What do (some of) our association measures measure (most)? Association?

Stefan Th. Gries
University of California Santa Barbara, USA

This paper discusses the degree to which some of the most widely-used measures of association in corpus linguistics are not particularly valid in the sense of actually measuring association rather than some amalgam of a lot of frequency and a little association. The paper demonstrates these issues on the basis of hypothetical and actual corpus data and outlines implications of the findings. I then outline how to design an association measure that only measures association and show that its behavior supports the use of the log odds ratio as a true association-only measure but separately from frequency; in addition, this paper sets the stage for an analogous review of dispersion measures in corpus linguistics.

Keywords: association, frequency, dispersion, log-likelihood, *t*, *MI*, generalized additive modeling

1. Introduction

In some way, just about any statistic in corpus linguistics is ultimately based on frequency of occurrence and/or co-occurrence: We report frequencies of tokens and/or types per se, we use frequencies to compute dispersion measures (DMs), or we use co-occurrence frequencies to compute association measures (AMs). For each of these three dimensions of statistical information, theoretical, cognitive, and psycholinguistic research has discussed cognitive/psycholinguistic mechanisms underlying these dimensions. For instance,

- **token frequency** has been related to matters of (cognitive) entrenchment (Schmid 2010) and/or baseline activation levels in psycholinguistic models of the mental lexicon (see discussion by Baayen et al. 2016);
- **dispersion** has been considered as a proxy towards the commonness of a word (I am using *commonness* here as a 'technical term' that, while usually

https://doi.org/10.1075/jsls.21028.gri | Published online: 12 November 2021
Journal of Second Language Studies 5:1 (2022), pp. 1–33. ISSN 2542-3835 | E-ISSN 2542-3843

operationalized using frequency, is not the same as frequency, see Savický & Hlaváčová 2002) and has also been related to recency (e.g., Gries 2019a);
– **association** has been related to contingency and associative learning in, say, the Competition Model or in Ellis's CREED model (e.g. Ellis 2007a, b), but has also played an important role in second language studies or learner corpus studies as in explorations of collocational knowledge (see Ellis et al. 2008; Durrant & Schmitt 2009, Bestgen & Granger 2014, or Siyanova-Chanturia 2015 for examples).

When corpus linguists want to quantify, say, the association of two elements in a corpus or the dispersion of an element in a corpus, they have to choose what (type of) AM or DM to use simply because for both association and dispersion many different measures have been proposed. For association, Evert (2009) and Pecina (2009) alone reviewed more than in 80 measures, for dispersion, Gries (2008, 2010, 2020) reviewed and compared about a dozen or so measures, and for both domains new measures have been proposed since, which of course raises the issue of which measure(s) to choose.

One of the most central aspects that should feature in any researcher's decision for a measure is of course **validity**, which can be approached from two important yet complementary perspectives. The first perspective is concerned with the desideratum that a measure *m* should really measure what it is intended to measure; that means an AM should be designed in such a way that it measures association and a DM should be designed in such a way that it measures dispersion. There are probably few who would disagree with this seemingly trivial statement, but there is another, complementary aspect to it which is less often considered: The values of a measure should measure, or 'react to', what they are intended to measure or 'react to', but also not measure or 'react to' much else, so that we can take/interpret the computed values at face value (no pun intended).

The second perspective is concerned with the fact that the results of some such measure should ideally be correlated (well (enough)) with the kind of external evidence that the measure is supposed to measure. For example, if an AM is truly a measure of the degree to which, say, two words are associated with each other *and* if one independently assumes that the association of a word pair is related to how tightly connected the two words would be in subjects' minds (e.g. in an associative test), it follows that a good AM should also correlate with such external data (e.g. associative test data).

Interestingly enough, much work in corpus linguistics using AMs and DMs has not concerned itself enough with both of these two perspectives (maybe especially by neglecting the first – often for good reasons, see below, and I have done so myself too often), which can then also impact the second. Put differently and

to say it out loud, if one's AM or DM is not *and* not only measuring what it is supposed to measure, then we are already beginning to fail the most basic test criterion, that of validity and then it's not a huge surprise that our measure might not correlate well with the kinds of external evidence we want to correlate it with or validate it against. As just mentioned, neglecting the first perspective – measuring what one wants to measure and nothing else – has often been done for probably just one single reason: simplicity and sortability along one dimension *d*: we all like to just click "Sort" and be done with it. If one computes AMs for how much some collocates are attracted to a node word but one's AM conflates or, to put a more positive spin on it, 'conveniently integrates' information from various dimensions – hopefully with at least one of them being association – then this might be (!) sufficient for a variety of lexicographic, applied, and maybe some descriptive purposes (and for many of those purposes the second perspective might not be relevant because, for instance, lexicographers don't need external psycholinguistic validation). In fact, sometimes the conflation of measures often returns 'intuitively satisfying' results precisely because the ranking one observes is actually not so much due to the dimension of information *d* one says one is using but more due to another dimension. This happens most often when, for example, frequency 'supports' the AM/DM *m* and, thus, makes *m* return results with a treacherously high(er) appeal. And that higher post hoc interpretive appeal has often made us ignore the fact that that appeal is not so much because *m* is so great and precise at capturing the dimension *d* we imply it captures (by its name) and the results are so great precisely because dimension *d* is exactly what matters, but because *m* actually reflects more than we say it does and it is actually everything that *m* uses above and beyond *d* that makes the results seem so great. More concretely, we might be calling something an *AM* and, correspondingly, interpret its results in terms of association when, figuratively as well as statistically speaking, $^{2}/_{3}$ of what it returns is just re-packaged frequency information, same for DMs. This can even lead to the treacherous situation that corpus results based on some AM fit external evidence well mostly because of the particular AM used is actually correlated more with frequency than with association. In a way, in such a situation, it might be the fact that we are violating the first perspective (our AM is more determined by frequency than association effects) that makes the measure seem to pass the second perspective (its result correlate well with external evidence).

Again, oftentimes this conflation is not necessarily a problem: Somewhat simplistically, the more descriptive the study, the less of a problem the conflation of different dimensions causes. But, as soon as the goal is more linguistic, theoretical, and/or psycholinguistic in nature than the simplest of descriptions, however, this kind of threat to validity becomes problematic and then addressing both perspectives is becoming more and more relevant: With interpretive goals, we need

'clean'/precise diagnostic tools (measures) – not tainted/conflating ones – that we can then maybe also relate to external evidence.

In this paper, the first one of a 'two-paper paper', I want to discuss the notion of association and how it is often computed and then used in corpus linguistics. I will focus on what I consider the most widely-used AMs, the log-likelihood value (G^2, see Dunning 1993), but also pointwise Mutual Information (*MI*) and the *t*-score, and I will focus specifically on the question of how cleanly they actually measure association and just association. I will argue that nearly at least two of the most widely-used AMs – G^2 and *t* – are problematic precisely in the sense that they are not 'clean' at all: They do not only measure association but also frequency; in fact, they react more to frequency than they do to true association (which I am using in the sense of 'quantifying contingency'), and in the sister publication to this paper, I will argue that the same is true of nearly all dispersion measures. Now, the fact that especially the *t*-score is correlated quite strongly with frequency is of course well-known: Stubbs (1995: 36) already stated that "*T* takes into account mainly (in some cases only) the absolute frequency of joint occurrence of node and collocate" and Thanopoulos et al. (2002) state that "the *t*-score produces exactly the same hits (ranked slightly different) as plain frequency" (see also Siyanova-Chanturia 2015: 153). However, the extent to which G^2- and *t* do not reflect association is usually not explicitly stated in the discussion of results, and it is even less often expressly quantified and, maybe because of that, both G^2 and *t* are still widely seen as perfectly valid measures of association.[1]

This study pursues the following goals. First, I want to shed some more light on the behavior of the most widely-used AMs: G^2 in general corpus linguistics and the frequent combination of *MI*/*t* in second language/learner corpus studies. I will begin with G^2 and show that, in some sense and for some applications, it is not really a good AM in how it combines frequency and association (rather than reflect association only) and in how that can lead to quite counterintuitive findings (using an example of G^2 in keywords analyses, which, statistically at least, are just association studies). Using generalized additive modeling, I will then extend that discussion to a small collocation case study (speed adjectives in the BNC) and show that not only does G^2 not just combine frequency and association, but that (i) it also reflects frequency more than it does association and that (ii) the degree to which it does reflect association is actually non-linearly

1. Also, some statements in the literature about properties of AMs are maybe a bit confusing. For instance, Siyanova-Chanturia (2015: 153) cites Hunston (2002: 73) as saying that "*MI* is not dependent on the size of the corpus, and is thus good for both larger and smaller corpora", but Stubbs (1995: 34) says that "[*MI*] […] also takes into account the size of the corpus."

dependent on frequency, making proper interpretation of G^2 in terms of association even harder. I will then apply that same methodology to *t* and *MI* to evaluate to what degree each of these measures reflects frequency and association, followed by an interim summary/discussion of the findings and what they imply. The final section will then develop a new approach to measuring association that is guaranteed to not be tainted by frequency; this part will (i) ultimately turn out to validate the (log) odds ratio as a true association-only measure, (ii) propose that the general approach to be discussed here is maybe of much more general importance and applicability, and (iii) set the stage for a similar discussion of the problems of nearly all DMs in Gries (2022). In order to make it easier for people to follow along or apply the logic of this paper to their own work, the exposition below will regularly provide R code; note, however, that understanding the R code is not required to understand the paper and readers unfamiliar with R can feel free to gloss over the code – the code is really only meant as help for readers who might want to program the proposed measures.

2. The conflation of frequency and association

2.1 Hypothetical data and G^2's behavior

2.1.1 *A collocation/collostruction example*

To introduce how measures that supposedly measure dimension *d* (e.g., association/contingency or dispersion/commonness) can actually return values that, to large extent, reflect something else, I will use the AM called the log-likelihood value or G^2. G^2 has been in wide use ever since Dunning (1993) because it is supposed to be better at handling the kind of low expected frequencies we often face in corpus linguistics as a result of the Zipfian distributions of linguistic elements and probably also because it scored well in influential papers such as Evert & Krenn (2001).[2]

How is G^2 computed? Most people do so from co-occurrence tables, which can be schematically represented as in Table 1.

2. Note that G^2, or any other traditional AM for that matter, does not address another major problem of such data, namely the fact that the observations summarized in the usual 2×2 tables are not independent of each other.

Table 1. Schematic co-occurrence table of a word *w* and a construction *c*

	Construction: *c*	**Construction: other**	**Sum**
Word: *w*	*a*	*b*	*a+b*
Word: other	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

Let's define a hypothetical table of observed corpus results like `table.01.obs` as follows:

```
addmargins(table.01.obs <- matrix(c(50, 950, 350, 9998650), ncol=2,
   dimnames=list(WORD=c("w", "other"), CONSTRUCTION=c("c", "other"))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w       50     350      400
##   other  950 9998650  9999600
##   Sum   1000 9999000 10000000
```

Most people compute G^2 from the observed data in `table.01.obs` and the corresponding expected frequencies (expected from H_o, that is), which we can compute as follows:

```
(table.01.exp <- chisq.test(table.01.obs, correct=FALSE)$expected)
##        CONSTRUCTION
## WORD          c      other
##   w        0.04     399.96
##   other  999.96 9998600.04
```

From that, G^2 is computed as represented in the usual formula:

$$G^2 = 2\sum_{a}^{d} observed \times log\,\frac{observed}{expected} = 622.2269$$

```
2*sum(table.01.obs*log(table.01.obs/table.01.exp))
## [1] 622.2269
```

Alternatively and conceptually actually more generally, we can also compute G^2 from a binary logistic regression model that tries to predict the occurrence of the word *w* from the presence or absence of the construction *c* (or vice versa, because G^2 is bidirectional).

```
glm(table.01.obs ~ colnames(table.01.obs), family=binomial)$null.deviance
## [1] 622.2269
```

While G^2 is always referred to as an AM, its output is far from just that. A true association-only measure should react to association only, but the fact that G^2 reacts to frequency *and* association rather than just to association can be illustrated straightforwardly in two ways.

First, we can see how G^2 increases when only the sample size increases even though all ratios in the table stay the same (e.g., it is still the case that *w* occurs in and outside of *c* at a 1-to-7 ratio). Consider table `table.02.obs`, which is the result of merely increasing the frequencies of `table.01.obs` by a factor of 10 without changing the contingency:

```
addmargins(table.02.obs <- table.01.obs * 10)
##        CONSTRUCTION
## WORD        c    other       Sum
##   w       500     3500      4000
##   other  9500 99986500  99996000
##   Sum   10000 99990000 100000000
glm(table.02.obs ~ colnames(table.01.obs), family=binomial)$ null.deviance
## [1] 6222.269
```

In other words, that 10-fold increase of G^2 is a reaction to frequency, not to association, because the association was not changed. And note that association-only measures such as the (log of the) odds ratio (a.k.a. the (exponentiated) slope of the above glm; here I'm using a discounted version of the odds ratio, which involves adding 0.5 to every frequency first) or ΔP (Ellis 2007a, Gries 2013) do not increase in the same ways and, therefore, reflect only association/contingency:

```
# the results for table.01.obs
round(c("Log odds ratio"=log(odds.ratio(table.01.obs)), "Delta
Ps"=delta.ps(table.01.obs)), 4)
##       Log odds ratio Delta Ps.Delta P 1.w Delta Ps.Delta P 2.c
##               7.3236               0.1249               0.0500
# the results for table.02.obs
round(c("Log odds ratio"=log(odds.ratio(table.02.obs)), "Delta
Ps"=delta.ps(table.02.obs)), 4)
##       Log odds ratio Delta Ps.Delta P 1.w Delta Ps.Delta P 2.c
##               7.3164               0.1249               0.0500
```

Second, we can see that G^2 reacts to frequency *and* association by seeing how this measure increases when only the frequency of the word *w* increases even though its distribution across *c*/not *c* (100 to 700 in `table.03.obs`) remains at the same 1-to-7 ratio as the 50 to 350 ratio in `table.01.obs`:

```
addmargins(table.03.obs <- matrix(c(100, 950, 700, 9998250), ncol=2,
dimnames=list(
   WORD=c("w", "other"), CONSTRUCTION=c("c", "other"))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w      100     700      800
##   other  950 9998250  9999200
##   Sum   1050 9998950 10000000
glm(table.03.obs ~ colnames(table.01.obs), family=binomial)$ null.deviance
## [1] 1239.451
```

Of course, we find a similar increase of G^2 when only the frequency of the construction *c* increases even though its distribution across *c*/not *c* (100 to 1900 in `table.04.obs`) remains at the same 1-to-9.5 ratio as the 50 to 950 ratio in `table.01.obs`:

```
addmargins(table.04.obs <- matrix(c(100, 1900, 300, 9997700), ncol=2,
dimnames=list(
   WORD=c("w", "other"), CONSTRUCTION=c("c", "other"))))
##        CONSTRUCTION
## WORD       c   other      Sum
##   w      100     300      400
##   other 1900 9997700  9999600
##   Sum   2000 9998000 10000000
glm(table.04.obs ~ colnames(table.01.obs), family=binomial)$ null.deviance
## [1] 1258.769
```

Again, other measures that reflect association only are not affected as much at all:

```
# the results for table.03.obs
c("Log odds ratio"=log(odds.ratio(table.03.obs)), "Delta
P1"=delta.ps(table.03.obs)[1])
##       Log odds ratio Delta P1.Delta P 1.w
##             7.319296             0.124905
# the results for table.04.obs
c("Log odds ratio"=log(odds.ratio(table.04.obs)), "Delta
P2"=delta.ps(table.04.obs)[2])
##       Log odds ratio Delta P2.Delta P 2.c
##             7.47270336         0.04996999
```

2.1.2 *A keyness example*

The facts that (i) G^2 reacts to both frequency and association and that (ii) it seems to react more to frequency than to association can also lead to results that, when one looks at such a 2×2 table just intuitively, seem extremely *counter*intuitive. For instance, keyness scores in keywords analyses are often computed using G^2 (simply because keywords analyses are just associations of a word not to a another word or construction, but to a corpus). If one does a keywords analysis on the Clinton/Trump Corpus to identify the words that are (strongly) characteristic of, or key for, Hillary Clinton's campaign speeches compared to Donald Trump's campaign speeches, one will find the following frequency distributions for the words *hillaryclinton* (as part of the phrase hillaryclinton.com) and the word *about* (cited from Gries 2021):

Table 2. The distribution of *hillaryclinton* in the Clinton-Trump corpus

	Clinton corpus	**Trump corpus**	**Sum**
hillaryclinton	26	0	26
other words	117263	445730	562993
Sum	**117289**	**445730**	**563019**

However, I doubt many people would look at these two tables and say, "ah, obviously a clear case of two words with the same association to, or keyness for, the Clinton corpus" … Yet, if one has committed to thinking that association or keyness is what G^2 measures, then that is what one would have to conclude

Table 3. The distribution of *about* in the Clinton-Trump corpus

	Clinton corpus	Trump corpus	Sum
about	579	1386	1965
other words	116710	444344	561054
Sum	**117289**	**445730**	**563019**

because these completely different distributions return nearly exactly the same G^2-values (81.6 and 81.7), and they do so in spite of the facts that

- they come with completely different odds ratios (≈201.5 for Table 2 vs. ≈1.6 for Table 3);
- they come with completely different proportional reductions of error:
  - if I show you only the first row of Table 2 and then ask you, "I have a sentence here that contains *hillaryclinton*, whose speeches is that sentence from?" you will of course 'guess' "Clinton" and you would be right, but
  - if I show you only the first row of Table 3 and then ask you, "I have a sentence here that contains *about*, whose speeches is that sentence from?" you will of course 'guess' "Trump", but you are likely to be wrong.

This shows that one has to be extremely cautious in interpreting such values given how G^2 conflates two separate dimensions of information and it again seems as if, at least sometimes, G^2 reacts more to frequency than to what it is claimed to reflect, association; in the following section, we will discuss this on the basis of actual data.

2.2 Actual data and G^2's behavior

Let us look at a fairly straightforward collocation application, namely nominal collocates in the R1-slot of the adjective *fast* in the BNC. I retrieved all instances of *fast* from the BNC that were followed by a word token whose POS-tag begins with *NN* and computed for all *fast*+N collocations their G^2 and their log odds ratio; G^2-values for collocations observed less often than expected were multiplied by –1 to reflect the repulsion relation.

The following three panels of Figure 1 show some of the relations between G^2 (left panel), the log odds ratio (as a more likely true association-only measure, center panel), and (logged) co-occurrence frequency (right panel): Each word type is a grey point and the grey lines summarize the relations with generalized additive models (GAM), whose R^2-values are also provided in the plot.

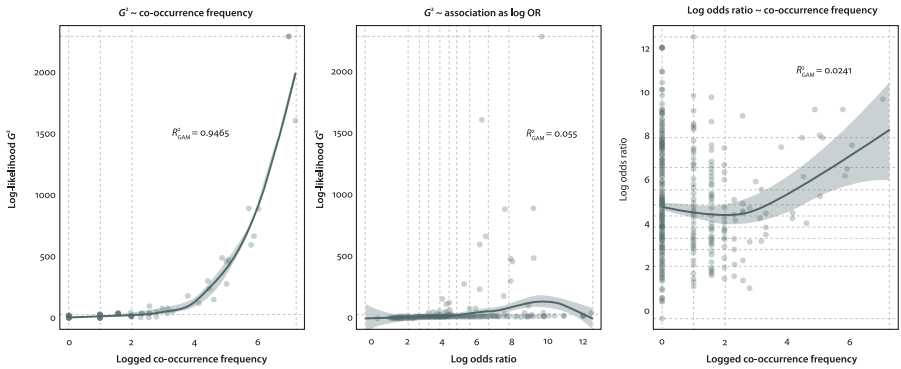

Figure 1. G^2 as a function of frequency and association (separately)

The findings are very obvious:

- the left panel shows that the 'AM' G^2 is nearly perfectly (curvilinearly) predictable from just the logged co-occurrence frequency;
- the center panel shows that the 'AM' G^2 is hardly at all predictable from what it supposedly measures, namely association as identified by the log odds ratio;
- the right panel shows that the association-only measure log odds ratio in turn is hardly at all predictable from logged co-occurrence frequency (as it should be, given that combinations of high or low association with low or high frequency respectively are perfectly conceivable, even if, and this is a crucial point, G^2 is by design hardly able to identify low frequency-high association collocations).

The results for three other speed adjectives are for all intents and purposes the same; all are summarily represented in Table 4.

Table 4. Predictability R^2_{GAM}s for G^2, the log odds ratio, and logged co-occurrence frequency

	R^2: G^2 ~ log freq	**R^2: G^2 ~ log odds ratio**	**R^2: log odds ratio ~ log freq**
fast	0.9465 (from above)	0.055 (from above)	0.0241 (from above)
quick	0.9625	0.028	0.0061
rapid	0.983	0.0273	0.0077
swift	0.9335	0.0589	−0.003

This case can be made even more clearly when we consider the results of regressing G^2 on both logged co-occurrence frequency and the log odds ratio (now on the data for all 4 speed adjectives combined so we have different adjective

frequencies in the data) to see what drives G^2-values more/most. Amazingly enough, even though the measures from all four adjectives were lumped together and the model was not told which adjective each triple of values (G^2, log OR, and logged frequency) belongs to, this GAM explains the G^2-values nearly perfectly: $R^2 \approx 0.95$. But the more interesting finding is of course that the logged co-occurrence frequency affects G^2 much more than association does, although we are always saying that G^2 is an AM. In Figure 2, the predictors are on the axes (frequency on the *x*-axis and association on the *y*-axis) and the excellent predictions of the G^2-values (recall the high R^2) are indicated with numbers that represent the predictions in 10 bins (0: lowest predicted G^2-values, 9 highest predicted G^2-values):

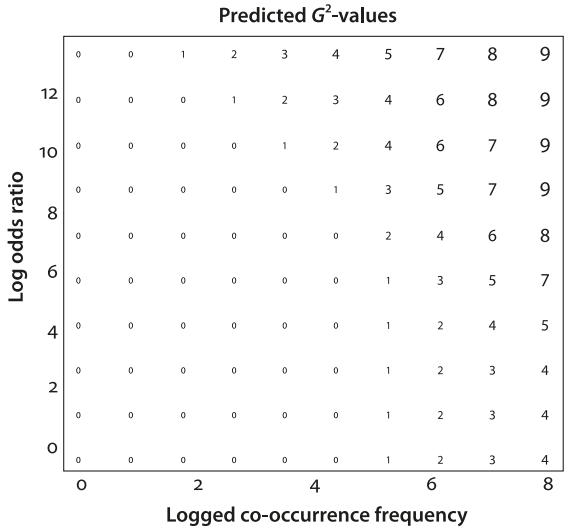

Figure 2. G^2 as a function of frequency and association (combined)

The nature of the effects is obvious and strong: Of course, G^2-values are predicted to be highest when both frequency and association are high (in the top right corner). But then,

– when co-occurrence frequency is low or intermediate (roughly in the left half of the plot), increasing association (by moving up in the plot) has only a small corresponding effect on the predictions (increasing them from 0 to between 0 and 3);

- when co-occurrence frequency is higher (in the right half of the plot), increasing association has a moderate effect on the predictions (increasing them by 4 or 5 (from, e.g. 1 to 5 or 3/4 to 8/9)); but now
- when association is low (in the lower half of the plot), frequency already has an effect on the predictions (increasing them from 0 to between 4 and 7);
- when association is high, frequency has a maximal effect on the predictions (increasing them from 0 to 8 or mostly 9).

This has two important consequences. The first is that what we call an 'AM', G^2, is a measure that is in fact much more responsive to co-occurrence frequency than it is to association. Pedantically speaking, one might even wonder whether G^2 should still be called an AM, given that *most* of what it reflects is actually not association. One might also wonder why, if these 'AMs' don't reflect association as much as their names suggest, the results they provided have often been (perceived as) insightful and why these 'AMs' performed well in some studies (e.g., Evert & Krenn 2001). One possibility is that this is because the G^2-values do reflect a bit of what one says one wanted (association) but also a lot of something else that is also often interesting though not what we say we aim for (frequency), so collocations with high G^2-values will have a high frequency and, therefore they are likely to also be more evenly dispersed in the corpus/language, which raises the probability of them being recognized or at least feeling familiar (both by linguists and collocation raters). In other words, if an AM reacts to frequency a lot, its results may seem more appealing from the second perspective above (external evidence and interpretability) precisely because of all the effects that high frequency might have or come with (recognizability, familiarity, dispersion) even though the AM then does not only measure what it is theoretically supposed to measure (undermining the first perspective). The perceived simplicity and utility of G^2's results reflecting two dimensions simply made us ignore the fact that the results were actually not mostly/exclusively reflecting the one dimension of information we said we were using and, thus, not strictly speaking valid; put differently, we're happy to interpret G^2's results as association even if they only are as good and interpretable as they are because G^2 mostly reflects frequency.

The second and more intricate consequence is that the above plot shows that G^2 is potentially even more misleading than what the previous paragraph suggested. That is because the plot shows that even the degree to which G^2 reflects association is not constant across co-occurrence frequency. Evert & Krenn (2001) already usefully separated high- and low-frequency collocations, but here we see a clearly graded nature of such an effect. It's not like what G^2 reflects is, say, $^2/_3$ frequency and only $^1/_3$ association, period – no, the plot above shows that with low

frequencies, G^2 does not co-vary much with association at all, with intermediate frequencies it does a bit more, and with high frequencies it then also reflects association most strongly, but, heuristically speaking, still only half as much as it does frequency! Put differently, the same increase in association (moving up the *y*-axis) does nothing when co-occurrence frequency is low but then more and more as co-occurrence frequency rises (moving right along the *x*-ais). Thus, even the smaller degree to which G^2 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.

Note that that also means that one must not compare the results of a collocational/collostructional study using an 'AM' such as G^2 (or $p_{\text{Fisher-Yates}}$ or *t*) against an association-only measure like the log odds ratio because that would mean comparing the results of a measure that pretends to use one dimension of information (association) but actually uses some combination of two against one that is a clean and precise measure of just one dimension. The results of the latter kind of measure *will* look 'less satisfying' because they might not contain enough familiar high-frequency expressions, but again, at least those results are interpretationally valid and speak to association. Thus and along the lines of Gries (2019b), G^2 would need to be compared to each word's tuple of {pure frequency, pure association}, which could be visualized as follows on the basis of data for *fast*:

This way, the fact that the highest association-only score for *fast* is found for *sealynx* is 'contextualized' for the analyst by its low co-occurrence frequency. At the same, this having two clean measures as opposed to one opaquely conflated one like G^2 also allows us to see how much information that latter one loses. For instance,

- the three red words in Figure 3 have virtually identical G^2-values – 43.8, 45.2, and 42.4 for *sealynx, chargers*, and *roads* respectively – although they clearly differ immensely in their distributional behavior with *fast* – in what way, with what contorted definition of *association*, does it make sense to say that *fast* is as associated with *roads* as it is with *sealynx*?
- the two blue words in Figure 3 have quite different G^2-values – 667.2 and 598.5 for *track* and *cars* respectively – although they clearly are very similar in terms of their distributional behavior with *fast*.

(As a side remark, Figure 3 also seems to reinforce the importance of dispersion: I am thinking most people would intuitively agree that *fast food* is probably a more interesting/noteworthy collocation than *fast bowler* (and probably more useful to teach to the admittedly rare language learner who doesn't already know that

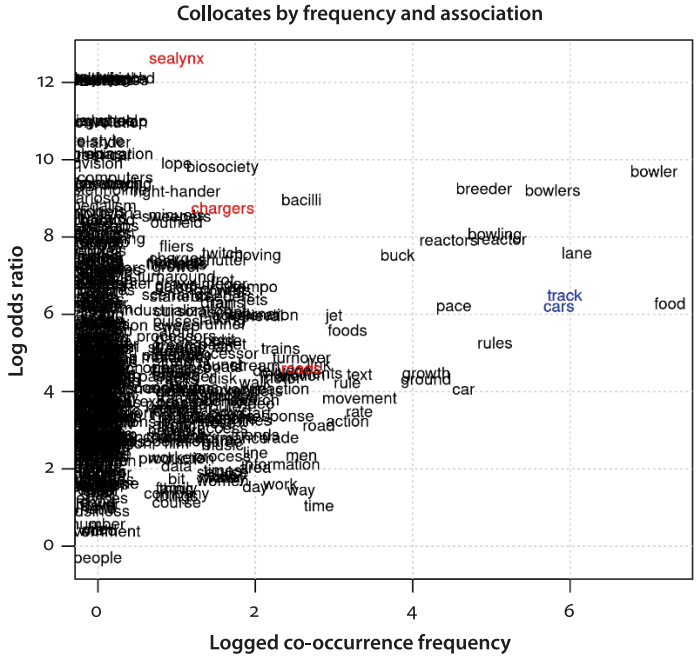

Figure 3. Collocates of *fast* by frequency and association

expression) but we can see that the latter is nearly as frequent as the former and exhibits a quite a bit stronger degree of attraction. However, the intuitively higher degree of importance we would attach to *fast food* would be recognized corpus-statistically if we checked the dispersion of the two collocations in the BNC: Using even just the primitive measure of *range* for the moment, while *fast food* is only a bit more frequent than *fast bowler*, it occurs in more than twice as many different files; this is something we will return in the sister publication to this paper on dispersion.

Having seen how, in a validity sense, poorly G^2 fares makes one wonder how well the other two most widely-used measures score when we look at them in terms of their validity as AMs; therefore, the next two sections will apply the above logic to *t* and *MI*.

2.3 Actual data and *t*'s behavior

Another frequently used measure is the *t*-score, which is often used together with the *MI*-score (e.g., Durrant & Schmitt 2009, Groom 2009, Siyanova-Chanturia 2015); that is because it is often said that the *t*-score returns frequent collocations and that *MI* returns infrequent collocations. What does the *t*-score reflect and

how consistently does it do so? Figure 4 shows the results for all 4 speed adjectives combined:

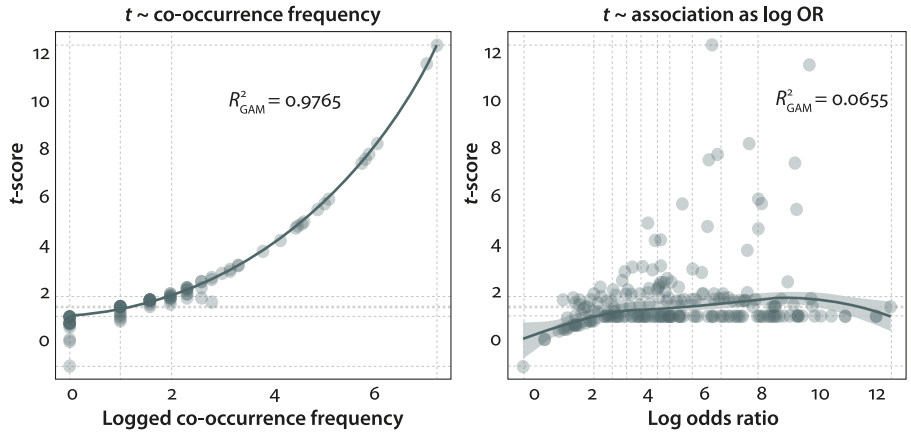

Figure 4. *t* as a function of frequency and association (separately)

As is clear, the *t*-score is even more perfectly (curvilinearly) related to the logged co-occurrence frequency (!) than G^2, and *t* again reflects association very little. Why the exclamation mark? Because it is important to realize that what *t*, like G^2, reflects is the co-occurrence frequency of the two words, not their overall frequencies. Stubbs (1995: 36f., 39) got that right, but, for instance, Bestgen & Granger (2014: 31) seem to have not: They state that "[the] *t*-score, which measures collocations *composed of very frequent words*" (my emphasis) and that "[the] *t*-score [...] brings out those [word sequences] *composed of high-frequency words*", (2014: 30, my emphasis). This is minimally misleading on two counts. First, because the *t*-score can *measure* any kind of collocation of any kinds of words; after all, we *can* compute a *t*-score for two words each of which occurs one time, namely with the other – what Bestgen & Granger presumably misexpress with "measure" is 'return high values for', but even corrected like that this is still wrong, because, to me, both quotes seem to imply that *t* is mostly affected by the marginal totals (i.e., the frequencies of the collocates in general) rather than the co-occurrence frequency of the two words. But the left panel of Figure 4 indicates that *t* reflects (logged) co-occurrence frequency (remember, $R^2_{GAM} = 0.9765$) and a GAM regressing *t* on the frequencies of the collocates in general, i.e. what Bestgen & Granger imply *t* does, returns an R^2_{GAM} of < 0.01 for *fast alone* and an R^2_{GAM} of < 0.025 for all four speed adjectives combined. Thus, *t* does not bring out [collocations] composed of *high-frequency* words in general, but collocations with a *high co-occurrence frequency* – that is not the same. And that of course is

why Evert & Krenn (2001) find that the *t*-score and G^2 perform fairly similarly with each other as well as with frequency and return the best results for their *high-frequency* PNV and AdjN data, or why Durrant (2014) finds very similar correlations between frequencies and *t*-scores from COCA and learner knowledge: G^2 as well as *t* are so highly correlated with each other (they can be more than 92% predicted from each other) and with co-occurrence frequency that they just don't offer much information above and beyond co-occurrence frequency, and they actually offer very little on association (see again the right panel of Figure 4). Thus and as before with G^2, a GAM regressing the *t*-scores on logged co-occurrence frequency and the logged odds ratio (for all 4 speed adjectives combined) supports this assessment with findings that are like those for G^2 but indeed even more extreme, as is shown in Figure 5.

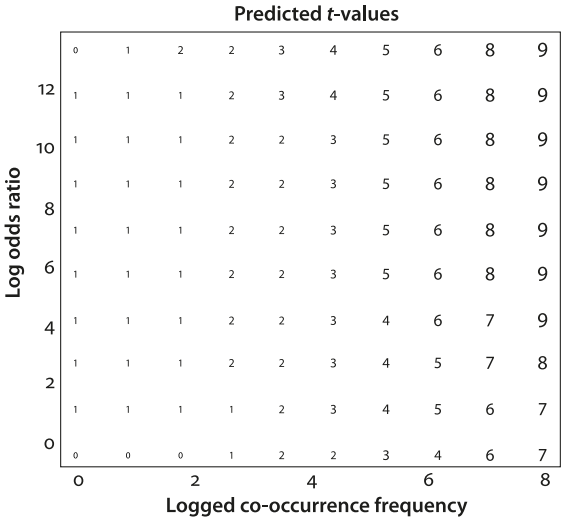

Figure 5. *t* as a function of frequency and association (combined)

Thus, in a sense, the *t*-score seems to deserve the label *AM* even less than G^2: it largely varies as a function of co-occurrence frequency and reflects actual association to just a small degree.

2.4 Actual data and *MI*'s behavior

What about *MI*? Siyanova-Chanturia (2015: 153) states that "*MI* is [...] not so strongly linked with raw frequency as other AMs" but the present results make this even seem a bit understated: In the present data (all 4 speed adjectives com-

bined), *MI* hardly reflects frequency at all and is pretty much identical to the log odds ratio:

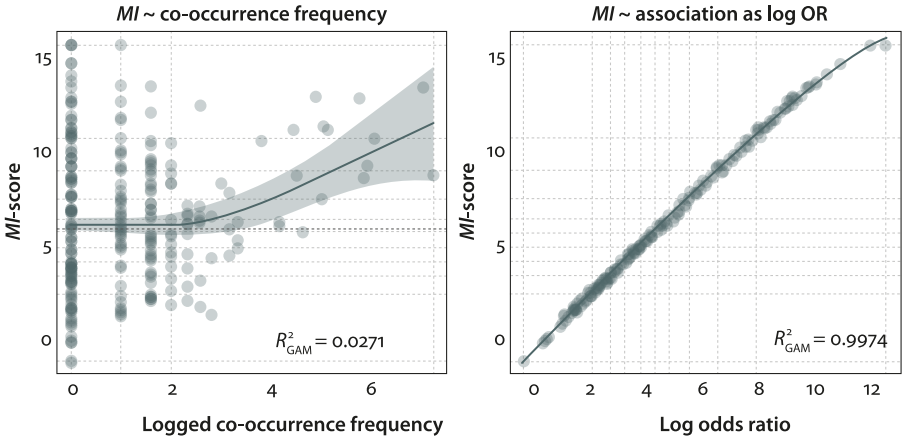

Figure 6. *MI* as a function of frequency and association (separately)

Note again some definitional confusion in Bestgen & Granger (2014: 31): "MI score, which measures collocations made up of infrequent words". This is again wrong for the same reason as above: *MI* can *measure* any kind of collocation – are we supposed to believe that one can not compute an *MI*-score for a collocation of frequent words?! It seems as if "measure" is again misused to mean 'return high values for'. But Figure 6 also suggests that even the frequent sentiment that *MI* returns low-frequency collocations ("MI, which tends to highlight word sequences made up of low-frequency words", Bestgen & Granger 2014: 30) is really not uncontroversially true: *MI* does not only return low-frequency associations. If that was the case, one might expect a noticeable correlation between *MI* and frequency such that, as co-occurrence frequency increases, *MI* decreases), but that is not what we find. For instance, for some reason, scholars often consider *MI*-scores of ≥ 3 as evidence of collocation,[3] but the left panel of Figure 6 clearly shows that there are many frequent collocations even with *MI*-scores twice as high, i.e. $MI \geq 6$ (represented by the dashed horizontal line at $y = 6$); in fact, the mean co-occurrence frequency of collocations is significantly *higher* for colloca-

3. The motivation of $MI \geq 3$ as a threshold value is not completely clear to me (and Section 6 will criticize the use of such universal cut-off points). Durrant & Schmitt (2009: 168), for instance, cite Hunston (2002) and Stubbs (1995) for this value; Stubbs in turn cites Church & Hanks (1993) for this, but the latter do not provide much of a justification for that value: why would observed frequencies need to be exactly eight times higher than expected frequencies for an 'interesting/noteworthy' collocation?

tions with $MI \geq 6$ than for collocations with $MI < 6$ rather than lower (W from a U-test = 719524, $p_{\text{2-tailed}} = 0.0035$)!

The fact that *MI* is much more affected by association than frequency is also reflected in the results of a GAM trying to predict *MI* from logged co-occurrence frequency and the log odds ratio: The GAM predicts the *MI*-scores perfectly and, as is obvious from Figure 7, the predictions vary solely as a function of association (along the *y*-axis), not as a function of frequency (along the *x*-axis):

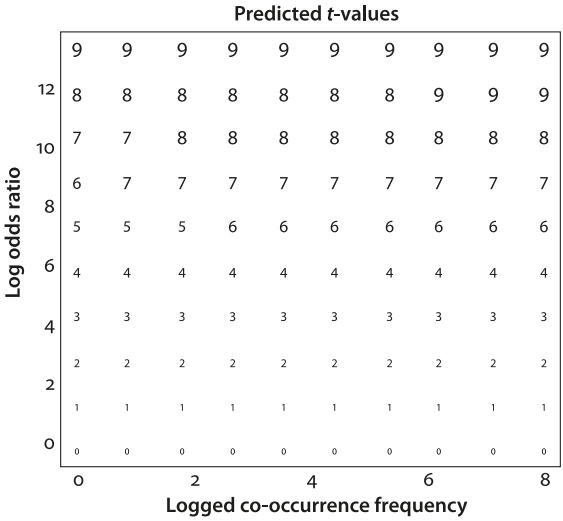

Figure 7. *MI* as a function of frequency and association (combined)

3. Interim discussion

3.1 Some general remarks

G^2, *t*, and *MI* are all called association scores and are all widely used in that capacity. However, G^2 and *t* at least do not really seem to deserve the name much – at least if we like calling spades spades – given that they reflect, or react to, co-occurrence frequency much more than they do to association. *MI*, on the other hand, is much closer to being a real AM since association is what it reflects. This has some implications regarding how people deal with and talk about these and other AMs.

First, obviously it means that one must be careful in terms of how much we want to interpret results from higher G^2- and *t*-values as strong associations because *t* reflects mostly frequency and very little association and, for G^2, the

degree to which it reflects association varies as a function of frequency. Clearly, this does not bode well for the validity of considering G^2 and *t* AMs. This is not to say they don't return something that can be useful – but association-only, that it is not. If readers think this is too harsh a statement, they need to ask themselves the following question: Would they be happy if they went to a lab to have their blood checked for their cholesterol level, gave a sample, paid the lab, and were sent an email with the sentence "Your HDL level is *x*" but then they find out that the value *x* they are given is only correlated with their HDL value with an R^2_{GAM} of 0.1 but it reflects their blood glucose level really well with an R^2_{GAM} of 0.9? I doubt they would. Yes, that value is also interesting from a general health perspective – just as frequency is generally interesting for many (corpus-)linguistic applications – but it's not quite the same now, is it?

Second, this in turn means that scholars who argue that they use both *MI* and *t* because they return different kinds of collocates (Durrant & Schmitt 2009, Bestgen & Granger 2014) – the former low-frequency ones, the latter high-frequency ones – are actually doing something subtly different from what they imply: Not only did we see in Figure 6 that *MI* can clearly return frequent collocations – the associations in case just need to be strong enough – but they are also not really using two AMs with different properties, one of which specifically targets/returns/filters low-frequency collocations and the other specifically targets/returns/filters high-frequency collocations. Rather, they are using (i) one AM (*MI*), which is nearly unrelated to co-occurrence frequency and can return higher-frequency collocations, and (ii) one other measure that is little else but a, so to speak, 'transformed-frequency measure' (*t*).

Now, I am aware that some of the above may seem like a pedantic distinction without a difference, but much ink has been spilled on discussing the pros and cons and performances of different AMs so, clearly, there is an interest in how these measures perform, which should include an interest in whether AMs actually measure association. I can't imagine that we want to be using terminology like *frequency* and *association*, which have both descriptive uses as well as theoretical implications, in a way that flies in the face of validity: If a measure is called an AM, it should not mostly reflect something else and association only to a degree to which that something permits in a statistical interaction (as in Figures 2 and 5). Of course, some scholars might now retreat to the position that they simply have a definition of *association* that is different from mine: Mine defines *association* in statistical terms of 'contingency' only and, thus, mostly orthogonally to mere co-occurrence frequency, while theirs is some amalgam of frequency and association/contingency … However, I would find that hard to accept for two reasons: First, because I have yet to see a paper that uses, say, G^2 or *t* and explicitly

admits that what they return is really mostly frequency and a small but actually not precisely-defined/weighted quantity of association – no, authors that use those AMs couch their discussion in terms of association. Second, because if one does not keep these two dimensions separate, one not only incurs the information loss exemplified with Figure 3 for G^2 above, but, because of that, one also deprives oneself of the possibility of exploring all possible combinations of frequency and association: If one's AM is so extremely correlated with frequency, then that practically means

- high frequency will by definition mean high 'association';
- low frequency will by definition mean low 'association';
- highvfrequency-low association and low frequency-high association collocations/collostructions will hardly ever be found.

Put differently, other than for the most practical/applied settings (which of course do exist), I can't see how it is in anyone's interest to label a measure an AM, implying by the name that it is something different from frequency, but then define it mathematically in such a way that it is pretty much inseparable from frequency.

3.2 *MI*, MI^2, and MI^3

The above discussion should already imply what I think of measures such as MI^2 and MI^3. These heuristic measures are computed by changing the numerator of the basic *MI* computation to the co-occurrence frequency to the power of 2 or 3 respectively, One sales pitch for this 'strategy' is to give a greater weight to the co-occurrence frequency, but what this really is is taking an AM that measures association nearly perfectly orthogonally to frequency (as it should) and then intentionally diluting it to make it reflect another dimension more (and more). Figure 8 shows how this manifests itself in practice: The upper three panels regress *MI*, MI^2, and MI^3 (on the *y*-axes) against logged co-occurrence frequency and we can see that the resulting 'association scores' of course correlate more and more with frequency; the lower three panels regress *MI*, MI^2, and MI^3 (on the *y*-axes) against the log odds ratio and we can see that the resulting 'association scores' of course correlate less and less with association. It does not seem reasonable to 'improve' the performance of an AM by making it reflect association less and unpredictably less and make it reflect another quantity more … Again the question for sceptical readers: would they be happy to be told "Hey, you didn't like the previous 'cholesterol value' we sent you? How about this: you can now do our

new and improved cholesterol level test: We worked hard on it and now it reflects cholesterol even less than before!"?[4]

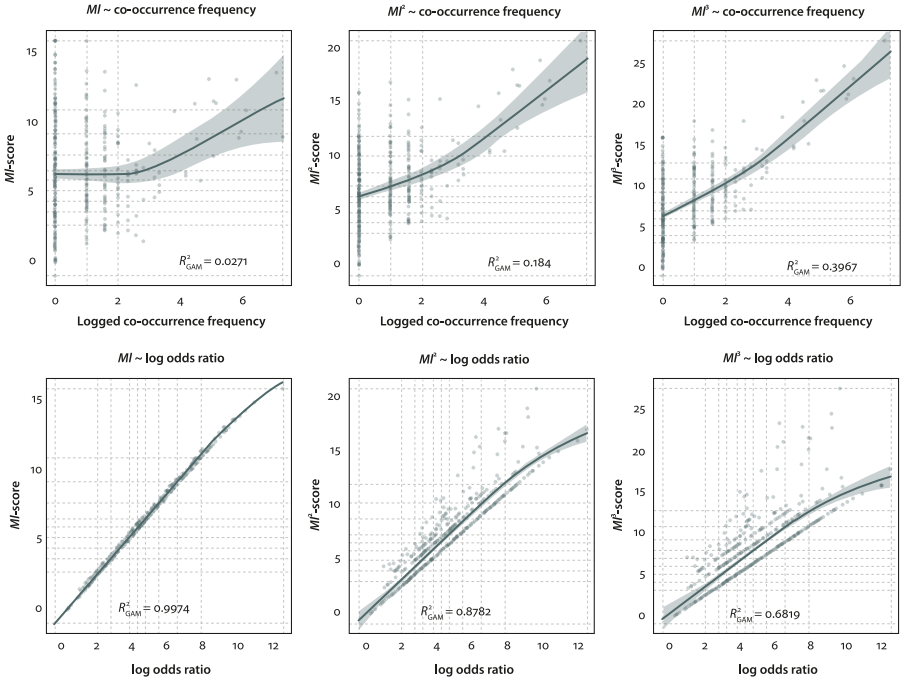

Figure 8. *MI*-, *MI*²-, *MI*³-scores as a function of frequency and association

3.3 A brief comment on (log) Dice

Another AM that is sometimes used (and that is implemented in, for instance, BNC Web) is the Dice score. This score is somewhat interesting in the present context because it is sometimes used in its raw form and sometimes in a logged version (the logging changes the distribution of the raw Dice values into a nearly normal distribution). Interestingly, as Figure 9 shows,

– the unlogged version of Dice is *more* strongly correlated with frequency than it is with association (the log odds ratio);

4. *MI*² has at least some theoretical justification because, unlike *MI, MI*² does not decrease with higher numbers of perfectly predictive co-occurrences, but I do not think that that outweighs the conceptual benefits resulting from keeping frequency and association as two conceptually clean and orthogonal measures.

– the logged version of Dice is *less* strongly correlated with frequency than it is with association (the log odds ratio):

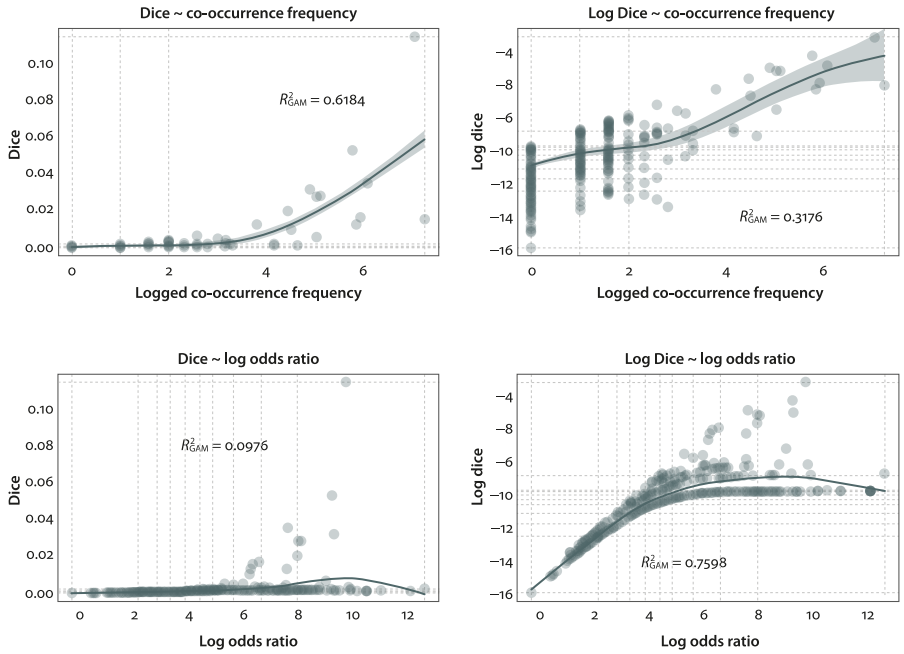

Figure 9. Dice/Log Dice as a function of frequency and association

Thus, at least the logged version of the Dice coefficient is much closer to being a 'true' AM than the more widely used G^2 or *t*-scores.

4. A new measure

4.1 Motivation and development

Much of this paper has been about the importance of (i) having measures that really mostly or even only measure what they are supposed to measure (validity) and (ii), thus, keeping dimensions of information clean *and* separate/orthogonal lest we conflate what can be very different kinds of information (see Gries 2019b for extended discussion). In this discussion, I routinely treated the (log) odds ratio as a pure association-only measure. That is probably in line with what many readers hopefully recall from basic statistical training, and I demonstrated in Section 2.1.1 that the (log) odds ratio is not affected by increases in the frequencies of the whole 2×2 table or the row/column sums $a+b/a+c$. In what fol-

lows, however, I also want to derive an AM that is *by design* not correlated with frequency to

- show that this new measure is not only not correlated with co-occurrence frequency but also very highly correlated with the measure that I have treated as an association-only measure, the (log) odds ratio (thus validating it) and to
- already set the stage for the sister publication on dispersion, where we are in a similar but worse situation: researchers using something they say is a dispersion measure when in fact their measures reflect frequency more than dispersion but where no apparent gold-standard like the (log) odds ratio for association is available and, thus needs to be developed.

Developing an AM that controls for any effect frequency might have on it involves the following main logic: We quantify the association for a certain collocation in the data (let's call this value `obs`). Then, we take the frequencies of the two co-occurring elements in question (i.e. the totals $a+b$ and $a+c$) and the corpus size (i.e. $a+b+c+d$), hold them constant (which virtually eliminates any way in which co-occurrence frequency can unduly boost/lower the resulting association-only measure), and *then* we determine

- the lowest possible association possible given the values we are holding constant (let's call this `low`); and
- the highest possible association possible given the values we are holding constant (let's call this value `upp`, for 'upper limit').

Because these maximal-attraction and minimal-attraction/maximal-repulsion values will exhibit different ranges (due to the marginal totals, see the examples below and Section 5), we then transform/normalize these three values (`low`, `upp`, and `obs`) such that they fit into the interval [0,1], and our new association-without-frequency measure becomes the value that corresponds to `obs` in that [0,1] interval.

Let's exemplify this with two small examples. We begin with the simplest possible case and use the collocation *rapid growth* and its distribution as shown here:

Table 5. The co-occurrence of *rapid* and *growth* in the BNC

	growth	**Other**	**Sum**
rapid	243	2463	2706
other	12545	98347067	98359612
Sum	**12788**	**98349530**	**98362318**

First, we compute obs to quantify how much *rapid* 'attracts' the noun *growth*; let's, for simplicity's sake, do that simply with the conditional probability $p(growth|rapid)$, which is $^{243}/_{2706}$, i.e. 0.0898004. Second, we compute the two most extreme distributions we might find given the actual frequencies of *rapid* and *growth*, which is how we control for frequency (by holding it constant) so it cannot affect the AM we are developing. Because of the distribution here, this particular case is straightforward:

- one extreme result would be that all 2706 instances of *rapid* are followed by *growth* (i.e., the distribution in the first row of Table 5 would become 2706 : 0); this is the upper limit of association strength that is possible given the frequencies of *rapid* and *growth* so we'll call this value upp;
- the other extreme result would be that none of the 2706 instances of *rapid* are followed by *growth* (i.e., the distribution in the first row of Table 5 would become 0 : 2706); this is the lower limit of association strength that is possible given the frequencies of *rapid* and *growth* so we'll call this value low:

```
## $'biased towards top left / a'
##        NOUN
## ADJ     growth    other      Sum
##   rapid   2706        0     2706
##   other  10082 98349530 98359612
##   Sum    12788 98349530 98362318
## $'biased towards top right / b'
##        NOUN
## ADJ     growth    other      Sum
##   rapid      0     2706     2706
##   other  12788 98346824 98359612
##   Sum    12788 98349530 98362318
```

Thus, we compute the same conditional-probability measure we used to compute obs for these two hypothetical distributions, which is trivial here: For the first table, upp becomes $p(growth|rapid)$, i.e. $^{2706}/_{2706} = 1$; for the second table, low becomes $p(growth|rapid)$, i.e. $^{0}/_{2706} = 0$.

The final step is to take the three measures we computed and transform them into the interval [0,1] such that

- the smallest value of the 3-element vector (typically, low) becomes 0 (if it isn't already);
- the largest value of the 3-element vector (typically, upp) becomes 1 (if it isn't already);
- the last value of the 3-element vector (likely obs) becomes whatever corresponds to its proportional position in the [0,1]-interval.

This computation here is trivial because the three values already constituted that [0,1] interval, which means that the final value of our new AM remains what we

already computed as obs. However, that will not always be the case, as we can see in the case of *rapid spread*:

Table 6. The co-occurrence of *rapid* and *spread* in the BNC

	spread	Other	Sum
rapid	16	2690	2706
other	1674	98357938	98359612
Sum	**1690**	**98360628**	**98362318**

For this distribution, obs is $p(spread|rapid)$ is $^{16}/_{2706} = 0.0059128$ and the most extreme distribution against co-occurrence is of course again this one, where $p(growth|spread)$ would be $^{0}/_{2706} = 0$:

```
##        NOUN
## ADJ     spread    other      Sum
##   rapid      0     2706     2706
##   other   1690 98357922 98359612
##   Sum     1690 98360628 98362318
```

2706/2706(

However, the most extreme distribution in favor of the co-occurrence cell *a* does not again yield $^{2760}/_{2760} = 1$ because *a*+*c* cannot exceed 1690; instead, it is this one and upp, therefore, becomes $^{1690}/_{2706} = 0.6245381$:

```
##        NOUN
## ADJ     spread    other      Sum
##   rapid   1690     1016     2706
##   other      0 98359612 98359612
##   Sum     1690 98360628 98362318
```

(Of course we would not do these computations manually; I am using a function most.extreme.2by2.tables that takes as input a 2×2-table and that returns as output those two most extreme distributions with the same marginal totals.) Now, low is 0, upp is 0.6245381, and obs is 0.0059128, which means the 0–1 transformation changes obs to 0.0094675, which is the value that we note down as our AM.

```
(what.we.have.4.spread <- c("low"=0/2706, "upp"=1690/2706, "obs"=16/2706))
##         low         upp         obs
## 0.000000000 0.624538064 0.005912786
zero2one(what.we.have.4.spread)
##         low         upp         obs
## 0.000000000 1.000000000 0.009467456
```

Again, the point of this is to determine for each observed frequency distribution (i) what the theoretically possible extreme results are while the frequencies of the collocating items are not allowed to vary from the observed corpus results (which is what keeps frequency in check), (ii) normalize them to the [0,1] interval (because low and upp will not always be 0 and 1 already), and then (iii) see

where, within that 0–1 normalized continuum of theoretically possible results, the observed result falls. (That of course also has the pleasant side effect that this new measure by definition falls into the [0,1] interval, which is useful for some applications.)

4.2 Application to *fast*

Let's now look at how these association-without-frequency values are related to logged co-occurrence frequency and to association as measured by the log odds ratio when we look at *fast* and its collocates in the BNC:

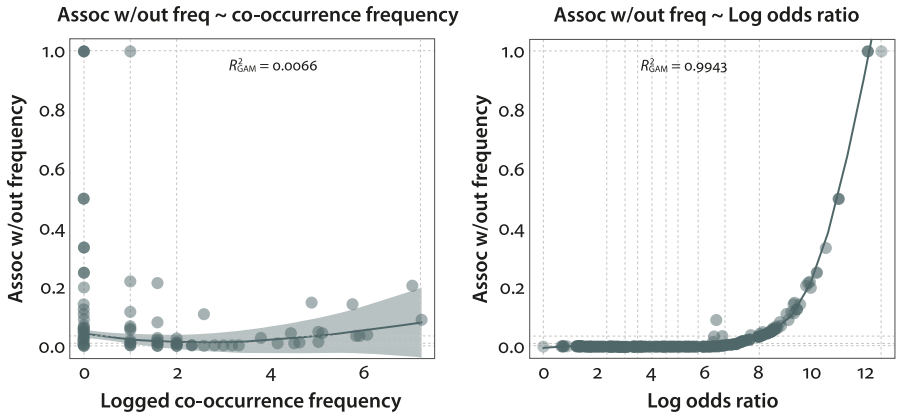

Figure 10. Association without frequency and its relations to frequency and association

The left panel shows that the new association-without-frequency measure behaves just as one would have hoped for, given its design: It is hardly related to frequency at all because it's computed for each word type by checking the range of possible results *while holding frequencies constant*, which explains the low R^2. Reassuringly, in the right panel we now also see that the new measure is virtually perfectly related to the (log) odds ratio, which means that, to measure association-only, we *could* use the new measure, but might as well stick with the simpler and more established alternative of the (log) odds ratio, which we have now seen behaves exactly like a measure that was designed to control frequency. For readers with enough statistical expertise, this characteristic of the (log) odds ratio might have been obvious, but I am fairly sure it wasn't for everyone and recall, again, that this process of designing a measure that eliminates the effect of frequency, which works of course for every measure, not just the conditional probabilities used here (see below), also sets the stage for the sister publication on dispersion where such an approach is in fact not just perhaps didactically useful but in fact

required to establish a gold standard dispersion measure – the corpus-linguistic study of dispersion does not yet have a gold standard like what the (log) odds ratio is for association.

5. A small excursus

In this brief excursus, I want to just emphasize the utility of the above approach on a more general level, where by "the general approach" I mean the notion of

- computing an observed value;
- computing the largest and smallest theoretically possible values given the marginal totals;
- relativizing the observed value against the theoretically possible range.

This is because this logic should make us, minimally, weary to uncritically use absolute cut-off points for AMs (like the widely-used value of 3 for *MI*. Why is that? Consider Tables 7 and 8 for the 2×2-tables for *quick advance* and *quick time*:

Table 7. The co-occurrences of *quick* and *advance* in the BNC

	advance	**Other**	**Sum**
quick	1	2661	2662
other	3582	98356074	98359656
Sum	**3583**	**98358735**	**98362318**

Table 8. The co-occurrences of *quick* and *time* in the BNC

	time	**Other**	**Sum**
quick	19	2643	2662
other	151820	98207836	98359656
Sum	**151839**	**98210479**	**98362318**

What happens if you compute *MI* for each of them? You obtain *MI* = 3.37 for *quick advance* and *MI* = 2.21 for *quick time*; obviously, using the traditional cut-off point of *MI* = 3, the former is 'interesting', the latter is not. However, what are the theoretically possible ranges of *MI* for these two collocations?

- we saw above that the lowest possible values would arise if, given the marginal totals, the two collocations were never observed. Since the log of 0 is not defined, let's apply the discounting logic from the odds ratio to just the *a* cell

and assume that observed *a* in both cases was 0.1; then the lowest possible *MI*-values for *quick advance* and *quick time*, given their marginal totals, are 0.04 and −5.36 respectively.
– we saw above that the highest possible values would arise if, given the marginal totals, the two collocations were observed as often as the less frequent of the two collocates is. Thus, the highest possible *MI*-values for *quick advance* and *quick time*, given their marginal totals, are 14 and −5.36 respectively.

```
(what.we.have.4.quick.advance.mis <- c("low"=0.04, "upp"=14.74, "obs"=3.37))
##   low   upp   obs
##  0.04 14.74  3.37
zero2one(what.we.have.4.quick.advance.mis)
##       low       upp       obs
## 0.0000000 1.0000000 0.2265306
(what.we.have.4.quick.time.mis <- c("low"=-5.36, "upp"=9.34, "obs"=2.21))
##   low   upp   obs
## -5.36  9.34  2.21
zero2one(what.we.have.4.quick.time.mis)
##       low       upp       obs
## 0.000000 1.000000 0.514966
```

This looks quite different: While the absolute value of *MI* for *quick advance* is higher than that for *quick time* and higher than 'the threshold' of 3, we can now see that given the range that *MI* can theoretically cover for each of *advance* and *time*, the observed *MI*-value is in fact more on the high side for *time* than it is for *advance*. This can also be shown visually, as in Figure 11 with one panel for *quick advance* (left) and one for *quick time* (right). On the *x*-axis, we see each possible value that might be in the *a*-cell of the possible 2×2-tables (from one extreme table to the other). As before, we can see that the *MI*-value for *quick advance* is higher than 3 and that that of *quick time*, but we now also see that the whole curve/range of possible *MI*-values given the marginal frequencies of *quick* and *advance* (indicated with the grey shading) is much higher than the whole curve of possible *MI*-values given the marginal frequencies of *quick* and *time* (also indicated in grey).

Thus, the fact that *quick advance* scores a higher *MI*-value than *quick time* is not just due to the association of the collocations, but also due to the range of values that is even just possible (based on the frequencies of the two words involved), but if we just stick to a universal cut-off value of $MI = 3$, then we don't see that 3 can mean different things, given the possible ranges of values for, here, just two collocations. I will leave it up to future research to examine whether, maybe, observed AMs in a study should always be contextualized against their possible range of values – if readers find this approach weird, I invite them to consider the fact that the probably most widely-used R^2-value in corpus linguistics – Nagelkerke's R^2 from logistic regression – is precisely that: it is the R^2 that results from adjusting Cox & Snell's R^2 to fall into the 0–1 range. In other words, this thought process is not

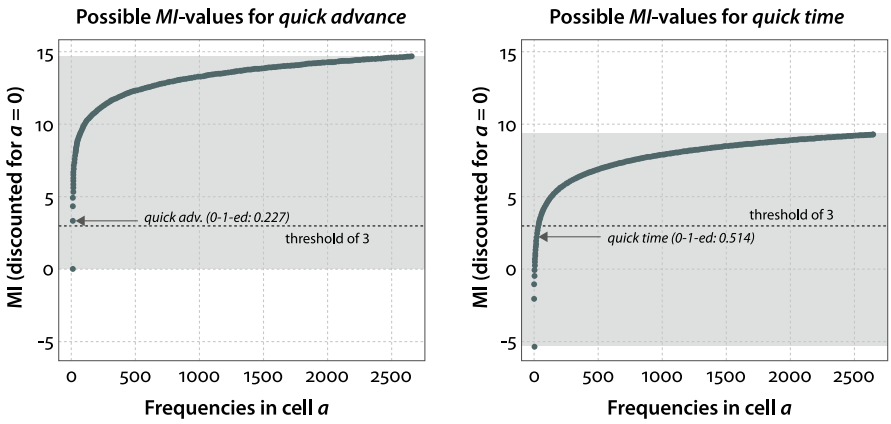

Figure 11. Contextualizing *MI* against its range

nearly as alien as it might seem but, again, I will leave more detailed exploration of this to future research.

6. Concluding remarks

While this was a long discussion and a lot of results, I hope to have shown that the notion of AM at least as used in many studies is problematic in different ways. Sometimes, properties of AMs are just described in misleading or even incorrect ways (recall the quotes on *MI*/*t* from Bestgen & Granger 2014) but often the problem is more fundamental and concerned with what many studies are calling AMs to begin with: Many studies use measures – G^2 in general corpus linguistics and *t* in learner corpus / SLA studies – and these measures react more to frequency than they do to association and they do so in inconsistent ways. This is problematic in how it might affect one's interpretation of one's results, if that interpretation involves more than just saying 'here's the top *n* elements'. For theoretical and/or psycholinguistic studies, for instance, results based on AMs might make us develop our theories such that they explain results in terms of association/contingency (because we think we are using a measure of association) when, if we use G^2 or *t*, we should instead make our theories explain the results in terms of frequency. One area where this might be particularly relevant is theories of first/second language acquisition because, there, we are dealing with speakers who may have much less input, but where association might still be very high so that co-occurrence provides strong cues (in the Competition Model sense of the term). If the co-occurrence of *x* and *y* is rarer in our data but highly or perfectly predictive (e.g., when *x* occurs, *y* does, too), measures such as G^2 and *t* might not return *xy*

as a notable co-occurrence because nearly all they see is *xy*'s relative rarity. Thus and as a reminder, association-only measures make it straightforward to identify all four possible combinations of things: (i) high frequency and high association, (ii) high frequency and low association, (iii) low frequency and high association, and (iv) low frequency and low association simply because the association scores they return are not by definition extremely correlated with frequency already – measures such as G^2, t, and p_{FYE} make that same thing very difficult.

Some researchers at least might counter this by pointing out that, in their areas/work, this is addressed by using two AMs, *MI* and *t*. And it is true that this can help in terms of retrieving collocations. Nevertheless, the (admittedly more abstract) problem remains that, minimally, that is terminologically sloppy or muddy in a way that we would probably not let our students get away with. To reiterate, using *MI* and *t* is really not so much using two AMs that react to 'different kinds of association', it's using one AM (*MI*) and then some derivative of co-occurrence frequency (*t*) and it would be more precise and better in terms of methodological validity to use co-occurrence frequency directly as one dimension and association ((log) odds ratio, *MI*, or the new measure defined above) as the other, as in Figure 3 above. There is in fact published precedence for using frequency and *MI* such as Ellis et al. (2008), Siyanova-Chanturia (2015), and Gries (2019b), who all show that this kind of clear separation of dimensions of information leads to interpretable and interesting results. For instance, Ellis et al. (2008: 381) found that the *MI*-scores of candidate formulae had a stronger effect than their frequencies on instructors' ratings of the 'chunkiness quality', cohesiveness of meaning, and worthiness of teaching; similarly, *MI*-scores of candidate formulae, not their frequencies, were found to predict native speakers' reaction time in a grammaticality judgment task and their voice onset times in a reading-aloud task. However, learners' reaction times and voice onset times were better predicted by frequency, not *MI* (p. 383, 385). This way, each measure measures what it purports to measure, which is much better in terms of terminological clarity and subsequent interpretation (than as when one pretends that the *t*-score is a good AM), and it is also much better than trying to 'fix' a good association-only statistic (*MI*) by injecting into it another dimension of information that should really be kept separate.

Now it's of course possible that someone reads this and says "But G^2 is nicely correlated with my external evidence (which has an association component) so why would you tell me not to use it as an AM?" One way to answer this is as follows: Ok, go ahead and use it, but don't pretend that your G^2 results reflect mostly association. Your *presentation* of the results of course has to state that your G^2 results are nicely correlated with the external evidence – but your *interpretation*

of the results should probably not (mostly) proceed on the basis of notions such as association, contingency, etc., because that is not what G^2 actually reflects most. And because G^2 reacts to a mix of (a lot of) frequency information and (some) association information, you won't even know how much of each contributes to your nice correlation. If, on the other hand, you separate the dimensions out as suggested here, then you can actually be more discernible and see what it is that is responsible (most) for your nice correlation.

Hopefully, the above discussion can provide some food for thought on what (some of) our favorite 'AMs' really do and how looking at two potentially orthogonal dimensions (as in Figure 3) counters information loss, especially because this exact problem of validity – the fact that, unlike what about 50 years of publications on AMs might make one expect, many of them reflect frequency more than they do association – is even greater when it comes to dispersion, i.e. the degree to which an element is distributed evenly throughout a corpus. That issue is the topic of the follow-up paper to this one and will require the development of a dispersion-measure gold standard for future work. But dispersion is not only interesting from that methodological perspective, it is also interesting for how it affects all corpus-linguistic statistics involving frequency and/or the keyness/association scores discussed here because considering frequency or keyness/association without dispersion runs the risk of getting misled by high(er) frequencies or association scores that are, however, extremely clumpily distributed in a corpus; we saw a hint of that above in the discussion of *fast food* vs. *fast bowler*, something to which we will return in the subsequent paper on dispersion.

References

Baayen, R. Harald, Petar Milin, & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiaology 30*(11). 1174–1220. https://doi.org/10.1080/02687038.2016.1147767

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

Chruch, Kenneth W. & Patrick Hanks. 1993. Word association norms, mutual information, and lexicography. *Computational Linguistics 16*(1). 22–29.

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

Durrant, Phil. 2014. Corpus frequency and second language learners' knowledge of collocations. *International Journal of Corpus Linguistics 19*(4). 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Durrant, Phil & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *Internationak Review of Applied Linguistics 47*. 157–177. https://doi.org/10.1515/iral.2009.007

Ellis, Nick C. 2007a. Language acquisition as rational contingency learning. *Applied Linguistics 27*(1). 1–24. https://doi.org/10.1093/applin/ami038

Ellis, Nick C. 2007b. The Associative-Cognitive CREED. In Bill VanPatten & Jessica Williams (eds.), *Theories of second language acquisition: an introduction*, 77–95. Mahwah, NJ: Lawrence Erlbaum.

Ellis, Nick C., Rita Simpson-Vlach, & Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly 42*(3). 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja. Kytö (eds.), *Corpus Linguistics: An International Handbook*, Vol. 2, 1212–1248. Berlin & New York: Mouton de Gruyter.

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. https://doi.org/10.3115/1073012.1073037

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. https://doi.org/10.1057/9780230245327_2

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13*(4). 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, 197–212. Amsterdam: Rodopi. https://doi.org/10.1163/9789042028012_014

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. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, Stefan Th. 2019a. *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Leiden & Boston: Brill. https://doi.org/10.1163/9789004410343

Gries, Stefan Th. 2019b. 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics 24*(3). 385–412. https://doi.org/10.1075/ijcl.00011.gri

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. Berlin & New York: Springer. https://doi.org/10.1007/978-3-030-46216-1_5

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: using frequency, and now also dispersion. *Research in Corpus Linguistics 9*(2). 1–33. https://doi.org/10.32714/ricl.09.02.02

Gries, Stefan Th. 2022. What do (some of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Pecina, Pavel. 2009. Lexical AMs and collocation extraction. *Language Resources and Evaluation 44*(1–2). 137–158. https://doi.org/10.1007/s10579-009-9101-4

Savický, Petr & Jaroslava Hlaváčová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Schmid, Hans Joerg. 2010. Entrenchment, salience, and basic levels. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, 117–138. Oxford: Oxford University Press.
Siyanova-Chanturia, Anna. 2015. Collocation in beginner learner writing: A longitudinal study. *System* 53. 148–160. https://doi.org/10.1016/j.system.2015.07.003
Stubbs, Michael. 1995. Collocations and semantic profiles: on the cause of the trouble with quantitative methods. *Functions of Language* 2(1). 23–55. https://doi.org/10.1075/fol.2.1.03stu
Thanopoulos, Aristomenis, Nikos Fakotakis, & George Kokkinakis. 2002. Comparative Evaluation of Collocation Extraction Metrics. Paper presented at LREC 2002.

Address for correspondence

Stefan Th. Gries
Department of Linguistics
University of California Santa Barbara
Santa Barbara, CA 93106-3100
United States
stgries@gmail.com

Publication history

Date received: 14 July 2021
Date accepted: 13 October 2021
Published online: 12 November 2021