Count regression models for keyness analysis

Lukas Sönning (University of Bamberg)

lukas.soenning@uni-bamberg.de

Abstract. A wide variety of measures have been used in previous work to assess the keyness of items in a particular domain of language use. The present paper explores an approach to keyword analysis based on regression modeling. Specifically, we use a form of negative binomial regression, which offers a number of advantages compared to existing techniques for identifying typical items in a target corpus. Thus, it is responsive to the multidimensional nature of keyness and can address multiple aspects of typicalness simultaneously, using a single statistical model. Further, metrics of interest can be enriched with confidence intervals, which allows us to isolate descriptive and inferential indicators of keyness. Finally, all quantities are based on a text-level analysis, which accounts for the fact that the target and reference corpus consist of text files and adjusts uncertainty estimates accordingly. As an illustrative case study, we rely on COCA to identify key verbs in academic writing and demonstrate how negative binomial regression may be used to this end. Our checks on the coverage rate of the 95% confidence intervals indicate that this model seems to be adequate for purposes of statistical inference. Due consideration will also be given to the limitations of this procedure, and we conclude by outlining the kinds of keyness analyses for which count regression models may be a worthwhile approach. The online supplementary material for this paper provides data and R code for the implementation of keyness regression.

1. Introduction

Keyness analysis is a corpus-based procedure that is used to identify items such (e.g. word forms, lemmas, or other structures) that are typical of a particular domain of language use (Scott 1997; for recent reviews, see Gabrielatos 2018; Rayson & Potts 2020). To this end, the text variety of interest is represented by what is referred to as the target corpus and compared to a reference corpus, which is selected to provide a meaningful basis of comparison. An important role is played in keyness analysis by keyness metrics, which aim to quantity typicalness. Based on these measures, candidate items are extracted from the target corpus and then ranked.

While keyness metrics have primarily been designed to respond to frequency differences between target and reference corpus (e.g. Scott 1997; McEnery & Hardie 2012: 245), other aspects of typicalness have recently received increasing attention. Thus, Egbert & Biber (2019) distinguish between the *distinctiveness* of an item and its *generality* in the target variety. While the former feature can be quantified using frequency comparisons, the *generality* of an item can be read from measures of dispersion, which indicate how evenly an item is spread across texts in the target (and reference) corpus (see Gries 2020 for a review). Accordingly, we can distinguish between frequency-oriented and dispersion-oriented approaches to keyness.

A wide variety of keyness measures have been proposed and used in previous work (for an overview, see Gabrielatos 2018; Rayson & Potts 2020; Sönning 2022a). The aim of the present paper is to explore an approach to keyword analysis based on regression modeling, which offers a number of advantages compared to the existing set of metrics. For one, regression is able to address multiple dimensions of keyness simultaneously, using a single statistical model. Further, all measures of interest come with indications of statistical uncertainty (i.e. confidence intervals). This facilitates an estimation approach (see Cumming 2012; Cumming & Calin-Jageman 2017; Gries 2022) to keyness analysis and allows us to keep apart descriptive ("effect-size") and inferential information. Finally, all quantities of interest are based on a text-level analysis that respects corpus design and the resulting structural features of corpus data (cf. Baroni & Evert 2009; Lijffijt et al. 2014).

The plan of the paper is as follows. To provide some background on keyness analysis, Section 2 distinguishes different dimensions of keyness and outlines some methodological matters that have received attention in the literature. In Section 3, we offer a brief introduction to count regression models and discuss how they may be applied to keyness analyses. Section 4 then demonstrates the use of negative binomial regression for the identification of key verbs in academic writing. After an audit of the inferential-statistical performance of this procedure (Section 5), we discuss, in turn, the advantages (Section 6) but also the drawbacks (Section 7) of keyness regression. Section 8 concludes with a summary and outlook.

2. Keyness analysis

This section deals with keyness as a multidimensional construct and sketches different dimensions of typicalness (Section 2.1). Some methodological issues surrounding the procedure are then discussed in Section 2.2.

2.1. Dimensions of keyness

While traditional approaches to keyword analysis have almost exclusively relied on frequency comparisons, keyness is in fact a multidimensional construct. This was stated explicitly by Egbert & Biber (2019), who distinguished between two features of typicalness: *distinctiveness* and *generality* (see also Baker 2004). An item is considered distinctive of the target variety if it reflects the aboutness of the discourse and is (strongly) linked to this domain of language use. This feature materializes in an overrepresentation in the target corpus and can therefore be determined using frequency comparisons. *Generality*, on the other hand, is concerned with the pervasiveness of an item throughout the discourse. Thus, if an item is to be considered key, it should be in widespread use and occur in a broad range of texts representing the target domain. This feature of typicalness manifests itself in a relatively even distribution of occurrence rates across target variety texts.

We can therefore make a general distinction between frequency-oriented and dispersion-oriented assessments of keyness (see also Gries 2021). Both directions can be further broken down depending on whether we evaluate typicalness (i) by considering the target corpus in isolation, or (ii) by comparing it to the reference corpus. Accordingly, we can delineate four dimensions of keyness (Sönning 2022a), which capture different aspects of typicalness and are therefore complementary in nature:

- *Discernibility*: The item is discernible if it is used at a perceptible rate in the target variety.
- *Distinctiveness*: Compared to the reference variety, the item is used at a noticeably higher rate in the target domain.
- *Generality*: The item enjoys widespread use in the target variety and can therefore be considered a common or general feature of this area of language use.
- *Comparative generality*: Compared to the reference domain, the item is more common and used more broadly in the target variety.

As Table 1 shows, these keyness dimensions allow us to form four linguistically meaningful classes of metrics. For reasons of space, we cannot provide details about the individual measures here, and we refer the reader to Gabrielatos (2018), Rayson & Potts (2020), Gries (2020), and Sönning (2022a, 2022b). The four-way arrangement in Table 1 offers a constructive point of departure for keyness analysis, since it requires the analyst to first consider which features of keyness to emphasize when looking for typical items in the target corpus. To this end, it is helpful to have a working definition of "keyness" and to sketch features of a prototypical key item. This allows us to recognize the relative importance of each dimension.

2.2. Methodological issues in keyness analysis

Keyness metrics can also be organized on statistical grounds, by considering (i) whether they express descriptive or inferential information and (ii) whether the analysis makes explicit reference to the individual text files in the corpus. These classifications, which are also denoted in Table 1, will be considered in turn.

The first division distinguishes metrics that involve statistical error probabilities (*p*-values) from descriptive ones, which express the magnitude of an observed association or difference. Consider, for instance, the *distinctiveness* of items in the target variety. Inferential metrics rank items based on the confidence with which we can state, given our sample (i.e. corpus), that the frequency difference holds in the underlying populations (i.e. the target and reference variety). This confidence assessment is based on a null hypothesis significance test and usually indicated with a *p*-value or (equivalently) a test statistic.

Descriptive measures, on the other hand, are not concerned with inferences from sample to population and instead provide sample-specific (i.e. corpus-specific) information. To express the *distinctiveness* of an item, for instance, they report on the degree to which it is overrepresented in the target corpus. An example would be an occurrence rate ratio of (say) 4, which tells us that the item occurs 4 times more often in the target corpus (cf. Kilgarriff 2009). In part driven by statistical reform movements across the (behavioral) sciences (see, e.g. Kline 2013), a number of "effect-size" measures have been proposed for keyness analysis (e.g. Gabrielatos & Marchi 2012; Brezina 2014). What these measures have in common is that they express keyness in descriptive terms.

Table 1. Dimensions of keyness and keyness metrics.

	Frequency-oriented	**Dispersion-oriented**
Target variety in isolation	*Discernibility* Descriptive ◐ Occurrence rate (e.g. pmw)	*Generality* Descriptive ○ Range [l] ● *TD* [m] ● D_{KL} [g] ● D, S_{adj}, D_2, D_P, D_A [b]
Comparison to reference variety	*Distinctiveness* Descriptive ◐ Rate ratio [a] ◐ Rate difference [b] ● *PS* [b] ◐ Log ratio [c] ○ Difference coefficient [d] ○ %DIFF [e] ◐ Odds ratio [f] ○ Signed D_{KL} [g] Inferential ○ χ^2 test [d] ○ Likelihood-ratio test [h] ● Wilcoxon test [j] ● t-test [j] ○ BIC [k]	*Comparative generality* Descriptive ● *TD* ratio [m] ● *TD* difference[b] ● D_{KL} difference [g] ● D, S_{adj}, D_2, D_P, D_A difference [b] Inferential ● *TD*-based likelihood-ratio test [m]

Note. Key to symbols: Keyness metric based on: ○ a bag-of-words analysis; ● a text-level analysis; or ◐ both kinds of analyses possible.

References: [a] Kilgarriff 2009; [b] Sönning 2022a; [c] Hardie 2014; [d] Hofland & Johansson 1982; [e] Gabrielatos & Marchi 2011; [f] Pojanapunya & Watson Todd 2016; [g] Gries 2021; [h] Dunning 1993; [j] Kilgarriff 1996; [k] Wilson 2013; [l] Rayson 2003; [m] Egbert & Biber 2019

A second way in which we can form statistical classes of keyness metrics depends on how corpus data are represented when measuring keyness. Thus, a broad distinction can be made between bag-of-words and text-level models (cf. Evert 2006; Baroni & Evert 2009; Lijffijt et al. 2014). The bag-of-words approach to keyness analysis treats a corpus as an unstructured set of word tokens and performs computations based on overall corpus counts. To quantify the keyness value of a specific item, data from the target and reference corpus are collapsed into a single two-by-two table. Using this four-number summary of the two corpora, various metrics may be computed. This whole-corpus approach has met with criticism in the literature (e.g. Brezina & Meyerhoff 2014; Lijffijt et al. 2014; Egbert & Biber 2019), the primary reason being the mismatch between corpus design and data analysis, which is particularly problematic when making statistical inferences (for more detailed discussions, see Lijffijt et al. 2014; Sönning 2022a). In Table 1, keyness metrics that rely on a bag-of-words representation of the corpus are marked with an open circle (○).

In contrast, a text-level analysis is carried out at the level of the individual text files in the corpus. This means that the analysis takes the number of texts (rather than overall words counts in the corpus as the relevant sample size, with each text contributing one data point to the analysis. For a text-level analysis, then, the data must be represented differently, as we are not counting occurrences in the corpus as a whole, but at the level of the individual text files. In Table 1, keyness measures that are based on a text-level analysis are marked with a filled circle (●). Note that while inferential measures are *either* bag-of-words (○) *or* text-level (●) metrics, many descriptive indexes can be computed in both ways (◐). The rate ratio, for instance, can express a comparison of two bag-of-words occurrence rates; or, alternatively, we may compare two averages based on text-specific rates.

From a linguistic viewpoint, the statistical groupings we have reviewed play a secondary role. This is because they merely distinguish different ways of measuring the *same* keyness dimension. *Distinctiveness*, for instance, can be expressed (i) descriptively or inferentially, and (ii) based on a bag-of-words or text-level representation of the data. The four dimensions of keyness summarized in Section 2.1, on the other hand, offer more elementary classes. They organize metrics based on what aspect of typicalness they intend to measure, i.e. which keyness dimension they express. And obviously, in any substantive area of research, the question *what* to measure takes precedence over *how* to measure it.

3. Count regression models for keyword analysis

We now consider the use of regression modeling for keyword analysis. We start out with some background on the relevant family of count regression models (Section 3.1) and then clarify, in Section 3.2, the different ways in which the term "dispersion" is used in corpus linguistics and statistics. Section 3.3 then introduces negative binomial regression, which will form the basis of keyness regression, and Section 3.4 discusses how measures of dispersion may be obtained from these models.

3.1. Count regression models

The term "regression" refers to a family of data-analytic procedures that are based on the generalized linear model (Nelder & Wedderburn 1972). Since many statistical tests and tools can be (re)formulated as regression models (e.g. the t-test, ANOVA, the chi-squared test, etc.), the generalized linear model offers a unified framework for data analysis (see Winter 2020 for an accessible introduction for linguists). Most regression models describe the relationship between an outcome variable and one or several predictors. These variables are measured on specific units, and in keyness analysis, these units are the text files. Thus, we may be interested in whether, for a specific lexical item, there is a relationship between its text-specific occurrence rate (outcome) and the domain of discourse (predictor), i.e. whether it appears at different rates in the target and the reference variety. The typical output in keyness regression would be an occurrence rate estimate for the reference corpus (a normalized frequency, e.g. per million words) and an estimate of the (multiplicative) difference between the two corpora (a rate ratio). In addition, the model returns confidence intervals (CIs), which indicate the statistical precision of these data summaries.

The generalized linear model can be applied to various types of outcomes including binary and categorical attributes (see Long 1997). The quantity of interest in keyword analysis is a count, i.e. the text-level frequency of an item. Since counts are discrete and non-negative, they fall into the remit of one particular family of generalized linear models: count regression models (see Winter & Bürkner 2021 for an introduction aimed at a linguistic audience). This family descends from the Poisson distribution, and its most basic variant is Poisson regression. For most data settings, however, the Poisson model is too simplistic and fails to capture existing variability among units.[1] This is also true for keyword analysis, where a Poisson model will underestimate the text-to-text variation of occurrence rates, thus offering a poor description (i.e. model) of the data. Before we discuss negative binomial regression as an extension that allows for more flexibility, we need to consider the notion of "dispersion".

[1] This is because the Poisson model assumes that a specific event (e.g. the occurrence of an item) has the same probability for *all* units in the group of data points that is modeled (e.g. texts in keyword analysis). Consider, for instance, the identification of keywords analysis in published academic writing. The Poisson model assumes that, in each published academic paper, a specific form (e.g. the verb *correlate*) has the same probability of being used. However, due difference in topic (or here: research methodology), the occurrence rate of *correlate* will fluctuate noticeably from text to text.

3.2. Dispersion: Corpus-linguistic vs. statistical sense

If the text-to-text variability in occurrence rates is higher than described by a statistical model, the data are said to show excess dispersion (or overdispersion). We should note that the statistical meaning of the term "dispersion" differs from its meaning in lexicography and corpus linguistics – in fact, the term is used in (seemingly) contradictory ways. To understand why this is the case, consider Figure 1. Panel (a) shows what is often referred to as a "dispersion plot" for two corpora, A and B. The long rectangles represent the string of words constituting the corpus, and the spikes inside of these locate occurrences of a specific item in the corpus. In corpus A, the item is spread out quite evenly. In corpus B, instances are more densely clustered, and there are large stretches of text where the item does not occur. In the corpus-linguistic sense, then, the dispersion of the item is higher in corpus A (see Gries 2008, 2020; Sönning 2022b).

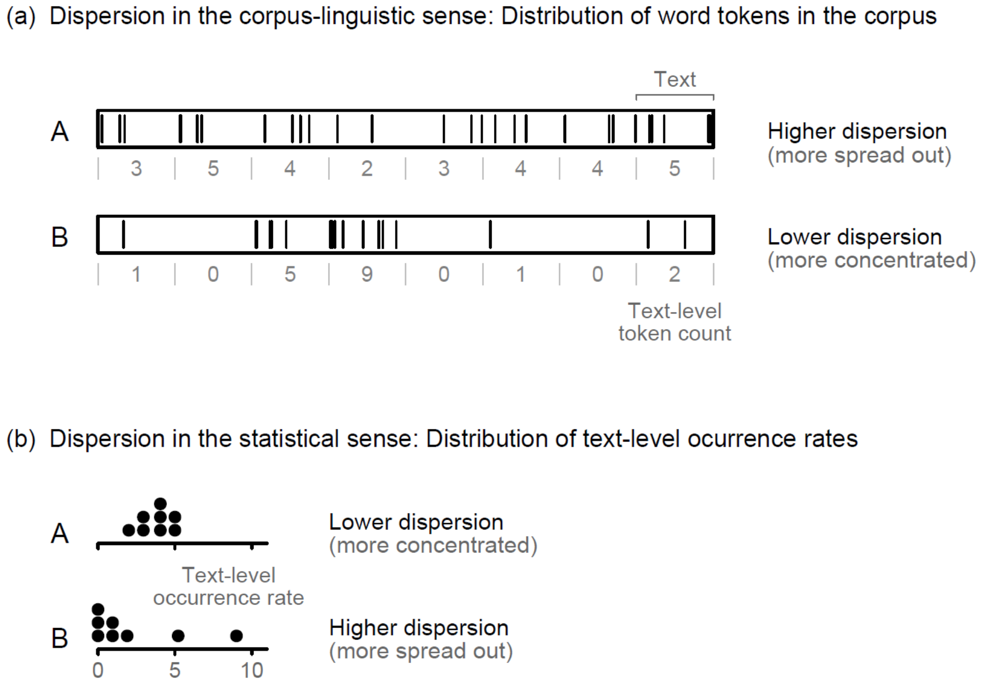

Figure 1. Usage of the term "dispersion" in corpus linguistics and statistics 🅭🅯 [2]

Panel (b) shows a different representation of these data. Instead of looking at the corpus as a string of words, we divide it into its constituent texts and count how often the item occurs in each document. Text boundaries are sketched in panel (a) using thin grey lines below the rectangles, and the numbers

[2] Images with the symbols 🅭🅯 in the figure caption have been published under the Creative Commons Attribution 4.0 licence (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/mc26t/.

denote text-level token counts. Let us now consider panel (b), where each text is represented by a dot. This dot marks how often the item appeared in the text. We can then look at the distribution of text-level occurrence rates in the two corpora. In corpus A, texts form a dense pile. In corpus B, occurrence rates are more widely spread out. At this level of description, then, it is corpus B that shows greater dispersion. In the statistical literature on count regression, the term dispersion is used in this sense, i.e. to refer to the variability of unit-specific (i.e. text-level) occurrence rates (e.g. Long 1997: 221; Gelman et al. 2021: 264–268).

An awareness of the different usages of "dispersion" is helpful when engaging with the statistical literature. In the following, we use the term in the corpus-linguistic sense, and refer to the perspective in panel (b) more concretely as the variation or variability in (text-specific) occurrence rates. The statistical term 'overdispersion', however, plays an important role in count regression and will appear in the following discussion. It describes data that show greater variability in text-specific occurrence rates than a currently entertained model is able to accommodate. Thus, count data are typically overdispersed relative to a simple Poisson model. The family of count regression models has therefore been extended to include more flexible members that manage to represent surplus variability. One offspring of Poisson regression is negative binomial regression, to which we now turn.

3.3. Negative binomial regression

The negative binomial model offers a simple remedy to the problem of overdispersion: It includes an additional parameter that describes the amount of variability among text-specific occurrence rates (see Long 1997: 217-238; Long & Freese 2014: 481-518; Hilbe 2014: 126-161; Hilbe 2011; see Mosteller & Wallace 1984 for an application to word frequency data). We will refer to this auxiliary parameter as the *shape* parameter; it is similar to a standard deviation in that it captures the variability of observations around some measure of central tendency. This means that it measures dispersion in the statistical sense.

Let us consider an example. Figure 2 shows data on the usage rate of *actually* (see Sönning & Krug 2021, 2022) in the SpokenBNC2014 (Love et al. 2017). Each dot denotes a speaker (668 in total), and the rate of *actually* varies between 0 and 8 ptw. If we describe these data using a Poisson regression model (for R code, see https://osf.io/tqchs), we obtain 1.54 ptw as an estimate of the typical rate, with a tight 95% CI [1.51; 1.56]. A negative binomial regression model yields virtually the same average rate but a much wider 95% CI [1.44; 1.62]. The increase in width stems from the fact that the negative binomial model is aware of the variability among speakers, which exceeds that accommodated by a Poisson distribution.

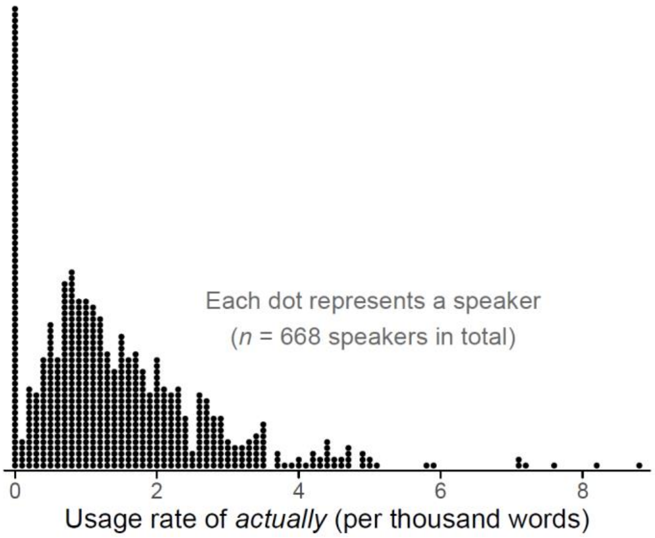

Figure 2. Speaker-specific usage rates of *actually* in the SpokenBNC2014 🅭🅯

The variability among speakers is captured by the shape parameter of the negative binomial distribution. This parameter defines the shape of a gamma distribution. For this reason, the negative binomial model is sometimes also referred to (more transparently) as a Poisson-gamma mixture model (e.g. Cameron & Trivedi 2013: 117; McElreath 2020: 373). The gamma distribution for the *actually* data appears in Figure 3. This distribution has a mean of 1, which is marked by the vertical line. The grey area represents the variation among usage rates in terms of their multiplicative deviation from the average occurrence rate. The x-axis therefore denotes factors: A factor of 2 indicates that a speaker's usage rate of *actually* is twice as high as the sample average. Figure 3 shows that, according to our current model, only few speakers in our sample exceed the average rate of 1.54 ptw by more than a factor of 2 – in terms of actual occurrence rates, this means that only few speakers show a usage rate greater than 3 ptw.

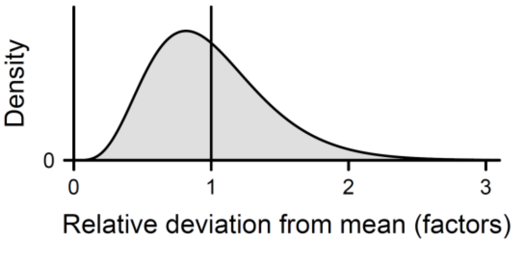

Figure 3. The gamma distribution expressing between-speaker variability in the usage rate of *actually* 🅭🅯

3.4. Dispersion measures based on the negative binomial regression model

We have seen that a negative binomial regression model preserves information about the variability among units (e.g. speakers or texts) using the gamma shape parameter. This parameter can be translated into a standardized dispersion index that ranges between 0 (concentrated distribution) and 1 (balanced distribution). This measure, which we will refer to as D_{NB}, is therefore on the same scale as other indexes of (lexical) dispersion (see Gries 2020; Sönning 2022b). The conversion formula is given in Appendix 1, and Figure 4 illustrates how the two quantities are related. For *actually*, the model outputs a shape parameter of 0.43. Looking at Figure 4, we note that this corresponds to a dispersion of roughly .90. This D_{NB} value indicates that the item is quite evenly distributed across the 668 speakers in the corpus and may therefore be considered a general feature of contemporary British speech.

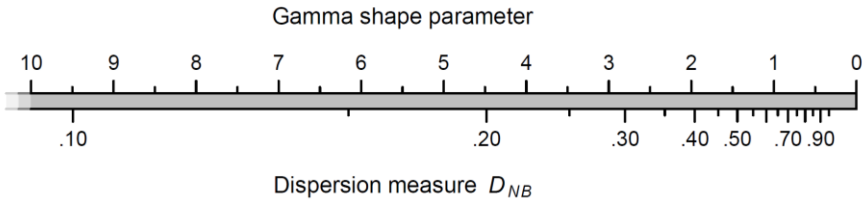

Figure 4. Mapping between gamma shape parameter values and the dispersion score D_{NB}

For the purpose of keyword analysis, we can also construct a dispersion measure that is similar to Egbert & Biber’s (2019) text dispersion (*TD*). *TD* indicates the proportion of texts in the corpus that feature at least one occurrence of the item. What is problematic about *TD* is its sensitivity to the length of texts (see Gries 2021; Sönning 2022b). Since the likelihood of observing a specific form is smaller in shorter texts, comparisons of *TD* values can be biased if texts in the target and reference corpus differ in average length. To sidestep these issues, we can construct a *TD* measure based on a negative binomial regression model. To this end, we take advantage of the model’s ability to make predictions. Thus, if we tell our model the length of a (future) target variety text, it can predict the probability that it contains at least one occurrence of the item. We can do the same for the reference variety, specifying the *same* (!) text length. This way we can adjust *TD* when comparing text collections that differ in typical length. The choice of text length is perhaps best made by considering a prototypical exemplar of the target variety text.[3] In the case of *actually*, we could consider, as a benchmark, a 10-minute

[3] Alternatively, we could use the typical (i.e. mean or median) text length in our target corpus, or across both corpora. It then makes sense to round generously, perhaps even to the nearest order of magnitude (i.e. 100/1000/10,000, etc.).

stretch of speech with roughly 1,000 words. Based on our model, the probability of at least one occurrence of *actually* in this 10-minute interval is .69 (or 69%). We will refer to this dispersion measure as the *negative binomial text dispersion* (*NB-TD*), with a subscript denoting the length of the envisaged text (e.g. $NB\text{-}TD_{1k}$).

4. Keyness metrics based on the negative binomial regression model

We now illustrate the use of count regression for keyword analysis. As we will see, an attractive feature of the negative binomial regression model is that, for a specific item, it can address all four dimensions of keyness, and deliver both descriptive and inferential indicators. In Section 4.1 we illustrate how such a model is specified for one particular item in academic writing. Section 4.2 then presents a more exhaustive key verb analysis. R code for running the analyses can be found in the accompanying OSF project (https://osf.io/mc26t/).

4.1. Modeling frequency and dispersion

A negative binomial regression model can be specified to address each of the four dimensions of keyness that appeared in Table 1:

- *Discernibility*: The occurrence rate (e.g. ptw) of the item in the target corpus
- *Distinctiveness*: The rate ratio (normalized frequency in the target corpus divided by that in the reference corpus)
- *Generality*: Either a standardized dispersion measure (D_{NB}), or the probability that a text of (say) 10,000 words features at least one occurrence of the item ($NB\text{-}TD_{10k}$)
- *Comparative generality*: Difference in *generality*, as indicated by D_{NB} or $NB\text{-}TD_{10k}$

To model both frequency and dispersion, we need to specify a negative binomial model of location and scale (e.g. Rigby & Stasinopoulos 2005), which is sometimes also referred to as a heterogeneous negative binomial model (Hilbe 2011: 319-323). In addition to the mean (i.e. the average occurrence rate of an item), such a model also uses predictor variables to describe variation in the text-to-text variability of occurrence rates. The output of such a model therefore not only includes (differences between) occurrence rates, but also (differences between) shape parameters. This means that location-scale models allow us to compare the distribution of a specific item in the target and reference corpus in terms of both frequency and dispersion.

To illustrate, consider the verb lemma DEFINE in published academic writing (target variety) vs. news texts (reference variety). The data, which are available via *TROLLing* (Sönning 2022c), are from the relevant sections of the *Corpus of Contemporary American English* (COCA) (Davies 2008-) and restricted to the year 2019. Our target corpus therefore includes 554 texts with an average length of about 9,000 words (4.96 million words in total), and the reference corpus includes 5,654 text files averaging 846 words in length (4.79 million words in total). We used the R package ‘gamlss’ (Rigby & Stasinopoulos 2005) to fit the model (for R code, see https://osf.io/tqchs).

Table 2 lists the estimates we obtain for these data.[4] Note that the coefficients are grouped into location and scale parameters. The first two columns give the model coefficients; for each parameter, we obtain an estimate and a standard error. The estimate is the model’s best guess at the value in the population represented by our data, i.e. the target and the reference variety of interest. The standard error is an indication of statistical uncertainty around this best guess. An approximate 95% confidence interval, for instance, covers parameter values within ±2 standard errors of the estimate (e.g. Gelman et al. 2021: 51).

Table 2. Negative binomial model of location and scale: Model coefficients and derived keyness metrics.

	Model coefficients		Keyness metrics	
Parameter	Estimate	(SE)		
Location			Rate (ratio)	95% CI
Academic writing	−8.3	(0.06)	246 pmw	[220; 276]
Newspapers	−10.4	(0.09)	31 pmw	[26; 38]
Difference (acad. − news)	2.1	(0.11)	7.8	[6.3; 9.7]
Scale			D_{NB}	95% CI
Academic writing	0.20	(0.10)	.56	[.49; .63]
Newspapers	1.88	(0.28)	.14	[.08; .23]
Difference (acad. − news)	−1.68	(0.29)	+.42	[.33; .53]

Let us consider the meaning of the estimates listed in Table 2. Since count regression models operate on the log scale, we cannot interpret these numbers directly. The value of −8.3 for academic writing, for instance, expresses the occurrence rate of DEFINE on the log scale; exponentiation ($e^{-8.3}$) yields, as a rate estimate, a probability of .000246, or 246 per million words. For newspaper writing, the occurrence rate is 31 pmw ($e^{-10.4}$ x 1,000,000). The difference between these two rates (2.1) is also on the log scale, so we again need to exponentiate ($e^{2.1}$) to obtain the rate ratio of 7.8. This indicates that the verb DEFINE is about 8 times more frequent in academic (compared to newspaper) writing. Table 2 also gives approximate 95% CIs for these frequency-related keyness metrics.

Next, we consider the gamma shape parameters. These are likewise expressed on the log scale, and exponentiation therefore yields estimates of 1.2 for ACAD ($e^{0.20}$) and 6.5 for NEWS ($e^{1.88}$). We can use Figure 4 to approximate the corresponding dispersion scores, which gives us D_{NB} values of roughly .60 (academic writing) and .15 (newspaper writing). These are quite close to the exact translations listed in Table 2. We obtain a difference of +.42 between these dispersion scores, which –

[4] To obtain the information listed in Table 2, in fact two regression models had to be specified. Thus, a regression model that contrasts two groups, i.e. including a single binary predictor, can only report two parameters – the third being implied by the other two.

though an abstract score – indicates an appreciably greater *generality* of the verb DEFINE in academic writing. Approximate 95% CIs for all estimates are also listed in Table 2.[5]

Let us finally obtain *NB-TD* estimates as a second measure of dispersion. As a yardstick, we consider a text of 10,000 words, a length that may be considered somewhat typical of academic research papers. As $NB\text{-}TD_{10k}$ estimates, we obtain proportions of .68 for academic writing and .16 for news writing. This suggests that the probability of at least one occurrence of DEFINE in a 10,000-word-long research paper is .68 (or 68%). For a newspaper text of the same length, this probability is .16 (or 16%). Uncertainty intervals for these measures are unfortunately quite hard to construct.

This gives us a set of regression-based metrics that address the four dimensions of keyness summarized in Section 2.1. Note that the measures we have developed are all descriptive in nature. Inferential information has (only) been given in the form of uncertainty intervals. While a purely inferential quantity similar to a *p*-value could also be found in regression output[6], we will not pursue this approach further here. For clarity of interpretation, we instead use descriptive indexes as ranking devices and draw our data visualizations around these. Confidence intervals are then added to offer an indication of the statistical uncertainty surrounding our estimates. We are now ready to use negative binomial regression for our case study, i.e. to identify key verbs in published academic writing.

4.2. Application to key verbs in academic writing

Instead of running regression models on all verb lemmas in the corpus, we restrict our inquiry to those verbs whose crude (i.e. bag-of-words, or whole-corpus) occurrence rate in the academic section of COCA (1990-2019) (i) exceeds 10 pmw, and (ii) is higher compared to that in the newspaper section of the corpus. In other words, we are setting lower bounds on the *discernibility* and *distinctiveness* of candidate items. This leaves us with a total of 578 verbs for consideration (data are available in Sönning 2022c). For illustration, we then looked at texts from the year 2019 only.

We started by running a regression model for each item and collected estimates and standard errors for model coefficients in a table. Based on these scores, we then constructed keyness metrics as described above. To rank items, we decided to give equal emphasis to the four keyness dimensions listed in Table 1. We started by computing ranks based on each keyness measure, which gave us four rank scores per item. We then averaged over these rank scores to obtain a mean rank for each item. Our list

[5] The CI for the difference between the dispersion scores cannot be based directly on the regression model since the standard error of the difference is only valid on the model scale, which in this case is the log gamma shape parameter. Once the shape parameter is translated into a dispersion score, the standard error for the difference cannot be straightforwardly applied to obtain a confidence interval on the dispersion difference. We used a simple inversion interval (see Newcombe 2013: 132) to construct a CI on the difference in dispersion. It takes as input the two dispersion values and their individual confidence limits.

[6] Thus, regression models report for each coefficient a test statistic (e.g. a t-value), which could be used as an inferential metric.

of 578 verbs was then ordered based on these average ranks. Note, however, that – rather than taking a simple average over the four metric-specific ranks – we could have weighted these dimensions differently if so required by our keyness analysis task.

When engaging with the list of ranked items, it is useful to have available information on each keyness dimension. We therefore use an information dashboard (cf. Few 2013) to display both frequency-related and dispersion-related scores. Figure 5 shows data for the top 30 verb lemmas; the corresponding values are given in Appendix 2. The purpose of a keyness dashboard is to have available, at a glance, the full array of information that helps us assess the (relative) typicalness of an item in the target text variety. To this end, we align four dotplots (see Sönning 2016), which show, in each row, the keyness report for an item and allow us to locate, in each column, exceptionally high (or low) scores on a particular dimension. The error bars denote approximate 95% CIs to give an indication of the statistical uncertainty surrounding our sample estimates.

Let us consider in some more detail how keyness information is arranged in Figure 5. The leftmost panel shows the rate ratio as an indication if an item's *distinctiveness*. A black vertical reference line marks a ratio of 1, which indicates the same rate of occurrence in both corpora. While the *x*-axis is log-scaled for better resolution, the numbers along the axis denote the original ratios to facilitate interpretation (see Schützler, to appear). The next panel reports on the *discernibility* of items, showing the log-transformed occurrence rate in the target corpus. Frequency is expressed in occurrences per 10,000 words of running text. Again, the *x*-axis is log-scaled and the numbers refer to the more tangible original rate ratios. Since a typical published research article is perhaps about 10,000 words long, a black reference line at 1 pttw marks an informative benchmark. Items to the right of this line have an expected frequency greater than 1 instance in such an article. The third panel shows D_{NB} as an estimate of dispersion. Recall that the scale is standardized to the unit interval [0;1], with greater values indicating a more even distribution across texts. For comparison, the light grey crosses indicate dispersion in the reference corpus. Finally, the rightmost panel shows the difference in D_{NB} scores. Since positive values indicate that an item is more evenly distributed in the target corpus, a reference line is added at zero, which marks the same level of *generality* in the two text varieties.

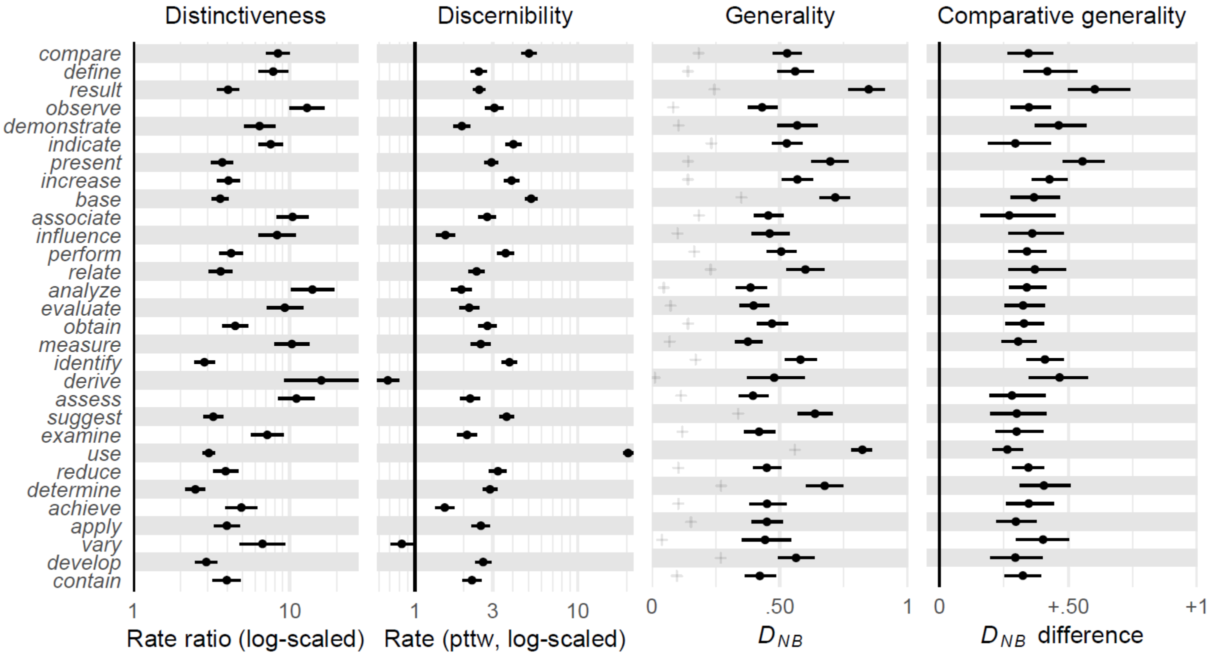

Figure 5. Keyness dashboard of the top 30 keyverb candidates in academic writing (compared to newspaper writing). Error bars denote approximate 95% CIs. 🅭🅯

To illustrate the richness of information provided by our keyness dashboard, let us consider two items: DERIVE (19th) and USE (23rd). Despite their similar overall rank, they show quite different keyness profiles. Thus, while DERIVE has higher *distinctiveness* (rate ratio: 15.9 vs. 3.0), its *discernibility* in the target variety is considerably lower (occurrence rate: 0.7 vs. 20.4 pttw). While its *generality* in the target variety is also lower (D_{NB}: .48 vs. 82), it exceeds USE in terms of *comparative generality* (+.47 vs. +.26). This is because USE is also quite common in the reference variety (.56); DERIVE, on the other hand, shows a very concentrated distribution in news writing (.01). Depending on the relative importance of these dimensions in our analysis task, our attention may be drawn more to particular features when studying a keyness dashboard.

5. Statistical properties of regression-based keyness indicators

We have argued that a key advantage of regression-based keyness indicators is that they come with indications of statistical uncertainty. In Figure 5, this information appears in the form of error bars reflecting approximate 95% CIs. In this section, we evaluate the quality of the error intervals provided by negative binomial regression models. Specifically, we test whether these intervals perform as intended, i.e. whether they protect against false conclusions at the stated level of probability. In Section 6.1, we describe our method and Section 6.2 then presents the results.

5.1. Method

In frequentist statistics, error probabilities describe long-run error rates of inferential procedures (Cox 2006: 8). A confidence interval therefore expresses a property of the method used to construct it (e.g. a regression model). To illustrate, let us assume that we are using an adequate model to study a simple quantity such as an occurrence rate (i.e. normalized frequency) in some text variety of interest (i.e. a population). If we draw 100 different samples from the same population, we obtain 100 different estimates of the true rate in the population. Each estimate comes with a 95% CI. If the model performs as intended, then roughly 95 of these CIs (i.e. 95% of them) will cover the true population rate. In the kinds of data settings for which a procedure is designed, i.e. which satisfy the assumptions of the statistical model, the method will work as intended. If the actual coverage rate of the procedure is lower, however, this would point to a mismatch between data and model and tell us that the tool we have chosen is inadequate for the inferential task at hand.

To assess the coverage properties of our procedure, we will turn to COCA as a sufficiently large corpus to simulate repeated keyness analyses of the same pair of text varieties. Our target variety will be published academic writing, and the reference domain will be the newspaper section of the corpus. We randomly divide the corpus, which covers a period of 30 years (1990-2019), into 100 subcorpora that are, however, balanced on year and genre; each text file then occurs in only one subcorpus (see Sönning 2022c for the data used, and Sönning 2022a for further methodological details). Our illustrative analysis is concerned with identifying typical verb lemmas in academic writing. For a selected set of 578 items, we then compute keyness indicators based on each subcorpus, along with 95% CIs. If our method works as intended for a specific item, we would expect around 95 of the 100 intervals to cover the true value in the population. Since we do not know the true value, we take the median over the 100 values as an approximation to it.

To illustrate, consider the occurrence rate estimate for the verb DEFINE in published academic writing. Figure 6 shows the set of 100 estimates (dots), one for each partition of COCA, with approximate 95% CIs. The occurrence rate estimates for DEFINE vary between roughly 1.5 and 2.5 per 10,000 words (pttw). The grey horizontal line represents the median over these 100 values (about 2 pttw) and will serve as a proxy for the occurrence rate in the population. Estimates whose interval does not cover the grey line are shown in black. Overall, 97 of the 100 intervals include the median rate, and the coverage is therefore at 97%, in the ballpark of the nominal level. This percentage summarizes the coverage property of this particular keyness measure (occurrence rate) for this particular item (DEFINE). If we repeat this procedure for all 578 items, and then do the same for the other keyness measures, we obtain, for each metric, a distribution of coverage rates.

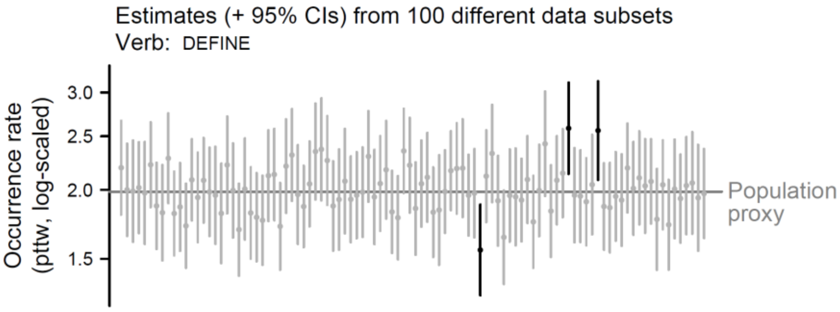

Figure 6. Illustration of the coverage property of the 95% CIs for the occurrence rate estimate for *define* in academic writing. Error bars reflect approximate 95% CIs. 🅭🅯

5.2. Results

Figure 7 shows the distribution of the observed coverage rates for six parameters of the negative binomial regression model. From top to bottom, these are the estimated occurrence rate in the target corpus (ACAD) and in the reference corpus (NEWS), and the rate ratio as a comparative measure. Listed next are the gamma shape parameters for each corpus and then again a contrastive measure. Each histogram summarizes 578 coverage rates (i.e. percentages), one for each verb in the data. The percentages hover around the nominal level of 95%, which is marked with a black vertical line. Despite a thin tail extending below 85% and a handful of deviant coverage rates below 70%, the results seem quite satisfactory overall. This suggests that the negative binomial regression model largely succeeds in generating uncertainty intervals with the intended performance, and that the model has some value for representing distributional features of verb lemmas in the two text varieties under study.

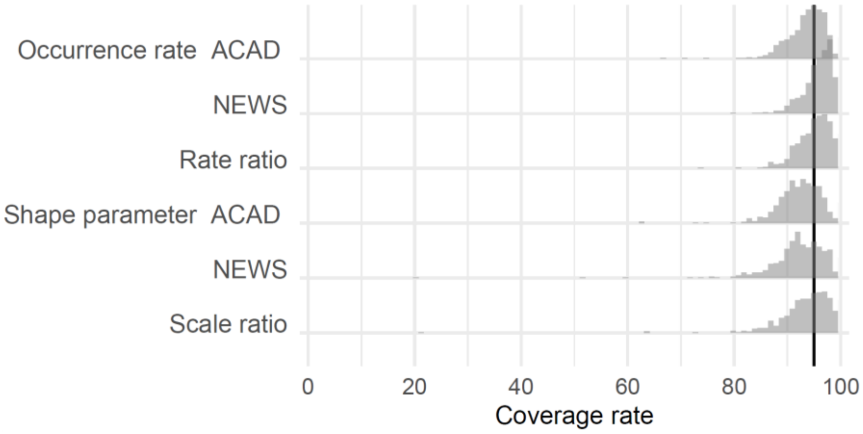

Figure 7. Distribution of the 578 coverage rates (one per verb lemma) for 95% CIs on each coefficient in the negative binomial regression model. The vertical line at 95 marks the nominal level of 95%. 🅭🅯

6. Advantages of keyness regression

Given that a large variety of keyness metrics already exists, we must ask what keyness regression has to offer over and above familiar alternatives. In this section, we recapitulate advantages of the approach.

Let us first consider the nature of the statistical model that is used for analysis. Thus, negative binomial regression is a text-level procedure, which means that the unit of sampling (texts or text excerpts) coincides with the unit of analysis (for further details, see Lijffijt et al. 2014; Sönning 2022a). The relevant sample size for generalizations from sample to population is therefore the number of texts (rather than the total number of words) in the corpus. This improves the quality of statistical inferences, and indeed, as our coverage analysis in the previsou section showed, we can put some faith in the statistical uncertainty indications returned by this model.

Since a regression model partitions a data summary into descriptive and inferential quantities (estimates and standard errors), it does not force us to opt for one particular form of evidence. Rather, we can use either kind of statistical information to rank candidate items and may then choose to incorporate both quantities when engaging with the list of candidate items. For a detailed study of keyness profiles, we have illustrated how keyness regression metrics can be visualized using information dashboards. Arguably, however, an in-depth analysis of candidate lists should center on descriptive indicators, since these are easier to interpret. This assists in making informed decisions about cut-off values and helps us appraise the relative typicalness of items (see Sönning 2022a).

As for questions of interpretability, we have seen that count regression models are able to offer tangible metrics, which are directly meaningful and accessible to a wider linguistic audience (cf. Egbert, Larsson & Biber 2020: 24-32). In particular, the rate ratio gives a transparent reflection of distinctiveness, and *NB-TD* also has a straightforward interpretation (see Sönning 2022a). D_{NB} as a measure of dispersion, on the other hand, is an abstract metric. Since it is standardized to the unit interval [0;1], however, its interpretation parallels that of other indexes of lexical dispersion (see Gries 2020; Sönning 2022b).

For the measurement of (lexical) dispersion, regression-based measures also show some favorable properties compared to existing indexes. Let us first consider D_{NB} as a standardized dispersion estimator. As a proof of concept, we apply D_{NB} to a set of 150 items in the BNC, which has been assembled by Biber et al. (2016) to study the behavior of dispersion indexes (see Sönning 2022b for further details; the data are available in Sönning 2022d). These items cover a broad range of frequency and dispersion levels. We apply D_{NB} along with other dispersion measures to the 150 items across the 4,098 text files in the BNC. Figure 8 shows the distribution of dispersion estimates for each metric. Note that D_{NB} covers the entire unit interval, with scores piling up at the extremes. For most keyness analysis tasks, this behavior is arguably desirable. This is because the most highly dispersed forms in the set of 150 items are function words (e.g. *a*, *but*, *with*), which are usually excluded from consideration in keyword analysis. At the lower end of the scale, we find infrequent items (e.g. *hm*, *nought*, *corp*), which may also not of primary interest due to their low level of *discernibility*. Thus, it seems that D_{NB} provides good resolution in those parts of the dispersion spectrum that are likely to be of greatest interest to keyword analysts.

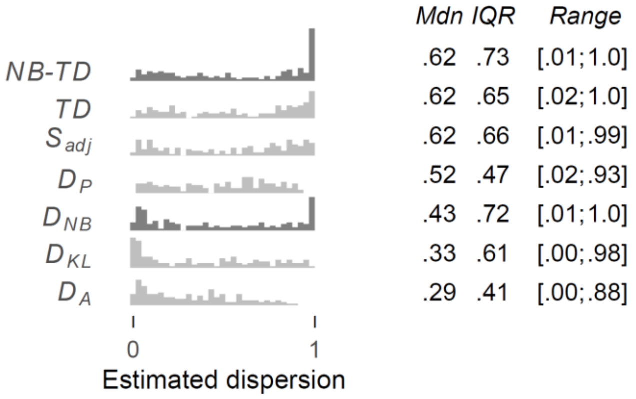

Figure 8. Performance of D_{NB}, *NB-TD* and other dispersion indexes on Biber et al.'s (2016) set of 150 items in the BNC. The items represent a broad range of frequency and dispersion levels. 🅭🅯

A general feature of dispersion measures is that they are correlated with the frequency of an item (see, e.g. Gries 2020). While this association cannot be overcome completely,[7] a multidimensional keyness analysis would give preference to dispersion measures that are better able to separate *discernibility* and *generality*. This is because, when decomposing keyness into the four dimensions discussed above, we would like each metric to ideally offer a pure reflection of a single dimension. To assess the strength of association between dispersion and frequency measures, we again turn to the set of 150 items assembled by Biber et al. (2016). Figure 9 shows, for a variety of dispersion indexes, the association between the log frequency of an item (*x*-axis) and its standardized dispersion score (*y*-axis). Each panel therefore includes 150 points, one for each item. While a positive correlation is apparent for all measures, we observe that the strength of association varies. To facilitate comparisons, dispersion metrics have been ordered according to the strength of correlation with frequency, which inreases from left to right.

[7] Consider, for instance, a simple setting: A corpus of 10 texts, each 1,000 words long. If an item occurs fewer than 10 times in the corpus, it cannot reach a dispersion score of 1, even if its occurrence rate is perfectly balanced across those texts in which hit does occur (i.e. 1 instance of the item in each of these texts).

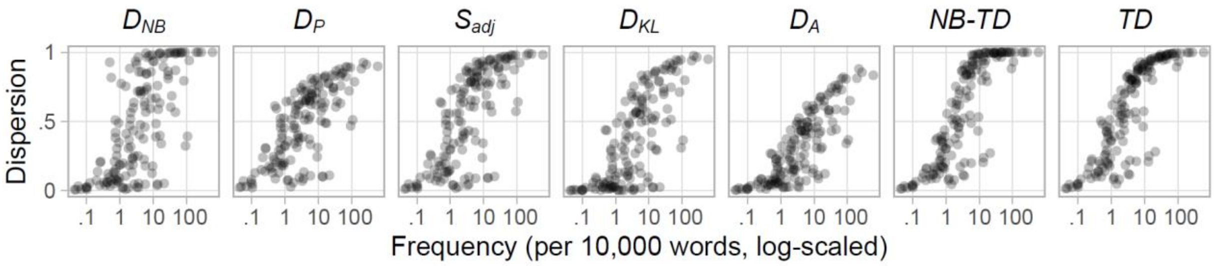

Figure 9. Scatterplot showing, for different dispersion measures, the association between the logged frequency and dispersion for Biber et al.'s (2016) set of 150 items in the BNC.

A summary of the association patterns appears in Figure 10, which reports Pearson correlation coefficients (with 95% CIs) for the measures. We note that D_{NB} shows the weakest link with frequency (r = .69), followed by D_P, S_{adj}, and D_{KL} ($r \approx .74$). D_A, *NB-TD* and *TD* are more strongly correlated with frequency ($r \approx .83$). Thus, if we were to look for a dispersion measure that tries to isolate an item's *generality* for these 150 items in the BNC, we would perhaps turn to D_{NB}. *TD*-related measures, on the other hand, blend frequency more noticeably into the assessment. This may be undesirable if keyness profiles are intended to keep apart information on frequency and disperison.

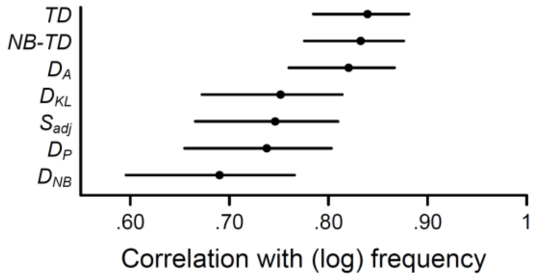

Figure 10. Summary of the association between logged frequency and dispersion for the set of 150 items in the BNC: Pearson correlation coefficients for different dispersion measures. Error bars reflect 95% CIs.

We also introduced *NB-TD*, a regression-based variant of *TD* that allows us to adjust for differences in text length. Since the value of *TD* depends in part on the length of text files (cf. Gries 2021; Sönning 2022b), scores are not directly comparable if text collections differ in average length. In our illustrative keyverb analysis, the average text length in the target corpus (about 9,000 words) exceeds that in the reference corpus (just under 900 words) by more than a factor of 10. It follows that the *TD* score for academic writing will be upwardly biased, which leads to systematic distortions in measures expressing *comparative generality*.[8] As a result, these scores cannot be taken at face value. It should

[8] The same applies to the TD-based likelihood-ratio test proposed by Egbert & Biber (2019).

be noted, however, that differences in average text length do not systematically alter the ranking of items. If it is the sole purpose of *TD* to serve as a ranking device, biased estimates are therefore a minor concern. However, if we study descriptive keyness profiles (see Figure 5) or introduce cut-offs on the *TD* (difference) score, this bias has the potential to alter our final set of key items. *NB-TD*, on the other hand, is adjusted for (potential) differences in text length, which protects our measures of (*comparative*) *generality* from bias.

7. Disadvantages of keyness regression

Our discussion of count regression models for keyness analysis would be incomplete without a consideration of their limitations. As we will see, these mainly concern the effort and expense involved in their implementation.

To carry out the calculations that are necessary to construct the keyness metrics we have developed in this paper, we must rely on specialized statistical tools. Thus, to run count regressions, a statistical software package is required. In fact, the location-scale model we have been using here is an extension of the generalized linear model that has only recently become available to data analysts. To have the ability to model both the average occurrence rate and the variability in text-specific rates, we must turn to purpose-built implementations such as the R package 'gamlss' (Stasinopoulos et al. 2017). This means that keyness regression can only be implemented with the help of a specialized suite of functions.

A further caveat is that a negative binomial model of location and scale requires (i) a sizeable amount of data to converge and generate sensible inferences; and (ii) quits with an error message if the number of instances per text does not exceed 1 in one of the corpora. For our keyverb regression analysis of the 2019 data from COCA, one item (*reside*) had to be excluded from the analysis because a regression model could not be computed for the reference corpus.[9] Similar problems were encountered in our study of the coverage properties using random subdivisions of COCA. Thus, about 0.1% of the regression models did not converge since the data failed to meet the demands of the procedure. This means that keyness regression is likely to be infeasible with smaller corpora, and especially for corpora consisting of relatively short texts.

A final admonition is the computational expense involved when running keyness regressions. Thus, for our set of 578 verb lemmas, we needed to run a sequence of 578 pairs of regressions. This took about 15 minutes on a personal computer, which, given today's standards in computing power, may severely try the patience of 21st-century corpus linguists.

[9] Thus, while *reside* occurred in 37 (out of 5,654) texts, it never appeared more than once.

8. Conclusion and outlook

The aim of the present paper was to explore the use of regression modeling for keyword analysis. We started by decomposing keyness into four recognizable dimensions, which provide complementary indications of typicalness. Based on these dimensions, we arranged existing keyness metrics into four classes. We then demonstrated how a negative binomial location-scale model can be used to build metrics addressing all four dimensions of keyness with accompanying 95% CIs as indications of statistical uncertainty. Arranged in a keyness dashboard, these quantities allow for a detailed study of keyness profiles, where the analyst may decide to vary the emphasis given to different dimensions.

While keyness regression does offer a number of attractive features, the caveats we have outlined weigh quite heavily. This especially concerns the routine implementation of this approach, since most software packages for corpus analysis require metrics that can be implemented without recourse to specialized statistical analysis suites. It has therefore become apparent that count regression does not qualify as a general keyness analysis procedure.

However, considering its attractive features, it should not be discarded entirely as a method for identifying typical items in a target corpus. Considering the computational burden imposed by keyness regression, a sensible general strategy would be to rely on simpler metrics as a screening device, to arrive at an intermediate, preliminary set of items (perhaps between 100 and 1,000 forms). These could then be analyzed in more detail using the methods described in this paper.

The feasibility of the regression approach also depends on the nature of the analysis task. Thus, it is unlikely to be of interest in settings where a keyness analysis plays a subordinate role in the overall line of argument, for instance if it is merely meant to provide supplementary observations in the context of, say, a multi-method discourse analysis. For other purposes, however, where a well-selected final set of items constitutes the primary objective of a study, the method may have more value. Thus, if considerable resources are devoted to the careful production of a custom-made vocabulary list, for instance (e.g. Thorndike 1921; Paquot 2010), analysts may appreciate the richness of information provided by a count regression models. Such studies will also welcome the favorable performance of inferential uncertainty intervals, since it is their aim of to make generalized statements about a particular domain of language use. In those settings, researchers will also have a clear idea of how to handle and combine statistical information relating to different dimensions. When constructing key verb lists for academic writing, for instance, some verbs will be quite strongly associated with certain methodological paradigms (e.g. MEASURE in Figure 5). A multidimensional approach to keyness allows us to recognize items with specialized usage, which will rank high in terms of *distinctiveness*, but show relatively low levels of *generality*. The relative weight assigned to different dimensions of keyness will then determine the overall level of typicalness of such items, and whether they will be included in a final set of items.

Finally, regression may be used to study the behavior of other keyness metrics, in order to understand to what extent they offer blended assessments of *discernibility*, *distinctivness*, and (*comparative*) *generality*. This can be observed by looking at the correlation between the item ranking generated by a specific metric and the ranking produced by regression-based metrics addressing different keyness dimensions. This allows us to see which distributional features a metric (primarily) responds to.

We hope to have illustrated the utility of regression modeling for keyness analysis. Considering the additional time and effort required to implement this approach, however, it would seem necessary to obtain further support for the favorable performance of this technique on corpus data. It would therefore be desirable for future research to test the technique on other language structures and corpus designs.

References

Baker, Paul. 2004. Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics* 32 (4), 346–59. doi:10.1177/0075424204269894

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

Biber, Douglas, Randi Reppen, Erin Schnur & Romy Ghanem. 2016. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics* 21(4), 439–464. doi:10.1075/ijcl.21.4.01bib

Brezina, Vaclav. 2014. Effect sizes in corpus linguistics: Keywords, collocations and diachronic comparison. Presented at the ICAME 2014 conference, University of Nottingham.

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. doi:10.1075/ijcl.19.1.01bre

Cameron, A. Colin & Pravin K. Trivedi. 2013. *Regression analysis of count data*. Cambridge: CUP.

Cox, David R. 2006. *Principles of statistical inference.* Cambridge: Cambridge University Press.

Cumming, Geoff. 2012. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, Geoff & Robert Calin-Jageman. 2017. *Introduction to the new statistics: Estimation, open science, and beyond.* New York: Routledge.

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. doi:10.5555/972450.972454

Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analysis. *Corpora* 14(1), 77–104. doi:10.3366/cor.2019.0162

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. doi:10.1515/zaa-2006-0208

Few, Stephen. 2013. *Information dashboard design*. Burlingame, CA: Analytics Press.

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

Gabrielatos, Costas & Anna Marchi. 2011. Keyness: Matching metrics to definitions. *Corpus Linguistics in the South* 1, University of Portsmouth, 5 November 2011. Available online at: http://eprints.lancs.ac.uk/51449

Gabrielatos, Costas & Anna Marchi. 2012. Keyness: Appropriate metrics and practical issues [Paper presentation]. CADS International Conference 2012, University of Bologna, Italy. https://www.researchgate.net/publication/261708842_Keyness_Appropriate_metrics_and_practical_issues

Gelman, Andrew, Jennifer Hill & Aki Vehtari. 2021. *Regression and other stories.* Cambridge: Cambridge University Press. Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*. New York: Springer, 99–118.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), 403–37. doi:10.1075/ijcl.13.4.02gri

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. doi:10.32714/ricl.09.02.02

Gries, Stefan Th. 2022. Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics* 1(1), 100002. doi:10.1016/j.rmal.2021.100002

Hardie, Andrew. 2014. Log ratio – An informal introduction. Post on the website of the ESRC Centre for Corpus Approaches to Social Science CASS. Available online at: http://cass.lancs.ac.uk/?p=1133

Hilbe, Joseph M. 2011. *Negative binomial regression*. Cambridge: Cambridge University Press.

Hilbe, Joseph M. 2014. *Modeling count data*. Cambridge: Cambridge University Press.

Hofland, Knut & Stig Johansson. 1982. *Word frequencies in British and American English*. London: Longman.

Kilgarriff, Aadam. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In L. J. Evett & T. G. Rose (eds.) *Language Engineering for Document Analysis and Recognition* (*LEDAR*). AISB96 Workshop proceedings, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK, 33–40.

Kilgarriff, Adam. 2009. Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference*, *CL2009*. Liverpool: University of Liverpool. Available online at: http://ucrel.lancs.ac.uk/publications/CL2009/171_FullPaper.doc

Kline, Rex B. 2013. *Beyond significance testing*. Washington, DC: American Psychological Association.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamaki & Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu064

Long, J. Scott. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.

Long, J. Scott & Jeremy Freese. 2014. Regression models for categorical dependent variables using Stata. College Station: Stata Press.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations. International Journal of Corpus Linguistics 22(3). 319–44.

McElreath, Richard. 2020. Statistical rethinking: A Bayesian course with examples in R and Stan. Boca Raton, FL: CRC Press.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mosteller, Frederick, and David L. Wallace. 1984. Applied Bayesian Inference: The Case of The Federalist Papers. New York: Springer.

Nelder, J.A. and Wedderburn, R.W.M. 1972. Generalized linear models. Journal of the Royal Statistical Society A 135(3), 370–384.

Paquot, Magali. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.

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. doi:10.1515/cllt-2015-0030.

Rayson, Paul. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University.

Rayson, Paul & Amanda Potts. 2020. Analyzing keyword lists. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*. Berlin & New York: Springer, 119–139.

Rigby, Robert A. & Mikis D. Stasinopoulos. 2005. Generalized additive models for location, scale and shape (with discussion), *Applied Statistics* 54(3): 507–554.

Schützler, Ole. To appear. Frequencies in corpus linguistics: Issues of scaling and visualization. In Lukas Sönning & Ole Schützler (eds.) *Data visualization in corpus linguistics: Reflections and future directions*, VARIENG.

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

Sönning, Lukas. 2016. The dot plot: A graphical tool for data analysis and presentation. In Hanna Christ, Daniel Klenovšak, Lukas Sönning & Valentin Werner (eds.), *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, 101–129. Bamberg: University of Bamberg Press.

Sönning, Lukas. 2022a. Evaluation of keyness metrics: Reliability and interpretability. *PsyArXiv preprint*. https://psyarxiv.com/eb2n9/

Sönning, Lukas. 2022b. Evaluation of text-level measures of lexical dispersion: Robustness and consistency. *PsyArXiv preprint*. https://psyarxiv.com/h9mvs/

Sönning, Lukas. 2022c. Key verbs in academic writing: Dataset for "Evaluation of keyness metrics: Reliability and interpretability", https://doi.org/10.18710/EUXSMW, DataverseNO, DRAFT VERSION

Sönning, Lukas. 2022d. Biber et al.'s (2016) set of 150 BNC items for the analysis of dispersion measures: Dataset for "Evaluation of text-level measures of lexical dispersion", https://doi.org/10.18710/MNVB36, DataverseNO, DRAFT VERSION. An anonymized version of the dataset is available at https://dataverse.no/privateurl.xhtml?token=a25d30a0-6067-4989-837a-19468c9fa661.

Sönning, Lukas, and Manfred Krug. 2021. Actually in Contemporary British Speech: Data from the Spoken BNC Corpora. https://doi.org/10.18710/A3SATC. DataverseNO, V1, UNF:6:rp13HUEAY75735Bcul7eCg== [fileUNF]

Sönning, Lukas & Manfred Krug. 2022. Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In Ole Schützler & Julia Schlüter (eds.), *Data and methods in corpus linguistics: Comparative approaches*. Cambridge: Cambridge University Press, 127–160. doi:10.1017/9781108589314.006

Thorndike, Edward L. 1921. *The teacher's word book*. Teachers College, Columbia University.

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. 2020. Statistics for Linguistics. New York: Routledge.

Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, e12439. https://doi.org/10.1111/lnc3.12439

Appendix

Appendix 1. Conversion of gamma shape parameter ϕ into a standardized dispersion measure.

Formula: $1 - e^{-\frac{1}{\phi}}$

R code:

```
1 - exp(-1/shape)
```

Appendix 2. Keyness profile for the top 30 verb lemmas.

		Frequency (rate, pttw)			Dispersion (D_{NB})		
	Verb	Acad.	News	Ratio	Acad.	News	Diff.
1	*compare*	5.02	0.60	8.4	.53	.18	+.35
2	*define*	2.46	0.31	7.8	.56	.14	+.42
3	*observe*	3.08	0.24	12.9	.43	.08	+.35
4	*result*	2.48	0.61	4.0	.85	.24	+.60
5	*demonstrate*	1.94	0.30	6.4	.57	.11	+.46
6	*indicate*	4.02	0.53	7.6	.53	.23	+.29
7	*present*	2.95	0.80	3.7	.70	.14	+.55
8	*increase*	3.92	0.97	4.1	.57	.14	+.43
9	*base*	5.17	1.44	3.6	.72	.35	+.37
10	*associate*	2.78	0.27	10.4	.46	.18	+.27
11	*influence*	1.54	0.19	8.3	.46	.10	+.36
12	*perform*	3.60	0.85	4.2	.51	.17	+.34
13	*relate*	2.39	0.66	3.6	.60	.23	+.37
14	*analyze*	1.93	0.14	13.9	.39	.05	+.34
15	*evaluate*	2.15	0.23	9.3	.40	.07	+.32
16	*obtain*	2.79	0.62	4.5	.47	.14	+.33
17	*measure*	3.80	1.33	2.9	.58	.17	+.41
18	*identify*	2.53	0.25	10.3	.38	.07	+.31
19	*derive*	0.68	0.04	15.9	.48	.01	+.47
20	*assess*	2.18	0.20	11.0	.40	.11	+.28
21	*suggest*	3.65	1.13	3.2	.64	.34	+.30
22	*examine*	2.09	0.29	7.2	.42	.12	+.30
23	*use*	20.40	6.71	3.0	.82	.56	+.26
24	*reduce*	3.23	0.83	3.9	.45	.10	+.34
25	*determine*	2.89	1.16	2.5	.68	.27	+.40
26	*achieve*	1.52	0.31	4.9	.45	.10	+.35
27	*apply*	2.54	0.64	4.0	.45	.15	+.30
28	*vary*	0.83	0.12	6.7	.44	.04	+.40
29	*develop*	2.62	0.89	2.9	.56	.27	+.29
30	*contain*	2.24	0.56	4.0	.42	.10	+.32