

Not just frequency

Keyness should integrate frequency,
association, and dispersion

Stefan Th. Gries

UC Santa Barbara | JLU Giessen

For decades, nearly all approaches to keyness analysis in corpus linguistics have been based on computing for each word type in question a single statistic – usually, the log-likelihood score G^2 – and ranking word types by how key that statistic made a word type for a target corpus T . In this paper, I discuss a new approach to keyness that (i) uses three dimensions of information (frequency in T , association to T , and dispersion in T relative to R and that (ii) measures both association and dispersion using the information-theoretic measure of the Kullback-Leibler divergence. I outline the computational steps and provide R code in a markdown document as well as a ready-made R function `Keyness3D` with which readers can conduct analyses of their own data. I exemplify the use of the function and its results using the learned text category in the Brown corpus against the rest.

Keywords: keyness, log-likelihood G^2 , frequency, association, dispersion, KL-divergence

1. Introduction

1.1 General introduction

One of the most central corpus-linguistic methods is keyness analysis, i.e. the identification of typically, but not necessarily, words that are **KEY** for a certain topic, register, genre, variety, etc.; thus, keyness is generally less of a concern for theoretical linguistics, but has been extremely important in many applied linguistics contexts, be it for:

- the development of vocabulary lists for students wishing to be able to focus their vocabulary studies on words that are particularly relevant to a particular topic;

- the comparisons of corpora representing different varieties or cultures, a field of study initiated by Hofland & Johansson’s (1982) and Leech & Fallon’s (1992) comparisons of keywords representative of British vs. American culture.

Computing keyness requires a target corpus T , which represents the topic, register, ... of interest and a reference corpus R , which typically is much larger and serves as a standard of comparison; often R is considered representative for ‘a language as a whole,’ but it can also be representative of a specific other topic, register, ... of interest to which T is supposed to be compared. Most keyness studies share a kind of methodological core and then differ with regard to how keyness is operationalized quantitatively. Specifically, the vast majority of applications are based on creating, for each word type represented in the combination of T or R , a two-by-two table of the kind represented in **Błąd! Nie można odnaleźć źródła odwołania.**, where a is the frequency of the word type in T , b is the frequency of the word type in R , $a+c$ is the size of T , and $b+d$ is the size of the reference corpus R .

Replace the bold Polish warning

Table 1. Schematic representation of the input to most keyness computations

	Corpus: T	Corpus: R	Sum
Word type w	a	B	$a+b$
All others, w	c	D	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d$

Replace the bold Polish warning

No doubt aided by the convenience of being able to use a ready-made software tool such as WordSmith Tools, the vast majority of studies have used the log-likelihood ratio G^2 as a measure of keyness. This measure involves the following steps: (i) one creates the table of observed frequencies such as **Błąd! Nie można odnaleźć źródła odwołania.**; (ii) one computes the table of expected frequencies from it (as one would for a chi-squared test); (iii) one computes G^2 using the formula shown in (1). These computations would be undertaken for each word type. The word types would then be ranked by their G^2 -values in decreasing order, and one would inspect/interpret the top x words.

(1) $G^2 = m \sum_a^d \left(obs \times \ln \frac{obs}{exp} \right), \text{ within } = \begin{cases} 2, & \text{if } a_{obs} > a_{exp} \\ -2, & \text{if } a_{obs} < a_{exp} \end{cases}$

Many other statistics, all to be run on such two-by-two tables, have been suggested – the odds ratio, the chi-squared statistics, the relative frequency ratio, the difference coefficient, a so-called %DIFF score; see Pojanapunya & Watson Todd (2018), Rayson & Potts (2020), and Gries (2024) for recent surveys. But there are also interesting alternatives, the two most interesting of which take dispersion into

deleted

Replace

Is this ln and not in?

consideration. One is Paquot & Bestgen's proposal to not use aggregate frequencies for both all of T and all of R , but to use file-/text-specific frequencies instead, which could then be compared with, say, a t -test. In their comparative evaluation, Paquot & Bestgen (2009) compute each word type's mean relative frequency per corpus part/file in T and again per corpus part/file in R and then quantifies its preference for either corpus with a standard t -test. Another recent proposal is Egbert & Biber (2019), who stick with G^2 on a table such as **Błąd! Nie można odnaleźć źródła odwołania.**, but propose to replace the frequencies in it by range values, i.e. the number of parts of T and R that a word type of interest occurs in at least once.

To my mind, all these proposals share one or more specific shortcomings. First, they very much underutilize the amount of information our corpora offer because they all provide only a single keyness value that is somehow supposed to comprehensively express the construct of keyness, which is strange given how much applied linguists in general acknowledge the multidimensional nature and corresponding complexity of their constructs.

Second, G^2 as a keyness statistic is actually very hard to interpret. Not only is it highly correlated with raw frequency of occurrence (either in the a -cell of **Błąd! Nie można odnaleźć źródła odwołania.**, the difference of $a-b$, or the overall frequency of the word type $a+b$), it also conflates frequency and association in a non-intuitive way. For example, the degree to which G^2 is actually just a function of $a-b$ is determined by how similar $a+c$ and $b+d$ are, and the correlation between any frequency values or their ratios on G^2 is highly curvilinear. This conflation is well-known from many quantitative applications in corpus linguistics trying to measure a complex construct with just one number.

1.2 Overview of the present paper

This paper attempts to address these two shortcomings and improve a previous suggestion along these lines (Gries 2021) in terms of both implementation and practicality. The proposed approach will be demonstrated hands-on on the basis of a tiny corpus but this paper also provides an R function allowing readers to very quickly do keyness analyses that are much more comprehensive than what has previously been available. Section 2 will introduce the overall method; §3 will showcase results from some small applications; §4 will conclude.

Replace the bold Polish warning

2. Methods

The proposal of this paper is to make keyness a three-dimensional construct – three dimensions because that provides the best answer to the question of ‘what makes a word w a good keyword for a target corpus T ?’ I submit the answer is: w is a good keyword for T (i) if w occurs frequently in T , (ii) if w is attracted to T (much like we use association in studies of collocation), and (iii) if w occurs very evenly throughout T and not at all or very clumpily in R . Previous work has sometimes recognized the relevance of these dimensions but the dominant use of G^2 has meant that most studies have only considered an implicit conflation of much of (i) and a bit of (ii) – utilizing all three dimensions explicitly and separately is a new approach. In what follows, I will outline and demonstrate how we measure and include these three dimensions in a way that measures each relatively independently of the others (i.e., each dimension makes its own meaningful contribution) but allow for conflation but in a principled way that provides a huge improvement over the current ways.

Replace <2025...zip> by the following:2026_STG_No

2.1 Data

We begin by creating two tiny artificial corpora, which are presented as a word type-by-corpus matrix in **Błąd! Nie można odnaleźć źródła odwołania.**; R resources and code are available at https://stgries.info/research/2025_STG_KeynessAs3D_QualicoVol.zip. This means, for example, that the ‘word’ b occurs 5 times in T and 4 times in R .

Replace

Table 2. A small word-by-corpus matrix (40 tokens in T , 40 tokens in R , 12 different types)

	<i>a</i>	<i>B</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>x</i>	<i>y</i>	<i>z</i>	Sum
<i>T</i>	5	5	3	1	2	2	3	3	4	9	0	3	40
<i>R</i>	5	4	3	3	4	3	2	1	2	10	3	0	40
Sum	10	9	6	4	6	5	5	4	6	19	3	3	80

2.2 The three components of keyness

2.2.1 The frequency component

The frequency component of keyness is very straightforward: it’s the frequency of each word type in T , add 1 (to avoid problems with os in the next step), take the binary log of the values, and min-max transform the value for all words with a custom-made function zerozone defined in the code file. The min-max trans-

a smallbis

formation converts all values proportionally to a range from 0 to 1, which means all words not occurring in T score a value of 0 and the most frequent word in T scores a value of 1; here, y scores 0 for T while x scores 1.

2.2.2 The association component

The association component is computed using the Kullback-Leibler divergence (KLD), an information-theoretic measure that quantifies how much one probability distribution (the posterior) diverges from another one (the prior); the KLD is computed as shown in (2).

(2)
$$KLD = \sum posterior \times \log_2 \frac{posterior}{prior}$$

Two possible KLD directions could be computed, but we will use the version that treats the proportional distribution of a word type over T and R as the posterior (e.g. 5 vs. 4 for b) and the proportional sizes of T and R as the prior (i.e., 40 vs. 40). Since the KLD falls into the interval $[0, +\infty)$, we normalize it with the odds-to-probability transformation ($KLD / (1 + KLD)$) and multiply it by -1 if the word types frequency in T is smaller than expected given the prior. Finally, we stretch the resulting values such that word types attracted to T receive values from $(0, 1]$, with values repelled by T receiving correspondingly stretched negative values. Here, z scores the highest value of 1 (all its occurrences are in T), a scores a value of 0 (its distribution across T and R corresponds to the prior), and y scores a value of -1 (all its occurrences are in R).

change

2.2.3 The dispersion component

For the dispersion component, we need a finer resolution. For each corpus, we need to know how often each word type occurs in each part and what the sizes of the corpus parts are; for T this is represented in Błąd! Nie można odnaleźć źródła odwołania..

Replace

Table 3. A small word-by-part matrix (T shown here transposed, 40 tokens in T , 11 different types)

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>x</i>	<i>z</i>	Sum
Tar1	2	4	0	0	0	0	0	0	0	2	1	9
Tar2	1	0	1	1	0	1	1	3	1	1	0	10
Tar3	1	1	1	0	0	1	1	0	2	2	1	10
Tar4	1	0	1	0	2	0	1	0	1	4	1	11
Sum	5	5	3	1	2	2	3	3	4	9	3	40

change to:

We can then quantify the dispersion of each word type over the parts of, here, T again with the KLD : the proportional distribution of each word type over the corpus part becomes the posterior and the proportions the corpus parts **makeup** of the corpus become the prior; with this, the KLD for a is ≈ 0.111 . We then again normalize with the odds-to-probability transformation ($KLD / (1 + KLD)$), min-max transform the values to the interval $[0, 1]$, and then subtract the value from 1 (so that high/low values mean even/uneven dispersion respectively). Then, we do the exact same for R and then subtract the dispersion values for R from those for T (where dispersions for word types not attested in a corpus are set to 0). This way, after everything, high values represent word types that are very evenly distributed in T and very clumpily distributed or unattested in R .

2.3 What to do with those values?

2.3.1 *Keeping dimensions separate*

In the best of cases, a researcher would then explore the three different dimensions to be able to shed light on which word types are key for T because of a strong association or a strong dispersion component or both. I will exemplify this in § 3.

2.3.2 *Amalgamations*

While keeping the different dimensions separate is most informative (in the sense of losing least information), one can try to combine the different dimensions. Two involve the notion of using the frequency component to weigh either the association component (see (3)) or both the association and dispersion components (see (4)).

$$(3) \text{ AMALGAM}_1 = \text{dispersion} + \text{association} \times \text{frequency}$$

$$(4) \text{ AMALGAM}_2 = (\text{dispersion} + \text{association}) \times \text{frequency}$$

A third possibility is to compute the Euclidean distance from the origin of the space of the three dimensions of frequency, association, and dispersion to the position of each word type in that space (see (5)).

$$(5) \text{ EUCLID} = \sqrt{\text{frequency}^2 + \text{association}^2 + \text{dispersion}^2}$$

One potentially particularly interesting aspect of these three approaches is that they allow the researcher to make motivated decisions about how to weigh each of these dimensions in the amalgamation. Unlike with any existing keyness statistic, a researcher can now decide to prioritize one dimension, e.g. association, for key-

ness by making it twice as important as the others by multiplying all the values on that dimension by 2; we will see an application of this in §3 below.

A final different possibility would be to compute every word type's Mahalanobis distance (see Manly & Alberto 2016:87–88) from the distribution of all word types in the two-dimensional space of association and dispersion. That distance could be better for keyness than the Euclidean distance because its distance computation would take the spread of the dimensions' values and their covariance in consideration, which in R can be computed very easily with a function called `mahalanobis`.

3. Case study: 'Learned' in Brown

changeand Frown co

This paper comes with the above mentioned downloadable zip archive, which includes a knitted Quarto document with all the code for §2 and §3, an R function `Keyness3D`, which performs all computations discussed here, and the `Brown` and `Frown corpora` as `.rds` files in an input that facilitates the use of `Keyness3D`. With that function, conducting the analyses proposed here is very simple. First, one sources the function into R and loads the whole corpus — i.e., *T* and *R* in one data frame — into R. Since the learned category is represented in the corpus by part names beginning with “j” so we make all of those *T* and everything else *R*, and apply the function:

insert line break before ref

```
source("Keyness3D.r") # load the function
BROWN.df <- readRDS("input/BROWN.df.RDS") # load the corpus
tar <- droplevels(BROWN.df[substr(BROWN.df$PART, 1, 1)=="j",])
ref <- droplevels(BROWN.df[substr(BROWN.df$PART, 1, 1)!="j",])
results <- Keyness3D(tar, ref)
```

A user could then focus on the word types that are key for *T* because of their association to *T*. Since we are measuring association in a way that is not already confounded with frequency, thousands of words have a perfect association score of 1 (because they only occur in *T*). To simulate what happens if one really only looks at association, I am showing a random sample of 50 word types with that score in (6).

- (6) brucellosis, biopsy, respondent's, height-to-diameter, optics, zero-magnitude, unpaired, gyro-stabilized, ebb, classifying, synergistic, nonequivalent, celso, butchered, iodinate, volts, jurisprudentially, exogamy, bereavements, argon, 2.405, rumscheidt, electrolysis, epitomize, nakamura, poland's, agriculture's, haupts', dubin, proteolytic, categorizing, nonspecifically, misnamed, oxygens, plastering, echelons, 3,450, **zq, no-valued, cardiomegaly, geatish, glycerolized, interference-like, disentangle, solvents, discolors, torsion, scalar, tangent

Clearly, if one measures association such that it is not confounded by frequency (or dispersion) and then focuses on it alone for keyness, the results are 'correct' (each of these words *is* perfectly associated with *T*) but also not **useless** because most of these words are too rare and, thus, specialized rather than generally useful for learned discourse — more information than just pure association needs to be included. change to: u

If we turn to word types that are key for *T* because of their dispersion in *T* relative to *R*, we do not need a random sample because these word types all differ in their values; see (7). Interestingly, none of these word types have a perfect association keyness score but they show that distinguishing between association and dispersion is worth it. Not only is the correlation between the association and dispersion components of keyness fairly low (Spearman's $\rho = 0.227$), but most of these dispersionally key words are intuitively indeed much more generally useful for learned discourse than the associationally key words in (6).

- (7) results, such, may, these, 1, 2, relatively, various, possible, similar, method, amount, conditions, however, distribution, assumed, basis, due, types, essentially, therefore, appears, af, whereas, differences, are, methods, per, has, cases, thus, considerable, described, which, extent, used, ratio, addition, defined, related, values, permit, isolated, cannot, necessary, latter, 3, experimental, same, certain

Finally, here are words with high amalgamation₁ scores uniting all dimensions (see (8)):

- (8) results, af, 1, distribution, 2, such, relatively, these, various, may, conditions, method, assumed, differences, similar, experimental, essentially, types, whereas, defined, possible, appears, values, amount, methods, isolated, however, described, measurements, basis, therefore, analysis, cases, systems, calculated, data, due, thus, occurring, parameters, q, related, sample, follows, thermal, variables, detected, 3, extent, proportional

A few of these have a perfect association with *T* (*proportional*, *q*, and *parameters*), one has a perfect dispersion score (*results*), and again I think these words, which score high on 'unified keyness', are very useful because of their general utility for learned discourse. Nearly none of these words are specific to a certain (domain of) science and these top 50 seem much better than those of the most widely used measure, G^2 , whose top 50 words in (9) certainly include many useful ones, but also suffer from several drawbacks: (i) they also include quite a few function words that most users would probably not be interested in; (ii) they include several very domain-specific rather than generally useful words; (iii) since G^2 does not consider dispersion, it returns *staining*, ***zg*, or *bronchial* as keywords for






learned discourse, words that occur 37, 32, and 29 times in the whole corpus, but with all occurrences in each just a single file.



- (9) af, of, is, anode, t, 1, data, index, the, 2, surface, cells, system, stress, function, by, q, dictionary, rate, reaction, temperature, in, platform, sections, information, analysis, results, values, staining, which, binomial, elections, cell, are, sample, be, onset, c, shear, systems, number, these, **zg, emission, wage, curve, bronchial, used, questionnaire, operator

4. Concluding remarks

The advantages of the present approach are that (i) it uses more dimensions of information to determine keyness than any existing approach and (ii) it measures them in ways that try to avoid conflation of information. That in turn means researchers can concentrate on what different dimensions of information contribute to keyness, avoid dispersion artefacts, and decide to weigh dimensions as their specific application may desire, and the now easy availability of Keyness3d will hopefully stimulate further exploration of a more multi-faceted view of keyness.

References

-  Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1). 77–104.
-  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.
-  Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: John Benjamins.
- Hofland, Knut & Stig Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.
- Leech, Geoffrey & Roger Fallon. 1992. Computer corpora – What do they tell us about culture? *ICAME Journal* 16. 29–50.
-  Manly, Bryan F.J. & Jorge A. Navarro Alberto. 2016. *Multivariate statistical methods: A primer*. 4th. ed. Boca Raton: CRC Press.
-  Paquot, Magali & Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Andreas 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: Keynes statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 133–167.
-  Rayson, Paul & Amanda Potts. 2020. Analysing keyword lists. In Magali Paquot & Stefan Th. Gries (eds.), *Practical handbook of corpus linguistics*, 119–139. Berlin: Springer.