Commentary

Stefan Th. Gries*

More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013)

DOI 10.1515/cog-2014-0092
Received December 1, 2014; revised January 25, 2015; accepted January 29, 2015

Abstract: Ever since the first studies introducing collostructional analysis in general and collexeme analysis in particular, these methods have been widely used for the analysis of constructions' semantic and functional characteristics. However, the more recent past has seen two publications, Bybee (2010) and Schmid and Küchenhoff (2013), which criticized several aspects of these methods. This paper briefly recaps my response to Bybee (2010) (published as Gries 2012) as a prelude to its main contribution, viz a rebuttal of various claims and problems of Schmid and Küchenhoff (2013).

Keywords: collostructional analysis, collexeme analysis, association measures, Fisher-Yates exact test, attraction and reliance

1 Introduction

1.1 General introduction

Over the last 20 or so years, Cognitive Linguistics has undergone a variety of changes. The field has changed from one that, in the 1980s, was dominated by the theoretical contributions of in particular George Lakoff (especially the theory of conceptual metaphor) and Ron Langacker (Cognitive Grammar) to a more diverse field in which many scholars contribute to and fuel the theoretical development towards an explicitly usage-/exemplar-based paradigm. In addition, the field has changed a lot methodologically: While much early work has not been particularly empirical, over the years the field has undergone a shift towards empiricism; many studies now use experimental data of varying degrees

***Corresponding author: Stefan Th. Gries,** Department of Linguistics, University of California, Santa Barbara, CA 93106-3100, E-mail: stgries@linguistics.ucsb.edu

of complexity (often using methods from psycholinguistics) or observational data (often using methods from corpus linguistics).

In the early 2000s, Stefanowitsch and I contributed to the cognitive-linguistic toolbox by tweaking a corpus-linguistic method – the use of statistical association measures to study collocations, the co-occurrence of words – so that it could be used to study the co-occurrence of words and constructions (in the Construction Grammar sense of the term). In a series of papers (Stefanowitsch and Gries 2003, 2005; Gries et al. 2004a, 2004b), we developed, extended, and exemplified this family of methods called *collostructional analysis*; an overview of the methods is represented in Figure 1.[1] These methods were referenced and adopted in a variety of subsequent studies and many studies have used not only the methods per se to study the co-occurrence of words and constructions, but also an R script (<coll.analysis.r>) provided by myself (Gries 2007, 2014; cf. http://tinyurl.com/collostructions).

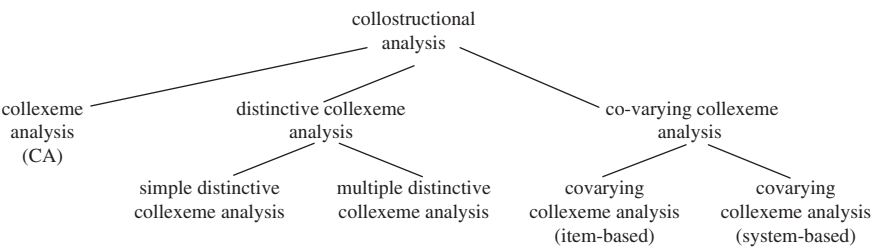

Figure 1: Overview of the members of the family of collostructional methods.

Recently, two publications have attempted to criticize the approach. The first of these is Bybee (2010, henceforth JB; cf. the appendix for a list of all frequently-used abbreviations); the second of these is Schmid and Küchenhoff (2013, henceforth S&K). Gries (2012) is a response to JB, and this paper is a response to S&K.[2] However, in order for this discussion to do justice to S&K's problems and omissions, it is necessary to briefly recapitulate some of the timeline of the development and discussion of collostructional analysis in general and collexeme analysis (CA) in particular. Thus, the remainder of this paper is structured as follows.

1 In the very first publication on collostructions (Stefanowitsch and Gries 2003), we used *collostructional analysis* with reference to the method of collexeme analysis. While later overview articles (in particular Stefanowitsch and Gries 2009) have been promoting a terminology conforming to Figure 1, some authors are still using *collostructional analysis* in this old way.

2 Note that both publications state they criticize collostructional analysis when in fact they only (seem to) take issue with collexeme analysis (see again n. 1).

Section 1.2 is a very brief recapitulation of the goals and method of CA as discussed in Stefanowitsch and Gries (2003) and then, most recently, in Gries (2012). Section 2 deals with the perceived problems of CA emerging from JB's discussion: Section 2.1 briefly mentions JB's main points, and Section 2.2 rebuts these (largely on the basis of Gries 2012); while this may appear redundant, it will become obvious how this is in fact essential for the subsequent discussion of S&K. Section 3 then deals with S&K: Section 3.1 summarizes their main points of critique; then, Section 3.2 discusses these points in detail. Section 4 concludes. As the reader will notice, the discussion below is sometimes a bit technical and rather detailed, but this is necessary because the many problems of the points S&K mention are not obvious to readers without certain background knowledge of both the method per se and the particular history of this paper. In an attempt to clarify (i) what the most common application of CA has to offer, (ii) why it has these things to offer, and (iii) what the relevant next steps would involve, a certain degree of deconstruction of S&K is indispensable; it is in the hope of setting straight what CA can and cannot do that I hope to contribute to the debate on methods in usage-/exemplar-based linguistics.

1.2 Overview of collexeme analysis

As mentioned above, CA is basically little more than the extension of the quantitative study of collocation (co-occurrences of words) with association measures (AMs) in corpus linguistics to the study of colligation (co-occurrences of words and grammatical patterns or constructions, hence *collostruction*) in Construction Grammar. Why would one want to study such co-occurrence phenomena? Because of the so-called *distributional hypothesis*, the assumption and finding that the similarity linguistic elements exhibit in terms of functional characteristics (semantic, pragmatic, ...) will be reflected in their distributional patterning in language (corpora) (cf. Firth 1957; Harris 1970: 785–786), i.e., the frequencies with which linguistic elements of interest co-occur with other linguistic/contextual elements. Given that (i) corpora differ in sizes and (ii) some linguistic elements' (e.g., *the*) frequencies are so high, corpus linguists have for decades now not used raw frequencies of co-occurrence as a measure of which co-occurrences are marked enough to be considered important but rather AMs that downgrade words that are highly frequent and omnipresent.

How are CAs, and thus AMs, computed and interpreted? The vast majority of them (cf. Pecina 2009) are computed and used according to the following four-step procedure.

i. retrieve all instances of a linguistic element *e* in question (with collocates, a word *w*; with CA, a construction *cx*);
ii. compute an AM for every relevant element type *ty* that co-occurs with *e* (with CA, those are often the words $w_{1\text{-}n}$ in a slot of construction *cx* and are referred to as *collexemes*);

Most AMs are then computed on the basis of a 2×2 co-occurrence table that cross-tabulates token (non-)occurrences of *e/w/cx* against every single co-occurring element/type *ty* as schematically represented in Table 1; thus, for instance, *a* is the number of times *e* co-occurs with *ty*, etc.

Table 1: Schematic frequency table of elements *e* and *ty* and their co-occurrence.

	***e* is present**	***e* is absent**	**Totals**
ty is present	a	b	$a+b$
ty is absent	c	d	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d=N$

iii. rank all co-occurring elements/types by the value of the AM;
iv. explore the top *n* (often 10–50) co-occurring elements/types for functional patterns.

Many AMs have been used in corpus-linguistic studies (cf. Evert 2009) but the most common ones are Mutual Information (*MI*), *t*, *z*, G^2, and $p_{\text{Fisher-Yates exact}}$ (FYE, cf. Pedersen 1996). For a CA, which follows the exact same four-step procedure "in principle, any of the measures proposed could be applied" (Stefanowitsch and Gries 2003: 217), but most papers' CAs use the negative $\log_{10}$ of the *p*-values of the FYE (which are multiplied by –1 in the case of elements/types repelled by *e*) because FYE

– is an exact test (rather than asymptotic) which makes no distributional assumptions;
– can therefore handle small and skewed frequencies better than, say, *MI* or chi-squared;
– as a significance test, can distinguish between identical effect sizes by weighing those that are based on more data more heavily;
– is not a linear function of the observed frequencies (in particular *a*) (cf. Evert 2009 for more discussion).

CA and other members of the family of collostructional analysis have been used successfully in a large variety of contexts: the synchronic analysis of

constructions and alternations, L1 and L2 acquisition, and language change, and related concepts have found their way into psycholinguistics (e.g., studies of syntactic priming or the current use of the notion of surprisal; cf. Gries 2005b; Jaeger and Snider 2008, 2013).

2 Bybee (2010) and Gries (2012): a short reminder

2.1 Bybee (2010)

As mentioned above, the first rather critical discussion of CA was Bybee (2010). In Section 5.12 of her book, she mentions several issues she views as problematic. The central ones of these are concerned with

(1) *normalization and the use of a significance test*: She argues against CA relativizing the co-occurrence frequency *a* (of a word *w* and a construction *c*) against the overall frequencies of *w* and *c* in the corpus because "lexemes do not occur in corpora by pure chance" and "it is entirely possible that the factors that make a lexeme high frequency in a corpus are precisely the factors that make it a central and defining member of the category of lexemes that occurs in a slot in a construction" (JB, p. 97).

(2) *cell* d *in the co-occurrence table*: "There is no known way to count the number of constructions in a corpus because a given clause may instantiate multiple constructions" (JB, p. 98).

(3) *discriminatory power*: Bybee claims a CA "works only with numbers" (JB, p. 100) and thus has less discriminatory power than her approach, which works with "simple frequencies analysis with semantic similarity" (JB, p. 100).

(4) *a lack of a cognitive mechanism justifying the normalization*: "By what cognitive mechanism does a language user devalue a lexeme in a construction if it is of high frequency generally? This is the question [CA] must address" (JB, p. 100f.).

Several of these comments are somewhat surprising because JB ignores both the useful results of many papers that make use of collostructional analyses and the published experimental work that was done to compare the results of a CA to that of "simple frequency analyses". In addition to this, her points of critique also exhibit specific problems, which will be briefly discussed in the next section.

2.2 Gries (2012)

The above-mentioned issues raised by Bybee I first discussed in a plenary address at the 6th International Construction Grammar Conference in Prague in 2010 as well as in a variety of other venues and then in a written response in Gries (2012) (available from my website since early-to-mid 2012), selected aspects of which are relevant for the discussion of S&K and will therefore be briefly recapitulated in this section.

2.2.1 Normalization and the use of a significance test

With regard to the first point of critique, I pointed out that any criticism of the logic of CA as a whole on the grounds that most studies use FYE as the association measure miss the point because Stefanowitsch and Gries (2003: 217) specifically stated that *any AM* can be used, significance-based or otherwise; in fact, the R script that most scholars doing collostructional analyses have used has always offered five or more AMs including *MI* and the odds ratio. Thus, anyone disliking this aspect of CA could just use any of the AMs that do not involve a significance test and still use and benefit from the overall logic of this approach. Second, I noted that the notion of normalizing observed co-occurrence frequencies of words/constructions against their overall frequencies is not only a standard approach in corpus linguistics – if Bybee was right, 50 years of research on collocations in corpus linguistics would need to be scrapped – but also separates the wheat (interesting frequent co-occurrence) from the chaff (uninteresting frequent co-occurrence; cf. Evert 2009: 1224); the example I discussed was the *as*-predicative in (5):

(5) a. V NP_{DO} *as* complement
b. I never saw myself as a costume designer.
c. Politicians regard themselves as being closer to actors.

The most frequent verb in this construction in the British Component of the International Corpus of English (ICE-GB) is *see*, but the verb with the highest collexeme strength (FYE) is *regard*. However, given (i) the high general frequency of *see*, (ii) its constructional promiscuity (that it can occur in many more different constructions), and (iii) the perfect match between the verb semantics of *regard* and what arguably the *as*-predicative is typically used for, it seems more appropriate to consider *regard* the most prototypical verb for the *as*-predicative, something that a simple frequency analysis would not reveal (but

see below).[3] A final example involves the comparison of three different AMs in a CA of the ditransitive construction: FYE, log odds, and *MI*, and while all three measures rank *give* highest, it is only FYE that ranks it highest as strongly as one might expect and it is only FYE that ranks next-highest verbs that are highly ditransitive and correspond to Goldberg's (2006) ditransitive senses such as *tell*, *send*, *offer*, etc. (rather than rare verbs like *award*, *allocate*, or *profit*).

2.2.2 The cell *d* in the co-occurrence table

While it is true that it is very difficult to determine one uniformly agreed-upon number of constructions in a corpus, which means that it is difficult to determine what number to use for *d*, this has not been problematic in any CA study I have ever seen.[4] As pointed out in Gries (2012: 488), the obvious solution – one that many statistical methods require – is that one chooses a level of resolution on which to count constructions that is close to the phenomenon in question: If one does a CA on an argument structure construction, then obviously using the number of letters of a corpus is useless, as is using the number of files – using the number of verbs or lexical verbs is probably more useful; cf. Roland, Dick and Elman (2007: 353) for a similar point. Second, JB (p. 98) herself reports that they "used several different corpus sizes" and different "corpus sizes yield similar results". Finally, Gries (2012) shows that halving all observed token frequencies and/or increasing the corpus size by one order of magnitude does not affect the resulting rank orders much.

2.2.3 Discriminatory power

Gries (2012) makes four comments with regard to Bybee's objections. First a more theoretical one, namely that Bybee's assumption that "[s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" is at odds with her exemplar-based approach as well as the large body of evidence from the distributional learning literature. This is because

3 An additional counterargument against non-normalizing on the basis of overall frequencies of occurrence applies in the cases of distinctive and covarying collexeme analyses; cf. Gries (2012: 487).

4 Cell *d* constituting a problem could have manifested itself in different ways: authors might have noted it as a problem or results could have been inexplicable for reasons that might have to do with cell *d* (e.g., with unintuitively high/low expected cell frequencies for cell *a*).

an exemplar-based approach such as hers – with which I agree – assumes that linguistic and contextual information of uses of linguistic elements (i.e., also semantic information) will be(come) associated with these linguistic elements, which in turn means that one might be able to infer semantic information from uses. Second, I discuss how the findings from many different CAs all revealed functional patterns and meaningful correlations with other data, and how cluster-analytic methods show fine semantic differences between collexemes. Third, I show that JB does not compare the results of her simple-frequency-analysis approach to a CA, but to a caricature of a CA that in fact uses maximally one of the four above-mentioned steps of a CA. Finally, I show how the experiments of Gries et al. (2005, 2010), in which frequency is pitted against FYE return significant effects for FYE but not for frequency (in a sentence-completion task and a self-paced reading task).[5]

2.2.4 Cognitive mechanisms

With regard to the alleged lack of discussion of the cognitive mechanisms that CA implies, I show in Gries (2012) how previous collostructional work – even the work cited by JB herself such as Stefanowitsch and Gries 2003) – relates to notions frequently discussed in cognitive-linguistic work such as conditional probabilities, cue validity and reliability, associative learning measures (such as Ellis's 2007 ΔP), and entrenchment.

2.2.5 Implications

In the final section of Gries (2012), I discussed a variety of implications. The discussion centered on three issues. The first was concerned with the question of the real complexity that studies of the associations of words and constructions need to face (at some point). To that end, I developed a cline of co-occurrence complexity, which is schematically represented in Figure 2. The top panel represents *simple frequencies/percentages* of some words $w_{1\text{-}3}$ in a construction c_1. It thus represents the kind of information that an account that was based only

5 The effect for FYE is highly significant in the sentence-completion task in a two-tailed test; the effect for FYE is significant in the self-paced reading task in a one-tailed test, which is permitted given that we had a directional alternative hypothesis (that FYE would have a facilitatory effect).

	*c*1
*w*1	**80**
*w*2	**60**
*w*3	**40**
…	…

	*c*1	other	Sum
*w*1	**80**	*200*	**280**
other	1000	…	…
Sum	**1080**	…	sum

	*c*1	other	Sum
*w*2	**60**	*310*	**370**
other	1020	…	…
Sum	**1080**	…	sum

	*c*1	other	Sum
*w*3	**40**	*420*	**460**
other	1040	…	…
Sum	**1080**	…	sum

	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-15	Sum	types	*H*rel
*w*1	**80**	*90*	*45*	*35*	*25*	*5*	*0*	**280**	6	0.578
*w*2	**60**	*0*	*310*	*0*	*0*	*0*	*0*	**370**	2	0.164
*w*3	**40**	*30*	*30*	*30*	*30*	*30*	*270*	**460**	15	0.999
*w*4	40	407	1	1	1	1	9	460	15	0.182
*w*5	40	420	0	0	0	0	0	460	2	0.11
*w*6	40	1	407	1	1	1	9	460	15	0.182
*w*7	40	0	420	0	0	0	0	460	2	0.11
*w*8-20	…	…	…	…	…	…	…	…	…	…
Sum	**1080**	1080	10800	…	…	…	…	sum	15	…
types	…	…	…	…	…	…	…	…	…	…
*H*rel	…	…	…	…	…	…	…	…	…	…

	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-*n*	Sum	types	*H*rel
*w*1	…	…	…	…	…	…	…	…	…	…
*w*2	…		*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-*n*	Sum	types	*H*rel
*w*3	…	*w*1	…	…	…	…	…	…	…	…	…	…
*w*4	…	*w*2	…		*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-*n*	Sum	types	*H*rel
*w*5	…	*w*3	…	*w*1	…	…	…	…	…	…	…	…	…	…
*w*6	…	*w*4	…	*w*2	…	…	…	…	…	…	…	…	…	…
*w*7	…	*w*5	…	*w*3	…	…	…	…	…	…	…	…	…	…
*w*8-*m*	…	*w*6	…	*w*4	…	…	…	…	…	…	…	…	…	…
Sum	…	*w*7	…	*w*5	…	…	…	…	…	…	…	…	…	…
types	…	*w*8-*m*	…	*w*6	…	…	…	…	…	…	…	…	…	…
*H*rel	…	Sum	…	*w*7	…	…	…	…	…	…	…	…	…	…
		types	…	*w*8-*m*	…	…	…	…	…	…	…	…	…	…
		*H*rel	…	Sum	…	…	…	…	…	…	…	…	…	…
				types	…	…	…	…	…	…	…	…	…	…
				*H*rel	…	…	…	…	…	…	…	…	…	…

Figure 2: The cline of co-occurrence complexity that collostructional phenomena require. (Note: these panels are taken from Gries 2012, Figures 4–8).

on frequencies would take into consideration: w_1 occurs in the construction in question c_1 80 times, w_2 occurs in the construction in question 60 times, etc.

The next panel represents the information that an account based on *AM*s (such as CA) would take into consideration. For each of the words in the first panel, this means not just taking the frequencies of co-occurrence into

consideration (i.e., 80, 60, 40, ...) but constructing the above kind of 2×2 tables for each $w_{1\text{-}n}$ in each of which the *a*-cell then contains the co-occurrence frequencies (again 80, 60, 40, ...); as discussed above, this means that for each table, the overall frequencies of the words $w_{1\text{-}n}$ as well as the overall frequency of the construction c_1, and a frequency for cell *d* is required.

The third panel, then, recognizes that the 'other' rows and columns in the second panel are a very simplistic conflation of all the types not-w_1, not-w_2, or not-w_3. This is because, e.g., in the leftmost 2×2 table of panel 2 (**80**, *200*, 1000, ...), it is not clear at all

- how many different words other than w_1 occur in the construction c_1 and what the overall frequency distribution of these frequencies is; in the third panel, for instance, we can see that w_4 also occurs in c_1, namely 40 times, etc.;
- how many constructions other than c_1 do words w_1, w_2, ..., w_n occur in; in the third panel, for instance, we can see that w_1 also occurs in c_2, namely 90 times, and that w_1 occurs in altogether six constructions.

Thus, panel 3 represents a *full cross-tabulation* approach, in which for all words and constructions under considerations co-occurrence and overall frequencies are included in the analysis. A particularly useful side-effect of this is that this also yields the type co-occurrence frequencies of all words and all constructions involved (e.g., 6 for w_1) as well as the entropies $H_{1\text{-}n}$ of all their distributions; for instance, the relative entropy H_{rel} of the frequency distribution of w_1 over constructions $c_{1\text{-}15}$ (80, 90, 45, 35, 25, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0) is 0.578. Entropies are important because they are correlated with productivity, contextual distinctiveness/diversity, and learning. For instance, Goldberg et al. (2004) show how a more skewed distribution (with a higher entropy) is learned better than a more balanced one (with a lower entropy); cf. also Redington et al. (1998), McDonald and Shillcock (2001), Mintz et al. (2002), Recchia et al. (2008), and Baayen (2010).

Finally, the fourth panel represents the data that one would have to take into account if one recognized that everything represented in the third panel may in fact differ dramatically across different corpora / parts of a corpus. Thus, ideally one would figure the *dispersions* of all of the co-occurrence frequencies into the computation of the association of words and constructions because grammatical co-occurrence patterns can be very register-specific (e.g., *fold* and *process* 'like' to occur in the imperative only in 1 of 500 corpus files; cf. Stefanowitsch and Gries 2003; Gries 2008, 2012).

One obvious question that arises from all of Figure 2 is how CA can possibly work at all when what it does is based on the very simplistic panel #2 and when what one should be doing is represented in panels #3 and #4. A statistical

reason it works is that the token distributions in the 'other' rows/columns of the second panel will be Zipfian most of the time with high frequencies in the c_1 column (especially for the central verbs); another reason is that FYE is correlated highly with an interesting association measure called ΔP (in particular $\Delta P_{\text{construction} \rightarrow \text{word}}$) from the associative learning literature (cf. Ellis and Ferreira-Junior 2009; Gries 2013). I will return to this issue below.

3 Schmid and Küchenhoff (2013) and a rebuttal

In this section, I will briefly summarize the main claims of S&K.

3.1 The difficulty of evaluating and comparing CA

One of the first points of critique is this: "It has proven rather difficult to evaluate the theoretical background assumptions and cognitive underpinnings of collostructional analysis and to compare them to alternative ways of modelling lexicogrammatical attraction phenomena" (S&K, p. 531).

This first point of critique is reminiscent of JB's somewhat similar point of critique. That means, however, that it has already been dealt with in Gries (2012) and is in fact not particularly appropriate in the first place. As for the former, the theoretical underpinnings, the fact that CA has "been welcomed quite enthusiastically by many members of the corpus-linguistic and cognitive-linguistic communities" (S&K, p. 538) makes it hard to believe that the cognitive underpinnings were difficult to evaluate – why would so many authors adopt a method if they didn't understand, and agree with, its underpinnings? In addition and as mentioned above in Section 2.2.4, Gries (2012) already lists a variety of ways in which the initial CA studies – in particular Stefanowitsch and Gries (2003) – motivate the cognitive underpinnings of CA quite clearly. Finally, Gries (2012) has nearly nine pages on cognitive underpinnings of CA, as when I discuss the above-mentioned cline of co-occurrence complexity, CA's relation to type frequencies, entropies, Zipfian distributions, dispersion, and a new definition of the role that "sufficient frequency" plays for constructions.

As for the latter, comparisons between methods, comparisons between FYE and "alternative ways of lexicogrammatical attraction" have been very easy right from the start: Even early versions of <coll.analysis.r> have offered five different measures of collexeme strength: FYE was a recommended default, but G^2, *MI*, chi-squared, and a logged odds ratio were always available for users, and regardless of the user's choice, S&K's reliance, $\Delta P_{\text{construction} \rightarrow \text{word}}$, and

$\Delta P_{\text{word} \rightarrow \text{construction}}$ were always outputted, too ... It is therefore not clear to me how this is supposed to make it difficult to evaluate alternative ways. In sum, I don't think this initial criticism of S&K has any factual merit.

3.2 CA requires a more powerful computer

S&K also criticize a particular kind of result that can be returned by my <coll.analysis.r> script that most users of collostructional analyses have been using: Sometimes, collostructional analysis can return a value of 0 (or Inf, if the negative log is used, as most users have done), which "can only be interpreted as representing a maximum degree of collostruction strength which could only be rendered more precise by using a more powerful computer" (S&K, p. 537).

The claim that CA requires a more powerful computer, unfortunately, is just plain wrong: One does not need a more powerful computer at all – if one uses R, all one needs is to be able to use computational methods that extend the operating system's use of its numerical computing abilities by, say, using the Rmpfr library (Maechler 2014). Consider, for instance, the two panels in Figure 3.

	M	N	Totals
A	20	4	24
B	3	5	8
Totals	23	9	32

	M	N	Totals
A	2000	4	2004
B	3	500	503
Totals	2003	504	2507

Figure 3: Two 2 × 2 co-occurrence tables.

The table in the left panel can be easily tested for significance in R, both with functions from the base package and my own extension `(fisher.test.mpfr)`

```
> x <- matrix(c(20,3,4,5), nrow=2)
> fisher.test(x)$p.value
[1] 0.02331508
> sum(dhyper(20:24, 24, 8, 23))
[1] 0.02331508
> fisher.test.mpfr(x, precBits=2)
$p
1 'mpfr' number of precision 100 bits
[1] 0.023315079432988220529933544394054
```

But the right table cannot be tested for significance precisely with the functions from the base package, returning the value of 0 that S&K mention:

```
> y <- matrix(c(2000,3,4,500), nrow=2)
> fisher.test(y)$p.value
[1] 0
> sum(dhyper(2000:2004, 2004, 503, 2003))
[1] 0
```

However, on the very same computer (a 2.5-year old standard desktop computer with an Intel Core I7-2600CPU @ 3.4 GHz), computing the exact *p*-value is unproblematic and takes about a second:

```
> fisher.test.mpfr(y, precBits=2)
$p
1 'mpfr' number of precision 100 bits
[1] 2.90872379395496320493595395393675e-526
```

For the example of *give* in the ditransitive that S&K use to 'make their point', the relevant (logged) FYE-value becomes 697.72. Thus, no "more powerful computer" is needed – in fact, using a more powerful computer would do nothing without the right packages – just statistical programming knowledge: CA can very well handle even large frequencies and the 2014 version of <coll.analysis.r> (Gries 2014) provides that functionality.

3.3 Specific aspects of the computations of a CA

3.3.1 Issues regarding the randomness assumption

One of the first issues raised by S&K is concerned with the randomness assumption. First, they assert that "[s]ince there can be no doubt that all languages show [non-random] distributional patterning, it is clearly problematic if we proceed from that assumption" (S&K, p. 538; it would have been nice to quote JB, who made the exact same point on the exact same topic; see Section 2.2.1 above).

Several things can be said with regard to this point: (i) given the fact that AMs for CA and in general have a decades-long history in corpus linguistics and other fields, the prescriptive conceptual issues S&K raise cannot have too many practical ramifications. Also, (ii) just like discussed in Gries (2012) and

mentioned above, any criticism of CA based on the fact that most users used FYE really misses the point because CA does not *require* the use of a significance test. Specifically, this issue doesn't even arise if a user of <coll.analysis.r> either chooses another AM or simply ignores the AM part of the output and focuses only on the ΔP part of the output.

Then, (iii) the significance tests performed on the usual 2×2 co-occurrence tables do not assume complete randomness/independence of all the words in a corpus / in the language – the significance tests assume as a null hypothesis that a word would be as frequent in a construction as would follow from its overall frequency in a corpus, a much narrower assumption of randomness.

Finally, (iv) while S&K make this point a criticism of CA, it is actually much broader in nature. As is well-known and frequently discussed in statistical literature, the null hypothesis is really never true and the fact that that is the case is part of a much larger discussion in many empirical sciences to not use null-hypothesis significance testing at all (Cohen 1994 is a widely-cited discussion of this). This is therefore hardly anything unique to CA, but, if anything, something the authors could say of any study using inferential statistics. The fact of the matter is that *any* AM is a huge simplification, because, trivially, both S&K's *attraction* (p(word|construction)) and *reliance* (p(construction|word)) as well as Stefanowitsch and Gries's FYE fail to capture most of the information of the actual multidimensional co-occurrence space. Given that there is not a single study at this point that begins to approximate the real complexity of the distributional data, one needs to realize that we're all just approximating in very coarse-grained ways by using heuristics (such as statistical independence ≈ lack of interesting linguistic patterning; cf. also Gries 2012: 486–487).

A second point of S&K's critique related to the (lack of) randomness of the data involves the fact that speakers in corpora usually contribute multiple data points and that corpus-linguistic studies often do not take that into consideration. However, while this is true and can affect significance tests based on corpora (unless one uses mixed-effects/multi-level modeling; cf. Gries (to appear) for an introduction), S&K simply assert that this problem is "aggravated" (p. 538) with significance testing but do not provide any proof for that claim.[6] In the absence of more refined analysis, it is not clear why significance tests should suffer more from this than, say, S&K's percentage-based measures: isn't it a problem if nearly

6 S&K refer to "mixed models including random effects, e.g., for speakers and sources," but, strangely enough, then add that "it is not clear how these could be applied in order to improve measures of lexicogrammatical associations. Later in their paper, however, they cite Baayen (2011), who, on p. 315, clearly states how mixed-effects models *can* be applied to (distinctive) collexeme analyses.

all data points that lead to an *attraction* or a *reliance* value of 0.7 are from only one or two speakers? Any corpus measure not taking dispersion into consideration suffers from that problem, FYE, (logged) odds ratios, *attraction*, ΔP, ...

3.3.2 Issues regarding the use of FYE

Perhaps somewhat predictably following from the previous section, S&K's next major point of critique revolves around the use of FYE, i.e., a significance test. Specifically,

- they again note that the use of a *p*-value is an unusual choice because such a *p*-value is not strictly speaking an effect size but a measure of "the likelihood with which the assumption that there is no attraction, i.e., the null hypothesis, can be rejected" (p. 539);
- they claim "it is not quite clear in which way the Fisher Exact *p*-value indeed [...] incorporates the size of the effect" (p. 539).

It is true that a *p*-value is a somewhat unusual choice, but S&K's criticism is still not particularly useful. First, they do not engage with any previous cognitive- or corpus-linguistic literature on the topic. For instance, they do not discuss any general corpus-linguistic work such as Pedersen (1996, 1998) as the first to suggest FYE as an AM, or Evert (2009: 1,235), who states "Mathematicians generally agree that the most appropriate significance test for contingency tables is Fisher's exact test." Also, they do not discuss any of the desirable characteristics of FYE mentioned above in Section 1.2. They do not even discuss Stefanowitsch and Gries's (2003: 239) detailed discussion of another useful characteristic of FYE, namely that, precisely because it is a significance test, it assigns more weight to effects that are based on more data points: An FYE would assign more importance to a reliance value of 0.2 if it based on 100 out of 500 rather than 1 out of 5 instances because, in the former case, the finding is likely more robust and the association is more entrenched (if one considers the corpus data of representative of an individual's cognitive system) or conventionalized (if one considers the corpus data representative of 'a speech community's linguistic system'). Thus, any linguist interested in exemplar models or entrenchment should appreciate this; cf. (2012: 483) for relevant quotes.

Second, on top of ignoring the theoretical discussion available on the issue, they also ignore existing CA findings or comparisons such as the above-mentioned (Section 2.2.1) comparison of FYE to *MI* and to log odds of Gries (2012) using the ditransitive, which provided strong evidence that, of those three measures, FYE performed best.

Finally, and this is an issue I will return to in a moment, in published studies, the exact *p*-values have hardly mattered much beyond providing the words' ranking: If the reader recalls the above four-step procedure that underlies virtually all applications of AMs, then the reader will also recall that the exact AM-values are nearly always only used to rank the collexemes/types and to identify the, typically, top 10–50 collexemes – whether a collocation is significant in the traditional sense is often irrelevant (cf. Stefanowitsch and Gries 2003: 239, note 6 for an early CA-related mention of this).[7]

Thus, while the use of a *p*-value is indeed unusual from a statistics-textbook perspective, there are good theoretical reasons to use FYE as a heuristic, good empirical findings that this heuristic can do its job, and a great degree of what one might not like about it numerically will not matter since it doesn't affect the rank-ordering of collexemes.

The second claim of S&K is a bit perplexing. While it is true that *p*-values are not effect sizes, *p*-values by their very nature reflect a combination of different things including the size of the sample(s), the variability of the sample(s), and the effect size. This means, all other things being equal (i.e., given a particular sample size and a certain degree of variability of the data), a stronger effect will lead to a smaller *p*-value (and a larger negative log of the *p*-value).

For a more concrete example, consider the two co-occurrence tables in Figure 4, which both have the same sample size (4,320) and the same marginal totals (120, 4,200, 220, and 4,100):

	Constr. *c*	other	Totals
Word *w*	20	200	220
other	100	4000	4100
Totals	120	4200	4320

	Constr. *c*	other	Totals
Word *w*	40	180	220
other	80	4020	4100
Totals	120	4200	4320

Figure 4: Two 2 × 2 co-occurrence tables.

As can be seen, the left table has an odds ratio of 4 and a logged FYE-value of 5.74 whereas the right one, in which the co-occurrence of *w* and *c* is much higher (given identical marginal totals), has an odds ratio of 11.12 and a correspondingly higher logged FYE-value of 22.69. And this is not a hand-picked case: there is a highly significant and strong correlation between a pure effect size that is blind to sample size (such as odds ratios) and *p*-values from FYE when tables have

7 The irrelevance of whether a particular co-occurrence is significant or not is underscored by the fact that hardly anyone corrects all the *p*-values for multiple post-hoc tests as would be necessary to control the overall error rate (cf. Gries 2005a for exemplification).

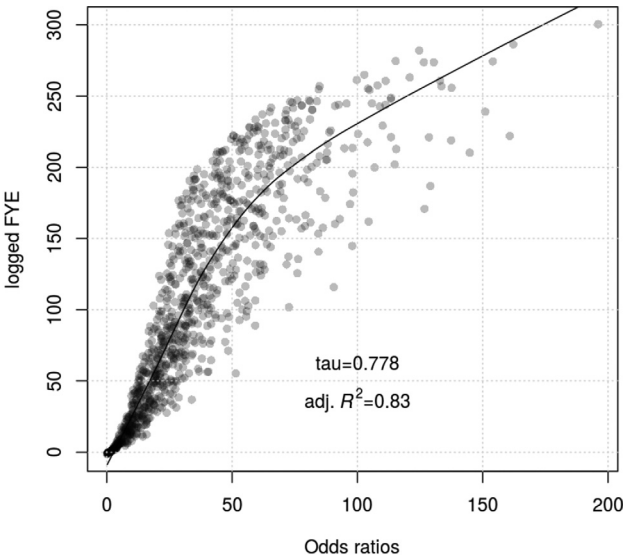

Figure 5: The correlation between FYE and odds ratios in 1,000 tables.

distributional characteristics typical for CA tables (namely $a < b$, $c <<< d$); this is represented in Figure 5 on the basis of 1,000 simulated pseudo-random co-occurrence tables.[8] (Recall also from above that Ellis and Ferreira-Junior find extremely strong correlations between FYE and $\Delta P_{\text{construction} \rightarrow \text{word}}$.) Thus, the claim that it is unclear how FYE incorporates effect size is somewhat questionable: yes, the two are by no means identical, but can be highly correlated.

8 The simulated values plotted here were computed in R (R Core Team 2014) as follows:

```
rm(list=ls(all=TRUE)); set.seed(1)
fye <- or <- rep(0, 1000)
for (i in 1:1000){
 aa <- sample(1:200, 1)        # create an a-value between 1 & 200
 bb <- sample(201:600, 1)      # create a b-value between 201 & 600
 cc <- sample(201:600, 1)      # create a c-value between 201 & 600
 dd <- 50000-sum(aa, bb, cc)   # create a d-value
 qwe <- fisher.test(matrix(c(aa,cc,bb,dd), nrow=2)) # compute FYE
 or[i] <- qwe$estimate # get the matrix' s odds ratio
 fye[i] <- ifelse (or[i]>=1, -log10(qwe$p), log10(qwe$p)) # log
}
cor.test(fye, or, method="kendall") # Kendall's tau for or & fye
summary(lm(fye~poly(or, 2))) # correlation between or & fye
```

3.3.3 Corpus size and cell *d*

Finally, S&K raise two other, related points: the issue of the sample size (and its impact on FYE-values) and, just like JB, the computation of the amount to be entered into cell *d* of the co-occurrence tables.

First, they comment on the fact that logged FYE-values are correlated with the sample size and that, given the large corpora that are often available, null hypotheses are rejected "more or less automatically" (p. 540); they refer to Kilgarriff (2005) in that connection. The first part of their statement is correct, the second one is not. Yes and as already discussed above, all other things being equal, *p*-values in general will decrease as sample sizes increase.

However, this does *not* mean that null hypotheses are rejected more or less automatically: In a response to Kilgarriff (2005) cited by S&K (p. 540), Gries (2005a) shows that, if one applies the required corrections for multiple post-hoc testing to association measures based on *p*-values, then the number of presumably false positives decreases drastically; in one case, when random word co-occurrences were tested, the post-hoc correction brought the proportion of false positives down to a value close to the traditional 5%. Thus, S&K's argument only applies if one does what one shouldn't do – not correct for multiple testing. In the case of a 'proper CA', that is not even necessary, because the FYE-values are only used for ranking, but if one studies the exact values, as S&K imply, then the typically required corrections make sure that not everything becomes significant. Incidentally, even the data S&K themselves discuss do not bear out their claim: They discuss the *as*-predicative data of Gries et al. (2005, GHS henceforth) and of the 107 verb types attested in that construction, 55.1% (59) are significantly associated with the construction without post-hoc correction, and only 23.4% (25) are significantly associated when a post-hoc correction (Bonferroni) is applied. While there are more powerful corrections, the results clearly show that significant results are definitely not returned automatically as S&K (and Kilgarriff) would have us believe.

As for the cell *d* issue, I have already discussed the points that Gries (2012) made before S&K bring up the same issue. And S&K are right that "[t]he size of the score in cell no. 4 has a strong effect on the *p*-values calculated by the test" (p. 544). However, since step iii. of the vast majority of CA/AM studies involves ranking the co-occurring types and since changing the corpus size will affect each type in a similar (though not identical) way, it is no surprise that even JB, who was otherwise quite critical of CA, too, found that corpus sizes didn't affect the overall results much, and Gries (2012) showed with simulation data how little the rank-ordering of collexemes varies as a function of corpus size.

More generally, S&K of course have a point when they caution that the identification of the cross-tabulated items – in a CA, the construction in question

and the words occurring in its relevant slot – must be identified and defended as carefully as possible, an obvious desideratum of any research project that applies here as well. Similarly, S&K are also correct in pointing out that deciding on a corpus size, and thus the cell *d*, may imply theoretical commitments an analyst may or may not be comfortable with. Again, though, this point is somewhat obvious: any operationalization in any empirical study requires that the analyst or annotator consider the best (or the most efficient) criteria to boost the validity of their study, so one, obviously, has to choose a hopefully useful level of granularity for the relevant corpus counts along the lines discussed above. Thus, these decisions about how to define the words and construction under consideration to get the frequencies *a*, *b*, and *c* are not at all particular to CA but apply to any contingency table and its study, be it with FYE, *MI*, *attraction*, *reliance*, (log) odds), ... and can thus not really be held against CA – it is only the cell frequency *d* that *attraction* and *reliance* need not consider, but then that number plays no decisive role for the rank-ordering of collexemes, which makes this point of critique rather toothless.

3.3.4 Directionality of association

A final point of critique regarding CA is concerned with the fact that FYE is a bidirectional AM. Put differently, it summarily states how much a word and a construction 'like' each other, but it cannot differentiate between the degree that the word 'likes' the construction and the degree to which the construction 'likes' the word. This is absolutely correct and precisely the reason why the script that most CA papers have used has been outputting directional ΔP-values for years in addition to whichever AM a user chose. Thus, this is again only an argument against FYE, but neither one against the tool that most people have used for CA nor one against CA in general, i.e., the idea that one can glean something from the however quantified degree to which words are associated to constructions.

3.4 S&K's suggestions and data

Section 4 of S&K deals with considering alternative approaches where, in the spirit of the above, it needs to be borne in mind that the discussion is largely (though not exclusively) about alternatives to FYE, i.e., the most frequently used measure of collexeme strength – not alternatives to CA. In this section, I will first discuss their theoretically-motivated discussion (Section 3.4.1) before I turn to the empirically-motivated discussion they provide (Section 3.4.3).

3.4.1 Theoretical considerations regarding directional measures

Following up on their (perfectly correct) observation that FYE does not include information about the directionality of association between a word and a construction, S&K discuss alternative measures. The first of these are the measures of *attraction* and *reliance*, which are essentially conditional probabilities as represented in Figure 6 and are correlated with the corresponding ΔP-values that <coll.analysis.r> has been providing since 2007.[9]

	Constr. *c*	other	Totals
Word *w*	*a*	*b*	*a* + *b*
other	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

$$attraction_{\text{word to construction}} = {}^{a}/_{a+c} = p\,(\text{word} \mid \text{construction})$$
$$\Delta P_{\text{construction} \rightarrow \text{word}} = ({}^{a}/_{a+c}) - ({}^{b}/_{b+d})$$
$$reliance_{\text{word on construction}} = {}^{a}/_{a+b} = p\,(\text{construction} \mid \text{word})$$
$$\Delta P_{\text{word} \rightarrow \text{construction}} = ({}^{a}/_{a+b}) - ({}^{c}/_{c+d})$$

Figure 6: Directional measures of collexeme strength.

S&K mention the following as advantages of these two conditional probabilities:

i. no estimate for cell *d* is necessary;
ii. they are "straightforward descriptive measures which allow for clear and unambiguous interpretation" (S&K, p. 550);
iii. "no assumptions about the stochastic structure and the random distribution in the corpus have to be made" (S&K, p. 550);

... and they mention the following two potential downsides:

iv. because cell *d* is not included, the number of competing constructions and the confidence one can have in the significance of the data is not factored in;
v. simple rank-ordering is impossible because one now uses two measures, not one.

In addition, S&K also note (p. 551) that *attraction* and *reliance* are extremely similar to the two ΔP-values but that, given the very high correlations between *attraction* and *reliance* and the corresponding ΔP-values and the fact that the former do not require filling cell *d*, *attraction* and *reliance* "may do the job just as well as the two ΔP measures" (p. 552).

As mentioned above, I think the fact that FYE is bidirectional is indeed a downside compared to other measures and I have argued for directional

9 In the *as*-predicative data discussed in GHS, the correlations between *attraction* and $\Delta P_{\text{construction} \rightarrow \text{word}}$ on the one hand and *reliance* and $\Delta P_{\text{word} \rightarrow \text{construction}}$ on the other are 0.987 and 0.999 respectively.

measures in other contexts myself (cf., for example, Gries 2013). However, some comments regarding S&K's discussion are nonetheless necessary. As for i., the effect that cell *d* has on the rankings of the collexemes is very small, as was admitted in JB in 2010, shown in the first review that S&K received in 2011, and in Gries (2012). Thus, this is not really a great advantage, if any. In addition, S&K miss one very important point. If one uses *attraction* and *reliance*, one loses the ability to determine whether a particular word occurs more in the construction than would be expected given the verb's frequency in the corpus. For instance, *attraction* and *reliance* for *give* and the *as*-predicative are 0.0044 and 0.0026 respectively, which doesn't tell the analyst that this is actually an instance of what in CA has been called *repulsion*: *give* occurs less often in the *as*-predicative than expected given *give*'s frequency in the corpus – the Δ*P*-values of −0.004 and −0.002 on the other hand reveal the repulsion by the negative sign.

As for ii., the ranking of FYE-values (rather than the actual FYE-values), which is what analysts are apparently mostly considering, is a ranking along a single dimension – how is that not straightforward and descriptive?

As for iii., S&K are distorting the picture a little here: First, yes, the null hypothesis is never true, but, as discussed above, the randomness assumption is only a (extremely widely used) heuristic and FYE does not test against a completely random distribution of all words in a corpus, but a proportionally random distribution of words in a constructional slot, a much more defensible proposition. Plus, even if a descriptive percentage makes no randomness assumption at all *statistically*, it makes one *conceptually*: The degree we would want to rely on the percentage is still dependent on (i) the corpus being a representative sample of the speech variety we're studying and (ii) the fact that hopefully not all data points entering into a percentage are provided by an extremely small number of speakers and their idiosyncrasies. In other words, computing a percentage indeed makes no stochastic assumptions of randomness but interpreting one relies on the assumption that the data figuring into the percentage are not useless/distorted. This doesn't invalidate *attraction*/*reliance* or the Δ*P*-values but let us be honest and explicit about how the different measures behave and what they require.

Now, as for iv., S&K are right that the significance is not factored in or returned, but given that few CA studies seem to invoke the notion of actual statistical significance, I don't think this needs to be held against *attraction*/*reliance* and the Δ*P*-values. As for v., yes, simple rank-ordering is harder now, but (i) the directional measures do provide more information, which may be informative as in Figure 7, which plots *reliance* against *attraction* for the *as*-predicative data of GHS (note the different scaling of the two axes). Note that one could still condense *attraction* and *reliance* into one value (for the sake of

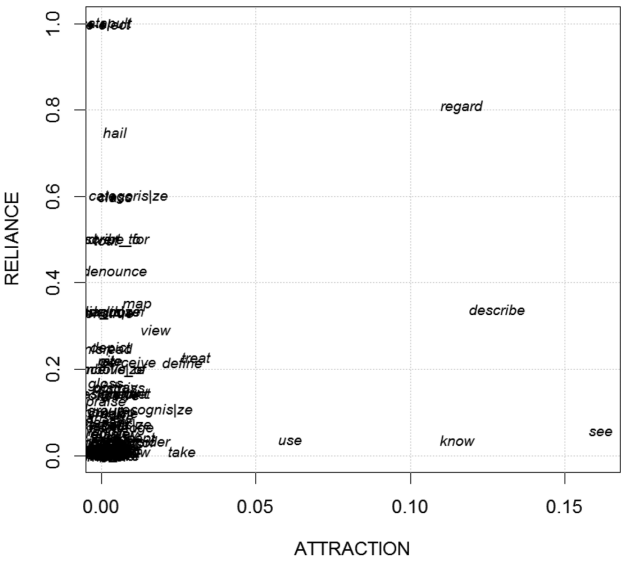

Figure 7: *Attraction* and *reliance* for the *as*-predicative data of GHS.

rank-ordering) by computing the Euclidean distance of a verb's position from the origin. For *regard* and *see* one obtains 0.816 and 0.171 respectively, which underscores the importance of *regard* for the *as*-predicative (compared to *see*'s). Even more interestingly, if one multiplies these Euclidean-distance scores with the verbs' frequencies in the *as*-predicative (let's call this frequency-adjusted *attraction*-and-*reliance* scores, FARS), the resulting values are nearly perfectly correlated with FYE (adj. $R^2 = 0.987$); this in turn reinforces something that CA has relied on for a long time: FYE combines mutual attraction and frequency in a very nice way, and this also means that scholars that want to avoid *p*-value AMs could use FARS to get the same result without the null-hypothesis significance testing baggage.

3.4.2 Theoretical considerations regarding unidirectional measures

S&K then turn to the odds ratio. They provide a long discussion of this measure, which is essentially quite simple: On the basis of Figure 6 above, it can be computed as $(^a/_b)/(^c/_d)$ or $(^a/_c)/(^b/_d)$. One advantage of the odds ratio according to S&K is that it is an effect size. That is, unlike FYE, which is based on a significance test, the odds ratio is not correlated with the sample size. However, S&K's discussion of the odds ratio is fraught with problems.

First, it is true that the independence of the sample size has advantages, e.g., when it comes to comparing results from different corpora – note, though, that one is often just as able to compare the rankings of FYE values across corpora. However, the odds ratio can easily return high values for rather small observed frequencies, something which FYE is less likely to do simply because the observed frequencies figure into its computation. This is more than just an academic issue because already the review that S&K received for their first submission as well as Gries (2012) compare ditransitive results of FYE and (log) odds and show that it is exactly for that reason that the FYE results seem more attractive: The top collexemes of the ditransitive according to FYE are *give*, *tell*, *send*, *offer*, *show*, *cost*, *test*, ... whereas those of log odds are *give*, *accord*, *award*, *allocate*, *profit*, ... I submit that the reader, just like many audiences of talks, would find the former list more appropriate. It is not clear to me why S&K refused to engage with this argument that was presented to them twice ...

The second problem is that S&K not only employ what appear to be double standards, but they also contradict themselves. As for the double standards, how is it that FYE is supposed to be a bad measure because it requires cell *d*, but the fact that the odds ratio does, too, is mentioned but apparently not really such a bad deal? As for contradicting themselves, S&K argue "Odds Ratio is superior to the Attraction and Reliance approach in that it is both frequency-adjusted and bi-directional" (p. 555). Let me retrace the logic here: Earlier, S&K argued that FYE is worse than *attraction* and *reliance* because FYE is bidirectional whereas *attraction* and *reliance* are unidirectional and, thus, more precise (and I even agree at least with regard to the directionality argument). But now, the odds ratio is "superior" (!) to *attraction* and *reliance* because it is bidirectional ... I have to admit the logic of this course of argumentation eludes me.

3.4.3 Empirical considerations

S&K then turn to empirical data and compare rank correlations between *attraction*, *reliance*, the corresponding ΔP-values $\Delta P_{\text{construction} \rightarrow \text{word}}$ and $\Delta P_{\text{word} \rightarrow \text{construction}}$, FYE, and the odds ratio. They find that, unsurprisingly the odds ratio is highly correlated with *reliance* and $\Delta P_{\text{word} \rightarrow \text{construction}}$, whereas FYE is more highly correlated with *attraction* and $\Delta P_{\text{construction} \rightarrow \text{word}}$. However, even from their very own data it emerges that, of the two bidirectional measures – FYE and the odds ratio – it is FYE that captures *both attraction* and *reliance* more than the odds ratio, which really only captures *reliance*.

More interesting and revealing than this comparison, however, is S&K's discussion of GHS. S&K make a correct observation: With hindsight, GHS should have probably not grouped the collexemes of the *as*-predicative into high- and low-frequency and collexeme strength groups given the information loss that that practice incurs.

Secondly, S&K criticize GHS for "a problem with the design of the sentence-completion task," (p. 561). Specifically, they say CA "essentially ranks lexemes [...] according to their potential to be attracted by a construction. It thus effectively proceeds from an Attraction perspective" (p. 561). That already is not strictly speaking true: Earlier in their paper, S&K went to great lengths to point out the FYE is not directional, but now they imply directionality in their formulation – what CA does is it ranks lexemes according to how much lexemes and the construction are attracted to *each other*. In addition, in their own paper (Table 9 on p. 556) they show that FYE is in fact highly correlated with both *attraction* and *reliance*. Yes, the correlation of FYE with *attraction* is higher (–0.88), but (i) the one with *reliance* is certainly also strong (–0.72) and (ii) in their own Table 13, FYE is then more strongly correlated with *reliance* than with *attraction*. Thus, even in their own analyses, FYE is not as uniformly biased in one direction as they imply at first.

The biggest problem arises, however, when they reanalyze GHS's (2005) data. While S&K criticized our earlier study for not including all potential variables and not all interactions between them (p. 559), what they do falls short on so many more levels: they

- reduce all numerical values to ordinal measures – which also incurs a loss of information just like what they criticized us for (though admittedly not as much as ours);
- run pairwise Spearman rank correlations on all variables – which means, their analysis does not even take all data into consideration at the same time, i.e., is not multifactorial in nature;
- ignore what CA was criticized for earlier in the paper, namely that the data points are not independent of each other. It's true they do not report *p*-values but even the interpretation of their correlations runs the risk of reflecting speaker-specific or verb-specific idiosyncrasies, which is why the state-of-the-art today would have been a multifactorial model with random effects as needed.

As if all that was not problematic enough, it is interesting to note which measures of collexeme strength is most strongly correlated with the subjects' sentence completions *in their own analysis*: FYE ... The difference from the next-ranked measure is probably not statistically significant, but it is revealing that,

after all the discussion of the presumed theoretical and empirical shortcomings of FYE, it is FYE that comes out best even in S&K's own evaluation.

Space does unfortunately not permit a detailed reanalysis of the experimental data here but in a first and as yet informal attempt to better analyze the complete experimental data, I first added a bunch of predictors to the data so that the final set of potentially relevant variables included logged FYE, *attraction*, *reliance*, $\Delta P_{\text{construction} \rightarrow \text{word}}$, $\Delta P_{\text{word} \rightarrow \text{construction}}$, the Kullback–Leibler divergences for *attraction* and *reliance*, and the log odds ratio. Then, two different methods were attempted.

First, in order to be able to use a regression-based approach, I applied a principal components analysis to all the CA measures. A screeplot suggested two principal components: one on which actually all variables loaded very highly and that accounted for 66.1% of the variance of all eight measures, and a second one that separated all attraction-related measures and FYE from all reliance-based measures and log odds ratios (and that accounted for an additional 30% of the variance). These two principal components' scores were then used as predictors in a generalized linear mixed-effects model that tries to predict the subjects' sentence completions (*as*-predicative: *no* vs. *yes*) on the basis of the two principal components and a predictor VOICE (whether the sentence fragment to be completed was for an *active* or a *passive* completion); in addition, the model had random intercepts for the items nested into the verbs (subject-specific intercepts were ns). The final model was somewhat uninteresting in that only the first principal component remained significant (in spite of a high classification accuracy: $C = 0.91$), and that component does not distinguish between the *attraction*- and *reliance*-related measures.

Thus, a second analysis was performed, a multi-model inferencing procedure (cf. Burnham and Anderson 2002; Kuperman and Bresnan 2012), a regression approach that generates a variety of models and weighs their regressions' coefficients proportionally to the degree to which each model deviates from the best model's performance/*AIC*. Models were fit with all eight measures and the above-mentioned random effects, and while the currently available implementation in R makes it difficult to include even pairwise interactions between all predictors, the two predictors that emerged as most important are FYE and log odds ratios, followed by the additional attraction-related measures $\text{KL-divergence}_{\text{attraction}}$ and $\Delta P_{\text{construction} \rightarrow \text{word}}$.

In sum, while further and more precise re-analyses of such data are required, criticism of CA based on the kind of re-analysis undertaken by S&K – with information loss, monofactorial pairwise correlations, and no consideration of the interdependence of data points they themselves mention as potentially problematic – can hardly be the answer, and if then FYE emerges as

the measure most correlated with the experimental results, it is hard to see why FYE is supposedly so bad.

3.5 Cognitive underpinnings

I have already noted above that, while S&K claim that it has "proven rather difficult to evaluate the theoretical background assumptions and cognitive underpinnings" (p. 531) of CA, both Stefanowitsch and Gries (2003), and Gries (2012) already discussed a variety of ways in which CA fits into a general cognitive-linguistic framework, namely by virtue of how the computation of an AM – in particular, but not necessarily FYE – relates to measures of verb-subcategorization preferences, conditional probability and thus cue validity and reliability, and entrenchment. Crucially, the already published Gries (2012) has nine pages on cognitive underpinnings, including a multi-page discussion of the above-mentioned cline of co-occurrence complexity and its relation to type frequencies and in particular entropy; recall Figure 2 from above. However, S&K's Section 6 newly 'develops' a 2–3 page argument toward full cross-tabulation, speaking for example of how one needs to take into consideration the frequencies of other words in the relevant construction and other constructions with a relevant word), and they represent this argument with what is shown here as Figure 8, which is in fact an exact visualization of the step from the 2nd to the 3rd panel of Figure 2 (i.e., the transition from Figure 4–5 in Gries 2012, their dotted arrows from cells *b* and *c* are the visual equivalent to how 'other' gets broken down into $w_{1\text{-}20}$ and $c_{1\text{-}15}$.). And how do they relate their argument to previous work? They say "[h]ints to some of these complications can be found in the literature" (p. 571) ...

There is one aspect of this plot that is potentially interesting, namely the notion of multiple attraction and reliance scores, but unfortunately there is no indication of how those would figure into the larger picture of the kind of multidimensional exemplar space that much usage-based work currently relies on.

4 Concluding remarks

I agree whole-heartedly with S&K that, given the widespread use and popularity that CA has enjoyed over the last decade, it is definitely useful to explore its assumptions and power; in fact, that is trivially true for any methodological innovation or application. On the whole, it seems that CA has been very useful

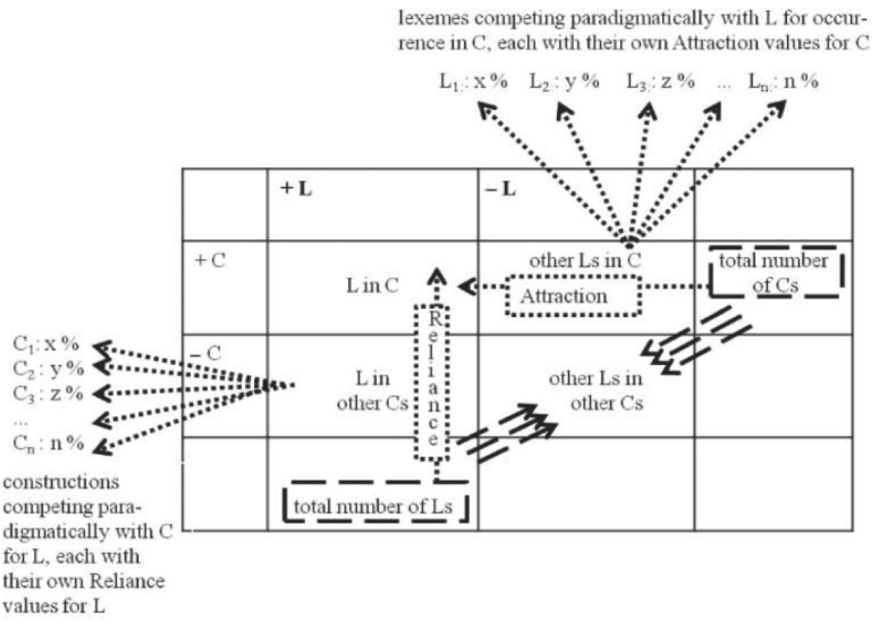

Fig. 1: Cotext-free and cotextual entrenchment as reflected in contingency tables (dashed arrows and boxes indicate scores and relations influencing cotext-free entrenchment; dotted arrows and boxes indicate scores and relations influencing cotextual entrenchment)

Figure 8: S&K's Figure 1 (p. 570).

for many researchers who have done insightful analyses and apparently have not found it hard to interpret the results they obtained. Now, if S&K were nevertheless right on track with all their criticism of CA, what would one expect them to present as evidence for all their claims? Two things: (i) examples that show how CAs yielded non-sensical/useless results and (ii) discussions that show how these bad results are due to the points of critique that they mention. However, S&K

- do not present a single example of a CA study whose results did not make sense (and do not acknowledge the presence of all the CA studies that make sense);
- therefore, do not show that CA produces bad results precisely because of the 'shortcomings' they discuss;
- ignore, on a variety of occasions, both published literature (JB) and Gries (2012) that discussed critical aspects of CA before them and they ignored empirical evidence presented to them even after their first submission;

- present as their own arguments in favor of full cross-tabulation even when that case was made years earlier in multiple presentations and one year earlier in print;
- present an argumentation that is fraught with misrepresentations and contradictions and when they reanalyze published data, the measure they criticize most (FYE) actually comes out best.

It is because of the above issues – not because I disagree with them but because of the quality of their argumentation – that I think that, apart from maybe some very general points, S&K do in fact not contribute much to the discussion in the best of cases and is misleading about existing discussion in the worst.

What is needed instead? I think two main strands of research/discussion are needed. On the theoretical side of things, it is obvious that any computation of conditional probabilities or AMs is a huge simplification and so we need to explore what other kinds of information co-determine speakers' multidimensional exemplar spaces of knowledge. In Gries (2012), I argued that we need to consider three notions that are related to each in complex obvious way:

- *type frequencies*: First, these provide context for a co-occurrence frequency of a word and a construction: All other things being equal (such as token frequencies and entropies), a co-occurrence frequency *a* is less informative the higher the frequencies of types summarized in cells *b* and *c* (but see below). Second, they are related to productivity, which means for any not fully lexically-specified construction, a certain type frequency must be observed.
- *Zipfian distributions*: The recognition of a category such as a not fully-lexically specified construction will require a certain type frequency of things in its slots and, as much work by in particular Goldberg and Ellis has shown, the token frequencies of these types in the slots will be Zipfian. That 'Zipfianness' will be multivariate, meaning that it will apply on many formal as well as functional co-occurrence characteristics and that in turn is correlated with its entropy (Gries 2012: Section 5.2).
- *entropies*: Above in Section 2.2.5 I mentioned how the learnability of constructions is related in some way to the entropy of its type-token distribution. However, there is arguably more involved here than discussed there and in Gries (2012). Imagine two constructions c_1 and c_2, each occurring 70 times in a corpus and each occurring only with six verb types $w_{1\text{-}6}$. Imagine further, the type-token distributions $w_{1\text{-}6}$ in c_1 and c_2 of 10, 50, 3, 2, 3, 2 and 10, 10, 11, 12, 13, 14 respectively. If we're interested in w_1, for example, then its frequency in both constructions is the same, the constructional frequencies of c_1 and c_2 are the same, and the type frequencies of words in c_1 and c_2 are the same. However, the same token frequency of w_1 of 10 plays very different roles in

both constructions: removing w_1 from the distribution of c_1 reduces the entropy of that constructions' word frequencies by 0.45 bits (31.6%) whereas removing w_1 from the distribution of c_2 reduces the entropy of that constructions' word frequencies by 0.26 bits (10.2%) only. In other words, token and type frequencies are not enough to distinguish what are intuitively obviously very different distributions – one needs more precise information. Some of that – overall entropy of a row/column – was already discussed in the full cross-tabulation approach suggested in Gries (2012), but the above shows that entropy and related notions such as surprisal need to be considered even more, esp. given the role they play for learning and processing. In fact, Gries speculates on whether the notion of sufficient frequency that is part of Goldberg's (2006) definition of a construction can be approached from this angle, namely when a point cloud in multidimensional exemplar space reaches a distribution that is multivariately Zipfian enough – maybe many types, but with many small and just a few large token frequencies – then that results in small enough uncertainty/large enough homogeneity to constitute a (constructional) category. For instance, the co-occurrence of a particular verb (such as *give*) in a particular slot of a particular pattern (such as the ditransitive) especially with an animate recipient and an inanimate patient and a particular scenario (such as 'transfer') may be of such a high frequency and mutual predictability that it can constitute a prototypical center of a category of a ditransitive construction. In a multidimensional exemplar space, this could correspond to a particularly dense point cloud in a region of semantic space concerned with transfer scenarios so that, when a transfer of an inanimate object to an animate recipient is to be communicated, *give* is the verb that, metaphorically speaking, *give* sticks out most out of the (Zipfian) distribution of candidate verbs or is most attracted by the construction (*attracted* in the attractor-basin sense of the word as in Rodd et al. 2004 account of lexical polysemy).

The above is certainly far from a full-fledged theory of how all these things are related, but I believe the above articulates a starting point for a real discussion of how to theorize about multivariate co-occurrence data in place of a reductionist two-percentages approach.

On the empirical side of things, we need more detailed studies of the corpus and experimental data that we have. We need to do all the things that are now state-of-the-art: multifactorial mixed-effects modeling of the relation between corpus and experimental data, aided by careful exploration of collinearity, subject- and item-specific effects (where item-specific effects may be related to other distributional notions such as entropy and others), maybe multi-model

inferencing, or even methods that allow for the analysis of causal relations such as Bayesian modeling or structural equation modeling. Again, the first sketches of more detailed analyses proposed above can hardly address all the intricacies of the data, but provide first steps of what to do instead of simplistic pairwise rank correlations.

I am fully aware that this paper as a whole has a certain degree of negativity in it. However, I hope it has become clear why that is the case: S&K's characterization of both the method and underpinnings of CA left a *lot* to be desired: As alluded to in Section 1.1 above, S&K's many problems directly undercut their alleged goal of clarification and exploration, but at the same time several of these problems are not immediately obvious to readers and, apparently reviewers, which is why a certain degree of deconstruction of S&K was required. No doubt, S&K will disagree with just about everything I say but it is my hope at least that this deconstruction will stimulate the kind of discussion of co-occurrence data, their theory, and method that usage-/exemplar-based linguistics deserves.

Acknowledgments: I am grateful to John Newman, an Associate Editor, and another member of the editorial board for their feedback. In addition, I would also like to thank (in alphabetical order) Sandra C. Deshors, Adele E. Goldberg, Beate Hampe, Viola G. Miglio, and Doris Schönefeld for feedback. Finally, I am grateful to the audience of the UK Cognitive Linguistics Conference 2014, where much of the above was discussed. Obviously, the usual disclaimers apply.

Appendix: list of abbreviations

AM	association measure
CA	collexeme analysis
FYE	Fisher-Yates exact test
GHS	Gries et al. (2005)
JB	Bybee (2010)
MI	Mutual Information
S&K	Schmid & Küchenhoff (2013)

References

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

Baayen, R. Harald. 2011. Corpus linguistics and native discriminative learning. *Brazilian Journal of Applied Linguistics* 11(2). 295–328.

Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*. 2nd ed. New York: Springer.

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

Cohen, Jacob. 1994. The earth is round ($p<.05$). *American Psychologist* 49(12). 997–1003.

Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220.

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.

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*. London: Longman.

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3). 289–316.

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

Gries, Stefan Th. 2005a. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2). 277–294.

Gries, Stefan Th. 2005b. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.

Gries, Stefan Th. 2007. Coll.analysis 3.2a. A script for R to compute perform collostructional analyses. http://tinyurl.com/collostructions.

Gries, Stefan Th. 2008. Dispersion 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. 2014. Coll.analysis 3.5. A script for R to compute perform collostructional analyses (major update to handle larger corpora/frequencies). http://tinyurl.com/collostructions.

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., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: CSLI.

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.

Harris, Zellig S. 1970. *Papers in structural and transformational linguistics*. Dordrecht: Reidel.

Jaeger, T. Florian & Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In B. C. Love, K. McRae & V. M. Sloutsky (eds.), *Proceedings of the Cognitive Science Society Conference*, 1061–1066. Washington, DC.

Jaeger, T. Florian & Neal Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition* 127(1). 57–83.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.

Kuperman, Victor & Joan Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66(4). 588–611.

Maechler, Martin. 2014. Rmpfr: R MPFR – Multiple Precision Floating-Point Reliable. (Version 0.5-6, dated 5 Sept. 2014). http://cran.r-project.org/web/packages/Rmpfr/index.html.

McDonald, Scott A. & Richard C. Shillcock. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(3). 295–323.

Mintz, Toben H., Elissa L. Newport & Thomas G. Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science* 26(4). 393–424.

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1–2). 137–158.

Pedersen, Ted. 1996. Fishing for exactness. *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*. Austin, TX, Oct 27–29.

Pedersen, Ted. 1998. Dependent Bigram Identification. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI, July 28–30.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Recchia, Gabriel, Brendan T. Johns & Michael N. Jones. 2008. Context repetition benefits are dependent on context redundancy. *Proceedings of the Annual Conference of the Cognitive Science Society* 30. 267–272.

Redington, Martin, Nick Chater & Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 22(4). 435–469.

Rodd, Jennifer M., M Gareth Gaskell & William D. Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science* 28(1). 89–104.

Roland, Douglas, Frederick Dick & Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57(3). 348–379.

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.

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.

Stefanowitsch, Anatol & Stefan Th. Gries . 2009 Corpora and grammar. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook, Vol. 2*, 933–951. Berlin & New York: Mouton de Gruyter.