Digital Scholarship in the Humanities, 2025, **40**, 668–681
https://doi.org/10.1093/llc/fqaf031
Advance access publication 2 May 2025
Full Paper

OXFORD

Key cluster identification in literary texts using and comparing multiple measures: an exploratory comparative study and its implications

Hongwei Zhan[1,*]

[1]School of International Studies, Hangzhou Normal University, 2318 Yuhangtang Road, Hangzhou, Zhejiang Province, China
*Corresponding author. Hangzhou Normal University, 2318 Yuhangtang Road, Cangqian Hangzhou 310036, China. E-mail: hwzhan78@163.com

Abstract

Various methods have been developed for identifying keywords/key clusters. Most of these methods use a reference corpus to identify keywords/key clusters in the target corpus although a few studies have employed methods for key word/cluster identification without the use of a reference corpus. However, little research appears to have been done comparing the effectiveness of these methods, especially when they are used for identifying key clusters, a relatively new concept than keywords. To address this research gap, this study compares the accuracy and effectiveness of the following five methods in identifying key clusters in a corpus of Charles Dickens's novels without the use of a reference corpus: *TF (Term Frequency*, a common frequency measure*)*, *DP_{norm}* (*Deviation of Proportions normalized*, a robust and effective dispersion measure), and *PPMI* (*Positive Pointwise Information*, a widely used association strength measure), and *TF-IDF* (*Term Frequency—Inverse document*, a blended method that considers both term frequency and inverse document frequency), and *TF-DP_{norm}* (*Term Frequency-DP normalized*), a self-developed blended method that factors in both frequency and normalized dispersion. With the top key clusters that Mahlberg (2007) identified in the same Dickens's corpus of novels as the benchmark, the results of the comparison show that, of the five methods, the self-developed *TF-DP_{norm}* method and the *TF* method are the most accurate and effective in identifying key clusters in literary texts when no reference corpus is used. Reasons for the differences across the methods are explored and research implications are also discussed.

Keywords: keywords; key clusters; corpus stylistics; vector space model.

1. Introduction

Corpus linguistic methods have become increasingly popular for analyzing literary texts, with Charles Dickens' novels in particular. For example, Schneider (2024) employed corpus-based topic modeling and distributional semantics to explore Dickens' works, uncovering insights related to social criticism, literary realism, and narrative structures. One corpus method commonly used in many existing studies is keyword analysis. Quite a few studies (e.g. Mahlberg 2007; Baker 2009; Culpeper 2009; Mahlberg, Smith, and Preston 2013, Mahlberg *et al.* 2019) show that keyword analysis can successfully identify linguistic patterns that convey the text's poetic function and reveal literary meanings that might not be immediately apparent when analyzed using other methods. However, there are various methods for computing keyness and extracting keywords (Siddiqi and Sharan 2015), including corpus frequency, dispersion measures, and word collocation strength. Each method has its own set of advantages and disadvantages, and scholars have continued to develop methods for keyword extraction and/or evaluate the respective strengths and weakness of the existing methods (e.g. Egbert and Biber 2019; Deng and Liu 2022). Additionally, studies have also started identifying and examining key clusters (a string of words often used together as a unit also known as ngrams in linguistics) thanks to their importance in language use, including in literary texts (Hori 2004; Mahlberg 2007; Fischer-Starcke 2009). Against this backdrop, this study examines the accuracy and effectives of five different methods in identifying key clusters, without the use of a reference corpus, in the aforementioned corpus of Dickens's novels used by Mahlberg (2007) against the top key clusters that Mahlberg (2007) identified in the corpus with the use of a reference corpus. The results of this study would help us better understand which of the method(s) may

© The Author(s) 2025. Published by Oxford University Press on behalf of EADH. All rights reserved.
For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

be more accurate and effective when used to identify key clusters in a corpus of literary texts without using a reference corpus.

2. Background: literature review

2.1 Keyness computation and keyword extraction

Keyness, which refers to the statistical significance of a word's occurrence in a specific text or corpus, is estimated by using a range of features and measures that can be obtained from a formal analysis of the source text (Scott 1997). Depending on the nature of the input data and the keyness characteristics of the target keywords, different keyword extraction methods have been developed that utilize various statistical measures, including frequency, collocation strength, and dispersion (Kilgarriff 2009; Pojanapunya and Todd 2018; Egbert and Biber 2019; Gries 2024).

2.1.1 Frequency

Frequency is a fundamental statistical feature commonly utilized in keyword extraction tasks because a term's representativeness within a text correlates with its frequency of occurrence. That is, the more representative a term is, the more often it tends to appear (Salton, Yang, and Yu 1975). There are several different frequency measures, including *raw term frequency* and *normalized term frequency* with the latter being used when comparing the frequencies of terms across corpora of different sizes (Salton and Buckley 1988; Aizawa 2003). The most common frequency measure-based methods used in keyness analysis thus far include *term frequency* (the frequency of a word/cluster in a corpus) and *corpus frequency*, which identifies keywords by compares the frequencies of words in a corpus against their frequencies in a reference corpus. In other words, this method considers the corpus as its primary unit of observation and it involves two units in its analysis: a target corpus and a reference corpus. Specifically, this approach entails counting the frequency of every word (term-frequency) in both corpora and calculating the keyness for each word in the target corpus. Various statistical measures, such as chi-square, log-likelihood, log ratio, and simple frequency difference, have been employed to quantify keyness or keywords (Paquot and Bestgen 2009; Pojanapunya and Todd 2018; Rayson 2022). Words that occur with statistically greater frequencies in the target corpus than in the reference corpus are identified as keywords or key terms.

More specifically, based on existing research (Scott 1997), the computation of the 'keyness' of a word in this approach is based on the comparison of relative frequency obtained from the following four kinds of data:

- The frequency of the word in the observed text
- The total word count of the observed text
- The frequency of the word in the reference corpus
- The total word count of the reference corpus

In this computation, the chi-square value derived from the chi-square test of four independent sample tables, adjusted with Yates continuity correction, is referred to as the 'critical value' (Scott 1997; Baker 2004; Gabrielatos 2018). Commonly used tools for corpus frequency analysis, such as AntConc or WordSmith, have built-in chi-square or log-likelihood algorithms that can automatically generate a keyword list, ranking keywords by their keyness.

However, this approach has a notable weakness: it treats the corpus as a homogeneous entity for keyness computation. According to Egbert and Biber (2019) while the resulting keyword lists may include words that are relatively frequent within the corpus, these keywords often lack widespread dispersion across the texts, leading to a misrepresentation of the target discourse domain.

2.1.2 Dispersion

To address the aforementioned weakness of the term frequency-based approach, Egbert and Biber (2019) proposed a dispersion-based approach that shifts the focus from term frequency to text dispersion, operating on the hypothesis that 'keyness could be measured without making any reference to word frequency by focusing entirely on the text dispersion of words' (p. 84). This method relies on *range*, the most basic original dispersion measure, which defines the dispersion of a word as the total number of texts in which a word appears at least once. More specifically, this method identifies keywords by using log-likelihood statistic (G^2, also known as the likelihood-ratio chi-square) to ascertain the significant difference between words' ranges in the target corpus and their ranges in the reference corpus. This log-likelihood statistical method is used because it is effective in estimating probabilities including in cases where counts are very low (see Dunning 1993; Kilgarriff 2005). The advantage of dispersion-based measures of keyness, such as this one, has also been affirmed by other scholars (Gries 2021; Sönning 2024).

However, according to Gries (2021), Egbert and Biber's (2019) range-based dispersion measure has a weakness since it overlooks two key factors: the sizes of the parts of a corpus and the frequencies with which words occur across those corpus parts. Gries (2021) argues that the parts of a corpus can vary in size, but Egbert and Biber's (2019) method focuses exclusively

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

on whether a word appears in a certain corpus part without considering the difference in the size of the various parts. In fact, a similar problem is also found in Juilland and Chang-Rodriguez' (1964) Juilland's *D*, a widely used dispersion measure that was once believed to be the most reliable measure for assessing lexical dispersion (Lyne 1985; Biber *et al.* 2016) because Juillan's *D* is also does not take into consideration the fact that the different parts of a corpus often vary in size. To address this issue, Gries (2008) developed a new dispersion measure known as *DP* (*Deviation of Proportions*). Biber *et al.*'s (2016: 439) comparison of Juilland's *D* and *DP* in terms of reliability and effectiveness finds that *DP* is indeed 'a more reliable and effective measure of dispersion in a large corpus divided into many parts'. Gries himself (2021) also recommends his *DP* as a more reliable and informative measure of dispersion for keyword identification. The formula for calculating *DP* is $DP = 0.5 \times \sum_{i=1}^{n} |\frac{v_i}{f} - s_i|$, which, in plain language, involves the following rather simple steps (Gries 2008: 415):

i) Calculate the expected percentages or proportions ('which take differently-sized corpus parts into consideration') and observed portions of the target words in each of the corpus parts.
ii) Compute the differences between the expected and observed portions for each corpus part,
iii) Add up the differences and divide the sum by 2.

The *DP* value 'can theoretically range from approximately 0 to 1', with low values close to 0 indicating high or even dispersions while high values close to 1 suggesting low dispersions (Gries 2008: 415). Later, Gries (2010, 2020) introduced an improved version of *DP*, that is, a normalized *DP* version named DP_{norm} with the formula: $DP_{\text{norm}} = \frac{DP}{1-min(s)}$. Specifically, DP_{norm} adjusts the *DP* values to a standardized scale by normalizing dispersion scores to make them more robust to corpus segmentation difference and make it easier to compare dispersion across corpora of varying sizes. Because of its greater robustness than *DP*, DP_{norm} will be adopted as the representative method of dispersion measures in this study.

2.1.3 Co-occurrence

Word co-occurrence serves as a statistical feature that evaluates the representativeness and exhaustivity of keywords. The core principle is to identify words that commonly appear together within specific contexts (Bullinaria and Levy 2007). Co-occurrence is measured using association metrics, such as *Co-Occurrence Frequency* and *Pointwise Mutual Information* (*PMI*, Church and Hanks 1990), to determine the strength of the statistical relationship between words, that is, to ascertain how often they appear together in a text compared to what would be expected by chance, highlighting the likelihood of those words forming a collocation or cluster based on their co-occurrence patterns within a corpus (Deng and Liu 2022). In association measures, a higher score signifies a stronger association between the words, suggesting a more likely collocation or cluster. Now we turn to how co-occurrence frequency and PMI (two of the most commonly used associations measures) are computed and determined.

Co-occurrence frequency simply counts how many times certain words appear together within a specific context, such as a fixed-size window, a sentence, a text/document, or a corpus. Since it counts frequency, it looks like the frequency measures mentioned in Section 2.1.1, but unlike the common frequency measures, which calculates the frequency of individual words, co-occurrence frequency counts the frequency of strings of words known variously as collocations, clusters, formulae, and lexical bundles, a very important language feature and an issue of this study that will be addressed below. This co-occurrence frequency method provides a basic measure of the proximity of the co-occurring words, but it does not account for the overall frequency of the individual words used in other contexts, that is, contexts where the words do not occur together.

Unlike *co-occurrence frequency*, *PMI* takes into consideration the overall frequency of the individual words in the other contexts by comparing the frequencies of the co-occurring words used together with the frequencies of the words when used with other words. Specifically, it calculates the statistical dependence between words, giving higher scores to word combinations that occur together more often than expected. *PMI* is computed as follows:

$$PMI(w_1, w_2) = \log_2\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

PMI can produce negative values when a word pair appears together less often than expected by chance, which is not always meaningful in collocation analysis. As a result, some linguistic researchers (e.g. Church and Hanks 1990; Niwa and Nitta 1994) have adopted a slightly different version of *PMI* called *Positive Pointwise Mutal Information* or *PPMI*, which turns all negative *PMI* values into 0 with the following formula:

$$PPMI(w1, w2) = \max(PMI(w1, w2), 0)$$

In other words, in the *PPMI* formula, if *PMI* is negative, it is replaced with **0** in *PPMI*.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Finally, it is also worth mentioning that there is another largely association-based method of keyword identification known as *Rapid Automatic Keyword Extraction* (*RAKE*) developed by Rose *et al.* (2010). It is especially useful for identifying multiword key technical terms or clusters (mostly noun phrases made up of nouns) by computing how often the words in the clusters co-occur, but this method relies heavily on the use of lists of stop words (such as articles and prepositions). This latter practice limits RAKE's value in keywords/cluster extraction in literary texts because important and useful keywords/clusters in literary texts are not restricted to noun phrases. Hence, *PPMI* will be adopted in this study as the representative method of co-occurrence association measures.

2.1.4 Blended methods

Additionally, there have also been methods that blend both corpus frequency-based and dispersion-based measures. For instance, Millar and Budgell (2008) combined corpus frequency keyword analysis with minimum text dispersion thresholds to refine the identification of key terms. Another approach, known as key keyword analysis (Scott 1997), uses the frequency generated to rank keywords based on the percentage of texts within the target corpus where these terms are deemed key. In short, these blended methods provide a more accurate and nuanced understanding of keyword significance by considering both frequency and distribution across texts. However, both these blended methods use a reference corpus, which will not work for the purpose of our study: identifying keywords without using a reference corpus.

Jone's (2004) *Term frequency—Inversed document frequency* (*TF-IDF*) method is arguably the only blended keyword identification method that, strictly speaking, makes no use of a reference corpus. In this method, *term frequency*, as already mentioned above refers to the frequency of a term in a document/corpus while *document frequency* (*DF*) means the number of documents out of the entire corpus (i.e., out of the total number of documents in the corpus) that contain the term, which is the measure of *dispersion* of the term in the corpus. The importance of a word in a document is directly proportional to *TF* but inversely proportional to *DF*, that is, the lower the dispersion of a term across other documents, the stronger its keyness in the document in question as will be illustrated below. Furthermore, each *TF-IDF* value involves three parameters: corpus, document, and term. Consider this scenario: the frequency of a term varies across different documents within the same corpus. To comprehensively consider these parameters, some conversions are necessary. The formula is $\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$ where *IDF* is the result of dividing the total number of documents in the corpus by the number of documents that contain the word and then calculating its base-10 logarithm. Let us look at two hypothetical cases in a corpus of ten documents: Term 1 occurs ten times in the corpus, but all its ten tokens are in one document, specifically Document 1; in contrast, Term 2 also occurs ten times in the corpus, but its ten tokens appear evenly across all the ten documents, with one token in each document, including Document 1. Regarding the *TF* measure, the higher the frequency of a term in a document, the higher its *TF* is, so Term 1 has a *TF* of 10 in Document 1 while Term 2 has a *TF* of only 1 in Document 1, with Term 1 being obviously far more important in Document 1. Based on the *IDF* formula, the fewer documents containing this term in the entire corpus, the greater its *IDF* value is and the more important the term will be as can be seen in the two hypothetical cases: the *IDF* of Term 1 is $lg_{10}\frac{10}{1} = 1$ while the *IDF* of Term 2 is $lg_{10}\frac{10}{10} = 0$, which clearly suggests that Term 1, with its *IDF* being 1, is significantly more important in Document 1 than Term 2 whose *IDF* is 0. In short, the greater the *TF-IDF* value of a term is, the more important the term is in the document in which the term appears.

From the perspective of the whole corpus, Term 2 is more important because it is more evenly distributed across the corpus. However, *TF-IDF* measure is document-oriented, and it is effective in identifying key terms that are important in a particular document. The importance of a term can be interpreted in two ways: corpus-oriented generality, document-oriented specificity. *TF-IDF*, as has been shown, is used for the latter purpose, that is, for identifying key terms in a document, not a corpus. Given that the purpose of the present study is to identify key terms (specifically key clusters) in the entire corpus of Dickens's novels, not in a specific individual novel, *TF-IDF* may not be an appropriate or good blended method for this study. Hence, for the purpose of identifying keywords/clusters in the corpus of Dickens's novels without using a reference corpus, I developed and used a new blended method, which will be introduced in the Section 3.4.1.

2.2 From keywords and clusters to key clusters

It is well known that words are the basic units of a text, but a text is not merely a simple collection of words. Studies (e.g. Biber *et al.* 1999, Biber, Conrad, and Cortes 2004; Carter and McCarthy 2006) have shown that a text contains many bundles or clusters of words which are often used together or cooccur frequently. These clusters are known variously as chunks, formulae, and multiword units. Because of their high frequency, clusters are the units of sentences and texts

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

and examining word clusters helps us better understand the principles of how words combine to form sentences and how sentences combine to form texts.

Because of this new understanding about language units, in recent years, the concept and scope of keywords have expanded from single words to multiword chunked clusters (Bondi 2010; Scott 2010). Specifically, and technically speaking, a key cluster is an n-gram composed of a group of keywords on the cluster tool interface, such as that shown in the AntConc interface Fig. 1. The collocates in a word cluster are not necessarily keywords, but they can reveal the typical appearance of keywords in the corpus and provide more thematic connotations than a single keyword.

It is also of interest and importance to note that key clusters often contain at least one noun and have the structure of $N + N$ (e.g. 'ongoing *climate change*') and $Adj + N$ (e.g. 'increased use of *artificial intelligence*') to highlight the effective disclosure of the subject content by nominal phrases (Hanks 2004; McEnery and Hardie 2011).

Because of the importance of key clusters, some studies have examined key clusters and their functions in literacy texts, such as Mahlberg (2007) and Mahlberg, Smith, and Preston (2013), both of which deal with key clusters in Dickens' novels. Mahlberg (2007) provides an excellent example of this line of research. In this study, Mahlberg (2007) first identified the clusters and their frequencies in a corpus comprising twenty Dickens's novels against a reference corpus. She then calculated the keyness and significance of the key clusters by employing the loglikelihood measure (G^2) provided by WordSmith. This resulted in a list of sixty-six positive clusters (i.e., those that occurred significantly more in the corpus of Dickens's novels than in the reference corpus) and seven negative key clusters (i.e., those that occurred significantly less frequently in the Dickens's corpus of novels than in the reference corpus) based on a predefined significance level of $P < .000001$. This threshold ensured that the identified clusters were not due to random variation but were meaningful and distinctive in the corpus of Dickens' novels compared to the reference corpus of nineteenth-century fiction.

Given that the purpose of her study was on key clusters used frequently in Dickens's corpus, Mahlberg (2007) focused her analysis on the sixty-six positive

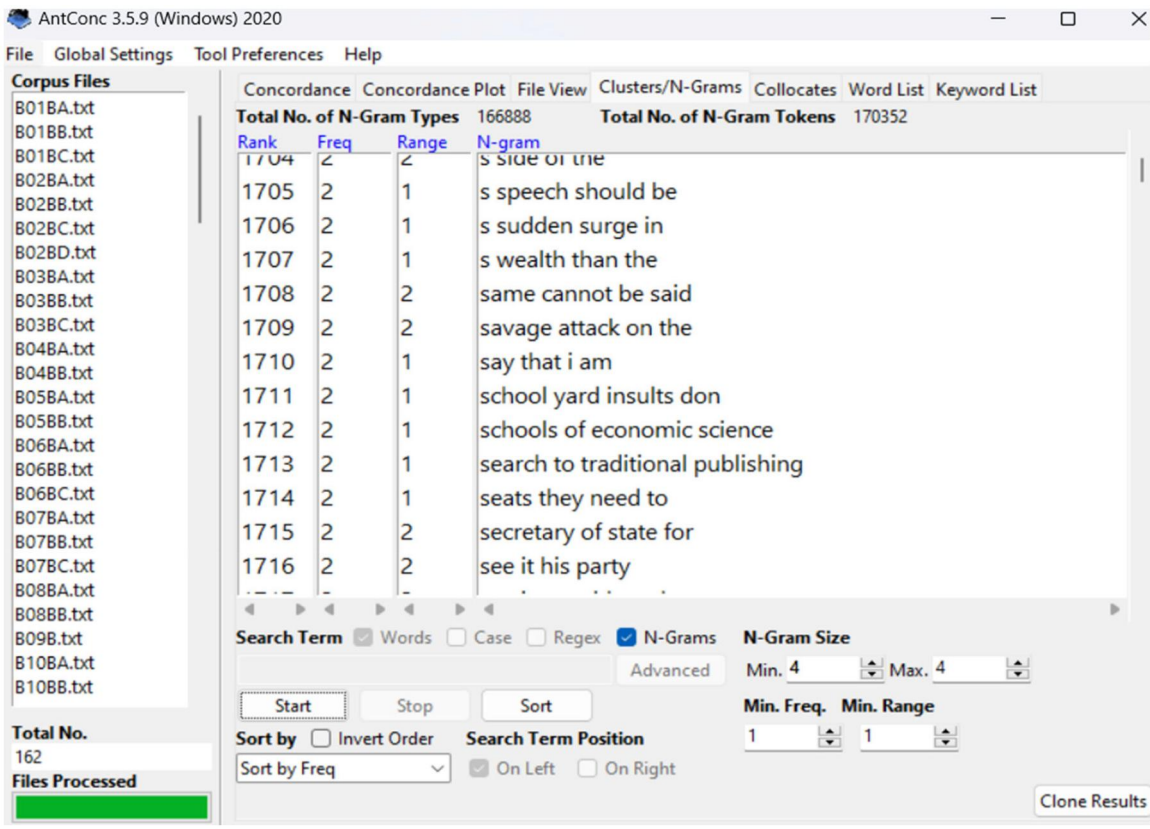

Figure 1. Screenshot of Cluster/Ngam research results on the AntConc interface.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

clusters. She classified these key clusters into five functional/structural categories: *label* (e.g. 'the father of the marshalsea'), *speech* ('do me the favour to'), *as if* (e.g. 'as if he would have'), *body part* ('his hands in his pockets'), and *time and place* ('on the opposite site of'). Mahlberg and associates (Mahlberg 2007; Mahlberg *et al.* 2019) have also explored the semantic and textual functions of key clusters to show how their uses reflect the style of the literary text and the social system being depicted because social conventions shape language expression patterns (Stubbs 2010).

Specifically, Mahlberg and associates' research (Mahlberg 2007; Mahlberg *et al.* 2019) reveals the following findings regarding the discoursal, semantic, and textual functions of the different types of key clusters. First, *label* key clusters play a significant role in character portrayal: more complex word clusters often depicted complex and mysterious characters. On the other hand, *as if* key clusters of denote superficial impressions and construct imaginative scenarios (sentences that guide the subjunctive mood). Compared with *label* and *as if* clusters, *speech* key clusters serve more complex and difficult-to-interpret functions. As an example, the 'I don't know what' cluster might reflect the personal style of a writer, the characteristics of a work, or the idiolect of a character within the work (Mahlberg *et al.* 2019). Concerning *body part* clusters, they are used mostly to describe characters' physical characteristics when speaking. Finally, *time and place* key clusters, as the name indicates, provide the situational contexts of the narrative events being presented. It is pivotal to point out that, according to Mahlberg (2007), these five functional categories are bottom-up, dynamic, and not prescriptive, that is, they can and need to be adjusted according to the actual corpus. Some categories may need further subdivision.

In short, the above review of the important issues related to keyness measures and keyword/key cluster identification/analysis has shown that there are largely four main types of measures used for identifying key clusters (*frequency-based, dispersion-based*, *co-occurrence association-based, and blended)*. However, little research appears to have compared the effectiveness of these identification methods when employed without the use of a reference corpus. Furthermore, key clusters are a relatively new concept, which has received limited research attention thus far. To address these research gaps, this study aims to examine accuracy and effectiveness of following five methods in identifying key clusters in the aforementioned corpus of twenty-three Dickens's novels without using a reference corpus: (1) *TF* (a frequency-based method), (2) DP_{norm} (a dispersion-based method), and *PPMI* (an association strength-based method), *TF-IDF* (a blended method), and *TF-DP*$_{norm}$ (a self-developed blended method to be introduced below). The first four methods are each a strong representative of the aforementioned four types of methods. The examination and comparison of the accuracy and effectiveness of the methods are conducted with those key clusters identified by Mahlberg's (2007) with the use of a reference corpus as a benchmark. The results of this study will help us understand the respective effectiveness (including strengths and weaknesses) of the five tested key-cluster identification methods respectively without the use of a reference corpus.

3. Methodology

3.1 Corpus used

The corpus for this study consisted of twenty-three Charles Dickens' novels. Following the titles provided by Mahlberg (2007), I used the R package {`gutenbergr`} (Robinson 2021) to create a metadata table (Table 1) that included novel titles, title abbreviations, and identifier codes (`gutenberg_id`). I then downloaded the e-books of twenty-three novels. The downloaded data were in a data frame format, comprising three columns: novel title, gutenberg_id, and text as shown as examples in Fig. 2.

3.2 Text cleaning

The original data format of the downloaded corpus showed that the e-books were divided into several lines with varying numbers of words per line, and there

Table 1. Metadata of the 23 Dickens's novels.

Doc ID	Gtbg ID	Full title	Short title
1	675	*American Notes*	*AmNote*
2	40723	*The Battle of Life*	*Batlle*
3	917	*Barnaby Rudge*	*Barnaby*
4	1023	*Bleak House*	*Bleak*
5	19337	*A Christmas Carol*	*Xmas*
6	653	*The Chimes*	*Chimes*
7	20795	*The Cricket on the Heath*	*Cricket*
8	766	*David Copperfield*	*DavCopp*
9	821	*Dombey and Son*	*Dombey*
10	1400	*Great Expectations*	*GreatExp*
11	786	*Hard Times*	*HardTm*
12	644	*The Haunted Man*	*Haunted*
13	963	*Little Dorrit*	*Dorrit*
14	968	*Martin Chuzzlewit*	*Martin*
15	564	*The Mystery of Edwin Drood*	*Edwin*
16	967	*Nicholas Nickleby*	*Nicholas*
17	700	*The Old Curiosity Shop*	*Curiosity*
18	730	*Oliver Twist*	*Oliver*
19	883	*Our Mutual Friend*	*MutualFrnd*
20	580	*The Pickwick Papers*	*Pickwick*
21	882	*Sketches by Boz*	*Sketch*
22	98	*A Tale of Two Cities*	*Tale2Cities*
23	914	*The Uncommercial Traveller*	*Traveller*

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Title	Gutenberg_id	Text
A Tale of Two Cities	98	Book the First--Recalled to Life CHAPTER I. The Period It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope,
Bleak House	1023	PREFACE A Chancery judge once had the kindness to inform me, as one of a company of some hundred and fifty men and women not labouring under any suspicions of lunacy, that the Court of Chancery, though the shining subject of much popular prejudice (at which point I thought the judge's eye had a cast in my
Dombey and Son	821	CHAPTER I. Dombey and Son Dombey sat in the corner of the darkened room in the great arm-chair by the bedside, and Son lay tucked up warm in a little basket bedstead, carefully disposed on a low settee immediately in front of the fire and close to it, as if his constitution were analogous to that of a
Great Expectations	1400	Chapter I. My father's family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip. So, I called myself Pip, and came to be called Pip. I give Pirrip as my father's family name, on the authority of his tombstone and my sister,—Mrs. Joe
Hard Times	786	CHAPTER I THE ONE THING NEEDFUL 'NOW, what I want is, Facts. Teach these boys and girls nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only fo
Little Dorrit	963	PREFACE TO THE 1857 EDITION I have been occupied with this story, during many working hours of two years. I must have been very ill employed, if I could not leave its merits and demerits as a whole, to express themselves on its being read as a whole. But, as it is not unreasonable to suppose that I may have held its
Our Mutual Friend	883	Chapter 1 ON THE LOOK OUT In these times of ours, though concerning the exact year there is no need to be precise, a boat of dirty and disreputable appearance, with two figures in it, floated on the Thames, between Southwark bridge which is of iron, and London Bridge which is of stone, as an autumn evening was
The Battle of Life	40723	PART THE FIRST [Illustration] Once upon a time, it matters little when, and in stalwart England, it matters little where, a fierce battle was fought. It was fought upon a long summer day when the waving grass was green. Many a wild flower formed by the Almighty Hand to be a perfumed goblet for the dew, felt
The Chimes	653	CHAPTER I—First Quarter. There are not many people—and as it is desirable that a story-teller and a story-reader should establish a mutual understanding as soon as possible, I beg it to be noticed that I confine this observation neither to young people nor to little people, but extend it to all conditions of
The Cricket on the Hearth	20795	INTRODUCTION The combined qualities of the realist and the idealist which Dickens possessed to a remarkable degree, together with his naturally jovial attitude toward life in general, seem to have given him a remarkably happy feeling toward Christmas, though the privations and hardships of his boyhood could
The Haunted Man and the Ghost's Bargain	644	CHAPTER I The Gift Bestowed Everybody said so. Far be it from me to assert that what everybody says must be true. Everybody is, often, as likely to be wrong as right. In the general experience, everybody has been wrong so often, and it has taken, in most instances, such a weary while to find out how wrong, that the
The Mystery of Edwin Drood	564	CHAPTER I. THE DAWN An ancient English Cathedral Tower? How can the ancient English Cathedral tower be here! The well-known massive gray square tower of its old Cathedral? How can that be here! There is no spike of rusty iron in the air, between the eye and it, from any point of the real prospect. What is the
The Old Curiosity Shop	700	CHAPTER 1 Although I am an old man, night is generally my time for walking. In the summer I often leave home early in the morning, and roam about fields and lanes all day, or even escape for days or weeks together; but, saving in the country, I seldom go out until after dark, though, Heaven be thanked, I love

Figure 2. Data frame of the Corpus of Dickens's Novels.

```
library(udpipe)
library(dplyr)

zmodel <- udpipe_load_model(file='english-ewt-ud-2.3-181115.udpipe')

dic23df <- zmodel %>%
  udpipe_annotate(dik23text$text,tagger='default',parser='none') %>%
  as.data.frame()
```

Figure 3. Tokenization and part-of-speech tagging.

were also blank lines. To prepare the text for analysis, I merged the lines in each text of the twenty-three novels according to their titles, using the `paste()` function, while deleting redundant blank lines and spaces.

3.3 Ngram-based cluster extraction

To extract clusters based on ngrams and their frequency, I used the R package {udpipe} (Wijffels 2022), a powerful toolkit for natural language processing, particularly well known for its dependency annotation capabilities. Extracting ngram-based clusters is one of its functions. Specifically, I first used the `udpipe_annotate()` function (illustrated in Fig. 3) to tokenize, lemmatize, part-of-speech tag, and dependency tag the corpus data.

Then, I used the {udpipe} function `txt_context()` (see Fig. 4) to extract ngrams, that is, clusters. The extraction involved the following steps. First, we had to set the parameters by deciding on the length of the clusters we wanted to extract. Currently, there is no consensus on the optimal length of clusters in the literature. However, the academic community generally agrees on some basic principles, such as short clusters are too frequent and flexible, making them difficult to

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

```
d23_punct$ngram <- txt_context(d23_punct$token,n=c(-2,-1,0,1,2),na.rm=FALSE)

d23_5gram <-  na.omit(d23_punct)
```

Figure 4. R script for extracting 5-g clusters.

analyze; long clusters are often limited to individual texts and lack generality (Mahlberg 2007). For this reason, I set the cluster length to five words, a length that was adopted by Mahlberg (2007). It is also important to note that five-word clusters have entailment relationships with 4-grams and 6-grams, indirectly reflecting the situations of longer and shorter related clusters. For example, the five-word cluster 'his hands in his pocket' contains the four-word cluster 'hands in his pocket' but is simultaneously part of the six-word cluster 'with his hands in his pocket'.

With the length of the target clusters set for 5 (as shown in Fig. 3), the extraction yielded a total of 5,604,731 tokens with 1,903,410 types of 5-gram clusters.

However, many of the 5-grams had punctuations in them, such as ', then? when serious[? When]', 'not forgotten, but treasured [up in]'. These clusters with punctuations in them are generally not structurally and semantically complete units and they sometimes cross sentence boundaries. As such, they are often not included as clusters in most keyword/key cluster studies (Boudin, Mougard, and Cram 2016; Firoozeh *et al.* 2020). It is important to note that Mahlberg (2007) did include some clusters with punctuations. Yet I decided to follow the generally established practice of not including such structurally and semantically incomplete ngrams as clusters. This exclusion of these ngrams resulted in 2,000,864 tokens and 27,095 types of five-word clusters, that is, it removed 3,603,867 tokens and 1,876,315 types of 5-grams with punctuations in them.

Then, the 27,095 5-grams were further screened by frequency, which was set at five by following Mahlberg's (2007) practice. This frequency-based screening was implemented to help ensure that the remaining 5-grams were indeed clusters that were often used together in the corpus. This screen process led to the removal of 22,832 ngram types, resulting in a total of 4,263 types of five-word clusters, which is listed in a table as Supplementary Material.

3.4 Identifying/ranking key clusters using four different methods

After the extracted ngrams/raw being screened by the aforementioned two steps, I used the following methods to identify and rank the 4,263 key clusters: the *term frequency* method, DP_{norm} dispersion method, *PPMI* association method, and a blended method that I developed and adopted. Since the *frequency*, DP_{norm}, and *PPMI* methods have already been introduced in Sections 2.1.1, 2.1.2, and 2.1.3, respectively above, only the blended method is introduced here.

3.4.1. The new blended method *TF-DP*$_{norm}$ developed for the present study

If we recall the discussion in Section 2.1.4 on blended methods, that is, those of Millar and Budgell (2008), Scott (1997), and Jones (2004), most existing blended methods considered both frequency and dispersion. However, as noted above, these existing methods either use a reference corpus (e.g. Scott 1997; Millar and Budgell 2008) or is designed for identifying keywords/clusters in a document, rather than a corpus. Following these existing blended methods, the blended method used in this study also considers both frequency and dispersion but makes no use of a reference corpus. Specifically, this new blended method identifies and ranks key clusters by using the algorithm that involves the formula of *Term frequency* × DP_{norm}, that is, by multiplying the term frequency of a cluster and its DP_{norm} value. The raw frequency has been processed with Min-max normalization. It is important to report that this blended method was adopted after a series of trials of algorithms with different formulas, such as *frequency*×*range* and *normed frequency*×*DP*, with the results showing that *frequency*×DP_{norm} yielded what appeared to be the best or most accurate and reliable results as will be shown in the Results section below. For clarity and consistency purposes, this new method used in this study is referred to as the *TF-DP*$_{norm}$ blended method.

3.5 Comparing the results of the four methods against those identified by Mahlberg (2007) with the use of a reference corpus

In this last step of the study, I compared the four lists of key clusters produced by the four different methods against the top twenty-one key clusters that Mahlberg (2007) identified by using a reference corpus. Specifically, I checked and marked where each of

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

twenty-one key clusters was located/ranked in each of the five lists and visualized the results using the R package {ggplot2} to highlight their similarities and differences.

4. Results and discussion

The full rank lists of the key clusters produced by each of the five different methods are included in a large table as Supplementary material.

4.1 The top twenty-five key clusters identified/ranked by the five methods

To make it easier for the comparison of the five lists, the top twenty-five clusters in each list are presented in Table 2. A similarity check of these top items among the lists of the five identification/ranking methods show noticeable similarity among the *TF*, DP_{norm}, *TF-*DP_{norm} lists since these three lists have eleven overlapping clusters: 'a quarter of an hour', 'as if he had been', 'at the bottom of the', 'his hands in his pockets', 'I do n't know what', 'in the course of the', 'on the part of the', 'the opposite of the', and 'What do you mean by'. Furthermore, of these three lists, the *TF* list is especially notable since it has additional clusters overlapped with the DP_{norm} and *TF-*DP_{norm} lists respectively. Specifically, the *TP* list has seven additional clusters overlapped with the *TF-*DP_{norm} list: 'as a matter of', 'as if it were a', 'at the end of the', 'I do n't know how', 'I do n't know that', 'I do n't think I', and 'what do you think of'. On the other hand, it also has six additional clusters overlapped with the DP_{norm} list: 'as if it had been', 'in the middle of the', 'on the opposite side of', 'on the other side of', 'the other side of the', and 'with his hands in his'. Given that *TF* is a frequency measure, DP_{norm} is a dispersion measure, and *TF-*DP_{norm} is a blended measure taking both frequency and dispersion into consideration, the observed overlaps among the three lists indicate that, at least in Dickens's novel corpus, high frequency clusters are often rather evenly dispersed across the novels. Hence, the two are very important in identifying key clusters as shown in previous research.

It is also of importance to mention that the lists of the two blended methods (*TF-IDF* and *TF-*DP_{norm}) also share two clusters: 'the father of the Marshalsea' and 'the person of the house', which rank first and second on the *TF-IDF* list and third and eighth on the *TF-*DP_{norm} list. One may wonder why these two items have made the top or close to the top of the two lists. A check of the complete *TF* list reveals that although these two clusters are not in the top 25 most frequent ones, their respective frequencies are still very high, with 'the father of the Marshalsea' ranked 29th and 'the father of the house' ranked 46th although their dispersions are low with each appearing in only one novel. This explains why the two are ranked the top two in *TF-IDF* (a blended measure that, if we recall, is designed to help identify key clusters in a document, not in a corpus of more than one documents) and 3rd and 8th in *TF-*DP_{norm} (a blended measure that takes into account both frequency and dispersion to identify key clusters in a corpus). These results involving the two blended methods show that *TF-*DP_{norm} is a better method than *TF-IDF* for identifying key clusters in a corpus since it does take dispersion into more consideration by not ranking the clusters 1st and 2nd, a point that will be shown more clearly below in the next section on the comparison of the lists with Mahlberg's (2007) reference corpus-based list.

4.2 Comparison of the results of the five methods against the top key clusters in Mahlberg's (2007) reference corpus-based list

It is important to first mention that of the top twenty-five key clusters that Mahlberg (2007) identified in the Dicken's corpus, four are those that include a punctuation in them. As noted earlier, ngrams (potential clusters) with a punctuation in them are excluded in this study due to their semantic and/or structural incompleteness. Hence, Mahlberg's (2007) four clusters with a punctuation are not included in the comparison analysis. In other words, only twenty-one key clusters from Mahlberg's (2007) reference corpus-based list are included as the benchmark for comparison. Table 3 presents the comparison results that consist of the value and rank order of each of Mahlberg's (2007) top 21 clusters in each of the lists yielded by the five identification/rank methods used. Figure 5 visualizes the rankings of Mahlberg's top twenty-one key clusters in each of the five lists.

It is clear from the results in Table 3 and Fig. 4 that the *TF* and *TF-*DP_{norm} lists provide a much better coverage of Mahlberg's (2007) top 21 clusters because all the 21 clusters are ranked within the top 375 and 404 in the two lists respectively while, in the other three lists, some of the twenty-one clusters are ranked as low as 4,005 in DP_{norm}, 2,485 in *PPMI*, and 3,663 in *TF-IDF*. This difference between the two groups of methods is clearly visualized in Fig. 4, in which all Mahlberg's (2007) top twenty-one clusters are located at the extreme left or the higher end of the slop in the *TF* and *TF-*DP_{norm} lists while, in the other lists, they scatter widely along the slope. It is thus safe to state that *TF* and *TF-*DP_{norm} are much better methods than the other three in identifying/ranking key clusters without using a reference corpus.

However, it is imperative to note that a closer examination also reveals that there are variations within the DP_{norm}, *PPMI*, and *TF-IDF* group. While DP_{norm} and

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Table 2. Top 25 key clusters identified/ranked by the five methods.

TF		*DP*$_{norm}$		*PPMI*		*TF-IDF*		*TF-DP*$_{norm}$	
Cluster	**Value**	**Cluster**	**Value**	**Cluster**	**Value**	**Cluster**	**Value**	**Cluster**	**value**
i do n't know what	219	his hands in his pockets	0.28	the dead march in saul	18.44	the father of the marshalsea	61.28	i do n't know what	40.02
i do n't know how	94	with his hands in his	0.30	improved hot muffin and crumpet	18.25	the person of the house	32.73	i do n't know how	19.99
his hands in his pockets	90	in the middle of the	0.32	the prison on the crag	17.93	the lady of the caravan	29.96	the father of the marshalsea	18.69
as if he had been	89	as if it had been	0.35	hollow down by the flare	16.93	mr. pickwick and his friends	25.87	i 'll tell you what	16.37
in the course of the	88	at the top of the	0.35	hot muffin and crumpet baking	16.93	i 'll tell you what	24.77	in the course of the	16.09
what do you mean by	73	as if they had been	0.36	imperfectly repressed by a belt	16.86	gentleman in the white waistcoat	24.51	as if it were a	14.98
as if it were a	72	on the other side of	0.36	the last of the patriarchs	16.67	the gentleman in the white	23.15	as if he had been	14.55
a quarter of an hour	72	the opposite side of the	0.37	of great britain and ireland	16.67	how not to do it	21.79	the person of the house	13.85
i do n't know that	71	to the top of the	0.37	all the king 's knights	16.43	it's of no consequence	21.79	what do you mean by	13.55
the opposite side of the	70	as if he had been	0.37	the ornamental painter's journeyman	16.39	my lovely and accomplished relative	21.79	i do n't know that	13.35
at the bottom of the	66	on the opposite side of	0.37	in the value of taunton	16.30	the six jolly fellowship porters	21.79	what do you think of	13.08
on the part of the	65	a quarter of an hour	0.38	him from top to toe	16.00	not to put too fine	20.43	on the part of the	12.06
what do you think of	63	in the shadow of the	0.38	united metropolitan improved hot muffin	15.97	put too fine a point	20.43	at the bottom of the	11.82
in the middle of the	61	the other side of the	0.39	was only to be equalled	15.92	to put too fine a	20.43	a quarter of an hour	11.74
with his hands in his	60	i should like to know	0.39	crumpet baking and punctual delivery	15.70	fine a point upon it	19.06	the opposite side of the	11.10
as if it had been	58	i do n't know what	0.40	and crumpet baking and punctual	15.58	mr. tupman and mr. snodgrass	19.06	his hands in his pockets	11.09
at the top of the	57	as if she had been	0.41	we will resume our studies	15.52	too fine a point upon	19.06	i do n't want to	11.01
i 'll tell you what	54	at the bottom of the	0.41	in the lap of luxury	15.42	of the name of guppy	17.70	up and down the room	10.60
i do n't think i	54	in the course of the	0.41	at all times and seasons	15.26	man with the wooden leg	15.91	at the end of the	10.35
on the opposite side of	54	in the course of a	0.42	near greta bridge in yorkshire	15.23	of the nuns ' house	14.98	i do n't think i	9.85
at the end of the	52	i do n't know why	0.42	of the six jolly fellowship	15.11	said the elder mr. weller	14.98	as a matter of course	9.81
on the other side of	52	do n't know what you	0.42	on the nineteenth of march	15.10	the child of the marshalsea	14.98	with the air of a	9.58
as a matter of course	51	what do you mean by	0.43	the six jolly fellowship porters	15.06	the youngest gentleman in company	14.98	very much obliged to you	9.32
as much as to say	50	i was going to say	0.43	and gentlemen and honourable boards	14.89	the man with the wooden	14.85	to the best of my	9.16
the other side of the	50	on the part of the	0.43	the member for the gentlemanly	14.65	be so good as to	13.95	as if he were a	9.14

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Table 3. Comparison of the results of the five methods against Mahlberg's (2027) top twenty-one key clusters.

Top twenty-one key clusters identified by Mahlberg with a reference corpus (2007)	*TF*		*DP_{norm}*		*PPMI*		*TF-IDF*		*TF-DP_{norm}*	
	TF value	Rank	*DP_n value*	Rank	*PPMI value*	Rank	*TF-IDF value*	Rank	*TF-DP_n value*	Rank
his hands in his pockets	90	3	0.28	1	12.59	103	5.46	1,045	11.09	16
the father of the marshalsea	45	29	1.00	4,004	13.41	65	61.28	1	18.69	3
the person of the house	37	46	0.93	3,864	8.85	549	32.73	2	13.85	8
do me the favour to	32	64	0.55	167	4.06	1,925	9.04	171	6.96	56
as if he would have	41	36	0.52	110	6.24	1,129	7.61	335	8.80	30
what do you mean by	73	6	0.43	23	5.54	1,314	7.77	324	13.55	9
with his hands in his	60	15	0.30	2	5.49	1,332	3.64	3,665	7.75	45
go so far as to	24	134	0.47	56	4.40	1,735	5.95	882	4.15	204
how do you find yourself	21	178	0.55	135	10.85	261	6.73	595	4.08	215
as if he were a	46	28	0.48	62	1.55	2,531	7.25	400	9.14	25
hands in his pockets and	15	353	0.59	227	1.45	2,538	5.43	1,059	2.76	404
with his hand to his	31	72	0.61	264	5.49	1,330	11.21	65	7.46	50
on the part of mr.	18	215	0.77	881	3.81	2,007	10.50	89	4.69	153
who had by this time	22	157	0.59	187	8.41	655	7.96	307	4.69	152
the lady of the caravan	22	148	1.00	4,005	14.23	36	29.96	3	7.94	43
on the top of his	21	182	0.50	77	3.27	2,195	5.93	891	3.74	244
the old man with a	14	375	0.82	1,391	3.95	1,966	9.28	163	3.44	292
on the part of the	65	12	0.43	25	1.93	2,485	8.53	199	12.06	12
as if he were going	32	63	0.55	166	5.96	1,188	9.04	170	6.96	55
upon my word and honour	25	111	0.70	594	11.61	173	11.47	62	6.55	65
as if it were a	72	7	0.48	64	3.05	2,274	7.66	333	14.98	6

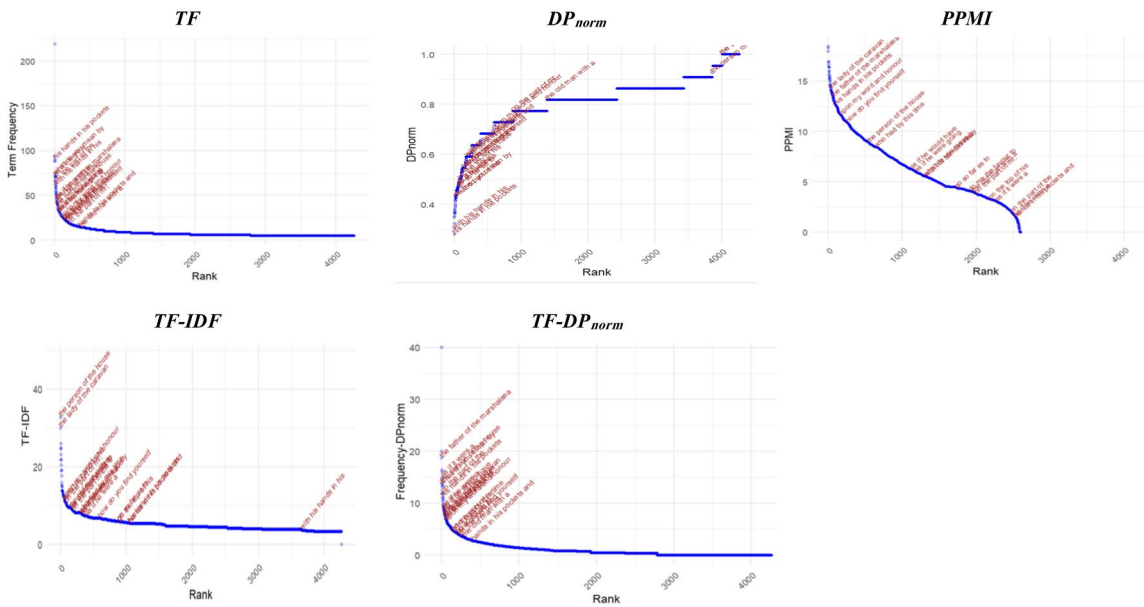

Figure 5. Visualization of the ranking of Mahlberg's (2007) top twenty-one key clusters in the lists yielded by the five identification/ranking methods.

Note: The slope of *DP_{norm}* figure looks the opposite of those of the other four figures because in *DP_{norm}*, the lower the value of a clusters is, the better or wider its dispersion is while in the other four measures, the higher the value of a cluster is, the better it is.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

TF-IDF each have only three (14%) of Mahlberg's (2007) top clusters ranked outside 1,000 (one seventh), *PPMI* has 14 (67%) of Mahlberg's (2007) top twenty-one clusters (two thirds) ranked outside of 1,000. Hence, *PPMI* is the least accurate and effective of the five methods. This result is understandable because as a co-occurrence association strength measure, *PPMI* focuses on how strongly the words in a cluster are associated, that is, how often these words co-occur without considering how frequently the co-occurring cluster appears in the corpus and how widely they are distributed in the corpus. As such, *PPMI* favors co-occurrences involving low-frequency words while disfavouring co-occurrences composed of high-frequency words, as has been found in both previous research (Role and Nadif 2011) and the results of this study. For example, among Mahlberg's (2007) top 21 clusters, 'the lady of the caravan' and 'the father of the Marshalsea' rank the highest on the *PPMI* list (at ranks 36 and 65, respectively), clearly due to 'caravan' and 'Marshalsea' (especially the latter) being low-frequency words. On the other hand, 'as if he were a' and 'on the part of the', two of Mahlberg's (2007) other top clusters, rank very low (2,531 and 2,485 respectively) on the *PPMI* list because they are composed entirely of high frequency words.

6. Conclusion

By comparing the lists of key clusters generated from a corpus of Dickens's novels by five statistical methods (*TF, DP_{norm}, PPMI, TF-IDF*, and *TF-DP_{norm}*) without the use of a reference corpus against the list identified by Mahlberg (2007) with the use of a reference corpus as a benchmark, this study has examined the accuracy and effectiveness of the five statistical methods for identifying key clusters in literary texts with no use of a reference corpus. The results indicate that *TF* and *TF-DP_{norm}* are more accurate and effective than the other three methods while *PPMI* is the least accurate and effective. These results suggest that *frequency* and *dispersion* are more important measures than co-occurrence association and that blended methods that take into sound consideration of both *frequency* and *dispersion* may have the potential to perform better than methods that consider either frequency or dispersion alone, as shown in in the case of my self-developed *TF-DP_{norm}* method.

The results also show that co-occurrence association measures are not particularly relevant in identifying key clusters, especially when no reference corpus is used. This is because, with their tendency to favor low frequency words, association measures often identify and rank highly clusters which have a general low frequency and hence may not play a key role in a text or corpus. This result also supports the findings of previous studies (Deng and Liu 2022; Biber *et al* 2016) that the effectiveness of statistical analysis methods in corpus research often varies across studies with different research purposes/objectives and that selecting which method(s) to adopt in a study depends on the specific purpose of the study.

This study has three limitations, however. First, it examined only literary texts. As a result, the results may not generalize to texts of other genres. Future research will need to investigate the accuracy and effectiveness of statistical methods for identifying key cluster in other genres. Second, this study compared only five specific methods, that is, *TF, DP_{norm}*, and *PPMI* and two blended methods *TF-IDF* and, *TF-DP_{norm}*, with the first three each representing the method types of frequency-based, dispersion-based, and co-occurrence association strength-based. Other methods from the different method types need to be included in future studies. Third, although this study finds my self-develop *TF-DP_{norm}* method to be highly accurate and effective, it is necessary and worthwhile for future research to explore and develop other blended methods.

Finally, despite the aforementioned limitations, this study, being the first one on comparing the accuracy and effectiveness of different statistical methods for identifying key clusters without the use of a reference corpus, has enhanced our understanding of the various methods used in the identification and ranking of key clusters. As a result, this study has made an important contribution to the research on keyword/key cluster identification and analysis.

Author contributions

Hongwei Zhan (Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Visualization, Writing—original draft, Writing—review & editing)

Supplementary data

Supplementary data is available at *DSH* online.

Funding

None declared.

References

Aizawa, A. (2003) 'An Information-Theoretic Perspective of TF-IDF Measures', *Information Processing and Management*, 39: 45–65.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Baker, P. (2004) 'Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis', *Journal of English Linguistics*, **32**: 346–59.

Baker, P (2009) 'The Question is, How Cruel is it?' keywords, Fox Hunting and the House of Commons', in D. Archer (ed.) *What's in a Word-list? Investigating Word Frequency and Keyword Extraction*, pp. 125–36. Aldershot: Ashgate.

Biber, D., Conrad, S., and Cortes, V. (2004) 'If You Look At: Lexical Bundles in University Teaching and Textbooks', *Applied Linguistics*, **25**: 371–405.

Biber, D. *et al.* (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D. *et al.* (2016) 'On the (Non)utility of Juilland's *D* to Measure Lexical Dispersion in Large Corpora', *International Journal of Corpus Linguistics*, **21**: 439–64.

Bondi, M. (2010) 'Perspectives on Keywords and Keyness: An Introduction', in M. Bondi, and M. Scott (eds) *Keyness in Texts*. Amsterdam: John Benjamins.

Boudin, F., Mougard, H., and Cram, D. (2016) 'How Document Pre-Processing Affects Keyphrase Extraction Performance'. https://doi.org/10.48550/arXiv.1610.07809

Bullinaria, J. A., and Levy, J. P. (2007) 'Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study', *Behavior Research Methods*, **39**: 510–26.

Carter, R., and McCarthy, M. (2006) *Cambridge Grammar of English. A Comprehensive Guide*. Cambridge: Cambridge University Press.

Church, K. W., and Hanks, P. (1990) 'Word Association Norms, Mutual Information, and Lexicography', *Computational Linguistics*, **16**: 22–9.

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**: 29–59.

Deng, Y., and Liu, D. (2022) 'A Multi-Dimensional Comparison of the Effectiveness and Efficiency of Association Measures in Collocation Extraction', *International Journal of Corpus Linguistics*, **27**: 191–219.

Dunning, T. (1993) 'Accurate Methods for the Statistics of Surprise and Coincidence', *Computational Linguistics*, **19**: 61–74.

Egbert, J., and Biber, D. (2019) 'Incorporating Text Dispersion into Keyword Analyses', *Corpora*, **14**: 77–104. https://doi.org/10.3366/cor.2019.0162

Firoozeh, N. *et al.* (2020) 'Keyword Extraction: Issues and Methods', *Natural Language Engineering*, **26**: 259–91. https://doi.org/10.1017/S1351324919000457

Fischer-Starcke, B (2009) 'Keywords and Frequent Phrases of Jane Austen's Pride and Prejudice', *International Journal of Corpus Linguistics*, **14**: 492–523.

Gabrielatos, C. (2018) 'Keyness Analysis: Nature, Metrics and Techniques', in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse*, pp. 225–58. Routledge. https://www.routledge.com/

Gries, S. T. (2008) 'Dispersions and Adjusted Frequencies in Corpora', *International Journal of Corpus Linguistics*, **13**: 403–37.

Gries, S. T. (2010) 'Dispersions and Adjusted Frequencies in Corpora: Further Explorations', in S. T. Gries, S. Wulff, and M. Davies (eds) *Corpus Linguistic Applications: Current Studies, New Directions*, pp. 197–212. Amsterdam: Rodopi.

Gries, S. T. (2020) 'Analyzing Dispersion', in M. Paquot and S. T. Gries (eds) *A Practical Handbook of Corpus Linguistics*, pp. 99–118. Cham, Switzerland: Springer.

Gries, S. T. (2021) 'A New Approach to (Key) Keywords Analysis: Using Frequency, and Now also Dispersion', *Research in Corpus Linguistics*, **9**: 1–33.

Gries, S. T. (2024). *Frequency, Dispersion, Association, and Keyness: Revising and Tupleizing Corpus-Linguistic Measures*. Amsterdam: John Benjamins.

Hanks, P. (2004) 'Corpus Pattern Analysis', in *Proceedings of Euralex 2004*, pp. 87–97. France: Lorient.

Hori, M. (2004) *Investigating Dickens' Style. A Collocational Analysis*. Basingstoke: Palgrave Macmillan.

Jones, K. S. (2004) 'A Statistical Interpretation of Term Specificity and its Application in Retrieval', *Journal of Documentation Volume*, **60**: 493–502. https://doi.org/10.1108/00220410410560573

Juilland, A., and Chang-Rodriguez, E. (1964) *Frequency Dictionary of Spanish Words*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783112415467

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

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, CL2009, Liverpool.

Lyne, A. A. (1985).*The Vocabulary of French Business Correspondence: Word Frequencies, Collocations and Problems of Lexicometric Method*. Genève: Slatkine-Champion.

Mahlberg, M. *et al.* (2019) 'Speech-Bundles in the 19th-Century English Novel', *Language and Literature*, **28**: 326–53.

Mahlberg, M. (2007) 'Clusters, Key Clusters and Local Textual Functions in Dickens', *Corpora*, **2**: 1–31.

Mahlberg, M., Smith, C., and Preston, S. (2013) 'Phrases in Literary Contexts: Patterns and Distributions of Suspensions in Dickens's Novels,' *International Journal of Corpus Linguistics*,**18**: 35–56.

McEnery, T., and Hardie, A. (2011)*Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Millar, N., and Budgell, B. S. (2008) 'The Language of Public Health—A Corpus-Based Analysis', *Journal of Public Health*, **16**: 369–74.

Niwa, Y., and Nitta, Y. (1994) 'Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries', in *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*. COLING 1994, Kyoto, Japan. https://aclanthology.org/C94-1049/

Paquot, M., and Bestgen, Y. (2009) 'Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction', in A. H. Jucker, D. Schreier, and M. Hundt (eds), *Corpora: Pragmatics and Discourse*, pp. 247–69. Leiden: Brill.

Pojanapunya, P., and Todd, R. W. (2018) 'Log-Likelihood and Odds Ratio: Keyness Statistics for Different Purposes of Keyword Analysis', *Corpus Linguistics and Linguistic Theory*, **14**: 133–67.

Rayson, P. (2022) 'Corpus Analysis of Key Words', in C. A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, pp. 1–7. New Jersey: John Wiley & Sons, Ltd.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

Robinson, D. (2021) 'gutenbergr: Download and Process Public Domain Works from Project Gutenberg', R package version 0.2.1, https://CRAN.R-project.org/package=gutenbergr.

Role, F., and Nadif, M. (2011) 'Handling the Impact of Low Frequency Events on Co-occurrence based Measures of Word Similarity—A Case Study of Pointwise Mutual Information', in J. Filipe and A. L. N. Fred (eds), *KDIR 2011—Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, Paris, France, 26–29 October, 2011, pp. 226–31. Setúbal: SciTePress.

Rose, S. *et al.* (2010) 'Automatic Keyword Extraction from Individual Documents', in M. W. Berry and J. Kogan (eds) *Text mining: Applications and theory*, pp. 1–20. New Jersey: John Wiley & Son Ltd.

Salton, G., and Buckley, C. (1988) 'Term-Weighting Approaches in Automatic Text Retrieval', *Information Processing and Management*, **24**: 513–23.

Salton, G., Yang, C. S., and Yu, C. T. (1975) 'A Theory of Term Importance in Automatic Text Analysis', *Journal of the American Society for Information Science*, **26**: 33–44. https://doi.org/10.1002/asi.4630260106

Schneider, G. (2024) 'Digital Dickens: An Automated Content Analysis of Charles Dickens' Novels', in S. Buschfeld ***et al.*** (eds) *Crossing Boundaries through Corpora*, pp. 62–98. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/scl.119.04sch

Scott, M. (1997) 'PC Analysis of Keywords-and Key Keywords', *System*, **25**: 233–45.

Scott, M. (2010) 'Problems in Investigating Keyness, or Clearing the Undergrowth and Marking out Trails', in M. Bondi and M. Scott (eds) *Keyness in Texts*, pp. 43–58. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.41.04sco

Siddiqi, S., and Sharan, A. (2015) 'Keyword and Keyphrase Extraction Techniques: A Literature Review', *International Journal of Computer Applications*, **109**: Article 2. https://doi.org/10.5120/19161-0607

Sönning, L. (2024) 'Evaluation of Keyness Metrics: Performance and Reliability', *Corpus Linguistics and Linguistic Theory*, **20**: 263–88.

Stubbs, M. (2010) 'Three Concepts of Keywords', in M. Bondi and M. Scott (eds) *Keyness in Texts*, pp. 21–42. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.41.03stu

Wijffels, J. (2022) 'udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit', R package version 0.8.9, https://CRAN.R-project.org/package=udpipe.

Downloaded from https://academic.oup.com/dsh/article/40/2/668/8124013 by Hangzhou Teacher's College user on 31 May 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of EADH. All rights reserved.
For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Digital Scholarship in the Humanities, 2025, 40, 668–681
https://doi.org/10.1093/llc/fqaf031
Full Paper