ARTICLE

Cultural Keywords in Varieties Research

Stefan Th. Gries

Department of Linguistics, University of California,
Santa Barbara, United States

ABSTRACT: One of the four most central corpus-linguistic methods is keywords/keyness analysis, which is generally the identification and interpretation of word types of a target corpus (*T*) that, when compared to their occurrence in a reference corpus (*R*), are key/characteristic for *T*. In terms of the target of the comparison/analysis, *T* and *R* usually differ regarding their topic, registers, or genres, but ever since Hofland and Johansson (1982) and Leech and Fallon (1992), studies have also used keywords analysis to compare corpora representative for different cultures—there, American English versus British English—and Mukherjee and Bernaisch (2015) or Collins (2021) apply similar techniques to data that include outer-circle varieties of English. In terms of statistical method, many different methods have been suggested, but most studies used log-likelihood ratios comparing the token frequencies of each word type in *T* and *R*. Recent work, however, has seen two major proposals to put keyness analysis onto a more solid statistical footing by incorporating dispersion—with Egbert and Biber (2019) proposing to use dispersion rather than frequency and Gries (2021) proposing to use frequency, association, *and* dispersion—but these methodological suggestions have not yet been applied to cultural keywords. In this article, I will (a) exemplify the use of Gries's methods to the study of three outer-circle varieties of English to identify keywords in a bottom-up fashion and (b) use the results from that first application to advance two suggestions how to extend keyness analyses to better understand the keywords from the first step: *key collocates*, which involves applying keyness to contexts of keywords; and *deep key collocates*, which involves distributional semantics methods like word2vec, GloVe, BERT, etc. to keywords. I will use Mukherjee and Bernaisch's (2015) keyness analysis as a launchpad to identify keywords from comparisons of Indian, Pakistani, and Sri Lankan Englishes (IndE, PakE, and SriE, respectively) and zoom in on the variety-specific differences of the keyword of *terror*. The results not only indicate what terms are key for which of the three varieties; they also allow for a previously nearly impossible degree of granularity in how keywords use differs across varieties and possibly cultures. For example, in the PakE data, newspaper coverage of *terror* is mostly discussed with regard its financial aspects and implications and matters of communication, whereas in IndE and SriE, *terror* is much more approached from a military and a religious perspective, respectively.[1]

KEYWORDS: association, dispersion, frequency, keyness, varieties of English

JRDS 1:1 (June 2025)
© University of Toronto Press, 2025 | doi: 10.3138/jrds-2024-0001

UNIVERSITY OF TORONTO PRESS

Introduction

Keyness as a One-Dimensional Measure

One of the four most central corpus-linguistic methods is keywords/keyness analysis,[2] which is the identification and interpretation of word types of a target corpus (*T*) that, compared to their occurrences in a reference corpus (*R*), are key/characteristic for *T*. I distinguish two scenarios:

- **Scenario 1** is the by far more widespread one where researchers are interested in determining keywords from corpora in a completely **bottom-up** fashion by doing some kind of keyness calculation for every word type attested at least once in *T* and/or *R* (or, alternatively, for every word type attested at least *x* times in *T*);
- **Scenario 2**, on the other hand, is a much less frequent one where researchers do keyness calculations for a set of *a priori*/**top-down** selected word types they are interested in to determine, for example, how each type from that set differentiates between *T* and *R*.

What are keywords often used for? In the case of the by far more frequent scenario 1 kind of bottom-up application, *T* and *R* usually differ regarding their topic areas (e.g., to identify expressions relevant for topic areas in language teaching contexts) or registers/genres (e.g., to identify expressions typical of certain communicative contexts/situations). However, ever since Hofland and Johansson (1982), studies have also used keywords analysis to compare corpora representative for different cultures. One particularly influential study is Leech and Fallon (1992), who compare British and American written English based on the Lancaster-Oslo-Bergen (LOB) and Brown corpora and distinguish two or three different "stages" in their analysis:

- stage 1 (pp. 33–35): the **bottom-up identification of keywords** (by comparing frequency lists) to find word types that are (substantially and/or significantly) overrepresented in *T*;
- stage 2a (pp. 35–37): the **disambiguation of senses of candidate keywords** (by looking at concordances) to determine what semantic category a keyword belongs to (they discuss the examples of *stress* and *pressure* and their ambiguity between psychological and physical senses);
- stage 2b (pp. 37–38): the **dispersion of the candidate keywords** (by looking at concordances) to "check that the high frequency of an item was not due to any obvious skewing of its distribution in the corpus."

All keyness analyses require some way to implement stage 1, but stages 2a and 2b have been considered much less, which is in part what motivated recent proposals in keyness analysis as well as in this article. But then how is stage 1

	Target node *N* (e.g. *strong*)	**Other words**	**Sum**
Collocate *C* (e.g. *tea*)	a	b	$a+b$
Other collocates	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d=N$

Table 1: Input to Compute the Mutual Association of N and C in Collocational Applications

implemented, and how are keywords identified in corpus-linguistic applications? Most of the time, determining how much a word type is associated with a target corpus T (which may represent a dialect, variety, register, or topic) compared to a reference corpus R (which represents a "standard of comparison" against which T is explored) involved a frequency-based association measure (AM) from collocation research: the log-likelihood ratio G^2, which is typically computed from a table such as Table 1.

In keywords analyses, the same logic was applied to an only slightly different frequency table, which is exemplified in Table 2 (based on data from Brown, 2016).

That is, for Table 1, G^2 would be understood as indicating how much the co-occurrence frequency of C with N differed from its null hypothesis expectation (given N's relative frequency) and collocate types would be considered interesting if the observed co-occurrence frequency a was higher than expected a with a high G^2-score. Correspondingly, for Table 2, G^2 would be understood as indicating how much the frequency of occurrence of W in T differed from its null hypothesis expectation (given the sizes of T and R) and word types would be considered key for T if observed a was higher than expected a with a high G^2-score. In either case, G^2 is computed as in Equation 1:

$$G^2 = 2 \sum_{\text{from } a}^{\text{to } d} \text{observed} \times \ln \frac{\text{observed}}{\text{expected}}. \tag{1}$$

Often, such G^2-scores are signed, meaning they are multiplied by –1 if observed a is not greater than expected a, specifically if a word is key not for the target

	Target corpus *T* (e.g., Clinton speeches)	**Reference corpus *R* (e.g., Trump speeches)**	**Sum**
Word type of interest *W* (e.g. *college*)	a	b	$a+b$
Other words	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d=N$

Table 2: Input to Compute the Association of W to *T* or *R* in Keyness Applications

corpus *T* but for the reference corpus *R*, which can be represented as in Equation 2 or, maybe more easily understandably, 3 (for information on how to compute this easily in R, see Gries, 2024, Section 2.3.1):

$$G^2_{\text{signed}} = G^2 \times \left(2 \times 1_{\{\text{observed } a > \text{expected a}\}} - 1\right) \tag{2}$$

$$G^2_{\text{signed}} = \begin{cases} -G^2 & \text{if } a_{\text{observed}} < a_{\text{expected}} \\ G^2 & \text{otherwise} \end{cases} \tag{3}$$

Thus, in the vast majority of keyness analyses, keyness is a one-dimensional construct allowing us to sort word types by their position on a continuum from "most/highly key for *T*" via "not key for either corpus" to "highly/most key for *R*." However, other keyness measures have also been used, such as the odds ratio, the chi-squared test statistic, the relative frequency ratio (see Edmundson & Wyllys, 1961; Damerau, 1993), Yule's Difference Coefficient, or the *t*-test; see Baron et al. (2009), Paquot and Bestgen (2009), Rayson and Potts (2020), and Gries (2024, Section 2.4) for overviews and comparisons.[3]

While most studies have used G^2—no doubt facilitated by the fact that some software applications made the computations very easily available —and have produced insightful results, it has also been argued that such work could be too simplistic. This is because, first, while G^2 is touted as a measure that quantifies the association between a word type and *T*, it is in fact more highly correlated with frequency than with association (Gries, 2024, Sections 4.2–4.3), and second, many studies have not addressed Leech and Fallon's stages 2a (sense disambiguation) and 2b (dispersion). I am not aware of any proposals to address 2a in readily quantifiable and especially nonmanual ways, but there have been several interesting proposals regarding the inclusion of dispersion in a readily quantifiable, nonmanual way.

The first kind of suggestion involves the notion of **key keywords**, specifically "words that are key in a large proportion of texts in a corpus." That is, a separate keyness analysis is done for each part of a corpus and word types that are key for *T* in many texts are considered key keywords (e.g., Scott, 1997; Baker, 2004); a somewhat related proposal is to consider words key only if they meet a certain range threshold in *T* (e.g., Millar & Budgell, 2008). However, Baker (2004) already concludes the former approach is too conservative to be useful (not to mention cumbersome), and (range or other) thresholds are very hard to motivate *a priori*.

A second proposal to **take dispersion more seriously** for keywords is to not use aggregate frequencies for all of *T* and all of *R*, but to use file-/text-specific frequencies instead. Paquot and Bestgen (2009) propose an excellent definition of keyness that utilizes dispersion—"[k]eywords of a specific corpus are lexical items that are evenly distributed across its component parts (corpus sections or texts) and display a higher frequency and a wider range than in a reference corpus of some kind" (p. 64)—and recommend the *t*-score from a *t*-test comparing each word type's mean relative frequencies in the parts/files of *T* and *R*. This is progress because of how it implicitly makes the dispersion of word types part of keyness (via the number of mean relative frequencies that are >0) but still only returns one number: from a certain *t*-value contrast between *T* and *R*, it will be very difficult to see whether a word type is key due to association or dispersion, or both.

Another interesting suggestion was made by Egbert and Biber (2019), who propose not to *augment* frequency-based association to *T* with dispersion, but to *replace* it with dispersion-based association to *T*. It involves computing G^2 on tables such as Table 2, but the four cell frequencies *a*, *b*, *c*, and *d* are not frequencies of occurrence of *W* and not-*W* in *T* and *R*; instead, they are the numbers of texts/parts of the two corpora that *W* and not-*W* are attested in at least once (i.e., the ranges of *W* and not-*W* in *T* and *R*). In other words, Egbert and Biber are replacing frequencies by *range*. While this is an interesting proposal in how it "downgrades" the role of frequency (which Egbert and Biber found makes keywords lists worse; Egbert, p.c.) and "upgrades" the role of dispersion, I still find it a bit suboptimal for the following reasons:

- *range* as a dispersion measure is very coarse-grained because it considers only the number of texts in a corpus that a word type is attested in but neither the frequency of the word type in the corpus part nor the corpus part's size;
- that also means that the numeric value of "frequency of occurrence" is reduced to a binary value: *yes* (*W* is in a corpus part) versus *no* (*W* is not in a corpus part)—how often *W* occurs somewhere is not considered anymore;
- *range* as a dispersion measure is still so very highly correlated with frequency that, on the whole, its results are still more informed by frequency than by actual dispersion: Gries (2022) shows that, across six widely used corpora, *range* is correlated with frequency with R^2 values from GAMs between 0.9104 and 0.9619;
- finally, this approach also considers (and outputs) only one dimension of information for each word type—it's just dispersion, whereas in most other approaches it's mostly frequency.

It is against this background that Gries (2021, 2024) proposes a three-dimensional approach to keyness, to which we now turn.

Keyness as a Three-Dimensional Measure

A very recent proposal to improve keyness measures is outlined in Gries (2021, 2024). He attempts to implement Leech and Fallon's stages 1 and 2b and the spirit of Paquot and Bestgen's definition cited earlier in a way that preserves the multiple dimensions of information that are arguably relevant to keyness: word frequency, association to *T* (versus *R*), and dispersion. In his approach, each word type is not characterized by a single keyness value, but by a tuple of three values (one for each of the above dimensions), which *can* be amalgamated into a single keyness value, if absolutely necessary (and in that amalgamation researchers can specify weightings, e.g., to downgrade the effect of a dimension (such as frequency)). In other words, he does not *replace* frequency with range/dispersion; he *augments* frequency with dispersion.

Another central aspect of Gries's proposal is that he uses the **KL-divergence D_{KL}** for the association- and dispersion-based components of keyness. This is because (a) D_{KL} is an extremely easy-to-compute and directional measure of how one vector of proportions (a so-called **posterior distribution**) differs from another one (a so-called **prior distribution**), and because (b) Gries (2022) shows that, of all dispersion measures compared there, D_{KL} is the one least correlated with frequency and, thus, most likely to make an original contribution above and beyond frequency (as opposed to just replicating frequency in disguise as G^2 is prone to do).

From these two characteristics, a third one follows, namely that this multiple-keyness-components approach is currently the only one able to distinguish different "reasons" why words are key for *T*. For example, a word type might be key for *T* because

- it is very strongly associated with *T* over *R* even though it is not that frequent and not widely dispersed in *T*; these could be word types that are quite specialized even in the topic domain or register of *T* itself and that never occur in *R*. For instance, the word types *colorectal* or *Washington Law Review* might be not that frequent and specialized even in a target corpus of academic writing but completely unattested in a reference corpus of general American English;
- it is somewhat associated with *T* over *R* but extremely evenly dispersed in *T* and very clumpily or not at all dispersed in *R*; these could be word types that are in very widespread use in the topic domain or register of *T* and used only a bit in *R*. For instance, the word types *characteristics* or *limitations* might be frequently and widely used in a target corpus of academic writing and pop up only occasionally in a reference corpus of general American English.

That way, the present approach (a) meets the desideratum of being able to distinguish global and local keywords and (b) addresses Leech and Fallon's stage 2b by avoiding flagging a word as key even if its keyness is only due to a single file/text (see Leech & Fallon's, 1992 example of *torsional* versus *emotional* in LOB, or Baker's (2004, p. 35) example of *wuz* "was" in gay male erotic narratives).[4] In what follows, I briefly characterize how each of the three dimensions is computed in this particular study, which is building on the logic of Gries (2021, 2024).

Frequency

The overall frequency of a node word type *W* in both *T* and *R* combined is important because without it, it is more difficult to assess how much importance one should assign to a word's association-based or dispersion-based keyness. If 100% of *W* occur in *T* then that should be considered more important when *W*'s frequency is high than when it is low. Thus, for the first dimension, frequency, one can use the binary log of *W*'s combined frequency in both corpora (i.e., log_2 of each word's value of $a + b$ in Table 2). Once all word types' logged frequencies have been obtained, the vector of all those logged frequencies can be min-max transformed.[5]

Association-Based Keyness

The second dimension of information, association-based keyness, is computed from the same kind of table as traditional keyness analyses, specifically tables like Table 2, but since D_{KL} is directional, one can actually compute two D_{KL} values from any such table:

- the direction word-to-corpus, which quantifies how much each word type changes the proportional distribution of the two corpora—that is, typically how much a word type of interest changes the probability that one is looking at *T*:
- the direction corpus-to-word, which quantifies how much the corpus being *T* changes the proportional distribution of each word—that is, typically, how much *T* increases the probability of occurrence of a word type.

I am using the former here, which is conceptually related to the question of how much better you can predict which corpus/variety you're looking at when you sampled a certain word; also, the D_{KL} values for that direction of keyness is less correlated with frequency than the latter, which helps keeping the dimensions of keyness nicely conceptually separate. For this scenario,

- the posterior distribution is the proportional distribution of *W* across the two corpora—that is, it is the vector of $\frac{a}{a+b}$ and $\frac{b}{a+b}$;
- the prior distribution is the corpus part sizes as proportions—that is, it is the vector of $\frac{a+c}{a+b+c+d}$ and $\frac{b+d}{a+bc+d}$.

	IndE part of SAVE 2020	PakE & SriE parts of SAVE 2020	Sum
Delhi	3,692 (0.866)	571 (0.134)	4,263
Other words	3,066,598	6,140,907	9,207,505
Sum	3,070,290 (0.333)	6,141,478 (0.667)	9.211,768

Table 3: Input to Compute the Association of *Delhi* to IndE Over PakE/SriE

Consider as an example Table 3 with the absolute frequencies and row percentages (rounded to three decimals) for the word *Delhi*. With these, the D_{KL} for *Delhi* used here is computed like this:

$$D_{KL} = 0.866 \ \times \log_2 \frac{0.866}{0.333} + 0.134 \ \times \log_2 \frac{0.134}{0.667} \approx 0.8829775. \qquad (4)$$

As readers might recognize, this is conceptually similar to G^2, but where G^2 uses frequencies, D_{KL} uses proportions, which is why it is much less correlated with the actual frequencies (both the frequency of W in T (a) and the overall frequency of the word type of $a + b$) and why it is much more separately informative in its own right. These computations are then made for each word type of interest. As a measure of divergence of one distribution from another, D_{KL} is by definition ≥ 0 (just like G^2 or chi-squared) but to make the measure more comparable with the other values collected, it gets transformed as follows:

- the D_{KL} values are transformed to the interval [0, 1] with the odds-to-probabilities transformation widely used in regression modeling contexts ($D_{KL\,\mathrm{norm}}$=D_{KL} / 1+D_{KL})[6];
- these values are then signed such that word types key for T are left positive and word types key for R are made negative;
- these values are then stretched by dividing them by their maximum.

This makes sure that the word types key for T occupy the whole space from a bit above 0 (the word types most weakly associated with T) to exactly 1 (the word type most strongly associated with T); for more explanation and exemplification in R, see Gries, 2024, Sections 3.1 and 3.2.3).

Dispersion-Based Keyness

The final dimension of information, dispersion-based keyness, is computed in three steps. Step 1 is to compute for each word in each corpus its D_{KL} dispersion in the corpus. For each corpus,

- the posterior distribution is the proportional distribution of a word across all the parts of the corpus;
- the prior distribution is the corpus part sizes as proportions.

This will return for each word in each corpus a value between 0 (if the word is distributed across the corpus parts in perfect correspondence to the sizes of the corpus parts) and, theoretically at least, $+\infty$ (the maximal value for a corpus is observed if all instances of a word show up in the smallest corpus part).

Step 2 is again to transform these values into a value range that is more useful for the current analysis using the same normalization as above so that the resulting values fall into the interval [0, 1], but now also flipping its orientation such that small and large values indicate "clumpy distributions" and "even distributions," respectively. Words that are not attested in T or R get their dispersion value set to 0.

Step 3 is to compute for each word the difference $D_{KL\text{norm}}$ in T minus $D_{KL\text{norm}}$ in R and "stretch these values" again to make sure that the word type most key for T in terms of dispersion scores the maximum value of 1. As a result,

- positive values indicate dispersion-based keyness for T;
- negative values indicate dispersion-based keyness for R;
- values around 0 indicate no dispersion-based keyness for either corpus; and
- the more the value differs from 0, the stronger the keyness preference.

Cultural Keywords in World Englishes

The concept of "cultural keyword" was first developed in Williams (1976) and has been of interest to linguists for quite a while. In a corpus-linguistic context, Stubbs (2002, p. 145) defines them as words "whose meanings give insight into the culture of the speakers" and Mukherjee (2009, p. 69f.) considers them to be "high-frequency content words included in a large and representative corpus of a language or a language variety" (cited in Mukherjee & Bernaisch, 2015, p. 417); Peeters (2020) prefers the term *culturally salient words*. Much exploration of such words has proceeded in not particularly corpus-based ways. For instance, in Wierzbicka's and Goddard's Natural Semantic Metalanguage (NSM) framework, cultural keywords are seen as words that are "particularly culture-rich and translation-resistant words that occupy focal points in cultural ways of thinking, acting, feeling, and speaking" (Wierzbicka 1997, cited in Goddard, 2018, p. 159). This kind of work is often scenario 2, or top-down and starting from a word form, and while it can be corpus-based (e.g., Cramer, 2015 on *Ordnung*), it often is not, and corpora might be little more than a repository from which supposedly illustrative examples are picked for qualitative interpretation.

In corpus-linguistic contexts, however, cultural keywords are defined and explored differently: much of the work there is scenario 1, or bottom-up, and quantitatively informed in at least some way. As mentioned earlier, this research area can be dated back to Hofland and Johansson (1982) and, maybe even more influentially, to Leech and Fallon's (1992) analysis, which compares the Brown and LOB corpora to identify "varied social, political, and cultural aspects of the two most populous English-speaking countries" (p. 29), specifically the "[n]on-linguistic contrasts [..., i.e., not just spelling variation] which could not be easily explained as matters of linguistic code or variety, but where one had to postulate a difference of subject-matter" (p. 34). Ultimately, they then interpreted the results and arrived at "one wild generalization" (p. 44) that, while "much of a caricature," is also "not an unconvincing portrayal" (p. 45) of the two cultures being compared.

Leech and Fallon (1992) prompted much at least conceptually similar work on both inner- and outer-circle varieties. Well-known studies include Oakes and Farrow (2007), Mukherjee and Bernaisch (2015), and Collins (2021). However, Oakes and Farrow's work is not particularly informative when it comes to cultural conclusions because, according to Collins (2021, p. 31), it returns "essentially predictable findings – of a type that would be dismissed in cultural keyword studies as 'uninteresting.'" Collins (2021) itself is more culturally focused and somewhat of a replication of Leech and Fallon (1992), but from a scenario 2/top-down perspective because he starts from "sets of culturally significant lexical items representing" (p. 7) Leech & Fallon's 15 semantic domains, and his data involve eight varieties in the web-based GloWbE corpus (AmE, BrE, AusE, NZE, SingE, HKE, IndE, and KenyE). Unfortunately, while Collins (2021) is based on much more data, it is actually much less methodologically sophisticated than Leech and Fallon's classic: unlike them, Collins does not use any statistical measures or tests (chi-squared or otherwise) to separate or rank-order keywords (relying instead on different shades of blue in GloWbE's web interface!) and frequencies per million words, plus he also does not seem to consider the dispersion of the words across the varieties.

Mukherjee and Bernaisch (2015) fares much better in this regard, and their work is interesting because it focuses on a culturally more interesting cluster of varieties and because it combines scenario 1/bottom-up and scenario 2/top-down kinds of explorations. Theirs is a comparative study of cultural keywords of IndE, PakE, and SriE based on the South Asian Varieties of English (SAVE) corpus containing newspaper data from 2003–2006. In the bottom-up part of their study, they identify keywords for the three varieties compared to the BNC news section and then focus on a set of 90 shared nominal keywords and verbs following them in a context window of 5 words. In addition and in the top-down part of their study, they conduct a more detailed analysis of three specific headwords (*government*, *terror*, and *religion*) involving both scrutiny of specific concordance lines and more bigger-picture results involving sentiment analysis and semantic prosody.

The Present Paper

To recap, this article pursues two goals. First and with regard to the bottom-up approach of scenario 1, I exemplify an application of Gries's (2021, 2024) proposals, which follow Paquot and Bestgen's (2009) definition of keyness and address Leech and Fallon's (1992) stages 1 (identification of keywords) and 2b (dispersion) and I will do so by "replicating" Mukherjee and Bernaisch's (2015) study of IndE, PakE, and SriE, but with (a) a more recent corpus (the new SAVE 2020 corpus; see Bernaisch et al., 2021) and (b) the three-dimensional approach to keyness. Second and with regard to the top-down-up approach of scenario 2, I propose two ways in which keywords identified in a bottom-up kind of analysis can be studied in much more detail and in ways that help accommodate Leech and Fallon's (1992) stage 2a (distinguishing senses) and I will exemplify the two proposals in more detail using *terror* in the same three varieties. The last section presents conclusions.

Scenario 1/Bottom-Up: If You Want Keywords

I already mentioned Mukherjee and Bernaisch's (2015) cultural keywords analysis of journalese data for three English varieties (IndE, PakE, and SriE) using the first SAVE corpus. That study inspired the current one, and I follow their reasoning that "all three [varieties] originate in a largely uniform proto-South Asian variety of English" (p. 414). With IndE and SLE as the two oldest postcolonial Englishes in South Asia and with all three varieties' historical, geographical, and political interrelations,[7] a keywords follow-up with more recent data would appear a promising endeavor to see if and how this might be reflected in the use of English. I will therefore exemplify the use of Gries's proposals for scenario 1/bottom-up keyword identification using data representing the same varieties from the SAVE2020 corpus of journalese (Bernaisch, Heller, & Mukherjee, 2021). In SAVE2020, each variety is presented by approximately 1.5m words from two different newspapers:

- IndE: *The Statesman* and *The Times of India*;
- PakE: *Daily Times* and *Dawn*;
- SriE: *Daily Mirror* and *Daily News*.

To implement 3D-keyness, I used an R script to generate a three-column data frame with a separate row for each word type and

- a first column called VARIETY indicating the variety the word/lemma was used in: *IND* versus *PAK* versus *SRI*;

- a second column called WORD indicating the word or lemma (I used the latter);
- a third column called PART indicating the corpus part/file the word/lemma was used in.

I then looped over this data frame three times, each iteration of which used one variety as *T* and the other two as *R*, and performed the computations discussed earlier using a function called Keyness3D.batch, which returns all statistics discussed above in one spreadsheet; the function will be available to readers from my website and requires as its main input only

(1) a data frame x.tar with *T* in a data frame with columns 2 and 3 of the above data frame;
(2) a data frame x.ref with *R* in the same format;
(3) an argument normkld, which specifies how the D_{KL} values are normalized into $D_{KL\mathrm{norm}}$ (I used the above odds-to-probabilities approach);
(4) an argument dir4freq that specifies whether one computes the association-based keyness from the word to the corpus (row2col, which I did) or from the corpus to the word (col2row).

Let us first visualize the kind of results one can get from this approach, in particular if one represents association and dispersion on the (*x*- and *y*-)axes of a scatterplot and represents the third dimension of frequency (frequency) in the types' font size (here, the font sizes are all fairly similar for print publication purposes, but in one's own exploration, stronger downgrading of rarer word types or even filtering them out can be advantageous). One can then zoom into areas where "useful" words are located—"useful" as in "words with high keyness scores in at least one, but also more, dimensions"—for instance, into the high-association part of the plot or into the high/even-dispersion part. Using IndE as *T*, Figure 1 is an example of the high-association plot, whereas Figure 2 is an example of the even-dispersion plot; the dotted rectangle in Figure 2 shows the part of the data represented in Figure 1.

Many words that are most key in terms of both association and dispersion clearly reveal that *T* is IndE, which means the method is successful, even if they are maybe not that culturally revealing/insightful:

- names of people and institutions: *Lok* and *Sabha* (lower house of Parliament), *BJP* (a political party), *Singh*, *Gandhi*, *Narendra*, …;
- locations: *Uttar* and *Pradesh*,[8] *Delhi*, *Maharashtra*, *Kolkata*, *Bihar*, *Karnataka*, *Mumbai*, …;
- unit terms: *lakh* or *crore*.

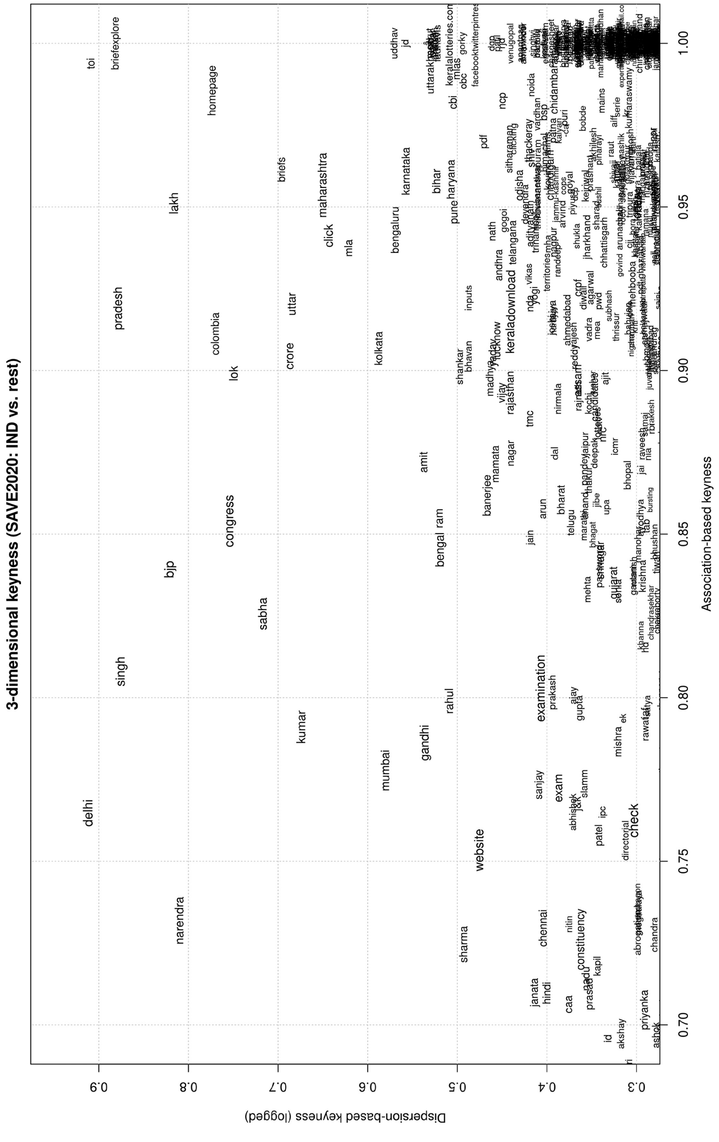

Figure 1: IndE: Association-based keyness (hi), dispersion-based keyness, and frequency.

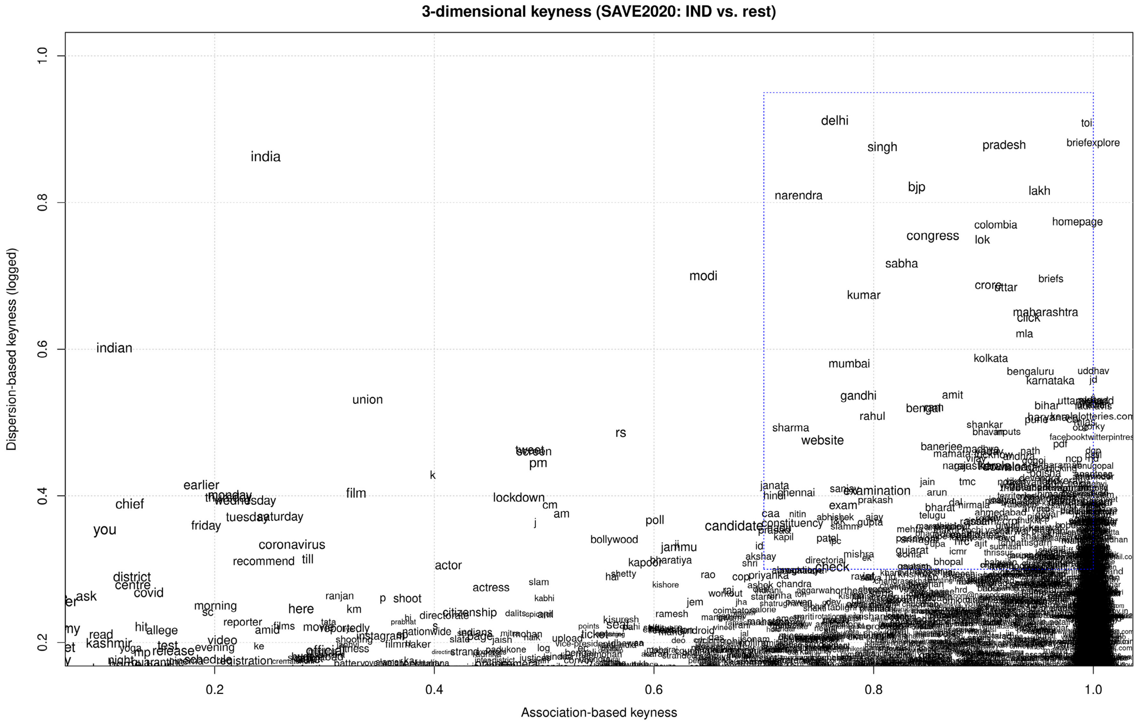

Figure 2: IndE: Association-based keyness, dispersion-based keyness (hi), and frequency.

Of course, it is generally more interesting to determine more general words highly characteristic for and hardly characteristic for a variety. The following are terms characteristic for IndE in a variety of semantic domains (heuristically defined, as in Leech and Fallon):

- political terms: *congress*, *candidate* and *candidates*, *constituency*, *territories*, *PM*, *poll* and *polling*, *union*, *seat …*[9];
- technological terms: *galaxy* (the Samsung product line), *aperture* (of cellphone cameras), *sensor*, *pixel*, *handset*, *processor*, *64gb*, *screen*, *qualcomm*, *android*, …;
- sex-related terms: *condom* (not attested in PakE and SriE at all), *gangrape*, and, to a lesser extent, *rape*, *sexy*, *Unnao* (a province in which a widely publicized rape took place)—interestingly, *prostitute* and especially *prostitution* are keywords “against” the IndE data—that is, word types more key of the combined reference corpus of PakE and SriE;
- crime- and weapons-related terms such as institutions such *CBI*, *NIA* and *NCRB*, police ranks like *DCP* and *ACP* (Deputy/Assistant Commissioner Police), *gunfight*, *rifles*, and, to a lesser extent, *rifle*, *Ponzi*, or *crimes—violence*, on the other hand, is not key for IndE;
- weather-related terms: *cyclone*, *cyclonic*, and *Fani* (a particularly severe cyclone in 2019), or *landfall*, but also an institution term like *IMD* (India Meteorological Department);
- sports-/nutrition-related terms: *workout* and *gym*, *f1*, *chess*, *diet*, and some of its derivatives, *carb* and *calory*, and, to lesser degrees, *tennis* and *badminton*.

By contrast, words that are particularly *un*characteristic of the IndE data relative to the other two varieties—I am excluding obvious geographical terms such as *Sri Lanka*, *Colombo*, *Karachi*, and the like—include the following:

- geographical/political terms: *provincial* and *province*, *federal*, *country*, *senate*, *candidacy* (strangely enough, given that the other words from that word family are very key for IndE), *island*, and others;
- economic/trade-related terms: *economic* and many hyphenated versions including *macro*- and *socio*- (with and without hyphens), *geo*- or *politico-economic*, *hydropower*, *devaluation*, *depreciation*, *profitability*, *chairman*, *patronage*, *macroeconomic*, *business*, *competitiveness*, *financing*, and many more;
- religious terms: *clergy*, *interfaith*, *blasphemy*, *seminary*, *buddhism*, *archbishop*, *buddhist*.

Word	Frequency in *T*	Association-based keyness	Dispersion-based keyness	Amalgam
government	7,053	−0.001	−0.28	−0.281
religion	213	−0.101	−0.316	−0.366
terror	737	0.17	0.027	0.115

Table 4: Keyness Values for IndE for *Government*, *Religion*, and *Terror*

While the results can only be suggestive at this level of analysis, their post hoc sense clearly validates the 3D approach, and some clear patterns emerge that a more in-depth analysis could explore. In addition, the present results can be nicely compared to Mukherjee and Bernaisch's words of interest, *government*, *terror*, and *religion*, as shown in Table 4. Of those three, only *terror* is key for IndE (quite key in terms of association and just a bit in terms of dispersion). Interestingly, while many political terms are strongly attracted to IndE, *government* is not at all.[10]

How do these three words fare in PakE and SriE? The PakE data and the SriE results are shown in Table 5 and Table 6, respectively. The word *government* is quite key for the PakE data (especially in terms of dispersion, its use is much more widespread than in the other varieties), and *religion* is a bit key for PakE; *terror*, however, is not and especially so in terms of association.

In SriE, by contrast, *religion* is a keyword (both in terms of association and dispersion), but the other two are not.

Looking at "politically or ideologically loaded" words might also be interesting, if only to validate the approach. Take, for example, *Kashmir*, a word that one would expect to be key to IndE and PakE but much less relevant to SriE,

Word	Frequency in *T*	Association-based keyness	Dispersion-based keyness	Amalgam
government	9,219	0.035	0.233	0.258
religion	381	0.002	−0.008	−0.007
terror	310	−0.069	−0.032	−0.068

Table 5: Keyness Values for PakE for *Government*, *Religion*, and *Terror*

Word	Frequency in *T*	Association-based keyness	Dispersion-based keyness	Amalgam
government	5,790	−0.026	−0.11	−0.128
religion	487	0.068	0.083	0.117
terror	347	−0.037	0.001	−0.018

Table 6: Keyness Values for SriE for *Government*, *Religion*, and *Terror*

Variety	Frequency in *T*	Association-based keyness	Dispersion-based keyness	Amalgam
IndE	2585	0.103	0.2	0.264
PakE	2657	0.119	0.127	0.201
SriE	121	−0.505	−0.304	−0.618

Table 7: Keyness Values for PakE for *Government*, *Terror*, and *Religion*

given political realities. The present approach confirms this nicely, especially in the association dimension, where IndE and Pak score robustly positive values while SriE scores extremely low, as is shown in Table 7.

In sum, the proposed approach can indeed be used to determine which words are key for which variety and, as opposed to nearly all previous work, also why (association and/or dispersion). At the same time, it implements very comprehensively multiple authors' pleas to take dispersion more seriously: one aspect of Leech and Fallon's stage 2 was checking the dispersion of key words, but they did so only with a simplistic clustering threshold (requiring three instances per document) based on actually checking concordance lines. By contrast, the proposed method here incorporates frequency, association, and a very comprehensive measure of dispersion (differences between *T* and *R*), but also allows to separate them; it is, therefore the most complete way to operationalize Paquot and Bestgen's definition, and it protects very robustly against outliers (e.g., words with high frequencies in few articles/files).

Scenario 2/Top-Down: If You Have Keywords

While most keywords analyses are scenario 1/bottom-up—as in the above, *all* word types are considered for their potential keyness—one might also be interested in something less bottom-up. For example, it's possible that one did an exploratory keywords analysis that yields keywords one wants to follow up on in more detail. This kind of application is conceptually a bit similar to what Scott (1997, pp. 238–240) calls "finding associates," and versions of it are exemplified both in Mukherjee and Bernaisch (2015), who, after their first keywords analysis, explore *government*, *terror*, and *religion*, and in Cvrček and Fidler (2022), who propose to use association rules to "recontextualize" keywords for the topic of migration in Czech internet media. An even more top-down approach could involve, for instance, varieties researchers hypothesizing findings regarding specific word types of interest in an *a priori* fashion. In this section, I propose two approaches for such more targeted follow-ups. For lack of space, I do so only briefly and show how, once a word (e.g., *terror*) has been decided on as a word for further exploration, additional information can be gleaned from the corpus data that is not only surprisingly precise but also useful for Leech and Fallon's stage 2a, disambiguation (recall the mention of *galaxy* and *aperture* earlier).

Key Collocates: Three-dimensional Keywords on Concordances

The first approach is actually a straightforward extension of the keyness approach outlined earlier and, thus, methodologically very convenient. Rather than applying the keyness algorithm to both *T* and *R* like we did for the bottom-up approach in the previous scenario/section, one now applies it to the concordance contexts of the word of interest in *T* and *R* as suggested in Mahlberg and O'Donnell's (2008) study of the word *fresh* in text-initial sentences (*T*) versus non-text-initial sentences (*R*)[11] or in Hoey and O'Donnell (2015), but, crucially and differently from this earlier work, here this is done with the same three-dimensional approach as before, not just with relative frequencies. More specifically, one generates

- a concordance of the word/lemma in question in *T* using a user-defined context window (e.g., 10L to 10R);
- a concordance of the word/lemma in question in *R* with the same context window;
- performs a keyness analysis of the same type as above but now;
    - the new *T* becomes the context window of the word of interest in the concordance run on the previous *T*;
    - the new *R* becomes the context window of the word of interest in the concordance run on the previous *R*.

Thus, this approach returns **key collocates** of words of interest, and the procedure just suggested is what might be considered its prototypical application: The scenario 2 exploration of a target word *W*—possibly a word of *a priori*/top-down interest or a keyword resulting from a first scenario 1/bottom-up analysis—involves the same *T* and *R* that the keyness analysis was conducted on, only this time each corpus is reduced, so to speak, to the context window of the word(s) in question. To exemplify this very briefly on the basis of the present data, recall that Mukherjee and Bernaisch were interested in, for example, *terror* and that the above analysis returned that word type as key for IndE. However, one might of course also be interested in knowing more about how this word is talked about differently in the three varieties. Thus, one could conduct three analyses of the above type, each with a different variety as *T* being compared to the other two to see whether this results in any interesting differences. Here, I generated concordances of the word *terror* with a context window for 6L to 6R for each of the three varieties and, for the analysis of IndE versus PakE and SriE made all collocates of *terror* in the context window in IndE the target corpus and all collocates of *terror* in the context window in PakE and SriE the reference corpus, both of which were given to the same Keyness3D.batch function as arguments.

Space does not permit a detailed analysis here, but there were interesting initial patterns. For example, the key collocates of the word *terror* in PakE are extremely strongly associated with financial aspects of terrorism: the top 20 include *financing* (rank 1), *laundering* (2), *money* (3), *launder* (4), *anti-money* (8), and *finance* (13) / *financial* (31) whereas most of these words are among the most strongly "rejected" in both IndE and SriE. In IndE, by contrast, *terror*'s key collocates mostly revolve around concrete terror *attack(s)* in certain regions or locations (e.g., *Pulwama* [the location of a bombing attack of a vehicle convoy transporting Indian security personnel in February 2019] or *Jammu*), their evaluative descriptions (e.g., *deadliest*), police actions and victims (*CRPF* and *jawan*), and the (potentially retaliatory) *strike* of Indian forces against a *camp* in *Balakot*, *Pakistan* (also in February 2019); in fact, the word *Pakistan* is the ninth most key collocate in all the IndE data around *terror* whereas *India* is not in the top 450 key collocates of *terror* in the PakE data! In a very surprising addition to that fact, *Balakot* is in fact not discussed in the PakE data at all: that word is in fact in the *most rejected* single percentile of all the PakE data, which is surprising given that it is a city in Pakistan and would arguably be newsworthy for just the same reasons that it is in the IndE data. The SriE data, finally, seem to be mostly affected by news coverage of the 2019 Sri Lanka Easter bombings; the approach is precise enough to return *April* and *21* as key collocates 5 and 7(!) and *massacre*, as is suggested by a group of top key collocates including *Easter*, *Sunday*, *Sri Lanka*, *tourism*, *hotel*, *attack*, with *ISIS*, *LTTE*, *catholic*, and *church* as additional compatible collocates.

In sum, not only does this key collocates extension make for a both conceptually and methodologically very straightforward exploration keywords, but it is also an extension that is appealingly compatible with assumptions about keywords in non-corpus-based and largely qualitative NSM work. This is because this key collocates approach is essentially one way of operationalizing Wierzbicka's assumption that keywords often participate in clusters (quoted in Goddard, 2018), which would of course show in the context window of a node word. Note also that the above constitutes a first nice operationalization of the other important methodological aspect of Leech and Fallon's stage 2, namely stage 2a of determining the senses in which key words were used from concordance displays. The approach is so precise that it returns even the exact dates at which events talked about happened (the April 21 attack in Sri Lanka and the February 14 attack in Pulwama). However, this can be done perhaps even better, and the following section develops a more demanding, but also much more powerful, approach.

Deep Key Collocates With Word2vec

The other approach for post hoc exploration can be a little more demanding, both in terms of required corpus sizes and the technicality of the approach, yet the technicality has immense advantages for the results. Over the last few years, distributional semantics has grown exponentially in especially NLP, and a flurry of new approaches—word2vec, GloVe, fastText, BERT, and many more—have been taking NLP and neighboring fields by storm. The quantitative sophistication of many of these methods notwithstanding, with enormous simplification they all amount to what might informally be called collocations on steroids. The general logic usually involves processing the co-occurrences of tokens with other tokens in a context window and reducing the high dimensionality of such data to a much smaller one. The usual result is a model of a corpus expressed in a matrix with as many rows as the corpus has word types and a much smaller number of columns (often 300), and this model allows us to compute how related two word types are to each other in terms of co-occurrence (usually expressed as the cosine of the angle between the two word types' 300-dimensional vectors). Often, the word types most related to a word type of interest *W* can be interpreted semantically or for disambiguation. For instance, the top 50 word types with the highest cosines compared to *vehicle* in the IndE data include

- hyponyms (e.g., *truck*, *car*, *bus*, *tractor*, *motorcycles*, *bike*);
- meronyms (e.g., *motor*, *plate*, and *generator*);
- frame-related participants (e.g., *owners*, *bikers*, *driver*/*drivers*, and *motorists*);
- frame-related action verbs (i.e., things that people do with vehicles, e.g., *speeding*, *rammed*, *laden*, *towed*, *parked*).

In other cases (e.g., with basic-level terms), one also usually finds synonyms, hyperonyms, and co-hyponyms. Similarly, the four word types most related to *Gandhi* in the IndE data are *Rahul*, *Priyanka*, *Vadra*, and *Sonia*. It is this method that I used to clarify/disambiguate the uses of *galaxy*, *aperture*, and others earlier: the top 50 words with the highest cosines compared to *galaxy* include different spellings of *Samsung*, *smartphone*, *handset*, many smartphone model names, *Exynos*, and phone memory and phone camera specifications, which are all indicating very clearly in which sense *galaxy* is used in the data. With this method, disambiguating the most likely majority sense of a word—in other words, addressing Leech and Fallon's stage 2a—can literally take just a few seconds or, if done for thousands of words at the same time, a few minutes.

Given the above, I suggest that (key)words of interest can be compared across varieties by comparing their most-related words as determined by such computational models, and I will call such most-related words their **deep key collocates**. I will exemplify the approach by returning to *terror* as follows. I first trained word2vec models (with the R package wordVectors; see Schmid & Li, 2022) separately on the IndE, PakE, and SriE components of SAVE2020 and with identical parameter settings: a minimum required word frequency of 3, a context window of 4, and 35 iterations for a skip-gram model with 300 dimensions. Then, I retrieved the top 300 word types most similar to *terror* and exported them for a brief qualitative analysis.

The proposed method yields some easy-to-follow-up and interesting results. First, and most obviously, one can determine which deep key collocates of *terror* are shared by the three varieties, the "shared conceptual core" of *terror* across IndE, PakE, and SriE so to speak, which is here represented by these word types: *terrorist*, *attack*, *Pulwama*, *Jaish*, *dastardly*, *mastermind*, *terrorism*, *islamist*, *ISIS*, *TISIS*,[12] *violence*, *militant*, *bom*,[13], *orchestrate*, *heinous*, *fundamentalist*, *Qaeda* (as part of *Al Qaeda*), *bombers*, *extremist*, and *condemns*.[14]

Second, and to determine how the varieties differ, I looked at the word types that, according to the embeddings, were only associated with *terror* in one variety. I grouped these word types into several semantic categories, which included the following (with some associated word type examples):

- military: *bombs*, *bombings*, *paramilitary*, *raid*, *combat*, *bogey*, *gunfire*, *…*;
- crime: *conspiracy*, *extort*, *illicit*, *mugging*, *piracy*, *abuser*, *murderous*, *…*;
- religion: *churches*, *Muslims*, *preacher*, *Salafi*, *synagogue*, *worshipper*, *Shariah*, *prayer*, *…*;
- communication: *bluff*, *threat*, *indoctrinate*, *intimidation*, *misinformation*, *propaganda*, *condemn*, *alleged*, *…*[15]

In a final step, I then determined which of the semantic categories were overrepresented in which variety when talking about *terror* and found that the use of *terror* in

- IndE involves an overrepresentation of military terms: *artillery*, *combating*, *commando*, *drone*, *enemy*, *grenade*, *occupier*, *squads*, and *submarine* are some examples, which suggests that the Indian English news coverage of *terror* might be concerned a lot with how terror is performed and what (military) reactions it prompts;
- PakE was unique in its overrepresentation of communication terms, especially terms that involve misleading communication such as *indoctrination*, *intimidation*, *misinformation*, *propaganda*, *provocation*, *ruse*, *spout*, and *stereotyping*, but also the words related to the financing of terrorism that the previous section already uncovered: *financing*, *laundering*, *AML* and *CFT* (for Anti-Money Laundering and Counter Financing of Terrorism, respectively), *launder*, and *money*;

- SriE involves an overrepresentation of religious terms, in particular general terms (e.g., *churches*, *cult*, *fanatic*, *Muslim*, *Shiite*, *Sunni*, *Taslim*, *religiosity*, *Salafi*, *Salafist*, and *extremists*) but also virtually the complete word family of *jihad*, which suggests that the Sri Lankan English news coverage considers (extreme) religiousness as relevant to terror (or is simply more open to discussing that assessment).

It was also interesting to see that *JeM* was extremely strongly related to *terror* in IndE (it was the 2nd most similar word to *terror*) and also quite closely related to *terror* in SriE (the 30th most related word), but in a first analysis of just the top 300 words most similar to *terror* in PakE, *JeM* did not show up (it finally occurred in position 518 only). Such findings are hard, if not impossible, to come by with just a simple scenario 1 / only bottom-up keyword analysis. To find more subtle patterns and utilize much less direct co-occurrence distributions, the kinds of follow-up discussed in this section—both the "simple" key collocates and, especially, the more powerful deep key collocates—seem ultimately indispensable.

Excursus: Onomasiological Variation

As a slight excursus following up on the previous section, the methods discussed here also offer a potentially interesting window into spelling variation, which may or may not be culturally interesting, but also onomasiological variation, which can be very illuminating. For instance, consider Table 8 for the distribution of several names for the group of fighters that in Western media is often referred to as *ISIS*.

Using chi-squared residuals or a correspondence analysis as in Figure 3, one can see a fairly clear connection between SriE and the *IS*… expressions, between IndE and *Jaish*, and between PakE and *Daesh*, and these are so strong that knowing which word form is used to refer to ISIS improves one's prediction of which variety one is looking at by nearly 40% (Goodman and Kruskal's lambda for Table 8 is 0.393). This is obviously relevant even just for the technical task of identifying the

	IndE data	PakE data	SriE data
ISIS	51	30	189
ISIL	2	5	12
IS	25	66	143
Daesh	2	20	1
Daish	0	0	0
Jaesh	0	0	0
Jaish	142	59	3

Table 8: Frequencies of Words Referring to ISIS

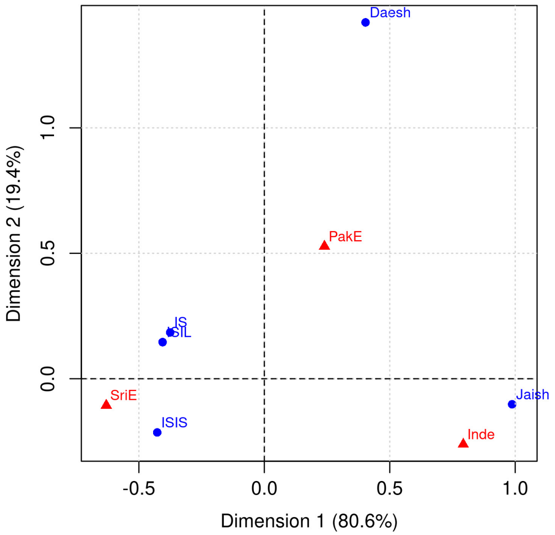

Figure 3: A correspondence analysis of ways to refer to ISIS across varieties.

concepts referred to by keywords and key collocates in corpora, but it is also relevant interpretationally given the undertones communicated by these apparently lexical choices. According to a 2015 article by the BBC, the Daesh/Daish spellings have "been used as a way of challenging the legitimacy of the group due to the negative connotations of the word. Daesh is essentially an Arabic acronym […]. Although it does not mean anything as a word in Arabic, it sounds unpleasant and the group's supporters object to its use" (https://www.bbc.com/news/world-middle-east-27994277).

Also, from that same article we learn that "[i]n the Arabic-speaking world, where the use of acronyms is otherwise uncommon, Daesh is used widely but with pejorative overtones. The label has gained currency despite or perhaps as a direct consequence of the irritation it causes the group, and is now used widely across the world by politicians and in the media."

Other 2015 coverage of this on NBC News (Garrity, 2015, last accessed January 22, 2024) seems to support this assessment. Thus, if this kind of onomasiological variation in the media was indeed representative of how ISIS is referred to, I might consider this at least prima facie evidence of localization or lexical indigenization (Mughees & Raza, 2015 might include this under their definition of *islamization*, the "inclusion of Islamic terms into [the] English lexicon," for which they give the example of the use of *mujahideen* in an otherwise English-only sentence); at present, I am unsure whether I would want to push this interpretation, but the fairly strong patterning of Table 8 is at least suggestive.

Concluding Remarks

While this article was quantitatively relatively dense, I hope it has offered ideas and food for thought. First, I exemplified how to utilize Gries's (2021, 2024) more comprehensive definition of keyness: not just one association measure based on frequencies of occurrence that is then also very highly correlated with frequency anyway (usually, G^2), but a combination of three different dimensions: frequency of occurrence, an association-based keyness component, and a dispersion-based keyness component. This was exemplified briefly both in a bottom-up way (applied to IndE) and in a more top-down way (by looking at specific words, e.g., *government*, *terror*, *religion*, and *Kashmir*); the results here were interpreted largely based on the numerical outputs.

Second, I suggested two ways in which top-down approaches can be combined with more qualitative work. Given a certain set of words, I proposed to get more interpretable results using either (a) *key collocates*, i.e. a second three-dimensional keyness analysis on keywords' collocates in the three varieties; or (b) *deep key collocates*, a distributional semantics kind of approach (here exemplified with word embeddings from word2vec; more advanced approaches are of course possible). I exemplified both by illustrating what the three varieties share and how they differ in their coverage of terror(ism), and both methods yielded useful and complementary results.

Given what I have seen in this dataset, I would go so far to recommend as a general practice for keyness analyses to try and identify for each keyword or key collocate the most similar words in an embedding model of the target corpus. Once one has such a model, this information is very easy to obtain and immensely helpful for generally getting a feel for how any specific word is used, for finding synonyms one might want to group for interpretation (see the *ISIS* example that follows), for identifying first or last names of people (recall the *Gandhi* example earlier), and for disambiguating a word (the two words closest to *BJP* in the IndE data were *congress* and *party*).[16]

Clearly, more needs to be done. We need more, and more than replications, to be able to see more conclusively how well my suggestions deal with different questions, corpora, corpus sizes, and so forth, and we need to see how parameter choices such as sizes of context windows, weightings of dimensions, and others affect (the interpretation of) the results. In addition, I would like to argue that keyness analyses can benefit from the prior filtering involved in lemmatization and POS-tagging used here but, maybe even more importantly, should perhaps always be performed on texts that have been tokenized in a way that allows for flexible multi-word unit recognition, especially for names and locations, but also other expressions.[17]

Finally, we also need to remain alert of the general question of how far we can push the interpretation of cultural keywords as obtained in scenario 1/bottom-up studies. It goes without saying that such analyses in general—simple G^2-based ones or the ones discussed here—need to find reliable ways to distinguish corpus-sampling results from cultural results: Are the many technology-related keywords in IndE indeed attributable to something in the Indian culture (relative to Pakistani and Sri Lankan culture) or are they also/more attributable to emphases of the sampled newspapers? Since keyness applications are entirely lexically based—as opposed to studies of, say, grammatical phenomena or alternations—they will require much care and background knowledge in the qualitative interpretation even after the most sophisticated statistical analyses. I would assume, however, that scenario 2 kinds of studies are a little less at risk simply because they are much more focused. Still, while quantitative methodological questions were at the heart of our discussion here, any discussion of empirical findings resulting from such methods obviously need to be carefully contextualized. For instance, if the corpora to which any of the methods proposed here are applied are largely news coverage, what role do editorial policies—on the level of an article, a news outlet, or even a society (like when governments constrain news coverage)—play for what findings are even possible and interpretable? And what effects do editorial policies on whatever level—for example, the choice how to refer to ISIS—have on a culture? Who even would be the target group of such policies (the domestic readers of news outlets or international politicians), and are the target readers, and thus "the culture," influenced as intended or do they have the opposite effect? And if the corpora used are largely fictional writing, how far can one generalize the findings to a culture as opposed to a stereotyped kind of culture that authors hoped would entertain and sell well? (A particularly relevant notion for social media, where something might be a keyword just for trolling effects.) Finally, we need to ask ourselves how strong and converging must evidence be to lead to robust conclusions regarding cultural differences: How quickly do we want to jump from observing an overrepresentation of *good on you* in AusE (with the meaning of what would be *good for you* in AmE) to a discussion of "the spirit of good-natured egalitarian companionship which Anna Wierzbicka has described as the essence of Australian culture" (Mair, 2007, p. 449)?

Thus, while I do think that methodological expertise is important and needs to be developed and broadened—which was the whole point of this article—it is important to bear in mind both the many general pitfalls that come with any kind of study of the language–culture interface and the particular bigger picture to which a specific study may aim to generalize with corpus-linguistic methods. Be that as it may, *if* good corpora are available, I hope to have demonstrated a few ways in which the study of cultural keywords can be moved to the proverbial next methodological level; hopefully this article will stimulate follow-up work improving these suggestions even more.

Notes

[1] I wish to thank two reviewers, who turned out to be Benedikt Heller and Jesse Egbert, for their thoughtful and valuable feedback, which helped me bring things into better perspective; all remaining inadequacies are of course solely my own.

[2] I will use both terms interchangeably, and from the term *corpus-linguistic* in the sentence it should be clear that I am not referring to "keywords" as used by, e.g., Garrett et al. (2005).

[3] Very recent developments include Clarke et al. (2021) and Cvrček & Fidler (2022) proposing, respectively, to use multiple correspondence analysis or association rules for keyword identification; especially the former strikes me as a very interesting approach. Model-based keyword identification, on the other hand, seems to not be used in corpus linguistics at all, let alone in the context of cultural keywords; see Monroe et al. (2008) for a Bayesian modeling approach to politically partisan keywords and Lim et al. (2024) for a cross-linguistic application with an NSM background.

[4] In his review of this manuscript, Egbert pointed out that the use of range (rather than something more sophisticated) and the use of range alone (rather than more dimensions) may be justifiable on the grounds of parsimony if its results were comparable to those of more complex methods—and I agree! If indeed the results were comparable, of course the more parsimonious approach is to be preferred. However, there are two considerations that have so far prevented me from following this to the conclusion that text dispersion with range is the final word. The first is this just discussed ability of the current three-dimensional approach to keywords to see distinctions in the data that an approach using just one single metric cannot see. The multidimensional keyness approach in Gries (2021) can distinguish "high keyness because of high association (rather than dispersion)" from "high keyness because of high dispersion (rather than association)" that, in that study, for example, lead to different kinds of words key for academic writing, namely very specialized, domain-specific keywords for academic writing versus domain-general keywords for academic writing. Thus, a higher-dimensionality definition of keywords does provide findings that a simpler one-dimensional approach does not.The second consideration is that concerned with the granularity, or resolution, of the dispersion metric used. I could not help but be a bit surprised at the range-based operationalization of dispersion in the text dispersion approach to keyness, because, not all that long ago, Egbert was part of a group of authors who argued in this very journal for a quite complex approach to dispersion (see Burch et al., 2017), one that was computationally orders-of-magnitude more demanding than other methods while yielding results that were 0.99 correlated with other methods (see Gries, 2020). This is of course not to say that demands for computational complexity may differ from application to application, but it does point to the fact that, not long ago, Egbert and colleagues were also open to more comprehensive ways to computing dispersion than range. Thus, it is at least not super obvious that dispersion is best measured with a statistic that disregards corpus part sizes and frequency of occurrence in a corpus part, and I hope this discussion will stimulate more exploration in this

area, e.g., as Jesse pointed out correctly, comparisons of keyword lists from multiple methods (see also Gries, 2022 and, for more general discussion, Gries, 2024).

5. The min-max transformation is a simple step that transforms a vector of numbers num into the interval [0, 1]: "/"(y <- num-min(num), max(y)). For example, the vector c(0.1, 0.5, 0.9) would become c(0, 0.5, 1) and the vector c(1, 3, 7) would become c(0, 0.333, 1).
6. Other transformations are available, but for most practical intents and purposes their differences are irrelevant.
7. See Mukherjee and Bernaisch (2015, p. 416f) for a brief sketch of India's, Pakistan's, and Sri Lanka's history and relations.
8. The preprocessing of the corpus made these two different word tokens, but this of course refers to the northern Indian state; this points to the relevance of tokenization, an issue I will take up briefly later in the article.
9. The reader might wonder how I determined that the uses of *seat* were used politically (as opposed to car seats, stadium seats, etc.) or how, in the next paragraph, *galaxy*/*aperture* etc. were classified as referring to Samsung phones/cellphone cameras rather than as astronomical or other terms. I did so using the methods developed in detail in this article.
10. This is not meant as a contradiction/falsification of Mukherjee and Bernaisch's results, which were conducted not only with a different keyness method but also on a different corpus from 15 years earlier.
11. I am grateful to Jesse Egbert for pointing out this study to me.
12. Post hoc exploration suggests these are mistokenized versions of *ISIS*.
13. Post hoc exploration suggests these are mistokenized versions of possibly hyphenated *bombing*.
14. I note that this approach considers the word *Pulwama* a deep key collocate of *terror* shared by the varieties whereas the more traditional approach in the previous section considered it an IndE key collocate. This is to be expected; even more traditional measures of keyness will not always agree, and the disagreement between the two methods here is actually only one of degree: *Pulwama* is the 4th highest deep key collocate of *terror* (i.e., the 4th most similar word to *terror* in terms of its IndE model-based cosine (0.511), but it is actually also the 62nd highest deep collocate of *terror* in PakE (0.786) and the 60th highest in SriE (0.711). In other words, the word2vec model overall agrees with the previous approach that *Pulwama* is much more similar to *terror* in IndE than in the other varieties.
15. Note again that the classification is heuristic: Categories overlap, can be of different parts-of-speech, and can operate on different levels of characterization. For instance, a word such as lid was characterized only as 'thing or concrete object' but drone, obviously also a concrete object, and bombings were classified as 'military'. The point was not to develop a new and general semantic taxonomy or ontology of 'things', it was to come to grips with how words profile different semantic domains.
16. Wulff & Gries (2024) use the same word2vec-based approach to identify dozens of authors who are key to one corpus-linguistic journal compared to two others: When given the first name of an author, the model nearly invariably returns the last names of all authors with that first name as the closest neighbors (e.g., for Manfred the model returns Krifka, Krug, and Stede).

17. See Ben Youssef (2024) for a multi-dimensional algorithm exemplifying exactly this: When applied to the Brown corpus, the unsupervised method developed there recognizes expressions such as Los Angeles, Hong Kong, dolce vita, Ku Klux Klan, El Dorado, Notre Dame, respiratory infections, Herald Tribune, and many names of people such as Dag Hammarskjold, Theodore Roosevelt, Benjamin Franklin, Santa Barbara, etc. This is not just useful especially for corpora that contain many non-English names of people or locations and other expressions that many readers (including the present author) may not always be familiar with, it is also useful to easily identify/disambiguate single-word units (when does Jaish refer to 'ISIS', when is it part of Jaish-e-Mohammed?) and it can help implement Halliday's (2003:408) second way of studying lexical variability (studying "not just new words but new word clusters"); I feel that this can raise the interpretability of many keyness analyses to a whole new level.

References

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

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and keyword statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, *20*(1), 41–67.

BBC. (2015). ISIS, ISIL, IS or Daesh? One group, many names. Retrieved January 22, 2024 from https://www.bbc.com/news/world-middle-east-27994277

Ben Youssef, C. (2024). mMerge: A corpus driven multiword expressions discovery algorithm. Unpublished PhD dissertation. University of California.

Bernaisch, T., Heller, B., & Mukherjee, J. (2021). Manual for the 2020:Update of the South Asian varieties of English (SAVE2020) corpus. Version 1.1. Justus Liebig University, Department of English.

Brown, D. (2016). Clinton-Trump corpus. Retrieved 2016 from http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus

Burch, B., Egbert, J., & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 189–216.

Clarke, I., McEnery, T., & Brookes, G. (2021). Multiple correspondence analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. *Register Studies*, *3*(1), 144–171.

Collins, P. (2021). Cultural keywords in World Englishes: A GloWbE-based study. *ICAME Journal*, *45*, 5–35.

Cramer, R.K. (2015). German *Ordnung*: A semantic an ethnopragmatic analysis of a core cultural value. *International Journal of Language and Culture*, *2*(2), 269–293.

Cvrček, V., & Fidler, M. (2022). No keyword is an island: In search of covert associations. *Corpora*, *17*(2), 259–290.

Damerau, F.J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, *29*(4), 433–447.

Edmundson, H.P., & Wyllys, W. (1961). Automatic abstracting and indexing: Survey and recommendations. *Communications of the ACM*, *4*, 226–234.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77–104.

Garrett, P., Williams, A., & Evans, B. (2005). Accessing social meanings: Values *of* keywords, values *in* keywords. *Acta Linguistica Hafniensia*, *37*(1), 37–54.

Garrity, P. (2015). Paris attacks: What does 'Daesh' mean and why does ISIS hate it? Retrieved January 22, 2024 from https://www.nbcnews.com/storyline/isis-terror/paris-attacks-what-does-daesh-mean-why-does-isis-hate-n463551

Goddard, C. (2018). Words as carriers of cultural meaning. In C. Goddard (Ed.), *Ten lectures on natural semantic meta language* (pp. 159–193). Brill.

Gries, S.T. (2020). Analyzing dispersion. In M. Paquot & S.T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 99–118). Springer.

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

Gries, S.T. (2022). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, *5*(2), 171–205.

Gries, S.T. (2024). *Frequency, dispersion, association, and keyness revising and tupleizing corpus-linguistic measures*. John Benjamins.

Halliday, M.A.K. (2003). Written language, standard language, global language. *World Englishes*, *22*(4), 405–418.

Hoey, M., & O'Donnell, M.B. (2015). Examining associations between lexis and textual position in hard news stories, or according to a study by… In N. Groom, M. Charles, & S. John (Eds.), *Corpora, grammar and discourse: In honour of Susan Hunston* (pp. 117–144). John Benjamins.

Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. Norwegian Computing Centre for the Humanities.

Leech, G., & Fallon, R. (1992). Computer corpora: What do they tell us about culture? *ICAME Journal*, *16*, 29–50.

Lim, Z.W., Stuart, H., De Deyne, S., Regier, T., Vylomova, E., Cohn, T., & Kemp, C. (2024). A computational approach to identifying cultural keywords across languages. *Cognitive Science*, *48*(1), e13402.

Mahlberg, M., & O'Donnell, M.B. (2008). A fresh view of the structure of hard news stories. In S. Neumann & E. Steiner (Eds.), *Online Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop*. Retrieved February 22, 2025 from https://publikationen.sulb.uni-saarland.de/handle/20.500.11880/23572

Mair, C. (2007). Varieties of English around the world: Collocational and cultural profiles. In P. Skandera (Ed.), *Phraseology and culture in English* (pp. 427–468). Mouton de Gruyter.

Millar, N., & Budgell, B.S. (2008). The language of public health – a corpus-based analysis. *Journal of Public Health*, *16*(5), 369–374.

Monroe, B.L., Colaresi, M.P., & Quinn, K.M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

Mughees, S.M., & Raza, A. (2015). The indigenization of English in Pakistan. Jun 27. Retrieved February 1, 2023 from https://www.slideshare.net/SheikhMuhammadMughee/presentation-49905236

Mukherjee, J. (2009). *Anglistische Korpuslinguistik: Eine Einführung*. Erich Schmidt.

Mukherjee, J., & Bernaisch, T. (2015). Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes. In P. Collins (Ed.), *Grammatical change in English world-wide* (pp. 411–435). John Benjamins.

Oakes, M.P., & Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, *22*(1), 85–99.

Paquot, M., & Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 247–269). Rodopi.

Peeters, B. (2020). Culture is everywhere! In B. Peeters, K. Mullan, & L. Sadow (Eds.), *Studies in ethnopragmatics, cultural semantics, and intercultural communication: Meaning and culture* (pp. 1–14). Springer.

Schmidt, B., & Li, J. (2022). wordVectors: Tools for creating and analyzing vector-space models of texts. R package version 2.0, http://github.com/bmschmidt/wordVectors.

Scott, M. (1997). PC analysis of key words – and key words. *System*, *25*(2), 233–245.

Stubbs, M. (2002). *Words and Phrases: Corpus studies of lexical semantics*. Blackwell.

Williams, R. (1976). *Keywords: A vocabulary of culture and society*. Croom Helm.

Wulff, S., & Gries, S.T. (2024). CLLT 'vs.' Corpora and IJCL: An only half serious keyness analysis. *Corpus Linguistics and Linguistic Theory*, *20*(3), 461–479.

About the Author

Stefan Th. Gries is a Professor in the Department of Linguistics at the University of California, Santa Barbara, and Chair of English Linguistics (Corpus Linguistics with a focus on quantitative methods, 25%) in the Department of English at the Justus-Liebig-Universität Giessen. He earned his MA and PhD degrees at the University of Hamburg and his Habilitation/Venia Legendi at the University of Marburg. Email: stgries@linguistics.ucsb.edu~~.~~