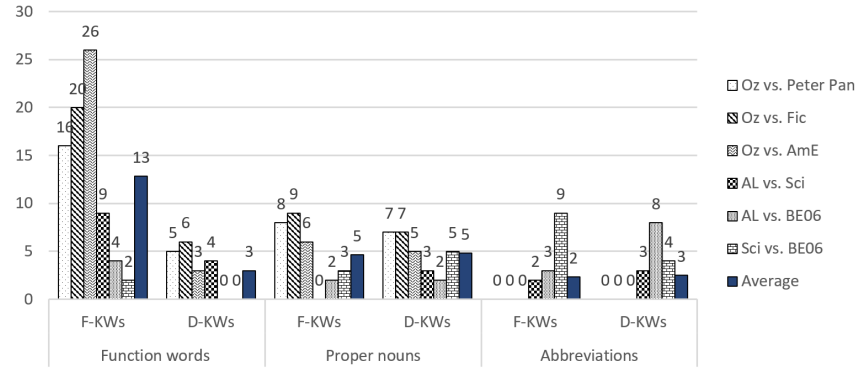

Figure 2 Proportions of categories of keywords according to content distinctiveness and generalisability

In summary, in terms of content distinctiveness, the analysis measured frequency and frequency ratio, dispersion and dispersion ratio, and the number of function words. Overall, the results show that D-KWs exhibit higher frequency ratios and dispersion ratios, suggesting that they have a stronger relationship with the content of the target domain represented by the target corpus compared to the domain represented by the reference corpus, unlike the F-KWs.

Content generalizability was measured through dispersion and dispersion ratio, as well as the number of proper nouns and abbreviations. The results show that D-KWs have a higher dispersion ratio, while the average number of proper nouns and abbreviations is approximately the same. This suggests that D-KWs are more representative of texts in the entire target corpus compared to F-KWs. However, the number of proper nouns and abbreviations is likely to be more influenced by the nature of the content and genre of the target corpus. For example, proper nouns are identified as keywords when using a classic literature corpus (Oz) as the target corpus, while abbreviations are found more in the comparisons where research article abstracts are used as the target corpus.

Rank difference

Average difference in ranks between ranks of 100 F-KWs and 100 D-KWs when they are in the other keyword list are shown in Figure 3.

	Oz vs. Peter Pan	Oz vs. AmE-Fic	Oz vs. AmE	AL vs. Sci	AL vs. BE06	Sci vs. BE06
Based on f-kws	350.52	408.12	359.64	113.33	286.85	1222.64
Based on d-kws	52.66	71.27	73.33	31.35	92.87	152.9

Figure 3 Average difference in ranks when keywords are sequenced by the two approaches

Mean rank difference based on top 100 F-KWs

Overall, the top 100 F-KWs in the D-list ranked much lower in terms of dispersion (indicated by high rank values), suggesting that these keywords may not be widely dispersed across the texts in the corpus and are therefore less likely to be identified as top 100 keywords.

Mean rank difference based on top 100 D-KWs

The top 100 D-KWs generally had smaller rank differences, suggesting that these keywords tend to be ranked higher on the list when sorted by frequency.

Across the six keyword lists, regardless of the number of texts in the target corpus (24, 100, or 200) and the comparison types (specific vs. specific corpus or specific vs. general corpus), the ranks of F-KWs in D-lists are generally higher than those of D-KWs in F-lists. This trend indicates that D-KWs also rank highly on F-based lists.

Shared and unique keywords between frequency and dispersion keyword lists

Keywords in both F-KWL and D-KWL were identified for shared keywords, those occurred in both lists, and unique keywords, those appeared only in one of the keyword lists. Figure 4 shows the number of shared and unique keywords.

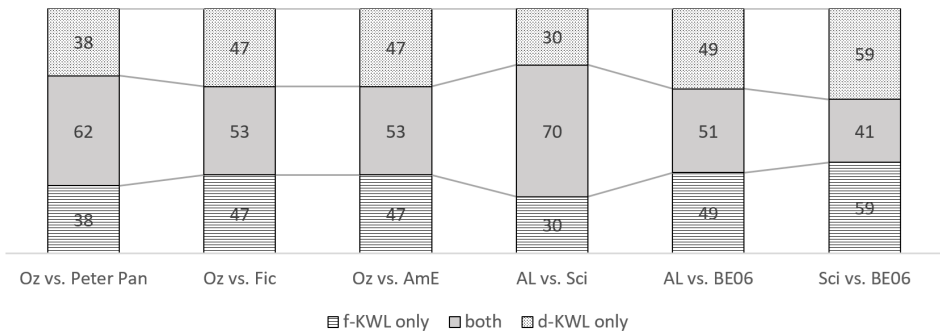

Figure 4 The number of shared keywords and different keywords

Figure 4 shows that many keywords are identified as keywords by both approaches, ranging from 41 to 70 percent. It is noticeable that the choice of reference corpus seems to affect the quantity of shared and unique keywords. Comparisons of keyword lists between those generated from specific and specific corpora reveal a higher number of shared keywords (such as, Oz vs. Peter Pan shows more shared keywords than Oz vs. Fic and AL vs. SCI reveals more shared keywords than those in AL vs. BE06).

DISCUSSION AND RECOMMENDATIONS

This study investigated D-KWs compared with F-KWs to determine whether this approach can be used in conditions where a target corpus is small, containing only a few texts, and in

conditions where a specific corpus serves as a reference corpus. Overall, the findings from the comparisons of keywords from both approaches suggest that the D-KWA produced useful keywords in the observed conditions. This section discusses the characteristics of keywords from the two approaches in terms of their relevance to the target texts, as well as their differences and similarities. These aspects could serve as input for discourse analysts and corpus analysts to consider when choosing between the frequency or dispersion approach to suit their research. It is important to emphasize that all conclusions and recommendations are based on the findings of the analysis of specific settings, specifically small target corpora with a maximum of 200 texts and small reference corpora with a maximum of 500 texts.

Relevance of F-KWs and D-KWs to the target corpus

D-KWs from the comparisons under the observed settings are as useful, particularly in terms of relevance to target texts, as those reported in previous studies (Egbert & Biber, 2019; Xu & Jhang, 2020). D-KWA produced KWs with higher relative frequency and relative dispersion (although less frequent and less dispersed in the target corpus than the F-KWs). It produced fewer function words, and approximately equal numbers of proper nouns and abbreviations. D-KWs, while less frequent in the reference corpus, are particularly distinctive to the target corpus. Considering the average proportion of proper nouns and abbreviations, content generalizability is similar for keywords from both approaches, with proper nouns and abbreviations appearing to be due to the nature of the target corpus. For instance, research articles often feature more abbreviations, while a corpus of novels features more proper nouns.

These findings suggest that the D-KWs perform better than the F-KWs in terms of distinctiveness and generalizability as in Egbert and Biber (2019). This study supported the arguments for the D-KWs to perform better in distinguishing keywords of the target from the reference corpus, making them more representative to the texts and more straightforward to interpret in relation to the content of the target texts.

Differences and similarities in rank and word forms in F-KWLs and D-KWLs

The six case studies show that the F-KWs tend to rank much lower in the D-KWL, supporting the findings of the dispersion measure that they are less widely distributed across the target corpus relative to the reference corpus. Conversely, the D-KWs generally show smaller rank differences in the F-list, suggesting they also rank high when sorted by relative frequency. This analysis indicates that D-KWA frequently generates keywords that are also identified as top keywords by F-KWA. Conversely, reliance solely on F-KWA may lead to the exclusion of certain keywords that would otherwise be identified by the D-KWA. Therefore, it is advisable for analysts to consider the potential limitations of F-KWA and to explore the use of D-KWA as a complementary approach when F-KWA results appear insufficient. Combination might be worth considering (as suggested by Xu and Jhang, 2020). Using both frequency and dispersion analyses can provide a more balanced view. Frequency analysis offers insights into frequent and dispersed words, while dispersion analysis highlights distinctive and relevant keywords specific to the target corpus.

Based on the findings of shared and unique keywords, both approaches produced over 40 percent of the same keywords. While the shared keywords are more prevalent in comparisons between specific corpora, comparisons between specific and general corpora show roughly equal or fewer shared keywords. Unique words from each approach seem useful, though the F-KWA approach includes a higher proportion of function words. It might be reasonable to say that analysts can use either method to identify keywords that are useful for interpreting and understanding a target corpus, especially when comparing between specific and specific corpora, with awareness of advantages and limitations of each method.

Choosing F-keyness or D-keyness

In addition to questions about the similarities and differences of the keywords from the two approaches, questions such as 'Given the similarities and differences in the keywords identified by both approaches, should researchers choose to use frequency or dispersion?' might also be expected to be addressed.

Previous studies have supported different principles, offering reasons that favor frequency or dispersion (Egbert et al., 2020; Egbert & Biber, 2019), and even combinations of both methods (Gries, 2021; Xu & Jhang, 2020). The primary consideration will always be the objectives and research questions of the research as they guide decisions on methodology including the choice of observational units, the operationalisation of variables, and the research design (Egbert et al., 2020). Rather than suggesting which approach to choose, it would be more reasonable to highlight some key considerations when deciding between frequency or dispersion approaches.

The only specific recommendation is that D-KWA requires multiple texts in both the target and reference corpora to be kept separate and cannot be conducted on corpora saved as single text files.

Based on my review of keyword analysis methodologies and applications, I have observed that the objectives of keyword analysis and the types of corpora used can be classified into three groups (see also Jeaco, 2020; Pojanapunya, 2017; Pojanapunya & Watson Todd, 2018). I will use these classifications to outline key considerations for choosing between the frequency or dispersion approach for keyword analysis.

1. Identifying differences between two specific corpora (Specific vs. Specific)

In comparing specific corpora, over half of the keywords were shared. Unique words included some function words, which were more common in F-KWL. Proper nouns and abbreviations are more influenced by the target corpus's content. Analysts can choose either frequency or dispersion. Both methods offer similar content generalizability. However, the choice may depend on the nature of the target corpus. If the target corpus content includes specialized terminology, such as in scientific research, F-KWA returns more function words than D-KWA which are less relevant to the target corpus.

2. Characterizing the target texts comparing the specific corpus against a general corpus (Specific vs. General)

This comparison type produced fewer shared words. Unique words from F-KWL contained more function words. In this case, analysts might consider prioritizing overall frequency or word dispersion across texts in a corpus, depending on the purpose. They should consider whether their observational unit can be described at the level of the text or the corpus (Egbert et al., 2020) and then they choose the method accordingly. Since the purpose of the research is to characterize the target texts, analysts might expect the resulting keywords to offer a broad and comprehensive understanding of the texts for interpretation. Therefore, it may be worth considering the use of both approaches, as suggested by Xu and Jhang (2020). While the D-KWA method ensures that the keywords are truly specific to the target corpus by distinguishing them from the reference corpus more effectively than the F-KWA, the F-KWA method highlights the keywords occur with high frequency overall.

3. Identifying key terms out of a corpus for creating a word list

Suggestions provided for the purpose of characterising a corpus can also be applied to this purpose of creating a word list, especially the academic word lists. In addition to the type of reference corpus used for comparison (either a specific or general) that affects the keyword outputs, the choice of keyness approach depends on the scope of the word list. In other words, it depends on the extent to which each word list is intended to represent. For example, choosing the approach for creating a word list for an academic discipline from a corpus consisting of several sub-corpora of texts from various sub-disciplines would depend on whether the list represents words that frequently appear in the entire corpus or words that are used across multiple sub-disciplines. To ensure that the word list includes both frequent and well-dispersed words across sub-disciplines, using both approaches could be one option. However, to ensure that a word list contains words common across sub-disciplines with less bias towards any specific sub-discipline in a corpus, D-KWA seems to outperform F-KWA, based on the assumption of Egbert and Biber (2019) that words which are well-dispersed across texts in a corpus will occur with a certain frequency.

In summary, this research demonstrates that D-keyness can be effectively used when the target corpus is small, consisting of no more than 200 files. Both methods provide useful keywords and yield conclusions about content distinctiveness and generalizability that are consistent with previous research. When comparing keywords from the D-KWA with those from the F-KWA, over 40% of the keywords are shared. Most unique keywords from both approaches are content words that enhance our understanding of the target texts.

While there is no single best approach, the findings from this research provide information on the characteristics of F-KWs and D-KWs that analysts should consider when selecting a method based on their specific research objectives. These findings can also help them become aware of the limitations of their chosen approach and the benefits of alternatives. For analysts new to keyword analysis, this study serves as a useful resource and guideline for future applications.

THE AUTHOR

Punjaporn Pojanapunya is a researcher at the School of Liberal Arts at King Mongkut's University of Technology Thonburi (KMUTT), Thailand. Her research interests include the methods and applications of corpus linguistics, with a particular focus on keyword analysis within the field of applied linguistics.
punjaporn.poj@mail.kmutt.ac.th

REFERENCES

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software.

Bailey, A. (2018). Dementia and identity: A corpus-based study of an online dementia forum. *Communication & Medicine, 15*(3). https://doi.org/10.1558/cam.36150

Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (Ed.), *Research methods in linguistics* (pp. 95–113). Continuum.

Baker, P., Hardie, A., & McEnery, T. (2013). *A glossary of corpus linguistics*. Edinburgh University Press. https://doi.org/10.1515/9780748626908-002

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords. *Journal of English Linguistics, 32*(4), 346–359. https://doi.org/10.1177/0075424204269894

Bancroft-Billings, S. (2020). Identifying spoken technical legal vocabulary in a law school classroom. *English for Specific Purposes, 60*, 9–25. https://doi.org/10.1016/j.esp.2020.04.003

Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). *#LancsBox: Lancaster University corpus toolbox* (Version 6.0) [Computer software]. Lancaster University. https://corpora.lancs.ac.uk/lancsbox

Clarke, I., Brookes, G., & McEnery, T. (2022). Keywords through time. *International Journal of Corpus Linguistics, 27*(4), 399–427. https://doi.org/10.1075/ijcl.22011.cla

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29–59. https://doi.org/10.1075/ijcl.14.1.03cul

Dayter, D., & Messerli, T. C. (2022). Persuasive language and features of formality on the r/ChangeMyView subreddit. *Internet Pragmatics, 5*(1), 165–195. https://doi.org/10.1075/ip.00072.day

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104. https://doi.org/10.3366/cor.2019.0162

Egbert, J., & Burch, B. (2023). Which words matter most? Operationalizing lexical prevalence for rank-ordered word lists. *Applied Linguistics, 44*(1), 103–126. https://doi.org/10.1093/applin/amac030

Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Press. https://doi.org/10.1017/9781108888790

Gabrielatos, C., & Marchi, A. (2012, September 14). *Keyness: Appropriate metrics and practical issues* [Paper presentation]. Critical Approaches to Discourse Studies 2012, Bologna, Italy. http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26MarchiKeyness-CADS2012.pdf

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction.* Routledge. https://doi.org/10.4324/9781315746210

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1–33. https://doi.org/10.32714/ricl.09.02.02

Jeaco, S. (2020). Key words when text forms the unit of study: Sizing up the effects of different measures. *International Journal of Corpus Linguistics, 25*(2), 125–155. https://doi.org/10.1075/ijcl.18053.jea

Ji, T., & Li, K. (2024). A hidden population: A rhetorical genre analysis of the posts in the Baidu depression community. *Social Science & Medicine, 353*, Article 117036. https://doi.org/10.1016/j.socscimed.2024.117036

Kilgarriff, A. (2009, July 20-23). Simple maths for keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool.

Lam, J. C., Cheung, L. Y., Wang, S., & Li, V. O. (2019). Stakeholder concerns of air pollution in Hong Kong and policy implications: A big-data computational text analysis approach. *Environmental Science & Policy, 101*, 374–382. https://doi.org/10.1016/j.envsci.2019.07.007

Langenhorst, J., Frommherz, Y., & Meier-Vieracker, S. (2023). Keyness in song lyrics: Challenges of highly clumpy data. *Journal for Language Technology and Computational Linguistics, 36*(1), 21–38. https://doi.org/10.21248/jlcl.36.2023.236

Lexical Computing. (n.d.). *Sketch Engine* [Computer software]. Retrieved May 2025, from https://www.sketchengine.eu/

Li, P. W., & Lu, C. R. (2020). Articulating sexuality, desire, and identity: A keyword analysis of heteronormativity in Taiwanese gay and lesbian dating websites. *Sexuality & Culture, 24*(5), 1499–1521. https://doi.org/10.1007/s12119-020-09709-5

Millar, N., & Budgell, B. S. (2008). The language of public health—A corpus-based analysis. *Journal of Public Health, 16*(5), 369–374. https://doi.org/10.1007/s10389-008-0178-9

Pojanapunya, P. (2017). *A theory of keywords* [Doctoral dissertation]. KMUTT Library Network. https://opac.lib.kmutt.ac.th/vufind/Record/1370763

Pojanapunya, P., & Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory, 14*(1), 133–167. https://doi.org/10.1515/cllt-2015-0030

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

Rayson, P. (2013). Corpus analysis of key words. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell Publishing Ltd.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, 9*, 1–6. https://doi.org/10.3115/1117729.1117730

Scott, M. (1997). PC analysis of key words – And key words. *System, 25*(2), 233–245. https://doi.org/10.1016/S0346-251X(97)00011-0

Scott, M. (2024). *WordSmith tools version 9* (64 bit version). Stroud: Lexical Analysis Software. https://lexically.net/wordsmith/

Sönning, L. (2022a). *Evaluation of text-level measures of lexical dispersion: Robustness and consistency* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/h9mvs

Sönning, L. (2022b). *Evaluation of keyness metrics: Reliability and interpretability* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/eb2n9

Xu, L., & Jhang, S. E. (2020). Keyword analyses of English charter parties. *Linguistic Research, 37*(2), 267–288.