15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures)

Stefan Th. Gries
University of California, Santa Barbara &
Justus Liebig University Giessen

Abstract
This paper discusses a variety of potential shortcomings of most of the most widely-used association measures as used in collocation research and collostructional analyses. To address these shortcomings, I then discuss a research program called *tupleization*, an approach that does away with the usual kinds of information conflation by keeping relevant corpus-linguistic dimensions of information – e.g., frequency, association/contingency, dispersion, entropy, etc. – separate and analyzing them in a multidimensional way; I conclude with pointers towards how these dimensions could, if deemed absolutely necessary, be conflated for the simplest kinds of of rankings as well as strategies for future research.

1 Introduction

Approximately 15 years ago, Anatol Stefanowitsch and I 'developed' a family of methods that became known as *collostructional analysis* (CA). These methods are all based on what is maybe the most fundamental of corpus-linguistic assumptions, the distributional hypothesis. Corpus linguists usually cite Firth's (1957:11) famous dictum "[y]ou shall know a word by the company it keeps," but I think Harris's (1970:785f.) statement makes the same case much more explicitly:

> [i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

The reason for the single quotes around *developed* above is that the basic idea of CA is really only an extension of the notion of *collocation*, i.e. the co-occurrence of (most often) two lexical items, to one sense of the notion of *colligation*, namely the co-occurrence of words with patterns (Hunston & Francis 1999) or constructions (Goldberg 1995, 2006). The family of methods of CA distinguishes three different approaches:

> *collexeme analysis* (Stefanowitsch & Gries 2003), whose purpose it is to quantify how much words that occur in a syntactically defined slot of a construction are attracted to or repelled by that construction; examples include the verbs that occur in the ditransitive, the imperative, or the nouns that occur in the *accident*$_{N}$ *waiting-to-happen* construction;
> *distinctive collexeme analysis* (Gries & Stefanowitsch 2004a), whose purpose it is to quantify how much words prefer to occur in slots of two functionally similar constructions; examples include the ditransitive vs. the prepositional dative or the *will*- vs. the *going-to* future (an extension of this approach, multiple distinctive collexeme analysis, extends this to >2 functionally similar constructions);

co-varying collexeme analysis (Gries & Stefanowitsch 2004b, 2005), whose purpose is to quantify how much words in one slot of a construction are attracted to or repelled by words in a second slot of the same construction; examples include the two verb slots in the *into*-causative (*trick*$_{V1}$ *someone into buying*$_{V2}$ or *force*$_{V1}$ *someone into accepting*$_{V2}$) or the verb and the preposition of the *way*-construction (*weave*$_{V}$ *your way through*$_{Prep}$ *the crowd* or *make*$_{V}$ *your way to*$_{Prep}$ *the top*).

As nearly all measures of association/contingency, these methods are all based on 2×2 co-occurrence tables of the type schematically represented in Table 1 and only differ in the nature of the two elements:

for collexeme analysis, element 1 in the rows might be a word (e.g., *give*), element 2 in the columns a construction (e.g., the ditransitive);
for distinctive collexeme analysis, element 1 is a word (e.g., *give*), the columns feature the two similar constructions (e.g., the ditransitive and the prepositional dative);
for co-varying collexeme analysis, element 1 is a word in one slot of the construction (e.g., *force*), element 2 is a word in another slot of that construction (e.g., *accepting*).

One can then compute any association measure (AM) for the 2×2 table; in the context of CA, the most widely-used measure is the *p*-value of a Fisher-Yates exact test (p_{FYE}) followed by, and very highly correlated with, G^2 (the log-likelihood ratio, which is computed as shown in (1)).

(1) $2\sum_{a}^{d} obs \quad log\frac{obs}{exp}$, where expected values are computed from row and column totals

Table 1: Schematic co-occurrence table for AMs

	element 2	other elements	Sum
element 1	*a*	*b*	*a+b*
other elements	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

CA has been widely used in a variety of applications: on different languages, with native speakers (see references cited above) as well as foreign- or second-language speakers (Gries & Wulff 2005, 2009, Deshors 2010, Ellis, Römer, & O'Donnell 2016), with synchronic and diachronic data (Hilpert 2012a, b, Gyselinck 2018), in theoretical linguistics, applied linguistics (Schmid & Ungerer 2011, Matthys 2014), and in psycholinguistic applications (Gries 2005, Szmrecsanyi 2006, Wiechmann 2008, Bernolet & Colleman 2016, Ellis, Römer, & O'Donnell 2016) etc. On the whole, it is probably fair to say that it has mostly yielded instructive results. This is because of the distributional hypothesis and the fact that the identification of the words in the relevant slots – usually called *collexemes* – is often done semi-manually, avoiding much of the 'noise' that would result from using a blunter window/span approach that is blind to syntactic structure. However, there are of course also ways in which the method is perhaps (much) less than ideal simply because CA inherits most of the problems of the traditional association-measures approaches to collocation.

First, CA has been criticized for the fact that most practitioners have been using the

association measure mentioned above (Schmid & Küchenhoff 2013), p_{FYE}, because that means CA inherits potential problems of the null hypothesis significance testing (NHST) paradigm including, importantly, the fact that *p*-value-based collocational statistics conflate both effect size and sample size. Second, AMs computed from such tables ignore type frequencies – e.g. how many different word types are attested in a construction? – and frequency distributions – e.g. what are the co-occurrence frequencies of other types with and without a construction in question? Third, most widely-used AMs are bidirectional/symmetric, i.e. in the case of CA, they quantify only the mutual attraction of the two elements involved and do not distinguish directions of attraction/repulsion. Finally, CA and all other association measures are sensitive to underdispersion, i.e. they can return misleading results when the co-occurrences captured in tables such as Table 1 are concentrated in small parts of a corpus.

The current paper makes a set of suggestions to address these problems. Some of these suggestions have been hinted at in previous work (e.g., Gries 2012, 2015), but for the most part those papers discussed these issues only from a theoretical perspective and to address misconceptions about CA. In this paper, however, I will show how these can be address not only in a principled way, but also practically and with implications for *any* kind of work based on association measures. This will be relevant for usage-based corpus-based construction grammarians and applied linguists alike because refining the method will allow us to more precisely identify the words that are most central to constructions' slots. Not only will be practically relevant (because we have better data for applications to diachronic change, language processing, language teaching, etc.), but it will also be theoretically/cognitively relevant, because the methodology proposed here is cognitively more comprehensive: we know that frequency and association are not the only things that matter for processing, learning, acquisition, etc., and the present proposal adds additional cognitively relevant dimensions to CA. In each subsection of Section 2, I will first briefly discuss the problem and then propose suggestions for improvement theoretically, which will then be collated and practically exemplified in Section 3; Section 4 concludes.

2 Problems of association measures and towards the solution of tupleization

2.1 AMs and the conflation of frequency and effect size of association/contingency

One issue discussed heatedly with regard to collostructional analysis is that of which association measure to use. Bybee (2010: Ch. 5) or Schmid & Küchenhoff (2013) have argued against the use of an approach based on *p*-values. However, as I have discussed elsewhere (Gries (2012, 2015), many of their claims were highly problematic, to say the least. For instance, counter to Schmid & Küchenhoff (2013:516), collostructional analysis does not "[require] a more powerful computer" to handle input data with high frequencies – instead, all that is needed is extending the relevant computer's ability to handle large numbers using a Multiple Precision Floating-Point Reliable Library. Similarly, the ranking of collostruction strengths is much less sensitive to how all constructions are counted in the corpus than they suggest. Finally, Schmid & Küchenhoff (2013) contradict themselves when (i) they criticize p_{FYE} for requiring an estimate of the number of constructions in the corpus while apparently not minding that very same fact when they promote the odds ratio as a measure, which does, too, and (ii) they analyze a data set of their own to show the inferiority of p_{FYE} and end up with a result in which precisely that measure is correlated best with the experimental reference data.

That being said, the point worth addressing here again is that AMs based on the NHST conflate at least two pieces of information, namely the size of the effect – attraction or repulsion – and the sample size, i.e. the sum of the four cells *a* to *d* in Table 1. The fact that this is so can be easily seen from the data in Table 2, which contains co-occurrence frequencies of the verb *regard* and the so-called *as*-predicative (V NP_{DO} *as* XP):

Table 2: Co-occurrence data on *regard* and the *as*-predicative (Gries, Hampe, & Schönefeld 2005)

	as-predicative	other constructions	Sum
regard	80	19	99
other verbs	607	137958	138565
Sum	687	137977	138664

Computing several of *p*-value-based AMs yields the following results: G^2 / log-likelihood ratio = 762.2, t = 8.89, z = 113.53, and $-\log_{10}p_{FYE}$ = 166.48. However, if one multiplies all values *a* to *d* in Table 2 by 10 and recomputes the AMs, they change considerably: G^2 / log-likelihood ratio = 7621.96, t = 28.11, z = 359.01, and $-\log_{10}p_{FYE}$ = 1656.55 – other measures, such as *MI* or the log odds ratio, stay the same (7.35 and 6.86 respectively).

This characteristic, the sensitivity to sample size, can be seen as either a 'bug' or a 'feature', depending on one's goals. Given the goals of most applications of CA with p_{FYE}, I would argue that the conflation is a feature, i.e. useful: The measure, and thus CA, is coming from a background of usage-based linguistics/construction grammar where the information inherent in frequency of (co-)occurrence is important, and it is useful to know whether a certain association (effect) is found in generally smaller or larger data sets. This is especially the case, I submit, when (i) the main or even only goal of the CA is to obtain a one-dimensional ranking of, say, verbs in a construction and (ii) in the absence of an AM approach that can handle effect sizes, frequencies, and maybe other dimensions quasi-separately, and I think it is a realistic assessment of most work using CA that they exhibited both of these features.

However, in cases where the analysis has a more cognitive and/or psycholinguistic orientation, the conflation is probably less ideal precisely because of the simplification it entails: With enough statistical sophistication, frequency and effect size can, or maybe should, be kept separate. As innocent as that sounds, it complicates things considerably because it means that, rather than having one AM for, say, each verb in a construction by which these verbs can be ranked – the overwhelming practice in nearly all corpus-linguistic work on co-occurrence – we would then have two: (i) the frequency of co-occurrence (i.e. 80 for the above example of *regard* in the *as*-predicative) and (ii) the AM measuring only the effect of the association (e.g., the log odds ratio of 6.86 for the above example). This means that the verbs would now be ranked according to two dimensions, which is tricky because these will of course not always lead to the same ranking. For instance, if we keep frequency and effect size separate, we see above that the verb *regard* occurs 80 times in the *as*-predicative and yields a log odds ratio of 6.86. How are we going to relate this to the behavior of the verb *see* in the *as*-predicative and elsewhere? In Gries, Hampe, & Schönefeld's (2005) data, *see* occurs 111 times in the *as*-predicative (i.e. more often than *regard*), but its log odds ratio with the *as*-predicative is only 2.64. Thus, if the goal is to determine whether *regard* or *see* is 'more attracted to', or 'more prototypical of the construction's semantics' in a usage-based theoretical context in which frequency is generally considered

important, then do we prioritize frequency (leading to *see*) or association strength (leading to *regard*)? In addition, this situation is only becoming more complex once we add other dimensions of information, such as type frequencies/distribution and dispersion.

2.2 AMs and type frequencies/distributions

The second problem is one that was already briefly discussed in Gries (2012, 2015), namely the fact that nearly all AMs do not include any information about what in Table 1 are the 'other elements' row and column. More concretely and specifically, Table 2 does not reveal

> how many verb types other than *regard* occur in the remaining 607 *as*-predicative tokens and with what frequencies?
> how many other construction types *regard* make up the 19 tokens it is not used in the *as*-predicative and with what frequencies?

The way that AMs are usually computed, the former information is usually available from the concordance that led to Table 2: One just looks at all 687 *as*-predicatives and counts how often each verb in the verb slot ever occurs in it; in Gries, Hampe, & Schönefeld's (2005) data, there are 107 different verb types in the *as*-predicative with a very typical Zipfian distribution: Nearly half of the 107 types (52/48.6%) occur in it only once and a mere 3 verb types (*see*, *describe*, and *regard*) account for 40.6% of all tokens. In fact, one can defend *regard* as the *as*-predicatives prototype – rather than the more frequent *see* and *describe* – by pointing out that (i) the *as*-predicative is by far the most frequent use of *regard* (80>>19) even if we don't know how many different construction types these 19 non-*as*-predicatives instantiate whereas (ii) *see* and *describe* are constructionally much more promiscuous and can and do occur with many more construction types than *regard*; in other words, *regard* is more 'focused on' or 'unique to' the *as*-predicative than *see* and *describe*.

The latter kind of information – which other constructions is each verb occurring in the *as*-predicative also attested in how often? – is hardly ever available simply because our corpora are usually not constructionally tagged (and it is not obvious how or at what level of generality to do that, how to deal with error rates etc.), and the rare cases of useful databases such as Roland, Elman, & Dick (2007) may not provide all the constructions one requires; for instance and unexpectedly, their spreadsheets do not contain data on a construction as prominent in the linguistic literature as the ditransitive. In other words, if we study the association of $word_1$ and $construction_1$, ideally we would need to take a table of the kind shown in the top of Figure 1 and 'zoom into' the constructions representing the **200** uses of $word_1$ outside of $construction_1$ and the *1000* uses of $construction_1$ without $word_1$ (Gries 2012:497-298).

Now why should one care about type frequencies and even their distributions? That question can be answered both just on the basis of the above (fictitious) data and on the basis of published work. As for the former, note that all typical AMs would return the same value for the association between $word_3$ and $construction_1$ on the one hand and the association between $word_4$ and $construction_1$ on the other because their 2×2 tables would involve the same frequencies a=40, b=460, c=1040, and whatever d results from in the remaining cells. That is hardly the best way to go given, for instance, the facts that (i) $construction_1$ makes up only a small portion of $word_3$'s uses but is also the most frequent construction $word_3$ is used with whereas (ii) $construction_1$ makes up the same small portion of $word_4$'s uses but the frequency of co-occurrence of $construction_1$ and $word_4$ is one magnitude less than the frequency of co-occurrence

of construction$_2$ and word$_4$.

	construction$_1$	other c.	Sum
word$_1$	80	**200**	280
other w.	*1000*	…	…
Sum	1080	…	…

	construction$_1$	construction$_2$	construction$_3$	construction$_4$	construction$_5$	constr.$_{6\text{-}15}$	Sum
word$_1$	80	**90**	**45**	**35**	**25**	**5**	280
word$_2$	*60*	0	310	0	0	0	370
word$_3$	*40*	30	30	30	30	300	460
word$_4$	*40*	407	1	1	1	10	460
word$_5$	*40*	420	0	0	0	0	460
words$_{6\text{-}20}$	*810*	…	…	…	…	…	…
Sum	1080	…	…	…	…	…	…

Figure 1: Zooming into the 'other' row and column of a traditional 2x2 co-occurrence table

As for the latter, we know that type frequency and their distributions are correlated with many important aspects of processing and learning, as much recent work involving notions such as surprisal and entropy have shown (Levy 2008, Jaeger & Snider 2008, Lester & Moscoso del Prado Martín 2016). Surprisal is essentially a logarithmic transformation of a conditional probability as defined in (2)a, while the entropy of a probability distribution is defined in (2)b.

(2) a. $log_2 p(some\ form/function\,|some\ form/function/context)$

b. $\sum_{i=1}^{n} p\,(x)\quad log_2 p(x)$, with $log_2 0 = 0$

Linzen & Jaeger (2015) find that the entropy reduction of potential parse completions is correlated with reading times of sentences involving the DO/SC alternation. Lester & Moscoso del Prado (2017) find that entropies of syntactic distributions affect response times of nouns in isolation and the ordering in coordinate NPs. Lester et al. (2017) find that words occurring in similar distributions of syntactic constructions prime each other. Nouns' representations appear to be connected to syntactic structures in proportion to how often they occur in them. Goldberg, Casenhiser, & Sethuraman (2004) find that subjects, when presented with two different distributions of 5 novel verbs in 16 tokens, learn the verbs in the lower entropy distribution better (i.e., the latter distribution). Finally, Ellis (2011) shows how much language learning is in general influenced by learners'/processors' statistical 'analyses' of the input.

In sum, in addition to the first potential desideratum from Section 2.1 – keeping the dimensions of frequency and effect size separate – we now add a third, namely adding more information on the distribution of two elements by breaking up the frequencies *b* and *c* into as many columns and rows as there are other constructions for the relevant verbs and other verbs for the relevant constructions.

2.3 Directionality of AMs

As mentioned above, most widely-used AMs conflate information that should not be conflated. Above it was frequency and the effect size reflecting the degree of association/contingency; here it is that most AMs quantify the *mutual* association/repulsion of two elements rather than separating the two directions of association. As I have discussed in more detail elsewhere (Gries

2013), this common practice leaves a lot to be desired (i) theoretically since there is no reason to assume that learning and processing are bidirectional and symmetric and (ii) practically/empirically since it cannot distinguish the following kinds of cases:

> 2-grams where word$_1$ attracts word$_2$ but not vice versa such as *according to*, *upside down*, *instead of*, *ipso facto*, …;
> 2-grams where word$_2$ attracts word$_1$ but not vice versa such as *of course*, *at least*, *for instance*, *in vitro*, *de facto*, …;
> 2-grams where both words attract each other (nearly perfectly) such as *Sinn Fein* or *bona fide*.

Most traditional AMs would flag 2-grams from all these groups as 'strongly attracted to each other' (they all have G^2-values greater than 150 in the spoken BNC) without alerting the user that only in the third group is the attraction truly mutual. Thus, a more comprehensive and cognitively realistic approach would also keep the directions of association separate, too, and ideally with an AM that keeps frequency and effect size separate; one such measure is ΔP, a difference between conditional probabilities ranging from -1 (perfect repulsion) to +1 (perfect attraction). For the data in Table 2, it shows that *regard* attracts the *as*-predicative very much whereas the *as*-predicative attracts *regard* much much less:

(3) a. $\Delta P_{verb \rightarrow construction} = \frac{a}{a+b} \quad \frac{c}{c+d} = 0.8037$

b. $\Delta P_{construction \rightarrow verb} = \frac{a}{a+c} \quad \frac{b}{b+d} = 0.1163$

2.4 Underdispersion of co-occurrence data

The fourth and final problem mentioned above is underdispersion, i.e. the fact that all frequencies in tables of both types in Figure 1 are nearly always frequencies from a complete corpus or a complete register or learner group. This, however, means that any AM calculations based on such tables (or any key words calculations, for that matter) can be affected by the fact that the observed co-occurrence frequency in cell *a* may only be due to a very peculiar corpus part. For AMs, Stefanowitsch & Gries (2003) pointed that out already in their very first CA paper, namely in their case study of the English imperative (using p_{FYE} on the British component of the International Corpus of English). While that analysis returned many verbs one might have intuitively expected – *let*, *see*, *look*, *listen*, *worry*, etc. – it also returned *process* and *fold* with high collexeme strengths. Crucially, these results were due to these two verbs occurring in the imperative a lot of times in only one of the 500 ICE-GB files. In other words, the moderate frequencies of co-occurrence of *fold* and *process* with the imperative resulting in high AMs did not reflect that the words were completely underdispersed, something which a dispersion measure would immediately reflect. One dispersion measure, *DP* (Gries (2008), is computed as shown in (4), where $p_{1\text{-i}}$ are the percentages of occurrence of the element in a corpus over all corpus parts and where $s_{1\text{-i}}$ are the sizes of the corpus parts in % over which a dispersion measure is computed.

(4) $0.5 \times \sum_{i=1}^{n} \left| \frac{p_i}{s_i} \right|$

DP ranges from near 0 (an element is distributed as the corpus part sizes would lead one

to expect) to near 1 (an element is distributed extremely unevenly in a corpus). That is, high values indicate high levels of clumpiness or uneven dispersion – if high values are meant to indicate even dispersions, one can just use 1-*DP*. *DP* for *fold* and *process* in the imperative in the ICE-GB are >0.995, indicating that relying only on an AM without also taking (under)dispersion into consideration can lead to misleading results.

However, there are also theoretical arguments to consider dispersion information. For instance, Schmid (2010:115) states "frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse," and one way in which recency can be operationalized in corpus linguistics is via dispersion: If something occurs (more) regularly, it is more likely to have been seen recently. Also, we know that

> learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition. (Ambridge et al. 2006: 175)

"Distribution over sessions" in a corpus is dispersion. Finally, studies such as Adelman, Brown, & Quesada (2006) or Gries (2010) show that dispersion measures can outperform token frequency as a predictor of word naming and lexical decision times and are therefore just as relevant as the token frequencies that typically only go into computing AMs.

2.5 Against conflation, towards tupleization

As this section has hopefully made clear, the computation of most AMs is as problematic as it is widespread. Unfortunately, I believe that this practice is so widespread for the sole reason that researchers like to have one AM-value so they can sort by it and do not have to worry about the complexities that arise if suddenly every co-occurrence item is characterized by a tuple of multiple values. Research on both lexical co-occurrence (i.e. collocations), on lexico-grammatical co-occurrence (i.e., collostructions), or on key words routinely conflates many things, most of which should most likely not be conflated:

> frequency and effect size in the choice of AM;
> the 'other' categories in both the rows and the columns of the traditional 2×2 tables;
> the directions of association/repulsion of the two elements involved;
> frequencies from whole corpora regardless of the elements' dispersions.

If one, instead of considering all this information, just uses G^2, one essentially just 'hopes and prays' that that one G^2-value will somehow still capture all of these dimensions we know to be cognitive relevant well enough. That is to say, one hopes the G^2-value is good enough as an approximation but one does of course not know at all how much each of the dimensions – frequency, mutual association, two unidirectional associations, dispersion, maybe entropies – enter into G^2: In case study X, does G^2 react reflect more the co-occurrence frequency or the association direction from, say, the verbs to a construction or the association direction from the construction to the verbs? And are the weights of these things on G^2 the same in case study Y? Probably not ...

Note again that this conflation problem is in fact not restricted to AMs: We run into the exact same kind of conflation problem when linguists compute adjusted frequencies for words in

corpora, i.e. a frequency of a word that is 'downgraded' to a lower value if the word is underdispersed (see, e.g., Davies & Gardner 2010 or Gardner & Davies 2014). As I have shown elsewhere (Gries to appear), this is not a good idea: If a researcher reports an adjusted frequency of 35 for a word, one does not know whether that word occurs 35 perfectly evenly distributed times in the corpus (i.e., frequency=35 and, say, Juilland's *D*=1) or whether it occurs 350 very unevenly distributed times in the corpus (i.e., frequency=350 and, say, Juilland's *D*=0.1). While this example is of course hypothetical, it is not unrealistic. For instance, the products of observed frequency and 1-*DP* for the two words *pull* and *chairman* in the spoken BNC are very similar – 375 and 368.41 respectively – but they result from *very* different frequencies and dispersions: 750 and 0.5 for *pull* but 1939 and 0.81 for *chairman*. Not only is it the dispersion value, not frequency, that reflects our intuition (that *pull* is more basic/widely-used than *chairman*) much better, but this also shows that we would probably not want to treat those two cases as 'the same' as one implicitly does when one simply computes and reports one conflated adjusted frequency. The same is true of key words, as mentioned above: key-word statistics based on 2×2 tables with one word (present vs. absent) in the rows and, say, two corpora in the columns have virtually always neglected to take into consideration how evenly dispersed in the two corpora the two words whose frequencies are listed in cells *a* and *b* are, a flaw that undermines parts of every single key words analysis.

Thus, I think that the diagnosis is fairly straightforward and uncontroversial, but not often stated. In what follows, I want to make a proposal regarding how to proceed abstractly and then offer some first modest examples of what such data and their analysis might look like. As for the proposal, I will argue that we should abandon conflation in most non-applied cases in favor of what, for lack of a better term, might be called *the tupleization of corpus linguistics*, namely (i) the collection of multiple values per event type, where event type can refer to an individual element or, more the focus here, the co-occurrence of elements and (ii) the use of as many of those values as possible in the analysis/interpretation part. As for brief exemplification, Sections 3.1 and 3.2 discuss collexeme-analysis applications to the ditransitive and the imperative in English respectively, whereas Section 3.3 is a distinctive-collexeme-analysis application to transitive phrasal verbs in English. Space does unfortunately not permit a detailed discussion of the many kinds of results tupleization provides and much of the exposition necessitates the use of plots, but I hope the overall logic will nonetheless become clear.

3 Case studies of tupleization

3.1 A collexeme analysis of the ditransitive

3.1.1 A 'traditional' collexeme analysis: 1-tuples

For the analysis of the ditransitives, I used release 2 of the ICE-GB. Using a small R script, all ditransitives were extracted from the complete corpus and lemmatized; also, for each verb lemma attested in the ditransitive at least once, its overall frequency in the corpus was determined as well. From those data, a collexeme analysis was computed using G^2, which shows the expected result, namely that *give* and the ditransitive attract each other most strongly, i.e. *give* is the prototypical ditransitive verb, followed by many other verbs denoting literal or metaphorical transfer (e.g., communication); Figure 2 shows the top 30 collexemes.

3.1.2 Separating frequency and association/contingency: 2-tuples

However, G^2 is a *p*-value-based measure conflating frequency and effect size, so let's break this up: Figure 3 represents observed frequency on the *x*-axis (logged) and the association as an effect size (logged odds ratio with 0.5 added to all cell frequencies) on the *y*-axis. Admittedly, here the differences are not massive, but it is interesting to note, for instance, that *give*'s attraction with the ditransitive is actually not the strongest of all verbs involved (at least when measured with the adjusted odds ratio): *tell*, *convince*, and *assure* all have higher pure association values than *give*, but are less frequent in the ditransitive, which is why *give* comes out on top in the traditional approach of Figure 2. Again, from a perspective that I myself have adopted at times, this has the advantage that G^2 provides information on both dimensions, but also the disadvantage that it loses information that *may* be interesting.

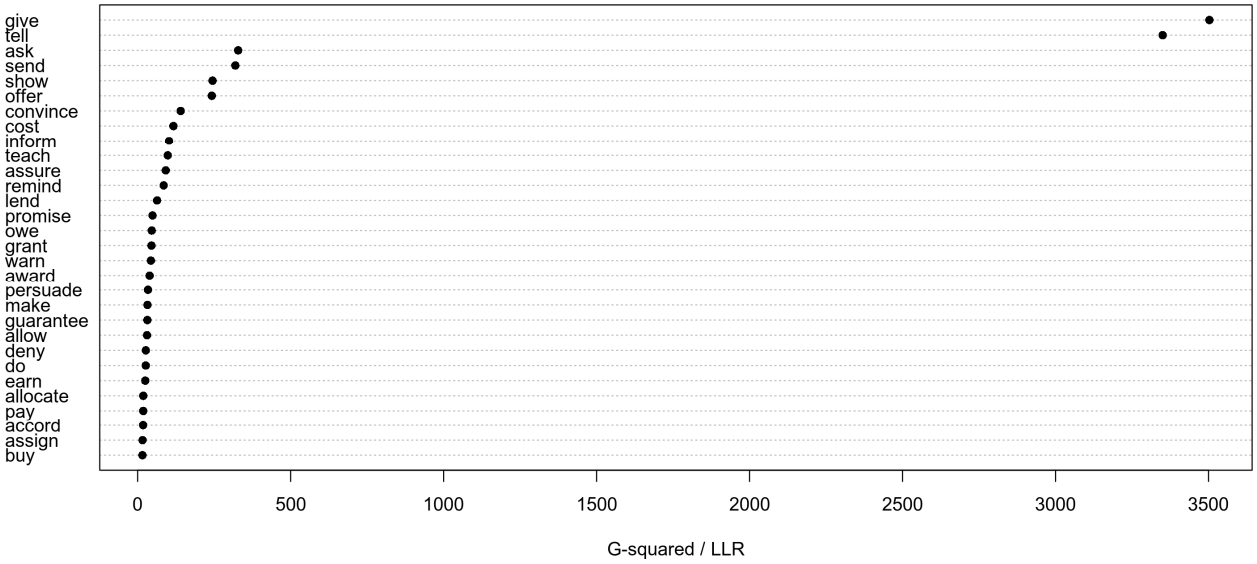

Figure 2: 1-: a traditional collexeme analysis of the ditransitive using G^2

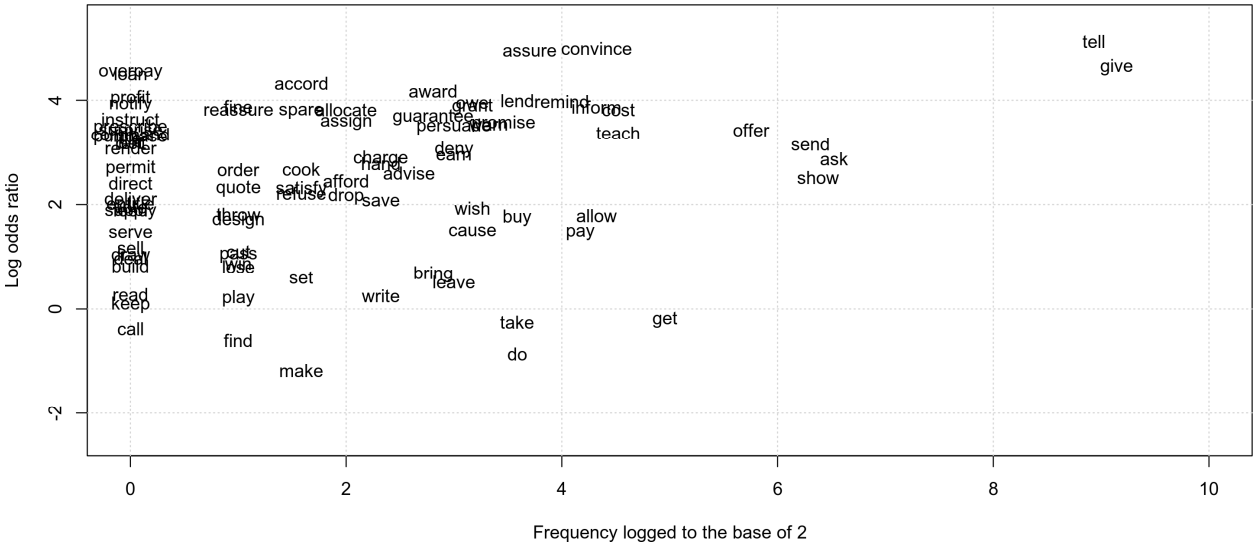

Figure 3: 2-tuples: separating frequency in the ditransitive (*x*-axis) and association (*y*-axis)

3.1.3 Adding dispersion to frequency and association: 3-tuples

However, we should still do better than this. For instance, we can add the dimension of dispersion. For each verb in the ditransitive, I computed its dispersion in the ditransitive against the baseline of the overall frequencies of these verbs in the corpus files. The dispersion values computed was *DP*, for plotting I am using 1-*DP* to have high values correspond to high/even dispersion throughout the corpus. Figure 4 shows the resulting 3-dimensional cube from two different perspectives. The left panel of Figure 4 essentially has Figure 3 'on the floor' (the log odds ratio axis has tick marks from -1 to 5 and the frequency axis has tick marks from 0 to 8), and it adds dispersion going 'up towards the ceiling' (with tick marks from 0 to 0.5), the right panel has the same plot rotated to see more of the structure from a different angle.

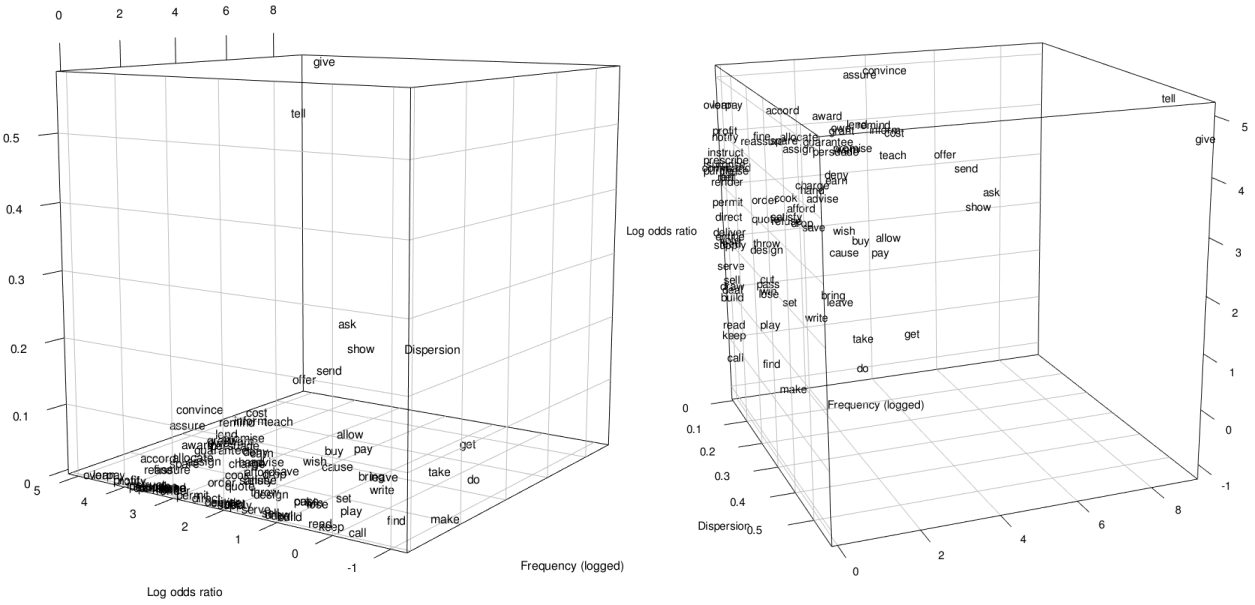

Figure 4: 3-tuples: frequency (*x*-axis), association (*y*-axis), and dispersion (*z*-axis)

Interestingly, we see here that, in a sense, *give* occupies the top slot again. We saw above that its attraction to the ditransitive is not quite as high as those of *tell*, *assure*, and *convince*, but we also saw that its frequency in the ditransitive is highest, and now we find that it is also most evenly dispersed in the ditransitive (and much more so than *convince* and *assure*), as one might expect a prototype to be.

3.1.4 Distinguishing directions of association: 4-tuples

Recall from above that association is mostly treated as bidirectional/symmetric/mutual even though we usually have no reason at all to make that assumption – after all, learning is grounded in time and events leading to learning are often not exatl contemporaneous but sequential, i.e. directional. Thus, let us explore what happens when, instead of the bidirectional log odds ratio, we use the directional Δ*P*s, which can distinguish how much a verb is attracted to the ditransitive and how much the ditransitive is attracted to a verb. That is to say, every verb's distributional behavior, which is usually only expressed in one number (e.g., G^2 or *MI* or …), is now characterized by a 4-tuple: {frequency, dispersion, $\Delta P_{\text{constr}\rightarrow\text{verb}}$, $\Delta P_{\text{verb}\rightarrow\text{constr}}$}, which we can represent in quasi-4-dimensional plots as in Figure 5: The left panel has frequency on the axis

with tick marks from 0 to 8, $\Delta P_{verb \rightarrow constr}$ on the axis with tick marks from 0 to 1, $\Delta P_{constr \rightarrow verb}$ on the vertical axis with tick marks from 0 to 0.3, and dispersion represented by the font size (bigger letters for more even dispersion). The right panel, by contrast, has the same two ΔP axes, a vertical axis with tick marks from 0 to 0.5 for dispersion, and uses font size for frequency.

While the interpretation of such 4-dimensional data using 2-dimensional plots is challenging – interactively rotatable 3D plots are more useful, but cannot be printed – some interesting observations can be made. For instance, it is very obvious that G^2 simplifies things considerably but not in a good way: The high G^2-values of many verbs notwithstanding, all verbs but *give* and *tell*, in spite of quite some frequency differences, exhibit quite low attractions from the ditransitive (all those $\Delta P_{constr \rightarrow verb}<0.05$), but *give* and *tell* exhibit much higher values, with *give* scoring highest (for *give*, $\Delta P_{constr \rightarrow verb}=0.307$; for *tell*, it is 0.267. On the other hand, the values for $\Delta P_{verb \rightarrow constr}$, i.e. the other direction of association, include some very high values (e.g., 0.987 for *overpay*, 0.709 for *assure*, 0.701 for *convince*), but then these verbs $\Delta P_{constr \rightarrow verb}$ are minuscule. Also, we can see that *give* seems to 'win' by scoring the highest values on frequency, dispersion, and $\Delta P_{constr \rightarrow verb}$ – only on the dimension of $\Delta P_{verb \rightarrow constr}$ does *tell* score higher.

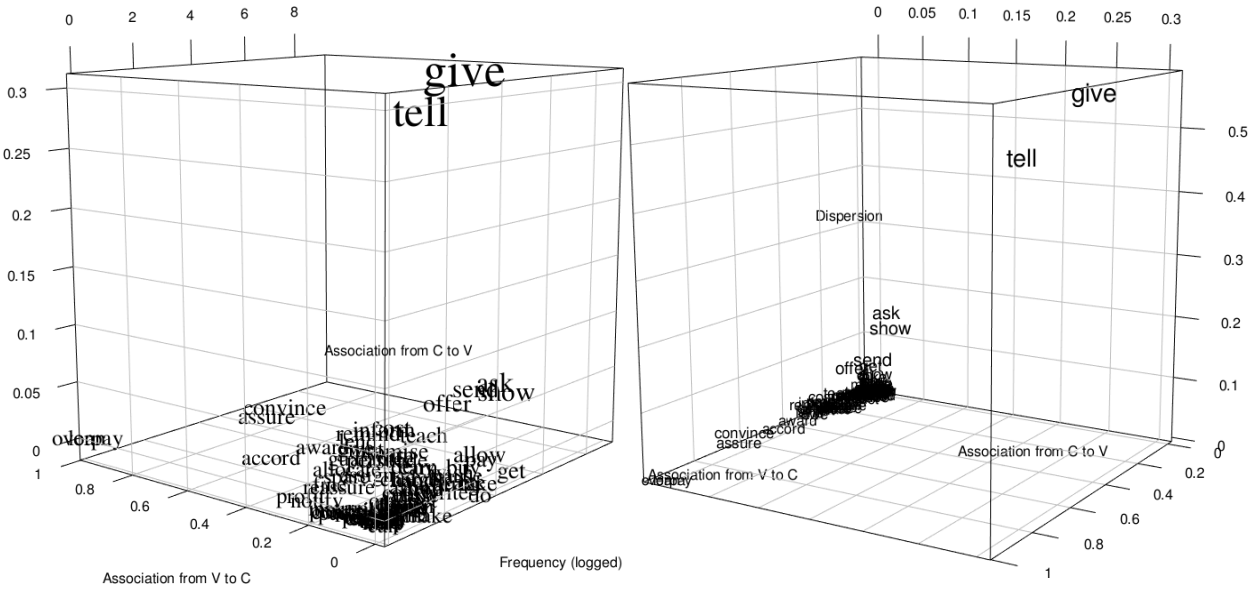

Figure 5: 4-tuples: frequency, verb-to-construction association, construction-to-verb association, and dispersion

In sum, AMs such as G^2 can be useful in the sense that the top two or three verbs might be identified properly, but they mask a huge amount of potentially interesting variability of the data that comes out once 4-tuples are used – ideally, we would also have entropy data for the distributions of ditransitive verbs in other constructions, but these are not straightforward to extract from the ICE-GB so this awaits future research. Clearly, any account of acquisition, learning, or processing of verbs' subcategorization patterns would benefit from being able to not just have one G^2-value, but to recognize (i) which verbs attract a certain construction ($\Delta P_{verb \rightarrow constr}$), which are attracted by a certain construction ($\Delta P_{constr \rightarrow verb}$), how likely, say, children or learners are to encounter the construction (frequency and *DP*).

3.2 A collexeme analysis of the imperative

In this section, we will look at the collexeme analysis of the imperative in English because it is interesting to see how the present tupleziation approach towards co-occurrence data solves the problem of underdispersed elements. All imperative forms were retrieved from the ICE-GB and for each verb lemma attested in the imperative, all occurrences in other constructions were counted as well; then, for each verb lemma, a collexeme strength value was computed (again using G^2). The top 30 collexemes of the imperative are shown in Figure 6.

As is obvious, most of the verbs make a lot of sense in the imperative, but here *fold* is even higher up in the list than in the p_{FYE}-based list of Stefanowitsch & Gries (2003) and *process* is also still ahead of verbs one might more straightforwardly expect in the imperative such as *hesitate* or *forget* (probably most often in *don't hesitate/forget to …*) and *shut* (probably most often in *shut up*). How do the results change when, again, G^2 is split up into (logged) frequency and the two directions of association/contingency offered by ΔP and when we add dispersion? Two versions of the resulting plot are shown in Figure 7, with a frequency axis with tick marks from 0 to 6, $\Delta P_{verb \rightarrow constr}$ on the vertical axis with tick marks from 0 to 1, $\Delta P_{constr \rightarrow verb}$ on the remaining axis with tick marks from -0.2 to 0.05, and dispersion represented by the font size (bigger letters for more even dispersion).

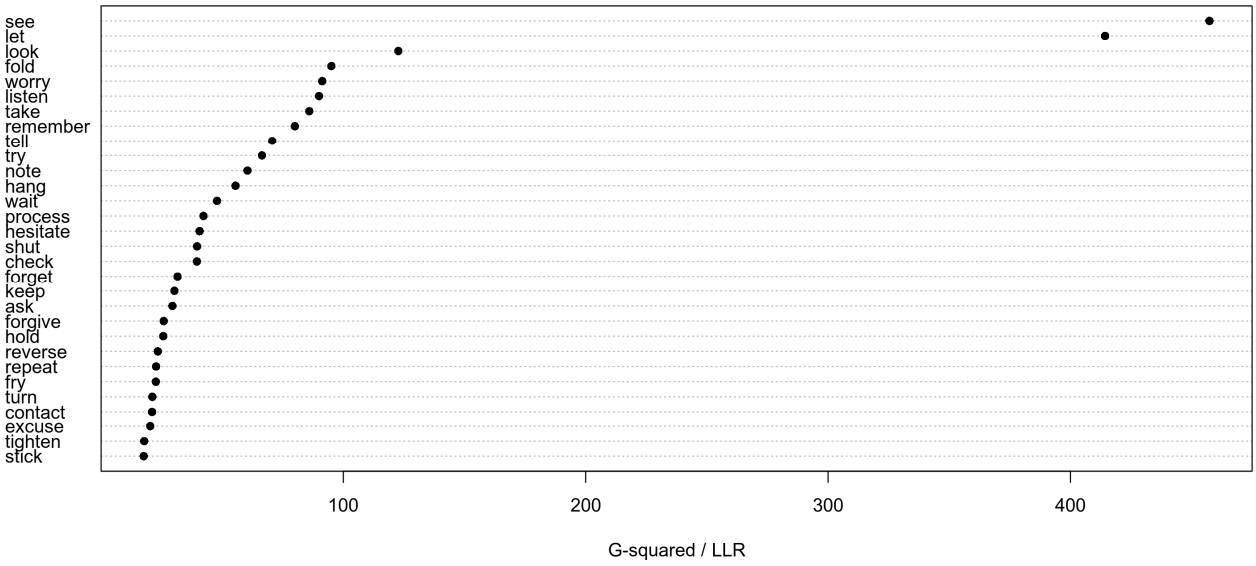

Figure 6: 1-tuples: a traditional collexeme analysis of the imperative using G^2/*LLR*

One clear finding is that with very exceptions, the attraction from the construction to most verbs is really very small: *see* and *let* are the only verbs with a noticeable attraction in this direction, and *be* is interesting for its clear negative value: the imperative strongly 'repels' *be* and less strongly *have*. Let us briefly compare *let* and *see*, given their prominent positions: *see* scores higher than *let* in terms of frequency, $\Delta P_{constr \rightarrow verb}$, and dispersion, but scores lower than *let* in terms of $\Delta P_{verb \rightarrow constr}$; in fact *see* scores lower than the average of that dimension (0.117), so many verbs – including *fold* and *process*, but also much rarer ones like *fry* or *reverse* – score much higher values on that dimension. Thus, the high rankings of *fold* and *process* are solely due to $\Delta P_{verb \rightarrow constr}$; all other dimensions of co-occurrence clearly flag those two verbs as not being strongly attracted to the imperative, which is of course what one would hope a multivariate

tupleization approach would be able to detect.

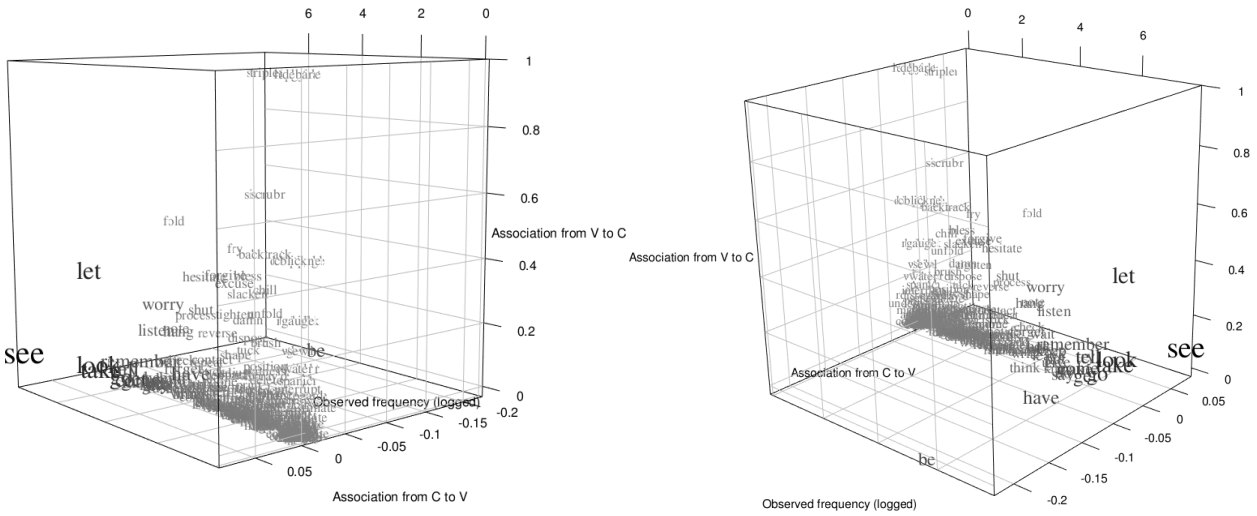

Figure 7: 4-tuples: frequency, verb-to-construction association, construction-to-verb association, and dispersion

3.3 A distinctive-collexeme analysis of transitive phrasal verbs

Finally, let us very briefly look at an example of a distinctive collexeme analysis, i.e. a case where for each verb we determine which of two functionally similar constructions they prefer; here, we are revisiting data from Gries & Stefanowitsch (2004) on the alternation of transitive phrasal verbs shown in (5):

(5) a. Captain Picard gave back the phaser. V-Part-DO
b. Captain Picard gave the phaser back. V-DO-Part

All transitive phrasal verbs were retrieved from the ICE-GB and for each verb lemma its frequencies of occurrence in each of the two constructions was determined. In this case, the tupleization analysis is a bit easier because one does not need to distinguish directions of associations: the ΔP-score of a verb for V-Part-DO is simply the negative of the AM for V-DO-Part, which means we are 'only' dealing with frequency, association, and dispersion, as shown in Figure 8: the logged frequency of the verb in V-DO-Part is on the axis with tick marks from 0 to 5, the association to V-DO-Part is on the vertical axis with tick marks from -0.4 to +0.4, and the dispersion score is on the bottom axis with tick marks from 0 to 0.1; the two red squares highlight the verbs with the highest G^2-scores for the two constructions.

As discussed in our previous paper, there is a patterning such that the uses of the verbs attracted to V-Part-DO are often idiomatic or metaphorical such that the particle does in fact not denote the spatial endpoint or resultant state of the referent of the DO: the uses of *carry out* refer to 'execute', not 'transport outside', etc. By contrast, the uses of the verbs attracted to V-DO-Part mostly do refer to spatial endpoints of motion events. The top G^2-scores of *carry out* and *get up* are mostly a reflection of the combination of frequency and the association score. However, in the distinctive collexeme analysis case, the dimensions are more correlated with each than in collexeme analyses because the overall sum of each 2×2 table is only the number of transitive

phrasal verbs, meaning the values of what in a collexeme analysis are the *d*-values are much more homogeneous here. That being said, the plot still shows quite some variability that G^2-values alone would fail to uncover and it still allows an analyst to restrict xyr attention to only those verbs that meet, for instance, certain frequency and dispersion thresholds.

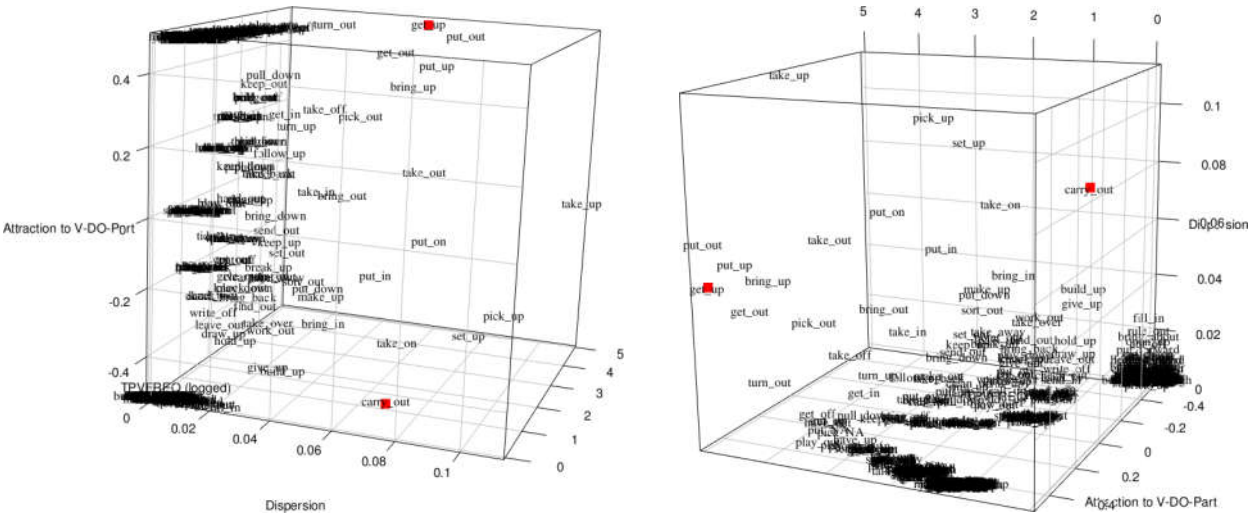

Figure 8: 3-tuples: frequency, verb-to-V-DO-Prt association, and dispersion

3.4 A brief discussion of type-token distributions and entropies

In the above examples, the dimension of type-token distributions and entropies was omitted for the practical reasons that it is not trivial how and at what level of resolution to extract such data from the ICE-GB. However, to give at least an idea of how such information might be added to the 4-tuples we explored above, I will briefly exemplify this on the basis of the data from Roland, Elman, & Dick (2007), who provide cross-tabulations of a set of verbs and a set of syntactic constructions for a variety of corpora; Table 3 is a small excerpt of their data, a parsed version of the British National Corpus.

Table 3: Excerpt of Roland, Elman, & Dick's (2007) BNC data on verbs and constructions

	Ditransitive	Intransitive	Transitive	PP	ToInfVP	27 more cs
accept	45	786	6153	702	14	...
acknowledge	9	322	1013	138	11	...
admit	8	1222	1119	681	148	...
advise	10	276	608	448	35	...
advocate	4	29	399	62	0	...
199 more vs	…	…	…	…	…	...

From such data, it is now possible to, for instance, identify the syntactically/constructionally most and least diverse verbs by computing the entropy of each verb's row using formula (2)b from above. Applying entropy computations to their data, we find that

the syntactically most diverse verbs are *help*, *prove*, *teach*, *advise*, *learn*, *call*, *find*; For instance, *help*'s entropy is 4.17 and to cover 75% of the uses of *help*, you need to include 11 of the 32 constructions in the database;
the syntactically least diverse verbs are *point*, *attempt*, *talk*, *respond*, *tire*, *hesitate*; for instance, *point*'s entropy is 2.45 and to cover 75% of the uses of *point*, you need to include only 2 of the 32 constructions (PP and V&Part).

That is, verbs and constructions differ in terms of both how often they take which constructions (i.e. their syntactic/constructional diversity) and which verbs (i.e. their lexical diversity) respectively and, as we have seen above in Section 2.2, for corpus-linguistic studies of acquisition and processing for instance, these things play important roles and would therefore ideally be included into the tupleization approach advocated here. In addition to the syntactic diversity of each verb and construction, however, we can also compute the closeness of each verb to the overall prototype of verbs respectively. Improving on earlier work by Milin et al. (2009) or Baayen et al. (2011), Lester (2017) proposes to

> operationalize the notion of a prototype of, say, all verbs' distributional behavior, as the frequency distribution of all constructions per verbs; that means computing the columns sums of Table 3, and then to
> determine the closeness of any verb to the overall verbal prototype by computing the Jensen-Shannon divergence (Lin 1991) of the relevant verb's row to the overall prototype, i.e. the column sums for all verbs.

This way, for each verb occurring in a construction we could add either a verb's constructional entropy as in (2)b or the verb's idiosyncrasy – the degree to which it deviates from the prototype – as element(s) to a tuple because, arguably and all other things being equal, a verb with a low constructional entropy, i.e. a verb that prefers to occur frequently in only a few constructions, forms stronger associations with the verbs that it does occur with than a highly promiscuous verb; similarly arguably, a verb that is very distinct from the overall prototype might also form stronger associations with the constructions it occurs in than a verb that is very typical, i.e. similar to all verbs' overall patterning. While these remarks are speculative at this point, this is informed speculation given the important role that in particular cognitive/psycholinguistic and quantitative corpus linguistics are assigning to matters of surprisal/entropy and prototypicality elsewhere, meaning these matters do merit more detailed research.

4 Discussion and concluding remarks

4.1 Interim summary

As the previous section has indicated, there is an alternative to the usual kind of conflation (of association measures, of adjusted frequencies, etc.) that corpus linguistics has over-relied on so far: (i) keeping dimensions of information separate, (ii) integrating them into a tuple, and (iii) perform different kinds of multidimensional analysis to understand the data better. Not all dimensions were explored – entropy/prototypicality were not discussed much, neither was the conflation of senses of words that often is implicit in collostructional/collocate analysis (see

Bernolet & Colleman 2016) – but we have seen the variability that the present approach reveals. In the next two sections, I discuss a few additional considerations arising from the above data.

4.1.1 What does G^2 (or p_{FYE}) actually do?

Given some of the results above, a sceptic might now ask what all this was good for if, for the ditransitive for instance, *give* comes out as the top verb for both G^2 and the tupleization approach. However, answering this question is straightforward. The first part of this answer is practical/empirical: One could not know in advance that an (oversimplified) one-dimensional measure of collexeme strength would return the same top position as the much more detailed analysis of all the verbs' tuples. Plus, let's not forget that the end of the previous section provided us with a much more nuanced understanding of the differences between, say, *give* and *tell* than could be gleaned just from G^2.

The second part of the answer is more theoretical: We know from previous literature – much of it cognitively and/or psycholinguistically informed – that the many dimensions studied above *are* relevant for acquisition/learning, processing, use, and change. Thus, while single-score-based rankings perhaps have a place in some contexts, any approach wanting to be more cognitively relevant/realistic needs to admit that, say, a G^2-approach returns only a single value for each verb without the analyst really knowing which of the many distributional characteristics of the verb and/or construction drive(s) a certain G^2-value most. I dare the reader to ask any corpus linguist reporting G^2-values what aspects of the data *exactly* are responsible for the scores being reported – which of course raises the question what G^2 is reflecting most. I did two stepwise regression analyses (bidirectional, based on *AIC*) for the ditransitive data. In both, G^2 was the dependent variable but in the first the predictors were frequency in the ditransitive (logged), the log odds ratio, the dispersion measure, $\Delta P_{constr \rightarrow v}$, and $\Delta P_{v \rightarrow constr}$; in the second, the predictors were all the previous ones but dispersion (because the values entering into dispersion are not the *a* to *d* frequencies from Table 1 that also enter into G^2). Both final models resulting from the selection processes achieved adjusted R^2s>0.98, but the analysis also shows how much each of the above dimensions is related to G^2. In the first model selection process, for instance, G^2 is most correlated with, or reflects most strongly (as measured by partial eta^2), the predictor $\Delta P_{ditransitive \rightarrow verb}$ (partial eta^2=0.99), followed by dispersion (partial eta^2=0.17); in the second model selection process, however, G^2 reflects most strongly $\Delta P_{ditransitive \rightarrow verb}$ (again, partial eta^2=0.99), followed by the log odds ratio (partial eta^2=0.13).

In other words, the high R^2-values show that whatever users of G^2 are looking for is in fact perfectly recoverable by the dimensions discussed here, but, obviously, the dimensions discussed here do this much more precisely and in a way that, if the goal is more than obtaining a simple ranking, is more useful and cognitively relevant/realistic. In addition and much more importantly, one needs to realize that the above results do not generalize to all studies using G^2: This is because, in the imperative data, the G^2-values can also be predicted nearly perfectly from the dimensions discussed above (adj. R^2>0.99), but not from the same predictors and not equally strongly: First, the statistical models are more complex and require pairwise interactions between predictors; second, the models return effect sizes that are quite different from those of the ditransitive: here, dispersion, $\Delta P_{imperative \rightarrow verb}$, their interaction, and the interaction of $\Delta P_{imperative \rightarrow verb}$ and the log odds ratio are the four strongest predictors. In other words, studies of the same type (simple collexeme analyses) on the same corpus (the ICE-GB) may use and report the same association measure (G^2), but in one case study (ditransitives) these G^2-values reflect mostly the association from the construction to the verb and not that much else, whereas in the

other (imperatives) the G^2-values reflect a much broader variety of distributional characteristics. That is, association measures such as G^2, p_{FYE}, t, z, while all called "*association* measures", reflect much more than association/contingency alone and they do so differently across data sets, which in turn means that one can actually not compare straightforwardly results and explanations from different data sets.

4.1.2 What if one really needs a one-dimensional ranking?

The previous sections have hopefully clarified that using G^2 can be too simplistic an idea no matter how widespread it is. However, I want to make a simple but also preliminary proposal as to how conflation *could* actually be achieved. This is not a contradiction to everything said so far, because the main problem of the conflations from G^2 or p_{FYE} is that the conflation is done 'uninformed', so to speak, because I think it is fair to assume that most corpus linguists would not have known how the influences of the different dimensions compare to each other in the different models and across the different constructions. In other words, a big part of the problem is that measures like G^2 or p_{FYE} offer a single number whose composition analysts don't know – but that can be changed. One possibility is to proceed as follows:

> choose the dimensions of information to include, e.g. the four dimensions of frequency, $\Delta P_{ditransitive \rightarrow verb}$, $\Delta P_{verb \rightarrow ditransitive}$, and dispersion;
> convert them all to an equal range, e.g., by transforming them to fall into the interval from 0 (for the minimum of each dimension) to 1 (for the maximum of each dimension). For instance, such a transformation would change the values {-2, -0.2, 0, 3, 6} into {0, 0.225, 0.250, 0.625, 1};
> represent the words by points at the values of the four dimensions in a four-dimensional unit hypercube and then measure the Euclidean distance of each word to the origin, which is just an application of the Pythagorean theorem.

Just like G^2, this approach yields a single AM for each verb and the ditransitive, but it is still better: First, because it forces the researcher to face the many dimensions underlying co-occurrence information. Second, because it forces the researcher to make and communicate an *explicit* decision about how each of the four dimensions should be weighted in the AM computation rather than just accept some non-linear effect combination of frequency and association embodied in G^2/p_{FYE}. The above approach weights the four dimensions equally, which means the researcher *explicitly* commits to saying 'frequency, both directions of association, and dispersion are all equally important to me and I want an AM that reflects that'. However, another researcher might say – for whatever (theoretical, empirical, experimental) reason – that the direction of association from the verb to the ditransitive is much more important, actually three times more important than each other dimension. Spatially, that re-prioritization corresponds to stretching the unit hypercube along that dimension such that it doesn't range from 0 to 1 anymore but from 0 to 3 and then recomputing all Euclidean distances. Figure 9 compares the top 30 collexemes according to both weightings.

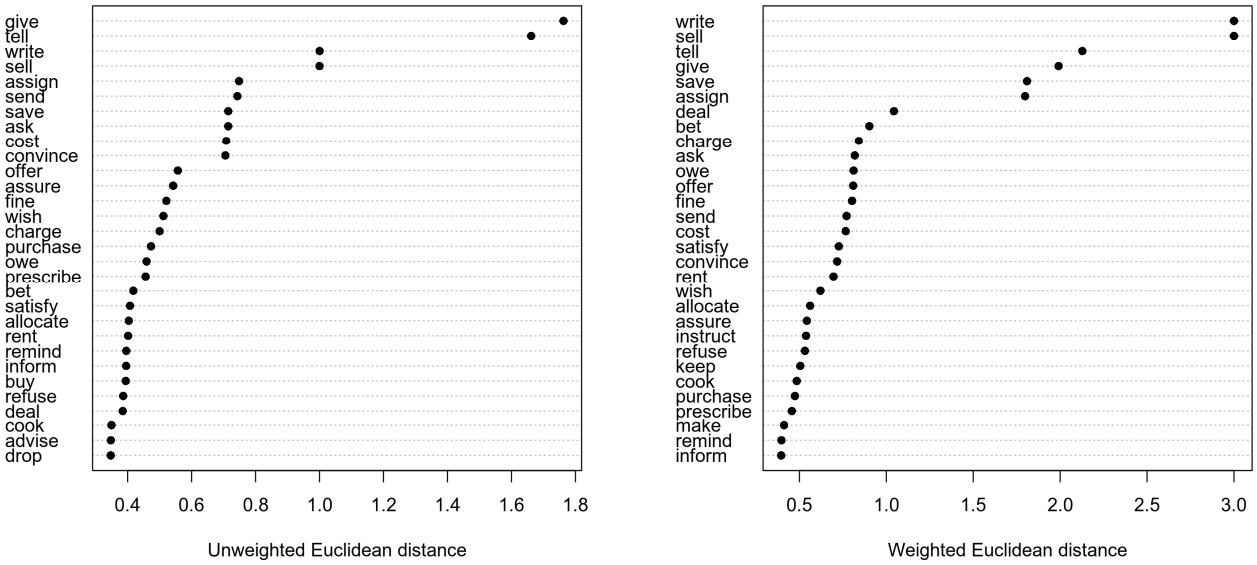

Figure 9: Informed conflation: equal weighting (left panel) versus prioritizing by a factor of 3 $\Delta P_{\text{verb}\rightarrow\text{construction}}$ (right panel)

The left approach – arguably the default – ends up assigning the top spot to *give*, followed by *tell*, then *write* and *sell*, etc., which intuitively makes a lot of sense. The right plot, which strongly prioritizes $\Delta P_{\text{verb}\rightarrow\text{ditransitive}}$ returns different results: *write* and *sell* win out (these verbs' AM-values change most), followed by *tell*, then *give*, etc., and a variety of other verbs such as *save*, *assign*, *deal*, and *bet* get higher values now that $\Delta P_{\text{verb}\rightarrow\text{ditransitive}}$ is prioritized whereas *give* and *tell* are downgraded, as are *ask*, *send*, and *convince*. Obviously, other prioritizations – e.g. one that says dispersion is more important than frequency – are also possible and worth exploring, as would be, more generally, the one dimension that was left out of this case studies (for lack of the pertinent data), namely the type-token distributions of the verbs and constructions (entropy and degree of prototypicality).

How does this apply to the imperative? The dominant role that *see* plays once all dimensions are first kept separate and then conflated with an explicit equal weighting is represented in the dotchart of Euclidean distances in Figure 10, which also shows that a variety of rarer verbs such as *season* make it into this top 30 group only because of their very high $\Delta P_{\text{verb}\rightarrow\text{imperative}}$ values, which points to a possible extension of the method, namely an exploration of further possibilities of how to downgrade such verbs: one possibility would be some weighting of the four dimensions, but another option might be to require minimum values on each dimension for a verb to be eligible for the final best list or, maybe more objective/automatable, one could choose to compute the Euclidean distances in the four-dimensional hypercube from only choosing the three highest or lowest dimension values. If the highest value is discarded for each, *let* and *see* still score highly, but *season* would then only be characterized by the three small values and would not make it high into the list. This procedure is less arbitrary than it sounds: we know from other areas that, in order for learning to happen, high frequency and dispersion, for instance, are not always required (i.e., these two dimensions might be dropped) as long as contingency and salience are high (as when some children learned the word *chromium* after just a single indirect/contrasting exposure, see Carey & Bartlett 1978).

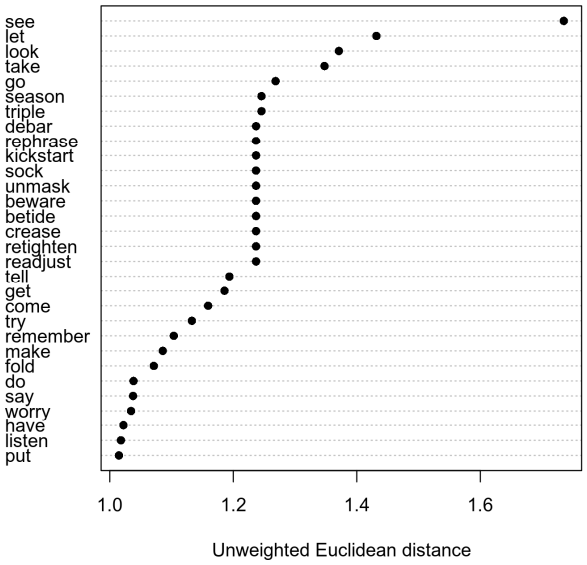

Figure 10: Informed conflation with equal weighting

In sum, the approach proposed here can also be used for a better form of conflation, namely one that does not leave the weightings of the different dimensions being conflated 'up to chance'. However, to make this as explicit as possible: The main thrust of this paper is the opposite, namely the separation/separate analysis of the different dimensions of information underlying co-occurrence data that I referred to as tupleization. The ideal outcome of this paper would be that co-occurrence phenomena – e.g., the co-occurrence of *give* and the ditransitive – are not summarized with a single value, but with a tuple of at least three values.

4.2 Where to go from here

A potentially particularly interesting approach following from this is concerned with the many attempts of correlating AM results from corpora with experimental results, which have often not produced the kind of good results corpus linguists (on the methodological side) and usage-based/cognitive linguists (on the theoretical side) have hoped for. I am convinced that part of the reason for this frequent lack of convergence is the huge degree of simplification we incur with our most frequently used measures. As I briefly discussed elsewhere (Gries 2013), correlating results from an inherently directional experimental task with a bidirectional association measure is bound to lead to suboptimal results, but as I discussed above, conflation is a much bigger problem for the discipline. When it comes to correlating experimental and corpus data, I hope that this paper will stimulate a greater degree of caution and precision: instead of regressing, say, reaction times to stimuli in a collocation production experiment on a single column of G^2-, *t*-, or *MI*-values of those collocations, the message of this paper is to instead look at the multivariate information and regress the reaction times on the four of five predictors discussed above to really see which dimensions of information are responsible for the experimental results.

Thus, I think the degree of precision resulting from the proposed tupleization of AMs and other corpus statistics can move all our analyses of co-occurrence data to a whole new level because we would be able to determine which (cognitively relevant) dimensions – frequency, association/contingency, recency/dispersion, etc. – really trigger subjects' responses, reactions times, sentence/VP-completions etc. After 50 years of ranking collocations and, later,

collostructions based on a single conflated score, maybe it's time to move things up a bit.

References

Adelman, James S. Gordon D.A. Brown, & Jose F. Quesada. 2006. Contextual Diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 19(9). 814-823.

Ambridge, Ben, Anna Theakston L., Elena V.M. Lieven, & Mike Tomasello. 2006 The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development* 21(2). 174-193.

Baayen, R. Harald, Petar Milin, Dušica Filipović-Đurđević, Peter Hendrix, & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438-482.

Bernolet, Sarah & Timothy Colleman. 2016. Sense-based and lexeme-based alternation biases in the Dutch dative alternation. In Jiyoung Yoon & Stefan Th. Gries (eds.), *Corpus-based approaches to Construction Grammar*, 165-198. Amsterdam & Philadelphia: John Benjamins.

Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Carey, Susan & Elsa Bartlett. 1978. Acquiring a single word. *Papers and Reports on Child Language Development* 15. 17-29.

Davies, Mark & Dee Gardner. 2010. *A frequency dictionary of contemporary American English: word sketches, collocates and thematic lists*. London & New York: Routledge, Taylor and Francis.

Deshors, Sandra C. 2016. *Multidimensional perspectives on interlanguage: Exploring* may *and* can *across learner corpora*. Corpora and Language in Use. Presses Universitaires de Louvain.

Ellis, Nick C. 2011. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1-24.

Ellis, Nick C., Ute Römer, & Matthew Brook O'Donnell. 2016. *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of Construction Grammar*. *Language Learning* 66 (Suppl. 1, Language Learning Monograph Series). New York: John Wiley.

Firth John R. 1957. A synopsis of linguistic theory 1930-55. Reprinted in F.R. Palmer (ed.), (1968) *Selected papers of J.R. Firth 1952-1959*. Longman, London.

Gardner Dee & Mark Davies. 2014. A new academic vocabulary list. *Applied Linguistics* 35(3). 305-327.

Goldberg, Adele E. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Gries, Stefan Th. Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365-399.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403-437.

Gries, Stefan Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-

based linguistics: some necessary clarifications. *Studies in Language* 36(3). 477-510.
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. 2015. More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505-536.
Gries, Stefan Th., Beate Hampe, & Doris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635-676.
Gries, Stefan Th. & Anatol Stefanowitsch. 2004a. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1). 97-129.
Gries, Stefan Th. & Anatol Stefanowitsch. 2004b. Co-varying collexemes in the *into*-causative. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture, and mind*, 225-236. Stanford, CA: CSLI.
Gries, Stefan Th. & Stefanie Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3. 182-200.
Gries, Stefan Th. & Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7. 163-186.
Gyselinck, Emmeline. 2018. The role of expressivity and productivity in (re)shaping the constructional network. Ph.D. dissertation, University of Ghent.
Harris Zellig S. 1970. *Papers in structural and transformational linguistics*. Reidel, Dordrecht.
Hilpert, Martin. 2012a. Diachronic collostructional analysis. How to use it, and how to deal with confounding factors. In Kathryn Allan & Justyna Robynson (eds.), *Current methods in historical semantics*, 133-160. Boston & Berlin: Mouton de Gruyter.
Hilpert, Martin. 2012b. Diachronic collostructional analysis meets the noun phrase. Studying many a noun in COHA. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*, 233-244. Oxford: Oxford University Press.
Hunston, Susan & Gill Francis 1999. Pattern Grammar: *A corpus-driven approach to the lexical grammar of English*. Amsterdam & Philadelphia: John Benjamins.
Jaeger, T. Florian & Neal Snider. 2008. Implicit Learning and Syntactic Persistence: Surprisal and Cumulativity. In *Proceedings of the Cognitive Science Society Conference*, ed. by Bradley C. Love, Kenneth McRae, K., Vladimir M. Sloutsky, 1061–1066. Washington, DC.
Lester, Nicholas A. 2017. The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition. Ph.D. dissertation, University of California, Santa Barbara.
Lester, Nicholas A. & Fermín Moscoso del Prado Martín. 2016. Syntactic flexibility in the noun: evidence from picture naming. Paper presented at CogSci 2016.
Lester, Nicholas A., Laurie B. Feldman, & Fermín Moscoso del Prado Martín. 2017. You can take a noun out of syntax …: syntactic similarity effects in lexical priming. Paper presented at CogSci 2017.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126-1177.
Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1). 145-151
Linzen, Tal & T. Florian Jaeger. 2015. Uncertainty and expectation in sentence processing:

evidence from subcategorization distributions. *Cognitive Science* 40(6). 1382-1411.
Matthys, Jana. 2014. Collostructional transfer in the dative alternation: An experimental study on the transfer of the dative constructions' verb biases by Flemish EFL learners. M.A. thesis, University of Ghent.
Milin, Petar, Dušica Filipović-Đurđević, & Fermín Moscoso del Prado Martín. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60(1). 50-64.
Roland, Douglas, Jeffrey L. Elman, & Frederick Dick 2007. Frequency of basic English grammatical structures: a. corpus analysis. *Journal of Memory and Language* 57(3). 348-379.
Schmid, Hans-Jörg. 2010. Entrenchment, salience, and basic levels. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, 117-138. Oxford University Press, Oxford.
Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531-577.
Schmid, Hans-Jörg & Friedrich Ungerer. 2011. Cognitive linguistics. In James Simpson (ed.), *The Routledge handbook of applied linguistics*, 611-624.
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209-243.
Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1-43.
Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin & New York: Mouton de Gruyter.
Wiechmann, Daniel. 2008. On the computation of collostruction strength. *Corpus Linguistics and Linguistic Theory* 4(2). 253-290.