Tupleization

Author
Stefan Th. Gries (stgries@linguistics.ucsb.edu)

Affiliation
UC Santa Barbara & JLU Giessen

Key points

Tupleization is the notion that corpus-linguistics should evolve from the use of only a single corpus corpus statistic (e.g. a single association measure or a single dispersion measure) and instead quantify findings with a tuple of multiple, ideally orthogonally computed, statistics (e.g. quantifying association using a tuple of three numbers: one quantifying co-occurrence frequency, one quantifying association, and one dispersion. Tupleization will lead to a much greater degree of precision as well as discriminatory power.

Abstract

This overview motivates and then discusses the notion of tupleization, i.e. the idea that corpus linguists should relinquish the decades-old practice to use one-dimensional corpus statistics (such as simple dispersion or association measures) and use tuples of multiple values instead. At the same time, this article argues that the different dimensions entering into a tuple should be statistically as independent, or potentially orthogonal, as possible. The combination of both tupleization and orthogonalization avoids the huge loss of information that comes with the current practice and can massively improve both the output of, and the input to, any kind of study using corpus statistics.

Key words
corpus statistics, frequency, dispersion, association, keyness

1 Introduction

1.1 Corpus-linguistic phenomena and their quantification

Corpus linguistics is an empirical discipline based on observational linguistic data, and much corpus-linguistic work is then based on quantitative analysis of these data. For most of corpus linguistics' history, four phenomena have played particularly important roles and have, therefore, seen much interest and debate:

- **occurrence in terms of token frequency**: the notion of token frequency captures how often a particular linguistic element *E* occurs in a text, parts of a corpus (which could be recordings/files/texts making up a register, a genre, a time period, …) or a complete corpus. (Note, *E* can be of any kind: a word, a morpheme, a syntactic construction, but in the requisite corpora also a phoneme, an intonation contour, etc.)
- **occurrence in terms of dispersion/distribution**: dispersion quantifies how evenly

distributed the occurrences of *E* are over (the parts of) a text, a register/genre, a time period, or a complete corpus, to give just a few examples. For instance, an intermediately high frequency of a word *W* in a corpus *C* tells us more about *C* if the occurrences of *W* are distributed over many different parts of *C* than if they are all in one small part of *C*.

- **keyness, or association in terms of occurrence**: these indicate how much a linguistic element *E* is associated with a target corpus representative of a language (or register, genre, topic, or other units) in such a way that, for instance, knowing the relative frequency of occurrence of *E* would allow an analyst to better predict whether they are looking at the target corpus/register than if they did not know *E*'s frequency. For instance, if an analyst looked at a text *T* and found frequent occurrences of the words *definition*, *similarity*, *significance*, *discussion*, and *limitations*, then they would be more likely to guess that *T* is an academic text than if they did not know that these words were frequent in *T*.
- **association in terms of co-occurrence**: these quantify how much the occurrence of an element *E* tells us about the frequency/probability of occurrence of another element F in (one or more parts of) a corpus. For instance, if an analyst picked a random word *W* in a corpus and found that it was *hermetically*, then they would be more likely to guess (correctly) that the following word is *sealed* than if they knew *W* was not *hermetically*.

Given that nearly every corpus-linguistic study needs to consider such information – even mostly qualitative corpus studies often need to know that *E*'s frequency is not 0 or that *E* is not used by only one speaker – it is not surprising that many different measures have been proposed for several of these dimensions. With regard to dispersion, for instance, there is more than a dozen different measures, with *range*, Juilland's *D*, Carroll's D_2, and Gries's *DP* being the most widely used measures.

Many different measures have been proposed with regard to keyness, too. There, the log-likelihood score G^2 is by far the most widespread statistic, but the chi-squared statistic, the odds ratio, the relative frequency ratio, or association rules have also been applied.

It is with regard to association that the number of proposed measures is by far the highest: Pecina (2010) already documented more than 80 measures and new ones have been proposed since then. Since keyness is essentially an association-based construct, some of the most widely-used association measures are those that are also used for keyness; examples include G^2, chi-squared, and the (logged) odds ratio, but also measures like Pointwise Mutual Information *PMI*, the *t*-score, p_{FYE} (the *p*-value of a Fisher-Yates exact test), and Delta *P* have been applied in many studies.

1.2 A problem shared by many approaches and measures

Corpus-linguistic work utilizing the above kinds of measures have yielded insightful results. That being said, there is one fundamental issue that, to put it strongly, undermines especially work that wants to do more than just describe, i.e. work that is interested in advancing theories with causal mechanisms or work that is applied in the sense of entailing real-world decisions and consequences. This fundamental issue is the contradiction between two facts:

- every phenomenon in linguistics is **multidimensional** (in the sense of encompassing several qualitative and quantitative dimensions of variation) and **multifactorial** (in the sense that there are multiple causes/predictors that jointly affect the phenomenon; yet
- every one of the above statistics reduces such a multidimensional phenomenon to just a

single number: a frequency (e.g., of a construction), a dispersion score (e.g., of a word), a keyness score (e.g., of a word for a topic), or an association score (e.g., of a verb in a construction).

The conclusion that this contradiction leads to is that our corpus-linguistic quantifications (i) definitely come with a huge **loss of information** because of how multiple dimensions of information are reduced into a single numeric score and (ii) likely come with some (sometimes systematic) **bias** because many of the regularly used numeric scores have sometimes well-known statistical characteristics that make certain outcomes more likely than others. For example, and as will be discussed in more detail below, dispersion measures like range or keyness/association measures like G^2 do not just quantify dispersion and keyness/association respectively, they are also extraordinarily sensitive to frequency of (co-)occurrence. More precisely, while range is supposed to represent the dispersion of, say, a word, but the range values of all words in corpora are often correlated with the words' frequencies with R^2-values greater than 0.94 (Gries 2022c:182-183), which begs the question of how much dispersion information such values really provide especially given that other measures do not behave the same 'duplicative' way (see Gries 2024: section 4.4.1 for an adjusted version of the Kullback-Leibler divergence (*KLD*) for dispersion or *PMI*/the odds ratio for association/keyness).

Given the undesirability of information loss and systematic bias, there has been a slow but increasing push towards addressing the traditional way of corpus-linguistic quantification and the remainder of this chapter discusses one such way - tupleization.

2 Tupleization and orthogonalization

2.1 Introduction

The general notion of tupleization is a simple one, but it is one that (i) till very recently, has not been explored in much detail and systematically with much rigor and that (ii) also needs at least one other notion to work well. **Tupleization** refers to the idea that one should not use a single statistic to corpus-statistically operationalize something linguistic, but should use a tuple, i.e. a sequence or vector of values describing an entity (e.g. an element in a row of a data frame/spreadsheet) arranged in a specific order for comparability across entities; often, this just means that an entity is described by multiple values (e.g., the values in additional columns for the same row in the spreadsheet). A simple non-linguistic example is that one's health can be described – incompletely, but still heuristically quite well – by a tuple of values representing the results of a metabolic panel conducted on a patient's blood sample; the tuple could be the sequence of values indicating a patient's blood glucose level, their cholesterol values, their chloride values, etc.

Given the above, one can define the notion of tupleization as an instruction to an analyst to not just use one value to operationalize something. However, this entails another important constraint feature without which tupleization is worthless. To use another non-linguistic example, it does not make sense to try to measure someone's monetary state, or affluence, with a 3-value tuple if the three values are the value of one's assets expressed in Euros, U.S. dollars, and British pounds because, obviously, these values are extremely highly correlated. Instead, a value tuple would make more sense if it was, for instance, the value of one's assets, the value of one's debt, and one's income (all expressed in Euros), because these three dimensions jointly defining affluence can vary much more than the three currencies can: People that have the same income expressed in Euros do not differ in their incomes in the other currencies (which are deterministically related via exchange rates), but people that have the same income expressed in Euros can differ wildly in their assets as well as their debts, etc. To return to corpus linguistics and generalize, this means that we also need something Gries (2024: Ch. 3) calls **orthogonalization**, which is the idea that, for maximum meaningfulness of any tuple, the different dimensions that the values of the tuple are covering should at least theoretically be allowed to vary fairly, if not completely, independently of each other. The problem is that this is often not the case; the following sections will exemplify these problems.

2.2 Association needs tupleization/orthogonalization

With some simplification, the most widely-used association measures fall into two groups: (i) significance test-based measures like G^2, chi-squared, t, z, and p_{FYE} and (ii) effect-size measures or other heuristics like *PMI*, the (logged) odds ratio, or the *KLD*. However, these measures exhibit several of the above-mentioned problems: All of them lose information because they cannot distinguish directions of association; i.e., they are not able to see that *according* attracts *to* much more than *to* does *according*. In addition, the first group of measures is so highly correlated with frequency of co-occurrence that one cannot help but wonder whether these kinds of association measures really measure association: Gries (2022b) shows that t as a collocation statistic is predicted by logged co-occurrence frequency with an R^2_{GAM} of 0.9765, and Gries (2024: Section 4.2) shows that G^2 is predicted by logged co-occurrence frequency of verbs and the ditransitive construction with an R^2_{GAM} of >0.991. Measures of the second group fare much better in that regard, meaning they provide values that do not simply regurgitate frequency in disguise but provide valuable additional information.

2.3 Keyness needs tupleization/orthogonalization

Given the conceptual closeness of keyness and association, the situation here is quite similar. Depending to some extent on the sizes of the target and the reference corpora, keyness statistics can be extremely highly correlated with either the frequencies of the words in question in general or the difference between the words' frequencies in the target and the reference corpus. Gries (2024: Section 4.3) shows how G^2 for keyness is predicted by frequency differences with an R^2_{GAM} of >0.972 when the target and the reference corpus are similarly large. When the target and the reference corpus are quite different in size, then frequency in the target corpus is vastly more predictive of G^2 than an effect size measure of association (the logged odds ratio). In addition, there is hardly any keyness research that distinguishes directions of keyness: Does knowing the word make it easier to guess the target corpus or does knowing the corpus make it easier to guess a word's occurrence? Two rare exceptions are Pojanapunya & Todd (2018) and Gries (2024),who show that treating keyness as a single simple bidirectional measure loses discriminatory power and makes it impossible to straightforwardly distinguish the predictive power that words might have for a topic from words that are being attracted by a topic in a way that statistically corresponds to the above question of whether *according* attracts *to* or vice versa or both. There is also not much work going in the direction of tupleization – the only work coming close until very recently was work like Paquot & Bestgen (2009), which tried to incorporate at least some dispersion information by utilizing how the occurrences of word types are distributed over the target corpus.

2.4 Dispersion needs tupleization/orthogonalization

Compared to keyness and especially association, dispersion has been researched much less and, because of the nature of the measure, directionality does not apply. However, as has been shown by Gries (2022c, 2024: Section 4.1), most dispersion measures are again so strongly correlated with frequency of occurrence that their value is at least not always obvious. Using the British Component of the International Corpus of English (ICE-GB) as an example, Gries (2024) shows that, of six dispersion measures, all are predicted by mere logged frequency of occurrence with R^2_{GAM}s of between 0.82 and 0.948. Similarly, in Gries (2022c) the most widely used dispersion measure, range, is predicted by logged frequency of occurrence in the spoken part of the British National Corpus with an R^2_{GAM} of 0.961. This also leads to the problem that words with extremely different distributions but identical frequencies score dispersion values that do not reflect dispersion but only frequency. Gries (2022c) discusses the example of *enormous* vs. *staining* in the Brown Corpus. Both occur 37 times in the 1m word corpus, but while the instances of *staining* are all in a single corpus part, *enormous* manages to spread its 37 instances out over 36 corpus parts. That huge difference notwithstanding, nearly all dispersion measures indicate that both words are extremely clumpily dispersed, something that flies in the face of the fact that *enormous* spreads out its occurrences nearly as much as possible, given its frequency.

3 Tupleization and orthogonalization in practice

It follows from the above that tupleization requires two steps: operationalizing relevant linguistic, cognitive, psycholinguistic, etc. dimensions in non-redundant ways and collecting such multiple dimensions of information per data point of interest (instead of relying on just one measure). In some studies, something along those lines has been pursued. For example, in the domain of association, Durrant & Schmitt (2009) or Groom (2009) attempted something similar to tupleization by considering two numeric dimensions: they collected two association measures for each collocation in question. In the domain of keyness, Millar & Budgell (2008) considered a numeric and a binary dimension, namely the association of word types to the target corpus and their dispersion: they focused only on word types that were numerically key enough and the binary variable of whether a word type's dispersion exceeded a minimal threshold or not (which of course loses the information of how dispersed the word type is). However, a more rigorous approach along the following lines would probably be more promising.

3.1 Step 1: Orthogonalization

As mentioned above, analysts would ideally make sure that the dimensions they enter into a tuple are as independently meaningful as possible and, thus, truly contribute independently valuable information. Sometimes, a simple approach to this could be to pick measures for the dimensions that one knows or suspects are not already correlated. The above-mentioned Durrant & Schmitt (2009) or Groom (2009) did so by collecting one association measure from each class of measures: *t* from the extremely frequency-related significance test-based measures, *PMI* from the effect-size measures or heuristics. This is generally not a bad step but still sub-optimal because if the *t*-score is used to 'reflect' the role of frequency, one could of course just use (logged) frequency itself rather than a measure that is highly, but not safely deterministically, related to frequency.

Another more advanced approach can help when there is no existing measure to fall back on (as when one falls back from *t* on actual frequency or from *t* on an actual effect size measure of association such as the (logged) odds ratio). Gries (2022b, c and especially 2024) discuss examples of how to 'remove the effect of frequency' from association and dispersion measures. In the case of association, one could proceed in the following way:

- one computes the observed statistic, which is the desired association measure as usual, e.g. a *t*-score;
- one computes the highest possible association score for words with the current frequencies observed in the data and for the current corpus size, e.g. the *t*-score when the two words in questions occur as often as they can in the current corpus;
- one computes the lowest possible association score for words with the current frequencies observed in the data and for the current corpus size, e.g. the *t*-score when the two words in question never co-occur.

One then min-max transforms the three values to the interval [0, 1] and the new, transformed version of `obs` is the new association score. This is then a t-score$_{\text{nofreq}}$ 'without frequency' because it has been computed and normalized for words with the exact same frequency, which is therefore 'held constant'. Gries (2022b) shows that these association scores are then not related to co-occurrence frequency anymore.

The same logic can be applied to remove frequency from dispersion scores:

- one computes the observed dispersion measure as usual, e.g. a *KLD*-value;
- one computes the highest possible dispersion score for words with the current frequencies observed in the data and for the current corpus size, e.g. the *KLD*-value when all occurrences of the word in question occur in the smallest corpus part;
- one computes the lowest possible dispersion score for words with the current frequencies observed in the data and for the current corpus size, e.g. the *KLD*-value when the word in question is spread out as evenly a word with that frequency can be in corpus files with the given sizes.

Again, Gries's (2022c) results show that this dispersion$_{\text{nofreq}}$ score is not much related to frequency anymore but is the best dispersion score to boost the prediction of lexical decision times and the logic of removing frequency like this can be applied to many measures.

3.2 Step 2: Tupleization

Once one has multiple corpus-statistical dimensions that are as orthogonal as possible, one can apply all of them at the same time to greatly exceed the amount of information extracted from the corpus data. For example, for association data, Gries's (2019) study of verbs in the ditransitive relies on verb-specific tuples of {frequency, log odds ratio (for association), dispersion (Deviation of Proportions)}, and his results show that at least the two values of frequency and the log odds ratio are not much correlated at all and, thus, provide useful separate information to the analyst.

For keyness, Gries (2024) develops a three-dimensional notion of keyness that has a frequency, but also an association and a dispersion component (each based on the *KLD*). In other words, every word type's keyness is a tuple {frequency, association to target, dispersion across target relative to dispersion across reference corpus}. In a case study of academic writing in the Brown corpus compared to the rest of the Brown corpus he shows that this approach can do what previous approaches cannot, namely identify words that are key for academic writing because of their association or because of their dispersion or both. For example, keywords that are mostly association-based are often lower-frequency, but topically highly specialized words (such as *anode*, *bronchial*, *hypothalamic*, or *tetrachloride*, etc.) whereas keywords that are mostly dispersion-based are often higher-frequency and generally academically useful words (such as *essentially*, *types*, *basis*, *distribution*, *conditions*, *method*, *similar*, etc.). This exemplifies how distinguishing different orthogonal levels of information in a tuple can lead to findings that decades-old traditional approaches are unable to produce systematically.

4 Conclusion

As this article has shown, the general idea of tupleization is not entirely new: there are studies that capture more than one of the four main corpus statistical dimensions (recall Durrant & Schmitt 2009), and there is work that makes distinctions even within one such dimension, as when Schmid & Küchenhoff (2013) or Gries (2013) make a case for distinguishing different directions of attraction for co-occurrence phenomena. However, such approaches are not common practice yet. It is probably fair to say that (i) the inclusion of more than two dimensions and (ii) the additional requirement of orthogonalization have not been explored systematically with much rigor. Gries (2019) is among the first major exploration, which then gets extended more systematically in Gries (2024). Regardless of which dimensions are included in one's analysis and regardless of which method(s) of orthogonalization are used, more work on tupleization seems indispensable if only to address the undeniably high degree of information loss coming with the traditional one-statistic-for-everything approach.

In addition, while this short overview focused on how *outputs* from tupleization are richer than their traditional counterparts, it is worth mentioning that tupleization can also improve the *input* to corpus-statistical methods. For instance, Gries (2022d) or Ben Youssef (2024) discuss a tupleization approach to the identification of multi-word units (e.g., names such as *Los Angeles*, technical terms like *oxygen transfer*, complex prepositions like *according to*, etc.), which can then help improve the results of keyness studies. For example, a traditional keyness analysis comparing American and British English corpora might not return *White House* as a keyword for American English because the usual tokenization process might not even recognize this as a single unit to be compared across both corpora (but use *white* and *house* separately). However, in Gries (2022d) and Ben Youssef (2024), tupleization on the basis of eight different statistical dimensions is used to tokenize the corpora *before* the keyness analysis so that such units can be identified as key. Thus, such an analysis would not only see that *White House* might be a keyword for American English, it would accordingly also produce more accurate results for *white* and *house* as single-word units (because the occurrences of *White House* would correspondingly reduce the frequencies of occurrence of *white* and *house*). Tupleization (with orthgonalization) is, therefore, an extremely promising strategy for both better input and output in many quantitative corpus-linguistic applications.

References

Ben Youssef, Chadi. 2024. mMERGE: a corpus driven Multiword Expressions discovery algorithm. Ph.D. dissertation, UC Santa Barbara.

Durrant, Phil & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics* 47. 157-177.

Gries, Stefan Th. 2013. 50-something years of work on collocations: what is or should be next … *International Journal of Corpus Linguistics* 18(1). 137-165.

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

Gries, Stefan Th. 2022a. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis* 19.

Gries, Stefan Th. 2022b. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(1). 1-33. [one figure caption correction]

Gries, Stefan Th. 2022c. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171-205.

Gries, Stefan Th. 2022d. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis* 19.

Gries, Stefan Th. 2023. Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. *Cognitive Semantics* 9(3). 351-386.

Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: revising and tupleizing corpus-linguistic measures*. Amsterdam & Philadelphia: John Benjamins, pp. 324.

Groom, Nicholas. 2009. Effects of second language immersion on second language collocational development. In Andy Barfield & Henrik Gyllstad (eds.), *Researching collocations in another language*, 21-33. Basingstoke, UK: Palgrave Macmillan.

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

Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Paquot, Magali & Yves Bestgen. 2009. Distinctive words in academic writing: [...]. In Andreas Jucker, Daniel Schreier, & Marianne Hundt (eds.), *Corpora: Pragmatics and discourse*, 247-269. Amsterdam: Rodopi.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2). 137-158.

Pojanapunya, Punjaporn & Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 133-167.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: […]. *Cognitive Linguistics* 24(3). 531-577.