How can we communicate (visually) what we (usually) mean by collocation and keyness? A visual response to Gries (2022a)

Stephen Jeaco

Xi'an Jiaotong-Liverpool University

The published Version of Record is:
Jeaco, Stephen. 2023. How can we communicate (visually) what we (usually) mean by collocation and keyness?: A visual response to Gries (2022a). *Journal of Second Language Studies* 6:1 pp. 29–60.
https://doi.org/10.1075/jsls.22019.jea

Abstract

Corpus linguistic methods can now be easily employed in a wide range of studies within sub-disciplines of linguistics and well beyond. In a two-part paper, Gries (2022a, 2022b) challenges some of the most widely used 'association measures' of what many might feel to be powerful aspects of text patterning: collocation and key words. While the additional association measure offers some new possibilities, this paper highlights the strong influence of another frequency parameter on odds ratio and Gries's suggested association measure, and questions the applicability of his cautions for many different kinds of corpus research. Nevertheless, having been inspired to look at different aspects of association and dispersion more carefully, the author presents some new visualizations which were designed to communicate some of the important lessons to be learned from Gries's papers, especially for learners and teachers using corpus tools in Second Language classrooms.

Keywords: association, frequency, collocation, keyness, dispersion, log-likelihood, range, Gries's DP

1. Introduction

Beyond the convenience of access to multitudinous examples of sampled language data, corpus linguistic methods might be admired for their almost miraculous means of showing relationships between words and their neighbours (collocation), between words and their texts (keyness) and between words and the distances of other instances of the same words (dispersion). With more than fifty years of developing, refining and re-calibrating statistical methods to help focus on *important* patterns of language in texts, a wide variety of measures and terminology have developed. Some of the terminological labels edge into the metaphorical; some nuances and distinctions may seem hard to justify, especially for those from a slightly different school or camp. In many ways, corpus linguistics was one of the first disciplines to embrace the age of big data; we had collocations of millions of words to analyze long before we had millions of items in archived shopping carts or of clicks and taps in browsing histories. In two joint papers, Gries (2022a, 2022b) makes claims about some of the most commonly used metrics for association (to measure collocation or key words) and dispersion. The main focus of this paper is a response to Gries (2022a), focusing on collocation and key words as a group of association measures. However, since Gries (2022b) further clarifies the other paper, and since his contributions on dispersion also inspired some of this response, the second paper will also be referred to. While the call for greater clarity about which inputs are (mainly) contributing to various measures is helpful, and while it should be acknowledged that the new proposal for measuring another aspect of association offers promising new avenues for research, this paper argues that a measure's suitability needs to be judged according to the perspective of the research and that some of the stronger concerns of Gries's papers might be better addressed through the development of visualizations of data, rather than complete abandonment of log-likelihood, T-score, MI3 or other popular metrics for collocation and key words.

This paper begins with a brief overview of log-likelihood's sensitivities and the hypothetical contingency tables and keyness examples from Gries (2022a). Within the complex and diverse contexts of different kinds of potential end-users of these measures and the different ways the kinds of corpora used might relate to their ultimate foci of study, it goes on to consider different interpretations of the core terminology surrounding association, and the role of frequency within definitions of collocation and keyness. After that, it takes each of the applications of measures which were partially explored in Gries (2022a), asking how free from frequency the proposed replacements are, and to what extent his descriptions provide opportunities for new (human) insights and new (machine learning) inputs. Having drawn out important principles which were inspired by (but sometimes in conflict with) his claims, examples of new data visualizations are presented and described. These developments are within the context of a corpus tool which was (primarily) developed for English language learning and teaching, but it is hoped that the message they have been designed to convey could be helpful for a wide range of corpus research contexts. Specifically, the paper introduces a scatter chart superimposed on the corpus Zipfian curve to indicate the scope of top-ranked collocations; twin double-logged charts to show the change from expected to actual frequency for key words; and a cone-shaped chart with a gradient colour-filling to represent three aspects of dispersion: diffusion, overall density and the uniformity of that density. Overall, the paper argues that log-likelihood, T-score, and MI3 are still fit for purpose for keyness and collocation studies, and that different dispersion metrics and related methods can help to show different aspects at different levels of study: the paragraph, the chapter, the text, sub-collections of text, and more.

As a response paper, there is less of a need for an extensive literature review, but when looking at the usefulness of measures of association and dispersion, a basic foundation is an understanding of the nature of the widely diverse range of raw word frequencies and the typical relationship between increases in sample size and increases in types. By plotting the logged frequency of a word against its logged rank, Zipf (1935) demonstrated that an almost straight diagonal line will be produced; while the top-ranking types for frequency are degrees of magnitude more frequent than most other types in a typical corpus, there will also be a long tail of items which occur just once or twice (see Croft et al., 2010; Oakes, 1998). As Croft et al. (2010) note, an extension of this relates to expected vocabulary sizes (the estimated number of types) in relation to the total size of the corpus, which can be estimated using Heaps' Law (1978), shown in equation 1.

Equation 1: Heaps' Law (1978), cited in Croft et al. (2010)

$$v = k \times n^{\beta}$$

$$10 \leq k \leq 100, \qquad \beta \approx 0.5$$

When comparing the strengths of association between different nodes and their candidate collocates or between two candidate key words, Heaps' Law can be helpful as a reminder of the way that many more new one-off rare events are likely to be encountered with increases in factors such as the number of events each candidate participates in (for both collocation and keyness), the size of the subset of in-window opportunities (for collocation), and the size of the reference corpus (for keyness). The level of exclusivity of a collocate or a key word may need to be contextualized by more than the item's frequency, and studies comparing collocations from different sized corpora or comparing key words from different sized pairs of corpora will feel the effects of Heaps' Law as more rare events seem to creep into the larger of the samples (c.f. the hapaxlegomena and double occurrences in Evert & Krenn 2001). These issues will be explored with reference to Section 2.1 of Gries (2022a) before re-examining the claims of Sections 2.2 and 4.2.

Gries (2022a, 2022b) questions the role of frequency for more targeted measures of association and dispersion and provides new metrics specifically designed to exclude the influence of frequency. This allows, he claims, for frequency and association to be treated as separate variables, and can then be combined to allow high-low, low-high and low-low rather than just high-high. The frequency referred to here is the frequency of co-occurrence of node and potential collocate, or the frequency of the candidate key word in the study corpus. Nevertheless, it was undoubtedly the problem of dealing with massive differences in raw frequency which led to the development of most collocation, keyness and dispersion metrics. Applications of corpus linguistics for language teaching have tended to focus on identifying the most frequent but less predictable collocation patterning, as from the vast amount of what could be taught, class time or study exercises need to focus on what is likely to be useful (Sinclair, 2004). Within corpus linguistics more broadly, concordance line analysis often focusses on the “central and typical”, which takes into account the most frequent patterns of the item (the typical) and also of the category (the central) (Hunston, 2002:42). Most statistics for measuring collocational strength take several frequency-related inputs: the frequency of a word with a potential collocate (co-occurrence frequency), the frequency of the collocate occurring in windows which do not contain the word of interest (remaining frequency of collocate), the frequency of the word where the collocate does not occur (remaining frequency of node), and the total sum frequencies of remaining items in the corpus (remaining frequencies of others). Gries (2022a) provides a schematic co-occurrence table where these parameters are conventionally labeled a, b, c and d respectively. As Dunning (1993:62) pointed out, statistical tests on these parameters need to consider the non-normal distribution of the frequencies of types in a corpus because otherwise, “When comparing the rates of occurrence of rare events, the assumptions on which these tests are based break down because texts are composed largely of such rare events.” Key words are also based on differences between frequencies in one corpus or context in contrast with another. Scott (1997:236) explains, “unusual frequency in a given text… does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind.” The same a, b, c and d labels can conventionally be used for key word contingency tables to represent four frequency parameters: the frequency of the candidate key word in the study corpus (co-occurrence of the node and the study corpus sample), the frequency of the candidate key word in the reference corpus, the sum frequencies of other items in the study corpus and the sum frequencies of other items in the reference corpus. However, with key word calculations, it is more likely that some candidates will actually have a stronger tendency to occur in the reference corpus than they do in the study corpus; in *WordSmith Tools* (Scott, 2020), for example, key word candidates where a/(a+c) is less than b/(b+d) are given a negative sign. The usual way log-likelihood contingency tables are formed for collocation means that negatively charged candidate collocates would not usually be found because it is hard to imagine a sample of language where the frequency of an item of interest can compete in this way with the frequencies of all other items.

In order to explore some of the claims Gries (2022a) makes against the suitability of including frequency information as part and parcel of association and against some of the sensitivities of log-likelihood in particular, it is necessary to briefly consider the relationship between the corpus being analyzed and the focus of a specific research project. Figure 1 shows a simple matrix of two dimensions – the completeness of the corpus record and the match between target language use and the texts of the corpus – which can be used to distinguish between many different kinds of corpus research, and consequently the appropriateness of different claims.

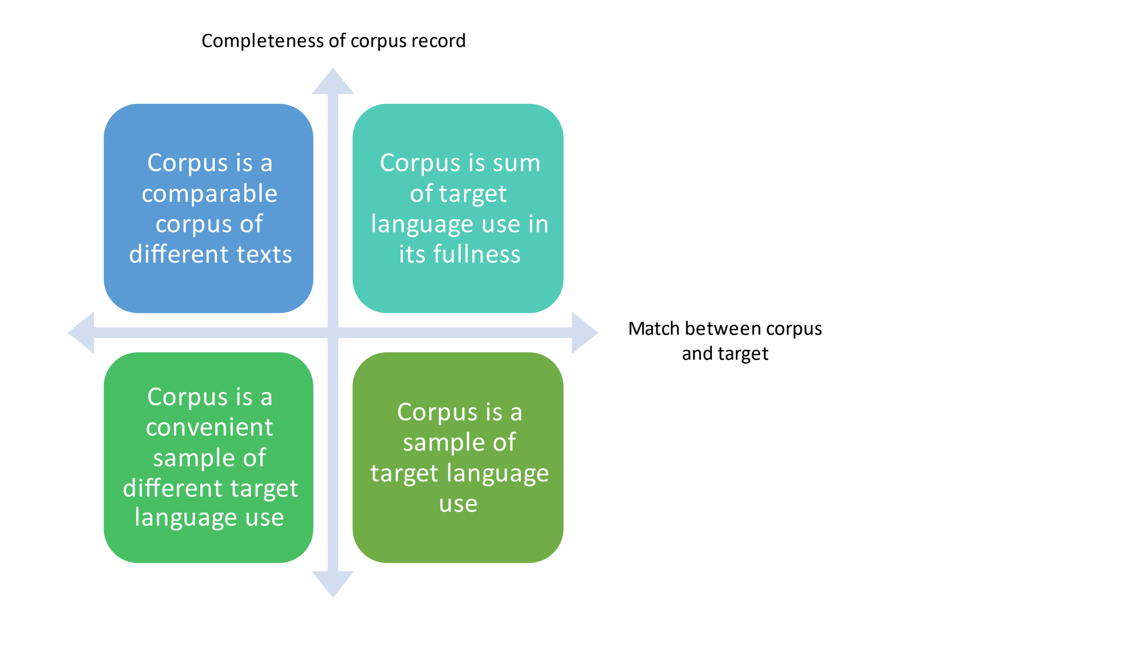


Figure 1: Relations between corpora and research foci

The top-right quadrant is probably the most ideal situation, with the corpus being analyzed actually containing the target language use in its entirety. This may seem quite remote and unachievable for most corpus work, but this would include, for example, corpus stylistics studies of the entire oeuvre of a specific author (Mahlberg, 2013). Moving down the y-axis to the lower-right quadrant, a corpus might be comprised of some but not all of the actual texts of the target language use. Conversely, moving up the y-axis could also entail taking fuller samples from the texts used for the corpus, so sample size might refer to both the number of potential texts included, but also the amount of each text included in the corpus if complete texts have not been used. The key point that distinguishes the sampling of the lower-right quadrant from that of the lower-left quadrant is that moving along the x-axis to the right entails having texts which have actually been produced by the language users within the genre and register under study. The application described by Gries (2022b) where frequencies from a reference corpus of general language use are used to train a machine learning algorithm (random forests) in order to generate a model which can predict language elicitation results from second language users would be located in the lower-left corner. The top-left quadrant is perhaps where the majority of corpus research would be located; to understand the patterns of language use for a group of interest, a corpus of similar, comparable texts has been constructed or selected. Clearly, this visualization is a simplification and summarizes two dimensions of corpus design which have been well established in the literature (e.g. Sinclair, 1991, 2005; Hunston, 2002; O'Keeffe, A. et al., 2007). Unless the research is firmly in the top-right quadrant, even for cases where a corpus can be said to closely represent a target population, "since no word count is ever based on a truly random sample of the target population the sample frequencies are generally overestimates of the true population frequencies, except that the highest-frequency items are underestimated …" (Oakes, 1998:191 citing Hann 1973). As will be discussed later, whether for collocation or for keyness, either the measure or the analyst needs to take into account the vast differences in raw frequency between what tend to be the mundane, grammatical and uninteresting words and items which seem to have just *slipped* into the sample yet are hardly likely to be of great interest because they have such low frequencies.

The first collocation/collostruction example from Section 2.1.1 of Gries (2022a) is used to challenge the fact that log-likelihood has a score which increases as the amount of evidence increases despite the fact that the candidate may be occurring at the same rate. In the second paper, he refers back to this example with this summary:

> "However, the problem with G2 is that it increases quite a bit when all frequencies of table.01.obs increase even though the ratios of the values in the

table do not change (which of course entails that the actual association between w and c is no different from before)" (2022b:6).

Returning to the matrix in Figure 1, the increase in the amount of evidence would constitute a movement up the y-axis, with an increasing coverage of the corpus data in terms of completeness. Gries (2022a) marks this clearly as a hypothetical example, but even hypothetically it is important to consider from where the new data containing the increased frequencies could have come. Within Gries's narrow definition of measuring association as "quantifying contingency" (2022a:4), it is reasonable to claim that the predictability of the collocate from the node remains the same and therefore the 'association' is the same. However, definitions of collocation may not be limited to contingency, and given the increase in the total amount of evidence (data) available, rather than entailing that the actual association is no different from before, in many kinds of corpus research, an increase in all the frequencies of the contingency table constitutes an increase in the confidence that the association measured in the smaller sample is representative; in other words, that claims about the association are firmer. When comparing two estimates based on Heaps' Law in terms of the number of times larger the number of types in a larger corpus would be, the constant k will cancel out, meaning that proportions are merely n to the power β. Based on Heaps' Law and following Croft et al. (2010) with β=0.5, if the sample size increases by factor of 10 (as in Gries's table.01.obs and table.02.obs), one might estimate a threefold increase in the number of types in the corpus ($10{,}000{,}000^{1/2}$ / $100{,}000{,}000^{1/2} \approx 31.62\%$). Remarkably, in this additional 90% of previously unchartered data, the proportion of co-occurrences in relation to the overall frequencies of node and collocate have remained exactly the same. Another way of estimating the number of different collocate candidates might be to assume smaller samples of possible 4 word span window slots. Even though Heaps' Law is less reliable on smaller corpora, it can still be used for demonstration purposes in our hypothetical situation. Again, this equates to an estimated threefold increase in the number of types in these windows. And yet, once again, remarkably in the hypothetical example the proportions have remained exactly the same. Given what we know about the way rare events constantly creep into new data, it seems reasonable to have a higher association score (in the broader definition) as a pointer towards the likely associations in the minds of language users given the additional evidence from the larger corpus. What Gries calls "a reaction to frequency" (2022a:7) could be seen as a sensitivity to frequency. From the perspective of the word, the proportion has not changed; but from the perspective of the corpus, this pair of words have endured the constant onslaught of additional new types vying to invade their relationship opportunities, and the new corpus data provide additional evidence of a strength of this relationship because it has prevailed against new words, unique names, spelling errors, typos, and the fuzziness of text as representative samples of naturally occurring language. Similarly, when distribution c / not c remains the same, but the overall frequency of the word of interest doubles (as in Gries's table.01.obs and table.03.obs), the association score (in its wider definition) as a pointer towards the likely associations in the minds of language users who created these texts, given the change in prominence of this specific word relative to the others does indicate a stronger rebuttal of expected alternative word combinations, given the increased estimates of types. This 'reaction' or sensitivity to distribution is another feature of the way log-likelihood works.

Furthermore, the importance of fully understanding the relationship between the corpus being analyzed and the focus of the research can be exemplified by re-examining another of the examples in Gries (2022a) – the distribution of *about* in two corpora of US presidential campaign speeches (Section 2.1.2). Gries asks the reader to only look at the top of a continency table and poses the question: "I have a sentence here that contains *about*, whose speech is that sentence from?" (2022a:9). He claims that the reader would guess Trump, but that the reader would be incorrect because the top of the contingency table only gives raw frequency information and knowledge of the overall corpus sizes is needed before the correct answer can be determined. However, on the one hand, based on his data, the correct answer is Trump; you are more than twice as likely to be right because even though Clinton used the word in presidential speeches in 2016 at a higher *rate of frequency* than Trump, the Trump corpus is about 4 times larger than the Clinton corpus and the actual number of instances of *about* is higher in the Trump corpus! Even though based on the log-likelihood keyness results *about* stands out for Clinton, it would be foolish of the reader to make a prediction based on raw frequency without any attention to the difference in size. The log-likelihood formula is sensitive to the difference in size, so judging the appropriateness of an emphasis on frequency by looking at the top of the contingency table alone is unfair. Looking back at Figure 1, the question could sit comfortably in either the top-right quadrant ("I have a sentence here from the 2016 election campaign speeches that contains *about*, whose speech is that sentence from?") or the lower-right quadrant ("I have just recorded a new speech from one of the candidates to add to what we know about the language use of these two candidates. I have a sentence here from that new speech which contains *about*, whose new speech is that sentence from?"). Based on the normalized frequency or the positive log-likelihood keyness result, guessing Clinton would be right if the reader is being asked to use the corpus to predict frequencies extending beyond the corpus.

Thus far, this paper concurs with Gries (2022a) on the sensitivities of log-likelihood as a measure for collocation or keyness as demonstrated in Section 2.1, with only a minor quibble about the need for clarification on the scope of the predictions in 2.1.2 in terms of it being within sample or beyond sample. Here it is also worth noting additional sensitivities of log-likelihood in relation to changes in the balance and magnitude of its parameters.

Equation 2: Log-likelihood formula, using matched a, b, c, d parameters

$$E1 = (a + c) \times \frac{a + b}{a + b + c + d}$$

$$E2 = (b + d) \times \frac{a + b}{a + b + c + d}$$

$$G^2 = 2 \times \left(\left(a \times \ln \frac{a}{E1} \right) + \left(\left(b \times \ln \frac{b}{E2} \right) \right) \right)$$

Formula from Read and Cressie (1988), cited in Rayson & Garside (2000), but adjusted here to ensure parameters c and d match the contingency table from Gries (2022a), and using *ln* to represent the natural logarithm.

As shown in equation 2 (where parameters c and d have been adjusted to match the labels from Gries 2022a Table 1), log-likelihood is also sensitive to a+b / a+b+c+d (Jeaco, 2020[1]). Furthermore, as noted earlier, if negative signs are given to key word candidates which have a higher proportion of occurrences in the reference corpus, for key word analysis the final log-likelihood score is also sensitive to underrepresentation in the study corpus. On the one hand, it is reasonable to take Gries's position and assert that these sensitivities can be a drawback, especially if the separate aspects such as co-occurrence frequency can be incorporated into the researcher's analytical procedure. After all, Gries is not claiming that frequency (separate from 'association') is unimportant. On the other hand, it could also be argued that when a researcher is taking a complete text or a collection of texts as the unit of study, the log-likelihood metric provides a useful way to move away from effects of the Zipfian frequency curve and the tendency for ever increasing rare events in larger samples, being duly sensitive to the frequency in the reference corpus, the frequency of the item overall, and the sizes of the study and reference corpora separately and combined.

2. Definitions

Having provided notes on the backdrop for research using association measures, given that Gries (2022a) demonstrates a strong influence of frequency on commonly used measures, it is important to consider definitions of collocation and keyness. The literature on collocation and keyness may use the term *association*, particularly when exploring the potential of various metrics as a means of using corpus texts (language produced in the output of language users) in order to predict connections between words and contexts which are likely to exist in the minds of language users. However, in mainstream corpus tools columns showing statistical measures of collocation and keyness are usually headed using the name of the measure, as in *WordSmith Tools* (Scott, 2020) and *CQPWeb* (Hardie, 2012), or a general label such as *Score*, as in *Sketch Engine* (Kilgarriff et al., 2004), *likelihood and effect* as in *Antconc* (Anthony, 2022), *or*

Stat as in *Lancsbox* (Brezina et al., 2015). The term *association*, then, is one that researchers might use (perhaps too loosely) when describing the results for a specific project using a specific metric. Definitions of collocation (and to a lesser extent keyness) may be closely related to the overall orientation of a research project. For example, the procedural aspect of a definition of collocation may differ according to the size or nature of the context window (cf. Hoey, 2005, 2014) or according to the absence of presence of grammatical restrictions (cf. Sinclair, 1991; Wermter & Hahn, 2006). Definitions of collocation for Second Language teaching and research in particular are often tuned to the context, measures, and the linguistic sophistication of the target users of the results. Hoey's definition of collocation makes specific reference to the frequency of the combination, and also makes a tentatively worded link between corpus data and psychological *association*:

> "So our definition of collocation is that it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora **more often** than is explicable in terms of random distribution"
>
> (Hoey 2005:5, emphasis added)

In English language textbooks and dictionaries, frequency also has a prominent place in definitions:

> "Learn new words in combination with other words that **often go with them**" (McCarthy, McCarten, & Sandiford, 2006a, p. vii, emphasis added).

> "words that are **often used** with a particular word" (Longman Dictionary of Contemporary English, 2009, p. xiii, emphasis added);

> "… **high-frequency word patterns**" (Collins COBUILD Advanced Dictionary of English, 2009, p. viii, emphasis added)

While textbooks and dictionaries are less likely to refer to keyness, the definition of keyness from the WordSmith Tools Manual is related to a candidate's frequency being "unusually frequent" (Scott, 2022). In order to link corpus frequencies to likely associations in the minds of readers of the texts of a corpus, the keyness definition from Scott and Tribble is at least conceptually related to the frequency of each candidate:

> "… keyness is a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, **avoiding trivia and insignificant detail**. What the text "boils down to" is its keyness, once we have steamed off the verbiage, the adornment, the blah blah blah." (Scott & Tribble, 2006: 55-56)

For these kinds of definitions, the claims about *association* from Gries (2022a, 2022b) are actually less akin to association in the minds of language users, evidenced by frequent patternings in corpus data, and more about exclusivity (promoting collocate pairs not found with other items except when one of the pair is more frequent than the other) or nicheness (promoting key words where the candidate occurs exclusively in the study corpus) or possibly even neediness (as the papers include anthropomorphisms to dramatize where words choose to occur). Many of the definitions of terms used to refer to linguistic phenomena in (corpus) linguistics are perhaps vague or visionary or optimistic or cyclical. However, it seems reasonable that referring to an association between items may include the notion that this association is to be measured given the size of the sample(s) and the frequencies of the words elsewhere. Gries (2022a) preempts some of this counter-argument stating, “Of course, some scholars might now retreat to the position that they simply have a definition of association that is different from mine…” (19). Measures of *association* for collocation are perhaps best captured in terms of weighing up the effects of the intensity of *stimulation of a collocate* (a/(a+b)) and the importance of the *relationship* for the node (a/(a+c)). Similarly, for keyness, it would be the intensity of *stimulation* of an item to be drawn to the study corpus (a/b) and the importance of the item, given the size of the study corpus (a/(a+c)). Measures of both collocation and keyness need to be sensitive to the amount of evidence for either item in relation to the corpus as a whole (a+b) and the amount of all evidence (a+b+c+d). If these aspects are not part of one’s definition of ‘association’ these aspects will often need to be considered before the results can be applied. If frequency were the only driver of log-likelihood collocations or key words, sorting the obtained shortlist of candidates by descending frequency rather than log-likelihood would provide results which were as intuitively satisfying, yet this kind of post calculation column re-sorting in this author’s experience tends to reduce the perceived usefulness of the ranked results drastically, and Evert and Krenn’s (2001) evaluation (cited in Gries, 2022a) compared raw frequency with the association measures and found log-likelihood outperformed frequency for Adj+N. The next section will consider to what extent Gries’s proposed new measure of association avoids information loss and avoids extreme correlation with other input parameters.

3. Re-examining the roles of inputs for collocation measures

In Gries (2022a), in Figures 1 and 10, the logged co-occurrence frequency – or log(a) – is plotted against the log-likelihood, the log odds ratio and the Association-Without-Frequency scores for collocations of *fast*+N in the British National Corpus (BNC, 2007). With R^2_{GAM} of 0.9465, 0.0241 and 0.0066 respectively, Gries demonstrates clearly the strong influence of co-occurrence (a) on log-likelihood. Given the low correlation between odds ratio and co-occurrence, and between Association-Without-Frequency and co-occurrence, the reader is invited to assume that these measures are not influenced by co-occurrence and therefore are not influenced by the separate notion of frequency and are therefore more suitable as measures of association only. However, using the same corpus and the same collocation, it is possible to demonstrate the strong inverse relationship between these two measures and the b parameter – the frequency of the candidate collocates when they do not occur immediately after *fast.* Table 1 shows the R^2_{GAM} values for the four *fast* adjectives, demonstrating a strong, but inconsistent relationship between both log odds and Association-Without-Frequency and the non-co-occurrence of the candidate collocates (where they do not occur immediately after each of the four adjectives).

Table 1: R^2_{GAM} for logged frequency of other occurrences of candidate collocates for four adjectives

	Log odds ratio	Association-Without-Frequency
fast	0.818	0.908
quick	0.791	0.615
rapid	0.772	0.979
swift	0.89	0.997

Figure 2 and Figure 3 show the combined results for all four adjectives, with strong negative correlations for both odds ratio (R^2_{GAM} 0.752) and Association-Without-Frequency (R^2_{GAM} 0.932). Plots and R^2_{GAM} values were generated using RStudio (2022) and the generalized additive modelling package *mcgv* (Wood, 2011).

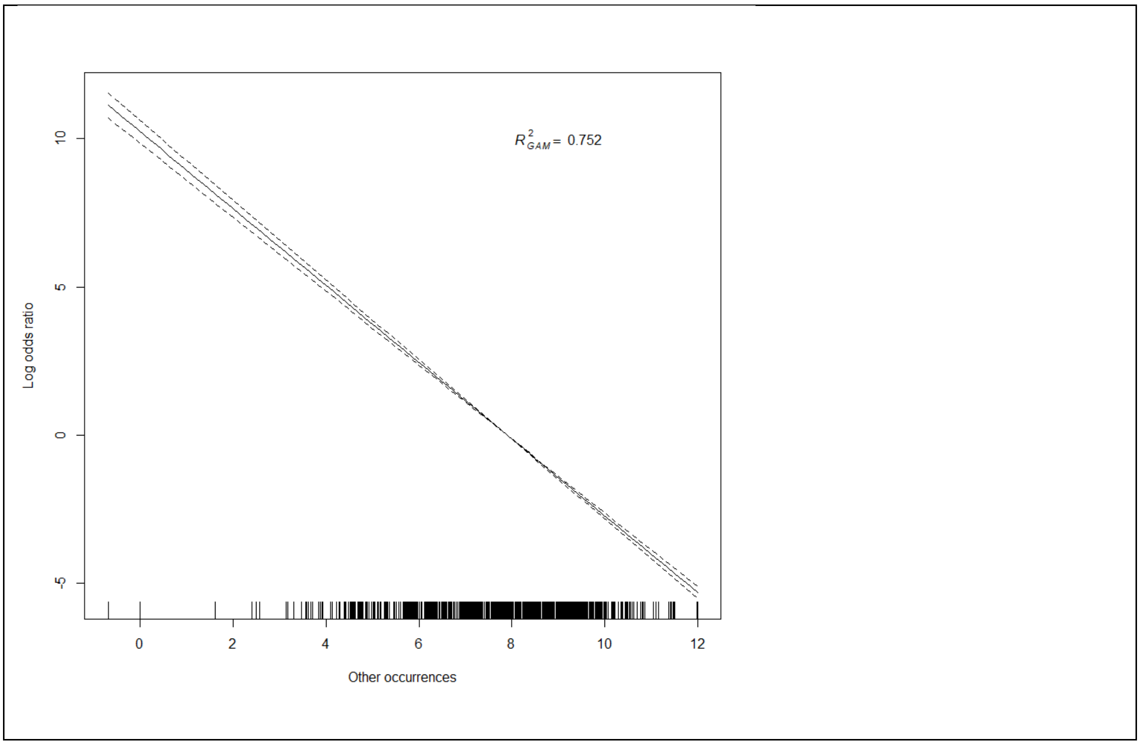


Figure 2: Plot of logged non-co-occurrence of the collocate against the Log-odds ratio scores for four adjectives + N in the BNC

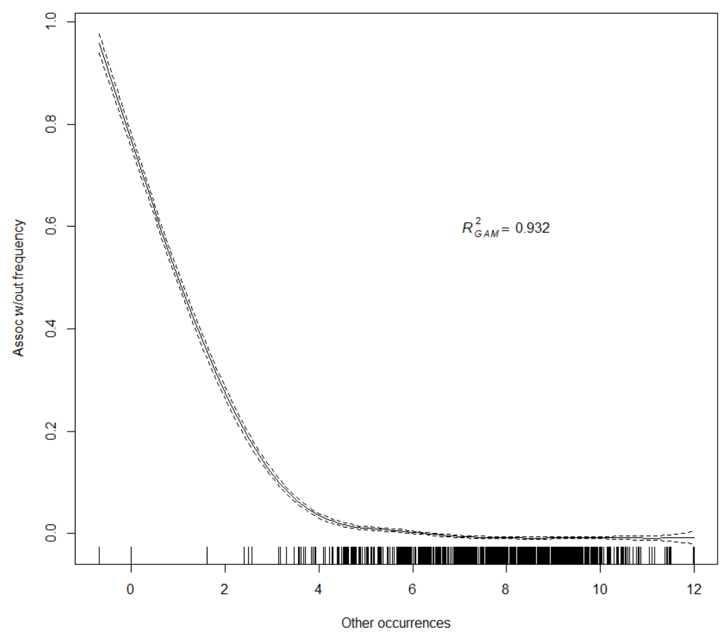


Figure 3: Plot of logged non-co-occurrence of the collocate against the Association Without Frequency scores for four adjectives + N in the BNC

These results raise the question of why Association-Without-Frequency is even more influenced by reference frequency than Odds Ratio. The idea of "relativizing the observed value against the theoretically possible range" (Gries 2022a:27) seems to provide a uniformity to the values and offers easier comparisons compared with other measures (c.f. logDice's theoretical maximum of 14 (Rychlý, 2008)). However, the apparent uniformity hides the fact that the space is sparsely populated. Figure 4 shows a histogram of the values for Association-Without-Frequency for the four adjectives, grouped in bins (and offers an alternative presentation of the tick-marks above the x-axis of Figure 3, which also shows the density of very low scoring Association-Without-Frequency scores).

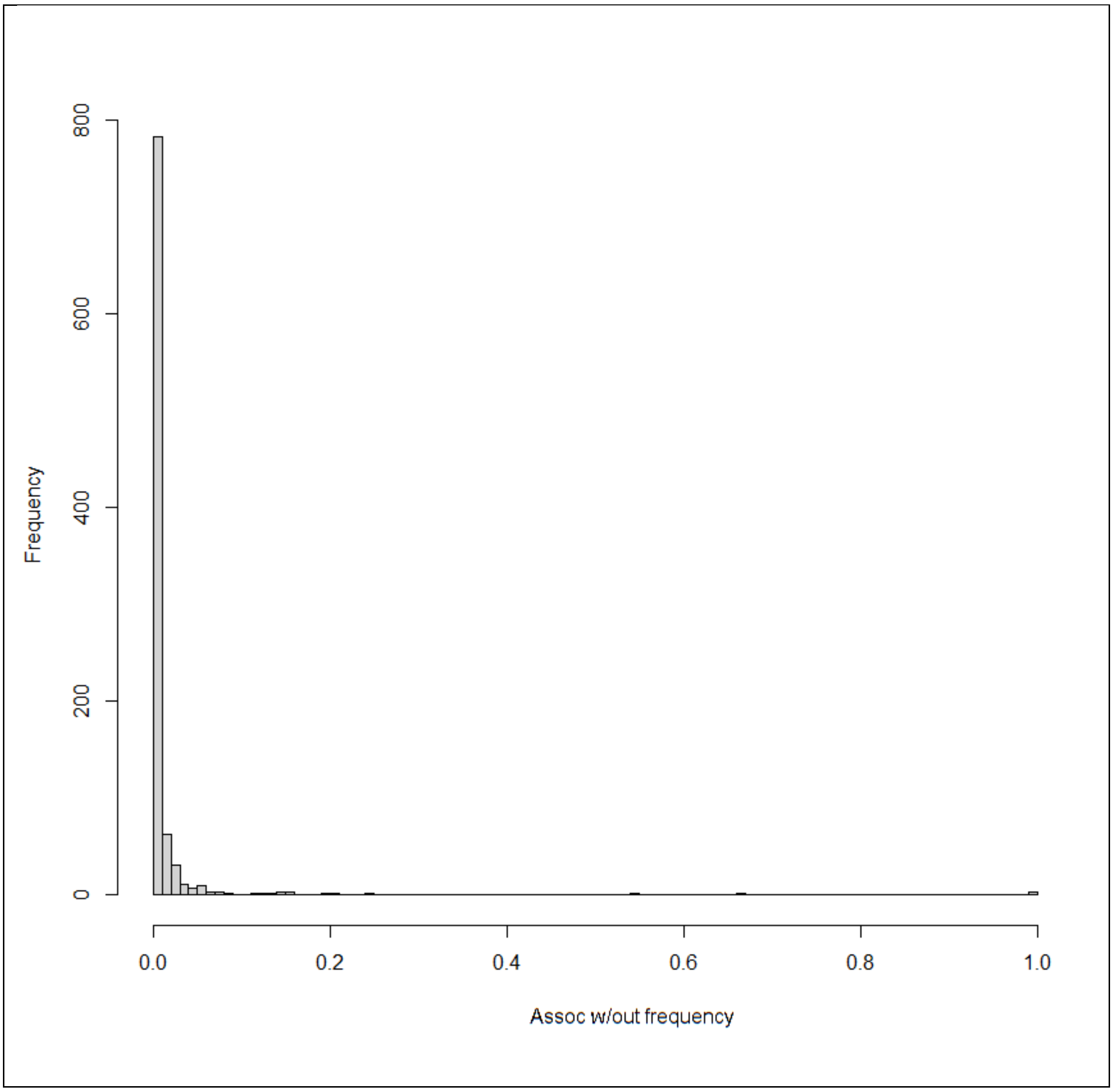
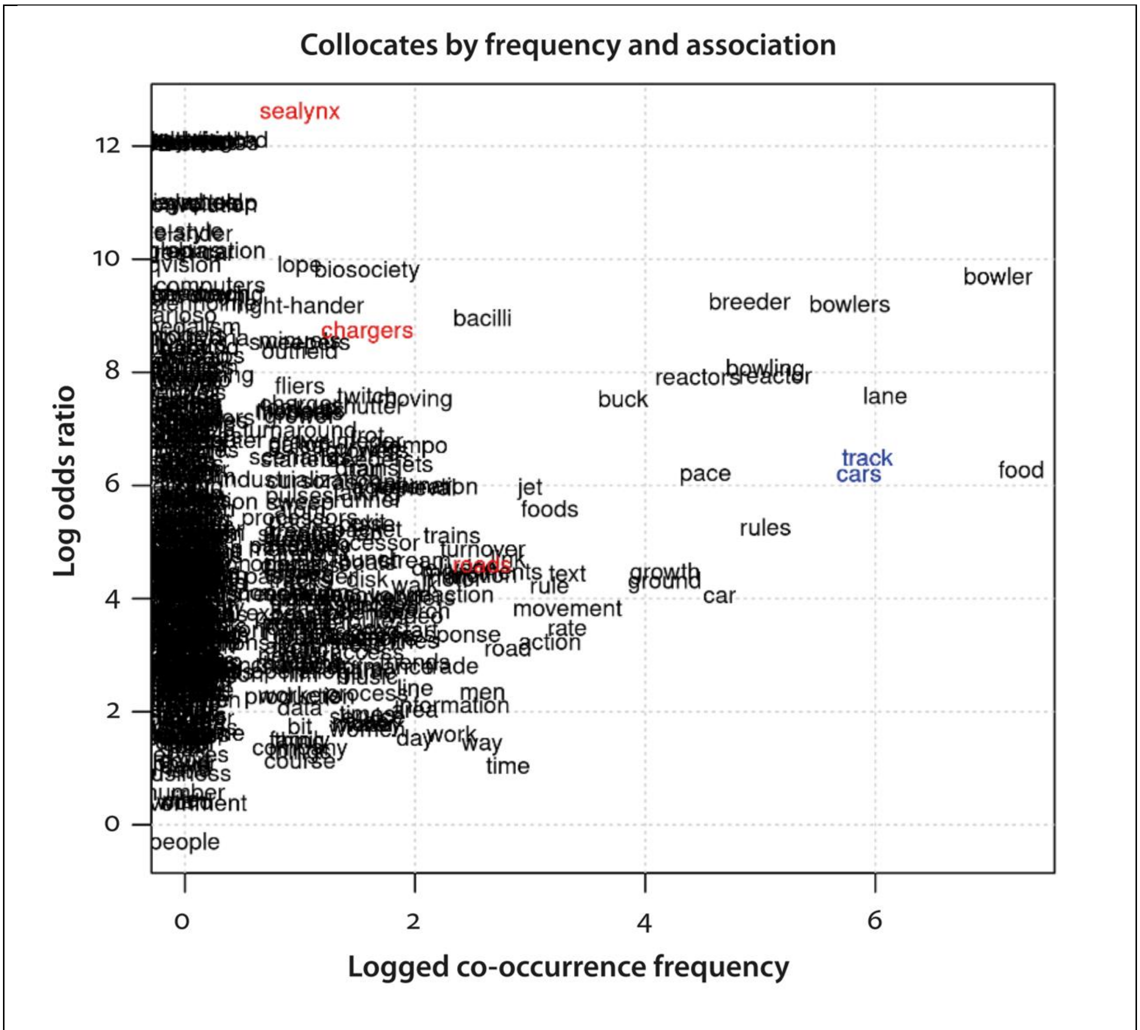


Figure 4: Binned Association-Without-Frequency scores for the four adjectives

As can be seen, the vast majority of the 925 candidates have scores between 0 and 0.125, while the remainder of the range has scarcely anything, excepting the very small number with a score of exactly 1. Association-Without-Frequency relies heavily on lack of counter-evidence for a candidate occurring in other contexts, and because of the relativizing stage in its calculation, collocate candidates with overall frequencies less than that of the node are favoured. It could be argued that the way to increase the strength of this kind of association is to remove from the corpus other items which occur with the collocate, and this is rather counter-intuitive to a notion of association.

In Figure 3 of his paper, Gries demonstrates that plotting logged frequency on the x-axis and logged odds ratio on the y-axis, it is possible for an analyst to explore these two factors separately. Figure 5 shows the original plot from Gries's (2022a:14) Figure 3 (top) and a new plot of simply log(a) against 12-log(b) (bottom), where 12 was used to flip the y-axis because the top logged reference frequency was approximately 12.



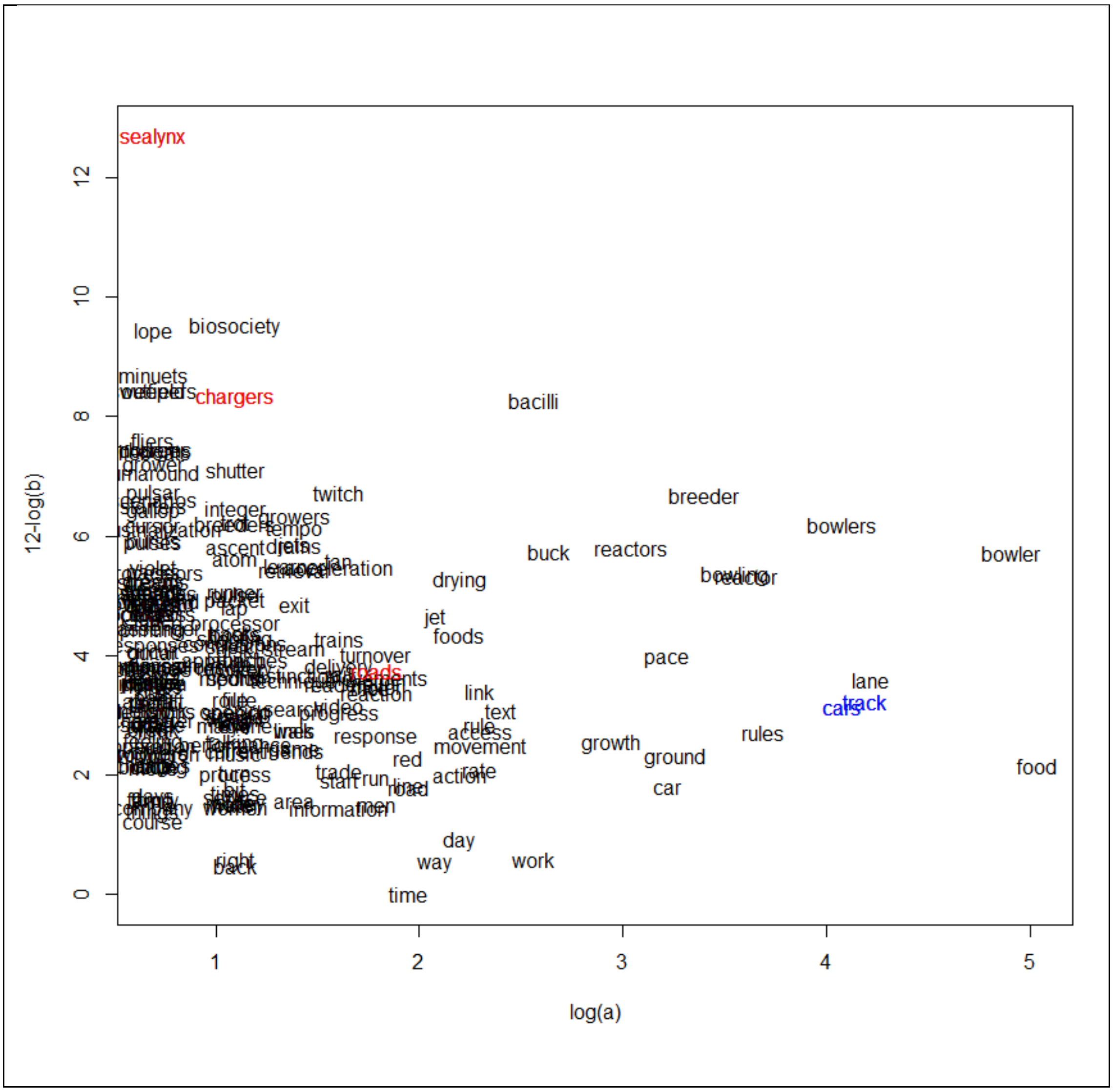
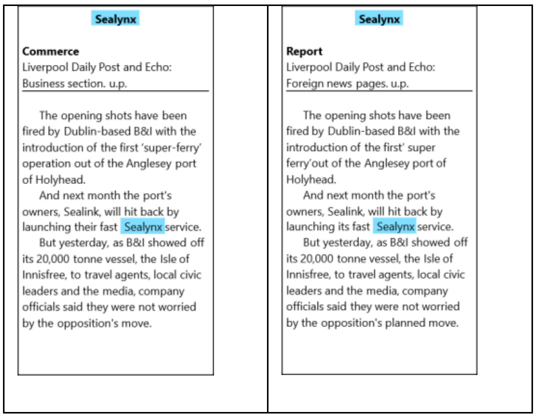


Figure 5: Comparing Gries (2022a) Figure 3 with two logged raw parameters

The reader's attention is drawn to the similarities in the positions of the five coloured items (which have been coloured in the same way in the bottom plot) and the positions of the other legible items. Given the R^2_{gam} shows an even stronger relationship between Association-Without-Frequency and log(b) than odds ratio, the similarity between the two plots in Figure 5 is particularly striking. These visualizations do demonstrate the potential for exploring more of the ground of relationships between co-occurrence frequency and non-co-occurrence frequency (or frequency and ‘association’), but with all the detail in terms of language teaching it seems overly optimistic that serendipity would lead a language learner to a useful position on the chart. Clearly, the challenge from Gries is to explore new ways of exploring multiple dimensions, but within the context of second language learning the cartography metaphor Sinclair (2004), borrowed from J. Borges seems at the very least partially fitting:

> the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that the vast Map was Useless.
>
> (J. Borges, “On Exactitude in Science”, in *The Maker*, 1960, quoted in Sinclair, 2004:286)

It would be completely unfair to suggest that it would be ‘useless’ to pay attention to low co-occurrence and even lower non-co-occurrence as well as high co-occurrence and high non-co-occurrence, and everything in-between, but it certainly seems the proposed measure is overly sensitive to non-co-occurrence and may be harder to apply as pedagogical prioritizing in many language learning situations. If most traditional metrics are “re-packaged [co-occurrence] frequency information” (Gries 2022:3) or parameter a, odds ratio and Association-Without-Frequency are to a large extent re-packaged non-co-occurrence frequency or parameter b. Gries’s overall message is that the new measure is not designed to be used for sorting results by just one column, but nevertheless, as he notes, the top collocate for *fast* would be *sealynx* (occurring just twice in the whole corpus and both times immediately after *fast*). He explains that “*sealynx* is ‘contextualized’ for the analyst by its low co-occurrence frequency.” (Gries 2022a:13). The item *sealynx* or *Sealynx* needs to be contextualized by more than low frequency, however. The wider context for these two hits shows both come from the same newspaper and one clearly draws on the other’s language (shown in Figure 6).



Sealynx

Commerce
Liverpool Daily Post and Echo: Business section. u.p.

The opening shots have been fired by Dublin-based B&I with the introduction of the first 'super-ferry' operation out of the Anglesey port of Holyhead.
And next month the port's owners, Sealink, will hit back by launching their fast Sealynx service.
But yesterday, as B&I showed off its 20,000 tonne vessel, the Isle of Innisfree, to travel agents, local civic leaders and the media, company officials said they were not worried by the opposition's move.

Sealynx

Report
Liverpool Daily Post and Echo: Foreign news pages. u.p.

The opening shots have been fired by Dublin-based B&I with the introduction of the first' super ferry'out of the Anglesey port of Holyhead.
And next month the port's owners, Sealink, will hit back by launching its fast Sealynx service.
But yesterday, as B&I showed off its 20,000 tonne vessel, the Isle of Innisfree, to travel agents, local civic leaders and the media, company officials said they were not worried by the opposition's planned move.

Figure 6: Concordance cards for top-ranking Association Without Frequency collocation candidate sealynx in the BNC.

It is unclear whether Gries’s (2022a) *fast + N* example was based on treating proper nouns separately, but the CLAWS (Garside & Smith, 1997) part of speech tag for *Sealynx* in these two concordance lines is NN1 rather than what might have been NP1 if the brand had been famous enough to warrant special treatment by the tagger. This can be demonstrated by using the free CLAWS online tagger (http://ucrel-api.lancaster.ac.uk/claws/free.html) to tag an invented example with similar structure, where the better known *Microsoft* is correctly tagged as NP1 (the names Microsoft and Microsoft Store are used purely for demonstrative purposes).

```
And_CC next_MD month_NNT1 the_AT outlet_NN1 's_GE owners_NN2 ,_,
Microsoft_NP1 Corporation_NN1 ,_, will_VM hit_VVI back_RP by_II
launching_VVG its_APPGE fast_JJ Microsoft_NP1 Store_NN1
service_NN1 ._.
```

The intertextuality of the two Sealynx examples supports Gries's point about dispersion also being a useful factor to consider for collocations (like his examples of *fast* + *food* and *fast* + *bowler*). The low frequency and narrow dispersion are part of the needed contextualization for *Sealynx*, but this specific example also highlights some issues when exploring low frequency phenomena, especially if relying on part of speech tagging; for low frequency, high 'association' items, the analyst best beware; in the words of other cartographers of legend: here be dragons!

Similarly, another high scoring Association-Without-Frequency collocate, *bacilli,* has a total frequency of 54 in the BNC and a relatively low frequency of 13 hits immediately after *fast*. These 13 co-occurrences are in just four texts, and 9 of these hits come from a single academic journal article. In contrast, *fast bowler*, the high ranking item when more traditional collocation measures are used (which Gries questions in terms of its usefulness in second language acquisition) occurs fairly frequently both in the Newspapers and Other Publications sub-corpora of the BNC; it might reveal more about a 1990s British fascination with reporting cricket or about the design choices of the BNC itself, but nevertheless it is a strong collocation on these measures as well as aysmetrical implementations (where *fast* collocates strongly with *bowler* and *bowler* collocates strongly with *fast*). Not only is it hard to see how Association-Without-Frequency can provide useful information for second language learners, by adjusting the proportion according to the maximum co-occurrence possible given an imbalance in the two frequencies, it also downgrades the asymmetry of the relationship between items and favours collocates which do not occur frequently with any other items. In a different paper, Gries (2013) explained the importance of asymmetrical measurements of collocation; the node *bacilli* would have a Delta-P value of 0.24 for the collocate *fast* when using a four word window without part-of-speech filtering and taking into account the ordering of the items, while the Delta-P score would be less than 0.00162 as a measure of association between the node *fast* and the association of *bacilli* as the collocate. Gries makes the following summary of his criticism of log-likelihood:

> Thus, even the smaller degree to which G2 reflects what everyone is using it for varies in a graded fashion according to co-occurrence frequency, which makes it an even less clean measure of association than if it reflected association less than frequency, but at least consistently so.
>
> (2022a:13)

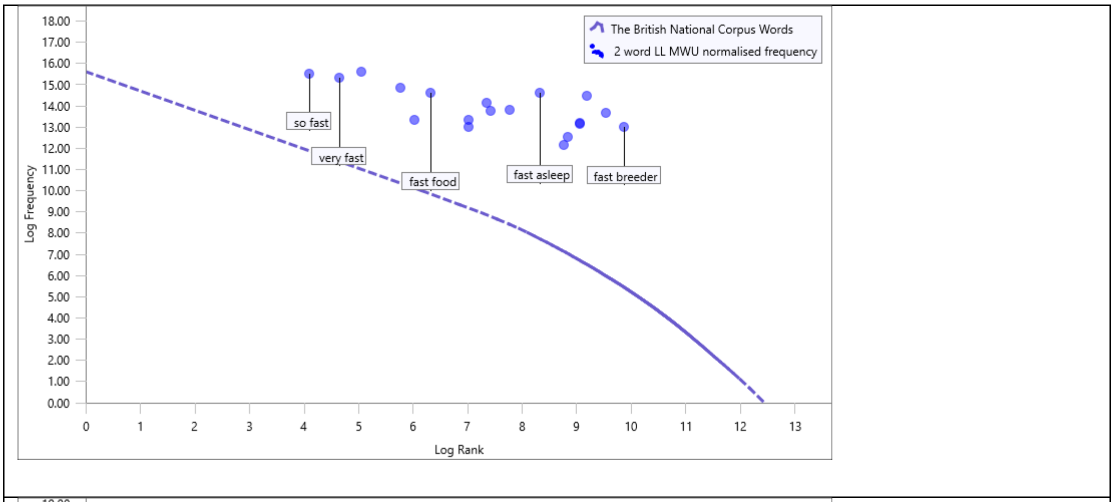
Based on the summary above, a similar criticism could be levelled at Association-Without-Frequency as follows: the strong relationship between Association Without Frequency and the (lack of) occurrences of a potential collocate with other words demonstrates it, too, is inconsistent, preferring extremely rare items on the one hand, yet favouring a more mixed set of potential collocates for words such as *quick* (Table 1). Given the Zipf distribution of items in a corpus, it is to be expected that most mid-frequency items would co-occur with a large number of extremely rare items, and as a consequence, these extremely rare items will be top-ranked even though they are unlikely to be of any pedagogic use. Not only will odds ratio and Association-Without-Frequency look (in Gries's words) less "intuitively satisfying" (2022b:3) – indeed they do – they cannot be said to be a precise measure of association when they are so heavily influenced by the non-co-occurrence of a candidate in other contexts. This would be particularly true of key word analysis, where (for example) the choice of reference corpus and long lists of proper nouns are likely to muddy the water.

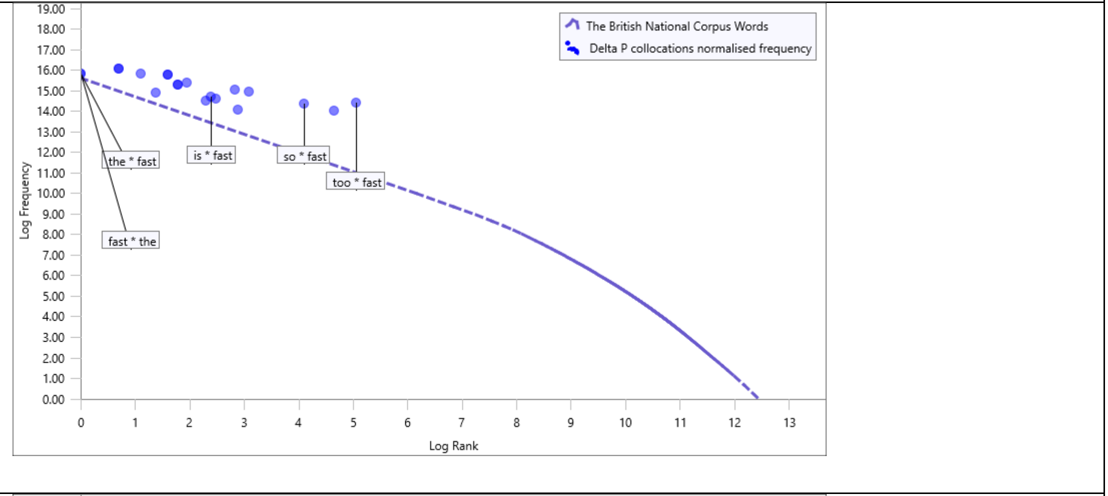
Ultimately, the usefulness of any collocation metric should be determined by the researcher within the context of the application of the resulting data. Gries (2022a) does provide a new way to combine co-occurrence and non-co-occurrence, and the metric may well prove useful. He also promotes an approach to collocation which goes beyond single-column sorting. Despite its strong correlation with reference frequency, the new association measure deserves attention as one of a range of existing association measures; it cannot be said to be independent from other raw data inputs, but it does provide an inspired new way to explore relationships which come under the larger umbrella of association. This is not then a response to Gries (2022a) intending to undo the potential or the ingenuity of the new approach; it is more a defensive response to some of the claims against log-likelihood (and other metrics) for collocation and keyness, and a positive response in terms of due attention to what might be missing if users of familiar metrics do not bear in mind their sensitivities, particularly when only looking on one dimension and only looking at the subset of top-ranked results.

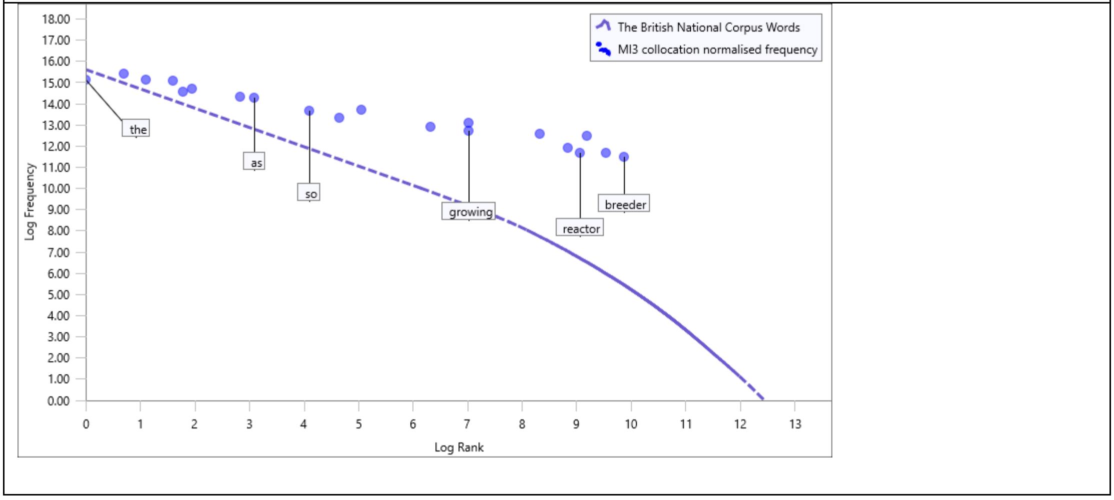
In practical terms, however, there are a number of caveats. While log-likelihood collocations and key words will have weighed up multiple factors including co-occurrence frequency, the proportion of c / not c, the proportion of b / c and the overall sample sizes, Association-Without-Frequency needs active consideration of how to weigh frequency against ‘association’, how to compare results from one corpus with another, and how to check dispersion, intertextuality and various errors when exploring low frequency rare events. But what other lessons can be learned from Gries’s discussions of Log-likelihood, T-score, and MI? As Gries notes, “brings out” (Bestgen

& Granger 2014:31 quoted in Gries 2022a:15) is too vague for a precise description of statistical methods for the inner workings of a software application or script, and "collocations composed of very frequent words" (Bestgen & Granger 2014:30 quoted in Gries 2022a:15) depends first on dismissing combinations of very frequent words which are not considered to be collocations. Nevertheless, within the context of the original papers these expressions certainly do describe what many researchers would experience if they used mainstream off-the-shelf corpus tools; web-based interfaces will tend to offer top-ranked hits on the first page, with lower scoring items theoretically being available through a large number of clicks; software installed on the user's machine may allow re-sorting of columns or a scrollbar to locate very low scoring items, but cut-offs and other settings are likely to have been used in the software algorithms to minimize the processing workload and to avoid grid-like components becoming sluggish in the user interface through having too many lines or results. As inspiration for teacher-researchers using corpus methods to analyze their students' essays, Bestgen and Granger (2014) imprecise explanation that within the results of statistically significant t-score collocations, what most users of most off-the-shelf corpus packages will obtain will be a first page of most highly ranked collocations which are composed of high-frequency words. Sorting by one column and focusing on top-ranked results does not seem unreasonable within the context of a relatively small corpus of learner essays, focusing on the word frequency ranges of highly ranked t-score results, especially as language teaching often draws on results of vocabulary profiling or intuitions of whether students in a class are likely to be more familiar with some vocabulary items than with others. One of the applications of CollGram noted by Bestgen and Granger is to "point to the collocations used by learners that are typically used by native speakers and those that are more rarely used by them, if at all" (2014:39). While in the description of the inner workings of CollGram the authors should have been more precise, in terms of the application of these results and the way some salient features might be displayed to the novice writers of the essays as part of L2 teaching, the emphasis on top-ranked matches when using different collocation scores is reasonable.

Therefore, Gries (2022a) provides a useful reminder of the importance of considering the points on the Zipf curve from which candidate collocates tend to come, when comparing the top-ranked collocations produced using different metrics. This has inspired a new visualization in *The Prime Machine* corpus tool (Jeaco, 2017) which has already been incorporated into the user-interface for the five collocation measures available: log-likelihood, Delta-P, MI3, T-Score and Dice. In addition to the collocation clouds and collocation tables previously available, the new visualization provides a Zipf log-rank log-frequency plot for the currently selected corpus and then superimposes the set of collocations which are visible in the clouds and tables as a series of dots. The position on the x-axis for each collocation is determined by the log rank of the collocate's frequency in the corpus overall. The position on the y-axis is determined by taking the ratio of co-occurrence with the node and extrapolating what the frequency would be if the relationship between the node and corpus were representative of the entire corpus.


The British National Corpus Words
2 word LL MWU normalised frequency
so fast
very fast
fast food
fast asleep
fast breeder
Log Frequency
Log Rank


The British National Corpus Words
Delta P collocations normalised frequency
the * fast
fast * the
is * fast
so * fast
too * fast
Log Frequency
Log Rank


The British National Corpus Words
MI3 collocation normalised frequency
the
as
so
growing
reactor
breeder
Log Frequency
Log Rank

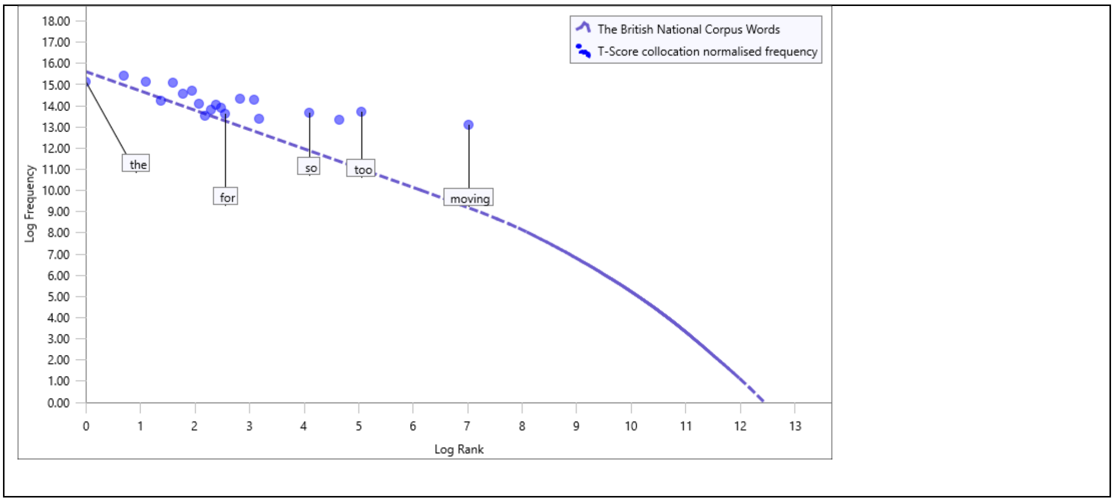


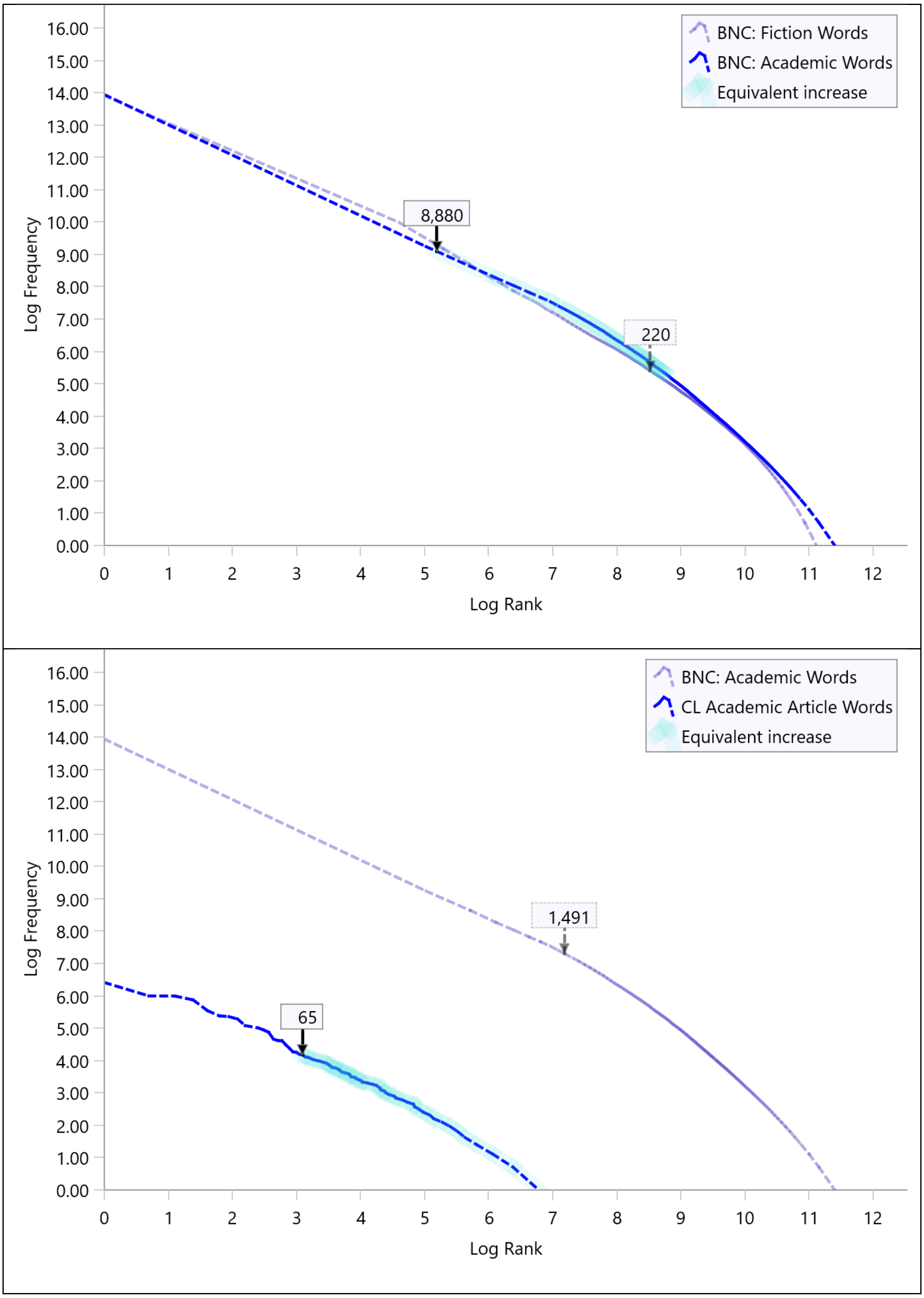
Figure 7: Scatter Charts for the node *fast* in the BNC, using four different collocation measures.

As can be seen in Figure 7, if the user only uses the top 20 collocations in each case, it really is the case that the different metrics provide collocates from quite different ranges of overall frequency in the corpus as a whole. In the images shown here, several of the dots have been clicked to reveal annotations showing the full collocation for log-likelihood and Delta-P results or just the collocate for the others. In this corpus tool, the former two measures are handled as asymmetric collocation measures which also take into account the ordering of the two words.

One of the benefits of this visualization is for introducing the effects of single-column sorting of commonly used collocation measures; the fact that these tendencies are for top-ranked single sorted collocations only is front-and-centre. The user of the application is invited to flick back to the collocation tables (sorted by a different single column) and to flick back to the concordance line sample which has been downloaded with these data; analysis of collocation in classroom settings can be well supported through engagement with concordance line data – not merely in isolation. This allows for the safety in numbers and focus on the typical which is often the focus of language teaching.

4. Keeping the Zipfian Curve in mind for key words

In *The Prime Machine* while log-likelihood is used for both collocations and key word functions, a different visualization has been developed to help users understand the relationships between frequencies in the study and reference corpus for a key word item. This was inspired by the importance of considering the relative sizes of the two corpora used for key word analysis, and by the importance of frequency for the notion of keyness. In *The Prime Machine*, key words can be generated by comparing frequencies in two readymade corpora, a Do-It-Yourself (DIY) corpus and a readymade corpus, or two DIY-corpora. As a server-client application, *The Prime Machine* gives users access to readymade corpora which are typically many millions of words in size and also provides tools to import texts to build DIY corpora which might be just one text of a few hundred or few thousand words, or a relatively small corpus of a few million words (constraints being memory on the device and the patience of the user as DIY corpora larger than a few million words see a performance hit). As noted in the definitions earlier, key words have a frequency which is unusually high in one corpus in comparison with another.


BNC: Fiction Words
BNC: Academic Words
Equivalent increase
8,880
220
Log Frequency
Log Rank
BNC: Academic Words
CL Academic Article Words
Equivalent increase
1,491
65
Log Frequency
Log Rank

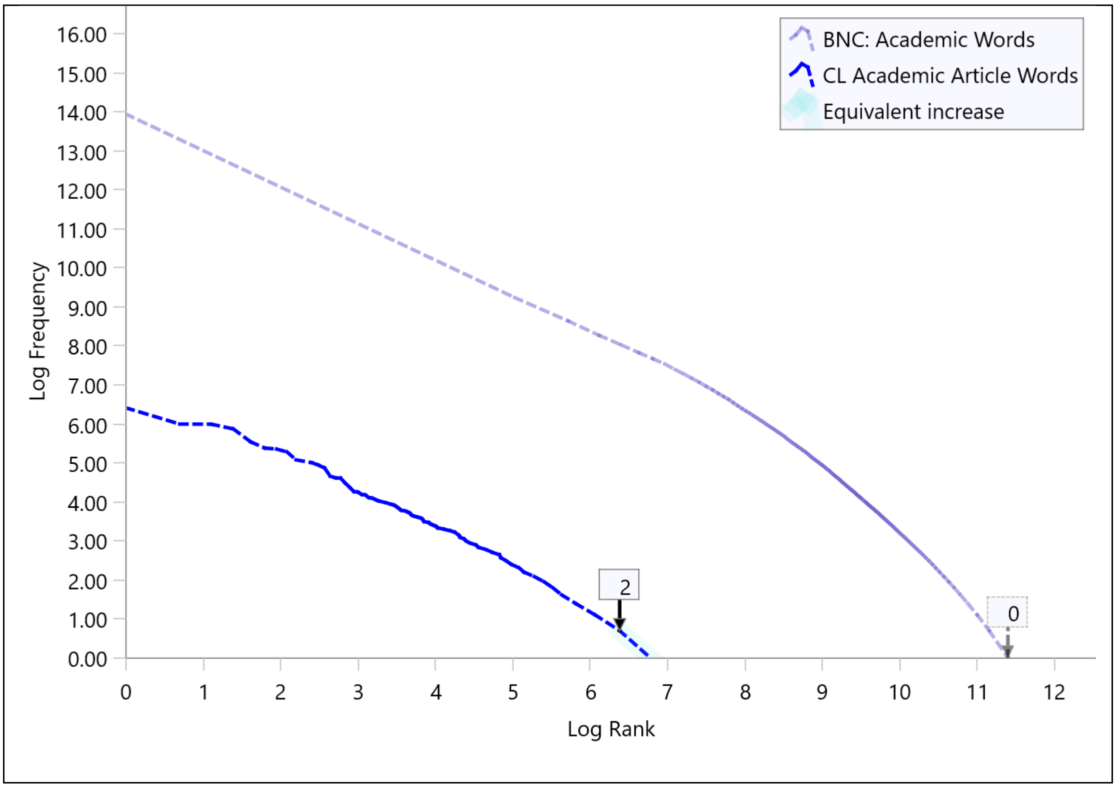


Figure 8: Twin double logged charts for two corpora of similar sizes (top: *development* in BNC: Academic vs. BNC: Fiction) and a small DIY corpus compared with a larger reference corpus (centre: *frequency* and bottom:*freq* in a Corpus linguistics academic article compared with BNC: Academic).

Figure 8 shows the visualizations for three different key words based on comparisons of two corpora of a similar size and of a smaller corpus (of approximately 10,000 words) with a larger reference corpus. As can be seen, in all three charts, the Zipf curve for each corpus is plotted and the point for the key word's log rank and log frequency is also marked on the curve. In addition, a turquoise marker highlights the equivalent increase between expected and actual frequency of the item in the study corpus, using the normalized frequency of the reference corpus to estimate where on the study corpus curve the key word would have been positioned if its rate of occurrence in the study corpus had been the same as it was in the reference corpus. As a tool for language learners and teachers, ranking key words in descending order by log-likelihood provides useful results. However, when ranking by log-likelihood, the top ranked items typically do not have the highest frequency of all key word items which meet the minimum statistical significance cut-off. This may lead some users to question why a key word with a lower frequency of occurrence in the study corpus is ranked higher than key words with a higher frequency. Comparing *frequency* (middle) with *freq* (bottom) shows the relative importance from the text perspective of *frequency* (which is not exclusive to the study corpus) compared to *freq* (which is exclusively used in the single article and not attested in the reference corpus). These additional graphs can provide a quick and convenient way for language teachers or language learners working independently to see a visual representation of the important relationships between the relative sizes of the two corpora used to generate key words (the distance between the curves), the differences between the item's rankings in both the study and reference corpus (the two points on the curves) and the degree of increase in the rate of occurrence (the estimated change of the item's ranking). It also clearly shows where the selected key word fits into the corpus as a whole, in terms of the frequency range from which it comes.

5. Re-examining issues of dispersion measures

Gries (2022b) highlights the fact that dispersion might be best considered not only from the perspective of the number of texts in which an item occurs, but also in terms of the rate of occurrence within the texts in which it occurs. While the new measure proposed by Gries (2022b) could be an additional metric, it is relatively expensive in terms of processing and has a similar bias towards very low-frequency items in a similar way to the Association-Without-Frequency metric discussed earlier. Returning to the matrix of Figure 1, and considering movement up the vertical axis, as a sample increased (more texts were added) with other measures of dispersion the addition of any texts not containing the item would weaken the dispersion score. With DP-nofreq, adding texts has no affect on this measure of dispersion – adding one text or one million texts not containing the item will not affect it. From the perspective of a low frequency word – as indicated by the exasperated tone of Gries's vivid description of what the word *enormous* might say if it could talk – adding texts which represent contexts in which a word would not occur does not need to affect the measure of its dispersion in the sense of to what degree it is evenly spread. However, if a corpus is increased in size and an item does not occur, for research considering dispersion from the corpus or text perspective, the word is quickly becoming less of an interest in the wider scheme of things. It could be argued that the new measure is also skewing the dispersion measure for items with a frequency which is less than the total number of texts.

Gries (2022b) has provided some new insights into dispersion, but within the context of a corpus tool for language learning and teaching, it seems more helpful to draw attention to different aspects of the meaning of dispersion in a more basic way, and this is what the new visualization for dispersion in *The Prime Machine* tries to accomplish; the Diffusion Charts try to capture three aspects of the way a word is spread across the texts of the corpus. Firstly, they give an indication of the size of the text or texts containing the item, relative to the whole corpus. The cone is always the same shape, but the higher the value on the scale, the greater the proportion of texts containing at least one instance of the item. This corresponds to Gries's $\text{range}_{\text{withsize}}$ measure of dispersion, and while different was greatly inspired by Gries (2022b).

Second, the depth of colour of the chart gives an indication of the density of the item – that is how frequently it occurs in each of its texts, relative to their size. The Density figure gives a normalized frequency per thousand words, but unlike the other frequency charts available in the app, the figures for Density on the diffusion charts are normalized based only on the texts in which the item occurs. For example, if a word only occurs in one text, the Density value will be its frequency divided by the number of running words in that text. If a word occurs in five texts of different sizes, the token counts for those five texts will be totalled, and the frequency will be normalized using this total. This measure is also based on a re-orientation following Gries (2022b), and contrasts well with the other frequency measures which are available in the app. Rather than providing a range from 0 to 1 for display in a table, this visualization is an attempt to capture the same aspect of dispersion through colour density. Finally, the variation of the depth of the colour of the chart gives an indication of how evenly spread (or otherwise) the item is within the texts in which it occurs. For example, if an item occurs in two texts, but has a low rate of frequency in one and a high rate of frequency in the other, the top part of the cone will be lighter and the bottom part of the cone will be darker. Items which have more consistent rates of occurrence will appear with little or no difference in the shading of the colour. This provides a visualization of the range of different rates using 6 slices of the cone. The way the colours are determined is through the following procedure: first, the texts are ranked by the rate of occurrence of the item, and along with each text's individual density, the cumulative total token count is kept; then the total token count for texts containing the item at least once is divided by 6 and the density score for each of the six slices is based on the density of the text which corresponds to the point. The density is converted into the alpha channel value for the fill colour, and each slice has a gradient filling style beginning with the lower slice density value and going up to the next slice's density value.

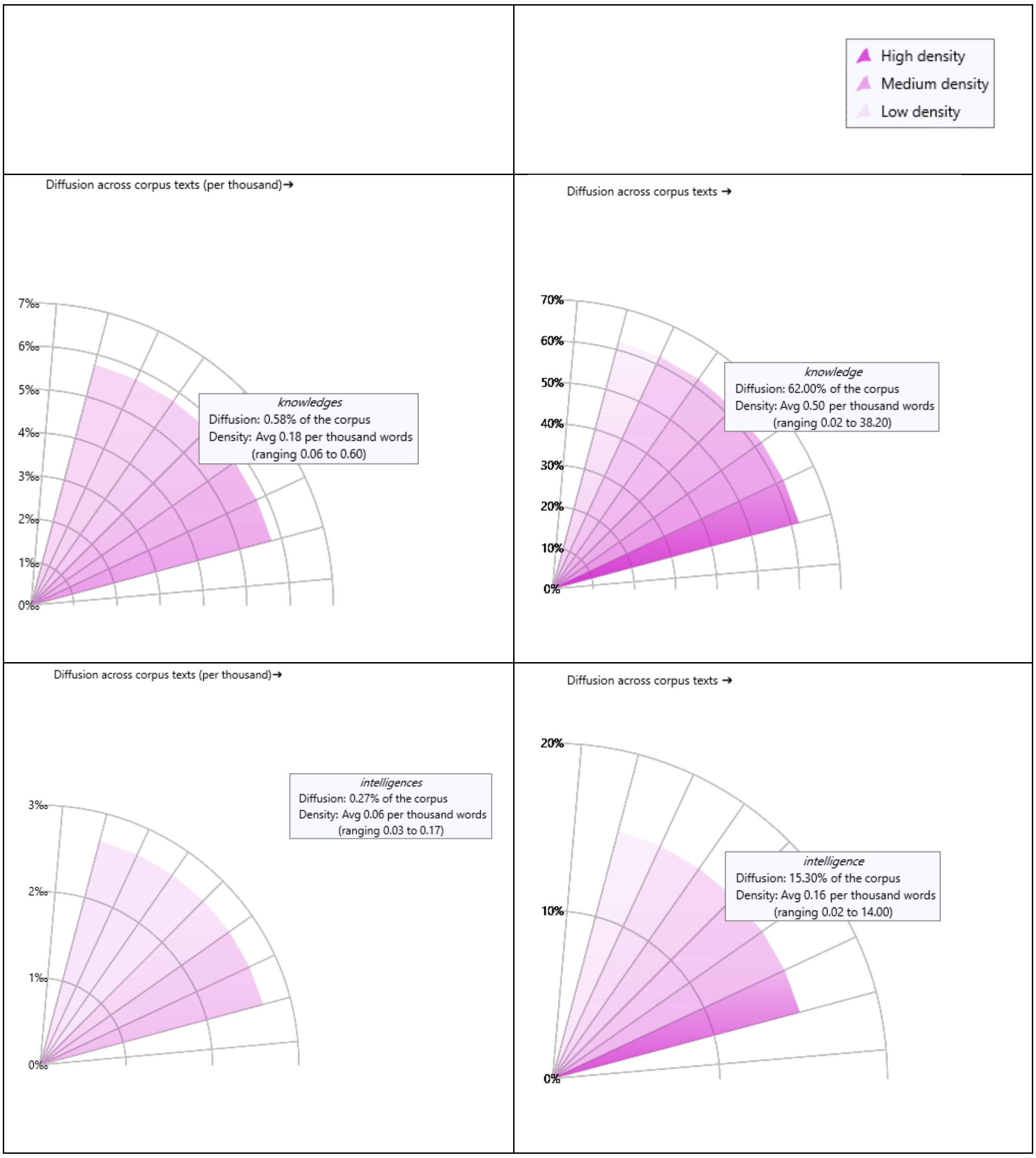


Figure 9: Diffusion Charts for *knowledges, knowledge, intelligences* and *intelligence* in the BNC: Academic Sub-corpus.

As can be seen in Figure 9, two pairs of words have been selected to illustrate how these three dimensions affect the visual representation of diffusion. In the top-left corner, it can be seen that *knowledges* appears to be a highly specialist form and is only in texts which represent 0.58% of the BNC: Academic corpus. Its rate is relatively low throughout these texts. In contrast, *knowledge* (top-right) is a word which will be found in 62% of the texts of the BNC: Academic corpus. Some texts have a relatively high rate of occurrence; it might be predicted that in these texts *knowledge* would be a strong key word. It also occurs in some texts at a lower rate (as low as 0.02 per thousand words). The second pair shows a contrast between the rather rare and specialist term *intelligences,* which has been plotted on a per thousand scale and occurs in 0.27% of the texts in the corpus, and *intelligence* as a word which is fairly widespread (over 15% of the texts in the corpus have it at least once). As with *knowledge,* the word *intelligence* occurs fairly often with a low rate of frequency in its texts, but there are some texts where it would be expected to be a strong key word too. This visualization allows the user to quickly get answers to questions such as the following:

- Taking into account the different lengths of texts in the corpus, how many of the texts would I need to read before I would be almost certain to come across at least one example of *intelligence*?
- If a text has the word *intelligence*, does it repeat it often?
- How similar are the rates of occurrence of the word *intelligence* within the texts where it can be found at least once?

For a teacher wanting to focus students' attention to the fact that attention to context is important and to foster consideration of the differences in distribution of different word forms, the examples here, along with the new visualizations, help carry the point. This approach is an example of how innovative new metrics from Gries (2022b) can inspire a new way of presenting the same underlying motivations.

6. Conclusion

In conclusion, Gries (2022a, 2022b) demonstrates the strong connection between frequency and commonly used metrics for collocation and keyness, and he takes the reader beyond single-column sorting by discussing issues from the perspective of individual words within the texts of the corpus which contain them. He presents graphs to highlight the degree of influence of frequency on commonly used statistical tests, along with proposals for alternative approaches. Moving forward, the word-centric measures proposed by Gries could well serve a purpose, especially in machine learning applications (because for machine learning, the weightings of ‘association’, frequency and other features can be fitted to the data automatically and would not require numerous decisions on how these should be combined). One might ask whether these could be extended further by thinking about dispersion (in all its senses) on more levels: Sub-corpus vs sub-corpus, Collection (author / publication / speaker), Text, Within-text and Within-paragraph / sub-section / section / chapter. Gries does provide new opportunities and opens the door for more diverse research in terms of relative weightings of frequency, dispersion and his ‘association’, but of course these new opportunities will require careful consideration of how to combine these aspects, and how to do it consistently.

A full answer to Gries’s article title questions needs to be research-contextually bound and application specific. For researchers using off-the-shelf tools, well-supported association measures in software packages for collocation and keyness and measures of dispersion typically measure a textual phenomenon which to some extent will reflect the happenstance of text selection, but will also represent some of the associations which must exist in the minds of language users. These measures typically try to deal with the problem of the Zipfian nature of text: having some words (types) occur extremely frequently (being hard to pin down in terms of typical patterning, and also usually of less interest to researchers, teachers and language learners), extremely low frequency items (the descriptions of the patterns of which are likely to be highly specific to a corpus and most subject to sampling error) and a large number of items which interact with words on all points of the curve, and usually form the focus of language learning and teaching. For new developments in machine learning applications, developers would do well to be aware of these special features of language data and inject into the algorithms elements in their composite forms. For other applications, especially where ranking by a single column of data is the main data inspection method employed, common measures of association and dispersion are still fit for purpose; they typically work best when working with text or collection of text (corpus) as the unit of study; they could be supplemented usefully by some newly proposed measures from Gries, especially if the focus of study is specific lexical items or perhaps pairs and groups of related lexical items in contrast. Researchers working with a specific, relevant corpus and wanting to extract for exemplification or teaching purposes examples based on the patterns that have been extracted by association and dispersion methods are likely to find the more traditional methods more helpful in the sense that their top-ranked results will typically be more representative of the corpus overall. Essentially, the question of whether they measure what they set out to measure, depends somewhat on understanding the way seemingly transparent terms have been used suppositionally in over 50 years of literature and being aware of the stated limitations and operational optimizations studies presenting these have employed. Those forming multidisciplinary research projects might do well to stick to definitions which describe the textual effects (that can be measured through corpus methods) and only hint at why these might be valuable to human inspection. For those moving from single-column sorting to more multivariate methods of analysis, more precision in descriptions will be needed. Gries has (yet again!) identified new perspectives on relations in text, and has provided new measures (c.f. 2013, 2015); these should form useful additions to some of his other innovations and to the broader range of metrics available. However, projects working with a convenient but questionably comparable corpus would need additional caution when employing measures which are so sensitive to fluctuations in low frequency, essentially rare events. In the meantime, new visualizations like those presented here should help raise awareness of the way choice of statistical measure can affect the kinds of phenomena which will be top-ranked, and help keep both the Zipfian curve and different nuances of dispersion front-of-mind.

Notes

1. The parameters are labelled differently, so in this earlier paper this relationship is referred to as "a+b vs. c+d" and is represents the "Relative sizes of study and reference corpora" (Jeaco 2020:137).

References

Anthony, L. (2022). AntConc (Version 4.0.1). Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software/antconc/

Bestgen, Yves & Sylviane Granger. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26. 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173.

BNC. (2007). The British National Corpus (Version 3 BNC XML ed.): Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. Boston: Addison-Wesley.

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

Evert, Stefan & Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, p, 188–195.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102-121). London: Longman.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137-165.

Gries. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. https://doi.org/10.3366/cor.2015.0068

Gries, S. (2022a). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1). https://doi.org/10.1075/jsls.21028.gri

Gries, S. (2022b). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*. https://doi.org/10.1075/jsls.21029.gri

Hann, M. N. (1973). The Statistical Force of Random Distribution. *International Journal of Applied Linguistics*, *20*, 31-44.

Hardie, A. (2012). CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, *17*(3), 380-409.

Heaps, H. (1978). *Information retrieval: Computational and theoretical aspects*. New York: Academic Press.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoey, M. (2014). Words and their neighbours. In J. R. Taylor (Ed.), *Oxford Handbook of the Word*. Oxford: Oxford University Press.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jeaco, S. (2017). Concordancing Lexical Primings: The rationale and design of a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 273-296). Amsterdam: John Benjamins.

Jeaco, S. (2020). Key words when text forms the unit of study: Sizing up the effects of different measures. *International Journal of Corpus Linguistics*, 25(2), 125-154.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (Vol. 4, pp. 1-13). Birmingham: Centre for English Language Studies, University of Birmingham.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). *The Sketch Engine*. Paper presented at the 2003 International Conference on Natural Language Processing and Knowledge Engineering, Beijing.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*: New York ; Routledge, 2013.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Rayson, P., & Garside, R. (2000). *Comparing corpora using frequency profiling*. Paper presented at the Workshop on Comparing Corpora, Hong Kong University of Science and Technology, Hong Kong.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. Boston, MA:PBC. Retrieved from http://www.rstudio.com/.

Rychlý, P. (2008). *A lexicographer-friendly association score*. Paper presented at the Recent Advances in Slavonic Natural Language Processing Conference, Masaryk University, Brno.

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

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Scott, M. (2020). WordSmith Tools (Version 8). Oxford: Oxford University Press.

Scott, M. (2022). WordSmith Tools online manual "KeyWords: calculation". Retrieved 31 October, 2022, from http://www.lexically.net/downloads/version7/HTML/keywords_calculate_info.htm

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2004). New evidence, new priorities, new attitudes. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 271-299). Amsterdam: John Benjamins.

Wermter, J., & Hahn, U. (2006). *You can't beat frequency (unless you use linguistic knowledge): A qualitative evaluation of association measures for collocation and term extraction.* Paper presented at the Annual Meeting of the Association for Computational Linguistics, Sydney.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* (B) 73(1):3-36.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston, MA: Houghton Mifflin.

Address for correspondence

Stephen Jeaco
Department of Applied Linguistics
HS431, Xi'an Jiaotong-Liverpool University,
111 Ren'Ai Lu, Suzhou Industrial Park,
Suzhou, Jiangsu Province,
P. R. China
steve.jeaco@xjtlu.edu.cn