RiCL Research in Corpus Linguistics

A new approach to (key) keywords analysis: Using frequency, and now also dispersion

Stefan Th. Gries
University of California, Santa Barbara / United States
Justus Liebig University Giessen / Germany

Abstract – A widely-used method in corpus-linguistic approaches to discourse analysis, register/text type/genre analysis, and educational/curriculum questions is that of keywords analysis, a simple statistical method aiming to identify words that are key to, i.e. characteristic for, certain discourses, text types, or topic domains. The vast majority of keywords analyses relied on the same statistical measure that most collocation studies are using, the log-likelihood ratio, which is performed on frequencies of occurrence in two corpora under consideration. In a recent paper, Egbert and Biber (2019) advocated a different approach, one that involves computing log-likelihood ratios for word types based on the range of their distribution rather than their frequencies in the target and reference corpora under consideration. In this paper, I argue that their approach is a most welcome addition to keywords analysis but can still be profitably extended by utilizing both frequency and dispersion for keyness computations. I am presenting a new two-dimensional approach to keyness and exemplifying it on the basis of the *Clinton-Trump Corpus* and the *British National Corpus*.

Keywords – Keyness; dispersion; frequency; association; *Clinton-Trump Corpus*; *British National Corpus*

1. INTRODUCTION[1]

1.1. General introduction

According to a recent introduction to corpus linguistics, there are four main ways, or methods, that corpus linguists use to extract information relevant to their research out of corpora: frequency lists, dispersion (the degree to which, say, a word is distributed evenly in a corpus), co-occurrence information (the degree to which, say, a word and a construction 'like' to co-occur), and concordances (Gries 2016: 12). In the more detailed discussion of these methods, Gries also mentions one particular use of frequency lists, namely the method of keywords, which he exemplifies there as "the

[1] I am grateful to Magali Paquot for discussion and input (in particular for Section 3); the usual disclaimers apply.

identification of words that are (significantly) overrepresented in one (target) corpus as compared to another (typically larger and more balanced reference) corpus" (2016: 14); a conceptually similar definition is provided in Egbert and Biber (2019: 77), who state that "[k]eyword analysis [is used] to identify the words that are especially characteristic of the texts in a target discourse domain" (see also Scott 1997: 236).

Applications of keywords analyses typically involve educational ones centering on language teaching, but some keywords analysis applications involve the analysis of text types or genres (see Scott and Tribble 2006: Ch. 5) or combine the two foci (e.g., Tribble 2002). An example of a language-teaching oriented application would be an applied linguist wanting to compile a list of important specialized – key – English vocabulary from the semantic domain of, say, engineering, and might therefore decide to compare the frequencies of use of words in a corpus of engineering textbooks and research articles to the frequencies of use of words in a corpus of more general (academic) English to arrive at a list of, for instance, 500 words that are particularly characteristic of engineering English and, thus, likely to be useful for learners of English who will have to read and write engineering English as part of their education or profession. On the other hand, an example of a more text type/genre-focused application, apart from those mentioned above, would be Xiao and McEnery (2005), who explore to what degree keywords analysis can be a useful alternative to Biber's Multidimensional Analysis (e.g., Biber 1988).

The majority of studies do keyword analyses – both for educational or genre studies – in a way that is essentially a blend of two corpus-linguistic methods: frequency lists and co-occurrence/association statistics. Specifically, keyword analysis typically involves the following steps: first, one compiles a frequency list of a target corpus *t* (e.g., a corpus of engineering English) and another frequency list of a reference corpus *r* (e.g., some corpus of general academic English). Second, for every word type observed in *t* or *r*, one generates a 2×2 table that is related to the one used in collocation/collostruction statistics. Association measures for collocation statistics are computed based on a table that contains co-occurrence frequencies of one target word type with another word type, association measures for collostruction statistics are computed based on a table that contains co-occurrence frequencies of one target word with a certain construction, and the association measures for a keyword analysis are

computed based on a table that contains frequencies of one target word in *t* and in *r*, as shown in Table 1.

	Target corpus *t* (engineering)	**Reference corpus *r* (general academic)**	**Sum**
Target word w (e.g., reactor)	a	b	a+b
Other words	c	d	c+d
Sum	a+c	b+d	N

Table 1: Schematic table to compute a keyness statistic for one word type

In Table 1, *a* is the frequency of the word in *t*, *b* is the frequency of the word in *r*, *a*+*c* is the size of *t* in words, and *b*+*d* is the size of *r* in words, and then many analyses proceed by computing the log-likelihood ratio (*LLR*) for this table (following Dunning 1993). For that, one first computes the expected frequencies for each cell from *a* to *d* using the equation in (1) (there, demonstrated only for *a*) and then one computes the log-likelihood score using the equation in (2).

(1) $a_{expected} = \frac{(a+b)\times(a+c)}{N}$

(2) $LLR/G^2 = 2 \times \left(a \times \log\frac{a}{a_{expected}} + b \times \log\frac{b}{b_{expected}} + c \times \log\frac{c}{c_{expected}} + d \times \log\frac{d}{d_{expected}}\right)$

For sortability and interpretability, one can 'manually' set the *LLR-scores* to negative values if $a<a_{\text{expected}}$ so that high positive values mean 'the word is attracted to *t* (relative to *r*)' whereas high negative values mean 'the word is repelled by *t* (relative to *r*).'

While the above is, so to speak, the default kind of analysis, which has been applied in many different papers (see Egbert and Biber 2019: 78–79 for a good overview of publications), it has been recognized that this mode of calculation is probably not ideal. This is why, by now, several alternatives or potential improvements have been explored; these improvements essentially try to add, in different ways, dispersion information to the analysis. The probably best-known suggestion for this is identifying not just keywords, but key keywords, which are "words that are key in a large proportion of the texts in a corpus" (Egbert and Biber 2019: 92). In their words:

> [t]o find key keywords, a separate frequency-based keyword analysis is performed to compare each text in the target corpus to the entire reference corpus. Key keywords are those that show up as key in a large number of texts from the target corpus.

An alternative approach proposed by Baker (2004) is to essentially discard keywords that do not meet a pre-defined dispersion criterion. In Baker (2004) that dispersion criterion is based on the simplest of dispersion measures, range, i.e. the number/proportion of texts in *t* that contain the word in question; obviously, this approach requires that the analyst defines a threshold range value, a requirement that is hard to do completely objectively – that fact, however, does not invalidate the idea per se.

While the above kind of keywords analysis was mostly based on word frequencies alone, recent work in corpus linguistics has begun to realize the importance that dispersion plays for such and other analyses; the next section discusses two such papers and how they motivate the present study.

1.2. Egbert and Biber (2019)

1.2.1. Overview

The main goals of Egbert and Biber (2019) are to

> (1) establish the importance of text dispersion in keyword analysis, (2) introduce text dispersion keyness, and (3) compare this new measure to four keyness measures that have been used in previous research (2019: 99).

The measure they develop, text dispersion keyness, "entirely disregards word frequency and instead generates keyword lists based solely on word dispersion across texts" (p. 83); crucially, their measurement of dispersion essentially also boils down to the measure *range*, because it "compares word use between the target and reference corpus in terms of the total number of texts where a word occurs at least once" (2019: 84) and then uses the *LLR*-score from above. Since they do not provide a numerical example and do not define their iterator *i* (2019: 84), it is instructive to briefly discuss one here. Imagine:

(i) a target corpus *t* that consists of three parts and the word in question *w* occurs at least once in the first and the second corpus part, but not in the third;
(ii) a reference corpus *r* that consists of eight parts and *w* occurs in six of them.

This situation can be represented in familiar 2×2 format that is used everywhere else in corpus linguistics, which is shown in Table 2.

	Target corpus *t*	**Reference corpus *r***	**Sum**
Corpus parts with *w*	2	6	8
Corpus parts without *w*	1	2	3
Sum	3	8	11

Table 2: Table to compute a text dispersion keyness statistic for one word type *w*

We can then apply (1) to Table 2 and compute the expected frequencies for each of the cells, which for cell *a* returns the result in (3).

(3) $\frac{(2+6)\times(2+3)}{11} = a_{expected} = 2.\bar{1}8$

Once that is done for all four cells, we can apply (2) and compute the *LLR*-score for this table, which amounts to 0.0745, which one could set to -0.0745 because a_{observed} (2) is less than a_{expected} (2.18182).

The authors then apply four more traditional keyness measures – ones that involve only frequencies and ones that involve frequency and dispersion – as well as their new measure to the *Corpus of Online Registers* (CORE; Biber and Egbert 2018). They find that "text dispersion keyness […] outperformed the other four keyness methods" (2018: 100) and that "[s]omewhat surprisingly, the two corpus frequency measures that account for dispersion in the form of a minimum text range (CF_R10, CF_R30) performed quite poorly on all of the metrics" and that

> [t]his suggests that there are fundamental problems with the corpus frequency approach that cannot be remedied with simple dispersion criteria. These problems seem to stem from the fact that the statistical procedure accounts only for frequency. (Biber and Egbert 2018: 100)

These findings are interesting and encouraging and, as someone who has argued for the relevance of dispersion for quite some time, I find it gratifying to see how the authors make first steps towards improving keywords analysis by utilizing dispersion. That being said, I also think that the authors are not going far enough with this and in what follows I make a few observations regarding the authors' arguments and implementation.

1.2.2. Dispersion in Egbert and Biber (2019): how is it measured?

First, Egbert and Biber adopt a resolution of dispersion that is very coarse. This is because, as already mentioned above, their measure of dispersion for keyness is range, i.e. it does actually not take much information into consideration: neither the sizes of the corpus parts (i.e. the overall frequency of all word tokens in a corpus part) nor the frequencies with which words occur in those corpus parts play any role – all that counts for their approach is whether in a certain corpus part, regardless of its size (!), a word has a frequency >0. In other words, they are reducing two numbers that characterize the results for each corpus part (ideally, a text) – (i) the size of the corpus part and (ii) the number of times a word occurs in there – to a simple binary *yes*/*no* decision:

(i) if the word occurs in the corpus part (no matter how often and no matter how big the corpus part), their approach says *yes* and adds 1 to cell *a*;
(ii) otherwise, their approach says *no* and adds 1 to cell *b*.

This, of course, loses a lot of information and is the equivalent of, in statistical modeling for instance, taking a numeric predictor (such as frequency or length or givenness) and reducing it to two categories, something that is usually not recommended at all (see, e.g., Altman and Royston 2006; Cumberland *et al.* 2014). Consider Table 3 for two hypothetical distributions of a word *w* in a ten-part target corpus. In the first/upper scenario, *w* occurs six times in the 31,000-word corpus, two times each in the three largest corpus parts; in the second/lower scenario, *w* occurs six times in the same 31,000-word corpus, but four, one, and one time in three of the smallest corpus parts – Egbert and Biber's (2019) formula reduces both scenarios to the number three – *w*'s range – for cell *a* of Table 1 and, subsequently equations (1) and (2) and can therefore not distinguish between the two scenarios.

It is at least not obvious that this is ideal because, even just intuitively, it seems that *w* is more evenly dispersed in the first/upper scenario, because (i) the six occurrences are more evenly distributed (2-2-2 vs. 4-1-1) and they are attested in larger corpus parts rather than smaller ones (and in general one would expect words to show up (more) in larger corpus parts). However, the measure Egbert and Biber are implicitly relying on, 'range', does not capture that. A dispersion measure that is more informative than *range*, such as *DP* (short for 'Deviation of Proportions', see Gries 2008, 2010;

Lijffijt and Gries 2012),[2] immediately shows this, however: *DP* ranges from 0 (very even dispersion) to 1 very clumpy/uneven dispersion) and *DP* for the first/upper example and the second/lower example are 0.5161 and 0.8065 respectively.[3] Thus, it stands to reason that a more fine-grained operationalization of dispersion could be advantageous.

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
# *w*	0	0	0	0	0	0	0	2	2	2
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
# *w*	4	1	1	0	0	0	0	0	0	0
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000

Table 3: Two hypothetical distributions of *w* in a ten-part target corpus

1.2.3. Frequency in Egbert and Biber (2019): how is it treated?

The above – the coarse-grained approach to dispersion they use – provides a useful segue into the second main point. Egbert and Biber essentially discard frequency information by reducing it to a binary variable. Of course, they are aware of the fact that their discarding of frequency information is not completely uncontroversial, which is why they discuss it (briefly). In particular, they state:

> We hypothesised that keyness could be measured without making any reference to word frequency by focussing entirely on the text dispersion of words. In part, this hypothesis was based on the fact that a word occurring in numerous texts will necessarily also have at least a moderate frequency (Egbert and Biber 2019: 84).

However, while their observation is partially correct, it also misses an important part of the picture. Yes, (logged) frequency and dispersion (e.g., *DP*) are highly correlated (a GAM regressing *DP* on logged frequency returns an R^2 of 0.924); see Figure 1 for data

[2] *DP* is calculated as follows: for each corpus part (e.g., a file), compute (i) how much of the corpus it constitutes (as a fraction of the whole corpus) and (ii) how much of the word in question it contains (as a fraction of the word's frequency). Then subtract all (i) values from all (ii) values, take the absolute values of those differences, sum them up, and divide by two.

[3] Interestingly, the dispersion measure D_A, which Egbert and Biber have been promoting in other work of theirs (Burch *et al.*2017) would also distinguish the two scenarios above (because, while it can take many orders of magnitude longer to compute than *DP* or another measure to be introduced below, D_A is highly correlated with *DP*), meaning that Egbert and Biber (2019) uses a dispersion measure that is much more coarse-grained than the one they discuss elsewhere.

from the spoken component of the *British National Corpus* (BNC): logged frequency is on the *x*-axis, *DP* on the *y*-axis, each grey point is a word type, and the blue ranges represent the range of *DP*-values in ten different frequency bins.

The most important point about this plot is not the correlation, but, as pointed out by Gries (2019a: 117–119), that the correlation between frequency and dispersion is really only very strong for the most frequent words, which are frequent and of course evenly dispersed, and for the rarest words, which are rare and of course very much underdispersed. However, the former are unlikely to be good keywords because they are often function words and the latter are unlikely to be good keywords because they are too rare. But in the middle range of values, i.e. exactly where the relatively frequent content words one might be interested in are located, that is where the correlation between frequency and dispersion breaks down. For example, the sixth frequency bin from the left includes words with frequencies between 2,036 and 5,838 (such as the words *council* and *nothing* represented by the *c* and the *n*) and *DP*-values between 0.23 and 0.86, i.e. a *DP*-range of 0.63 also noted in blue at the bottom of the scatterplot, and R^2 for the correlation between frequency and *DP* in the sixth bin is in fact 0.086.

Given the above, it is risky to argue that dispersion can replace frequency because the two are correlated when that very correlation actually breaks down exactly in the frequency bins that contain the words that keywords analyses would be most interested in.

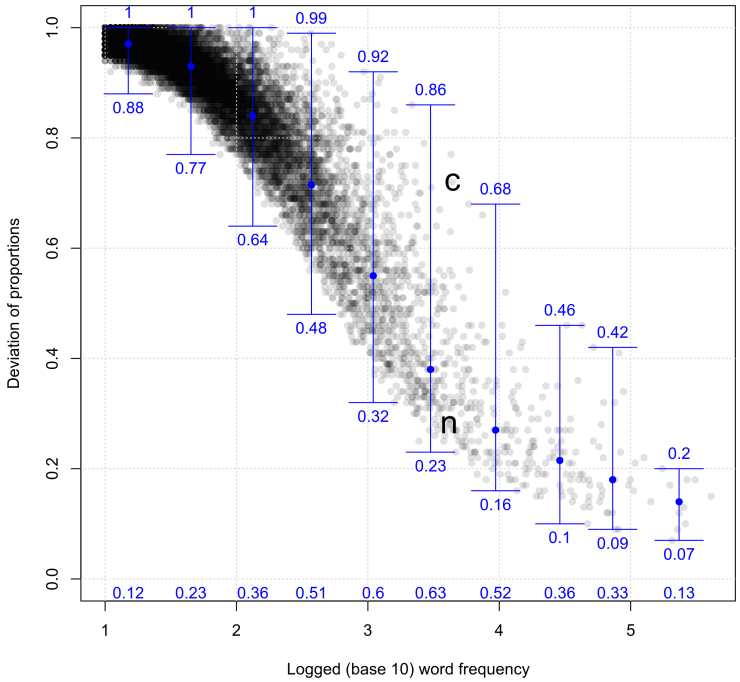

Figure 1: The correlation between frequency and *DP* in the spoken *British National Corpus*

1.3. Gries (2018) / (2019b)

There is a more general point to be made with regard to the approach advocated for by Egbert and Biber (2019) and how they replace frequency – as the fundamental measurement unit of keyness – with dispersion. That more general point was most recently and most pertinently for the present case discussed in Gries (2019b), who argues that all sorts of corpus-linguistic measures should be reconceptualized with an eye to avoiding 'conflation' of multiple separate dimensions of information into a single nicely sortable score and instead using 'tupleization', i.e. keeping multiple separate dimensions of information separate within a tuple. What does that mean in general and for here?

Corpus linguistics as a distributional discipline in general, and the more quantitative parts of it in particular, has a long history of quantifying the distributional patterns of linguistic units with statistical indices: corpus linguists report frequencies of occurrence (raw, normalized, and/or adjusted) and of co-occurrence, association measures quantifying co-occurrence patterns, dispersion scores, keyness scores, etc. Crucially, these statistical indices often serve the purpose of sorting the elements for which they are computed. For instance, in collocation studies we compute association scores to find the collocates most strongly attracted to our node word or the collocates that distinguish between multiple node words (e.g., near synonyms); in keywords analyses we compute keyness values (usually using association measures) to find the words most representative of a certain corpus/register type; in lexicography, we compute adjusted frequency values to find how much in use a word is, etc.

However, in the vast majority of applications, the measures we use for these purposes conflate different kinds of information:

(i) for collocation/collostructional work: many of the most widely-used association measures conflate two separate dimensions, namely frequency of the target word in question and the strength of its association to something else; this is particularly true of all measures that are related to, or derivative of, a significance test and thus affects measures such as the *LLR*, $p_{\text{Fisher-Yates exact test}}$, *t*, *z*, and others.

(ii) for keywords analyses: since these analyses are basically done with association measures (nearly always with *LLR*), the above point applies to them as well;

(iii) for adjusted frequencies in lexicography: these adjusted frequencies are computed with some combination of observed frequency and dispersion (such as multiplying the observed frequency of a word by Juilland's *D* for that word)[4] so that words with the same frequency but different dispersions receive different values.

However, Gries (2019b: 395) argues that this conflation of information is not a good idea because it, too, loses a lot of information; the following is worth quoting at length:

> For instance, the products of observed frequency and 1-*DP* [to make the dispersion value be small for underdispersed words] for the two words *pull* and *chairman* in the spoken BNC are very similar – 375 and 368.41 respectively – but they result from very different frequencies and dispersions: 750 and 0.5 for *pull* but 1939 and 0.81 for *chairman*. Not only is it the dispersion value, not frequency, that reflects our intuition (that *pull* is more basic/widely-used than *chairman*) much better, but this also shows that we would probably not want to treat those two cases as 'the same' as one implicitly does when one simply computes and reports one conflated adjusted frequency.

Gries (2019b: 395) goes on to extend this point to this paper's topic, keywords analyses:

> The same is true of key words, as mentioned above: key-word statistics based on 2×2 tables with one word (present vs. absent) in the rows and, say, two corpora in the columns have virtually always neglected to take into consideration how evenly dispersed in the two corpora the two words whose frequencies are listed in cells a and b are, a flaw that undermines parts of every single key words analysis.

Thus, Egbert and Biber (2019) and Gries (2019b) agree that keyness analyses are potentially deficient because of their not including dispersion information, but their recommendations as to how to deal with that problem are different:

(i) the former make a proposal where a single dispersion index for each word replaces the single association measure for each word (which, typically, is computed on corpus-wide frequency information and typically conflates frequency and association);
(ii) the latter makes a proposal where dispersion information 'augments' (i) the association information of how much a word 'likes' (or prefers) a corpus (over another) and (ii) the frequency information (how frequent is the word).

[4] Juilland's *D* is based on the variation coefficient of the percentages that the word in question makes up of each corpus part, with a normalization for the number of corpus parts, see Gries (2021).

1.4. Overview of the present paper

Given all of the above, the present paper is exploratory in nature and tries to address two goals:

(i) The first goal is to develop an approach to key words that, just like Egbert and Biber's proposal, goes beyond the corpus-frequency-based way, but then also extends and hopefully improves their approach in two steps. First, I will propose a new keyness measure that is also just based on frequency (i.e., does not yet include dispersion), but that, I believe, nevertheless constitutes a useful improvement of what is currently the default approach, *viz. LLR*, because of how it is less correlated with frequency than *LLR*.

(ii) Second, I will extend this improvement in two novel ways: on the one hand, 'extend' here means that, unlike Egbert and Biber (2019), dispersion information will be added to the frequency information, rather than replace it. On the other hand, the dispersion information in this approach will be computed in a way that is very similar to the way in which I propose to improve on the frequency information: it essentially relies on the same measure.

Section 2 will introduce and exemplify both proposed improvements on the basis of a small corpus, the *Clinton-Trump Corpus* (Brown 2016); Section 2.1 will briefly apply a traditional keywords analysis using *LLR* to the corpus, Section 2.2 will introduce the new frequency-based keyness measure, and Section 2.3 will introduce and add the dispersion-based keyness measure. Section 3 will apply the new method to a much larger example and one that is maybe more typical of keywords applications, namely academic-writing keywords in the BNC. Section 4 will conclude.

2. DEVELOPING A NEW APPROACH TO (KEY) KEYNESS

2.1. Introduction

In order to exemplify the improvements to be proposed, I will use the *Clinton-Trump Corpus*, which contains ≈117K words from 36 speeches of Hillary Clinton's 2016 presidential campaign and ≈446K words from 82 speeches of Donald Trump's 2016 presidential campaign. When that corpus is converted to lower case and tokenized at one or more occurrences of the Unicode category of non-letter characters (the PCRE

regex in *R* was "[^\\p{L}]+"), the corpus contains 563,019 word tokens / 10,317 word types. If one applies the traditional/default kind of keyword analysis to this data set, trying to identify words characteristic/key for Hillary Clinton's speeches using *LLR*, the top 50 keywords are those listed in (4); on the whole, those results seem not too bad, especially when compared to the corresponding top 50 from Donald Trump's speeches shown in (5).

(4) *he*, *his*, *donald*, *together*, *work*, *my*, *economy*, *who*, *college*, *families*, *young*, *election*, *president*, *help*, *kind*, *rights*, *america*, *kids*, *sure*, *to*, *as*, *stronger*, *someone*, *trump*, *com*, *am*, *for*, *hard*, *women*, *everyone*, *fairer*, *grateful*, *can*, *khan*, *commander*, *dad*, *each*, *i*, *should*, *small*, *insults*, *about*, *hillaryclinton*, *campaign*, *challenges*, *gun*, *family*, *senate*, *that*, *nuclear*

(5) they, *hillary*, *she*, *re*, *clinton*, *going*, *very*, *it*, *bad*, *s*, *folks*, *great*, *percent*, *ok*, *trade*, *t*, *don*, *obamacare*, *win*, *borders*, *money*, *her*, *mexico*, *nafta*, *ll*, *illegal*, *border*, *incredible*, *media*, *over*, *these*, *disaster*, *tremendous*, *politicians*, *deals*, *will*, *right*, *china*, *massive*, *look*, *dishonest*, *unbelievable*, *corrupt*, *deal*, *donors*, *administration*, *happen*, *never*, *hell*, *like*

However, recall from above that *LLR* as a measure combines the information of the overall token frequency of a word type (i.e., $a+b$) with association information; in other words, *LLR* increases

(i) when the word in question becomes more associated to a corpus and becomes stronger even if its overall frequency remains the same,[5] but also

(ii) when the word in question becomes more frequent even if the association to the corpus actually remains the same.[6]

---

[5] The reader can verify this easily by running the following code in *R*:

```
addmargins(lo.assoc <- matrix(c(100, 999900, 50, 999950), ncol=2))
  (100/999900) / (50/999950)
  2*sum(lo.assoc * log((lo.assoc/chisq.test(lo.assoc)$exp)))
addmargins(hi.assoc <- matrix(c(125, 999875, 25, 999975), ncol=2))
  (125/999875) / (25/999975)
  2*sum(hi.assoc * log((hi.assoc/chisq.test(hi.assoc)$exp)))
```

Lines 1 and 4 generate two tables called lo.assoc and hi.assoc that might result from comparing a word's frequency in two one million-word corpora. While the frequency of the word is the same in both tables (150), the *LLR*-values are of course very different: 16.99 for lo.assoc and 72.78 for hi.assoc.

[6] The reader can verify this easily by running the following code in *R*:

```
addmargins(hi.freq <- matrix(c(200, 999800, 100, 999900), ncol=2))
  (200/999800) / (100/999900)
  2*sum(hi.freq * log((hi.freq/chisq.test(hi.freq)$exp)))
addmargins(lo.freq <- matrix(c(100, 999900, 50, 999950), ncol=2))
  (100/999900) / (50/999950)
  2*sum(lo.freq * log((lo.freq/chisq.test(lo.freq)$exp)))
```

Let us therefore look at the results by representing both frequency and *LLR* in the picture: Figure 2 and Figure 3 both plot all word types in the corpus at coordinates reflecting their overall frequency in the corpus (*x*-axis, logged) and their signed *LLR*-value (on the *y*-axis), positive *LLR*s represent 'Clinton words', words from her top 50 are in blue, and for ease of visual scanning, Figure 3 zooms into Figure 2.[7]

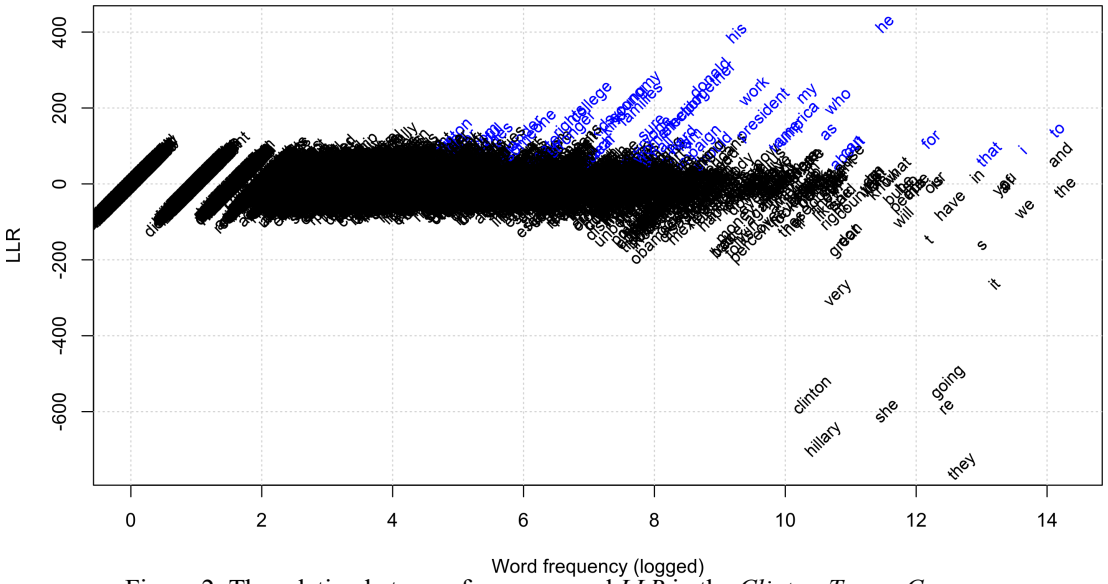

Figure 2: The relation between frequency and *LLR* in the *Clinton-Trump Corpus*

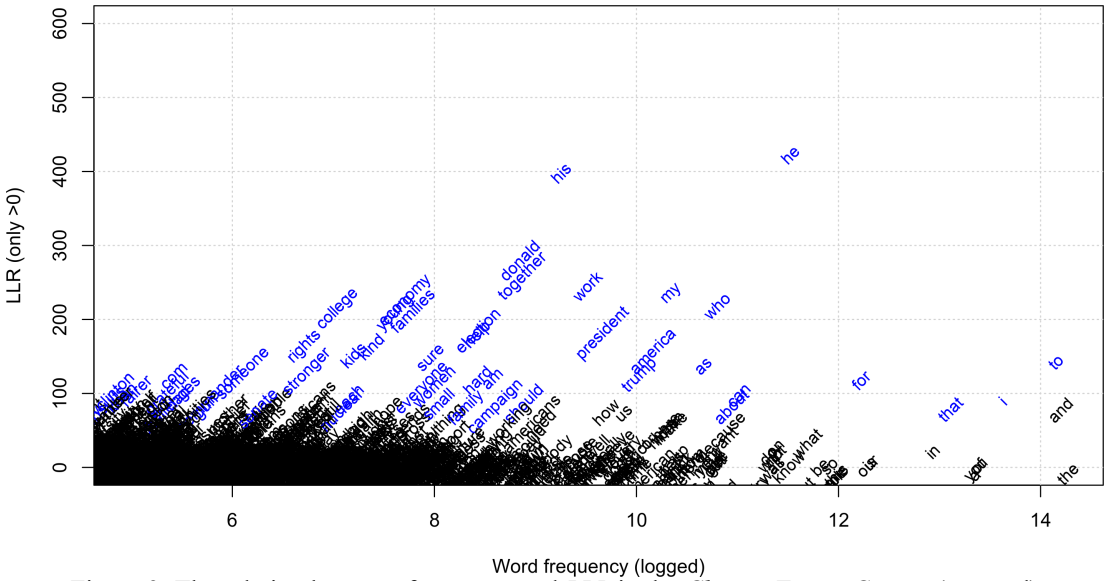

Figure 3: The relation between frequency and *LLR* in the *Clinton-Trump Corpus* (zoomed)

Lines 1 and 4 generate two tables called hi.freq and lo.freq that might result from comparing a word's frequency in two one million-word corpora. While the association of the word to the first (left-column corpus) is the same in both tables (odds ratio of ≈2), the *LLR*-values are very different: 33.98 for hi.freq and 16.99 for lo.freq.

[7] The visual representation of Figure 3 can be improved by logging the *LLR*-values but given the rarity of this kind of transformation I am not showing this plot here.

If nothing else, these plots show two things. First, there is a bit of a positive correlation between the absolute *LLR*-values and frequency: *LLR*-values are mostly only high when the word is 'reasonably' frequent (R^2_{GAM} regressing abs(*LLR*) on frequency = 0.293). Second, in spite of that correlation, *LLR* is still indeed a conflation: even restricting our attention to the top 50 words, one finds that:

(i) sometimes, words with fairly similar *LLR*-values also have fairly similar frequencies (see *donald* and *together* or *economy*, *young*, and *families*);[8]
(ii) sometimes, words with fairly similar *LLR*-values have very different frequencies (see *who* and *college* or *stronger* and *trump*);
(iii) sometimes, words with fairly different *LLR*-values have very similar frequencies (see *should* and *donald* and *together* or *everyone* and *economy*).

Clearly, *LLR*-values lose quite a bit of information: just from looking at a word's keyness *LLR*-value, it is quite hard to see to what degree the word owes its *LLR*-value to a high overall frequency in both *t* and *r* and, say, a moderate association or to a moderate frequency but a high association, as exemplified in footnotes 5 and 6. It is this information loss that the following sections are trying to combat.

2.2. Improvement 1: A new keyness measure using frequency information

As a first (smaller) improvement, I am proposing a different keyness measure. The first of its two main advantages is that it is less related to frequency and, thus, amounts to less of a conflation; the second advantage will be discussed below. This measure is an information-theoretic measure called the 'Kullback-Leibler (KL) divergence'. The KL divergence is written as D_{KL} (posterior/data || prior/theory), which in the present context refers to how much the probability distribution of the two corpora given the word we are currently looking at (the posterior) diverges from the percentage distribution of the corpus sizes (the prior). That means, it is computed from the same kind of 2×2 table as Table 1. Consider Table 4 for the frequency distribution of the word *college* in our corpus, with row percentages added for the first row and the column totals, and let's refer to the two column totals as cells *e* and *f*.

[8] This finding is not due to occurrences of *young families* as a collocation.

	Target corpus *t* (Clinton)		Reference corpus *r* (Trump)		Sum
Target word (i.e., *college*)	106	0.80303 ($=^{106}/_{132}$)	26	0.19697 ($=^{26}/_{132}$)	132
Other words	117,183		445,704		562,887
Sum	117,289	0.20832 ($=^{117289}/_{563019}$)	445,730	0.79168 ($=^{445730}/_{563019}$)	563,019

Table 4: Data to compute D_{KL} for the keyness of *college* for Clinton

D_{KL} (p(corpus|"college")‖p(corpus) is how much the probabilities of the two corpora, given we are looking at *college* (i.e. *a*=0.80303 and *b*=0.19697), diverge from the two overall probabilities of the two corpora (i.e. *e*=0.20832 and *f*=0.79168). It is computed using the probabilities – not the frequencies! – in the table's cells *a*, *b*, *e*, and *f*, as shown in (6).

(6) $D_{KL}\big(p(corpus|college) \parallel p(corpus)\big) = \left(a \times log_2 \frac{a}{e}\right) + \left(b \times log_2 \frac{b}{f}\right) \approx 1.168$

As a (directional) divergence, D_{KL} values range from 0 (the two probability distributions are identical) to, theoretically, +∞ so how do we interpret this? We might already guess that, for our current data, this value might be on the higher end of things simply because, while the Clinton part of the corpus is only ≈20 percent of the total corpus, she accounts for ≈80 percent of all uses of *college*; that should be 'noteworthy' (and the *LLR*-value for Table 3 is 308.423, i.e. quite high for this corpus). Second, just like in a traditional keyword analysis, we can compare this with other words: Table 5 contains the results for the word *instead*.

	Target corpus *t* (Clinton)		Reference corpus *r* (Trump)		Sum
Target word (i.e., *instead*)	26	0.19697 ($=^{26}/_{132}$)	106	0.80303 ($=^{106}/_{132}$)	132
Other words	117,263		445,624		562,887
Sum	117,289	0.20832 ($=^{117289}/_{563019}$)	445,730	0.79168 ($=^{445730}/_{563019}$)	563,019

Table 5: Data to compute D_{KL} for the keyness of *college* for Clinton

Obviously, I chose this example because here the two frequencies *a* and *b* are reversed, meaning that the distribution of the word *instead* across the corpora is nearly perfectly proportional to the corpus sizes; I invite the reader to determine that D_{KL} for Table 5 is ≈0.0006 (and LLR≈0.151). In other words, the distribution of *college* diverges much more from that of the corpus sizes than the distribution of *instead* does (*college*'s D_{KL}-value is >2000 as high as *instead*'s because *college* is so overrepresented in the Clinton data), and just like with LLR we can leave the sign of D_{KL} as positive when 'the word prefers Clinton' and set it to negative when 'the word prefers Trump'.

What happens if we apply this to all words and plot it again just like we did for the LLR-values above? The result, using signed D_{KL}, is shown in Figure 4 (already zooming in and showing only the words 'preferring the Clinton corpus').

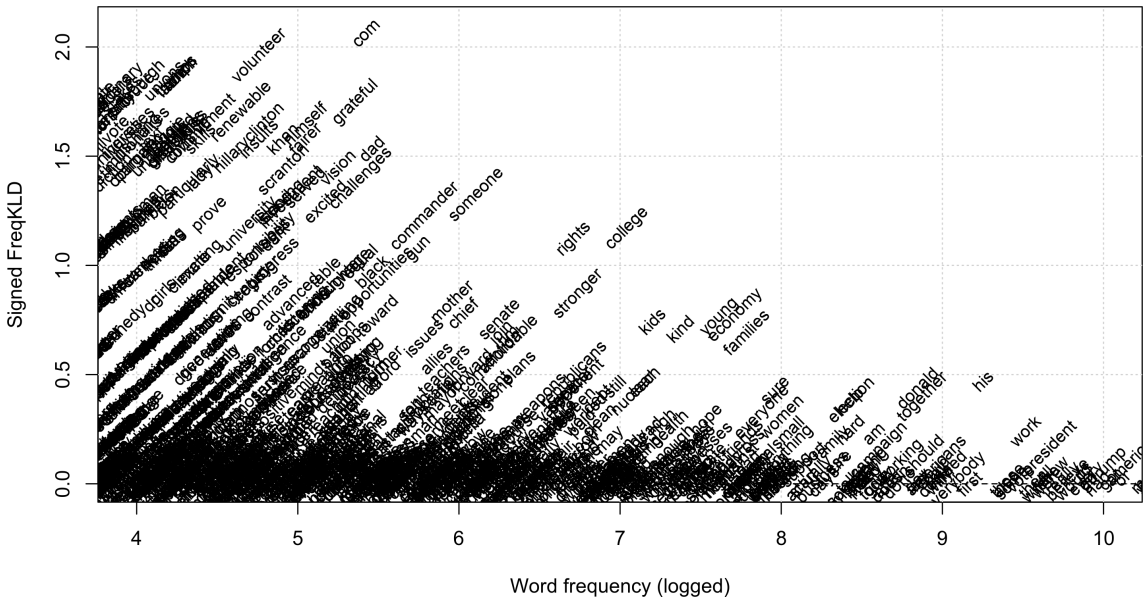

Figure 4: The relation between frequency and D_{KL} in the *Clinton-Trump Corpus* (zoomed)

What do the results show? First, and this is important, D_{KL} as a keyness measure is much less related to overall word frequency than LLR (R^2_{GAM} regressing abs(D_{KL}) on frequency = 0.0283), which means that D_{KL} is less of a conflation of frequency and association than LLR.[9] Thus, D_{KL} is better at capturing association to a corpus (i.e. keyness) 'above and beyond frequency' than LLR.

Second and in the spirit of 'tupleization', that of course also means that, to identify keywords, one would not really look at the top 50 words in terms of D_{KL} –

[9] This is of course not surprising: D_{KL} as computed here is mathematically equivalent to $LLR_{\text{cells } a,\ b}$ divided by $a+b$ (without the doubling).

instead, one would look at the upper and right margin of the word cloud in Figure 4, i.e. at words that have both a (relatively) high frequency of occurrence 'and' a (relatively) high D_{KL}-value. Some words that stick out like that are listed in (7) (with two uncertain ones parenthesized).

(7) *volunteer*, *skills*, *renewable*, *hillaryclinton*, *insults*, *khan*, *himself*, *fairer*, *grateful*, *com*, *vision*, *dad*, *excited*, *challenges*, *black*, *commander*, *gun*, *someone*, *mother*, *chief*, *(senate)*, *stronger*, *rights*, *college*, *kids*, *kind*, *young economy*, *families*, *sure*, *everyone*, *(women)*, *election*, *hard*, *donald*, *together*, *his*, *work*, *president*, *trump*

Obviously, there, is some overlap with *LLR* and, as a result, there is a bit of uncertainty there given the visual/heuristic identification of the keywords above. However, this is less reason for concern than one might think. As for the former, if anything, it is good that there is some overlap because it means that both measures are, if only to different extents, 'up to something', but the advantage of D_{KL} is that it separates association and frequency more cleanly than *LLR* does. In other words, we see that words like *work*, *president*, *trump* owe their keyness status more to high frequency than association, and we see that words like *volunteer*, *renewable*, *khan*, *insults*, and *fairer* owe their keyness status more to high association than to frequency; this kind of recognition is only possible if we keep frequency and association separate,

(i) minimally, by using a measure that conflates frequency and association (i.e. *LLR*) but at least also plotting frequencies as in Figure 2/Figure 3;
(ii) ideally, by using a measure that keeps frequency and association separate (i.e. D_{KL}) and plotting both frequency and association as in Figure 4.

As for the latter, the seemingly subjective choice of words in the margin should not be much of an issue for two reasons. First, if one is being honest, the interpretation of keywords using a sorted *LLR* list is also subjective in some respects at least. Let's face it: if one chooses to explore the top 100 *LLR* keywords, the choice of 100 is more due to our affection for the decimal system and round numbers than anything else, let alone scientific or objective criteria. The same happens when scholars choose a usually arbitrary *LLR* cut-off point, e.g., Scott and Tribble's (2006: 77), threshold value of the *LLR*'s *p*-value of 10^{-6}). Strictly speaking, one should:

(i) either use *LLR*=3.841 as a cut-off point (because that is the LLR-value denoting significance in a single 2×2 table); this would leave us with 2,597

keywords, a number of keywords far higher than those that most people ever explore/discuss);

(ii) or one should use an *LLR*-value that corresponds to a significant result when one corrects for the number of (posthoc) tests one is doing, i.e. the number of word types/2×2 tables for the data; given that the corpus has 10,317 different word types, Holm's correction would leave us with 567 keywords.

Alas, there are very few studies which adopt either one of these more objective standards, in particular the posthoc correction approach to keyness (argued for and somewhat validated in Gries 2005: 281–282) is hardly ever used.[10] Thus and with all due respect, users of the either one of the above two approaches would be well advised to recognize the issues of these approaches before considering to criticize the combination of D_{KL} and frequency, which, at least, uses a better/cleaner statistical measurement tool to separate frequency and association.

2.3. Improvement 2: A new keyness measure using frequency and dispersion information

2.3.1. Motivation

The first improvement proposed above consisted of a new keyness measure whose first advantage was that it offers a cleaner separation of frequency and keyness (i.e., association to a corpus) than most previous work. However, that first improvement does not yet consider dispersion although dispersion is an extremely important corpus statistic in general and although Gries (2018) and Egbert and Biber (2019) have shown it seems to be useful in a keywords context in particular. In this section, I will therefore discuss how to add dispersion to the keywords analysis, a goal that of course immediately raises the next questions, namely (i) how exactly to include dispersion in keyness conceptually and, provided this question can be addressed, (ii) which dispersion measure to use.

[10] Gries (2005) shows that (i) counter to Kilgarriff (2005), statistical significance testing on (word frequency) corpus data is *not* bound to 'almost always' leading to significant results, because (ii) when corrections for multiple testing are applied, it is possible to get a number of baseline false hits that is in fact close to the 0.05 threshold that significance testing typically relies on.

As for (i), the degree to which a word *w* is considered a keyword, or a key keyword, for a target corpus/text type *t* should increase with *w*'*s* more even dispersion in *t*. This way, one would rule out that, for instance, the name of an author of a quoted specialized article becomes a keyword for *t* even if that author is only mentioned in a tiny part of *t*. At the same time, however, *w* would also be a stronger keyword when it is not also very evenly dispersed in the reference corpus *r*. This way, we rule out that function words like determiners or prepositions, which will be evenly dispersed in *t*, become keywords – they will also be evenly dispersed in *r*, because they are in fact evenly dispersed in pretty much any corpus. Combining these two notions seems straightforward: one could compute the difference in dispersion of *w* in both *t* and *r*, and if *w* is evenly dispersed in *t* and unevenly dispersed in *r* (perhaps only occurring in a part of *r* that is topically similar to *t*), then *w* is most likely a key word. This implies that we might use a dispersion measure that can be compared across corpora so that, for instance, the fact that *r* is usually bigger than *t* does not affect the results, which rules out measures such as chi-squared – ideally, the measure might fall between, say, 0 and 1, to be most useful.

Thus, let us turn to (ii), the question of which dispersion measure to use. Just like for association measures/collocation statistics, a sizable variety of dispersion measures have been proposed (see Gries 2008 for the most recent comprehensive overview). The simplest one, 'range', I have already argued against above, both in general and in Egbert and Biber's version of using 'range' for *LLR*. An alternative measure would be *DP*, which was briefly discussed above and which indeed falls between 0 and 1 as might be desired. However, the current proposal actually follows Gries (2021) and uses the same measure we have used before, D_{KL}, this time as a measure of dispersion. The computation is essentially done as before: the posterior distribution becomes the percentage distribution of a word *w* across the parts of a corpus (*t* or *r*) and the prior distribution becomes the percentage distribution of the corpus part sizes (of *t* or *r*).

Let us look at an example for this, for which we return to the upper panel of Figure 1 from above, repeated here in the first two rows of Table 6; recall that *w* occurred six times and that the corpus contained 31,000 tokens.

	Part 1	**Part 2**	**Part 3**	**Part 4**	**Part 5**	**Part 6**	**Part 7**	**Part 8**	**Part 9**	**Part 10**
# *w*	0	0	0	0	0	0	0	2	2	2
part size	1,000	1,000	2,000	2,000	3,000	3,000	4,000	5,000	5,000	5,000
↓										
p	0	0	0	0	0	0	0	0.3333	0.3333	0.3333
q	0.0323	0.0323	0.0645	0.0645	0.0968	0.0968	0.129	0.1613	0.1613	0.1613
↓										
$\log_2(p/q)$	0	0	0	0	0	0	0	1.0473	1.0473	1.0473
↓										
$p\times\log$	0	0	0	0	0	0	0	0l.3491	0.3491	0.3491
↓										
$\Sigma\, p\times\log=$	1.0473		→		$1\text{-}e^{-DKL}=$	0.6491				

Table 6: The computation of D_{KL} as a dispersion measure in a ten-part target corpus

The then following two rows (*p* and *q*) convert the word frequencies and corpus part sizes into percentages: $^{2}/_{6}$=0.3333 and, e.g., $^{4000}/_{31000}$=0.129. The next row computes $\log_2(^{p}/_{q})$, which is set to zero if the fraction returns 0. The next row computes all products *p* times $\log_2(^{p}/_{q})$, and the final row sums that up into D_{KL}=1.0473. By default, D_{KL} does not fall into the range [0,1], but with a straightforward transformation ($1\text{-}e^{-DKL}$), we can normalize D_{KL} *to* fall into that range easily. Now we have a dispersion measure that ranges from 0 to 1 as desired, and this is the second advantage alluded to before in Section 2.2: rather than proliferate measures, we are using the same kind of information-theoretic measure to quantify a word's dispersion as we used before to quantify the same word's frequency difference in the target and the reference corpus. The next section will apply this tupleized two-part measure of keyness to the *Clinton-Trump Corpus* data.

2.3.2. Analysis and results

In order to exemplify the current approach to dispersion, we will need a plot representing minimally two dimensions:

(i) on the *x*-axis, we will represent the words' behavior with regard to frequency by plotting a signed normalized version of D_{KL}. This sounds complex, but only means that values in the range [-1,0) will represent words whose

frequency distribution makes them Trump keywords whereas values in the range (0,1] will represent words whose frequency distribution makes them Clinton keywords. The more a value deviates from 0, the stronger a word's frequency preference for either Trump or Clinton, i.e. the strongest Trump/Clinton words in terms of frequency will be far on the left/right respectively.

(ii) on the *y*-axis, we will represent the words' behavior with regard to dispersion by plotting the difference of a signed normalized version of a word's dispersion in DKL. Specifically, we will plot a word's D_{KL}-dispersion in the Trump corpus minus the same word's D_{KL}-dispersion in the Clinton corpus; that way, high values of these differences will represent words that are much more evenly distributed in the Clinton corpus than in the Trump corpus (see Table 7 for examples), i.e. the strongest Clinton/Trump words in terms of dispersion will be at the top/bottom respectively.

	***D*KL Clinton: 0**	***D*KL Clinton: 0.333**	***D*KL Clinton: 0.667**	***D*KL Clinton: 1**
***D*KL Trump: 0**	0	-0.333	-0.667	-1
***D*KL Trump: 0.333**	0.333	0	-0.333	-0.667
***D*KL Trump: 0.667**	0.667	0.333	0	-0.333
***D*KL Trump: 1**	1	0.667	0.333	0

Table 7: Differences of D_{KL}Trump - D_{KL}Clinton for different D_{KL}-values

This kind of representation will then allow us to see for each word type *w* whether or not it is over-represented frequency-wise in the Clinton corpus relative to the Trump corpus, but also how it behaves dispersion-wise in the Clinton corpus relative to the Trump corpus; see Figure 5 for an overview of all results and Figure 6 for a version zooming into the words key for the Clinton corpus.

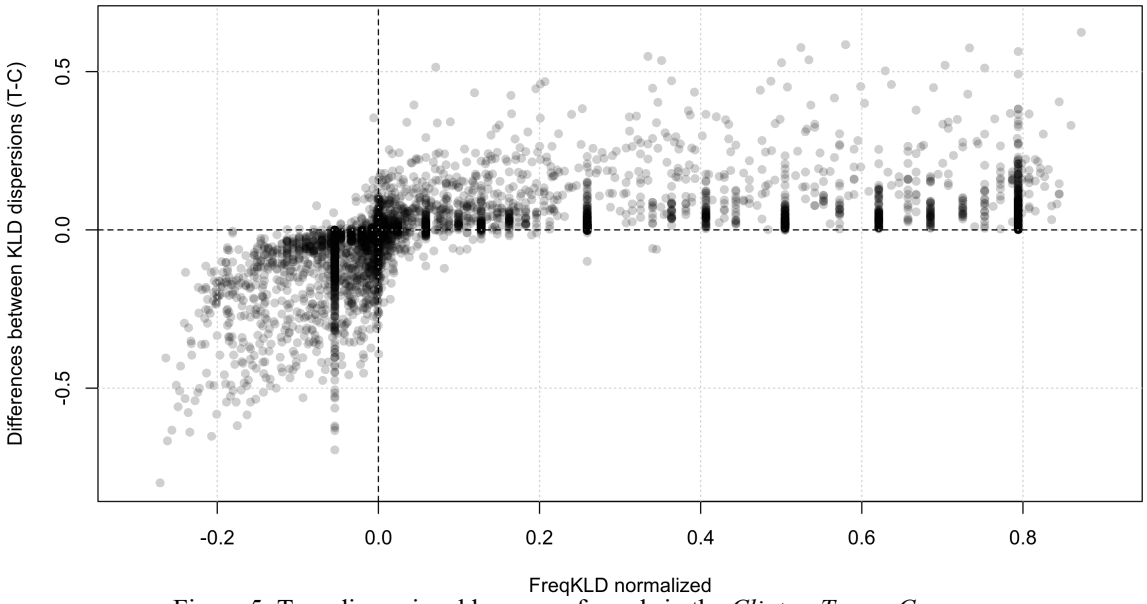

Figure 5: Two-dimensional keyness of words in the *Clinton-Trump Corpus*

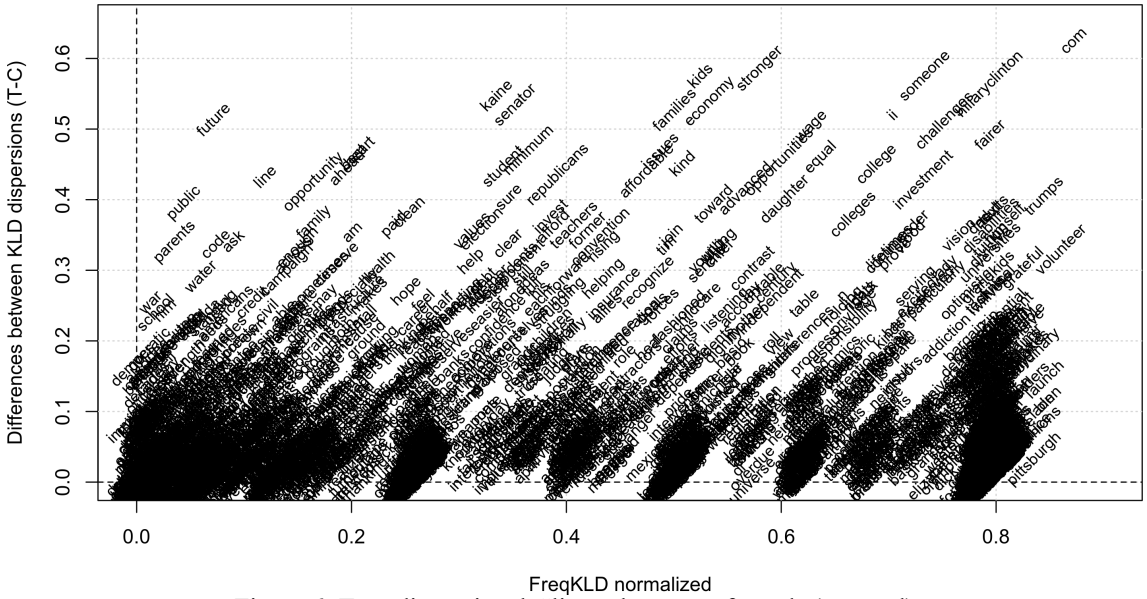

Figure 6: Two-dimensional Clinton keyness of words (zoomed)

On the whole, there is clearly a correlation: in Figure 5, most points are in the lower left and the upper right quadrants, meaning that words that are Clinton keywords in terms of their frequency patterning are also Clinton keywords in terms of their dispersion. However, it is also obvious that this is not always the case: there are some words that are Clinton keywords in terms of frequency but Trump keywords in their dispersion (words/points in the lower right quadrant such as *crisis*, *ready*, *compete*, *wealthy*, *trump* (!), *joe*, *nuclear*, *service*) and the other way round (words/points in the upper/left quadrant such as *great*, *poverty*, *story*, *congress*, *new york*, *state*, *decent*); an analysis that does not incorporate both frequency and dispersion would not find these. As for the

Clinton words represented in Figure 6, we can now explore them in more detail. The most key Clinton word is her providing the link to her website: *hillaryclinton* and *com*, the word types highest up and rightmost; other words scoring high on both dispersion and frequency are *fairer* and *challenges*, and maybe *college*, *colleges*, and *investment*. The word *fairer*, for instance, is used 33 times by Clinton (281 pmw) and in more than half of her speeches, but not once by Trump; the word *challenges* is used 36 times by Clinton (307 pmw) in about two thirds of her speeches, but only six times by Trump (13.5 pmw).

At the same time, there is a variety of words that are very key for Clinton in terms of dispersion, but decreasingly so in terms of frequency: *equal*, *wage*, *opportunities*, *stronger*, *economy*, *kids*, *families*, *republicans*, *minimum*, *student*, *senator*, *kaine*, *clean*, *paid*, *ahead*, *opportunity*, *line*, *code*, *ask*, *future*, *public*, and *parents*. In other words, these are words that are in many of Clinton's speeches (compared to Trump's), even if the frequency with which she uses them is not that high (compared to Trump's). For instance, Clinton uses the word *stronger* 77 times (656.5 pmw) in 32 out of 36 speeches (one speech has eight occurrences already), but Trump uses *stronger* frequently as well (30 times (67.3 pmw)), although not even in a quarter of his 82 speeches. Similarly, Clinton uses the word *economy* 145 times (1,236.3 pmw), which is a lot, but Trump also uses it 66 times (148.1 pmw); however, Clinton uses it in 90 percent of her speeches (32 out of 36) whereas Trump does so only in 44 percent (36 out of 82).

On the other hand, there are words that are quite key for Clinton in terms of frequency, but decreasingly so in terms of dispersion: *trump*, *volunteer*, *grateful*, *afraid*, *renewable*, *vladimir*, *fortunate*, *stakes*, *extraordinary*, *founders*, *launch*, *bruce*, *bin laden*. For just one example, Clinton uses *renewable* 23 times (196.1 pmw), whereas Trump does so only twice (4.5 pmw) – a relative frequency ratio of nearly $^{196.1}/_{4.5}$=44, the by far highest reported so far – but Clinton and Trump both do not use it in the majority of their speeches (14 out of 36 for Clinton and two out of 82 for Trump).

Let us finally make a brief – for considerations of space – comparison between the two-dimensional D_{KL}-based keyness and the traditional *LLR*-based approach. I retrieved all Clinton-favoring word types with an *LLR*-value of ≥50 from the data, which amounted to 101 different types. Then, I grouped those into six different groups (using a simple hierarchical cluster analysis so as to avoid me choosing six arbitrary values); the resulting groups were $50.2 \leq LLR \leq 65.89$, $70.29 \leq LLR \leq 86.09$, $89.92 \leq LLR \leq 128.66$,

135.8 ≤ *LLR* ≤ 184.8, 207.2 ≤ *LLR* ≤ 279.6, and 397.6 ≤ *LLR* ≤ 422.6. These 101-word types were then plotted in a reduced version of Figure 6 such that different colors indicate which word types belong into which *LLR*-clusters; this plot is shown in Figure 7.

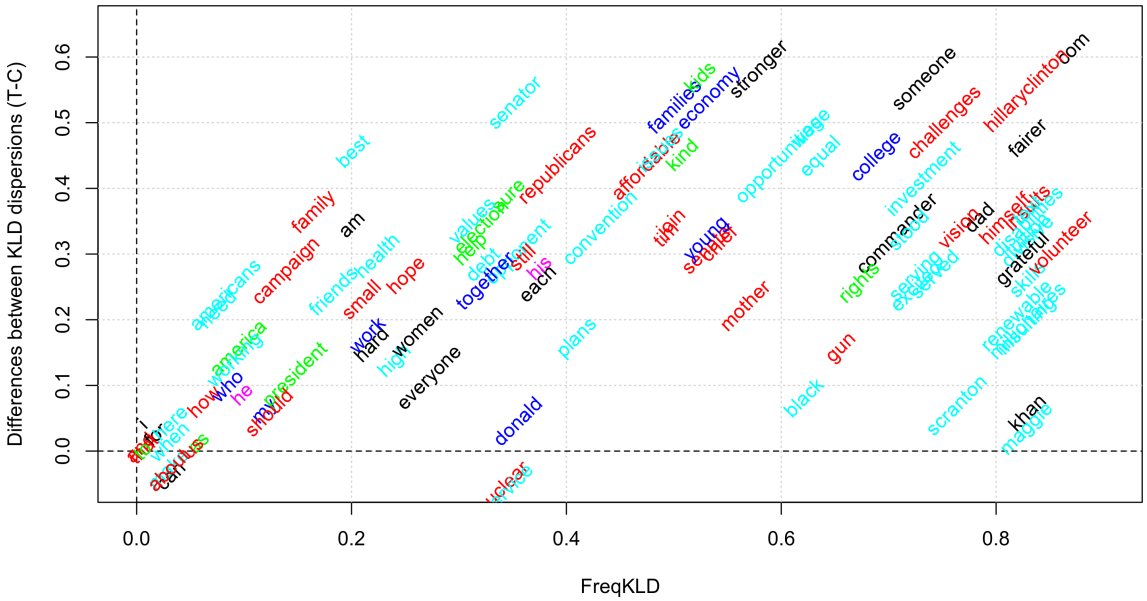

Figure 7: Two-dimensional D_{KL}-based keyness against *LLR*

Clearly, *LLR* loses a lot of information: for nearly every color, i.e. every group of relatively adjacent *LLR*-values, we find that the words are quite spread out over the plot. In other words, all red words are considered quite similar in terms of *LLR* even though we can plainly see that they can in fact be extremely different from each other. That is, from the *LLR*-value, it is nearly impossible to infer anything more specific about a word type's distribution in the corpora or, from the reverse perspective, words even with very similar *LLR*-values can behave completely differently. One of the most striking examples seems to be the word pair *hillaryclinton* (top right corner in red) and the word *about* (bottom left corner in red). Curiously enough, both words have for all practical intents and purposes the same *LLR*-value (nearly exactly 81.6±0.1) indicating 'Clintonness', but, in a way, they could not be distributionally less similar, as is obvious from Table 8, below.

	Clinton	**Trump**	**Sum**		**Clinton**	**Trump**	**Sum**
about	579	1386	1965	***hillaryclinton***	26	0	26
other	116,710	444,344	561,054	**other**	117,263	445,730	562,993
Sum	117,289	445,730	563,019	**Sum**	117,289	445,730	563,019

Table 8: Frequency distributions for *about* and *hillaryclinton*

The current approach shows that *hillaryclinton* is nearly perfectly key for Clinton's speeches: in terms of frequency of use, she uses it often (221.7 pmw) whereas it is not used by Trump at all (theoretically, this amounts to a relative frequency ratio of infinity); in terms of dispersion, she uses it in more than 60 percent of her speeches. However, *about* receiving the same *LLR*-value is a bit of a problem for the traditional keywords approach. In terms of frequency of use, Clinton uses it 4,936.5 pmw while Trump does so 3,109.5 pmw, which corresponds to a relative frequency ratio of not even 1.6; in terms of dispersion, both Clinton and Trump use it in every speech. Thus, *LLR* ranking *about* so highly is mostly only due to its high overall frequency, but neither to it being strongly preferred by Clinton frequency-wise nor to it being more widely used by Clinton. The extent of the problem of the traditional keywords approach is visualized in Figure 8.

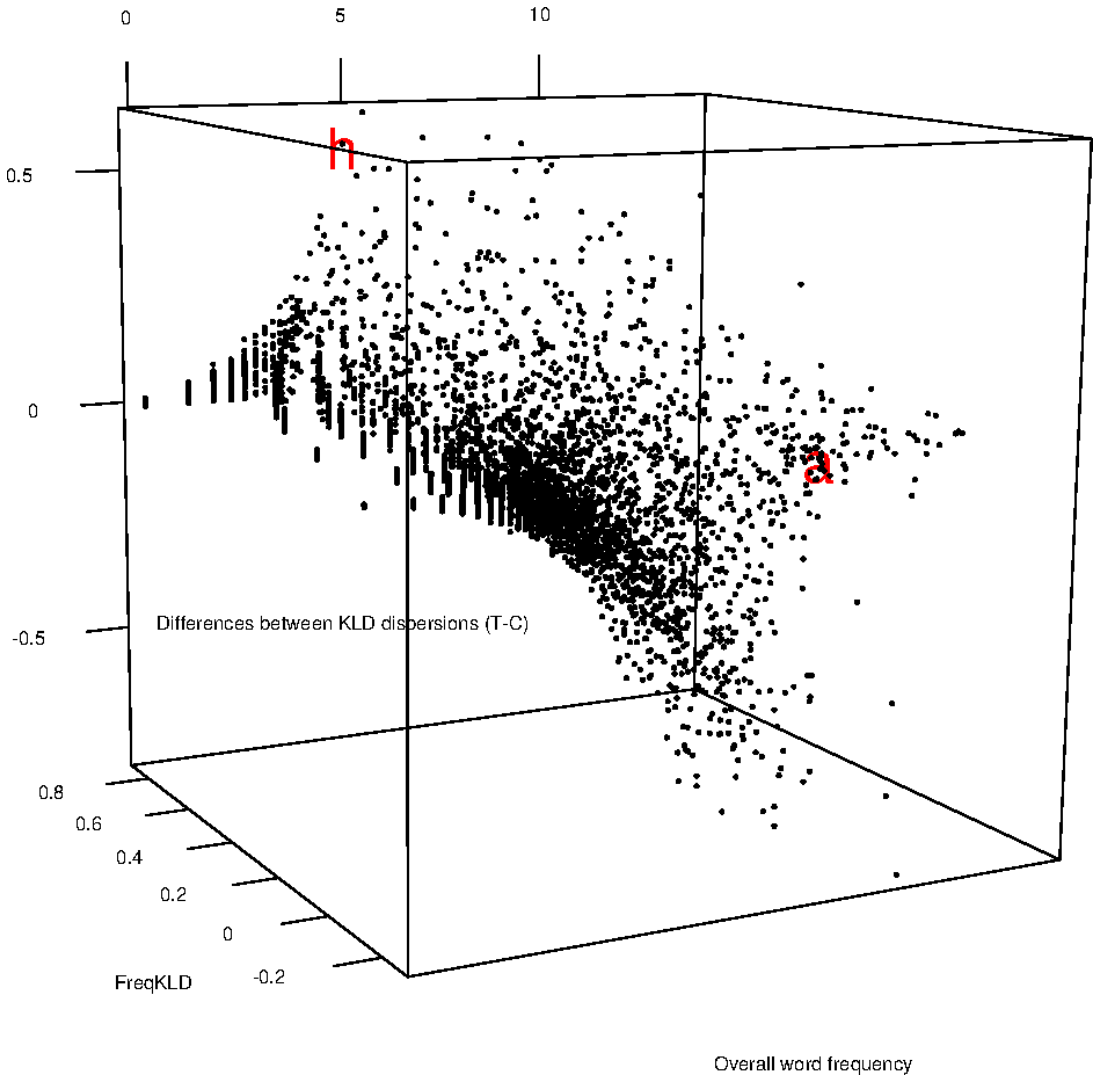

Figure 8: All words in the corpus in a space covering frequency and two-dimensional D_{KL}-based keyness

In Figure 8, the *a* and the *h* represent the positions of *about* and *hillaryclinton* in the three-dimensional space of overall word frequency and two-dimensional keyness. As one can infer, although both word types score about the same *LLR*-value, *about* only scores high on frequency (the *x*-axis) but, as we know, close to 0 on the other two dimensions, whereas *hillaryclinton* is high up and in the back of the plot, representing its high values on both keyness dimensions proposed here.

A potential counterargument to the above argumentation – in particular regarding *about* – might be that *about* is a function word that a keywords analyst would have excluded from analysis anyway, it would have been part of a stoplist or among the most frequent words in English in general. However, I do not consider this a good argument for two reasons. First, this might 'save' someone preferring the traditional method for this example – *about* and *hillaryclinton* – but not in other cases of the same general type. The fact remains that *LLR* is just very poor at distinguishing words with extremely different distributional characteristics; more polemically, but to make it really clear: the present example shows that *LLR* as a measure is so bad that it needs an analyst coming up with the right stopwords first, otherwise part of what it will return will be garbage – the method proposed here, however, works well without a stoplist: Table 8 showed clearly that, no matter what *LLR* says, *about* is not a keyword for Clinton.

Second, Egbert and Biber (2019) did not use a stoplist and also showed convincingly that even their 'range'-based approach not only does not rank many function words highly, but also that function words that are ranked highly, can be useful: in their case, the word *around* "is quite easy to interpret as a travel-related word" (2019: 95).

2.4. Interim conclusion

In conclusion, the proposed approach seems to work very well. I began by demonstrating that the traditional approach using *LLR*-values is problematic in how it (i) conflates word type frequency (*a*+*b*) and association in a not-so-helpful way and, of course, (ii) does not include dispersion information. I first introduced a new frequency-based keyness measure, the 'Kullback-Leibler divergence', that is well-grounded in information theory and much less correlated with frequency, allowing the researcher to

keep different dimensions of information separate for a more precise picture of how words are distributed across the target and the reference corpus.

I then developed the notion that keyness measures should include dispersion information. However, counter to Egbert and Biber (2019), I proposed that dispersion information should *augment*, not *replace*, frequency information, and I showed how that can be done using, again, the 'Kullback-Leibler divergence'. The results not only indicate that, with this finer resolution, words can be key because of their frequencies, their dispersion, or both; in addition, the proposed approach is able to tease apart distributional differences even between words whose *LLR*-values are virtually identical and may just be due to high overall frequency of occurrence (as opposed to anything having to do with keyness).

The next section will apply the same methodology to a different example, one that differs both in scale and in content/application: in the next section, the corpus used is the written part of the BNC (>150 times bigger than the *Clinton-Trump Corpus*) and the task will be to explore keywords of academic writing, a frequent application of keywords approaches and word lists.

3. Another application: (key) key words in the BNCs academic writing

3.1. Methods

For this case study, the data from the BNC were explored as follows. First, a data frame containing the whole written component of the BNC was created by looping over all files and extracting every word token (converted to lower case) using the XML word annotation (the PCRE regex in *R* was "<w [^<]*?(?=</w>)"), the file name it occurs in, and the corpus part, for which David Lee's *BNC index* was used.[11] Then, once every hapax word type was discarded, the resulting data frame contained approximately 87.6m word tokens (304.5k types).

Second, the corpus was split into two parts, a target part that contained all academic writing parts (humanities_arts, medicine, nat_science, polit_law_edu, soc_science, tech_engin, approximately 16m word tokens) and a reference part containing everything else (approximately 71.6m word tokens).

[11] See http://ucrel.lancs.ac.uk/bncindex/

Third, I computed for each of those 304.5k types the frequency-based D_{KL} keyness, i.e. how much the frequency distribution of each word type in the two corpus parts differed from the percentage distributions of the corpus part sizes (0.183 vs. 0.817). In addition, I computed for each type its D_{KL} dispersions in the target corpus and the reference corpus as well as the difference between the two so that a summary plot of the type of Figure 6 could be created.

In a final step and to facilitate interpretation and analysis, I also added a new analytical step to the procedure. In a first step, I selected all word types that had a positive value on both the frequency-based D_{KL}-value and the dispersion-based D_{KL}-difference, i.e. all word types labeled as key on both dimensions of the new keyness method. Then, both dimensions were transformed to fall into a range [0, 1] in order to make them symmetric/comparable. This transformation now also means we can straightforwardly measure the distance of a word's coordinates to the origin as a 'Euclidean distance', obtaining a single value summarizing – with some information loss! – both keyness dimensions into a single sortable score. Disclaimer: I am doing this here for didactic reasons – in general, the two-dimensional tuple is of course to be preferred since it does not incur the information loss resulting from such a conflation.

3.2. Results

The results are quite interesting in a way that supports the proposed two-dimensional mode of analysis. Like Figure 6, Figure 9 shows the frequency-based D_{KL} on the *x*-axis and the dispersion-based D_{KL}-difference on the *y*-axis, but with the coordinates resulting from the [0,1] transformation of the scores, meaning that, in it, we can more felicitously make visual comparisons of the horizontal and vertical distances of words from the origin.

What do these results show? The most interesting aspect of them is how nicely they result in two kinds of keywords, depending on which of the dimensions of keyness one focuses on: the keywords listed in (8) are the top 50 keywords that have an *x*-axis value of 0.6 (an arbitrarily-chosen value), meaning they are keywords that are much more evenly dispersed in the academic target part of the BNC than in the reference corpus (though not also necessarily much more frequent in the target corpus); I think it

is relatively uncontroversial to say these are typical key keywords that are generally useful to academic writing regardless of which discipline one is in.

(8) *defined*, *similarly*, *thus*, *degree*, *factors*, *significance*, *extent*, *related*, *analysis*, *therefore*, *specific*, *characteristics*, *determining*, *importance*, *discussion*, *limitations*, *requires*, *underlying*, *define*, *differ*, *example*, *relation*, *relative*, *suggests*, *appropriate*, *derived*, *consequence*, *context*, *basis*, *forms*, *differences*, *provides*, *furthermore*, *arise*, *necessarily*, *generally*, *defining*, *distinguish*, *whereas*, *relate*, *essentially*, *interpreted*, *relatively*, *argued*, *adequate*, *identified*, *conclusions*, *moreover*, *indicates*, *subsequent*

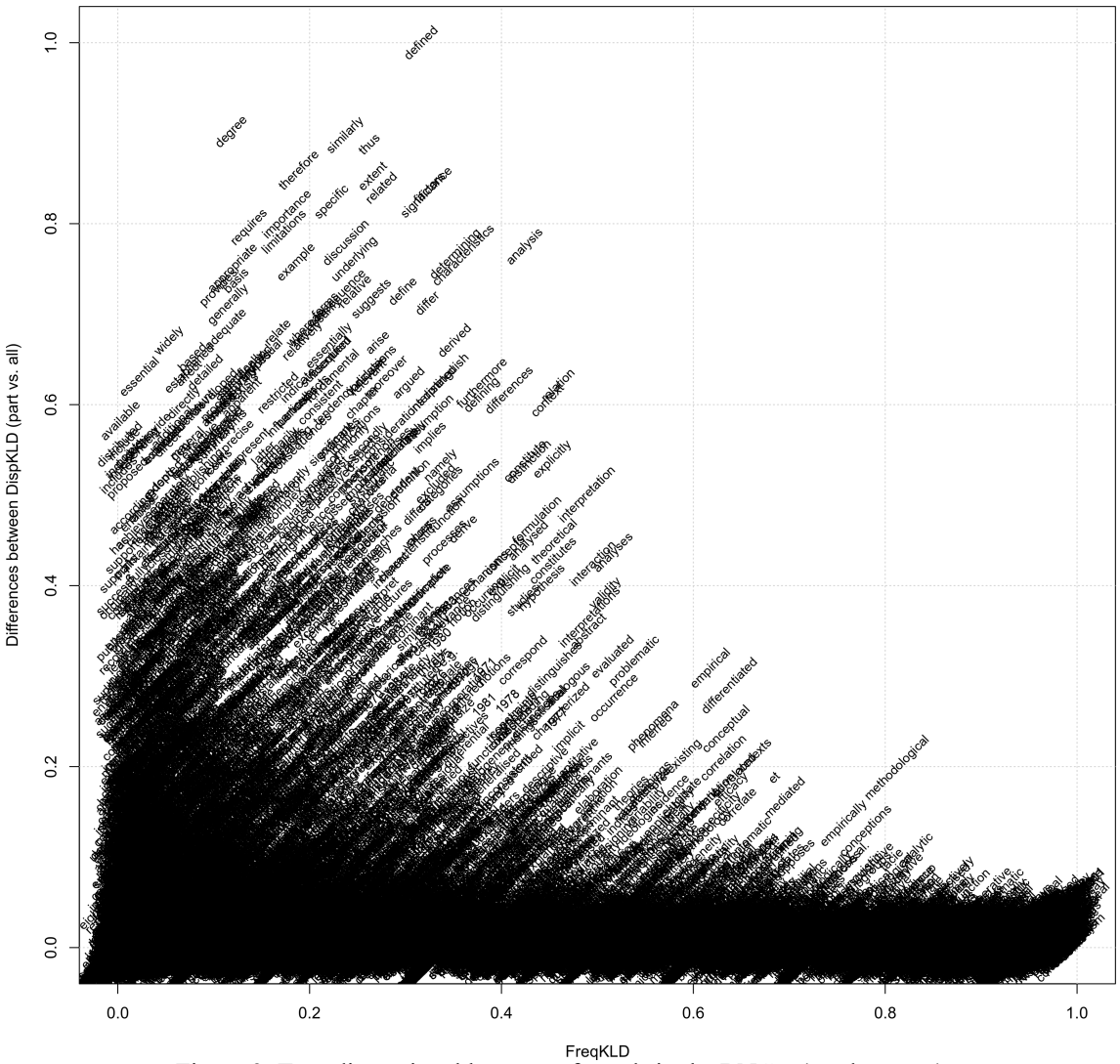

Figure 9: Two-dimensional keyness of words in the BNCw (acad vs. rest)

The keywords listed in (9), on the other hand, are the top 50 keywords that have a *y*-axis value of 0.6, meaning they are keywords that are more frequent in the academic target part of the BNC than in the reference corpus though not also much more dispersed in the target corpus.

(9) *w.l.r.*, *crohn*, *reg.*, *colorectal*, *χ*, *oesophageal*, *pylori*, *oesophagitis*, *colonic*, *labov*, *ileal*, *deixis*, *p<0.05*, *endoscopic*, *sclerosing*, *ulcerative*, *nsaid*, *ileum*, *cnut*, *antislavery*, *æthelred*, *pre-exposure*, *prednisolone*, *rugose*, *drafter*, *colitis*, *mg/kg*, *eadwine*, *p<0.001*, *mucosal*, *reflux*, *colonoscopy*, *gastrin*, *idiopathic*, *conventionalism*, *creatinine*, *antrum*, *µm*, *pou*, *amylase*, *deictic*, *thrombolytic*, *mucosa*, *gastro-oesophageal*, *tncs*, *thromboxane*, *antiracist*, *guilloche*, *carcinomas*, *guntram*

These keywords are much more specific to certain disciplines, or kinds of disciplines; clearly, many of these would not necessarily be relevant to a learner of overall academic English but to someone specializing in certain fields: learners in a field that requires them to know the words *colorectal*, *colonoscopy*, or *ulcerative* may not need to know about *labov*, *antislavery*, and *w.l.r.* (*Washington Law Review*), etc.

3.3. Interim conclusion

Again, the approach produces instructive results. In particular, it is interesting to see how the method produces different kinds of results. With a single procedure, we get both general academic words and domain-specific academic words, and the results obtained follow naturally from an approach that takes into consideration the relative frequencies as well as the dispersions of words in both the target and the reference corpus. It is then the researcher, or the applied linguist, who can choose which kind of keyword to focus on, general or specific ones or both.

4. DISCUSSION AND CONCLUDING REMARKS

4.1. Interim summary

I began with a brief review of keyword applications in general and Egbert and Biber's recent suggestion to improve keywords analyses by replacing the *LLR*-scores computed on word frequencies by *LLR*-scores computed on ranges. Given the degree to which *LLR*-scores conflate information, I first proposed to use the 'Kullback-Leibler divergence' instead and I showed that it is pleasantly less correlated with overall token frequency – something that distorts *LLR*-values considerably – but also leads to well interpretable results.

I then developed the additional proposal to explore keyness by adding dispersion information to frequency information rather than substituting dispersion for frequency

(as in Egbert and Biber 2019). For that, too, the 'Kullback-Leibler divergence' was used (in the form of a difference between the target and the reference corpus results), i.e. the same information-theoretic measure was applied to both frequency and dispersion data.

This proposal was then exemplified in two case studies, the *Clinton-Trump Corpus* and the written part of the BNC. In the former, simpler case, the results were meaningfully interpretable, and I demonstrated how words can be (key) key in different ways and in particular how *LLR* can return misleading results (especially visible in a three-dimensional plot that included token frequency). In the latter case, the results were again instructive and particularly interesting for how the proposed method returns both general academic words as well as domain-specific words in different quadrants of the results plots. Just about all of the above could be applied without many arbitrary choices: no stop list was needed, no frequency threshold other than hapaxes was used, without arbitrary range threshold (of, say, 5%, 10%, or 30% of the texts) was applied (and none of those would even take corpus part/file sizes into consideration in the first place), and there was no elimination procedure in place one would need to justify in some way (such as eliminating the 2,000 most frequent English words, as in Coxhead's (2000) *Academic Word List*).

4.2. Where to go from here

I can begin only by echoing Egbert and Biber's (2019: 102) conclusions:

> It is [my] hope that this study will raise awareness of the importance of text dispersion in corpus linguistics and discourse analysis. More importantly, [I] hope to see a trend in these fields in the direction of using the text – rather than the corpus – as the primary unit of analysis.

It is precisely studies like theirs that the field needs more of in order to develop a better understanding of what current methods do and do not do and, building on that, to develop more comprehensive methods. There is much talk in papers and conferences about how complicated the distributional data offered by corpora are (in terms of their diversity, their 'Zipfianness', often their ambiguities, etc.) but all too often researchers uncritically fall back on the same methods or statistics that are offered in some software and Egbert and Biber did well to push the envelope. Accordingly, it is my hope here that the proposed 'tupleization' – the idea to not conflate dimensions of information but

consider them separately and jointly, here developed for keyness, in Gries (2019b) for association measures – will also move the field along and offer us a better understanding of keywords in general and its application in discourse analysis, text type/genre/register studies, and educational applications. That being said, of course the proposed method here can also still be improved. The most pressing improvement that keyness approaches need is better input: ideally, we would not just apply our keyness computations to the individual words resulting from some sort of tokenization, but to the combination of individual words and multi-word units as defined by some, ideally, bottom-up algorithm, which would boost especially educational applications considerably: why not have a bottom-up algorithm find that statistically significant behaves like a word and then compute its keyness? The combination of something like this together with the above two- or three- dimensional approach to keyness should help us understand and use the richness of our data much more.

REFERENCES

Altman, Douglas G. and Patrick Royston. 2006. The cost of dichotomising continuous variables. *BMJ* 332(7549). 1080.

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

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.

Brown, David. 2016. *Clinton-Trump Corpus*. http://www.thegrammarlab.com/?nor-portfolio=corpus-of-presidential-speeches-cops-and-a-clintontrump-corpus

Burch, Brent, Jesse Egbert and Douglas Biber. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3/2: 189–216.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34/2: 213–238.

Cumberland, Phillippa M, Gabriela Czanner, Catey Bunce, Caroline J Doré, Nick Freemantle and Marta García-Fiñana. 2014. Ophthalmic statistics note: The perils of dichotomising continuous variables. *British Journal of Ophthalmology* 98/6: 841–843.

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

Egbert, Jesse and Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14/1: 77–104.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1/2: 277–294.

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

Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Th. Gries, Stefanie Wulff and Mark Davies eds. *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.

Gries, Stefan Th. 2016. *Quantitative Corpus Linguistics with R*. New York: Routledge.

Gries, Stefan Th. 2018. *Towards a Unified Tupleization of Corpus Linguistics*. Invited plenary talk at the *56th Annual Meeting of the Association for Computational Linguistics*. Georgia State University.

Gries, Stefan Th. 2019a. *Ten Lectures on Corpus-linguistic Approaches: Applications for Usage-based and Psycholinguistic Research*. Leiden: Brill.

Gries, Stefan Th. 2019b. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24/3: 385–412.

Gries, Stefan Th. 2021. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries eds. *Practical Handbook of Corpus Linguistics*. Berlin: Springer.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1/2: 263–275.

Lijffijt, Jefrey and Stefan Th. Gries. 2012. Correction to "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics* 17/1: 147–149.

R Core Team. 2020. *R* a language and environment for statistical computing. Vienna: *R* Foundation for Statistical Computing. http://www.R-project.org.

Scott, Mike 1997. PC analysis of key words – and key words. *System* 25/2: 233–245.

Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Tribble, Christopher. 2002. Small corpora and teaching writing: Towards a corpus-informed pedagogy of writing. In Mohsen Ghadessy, Alex Henry and Robert L. Roseberry eds. *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins, 381–408.

Xiao, Zhonghua and Anthony McEnery. 2005. Two approaches to genre analysis: Three genres in Modern American English. *Journal of English Linguistics* 33/1: 62–82.

Corresponding author
Stefan Th. Gries
University of California, Santa Barbara
Department of Linguistics
Santa Barbara
CA 93106-3100
United States
Email: stgries@linguistics.ucsb.edu

received: February 2020
accepted: June 2020