This is an Accepted Manuscript of an article published by De Gruyter in *Corpus Linguistics and Linguistic Theory* on 27 April 2023, available at https://doi.org/10.1515/cllt-2022-0116

Evaluation of keyness metrics: Performance and reliability

Lukas Sönning (University of Bamberg)

Abstract. The methodological debates surrounding keyword analysis have given rise to a wide range of keyness metrics. The present paper delineates four dimensions of keyness, which distinguish between frequency- and dispersion-related perspectives. Existing measures are then organized according to these dimensions and evaluated with regard to their performance on a specific keyword analysis task: The identification of key verbs in academic writing. To this end, the rankings produced by 32 different metrics are evaluated against an established academic word list. Further, the reliability of measures is assessed, to determine whether they produce stable rankings across repeated studies on the same pair of text varieties. We observe notable differences among metrics with regard to these criteria. Our findings provide further support for the superiority of the Wilcoxon rank sum test and text-dispersion–based measures, and allow us to identify, within each dimension of keyness, metrics that may be given preference in applied work.

Keywords: keyness, keyword analysis, frequency, dispersion, methodology

1. Introduction

The purpose of a keyness analysis is to identify (lexical) items that are typical of a particular text variety. This text variety is represented by what is referred to as the target corpus, and typical items are then detected by way of comparison to a reference corpus, which serves as a baseline to foreground relevant structures. While keyword analysis (Scott 1997) has become an essential tool in corpus-based work (for a recent review see Gabrielatos 2018), the question of how to go about identifying relevant items has been subject to considerable methodological debate. As a result, researchers can choose from a variety of metrics, i.e. computerized ways of detecting and ranking items. The default statistical maneuver, a likelihood-ratio test, has met with justified criticism due to its statistical inadequacy (e.g. Kilgarriff 2005). Some have rightly argued that keyness should be expressed using descriptive (rather than inferential) statistics, to reflect the actual degree to which an item is overrepresented in the target corpus (e.g. Gabrielatos and Marchi 2012; Hardie 2014). Finally, text-level analyses have gone further still by accounting for the fact that corpora consist of text samples (e.g. Brezina and Meyerhoff 2014; Lijffijt et al. 2014). This has also opened up novel perspectives on keyness, which concentrate on the pervasiveness of an item across texts rather than a comparison of its overall frequency (Egbert and Biber 2019), or a combination of these two aspects (Gries 2021).

The aim of the present paper is to contribute to the methodological literature in two ways: First, we consider keyness as a multidimensional construct, and build on Egbert and Biber

(2019) and Gries (2021) to identify four dimensions of keyness. While these dimensions have mostly been recognized only implicitly in previous work, they are grounded in the actual practice of keyword analysts and reflect recent methodological impulses (Egbert and Biber 2019; Gries 2021). These keyness dimensions emphasize different aspects of typicalness, which allows us to organize metrics on linguistic grounds and make informed choices among metrics in applied work. The second aim of the paper is to evaluate the performance of 32 metrics on a specific keyword analysis task: the identification of key verbs in academic writing. This evaluation is based on two criteria. The first considers the quality of the ranking produced by each metric, i.e. whether those verbs that are typical of academic writing do in fact surface at (or near) the top of the list. Our second criterion, reliability, reflects the degree to which a metric produces stable and replicable rankings across studies. This feature can be read from the sample-to-sample variability of keyword lists – that is, how much they fluctuate over different samples (i.e. corpora) from the same pair of text varieties.

The remainder of this article is structured as follows. Section 2 outlines different dimensions of keyness and the role of keyness metrics in keyword analysis. Section 3 then offers a systematic overview of existing metrics, dealing in turn with descriptive and inferential measures. After a brief review of existing approaches to the comparative evaluation of metrics (Section 4), Section 5 turns to our performance and reliability study. Section 6 closes with a summary and discussion of our findings.

2. Keyness analysis: Background

In the 25 years or so since keyword analysis was first introduced into corpus linguistics (Scott 1997), the notion of keyness has evolved into a multidimensional construct. This section sets the scene by delineating four dimensions of keyness that have (implicitly) informed corpus-based work (Section 2.1). We then discuss the role of keyness metrics in keyword analysis (Section 2.2).

2.1. Dimensions of keyness

Keyness analysis is used to identify items that are *typical* of a particular text variety. Traditionally, typicalness has been related to frequency of occurrence. The emphasis has therefore been on identifying forms that are used at a higher rate in the target (compared to the reference) corpus (e.g. Scott 1997), and this is reflected in current definitions[1] of the term "keyword". Egbert and Biber (2019: 78-79) extended the notion of keyness by considering two aspects of typicalness: content-distinctiveness and content-generalizability. A key item is *distinctive* of the target variety in two senses: It is (strongly) associated with it, and, in terms of topicality, it signals its aboutness. This first aspect reverberates partly in the traditional approach to keyness, since the distinctiveness of an item surfaces in a higher

[1] For instance, McEnery and Hardie (2012: 245) define a keyword as "[a] word that is more frequent in a text or corpus under study than it is in some (larger) reference corpus, where the difference in frequency is statistically significant."

usage rate in the target corpus. The second aspect accentuates how broadly an item is used in the target variety. A key item should have considerable generality and occur across a sufficiently wide range of texts (see also Baker 2004). This second aspect is linked to the corpus-linguistic notion of dispersion.[2]

We will adopt Egbert and Biber's (2019) criteria and distinguish, on methodological grounds, between frequency-oriented and dispersion-oriented approaches to keyness (see also Gries 2021). Note that these are complementary perspectives, since they capture two different, linguistically meaningful aspects of typicalness. In addition, we will make a distinction between keyness features that are reflected in attributes of the target variety as such, and contrastive features, which emerge when comparing the target variety against a reference variety. Table 1 gives an overview of the resulting four-way classification, which relates methodological choices (frequency vs. dispersion; isolated vs. comparative analysis) to the linguistic meaning carried by metrics. Its purpose is to systematize existing ways of measuring keyness (i.e. keyness metrics), and it therefore draws heavily on earlier methodological work (Baker 2004; Egbert and Biber 2019; Gries 2021).

Let us now consider these four dimensions in turn, starting with frequency of use. If an item is to be considered typical of the target variety, its occurrence rate (i.e. normalized frequency) should be sufficiently high, to make it a *discernible* feature of this text variety. Frequency-oriented measurements of typicalness can also be made contrastively, i.e. by comparison to a reference corpus. Thus, a key item should also be a distinctive feature of the target variety, in the sense that it is used at a higher rate than in the reference variety. Note that while discernibility and distinctiveness both rely on frequency of occurrence, they measure different aspects of typicalness.

Table 1. Dimensions of keyness.

Analysis	Frequency-oriented	Dispersion-oriented
Target variety in isolation	Discernibility of item in the target variety	Generality across texts in the target variety
Comparison to reference variety	Distinctiveness relative to the reference variety	Comparative generality relative to the reference variety

Dispersion-oriented measures, on the other hand, capture the *generality* of an item, i.e. (i) its breadth of usage across the text variety and/or (ii) the evenness of its distribution across texts. Thus, we would expect a typical item to be in broad use, with its probability of occurrence ideally being roughly balanced across texts. Generality can also be considered contrastively: A key item should be more pervasive and/or balanced across target (vs. reference) variety texts. We will refer to this as the *comparative generality* of an item.

[2] A procedure that also incorporates the notion of generality is key keyword analysis (see Scott 1997).

2.2. Keyness metrics

To aid the linguist with the identification of typical items in the target variety, keyword extraction tools produce an ordered list of forms, with more promising candidates ranking higher. These lists will usually include false hits, and the user must therefore manually identify elements they consider relevant. To optimize the set of candidates (and thereby reduce the amount of manual post-processing work), several parameters of the extraction algorithm can be manipulated. The linchpin, however, is the choice of metric, which determines the ranking of items. A good keyness metric will background items that are unlikely to be of interest, which obviates the need for informed yet subjective decisions by the researcher (cf. Egbert and Biber 2019: 81).

As we will see, a wide range of metrics have been proposed and used in previous work. It is helpful to group these based on the dimensions of keyness we have just outlined (see Section 2.1), as these offer substantive guidance for the choice among them. However, keyness metrics can also be classified on statistical grounds, into descriptive and inferential measures. Descriptive metrics capture distributional features of an item in the corpora under study. Inferential measures, on the other hand, quantify the statistical uncertainty arising from the fact that a corpus is just a sample from the text variety of interest. This is usually communicated using a *p*-value or test statistic, as explained in more detail below. Within each of the dimensions listed in Table 1, both types of measures are (in principle) available.

To summarize, when choosing among keyness metrics, we must first decide which dimension(s) of keyness we wish to examine. This decision is informed by linguistic considerations. Methodological recommendations can then be considered if various metrics are available to express a particular facet of typicalness.

3. A survey of keyness metrics

A wide range of keyness metrics have been proposed in the literature and used in applied work. This section provides a fairly extensive survey, which organizes metrics on statistical (descriptive vs. inferential) and linguistic grounds (dimensions of keyness). Table 2, which serves as an advance organizer, lists all metrics that are part of our survey and performance evaluation. To give an impression of how often these metrics are used in applied work, Table 2 also reports the results of a review of 47 research articles that include a keyness analysis.[3] The figures indicate the number of studies that used a specific metric as a ranking device, or to set thresholds (numbers in parentheses). It is immediately apparent that the vast majority of keyness analyses default to distinctiveness metrics for ranking candidate items[4], with about 80% of the studies in our sample relying on the likelihood-ratio test.

---

[3] Details on this survey can be found in web appendix 1 (https://osf.io/fhcrd).

[4] These findings are consistent with those of Pojanapunya and Watson Todd's (2018) review of 30 keyness analyses published between 2002 and 2013. Out of 20 articles that specified the metrics used, 13 relied on the likelihood-ratio test and 7 used the chi-square test to rank items; thresholds were based on the LR test (7), the occurrence rate (3), and *TD* (1).

Table 2. Overview of keyness dimensions and their associated metrics.

Dimension	Keyness metric	Reference[†]	Survey[‡]
Discernibility	Descriptive		
	◐ Occurrence rate		1 (11)
Distinctiveness	Descriptive		
	◐ Rate ratio	Kilgarriff 2009	4
	◐ Rate difference		
	● Probability of superiority (*PS*)		
	◐ Log ratio	Hardie 2014	2
	◐ Difference coefficient	Hofland & Johansson 1982	2
	◐ %DIFF	Gabrielatos & Marchi 2011	2
	◐ Odds ratio (OR)	Pojanapunya & Watson Todd 2018	2
	○ Signed D_{KL}	Gries 2021	1
	Inferential		
	○ Chi-square test	Hofland & Johansson 1982	1
	○ Likelihood-ratio test	Dunning 1993	3 (13) 7
	● Wilcoxon rank sum test	Kilgarriff 1996	
	● *t*-test	Kilgarriff 1996	
	○ BIC	Wilson 2013	2
Generality	Descriptive		
	◐ Range	Rayson 2003	1 (3)
	● *TD*	Egbert & Biber 2019	1 (2)
	● D_{KL}	Gries 2021	1
	● $D/S_{adj}/D_2/D_P/D_A$		
Comparative generality	Descriptive		
	● *TD* ratio	Egbert & Biber 2019	1
	● *TD* difference		
	● D_{KL} difference	Gries 2021	1
	● $D/S_{adj}/D_2/D_P/D_A$ difference		
	Inferential		
	● *TD*-based likelihood-ratio test	Egbert & Biber 2019	1

Note. Key to symbols (cf. Section 3.2): ○ metric based on a bag-of-words analysis; ● metric based on a text-level analysis; ◐ metric has been proposed as a bag-of-words measure, but both kinds of analyses are possible.

[†]These references indicate where (to our knowledge) the metric first appeared in the literature on keyness analysis or word frequency comparisons.

[‡]These figures are the number of articles in our survey (n = 47 in total) that use the metric as a ranking device; numbers in parentheses denote the number of studies that used the metric to define a threshold; only 19 studies reported using a threshold. Underlining indicates counts that include the methodological paper where the metric was originally proposed.

To illustrate the rationale underlying different metrics, we will use a small set of hypothetical data, with 10 texts each in the reference and target corpus. All of these imaginary texts contain 2,000 words and the text-level token counts for an item of interest are as follows:

- Target corpus: 0 0 4 6 6 10 18 20 30 36 (130 instances in total)
- Reference corpus: 0 0 0 0 0 2 2 2 6 10 (22 instances in total)

Since each text contains 2,000 words, dividing these token counts by 2 yields a rate of occurrence (i.e. normalized frequency) expressed in per thousand words (ptw) of running text.

Let us now illustrate the metrics in our survey using this hypothetical data set. We first consider descriptive measures (Section 3.1), and group them along the four dimensions of keyness. Section 3.2 then discusses inferential measures and Section 3.3 considers the question of how to choose among metrics.

3.1. Descriptive metrics

Metrics that describe and summarize distributional features of an item in a corpus are referred to as descriptive metrics. We can group them according to the dimensions of keyness they address and distinguish between frequency- and dispersion-oriented metrics.

3.1.1. Frequency-oriented metrics: Discernibility and distinctiveness

To measure the discernibility of an item, its usage rate in the target variety is expressed in relative terms, i.e. as a normalized frequency. Taking into account corpus structure (see Section 3.3), it can be calculated at the text level, using a measure of central tendency to summarize the set of observed text-specific rates. Figure 1 shows our illustrative data, where each dot denotes a text. A simple average over the 10 text-specific rates gives an occurrence rate of 6.5 ptw in the target corpus.

In applied work, the occurrence rate of items (or the observed absolute frequency) has primarily been used to set cut-offs for the inclusion of forms rather than as a ranking device (cf. Table 2). The choice of threshold depends on the context and purpose of the analysis. For our case study, we could consider, as a prototypical instance of this text variety, a research article of about 10,000 words and reason that a key verb should, on average, occur at least once in a document of this length. This corresponds to a rate of 1/10,000 (i.e. 0.1 ptw), and our average of 6.5 ptw would therefore indicate sufficient discernibility.

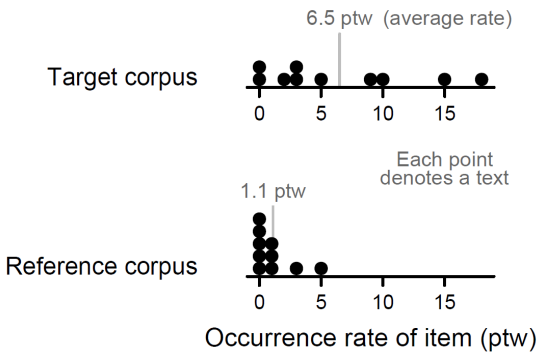

Figure 1. A hypothetical data set, which we use to illustrate the calculation and meaning of keyness metrics. ⓒⓘ[5]

To quantify the distinctiveness of an item in the target variety, occurrence rates can be compared in absolute terms (e.g. a rate difference, measured in points on the ptw scale), or in relative terms (as a rate ratio; Kilgarriff 2009). How much more prevalent an item has to be in order to be considered typical will again depend on our research task. Here, we could state that a key verb in academic writing must occur at least twice as often in the target corpus. In our illustrative data, we note a rate ratio of 5.9 (6.5/1.1), which is certainly indicative of a higher currency in academic writing. The log-transformed version of the rate ratio has also appeared in the literature on keyword analysis (Hardie 2014); on the $\log_2$ scale, our rate ratio of 5.9 would be 2.56 (cf. Table 3). Considering these two options, an advantage of the untransformed ratio is its direct interpretability.

A measure that does not seem to have been adopted for keyness analysis is the probability of superiority (*PS*). We include it into our survey because of its direct connection to the Wilcoxon rank sum test, an inferential procedure that has been evaluated favorably in the methodological literature on keyword analysis (e.g. Paquot and Bestgen 2009; Lijffijt et al 2014; Brezina and Meyerhoff 2014). *PS* gives the probability that a target variety text shows a higher occurrence rate than a text from the reference variety (see Grissom and Kim 2012: 149-151). Keyness candidates should show *PS* values above .50, and values closer to 1 signal greater distinctiveness.

Other measures have been proposed to express differences in frequency. These include the difference coefficient, %DIFF, the odds ratio[6], and Signed D_{KL}. A disadvantage of these measures is that they are more difficult to interpret (see Table 3), and may be less meaningful and accessible to a linguistic audience (cf. Egbert, Larsson and Biber 2020: 24-32). It may then be difficult to recognize which values provide relatively strong or weak indications of keyness, and to make informed decisions about cut-offs.

---

[5] Images with the symbols ⓒⓘ in the figure caption have been published under the Creative Commons Attribution 4.0 license (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project at https://osf.io/kcwus/

[6] For items with an occurrence rate below 10% (i.e. all lexical words), the odds ratio is very close to the rate ratio (see Zhang and Yu 1998).

Table 3 gives an overview of distinctiveness metrics, listing their possible range of values, the score they produce for our exemplary data, and their meaning. The order of metrics reflects the degree to which we consider them capable of interpretation by a wider audience, with the most transparent measure appearing at the top.[7] It should be noted that five of these metrics (rate ratio, difference coefficient, %DIFF, odds ratio, log ratio) produce the same rank order of items in a keyword list (see Gabrielatos 2018: 237). Further, open circles ○ in Table 3 denote metrics that are unable to handle situations where the item of interest does not occur at all in the reference corpus.[8]

Table 3. Distinctiveness (i.e. difference-in-frequency) metrics for our hypothetical set of data.

Metric		Range	Value	Meaning
Rate ratio	○	[0; +∞]	5.9	The item is **5.9** times more frequent in the target corpus.
Rate difference	●	[−∞; +∞]	5.4	On the ptw scale, the frequency in the target corpus is **5.4** points higher.
PS	●	[0; 1]	.80	The probability that a randomly sampled text from the target variety has a higher rate than a randomly sampled text from the reference variety is **.80** (or 80%)
Difference coefficient	●	[−1; +1]	.71	The rate difference (5.4 points on the ptw scale) amounts to a share of **.71** (or 71%) of the sum of the rates (7.6 ptw).
%DIFF	○	[−100; +∞]	491	The rate difference (5.4 points on the ptw scale) amounts to **491**% of the rate in the reference corpus (1.1 ptw).
Odds ratio	○	[−∞; +∞]	5.94	The odds of encountering the item are **5.94** times higher in the target corpus.
Log ratio				
Base 2	○	[−∞; +∞]	2.56	Rate ratio = $2^{\mathbf{2.56}}$ = 5.9
Base e	○	[−∞; +∞]	1.78	Rate ratio = $e^{\mathbf{1.78}}$ = 5.9
Base 10	○	[−∞; +∞]	0.77	Rate ratio = $10^{\mathbf{0.77}}$ = 5.9
Signed D_{KL}	●	[−1; +1]	+.28	Positive values indicate that the item is overrepresented in the target variety.

Note. The column "value" gives the score for our hypothetical set of data (see Figure 1).
Empty circles ○ denote metrics that, without data modification, cannot handle items that are absent from the reference corpus.

[7] See web appendix 3 for the rationale underlying the interpretability ranking of keyness metrics (https://osf.io/wbnzv).
[8] A workaround is to augment the data by adding imaginary counts to each corpus (see Kilgarriff 2009 for a discussion of various options focusing on the rate ratio). For a text-level analysis, we would add to each corpus one hypothetical text with a pre-defined, low occurrence rate.

3.1.2. Dispersion-oriented metrics: Generality and comparative generality

Within the context of corpus linguistics, dispersion indicates how widely and how evenly an item is distributed throughout a corpus (e.g. Gries 2020: 99). A highly dispersed form occurs regularly across text files – its status in the target variety is "generalizable" (Egbert and Biber 2019: 79). In contrast, low dispersion levels reflect occurrence in very few places of the corpus[9]; the distribution is then said to be concentrated or "bursty" (Church and Gale 1995: 168). Most dispersion measures are (or can be) standardized to extend from 0 (bursty or highly concentrated distribution) to 1 (widely dispersed, even distribution). Higher values therefore signal greater generality.

Nearly all dispersion measures require a corpus to be divided into parts. These parts can be linguistically meaningful units such as text files, or they can be equal-sized corpus sectors, i.e. arbitrary partitions that cut across or conflate such units. While the latter approach is dominant in software tools and (therefore) applied work, linguistically defined subdivisions are recommended in the methodological literature[10] (see Gries 2008: 420; Egbert and Biber 2019: 77; Egbert, Burch and Biber 2020: 89–90). In what follows, we will consider the sampling units, i.e. text files, as the relevant corpus parts for dispersion analysis.

Dispersion can be measured in different ways (see Gries 2008, 2020 for an overview of different measures). A simple index is the share of texts that feature at least one occurrence of the item (Egbert and Biber 2019; see also Baker 2004). We will follow Egbert and Biber (2019) and refer to this measure as *text dispersion* (*TD*). In our illustrative target corpus, *TD* is at .80 (8 out of 10 texts). *TD* is a useful metric due to its direct interpretability.[11] Considering guidelines for its interpretation, we could reason as follows for our case study: If a verb occurs in fewer than 10% of research articles, it may not be truly characteristic of this text variety.

A similar measure is *Range* (Rayson 2003: 93), which gives the number of corpus parts in which an item appears. Since *TD* is a relative frequency and by definition calculated at the level of the text files constituting the corpus, it makes sense to distinguish it from Range. Note that the two measures produce the same keyness ranking if Range is calculated based on text files (rather than equal-sized corpus parts).

While *TD* only considers the presence/absence of an item in text files, more fine-grained metrics compare occurrence rates across texts. While most measures were originally constructed for equal-sized corpus parts, they can be generalized to apply to units (i.e. texts)

---

[9] Note, however, that the dispersion measures proposed by Gries (D_P, D_{KL}) deviate from the convention in the literature and assign a value of 0 to a perfectly even distribution.

[10] From the viewpoint of corpus design, a division into arbitrarily defined corpus parts, and thus an indifference to the unit of sampling (i.e. texts), appears questionable. It also negatively affects the interpretability of estimates, since different speakers, texts, genres, or points in time may have been combined, producing fuzzy, potentially ambiguous quantities. Opinions may also vary on the appropriate number of corpus parts that should be formed, and it may be difficult to reproduce results from other studies if the partitions cannot be reconstructed exactly.

[11] A disadvantage is its sensitivity to the length of text files – the chance of observing a particular item is higher in longer documents. *TD* therefore partly depends on the average text length in a corpus.

of different length (cf. Gries 2020; Egbert, Burch and Biber 2020). For an assessment of the generality of an item, we will consider the following measures in addition to *TD*:

- *D* (Juilland et al. 1970; we use the formula given in Gries 2020: 103),
- D_2 (Carroll 1970),
- S_{adj} (Rosengren 1972; we use the formula given in Gries 2020: 103)
- D_P (Gries 2008; we use the modified formula given in Egbert, Burch and Biber 2020: 99),
- D_A (Wilcox 1973; we use the formula given in Egbert, Burch and Biber 2020: 98), and
- D_{KL} (Gries 2020, 2021).

Dispersion measures can also be used to contrast the level of generality in the target and reference variety. For purposes of comparison, we can calculate the difference (target corpus minus reference corpus); positive values then reflect greater generality in the target variety. If we rely on *TD*, for instance, this produces a difference between proportions (*TD* difference). Note that Egbert and Biber (2019: 90) focus on the relative difference, i.e. the *TD* ratio.[12] We will consider both variants. For the remaining measures, we can also compute absolute differences (see Gries 2021: 21). Given the standardized scaling of dispersion measures, these will range between −1 and +1, with positive values indicating greater dispersion, or generality, in the target corpus.

To conclude our discussion of dispersion measures, Table 4 reports (comparative) scores for our illustrative target and reference corpus. All measures suggest that the item is more widely dispersed in the target variety, yielding somewhat similar differences (ranging from +.15 to +.30). Corpus-specific values are quite distinct, however: for the target corpus, for instance, they range from .44 (D_A) to .80 (D_2).

Table 4. Generality (i.e. dispersion) metrics for our hypothetical set of data.

Metric		Target corpus	Reference corpus	Difference
D	Juilland et al. 1970	.69	.52	+.17
D_2	Carroll 1970	.80	.59	+.21
S_{adj}	Rosengren 1972	.70	.44	+.26
D_A	Wilcox 1973; Egbert, Burch and Biber 2020	.44	.24	+.20
D_P	Gries 2008; Egbert, Burch and Biber 2020	.56	.41	+.15
D_{KL}	Gries 2021	.52	.26	+.26
TD	Egbert and Biber 2019	.80	.50	+.30

[12] The interpretation of both the *TD* difference and the *TD* ratio should be alert to the sensitivity of *TD* to the length of texts. If the corpora differ in average text length, the *TD* score for the corpus with longer text files will be upwardly biased, which could be considered as yielding a misrepresentation of the actual difference. The ranking of items, however, is not systematically altered.

The keyness metrics we have discussed so far are descriptive – they capture distributional features of the data with no reference to statistical error probabilities.

3.2. Inferential metrics

Inferential keyness metrics express the statistical uncertainty that arises from sampling variation. The idea is to prioritize items for which we find a statistically dependable difference that is unlikely to be an artefact of the particular samples (i.e. corpora) at hand. Inferential metrics express statistical dependability with an error probability (a *p*-value[13]), which is used to rank candidates and/or set exclusionary cut-offs. Alternatively, a so-called test statistic can be used, which produces the same ranking. To obtain error probabilities, we must rely on a statistical model, which consists of a set of assumptions about the data. These assumptions can be understood as prerequisites for valid error probabilities. If they are untenable in light of our data, so are the resulting *p*-values. We can, in general, distinguish two types of statistical model: a bag-of-words model[14] and a text-level model (see Evert 2006; Baroni and Evert 2009; Lijffijt et al. 2014).

3.2.1. Bag-of-words model

The bag-of-words model considers a corpus as an unstructured bag of words, with each word having been sampled randomly and independently from the language variety. We could collect data in a way that will approximate this model, by first sampling texts and then drawing (only) one word from each text (cf. Evert 2006: 182). Corpora, however, are compiled by sampling texts, i.e. (excerpts from) coherent spoken or written material (see Baroni and Evert 2009: 796). In these stretches of text, the choice and arrangement of words will follow regular patterns that violate assumptions of independence or "randomness" (cf. Kilgarriff 2005). Due to, say, differences in style or topicality, the words occurring in a given text cannot be considered independent (see Winter and Grice 2021).

Despite its inadequacy for corpus data, the bag-of-words model forms the basis of the most commonly used inferential procedure, the likelihood-ratio test (cf. Table 2). Further, the chi-square test and a recently proposed metric, the BIC-based approximation to Bayes factors, also rely on this kind of data representation. Due to the mismatch between data and assumptions, the statistical error probabilities (or evidence measures) of these metrics are dubious. Since corpora are not compiled in accordance with the bag-of-words model, we

---

[13] *P*-values are conditional probability statements tied to the null hypothesis significance testing framework. For keyword analysis, the null hypothesis for a specific item would state that it is *not* a key item – it is equally typical of both text varieties. The probabilitiy *p*, then, is a conditional statement. It answers the following question: If the null hypothesis is true (condition 1), and if the statistical model we are using is adequate for the data at hand (condition 2), what is the probability of observing a difference in keyness at least this large, i.e. a difference that is at least as large as the one we have observed between our reference corpus and our target corpus?

[14] Analyses based on the bag-of-words model have also been referred to as "aggregate data methodology" (Brezina and Meyerhoff 2014: 3) or "whole-corpus" approach (Egbert, Larsson and Biber 2020: 17).

must represent the data in a way that accounts for the non-independence of lexical choices at the text level.

3.2.2. Text-level model

A second class of procedures considers a corpus instead as a bag of texts. This form of data representation assumes that each *text* (rather than each word) is sampled randomly and independently from the language variety. No assumption is made about "randomness" or independence of words within texts (see Lijffijt et al. 2014 for a detailed discussion). The bag-of-texts model better matches corpus design, since text files form the unit of analysis (Baroni and Evert 2009: 798). The relevant sample size for statistical inference, then, is the number of texts (rather than word tokens) in the corpus.

For keyword analysis, statistical comparisons between the target and reference corpus are then based on text-level occurrence rates. Figure 1 above illustrates this form of data representation: One occurrence rate for each text, with a sample of 10 texts each from the target and reference variety. These 20 data points can then be subjected to quite different statistical procedures. Among those that have been used in keyness analysis are the Wilcoxon rank sum test[15] (see Kilgarriff 1996: 3; Kilgarriff 2001: 103) and the *t*-test (see Kilgarriff 1996: 3; Paquot and Bestgen 2009: 252-253). Note that these metrics assess the distinctiveness of an item, i.e. they are comparative frequency-oriented measures. In contrast, the *TD*-based likelihood-ratio test proposed by Egbert and Biber (2019: 84) measures comparative generality.

While these text-level procedures all treat the text as the unit of analysis, they differ in the additional assumptions they introduce about the data. Thus, the *t*-test relies on data to be normally distributed, which rarely (if ever) happens for corpus-based occurrence rates. The Wilcoxon rank sum test, in contrast, is a nonparametric procedure that is designed to be robust against departures from normality (Snedecor and Cochran 1989: 142-144). While the relative adequacy of different text-level analysis strategies is worthy of our attention, it remains to be explored more fully in future studies. For the present, we note that, from an inferential viewpoint, any of these techniques should constitute an improvement over the currently dominant bag-of-words analyses (see also Lijffijt et al. 2014).

We conclude our overview of inferential metrics by applying them to our exemplary data. Table 5 shows that, considering the size of our corpora (10 texts/20,000 words), the bag-of-words error probabilities are suspiciously small (cf. Evert 2006). The *p*-values based on a text-level analysis seem more reasonable. For interpretation, we would perhaps give priority to Wilcoxon's rank sum test due to its robustness advantage; in the present case, however, it yields virtually the same conclusion as the *t*-test. The *TD*-based likelihood-ratio test provides the most conservative assessment since it utilizes less information in the data. Thus, while the *t*-test treats the ptw rates as a continuous variable, the Wilcoxon rank sum

[15] The Wilcoxon rank sum test gives the same results as the Mann-Whitney test (or *U*-test).

test reduces them to an ordinal scale (i.e. ranks), and the *TD*-based likelihood-ratio test reduces them yet further to a binary scale.

Table 5. Inferential metrics for our hypothetical set of data.

Inferential metric	Sample size	*p*-value
Bag-of-words model		
Likelihood-ratio test	20,000	< .0000000000000001
Chi-square test	20,000	< .0000000000000001
Text-level model		
Wilcoxon rank sum test	20	.022
t-test	20	.025
TD-based likelihood-ratio test	20	.155

We have reviewed a large and varied set of keyness metrics. In the next section, we briefly discuss how to choose among them.

3.3. Choosing among keyness metrics

To recapitulate the various metrics we have surveyed, the reader may now wish to refer back to Table 1. As we have mentioned above, in applied work the choice among metrics is (or should be) a two-step process. We must first decide which dimension(s) of keyness to emphasize. This decision is made on linguistic grounds, in light of the researcher's objectives and the text varieties under investigation. The relative weight assigned to questions of distinctiveness or generality, for instance, will depend on our working definition of a prototypical key item. The purpose of our analysis also determines whether keyness will primarily be read from comparative measures (difference in frequency/dispersion) or whether equal (or even greater) emphasis should be given to target-variety features in isolation.

To illustrate, consider the identification of key verbs in academic writing and let us assume our aim is to produce a list of verbs for teaching purposes. We start by considering the target variety in isolation. Given our pedagogical focus, we may restrict our attention to verbs that are used at least moderately frequently, so that our final list includes items that are "worth learning". As for measures of generality, we may wish to emphasize verbs that are used across a broad range of disciplines and research styles. This will accentuate verbs that are used for general scientific argumentation (e.g. *compare*, *indicate*, *suggest*) rather than specific methodologies (e.g. *measure*, *correlate*, *analyze*). For comparative measures, much of our reasoning depends on the choice of reference variety. If we choose fiction as a benchmark, comparative measures may foreground items that are extremely rare or in very restricted use in that genre (e.g. *derive*, *hypothesize*, *deduce*). Such items, however, may not show sufficient discernibility or generality in our target variety. Based on these considerations, we would start out by ranking candidate items in a way that assigns the

greatest weight to discernibility (.35), followed by generality (.30), distinctiveness (.25), and comparative generality (.10).[16]

Considering the wide range of areas where keyness analysis is used, general recommendations as to the relative weight that should be assigned to the different dimensions cannot be given. We wish to stress, however, that the choice of dimension (i.e. *what* to measure), should take precedence over the choice of metric within a dimension (i.e. *how* to measure it). Once we have set our priorities with regard to the four dimensions of keyness, we can pay attention to methodological considerations such as the relative performance of metrics.

4. Evaluation of keyness metrics: Previous work

Comparative analyses of keyness metrics have been chiefly concerned with inferential measures. A number of different evaluation methods have been applied. Some studies start with a concrete keyword extraction task and then compare the adequacy of different metrics via a detailed qualitative examination of their output, i.e. lists of candidate items (e.g. Kilgarriff 2001; Paquot and Bestgen 2009; Lijffijt et al 2014; Gabrielatos 2018; Egbert and Biber 2019). An issue that has been repeatedly addressed is the effect of setting different cut-off values for inferential measures (e.g. Baker 2004; Rayson et al. 2004; Oakes and Farrow 2007; Bestgen 2014; Gabrielatos 2018; Egbert and Biber 2019). Other work has used simulation to evaluate inferential procedures. The way in which synthetic data are formed is critical, of course. If the data are generated in accordance with the bag-of-words model (e.g. Rayson et al. 2004), implications for the analysis of corpora are at best unclear. A more useful strategy is pursued in permutation-based work using existing text files from a target and a reference corpus. These texts are (repeatedly) shuffled and divided into two unrelated sets. For such random subdivisions of units, comparative *p*-values will have a predictable distribution under conditions of perfect agreement between data and assumptions (see Section 3.2). The deviation from this expectation then reflects the trustworthiness of an inferential measure. Studies in this spirit have demonstrated that text-level procedures outperform bag-of-words models (Paquot and Bestgen 2009; Lijffijt et al 2014; Brezina and Meyerhoff 2014).

Considering the function of inferential metrics in keyness analysis, the question of which procedures yield valid *p*-values has perhaps received too much emphasis in previous work. After all, keyword analysts do not test hypotheses; rather, they are interested in a useful technique for filtering out and ranking relevant candidate items. The performance criteria we adopt in the present study aim to foreground this applied purpose.

[16] The provisional ranking of verbs would then be based on a weighted average of the dimension-specific ranks.

5. Corpus study: Performance of keyness metrics

We come now to our performance evaluation of keyness metrics. In Section 5.1 we describe the keyword analysis task. We then introduce the criteria we apply (Section 5.2) and provide details about our evaluation method (5.3). In Section 5.4 we present the results. The data are available via TROLLing (Sönning 2023) and R code for reproducing the analyses can be found in the OSF project associated with this article (https://osf.io/kcwus/).

5.1. Analysis task: Key verbs in academic writing

To illustrate and evaluate different metrics, we concentrate on identifying key verbs in published academic writing. We rely on lemmatized data, which are drawn from COCA (Davies 2008) and cover a period of 30 years (1990–2019). The section "academic", which contains research articles from peer-reviewed journals, represents our target variety (for more details on this COCA genre, see Egbert, Larsson and Biber 2020: 8–10). Our reference corpus will be the "fiction" section of COCA, which contains short stories, plays, movie scripts, and the first chapter of novels. Web appendix 2 gives an overview of the data (https://osf.io/vg9td). It lists, for each year and subcorpus, the total number of text files and words, and the average text length. Overall, each subcorpus (academic and fiction) includes about 140 million words and 26,000 texts.

5.2. Performance criteria: Coverage and reliability

We evaluate the performance of keyness metrics with regard to two criteria. The first, which we will refer to as *coverage*, aims to measure the quality of a ranking, i.e. whether the items it ranks high are indeed good candidates for academic key verbs. We use as a standard of comparison the list of 233 verbs in Paquot's (2010: 57) Academic Keyword List, which was established based on a thorough corpus study relying on both frequency and dispersion information. If a keyness metric assigns relatively high ranks to these 233 verbs, we consider it to perform well. This gives us a means of comparing the rankings produced by different metrics. To put our keyword analysis task on a similar footing to Paquot (2010), we used the same genre (i.e. fiction) as a reference variety.

The second criterion we will concentrate on is *reliability*. We will consider a metric reliable if the rankings it produces are stable across analyses of the same pair of text varieties. Reliability can be measured by repeatedly applying a metric to the same target and reference variety. For a reliable metric, the ranking of candidate items will show low sample-to-sample (i.e. corpus-to-corpus) variability; in other words, it is not much affected by the specific texts that were sampled from the varieties of interest. We will also consider, as a second aspect of reliability, the sample-to-sample variability in coverage – i.e. how stable the observed level of coverage is across repeated studies. In the next section, we explain in more detail how we quantified these criteria.

5.3. Method

To monitor the performance of keyness metrics across repeated analyses of the same pair of text varieties, we take advantage of the size of COCA and work with 100 subsets of the

corpus. To this end, we formed 100 random subdivisions of roughly 520 text files each (~260 academic and ~260 fiction texts). These subsets will represent, as it were, different studies with the same objective: To obtain a set of verbs that are typical of academic writing. In order for these subsets to be as similar as possible, we balanced them on year, genre, and subgenre: Each contains (almost) the same number of texts per genre/subgenre-year combination, but each text appears in only one subset.

Partly for expediency, we chose to restrict our attention to those verb lemmas whose overall (corpus) frequency in the academic section of COCA exceeds 10 pmw. This left us with 700 verbs. For each of the 100 corpus subsets, we ranked the 700 verb lemmas based on each of 32 metrics. To assess coverage, we were then interested in the ranks that a specific metric assigned to the verbs in Paquot's (2010) list. It turns out that only 223 of these (i.e. 95%) cleared the 10-pmw threshold, and we focus on these items.[17] To measure coverage, we determined how many of these 223 key verbs appeared in the top 223 ranks, and expressed this as a proportion. The procedure is illustrated in miniature in Figure 2, where there are only 10 (instead of 100) corpus subsets, only 8 (instead of 223) key verbs, and only 20 (instead of 700) verb lemmas. The coverage of a metric ranges between 0 and 1, with 1 indicating that the 223 items in Paquot's (2010) list occupied the top 223 ranks.

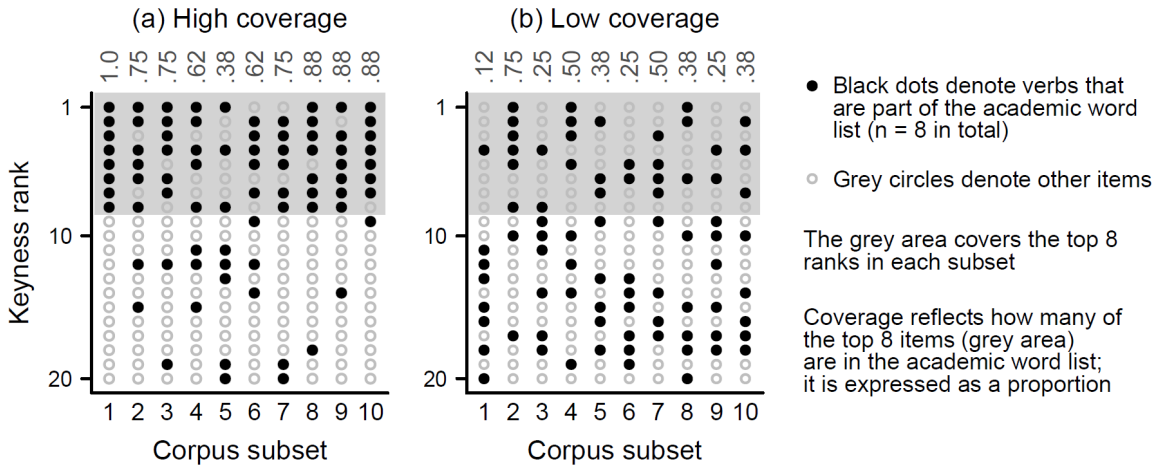

Figure 2. Quantification of coverage: Illustration using a hypothetical miniature analysis.

Since we are evaluating the performance of each metric on 100 corpus subsets, we obtain a distribution of coverage proportions, which reflects how the coverage rate of a metric varies

[17] The key verbs that did not clear the 10-pmw threshold are the following (overall ACAD occurrence rate in parentheses): *allocate* (9.8 pmw), *coincide* (8.0), *confine* (8.9), *conform* (9.7), *damage* (9.9), *effect* (7.6), *exemplify* (8.5), *neglect* (9.5), *tackle* (5.2), *term* (9.3).

from sample to sample. We will summarize this distribution using a 90% percentile interval, which includes the middle 90% of the distribution of coverage rates (i.e. 90 out of the 100).

We now turn to our second performance measure, reliability. A reliable keyness metric will yield very similar rankings across the 100 subsets (i.e. studies). This ensures that each of the 100 studies engages with a very similar set of candidate items, even though they have sampled different texts from the reference and target variety. We quantify reliability using a reliability coefficient (see Woods et al. 1986: 216), which ranges from 0 to 1, with 1 indicating perfect reliability.[18] The idea underlying such a coefficient is illustrated in Figure 3, again in miniature. We see that a high level of reliability reflects stable ranks across corpus subsets. As reliability decreases, the rank for a specific item fluctuates more and more across subsets.

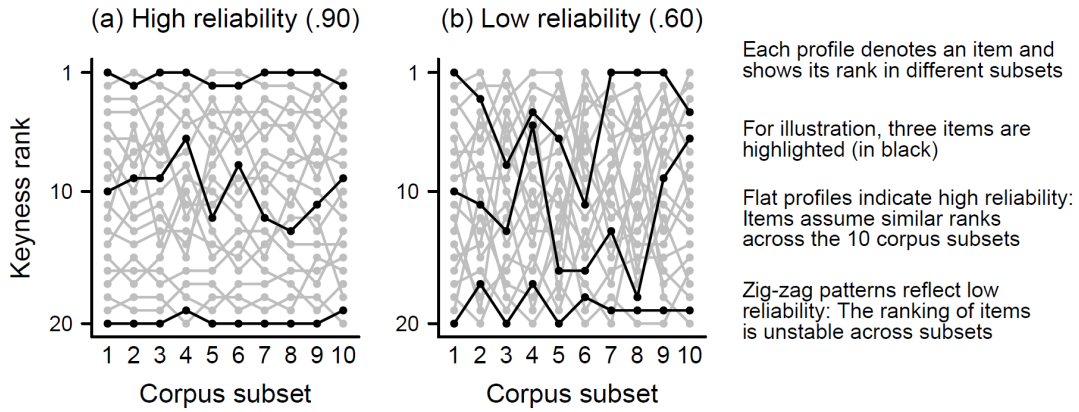

Figure 3. Quantification of reliability: Illustration using a hypothetical miniature analysis.

5.4. Results

Our results appear in Figure 4, where metrics are grouped by keyness dimension. The graph shows both performance measures. Filled circles ● denote the median level of coverage observed across the 100 subsets. Since coverage is arguably the more important criterion, it is used to order metrics within each keyness dimension. The error bars surrounding the filled circles reflect the variation of coverage rates across the 100 subsets; they include the middle 90% of the coverage rates. The reliability coefficients are shown using empty circles ○, and the corresponding numeric values for both measures are listed to the right of the display. Note that the vertical lines at the right margin reflect the performance limit (i.e. a reliability coefficient and coverage proportion of 1).

[18] A reliability coefficient is in fact similar to a correlation: If we compute a correlation matrix from the 100 rankings and calculate its average, we obtain the reliability coefficient.

When studying Figure 4, our main focus will be on the ranking of metrics within each keyness dimension. Apart from the question of which measures perform best, we will be concerned with two comparisons: (i) descriptive (black labels) vs. inferential metrics (grey labels) and (ii) metrics based on a bag-of-words (B) vs. text-level (T) model.

For discernibility metrics, we observe similar coverage and reliability levels for both the bag-of-words and the text-level occurrence rate. We were surprised to find that the bag-of-words version performed slightly better than the average across text-specific rates. For metrics reflecting generality, the differences in coverage are minor. In terms of reliability, however, *TD* performs best, with an index of .92.

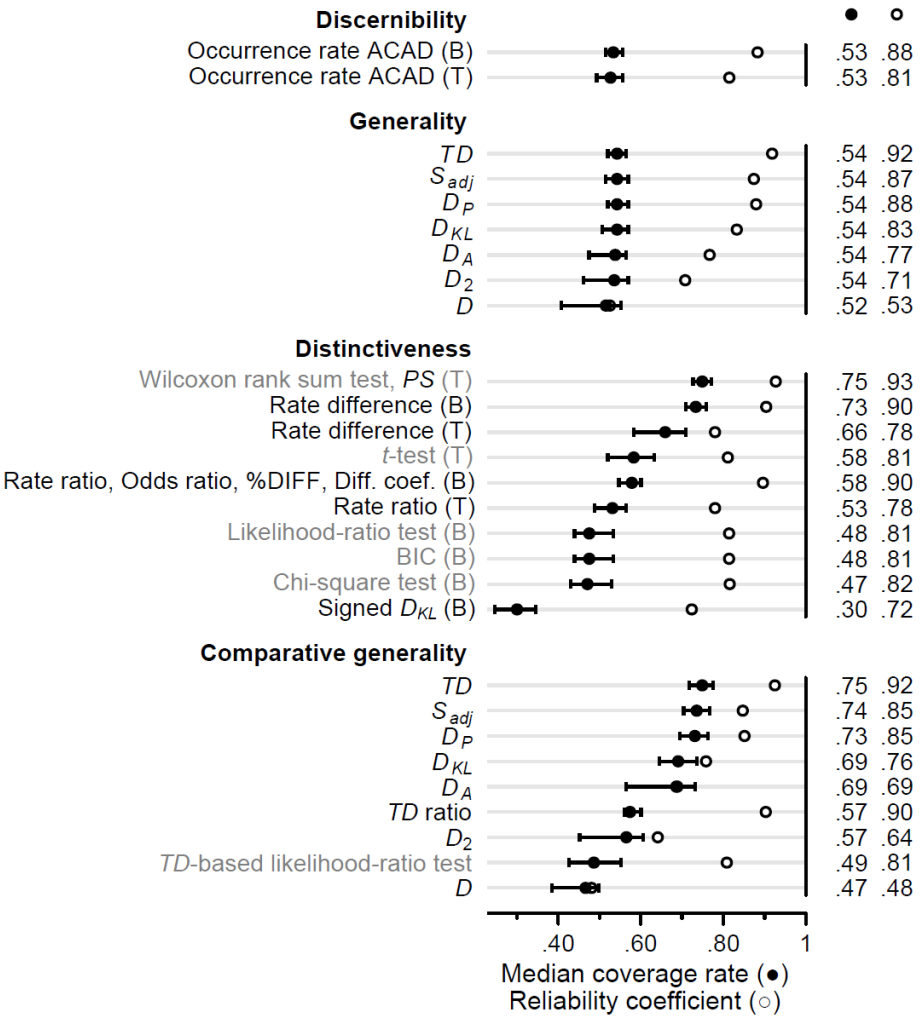

Figure 4. Performance measures for keyness metrics: Coverage and reliability. Inferential metrics have grey labels, descriptive metrics black labels; (B) denotes bag-of-words metrics, (T) text-level metrics; all (comparative) generality metrics are text-level measures. Error bars denote the variability in coverage rates across corpus subsets and contain the central 90% of coverage proportions.

Among distinctiveness metrics, Wilcoxon's rank sum test and *PS* outperform the other measures: The median coverage rate is .75, with error bars reflecting little sample-to-sample variability; the reliability coefficient is .93, the highest level obtained in our evaluation. The bag-of-words rate difference is the runner-up, with quite similar performance levels. Note that the likelihood-ratio test, which is by far the most widely used keyness metric (cf. Table 2), offers relatively poor coverage (.48) and reliability (.81). Somewhat surprisingly, the bag-of-words variants of two descriptive metrics, the rate difference and the rate ratio, outperform their text-level analogues by a considerable margin. It is also of interest to note that rate differences produce more consistent rankings than rate ratios. Signed D_{KL} comes in last, with a coverage of only .30.

The group of metrics expressing comparative generality is headed by the *TD* difference, which excels in both coverage (.75) and reliability (.92). S_{adj} and D_P also show quite favorable performance. Note that the *TD*-based likelihood-ratio test, the only inferential procedure in this group, performs considerably worse than descriptive metrics relying on *TD*; further, the *TD* ratio lags behind the *TD* difference in terms of both coverage and reliability.

6. Discussion

The aim of the present paper was twofold: to organize keyness metrics based on different dimensions of keyness, and to evaluate their performance on a specific task: the identification of key verbs in academic writing. Building on work by Baker (2004), Egbert and Biber (2019), and Gries (2021), we started out by delineating four dimensions of keyness (discernibility, distinctiveness, generality, and comparative generality), which offer genuinely complementary perspectives on typicalness. From a methodological angle, we therefore distinguished between (i) frequency-oriented vs. dispersion-oriented measurements, and (ii) an analysis of the target corpus in isolation vs. in comparison to the reference corpus. This allowed us to organize metrics on substantive grounds, i.e. according to the facet of typicalness they express. We underscored that the choice of keyness metric is primarily a choice among the four dimensions, with methodological trade-offs among measures being a secondary concern.

Our evaluation task focused on the applied nature of keyness analysis and assessed the quality (coverage) and reliability of the rankings produced by 32 metrics. Using data from COCA, we simulated 100 keyness analyses of the same pair of text varieties and recorded their agreement with Paquot's (2010) Academic Keyword List (coverage) as well as the stability of rankings across corpus subsets (reliability). We observed appreciable differences between metrics, especially for the dimensions of distinctiveness (i.e. difference in frequency) and comparative generality (i.e. difference in dispersion). Overall, we were able to identify, for each keyness dimension, metrics that showed favorable performance on our data:

- Discernibility: The bag-of-words occurrence rate
- Generality: *TD*, S_{adj}, and D_P
- Distinctiveness: Wilcoxon rank sum test/*PS* and the bag-of-words rate difference
- Comparative generality: *TD*, S_{adj}, and D_P

In our survey of keyness metrics (Section 3), we endorsed a text-level (rather than a bag-of-words) approach to the measurement of keyness. For frequency-based *inferential* measures, text-level metrics indeed performed better. Thus, the Wilcoxon rank sum test and the *t*-test outperform the likelihood-ratio test, the chi-square test, and BIC. These findings are consistent with earlier work on inferential metrics (Paquot and Bestgen 2009; Lijffijt et al 2014; Brezina and Meyerhoff 2014). However, for a number of frequency-based *descriptive* metrics (the occurrence rate, rate difference, and rate ratio), it was the bag-of-words version that yielded better coverage and reliability; this was clearly unexpected.

To make sense of these findings, let us consider, as a possible explanation, the (implicit) weighting of texts that is implemented when a metric is calculated. We start by noting that if all texts have the same length, the bag-of-words estimate and the text-level estimate coincide. This is because both approaches then assign the same weight to each text. If the length of texts in a corpus differs, the two versions no longer coincide: A simple average across text-specific rates in a corpus will still assign the same weight to each text, no matter how long it is. A bag-of-words calculation, on the other hand, will weight texts in proportion to their word count. This means that longer texts will have a greater influence on the result. Shorter texts, on the other hand, are backgrounded. These different weighting schemes may have given rise to the difference in performance. Thus, perhaps a problematic feature of the descriptive text-level metrics we have implemented is the fact that they put too much weight on short texts. Based on statistical theory (and common sense), estimates from smaller samples show greater sampling variability. For keyness analysis, this means that shorter texts yield more variable occurrence rates, and this may have negatively affected the performance of text-level descriptive metrics.

A further finding that we did not anticipate was the fact that the simplest – and coarsest – dispersion measure (*TD*) showed the best performance for (comparative) generality rankings. Other dispersion measures are able to capture more subtle differences in the text-to-text variability of occurrence rates, and we expected this increase in informativity to have a positive effect on performance. Our findings, however, seem to suggest that reducing the information in the data to a binary indicator variable (whether the item appears in a text or not) is particularly beneficial for the reliability of a metric, without detriment to its coverage. This might be due to the fact that coarseness brings with it the advantage of resistance to minor perturbations in the data. Thus, the underlying dichotomization may give *TD* a robustness advantage over other dispersion metrics.

Another aspect that is at variance with our pre-data conceptions is that, for two pairs of comparative metrics (rate difference/rate ratio; *TD* difference/*TD* ratio), absolute differences yielded better coverage than relative differences. Thus, for the comparison of occurrence rates, we would have given preference to the rate ratio, primarily on grounds of interpretability. A possible reason for the performance advantage of the rate difference is its association with distinctiveness: If an item has a high occurrence rate in the target corpus, the absolute difference from the rate in the reference corpus will also be greater, on average. This is borne out by our data: The average (Spearman) rank correlation between (bag-of-words) rate differences and (bag-of-words) occurrence rates is $\rho = .28$, indicating a moderate association. For the (bag-of-words) rate ratio, on the other hand, the corresponding average

is ρ = .01. This means that the rate difference not only measures distinctiveness, but also discernibility. For the *TD difference* and the *TD ratio*, a similar situation holds: The *TD* difference shows a stronger association with *TD* (ρ = .22) than the *TD* ratio (ρ = .02). This means that it offers information not only about comparative generality but also, in part, about generality. In both cases, then, absolute differences combine information about two dimensions of keyness. Considering that Paquot's Academic Keyword List rests on a consideration of different dimensions (most notably distinctiveness and generality), our approach to measuring coverage may unfairly benefit metrics that reflect different aspects of typicalness.

We should also bear in mind that the current investigation monitored the performance of metrics in a highly restricted setting, i.e. one particular word class in one specific target variety. To assess the degree to which our findings generalize to other analysis settings, further research is needed. A direct follow-up to this study could look at different word classes in the same pair of text varieties, perhaps using a similar research design, with Paquot's (2010) Academic Keyword List as a standard of comparison.[19] The scope could then be broadened to other, less constrained analysis tasks, eventually also focusing on other text varieties. Consideration should then be given to genres where the prototypical text differs in length and structure from a published research article. Thus, the average exemplar in our analysis task is about 5,000 words long. It would be of particular interest to observe the performance of metrics in genres with (much) shorter texts.

Despite these caveats, the present study has been able to report further support for the Wilcoxon rank sum test and *TD*-based metrics. We can also make additional recommendations for these measures: First, instead of an inferential *p*-value or test statistic based on the Wilcoxon rank sum test, analysis software could report *PS* as a descriptive and more informative metric. Second, the *TD* difference, as a descriptive metric, may be worthy of consideration as a potential replacement for the inferential *TD*-based likelihood-ratio test. Apart from these concrete suggestions for applied keyness analyses, the discussion of the unexpected parts of our findings has pointed to two relatively open areas for methodological research on keyness metrics: (i) the aspect of robustness, i.e. whether certain metrics are more sensitive to differences in text length and/or the sampling variability of text-specific occurrence rates; and (ii) the degree to which certain metrics may be offering a blend of different keyness dimensions, which may not only give them an (unfair) advantage over alternative measures, but would arguably also negatively affect their linguistic interpretability. It would seem that an engagement with these matters may help us better understand performance differences among keyness metrics.

Acknowledgements: I would like to thank the five anonymous reviewers for their constructive and helpful comments on earlier versions of this paper.

[19] We have made an effort to document our analyses in a way that will allow other researchers to implement and extend our methodological approach. All analysis scripts are commented, and available both as R scripts and html files in the associated OSF repository.

References

Baker, Paul. 2004. Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics* 32(4). 346–359.

Baroni, Marco & Stefan Evert. 2009. Statistical methods for corpus exploitation. In Anke Lüdeling & Merjya Kytö (eds.), *Corpus linguistics: An international handbook*, 777–803. Berlin: Mouton de Gruyter.

Bestgen, Yves. 2014. Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* 29(2). 164–170.

Brezina, Vaclav & Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1). 1–28.

Carroll, John B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour* 3(2). 61–65.

Church, Kenneth W. & William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1(2). 163–190.

Davies, Mark. 2008. *The Corpus of Contemporary American English*. www.english-corpora.org/coca.

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

Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analysis. *Corpora* 14(1). 77–104.

Egbert, Jesse, Brent Burch & Douglas Biber. 2020. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics* 25(1). 89–115.

Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge: Cambridge University Press.

Evert, Stefan. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–190.

Gabrielatos, Costas. 2018. Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor & Anna Marchi (eds.), *Corpus approaches to discourse: A critical review*, 225–258. New York: Routledge.

Gabrielatos, Costas & Anna Marchi. 2011. Keyness: Matching metrics to definitions. http://eprints.lancs.ac.uk/51449. (29 March, 2023.)

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. New York: Springer.

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33.

Grissom, Robert J. & John J. Kim. 2012. *Effect sizes for research: Univariate and multivariate applications*. New York: Routledge.

Hardie, Andrew. 2014. Log ratio – An informal introduction. http://cass.lancs.ac.uk/?p=1133. (29 March, 2023.)

Hofland, Knut & Stig Johansson. 1982. *Word frequencies in British and American English.* London: Longman.
Juilland, Alphonse G., Dorothy R. Brodin & Catherine Davidovitch. 1970. *Frequency dictionary of French words.* The Hague: Mouton de Gruyter.
Kilgarriff, Aadam. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In Lindsay J. Evett & Tony G. Rose (eds.), *Language Engineering for Document Analysis and Recognition*, 33–40. Nottingham: Nottingham Trent University.
Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.
Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.
Kilgarriff, Adam. 2009. Simple maths for keywords. In Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds.), *Proceedings of the Corpus Linguistics Conference, CL2009.* Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/CL2009/171_FullPaper.doc. (29 March, 2023.)
Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki & Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31(2). 374–397.
McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.
Oakes, Michael P. & Malcolm Farrow. 2007. Use of the chi-squared test to examine vocabulary differences in English-language corpora representing seven different countries. *Literary and Linguistic Computing* 22(1). 85–100.
Paquot, Magali. 2010. *Academic vocabulary in learner writing.* London: Continuum.
Paquot, Magali & Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Andreas H. Jucker, Daniel Schreier & Marianne Hundt (eds.), *Corpora: Pragmatics and discourse,* 247–269. Amsterdam: Rodopi.
Pojanapunya, Punjaporn & Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 133–167.
Rayson, Paul. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison.* Lancaster: Lancaster University dissertation.
Rayson, Paul, Damon Berridge & Brian Francis. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In Gérard Purnelle, Cédrick Fairon & Anne Dister (eds.), *Le poids des mots: Proceedings of the 7th International Conference on Statistical Analysis of Textual Data, Vol. 2*, 926–936. Louvain-la-Neuve: Presses Universitaires de Louvain.
Rosengren, Inger. 1971. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de Linguistique Appliquée (Nouvelle Série)* 1. 103–127.
Scott, Mike. 1997. PC analysis of key words – and key key words. *System* 25(2). 233–245.
Snedecor, George W. & William G. Cochran. 1989. *Statistical methods.* Ames: Iowa State University Press.

Sönning, Lukas. 2023. *Key verbs in academic writing: Dataset for "Evaluation of keyness metrics: Performance and reliability"*. DataverseNO, V1. https://doi.org/10.18710/EUXSMW

Wilcox, Allen R. 1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly* 26(2). 325–343.

Wilson, Andrew. 2013. Embracing Bayes factors for key item analysis in corpus linguistics. In Markus Bieswanger & Amei Koll-Stobbe (eds.), *New approaches to the study of linguistic variability*, 3–11. Frankfurt: Peter Lang.

Winter, Bodo & Martine Grice. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5). 1251–1277.

Woods, Anthony, Paul Fletcher & Arthur Hughes. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

Zhang Jun & Kai F. Yu. 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association* 280(19). 1690–1691.