BOOK REVIEWS

Gries, S. T. (2024). *Frequency, Dispersion, Association, and Keyness: Revising and tupleizing corpus-linguistic measures*. John Benjamins. 321 pp.

Reviewed by **William C. X. Platt** (Lancaster University)

Much of corpus linguistic research uses measures of association, keyness, or dispersion (e.g. for collocations, keywords, and lexicography). The use of frequency is even more ubiquitous. Based particularly on two sister papers (Gries, 2021a, 2021b), Gries' new monograph aims to expose critical flaws in the ways corpus linguists measure or use frequency, dispersion, association, and keyness, and to provide solutions to these problems. Secondarily, the work seeks to be a textbook introducing the most common measures and teaching some R programming. As a result, no familiarity with corpus-linguistic statistical measures is assumed, but some experience in R is recommended. Below is a summary of the main arguments and conclusions, followed by a critical reflection.

Gries identifies the impropriety of using a corpus frequency without a corresponding measure of spread, analogous to a mean without standard deviation (and a corpus frequency is the weighted mean of frequencies across parts of a corpus). Therefore, we ought to be utilising dispersion (which for Gries means evenness of distribution) concomitant with frequency. However, Gries uses some example corpora to show that all the most common dispersion measures (DMs) are strongly correlated with frequency (often over 90%) rather than providing an independent/orthogonal metric. This means they measure frequency much more than what they intend to measure. He furthermore finds that the range of theoretically possible dispersion values is very different for items of different frequencies; 0.2 could be a very low value for high frequency words but high for low frequency words. That means dispersion values are not comparable for items of different frequencies, even for the same corpus.

Regarding association and keyness, Gries highlights that these can be measured in two directions: how much the frequency table's row variable appears to affect the column variable, or the opposite column-to-row direction. Yet this bidirectionality is rarely recognised, with ΔP (Gries, 2015; Schneider, 2018) being one of the few metrics corpus linguists are likely to use that can measure each. Moreover, some common association measures (AMs) and all those that measure the column-to-row direction are very strongly correlated with co-occurrence

frequency (frequently with $R^2_{\text{GAM}} > 0.99$), and measures of keyness are highly correlated with the difference between frequency in the target corpus and that of the reference corpus. This means they are overwhelmingly measures of frequency, not association or keyness. Like with DMs, association/keyness measures are found to have varying ranges of possible values for items of different frequencies, meaning association/keyness values are not comparable for items of different frequencies, even with all else being equal.

Somewhat related to these issues of correlation with frequency, Gries argues that multiple dimensions of information should never be combined into a single dimension. For instance, he is vehemently opposed to combining frequency and dispersion to calculate an adjusted frequency. This is because doing so leads to a loss of information that is likely to mislead us.

Finally, Gries touches on the fact that we should be computing confidence intervals for our values, and that bootstrapping is the proper approach. But doing this for association on a million-word corpus, he finds the confidence intervals so large that we cannot say much about the association of words below frequency 16 (easily the vast majority of words). Fortunately, the situation seems less bleak for dispersion (Gries, 2022).

Because of the above problems, virtually all corpus linguistic work based on the ranking or comparison of dispersion, association, or keyness values is brought into question. That would include, for example, almost all corpus-based studies of collocations, colligations, and keywords.

Gries' solutions to these problems are as follows. First, he proposes that we use the Kullback-Leibler divergence (KLD) (Kullback & Leibler, 1951) as our measure of choice for dispersion, association, and keyness. His main arguments for preferring the KLD include the fact it can be applied to measure all of these (thus 'unifying' them), including both directions of association/keyness, and its information-theoretic pedigree. The use of a single DM or AM is also intended to simplify decision-making for corpus linguists.

Second, he proposes a method of removing ('partialing out') frequency from any measure. This is done by finding the minimum and maximum values the measure can give for a particular frequency and mapping that range to a desired interval, such as [0, 1]. As an illustration, if an item has frequency 12 and for items of frequency 12 the measure can only output values between 0.1 and 3.5, then 0.1 is mapped to 0 and 3.5 is mapped to 1. If the item was originally given a value of 3.2, then this is mapped to (3.2 - 0.1)/(3.5 - 0.1) = 0.912. In this

way Gries solves both correlation with frequency and the matter of output ranges varying for different frequencies.

Third and finally, instead of amalgamating multiple dimensions into one, Gries says we should 'tupleise'. In other words, use orthogonal, uncombined dimensions of information so that no information is lost. Gries exemplifies this by using the three dimensions of frequency, association (which in this case is what he elsewhere refers to as keyness), and dispersion to identify keywords without computing keyness scores.

Now let us examine the arguments in more detail. The book makes solid arguments that corpus frequencies should not be used without dispersions, that our measures should not be so predictable from frequency, and that we should be computing confidence intervals using a method appropriate for corpora. Additionally, greatly different ranges of possible output based on frequency can harm the comparability of outputs across frequencies.

At the same time, however, these variable ranges are not without reason, as for example words with only one occurrence in a corpus cannot vary in dispersion as much as words with many occurrences. It's also worth noting that Gries describes dispersion as evenness of spread, but in the literature it's commonly conceptualised more broadly as a measure of spread across a corpus (e.g. Hashimoto & Egbert, 2019: 842, 846), and Egbert and Burch (2022) have shown that we ought to distinguish dispersion as pervasiveness from dispersion as evenness. But pervasiveness appears to be intentionally correlated with frequency, with low frequency items being unable to pervade an entire corpus and deliberately having a lower maximum pervasiveness than higher frequency items.

Gries' concern that combined measures can be misleading is justified, but on the other hand, simplification of data has its benefits. High dimensional data (e.g. from each text in a corpus) may be infeasibly slow for computers to process, or even too large to store, as well as having undesirable properties, motivating dimensionality reduction (Ayesha et al., 2020). Having a sense of the data in our minds is essential for human reasoning and pattern recognition too, but this is difficult for more than a few dimensions (Cowan, 2001). This explains why combined metrics like BMI (body mass index) and summary statistics such as means, standard deviations, and corpus frequencies are popular. When the goal is scoring or ranking, simplification to a single dimension allows this to be done in a transparent way that avoids personal bias. Nevertheless, to avoid information loss, it may be wise to retain raw data for consultation alongside summary data. For example, an adjusted frequency based on frequency and dispersion could be used to select fifty words for which one then examines the underlying features to find and analyse the ten most important words.

Overall, Gries' presentation of the flaws in our measures is convincing. Nevertheless, for a long time our traditional measures have been producing ostensibly reasonable results, and Gries has suggested that integration of frequency information can often be beneficial (Gries, 2021a: 3). Consequently, further evidence may be worthwhile to bolster claims about the generality and severity of the problems. In terms of generality, the large uncertainties and correlations could be demonstrated on a wider variety of datasets—sometimes only a single corpus and research question taken from Gries' earlier writings are used. In terms of severity, while it's observed these issues affect keyword rankings and other important outputs, it's unknown to what extent this leads to incorrect research conclusions in practice.

Now we shall turn to the proposed solutions, beginning with the KLD measure. This measure stands out because it can be applied as a DM or AM. However, it is not totally unique in this regard (as Gries concedes for Theil's *U*), and it's not yet clear what tangible benefit this unification provides. In other respects, one may be forgiven for thinking the KLD is no superior to more common measures like Δ*P* and Juilland's *D*, which showed similar levels of correlation to frequency. Some additional complexities also hamper the simplicity of employing it: it cannot handle zeroes in the expected distribution without complicated smoothing, and there are two ways to choose the AM expected distribution (with noticeably dissimilar results, as below). For now, the KLD may be one of the better options, but like other measures it's highly correlated with frequency and lacks some desirable properties for a DM or AM, which may lead to inaccurate results.

For instance, Gries (p. 179) points out Tables 1 and 2 distinctly show the ditransitive construction attracts *assured* more than *offer*. However, he computes their KLD values using the row totals as the expected distribution, which determines that *assured* is attracted slightly less (0.025 vs. 0.028). If we instead use the second column as the expected distribution, *assured* is found to be attracted more (0.039 vs. 0.032), although the difference is still smaller than expected. Other examples can be found where the rankings appear incorrect for both expected distributions or when measuring dispersion.

Table 1. Frequency table cross-tabulating the ditransitive construction and the verb *assured*

		Construction is ditransitive	
		True	**False**
Verb is *assured*	**True**	10	2
	False	1,810	137,085

Table 2. Frequency table cross-tabulating the ditransitive construction and the verb *offer*

		Construction is ditransitive	
		True	**False**
Verb is *offer*	**True**	18	55
	False	1,802	137,032

Now consider the partialing out of frequency, which is based on the idea that dispersion, association, and keyness are dimensions of information orthogonal to frequencies. This lends itself to the idea that pure dispersion, association, and keyness ought not be correlated with frequency at all. While this makes some intuitive sense for evenness of spread, it's not quite so clear for association, and keyness in particular, where one often cares more about the outcome (e.g. a list of the words that are most characteristic of a corpus) than the method used (e.g. pure attraction to the corpus over a reference based on 2x2 frequency tables) (Scott, 1997: 235). Nevertheless, Gries's approach successfully fixes the issue of frequencies having wildly different theoretical ranges and thus being heavily correlated with frequency as well. This is exemplified in log-likelihood's R^2 correlation changing from 0.991 to 0.048, but in other cases the reduction is more modest: 0.916 to 0.423 for the KLD DM and 0.712 to 0.362 for t-score. Unfortunately, all the AMs continue to display narrow ranges of output for high frequencies, even if the R^2 is low.

While this approach is a very important contribution to the field, there's a need for further research to discover more sophisticated and less computationally expensive (Sönning & Egbert, 2024: 4) methods. For example, it's apparent that association/keyness values are still not comparable across frequencies, as rankings are incorrectly inverted when frequency is partialed out. This is most obvious when repelled elements are seen with higher association scores than many attracted elements (p. 218), but consideration of Tables 3 and 4 (p. 225, corrected to match the originals in Gries (2021a: 27)) shows the problem is more general. An analysis of the tables suggests *advance* is a considerably stronger collocate than *time*, concurring with the MI scores, but partialing out frequency from MI inverts their ranks (p. 225).

Table 3. Frequency table for the collocation *quick advance*

	quick	**Other**	**Sum**
advance	1	3,582	3,583

Other	2,661	98,356,074	98,358,735
Sum	2,662	98,359,656	98,362,318

Table 4. Frequency table for the collocation *quick time*

	quick	**Other**	**Sum**
Time	19	151,820	151,839
Other	2,643	98,207,836	98,210,479
Sum	2,662	98,359,656	98,362,318

Research proving the partialed-out measures work as intended will be crucial too. If popular measures were unintentionally measuring frequency much more than dispersion/association, what else could they be measuring, and how do we know they measure the feature of interest? One concern here is that Gries' approach is uniform for all measures, but it's likely some measures require a different transformation, as indicated by DMs correlated with corpus design features such as the number of corpus parts (Biber et al., 2016; Sönning & Egbert, 2024). Another sign we need to learn more about the issues at work is that each direction of association is supposed to provide unique information, but removing frequency from the column-to-row direction makes it perfectly predictable from the other direction.

Gries offers a few explanations for the correlations with frequency observed: some AMs are based on significance testing, and low frequency words are unable to attain high dispersion values. However, some correlations remain unexplained, such as those of the column-to-row AMs. Therefore, it would be interesting to explore more precisely what causes the correlations with frequency. More generally, we could do to understand why measures behave the way they do, which may aid us in showing whether a measure behaves as intended.

Lastly, it's worth noting there are simplifying assumptions to how we commonly measure dispersion, association, and keyness that are not mentioned in the book. These assumptions can greatly influence results, and we ought to consider them in how we compute our statistics. For instance, Gries presents dispersion as something that must be calculated from frequencies in the pre-defined parts of a corpus (and nothing else except the sizes of these parts). However, the way we define the parts of the corpus significantly affects the dispersion values that result (Biber et al., 2016; Sönning & Egbert, 2024). We may also wish to treat some parts as more important than others (unrelated to their sizes), and we often want to treat occurrences of an element a few words apart differently than those of different speakers in the same conversation or in separate texts within a linguistic register.

With respect to the second goal of being a textbook, only one chapter surveys existing knowledge of the dimensions and their common measures, but it succeeds with limited space, allowing the rest of the work to focus on the problems and solutions. The R teaching consists of code for calculating almost all of the results, with comments and explanation, but little in the way of exercises or programming concepts. The code included throughout makes Gries' methods transparent and reproducibility very straightforward.

In conclusion, *Frequency, Dispersion, Association, and Keyness* is an articulate presentation of a number of important problems with corpus linguistic measures and how they are used. The solutions proposed are a critical first step to addressing them and provide more avenues for research into dispersion, association, and keyness. This book is instructive for anyone using these measures to become better acquainted with them, including their problems and how we must be more careful in employing them.

Funding

This work was supported by the Economic and Social Research Council [grant number ES/Y001796/1]. Open access publication was funded through a Read & Publish deal with Lancaster University.

References

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/s0140525x01003922

Egbert, J., & Burch, B. (2022). Which words matter most? Operationalizing lexical prevalence for rank-ordered word lists. *Applied Linguistics*, *44*(1), 103–126. https://doi.org/10.1093/applin/amac030

Gries, S. T. (2015). 50-something years of work on collocations. In S. Hoffmann, B. Fischer-Starcke & A. Sand (Eds.), *Current Issues in Phraseology* (pp. 135–164). John Benjamins. https://doi.org/10.1075/bct.74.07gri

Gries, S. T. (2021a). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1–33. https://doi.org/10.1075/jsls.21028.gri

Gries, S. T. (2021b). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, *5*(2), 171–205. https://doi.org/10.1075/jsls.21029.gri

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

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning, 69*(4), 839–872. https://doi.org/10.1111/lang.12353

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Schneider, U. (2018). ΔP as a measure of collocation strength: Considerations based on analyses of hesitation placement in spontaneous speech. *Corpus Linguistics and Linguistic Theory*, *16*(2), 249–274. https://doi.org/10.1515/cllt-2017-0036

Scott, M. (1997). PC analysis of key words — And key key words. *System*, *25*(2), 233–245. https://doi.org/10.1016/s0346-251x(97)00011-0

Sönning, L., & Egbert, J. (2024). *Sensitivity of dispersion measures to distributional patterns and corpus design*. [Manuscript in Preparation]. https://doi.org/10.31234/osf.io/rz8qn

Address for correspondence

William C. X. Platt
Department of Linguistics and English Language
Lancaster University
Lancaster, LA1 4YD
United Kingdom
w.platt@lancaster.ac.uk