BRILL

COGNITIVE SEMANTICS 9 (2023) 351–386

COGNITIVE SEMANTICS
brill.com/cose

Overhauling Collostructional Analysis: Towards More Descriptive Simplicity and More Explanatory Adequacy

Stefan Th. Gries
University of California, Santa Barbara, USA and JLU, Giessen, Germany
stgries@linguistics.ucsb.edu

Received: 20 August 2022 | Accepted: 12 April 2023 |
Published online: 7 August 2023

Abstract

In this paper, I make two sets of suggestions of how collostructional analysis can be updated. One set of suggestions involves simplifying the analysis for descriptive/exploratory purposes while at the same time enriching it with bootstrapped confidence intervals. The other set of suggestions involves the idea that we should move away from a single kind of association measure for theoretical/exploratory purposes and instead quantify collostructional attraction as a tuple of, minimally, three ideally orthogonal dimensions, namely frequency, association, and dispersion, because only this kind of analysis will be able to address all the dimensions that are relevant to cognitive/usage-based approaches to constructions. In addition, I end with a (renewed) plea to take the notion of construction more seriously: Rather than looking at associations of constructions to forms, which many studies have basically amounted to, I would like us to 'go back to' looking at associations of constructions to constructions, i.e. to take the meaning/functional pole of constructions more seriously again and include sense/function in all kinds of collostructional analyses more.

Keywords

collostructions – collexeme analysis – chi-squared – residuals – bootstrapping

Published with license by Koninklijke Brill NV | DOI: 10.1163/23526416-BJA10056
© STEFAN TH. GRIES, 2023 | ISSN: 2352-6408 (print) 2352-6416 (online)

1 Traditional Collostructional Analysis

Construction Grammar is the perhaps central grammatical framework within the overall 'paradigm' of usage-based linguistics or cognitive linguistics, where its subject matter – constructions – is often defined with reference to Goldberg (2006:5):

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

One of the most widespread corpus-linguistic approaches/methods to study constructions is collostructional analysis, a family of three methods, all of which share applying an association measure-based approach towards the co-occurrence of constructions that usually differ in their levels of schematicity. The three methods differ in what the units are that are involved in these co-occurrence data:

1. **collexeme analysis** (Stefanowitsch and Gries, 2003): one looks at how much constructions (usually words/lemmas) (dis)like to occur in a slot of one usually more schematic constructions (as opposed to elsewhere) such as a syntactic, or argument structure, construction;
2. **distinctive collexeme analysis** (Gries and Stefanowitsch, 2004a): one looks at how much constructions (usually words/lemmas) (dis)like to occur in a slot of a usually more schematic construction as opposed to alternative, functionally similar constructions:
   a. distinctive collexeme analysis involves the comparison of two functionally similar constructions;
   b. multiple distinctive collexeme analysis involves the comparison of three or more functionally similar constructions;
3. **covarying collexeme analysis**: one looks at how much constructions (usually words/lemmas) in one slot of a more abstract construction (dis) like to co-occur with constructions (usually words) in another slot of the same more abstract construction:
   c. item-based covarying collexeme analysis: the standard case of co-varying collexeme analysis mentioned above (Gries and Stefanowitsch, 2004b);
   d. system-based covarying collexeme analysis: a correction to item-based covarying collexeme analysis that also considers the

frequencies of words outside of the construction in the corpus (Stefanowitsch and Gries, 2005).

Of these altogether 5 methods, collexeme analysis (1.) and distinctive collexeme analysis (2a.) have been by far the most widely-used, and the overwhelming majority of their applications involved computing a bidirectional association measure from the traditional 2 × 2 co-occurrence tables that are used for association measures in collocation or keywords research – as per the above, the three methods differ 'only' in terms of what the row and column units are. For a collexeme analysis (1.), for each word/lemma in a more schematic construction, one generates a table such as Table 1.
For a simple distinctive collexeme analysis (2a.), for each word/lemma in one of the more schematic two constructions, one generates a table such as Table 2. These methods have been applied in a wide variety of contexts, registers, and languages (e.g., Dutch, English, German, Italian, Mandarin Chinese, Standard Arabic, and many more). In addition, these methods have also been used in

TABLE 1 Input for a collexeme analysis to compute the attraction between w and c

	Construction *c* (e.g. the ditransitive)	**Other constructions**	**Sum**
Word/lemma *w* (e.g. *give*)	*a*	*b*	*a+b*
Other words/lemmas	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

TABLE 2 Input for a distinctive collexeme analysis to compute the attraction between w and c_1 vs. c_2.

	Construction c_1 (e.g. the *ditranitive*)	**Construction c_2 (e.g. the *prep. dative*)**	**Sum**
Word/lemma *w*, e.g. *give*	*a*	*b*	*a+b*
Other words/ lemmas	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

- psycholinguistic contexts (e.g. as a measure of verb-construction preferences that will help predict syntactic priming; see Gries, 2005a or Szmrecsanyi, 2006);
- second/foreign-language learning, e.g. to determine whether learners are acquiring, or have acquired, native-like verb-construction preferences; see Ellis and Ferreira-Junior (2009a, b), and especially Ellis, Römer, and O'Donnell (2016);
- work on language change; see Hilpert (2006, 2008).

There is also a growing body of work combining collostructional methods with experimental work, either for purposes of validation or to use collostructional results as predictors or controls. As for the former, in Gries, Hampe, and Schönefeld (2005, 2010), collexeme strengths ($-\log_{10} p_{\text{Fisher-Yates}}$) have a significantly higher degree of predictive power for sentence completions and self-paced reading times than raw frequencies or conditional probabilities alone; see also Backus and Mos (2011) or Flach (2020b). As for the latter, Gries and Wulff (2005, 2009) show how advanced German learners of English exhibit collostructional preferences similar to those of native speakers for both the dative alternation and the *to-/-ing* alternation (e.g. *I like* (*to swim*|*swimming*)) in both sentence completion tasks and acceptability ratings; see also Flach (2020a).

The relative simplicity and widespread use of (distinctive) collexeme analysis notwithstanding, these methods – in particular collexeme analysis – have also been discussed more controversially. Bybee (2010) criticized collexeme analysis on the basis of misunderstandings of the method's goals and results and Gries (2012) provided a detailed critcism of Bybee's claims. Schmid and Küchenhoff (2013) also criticized collexeme analysis (by repeating some of Bybee's misunderstandings and adding others), which are refuted in detail in Gries (2015). The present paper is, in a sense, a bit of a follow-up to (i) some of that discussion as well as to (ii) a recent first attempt of mine to make collostructional methods more useful (Gries, 2019). Specifically, I will discuss two sets of suggestions, which might seem contradictory at first glance, but which have to do with

1. the reason why one is performing a (distinctive) collexeme analysis – for mostly descriptive/exploratory purposes (at this point, probably the vast majority of cases) or for more theoretical/psycholinguistic reasons (up to this point, probably a small minority of cases)?
2. earlier criticism of how collexemic attraction has usually been measured, namely by means of a significance test (Fisher-Yates exact) performed on tables such as Table 1 or Table 2.

The first suggestion will be discussed in Section 2 and essentially amounts to a huge simplification (and speed-up) of how the analysis can be done; the proposal targets users who are interested in a (distinctive) collexeme analysis for descriptive/exploratory purposes. Specifically, I will outline how, instead of computing hundreds or thousands of potentially much more complex association measures (one for each word in a construction slot) one can instead compute a single, very simple test to get results that are essentially conceptually/interpretively very similar. I will explain the suggestion, exemplify it on the basis of three different brief applications, and then discuss one major benefit of the resulting speedup.

2 Collexeme Analysis: The Ditransitive

2.1 *Introduction*

To be able to explain and then evaluate the first suggestion, we need a data set to 'play with'. We will use a simplified data set (simplified from Stefanowitsch and Gries, 2003, that is) on the ditransitive construction from the ICE-GB. In order to set the stage for things to be discussed later, we will retrieve the input data for the collexeme analysis in a way that is a bit different from the usual format. Usually, the kind of input data that would in turn lead to tables such as Table 1 (from which one would then compute an association measure for each verb lemma) would look like Table 3 (esp. for the many users of Gries's R script coll.analysis<https://www.stgries.info/teaching/groningen/index.html>).
The frequency of the ditransitive would be the sum of the second column and the corpus size (for such applications usually the number of lexical verbs in the corpus) would be provided externally by the user. One notable disadvantage of this format is that it does not reveal anything about the distribution

TABLE 3 Input for collexeme analysis to compute the attractions between all lemmas w and the ditransitive

Word/lemma	Freq in ditr	Freq in corpus
w/l_1	m	n
w/l_2	o	p
w/l_3	q	r
...	s	t

of each lemma within and outside of the ditransitive across the parts/files of the corpus, which has at least two serious consequences, which I will address below. To already set the stage for this later goal of adding the distribution of the verbs in the constructions to the collostructional analysis, we will use a better input format, namely the one exemplified in Table 4, where each verb use, ditransitive or not, gets its own row, and each row also indicates which corpus file/part it is from and whether it is used ditransitively or not.

In this example, Table 4 will be based on the parse trees of the ICE-GB (viz. the annotation for a main verb ("mvb,v(")). The second step will then usually, though not necessarily, consist of adding a column that lists for every verb the corresponding lemma. In this particular case, I retrieved the lemmas by looking up each verb form in the BNC XML and extracting the corresponding lemma from there. The result was a table such as Table 5. If one cross-tabulates

TABLE 4 New input for collexeme analysis to compute the attractions between all forms f and the ditransitive

File	Verb form	Ditransitive
S1a-001	*see*	FALSE
S1a-001	*missing*	FALSE
...	...	...
S1a-001	*gave*	TRUE
...	...	...

TABLE 5 New input for collexeme analysis to compute the attractions between all lemmas *l* and the ditransitive

File	Verb form	Verb lemma	Ditransitive
S1a-001	*see*	*see*	FALSE
S1a-001	*missing*	*miss*	FALSE
...	...	...	...
S1a-001	*gave*	*give*	TRUE
...	...	...	...

TABLE 6 Frequencies of verb lemmas in and outside of the ditransitive

Word/lemma	Freq.: ditransitive.	Freq.: rest of corpus
accord	3	3
ask	92	428
be	0	32258
consist	0	47
do	12	2983
give	566	603
make	3	1932
stagnate	0	1
teach	23	69
tell	489	299

the verb lemmas and the constructional choices, one gets a table like Table 6, of which I only show a few rows.

2.2 *Computing Some Well-Known Measures*

To be able to assess the first suggestion to be made, we first compute a few different measures of collexeme strengths that cover different degrees of statistical complexity and exemplify different kinds of association measures (as per Evert, 2009):

- the negative $\log_{10}$ of $p_{\text{Fisher-Yates}}$; we will compute the version that most users have been working with (one-tailed p-values from R's default version using double precision numbers) but also a computationally more advanced version (two-tailed p-values using arbitrarily precise computations based on the Rmpfr package, see Maechler, 2022; see also footnote 1); we will use a signed version that uses positive and negative values to reflect attraction and repulsion respectively;
- the log-likelihood value G^2, which is actually the null deviance of a binary logistic regression predicting the constructional choice from the presence/absence of a verb and which has been proposed as an easier-to-compute approximation of p_{FYE} (see Dunning, 1993; Evert, 2009); again, we will use a signed version;
- the log odds ratio, which can also be traced back to a binary logistic regression predicting the constructional choice from the presence/absence of

a verb and which is less affected by the absolute frequencies in the input tables (Gries, 2022b);
- pointwise *MI*, a frequent collocational measure also less strongly predictable from the co-occurrence frequency;
- varying intercept adjustments of a glmer as collexeme strengths (following Baayen, 2011), which one would expect to be very close to the log odds ratios).

These results will be compared to the result of the first proposal shortly.

3 Suggestion Set 1: Simplification

As mentioned above, the first suggestion to be discussed in this paper essentially amounts to a huge simplification (and speed-up) of how the analysis can be done. In fact, it goes back to something I was already wondering about I was still a young padawan and had just begun to learn about statistical methods. Back then, I often wondered why people computed their association measures (for, back then, usually collocations) by generating so many separate 2×2 co-occurrence tables – one for each word – when the residuals of a lowly chi-squared test (or the corresponding computations in an HCFA) seemed to provide all one wanted, and the first suggestion is to do just that: In the present case, this would mean instead of computing 3368 G^2-values from 3368 2×2 co-occurrence tables such as Table 1 (one for each verb lemma in the corpus) or instead of computing 88 G^2-values from 88 2×2 co-occurrence tables such as Table 1 (one for each verb lemma used ditransitively at least once in the corpus), we compute the residuals of a single chi-squared statistic on the complete cross-tabulation of all 3368 verb lemma types and the two constructional choices (DITR: *no* vs. *yes*). Since the residuals of a chi-squared test indicate how the observed frequencies relate to the expected ones (Gries, 2021: Section 4.1.2.1), the residuals in the column for the ditransitive construction tell us a lot about how much and in what direction each lemma's observed frequency in the ditransitive differs from its expected one. Table 7 shows results for the same 10 sample verbs as above.

This indicates that
- verbs that never occur in the ditransitive (like *be* or *consist*) are repelled by it (their residuals in the middle column are negative);
- verbs that occur in the ditransitive only a few times but that are otherwise very frequent (like *do* or *make*) are also repelled by it;
- verbs that occur much more in the ditransitive than their overall frequency would make one expect (like *give*, *tell*, or *ask*) are quite strongly attracted to the ditransitive; their residuals in the middle column are positive.

TABLE 7 Chi-squared residuals of verb lemmas in and outside of the ditransitive

Word/lemma	Residuals: ditransitive.	Residuals: rest of corpus
accord	10.347	-1.2
ask	32.387	-3.757
be	-20.695	2.401
consist	-0.79	0.092
do	-4.403	0.511
give	139.73	-16.208
make	-4.477	0.519
stagnate	-0.115	0.013
teach	19.706	-2.286
tell	147.948	-17.162

And creating this is super simple. Once the right type of input (e.g., the right two columns of Table 5) is available in a data structure – let's call it x with the column names LEMMA and DITR – all one needs to do is this single line of R code, that's it:

```
sort(chisq.test(table(x$LEMMA, x$DITR), correct=FALSE)$residuals[,"TRUE"])
```

3.1 *Comparison of All Measures*

How does this new measure compare to the more established association measures? Figure 1 shows a variety of pairwise scatterplots, summarize the correlations with locally-weighted smoothers, and quantify them with Spearman's ρ.

What are the findings?

- there is one extremely highly correlated cluster ($\rho \geq 0.99$) : {*FYE*, G^2, and the chi-squared residuals};
- there is another extremely highly correlated cluster ($\rho \geq 0.99$) : {*LOR*, *MI*, and the varying glmer intercept adjustments};
- even the three measures from the second cluster are all ≥0.925 correlated with the chi-squared residuals.

In essence, the proposed approach is by far simplest and fastest to implement: It requires very little actual code/ data wrangling and no statistical knowledge above and beyond a one-line chi-squared test, which might be the one test nearly all corpus linguists know about; it requires no discounting (like,

e.g., the log odds ratio); it requires no additional steps to distinguish attraction from repulsion (like, e.g., p_{FYE} or G^2); it requires no additional corrections for observed frequencies of 0 (like, e.g., *MI*); yet it is correlated so highly with various completely different kinds of measures that it does indeed seem like a very efficient replacement of other, more cumbersome approaches. Given the simplicity of the approach, it would be nice if we could also apply this to (ideally both kinds of) distinctive collexeme analysis as well, which is what we will try next.

3.2 *Application 2: Distinctive Collexeme Analysis*

Let's revisit the transitive phrasal verb data from Gries and Stefanowitsch (2004), which again only need to come in a straightforward format as in this data frame x.tpv:

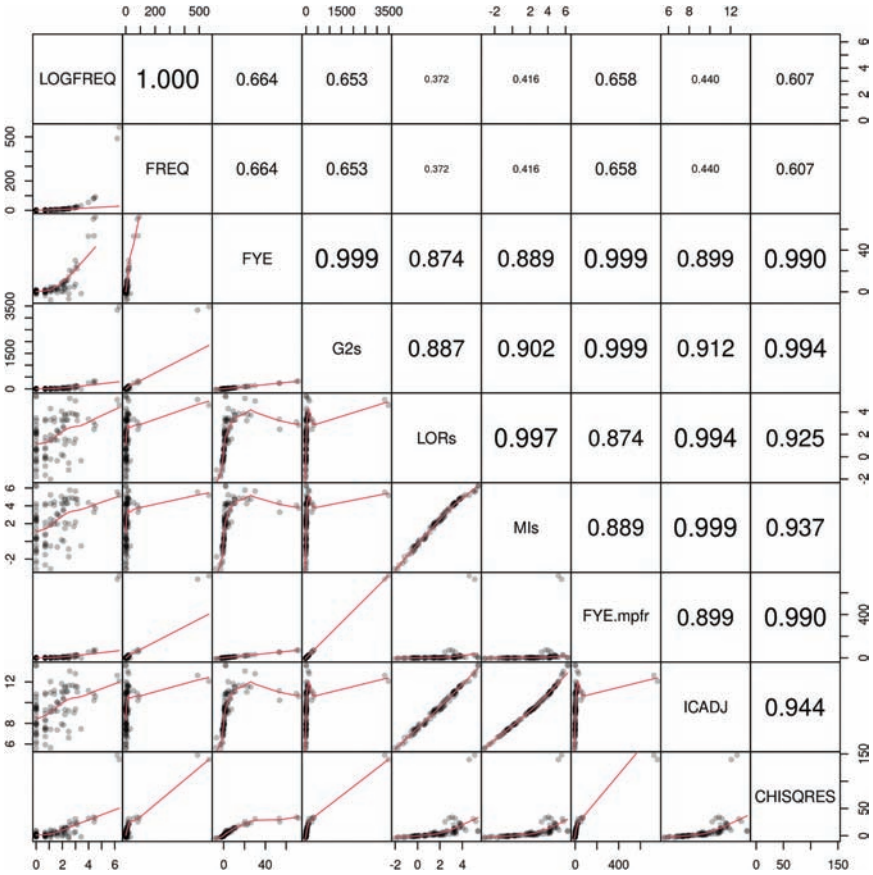

FIGURE 1 Ditransitive collexemes: comparing all measures

```
head(x.tpv, 3)
  FILE    VERBLEMMA PARTICLE CONSTRUCTION
1 S1A-003 take      away     VPO
2 S1A-006 bring     along    VPO
3 S1A-008 give      up       VPO
```

One would begin by merging each verb and its particle into a transitive phrasal verb but then, before this paper, you might have to do some sort of loop (or use apply) to compute an association measure for each of the transitive phrasal verb lemmas, meaning you might have computed 835 Fisher-Yates exact tests or 835 G^2-values etc.; maybe you would have computed 835 log odds ratios (with 835 discounting operations), … Now, you just cross-tabulate transitive phrasal verbs and constructions, …

```
x.tpv$TPV <- paste(x.tpv$VERBLEMMA, x.tpv$PARTICLE, sep="_")
TPV.by.CONSTR <- table(x.tpv$TPV, x.tpv$CONSTRUCTION)
            VOP  VPO
  point_out   3   43
  carry_out   1   49
  set_up      8   42
  find_out    5   49
  put_in     33   21
  pick_up    41   44
```

… compute the chi-squared residuals, and again that's it:

```
CHISQRES <- chisq.test(TPV.by.CONSTR, correct=FALSE)$residuals
```

We can then sort the table (which you could of course do in a spreadsheet software) and look at the top 15 transitive phrasal verbs for each construction:

```
TPV.by.CONSTR <- data.frame(
    VPO=TPV.by.CONSTR[,1], VPO=TPV.by.CONSTR[,2],
    CHISQRES.VOP=CHISQRES["VOP"],
    CHISQRES.VPO=CHISQRES["VPO"])
temp <- CHISQRES[order(CHISQRES["VPO"]),]
```

```
head(temp[,"VPO"], 15) # descending order
get_back     get_out      play_back    turn_off     ring_up
3.110239     2.918756     2.246138     2.216558     2.210167
get_on       get_together get_in       let_down     get_down
1.939575     1.939575     1.881734     1.795697     1.639240
have_back    have_on      play_forward play_out     trace_back
1.639240     1.639240     1.639240     1.639240     1.639240

tail(temp[,"VPO"], 15) # ascending order
play_down    wipe_out     make_out     bring_out    bring_about
1.670173     1.670173     1.803993     1.804269     1.840081
set_out      work_out     give_up      take_up      build_up
2.156184     2.199342     2.209769     2.626803     2.651500
take_on      set_up       point_out    find_out     carry_out
3.048146     3.240352     4.006376     4.059682     4.623747
```

The results replicate Gries and Stefanowitsch's (2004) earlier results nearly perfectly with a Spearman rank correlation between their p_{FYE}-based values and the chi-squared residuals of 0.941 (and 99.6% explained deviance from a generalized additive model).

3.3 *Application 3: Multiple Distinctive Collexeme Analyses*

Finally, let's see whether this also works with a multiple distinctive collexeme analysis, a method that so far has always required a quite different approach, namely as many exact binomial tests as there were combinations of words and constructions, which, with big data sets, could also be very time-consuming and run into small-numbers kinds of problems discussed in note 1. We revisit a part of the 'future alternation' data from Gries and Stefanowitsch (2004), which again only need to come in a straightforward format as shown here in x.fut:

```
head(x.fut, 5)
  FILE     VERB  FUT
1 S1A-001  be    goingto
2 S1A-001  be    goingto
3 S1A-001  go    goingto
4 S1A-001  be    will
5 S1A-001  be    will
```

Again, before this paper, you'd have to do some sort of loop(s) to compute 556 (verbs) times 3 ('futures') exact binomial tests (maybe with the notable additional complication of arbitrarily precise computations) – now, you again just cross-tabulate the verb lemmas and the 'futures', …

```
VERB.by.FUT <- table(x.fut$VERB, x.fut$FUT)
tail(VERB.by.FUT[order(rowSums(VERB.by.FUT)),], 3)
```

	goingto	shall	will
do	63	2	36
have	60	9	70
be	200	16	460

… compute the chi-squared residuals, and are essentially finished:

```
CHISQRES <- chisq.test(VERB.by.FUT, correct=FALSE)$residuals
```

For easily interpretable output, I sort the table (which you could of course do in a spreadsheet software) and look at the top 14 verbs for each 'future' construction:

```
head(CHISQRES[order(-CHISQRES[,"goingto"]),"goingto"], 14)
```

say	do	go	ask	stay	have	win
6.190304	5.837703	3.859386	3.046159	2.788076	2.738282	2.657354
talk	put	happen	get	buy	manage	use
2.569309	2.551254	2.486267	2.432006	2.354397	2.188665	2.177479

```
head(CHISQRES[order(-CHISQRES[,"shall"]),"shall" }, 14)
```

preclude	refer	see	determine	accrue	approve	comply
6.392609	6.392609	6.256779	5.097524	4.520257	4.520257	4.520257
devote	exceed	regret	rest	reveal	summarize	terminate
4.520257	4.520257	4.520257	4.520257	4.520257	4.520257	4.520257

```
head(CHISQRES[order(-CHISQRES[,"will"]),"will" }, 14)
```

depend	know	become	include	provide	receive	find	remain
1.671198	1.495712	1.375743	1.364527	1.364527	1.294504	1.288549	1.212230
want	continue	agree	end	follow	meet	reach	
1.212230	1.166429	1.141646	1.141646	1.141646	1.141646	1.141646	

These results cannot be straightforwardly compared to Gries and Stefanowitsch (2004), because (i) they did not do a multiple distinctive collexeme analysis

(but a distinctive collexeme analysis comparing just *going to* and *will*) and because (ii) I did not include negated forms (like *shan't* but especially *won't*) in this quick replication. However, their results for the most comparable 'future' construction of *going to* are still already quite similar to the present ones because their top 6 were *say*, *do*, *happen*, *have*, *go*, and *win*, all of which are in the top 10 for *going to* above as well.

3.4 *Interim Discussion*

The results are quite promising. While it is true that the availability of Gries's and Flach's R tools allowed for fairly easy computations already for many users, this is taking the ease of things to a new level. As mentioned above, the residuals-of-chi-squared approach to collexeme analysis provides results that are numerically extremely highly correlated with most traditional measures and conceptually extremely similar to what has been the default in most analyses: p_{FYE} or the log-likelihood measure G^2. Similarly, the results of the residuals-based approach to both kinds of distinctive collexeme analysis are also very similar to those of the traditional method yet considerably simpler to obtain (esp. for the multiple distinctive collexeme analysis) by requiring literally thousands of fewer operations. This makes the residuals-based approach a very attractive alternative:

- it requires extremely little prior knowledge of coding or statistics: once the input data are available in the right format of something like Table 5, literally 1–2 lines of code return the desired (bidirectional) association measure with already the right sign for attraction/repulsion;
- it is blazing fast to compute, providing instantaneous results;
- it does not ever run into 0/Inf problems like the traditional approach,[1] which also means it does not require the harder-to-program and extremely time-consuming mpfr approach (which could take hours on big data sets)

1 The 0/Inf problem is one that is common with data sets based on bigger corpora and that is misunderstood even by senior scholars. The problem manifests itself most often when one tries to compute a *p*-value of a Fisher-Yates exact test (e.g., with dhyper) and the individual values from the hypergeometric distribution become so small ($\approx 4.9\times10^{-324}$) that R's normal computation just rounds them down to 0, which means their $\log_{10}$ becomes -Inf. If this happens with multiple lemmas in one's analysis, one cannot rank-order these anymore since they all share the same output value (of -Inf). This has been misunderstood most prominently by Schmid and Küchenhoff (2013:537) as a weakness of collostructional methods because they think it means one requires "a more powerful computer". This is categorically false: The problem can instead be dealt with on any standard computer by incorporating a library in one's code that can work with arbitrary precision floating point numbers. Gries's coll.analysis script has incorporated such a solution for the last few years (using the Rmpfr package, see Maechler, 2022).

to resolve them nor does it require other kinds of statistical adjustments or preparations;
– unlike the traditional approach(es), the same approach can be applied seamlessly to collexeme, distinctive, and multiple distinctive collexeme.

The ease of application and the speed in particular open up additional avenues for following up on the initial results in ways that virtually no collostructional studies have offered. The first kind of follow-up has been theoretically possible for any kind of collostructional analysis, but would have been practically difficult to implement especially with large amounts of data, which have always been likely to lead to -Inf results and, thus, might have required the much much slower approach discussed in note 1. This first follow-up/extension is concerned with quantifying the uncertainty that comes with the collexeme strengths. Obviously any corpus is only a sample of the whole 'population of language' that it represents (hopefully well), which means that it would be nice to be able to quantify how variable the collexeme strengths based on a specific corpus are with some kind of confidence interval, which is the topic of the next two sections.

3.5 *Excursus: The Variability of Collostructional Results*

Computing confidence intervals can be in two main ways – parametrically or with bootstrapping – and in the case of the latter, it can be done in two ways again: based on (i) random sampling with replacement on the level of the individual data points or based on (ii) random sampling on a higher level of organization in the data. For reasons to be discussed below in Section 3.6, we will proceed with the latter way, bootstrapping on the basis/level of corpus parts/files.

3.5.1 The Uncertainty of Collexeme Strengths

To compute proper confidence intervals for collexeme strengths, one requires the new input format illustrated in Table 5 (because we need the information of where in the corpus files/parts ditransitives do and do not occur). In addition and just in terms of implementation, one would need also need a collector data structure that allows us to collect an association measure for each verb lemma for each bootstrapping iteration (say, 500), i.e. a matrix with as many rows as there are verb lemmas in the corpus and as many columns as there iterations in the bootstrap. Then, for each of the 500 iterations, we
– draw a random sample of all corpus files (with replacement!);
– compile them into a temporary bootstrapped corpus;
– do the chi-squared residuals computation on this sampled corpus and save the results in the column for the current iteration.

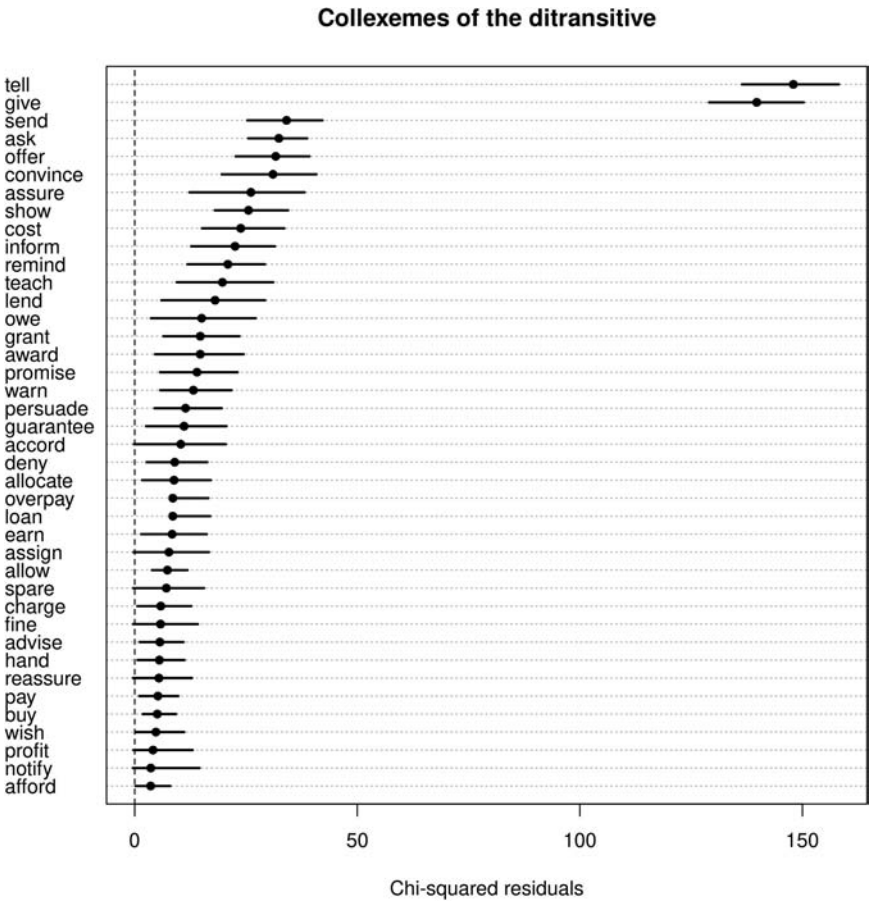

FIGURE 2 Ditransitive collexemes: association w/ confidence intervals

Based on the matrix with all collected bootstrapped residuals for all verbs, we can now derive confidence intervals simply by computing, for each verb, the 2.5% and the 97.5% quantiles, which we can then represent graphically, e.g., in a dotchart (here in Figure 2 for the top 40). The two maybe most obvious findings from this are the following. First, while every analysis of the English ditransitive would predict *give* being one of the top, if not the top, of the collexemes and base its analysis on the semantic similarity of *give* with the usually postulated prototypical 'transfer' meaning of the construction, here *tell* is actually first, but it also does not differ from *give* significantly. Second, even among the top 40 verbs and among verbs with highly fitting semantics, we already find some whose confidence interval suggests that they are not significantly different from chance in their attraction to the ditransitive (e.g., *accord*

or *assign*); of all 88 verbs attested in the ditransitive, 34 are not significant in the present sense.

While I do not mean to exaggerate the importance of significance decisions or cut-off points for collostructional studies, information assessing the variability of one's results is always useful (see more on this below). Also, this kind of computation can of course be done with any kind of association measure but, again, this is by far the fastest and simplest approach while still correlating extremely highly with everything that's much more complicated and time-consuming.

3.5.2 The Uncertainty of Distinctive Collexeme Strengths

The same logic can be applied to distinctive collexeme analyses. One would again require the data discussed above and would again create a collector matrix that allows us to collect an association measure for each transitive phrasal verb lemma for each iteration, i.e. a matrix with as many rows as there are transitive phrasal verb types in the corpus and as many columns as there iterations (say, 500) in the bootstrap. Then, for each of the 500 iterations, we

- draw a random sample of all corpus files (with replacement!);
- compile them into a temporary bootstrapped corpus of the two particle placement constructions;
- do the chi-squared residuals computation on this sample and save the results in the column for the current iteration.

As before, we can then use the matrix with all collected bootstrapped residuals to compute confidence intervals simply by computing, for each transitive phrasal verb, the 2.5% and the 97.5% quantiles, and then we plot the top 20 transitive phrasal verbs preferring *VPO* (see Figure 3) and the top 20 transitive phrasal verbs preferring *VOP* (see Figure 4).

To get a general impression of how all verbs together exhibit significant preferences, we can use a visualization like Figure 5.

Given how a distinctive collexeme analysis is much more discriminatory than a simple collexeme analysis (essentially, because it holds the constructions' function constant), it is not surprising that the percentage of transitive phrasal verbs significantly attracted to a construction (those whose confidence interval does not include 0, the blue data) is higher, but we can again see a variety of verbs whose distribution is such that their constructional association is not reliable (the grey data).

3.6 *Excursus on 'Significant Association'*

The bootstrapping approach towards collostructional uncertainty here has another interesting implication. In the discussion so far, I have treated the

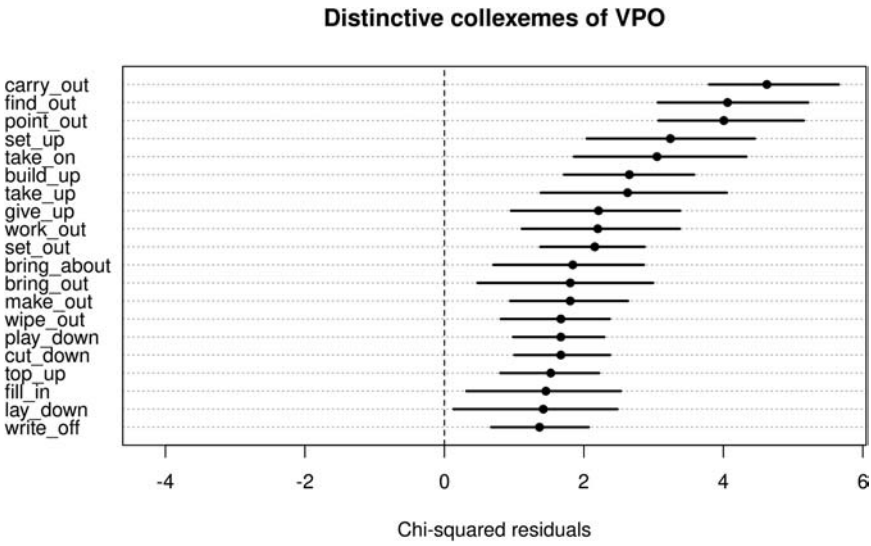

FIGURE 3 Transitive phrasal verbs: preferring VPO

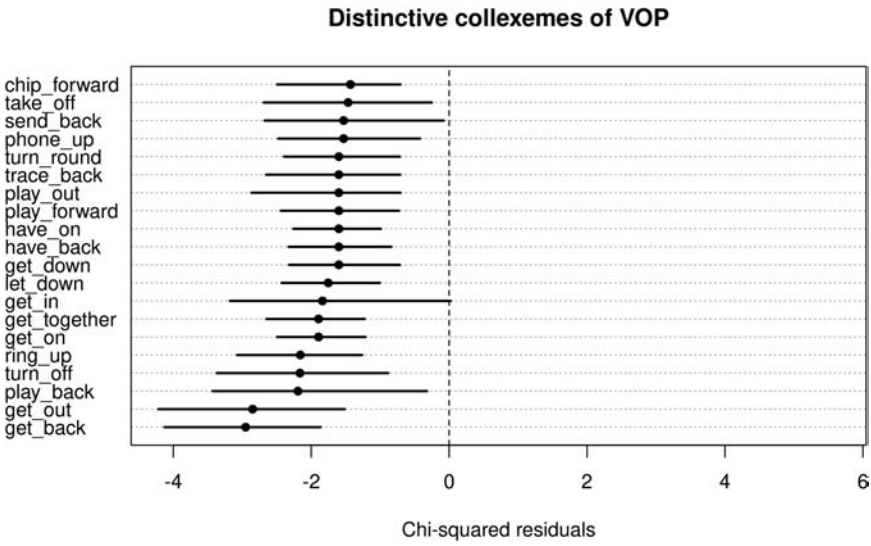

FIGURE 4 Transitive phrasal verbs: preferring VPO

results as equivalent to confidence intervals and most readers will know that, if a confidence interval does not include the value of 0, then we usually consider the result significant; that's why in the two case studies above, we talked about verbs as 'significant collexemes'. But a reader might of course now say that that

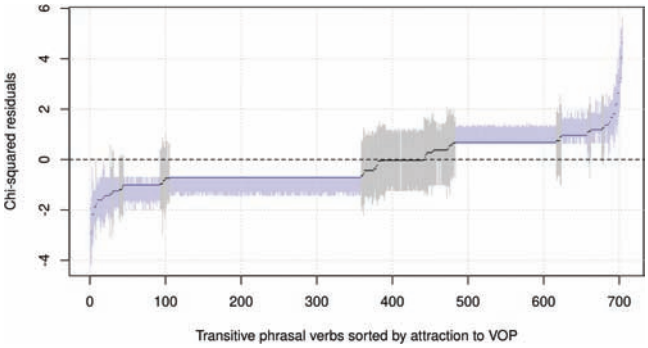

FIGURE 5 Transitive phrasal verbs and their preferences[2]

aspect of the new measure is not really required, because if we use an association measure that is based on a significance test (like p_{FYE} or G^2) we already have a significance test result there. But there is a problem in that reasoning, and not just the problem that virtually no association measure-based study corrects for multiple post-hoc tests, something which Gries (2005b) recommended to reduce the number of potentially falsely positive significant results. The problem is that all significance test-based measures – p_{FYE}, G^2, t, z, and others – compute their p-value involving a null hypothesis based on a model of complete independence of data points (a binomial distribution), which we know is not appropriate. That means that that traditional approach does not consider the division of the corpus into, here, 500 parts/files and would not consider the fact that the probability for any word to show up more than once (in a construction) in one text is higher than the probability of the word to show up more than once in a construction in a corpus as a whole (see the fitting sub-title of Church's (2000) famous paper: "The chance of two *Noriegas* is closer to $p/2$ than p^2."). Thus, even if the above association measures are significance tests, their p-value results cannot actually be interpreted as such because they are based on an incorrect null hypothesis, and the same would be true of any parametric confidence intervals computed for them.

This also has consequences for the kind of bootstrapping approach one might consider. For the same reason that we cannot use the usual p-values/

2 A reviewer asked about the verbs whose confidence intervals stretch out in only one direction; this happens most notably for transitive phrasal verbs with a positive residual score of ≈0.68 and those with a negative residual score of ≈-0.72. This result arises from the fact that the verbs are hapaxes: the sampling can only make their residuals deviate from 0 more (because the file in which they occur gets sampled 2+ times), but it cannot make their residuals get closer to 0.

confidence intervals of significance-based association measures for a real significance assessment – non-independence of data points due to the structure of the corpus into parts – we should not use a bootstrapping approach that randomly samples from all data points (e.g., verb uses). As Gries (2022a) has shown, such bootstrapping applied to corpus data can considerably exaggerate the significance of the differences between different results: While it doesn't so much affect the point estimate (i.e. the estimate used for the association measure), it affects the width of the confidence interval and, thus, whether it includes 0 or not. The proper way to proceed is what we did here, bootstrapping by sampling with replacement on the levels of the files (or, ideally, speakers) and, with that, we can more properly interpret the result (as a significance test or a confidence interval). Thus, while this logic can of course be applied to any association measure – not just the chi-squared residuals here – it lends supports to the current CI computation, which in turn is considerably facilitated and accelerated with a simple measure like the one proposed here.

4 Suggestion Set 2: Using Multiple Dimensions Separately

4.1 *Introduction and a Thought Experiment*

To reiterate an important qualification to the above suggestion (of using chi-squared residuals as a measure of collexeme strength), this suggestion is only (!) for those applications of collostructional studies that are descriptive/exploratory – it should probably not be used for studies with more explanatory or theoretical goals. This is because the amount of (statistical) information entering into the kind of collostructional study that is still vastly predominant approach is severely limited, as we can show easily with a thought experiment. Imagine for a moment you knew nothing about corpus-based association measures and in particular their application to the association to words and constructions; imagine also you, maybe as a cognitive or usage-based linguist, were tasked with developing the ideal association measure approach to study word-construction associations. What dimensions of information might be relevant to cognitive and psycholinguistic aspects of word-construction associations (especially if you follow Bybee (2010) and many others in assuming that much of linguistic knowledge and processing is grounded in domain-general abilities)? First and most uncontroversially, there is *frequency*, which would be relevant to word-construction associations because it is in fact one of the most central notions in cognitive or usage-based linguistics and psycholinguistics, given its robust correlation with reaction times and mental entrenchment. As Schmid (2010:115) says, correctly, I believe, "frequency is one major

determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse"; see also Ellis (2002a, b).

The second dimension of information you would probably want to include is *contingency/association*, which would be relevant because recognizing contingency, or correlation, is the cornerstone of associative learning. Ellis (2006:10) argues, "it [is] contingency, not temporal pairing, that generated conditioned responding in classical conditioning" and "human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like r, χ^2 and ΔP [...] and that probability theory and statistics provided a firm basis for psychological models that integrate and account for human performance in a wide range of inferential tasks" (Ellis, 2006:7). In addition, note that association can be measured bidirectionally, as it has in the vast majority of all collocation and collostruction studies, but it can also be measured unidirectionally – from the word to the construction or from the construction to the word – and this might be useful to distinguish given that many – or even most? – aspects of language learning are probably directional, from one unit to the other.

Then, as a cognitive/usage-based linguist, you wouldn't want to ignore a third dimension, namely *dispersion*, which would be relevant because of how it moderates, complements, or even replaces effects of frequency (Adelman et al., 2006; Brysbaert and New, 2009; Baayen, 2010; Gries, 2020, 2022), but also because of its correlation with everything having to do with learning; Ambridge et al. (2006:175) put it best: "Given a certain number of exposures to a stimulus, or a certain amount of training, 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."

More dimensions of information might be relevant, e.g. *salience* and *surprisal*. Salience would be relevant because it would be one way to explain fast mapping, i.e. the learning of a salient association even if it was encountered only very few or even just one time (see Carey and Bartlett, 1978); surprisal would be relevant because it can be related to salience and because we know it can moderate other processes that have been linked to learning (e.g., priming, see Jaeger and Snider, 2008). But frequency, association, and dispersion are probably among the most essential to a truly cognitive-linguistic, or usage-based, approach to lexical or lexicogrammatical co-occurrence. Determining the dimensions of information we would ideally always consider is only the first step, though, because we also need to determine how we measure each of them.

Frequency is pretty straightforward: we count things and usually log the counts, no problem there. Association is already a much trickier beast because, while dozens of association measures are available, many of them are really not that good at capturing association and only association – most of the most widely-used association measures including p_{FYE}, G^2, t, and z not only reflect frequency *and* association, but they also reflect frequency *more* than association; see Gries (2022b), and that is also true of the chi-squared residuals. That is precisely why such measures are good *for descriptive/exploratory studies* (*!*): They provide a heuristically useful amalgam of two kinds of information. But for explanatory studies, such measures are not that good: One needs to be able to assess the importance of contingency/association separately, which requires an association measure that is largely orthogonal to frequency. Gries (2022b) shows how to develop such a measure, but also demonstrates that, if one wants a bidirectional measure, the log odds ratio behaves as desired, which is what we will use here. Finally, adding dispersion is similarly tricky because nearly all dispersion measures in corpus linguistics – *range*, Juilland's *D*, and my own *DP* – are also >0.9 correlated with frequency (Gries, 2022c). Thus, the same logic applies: for explanatory studies, one needs to be able to assess the importance of dispersion separately and must therefore use a dispersion measure that is largely orthogonal to frequency and Gries (2022c) shows how *DP* can be adjusted to that end, which we will do here, too.

In sum, descriptive/exploratory collostructional studies can use an association measure that blends frequency and association, as most work has done, but theoretical/explanatory work must do better and should (i) study lexicogrammatical association not with just one score, but with a tuple of scores for, minimally, frequency, association, and dispersion (as argued already in Gries, 2019) and should (ii) measure each dimension in such a way that it is not already mostly pre-determined by frequency, which is how this study goes beyond Gries (2019), which used an unadjusted version of *DP* but probably shouldn't have. Let us now explore what this would look like for the ditransitive collexeme data.

4.2 *Ditransitives in the ICE-GB*

For this construction, we already have the (logged) co-occurrence frequencies of verbs in the ditransitive and the log odds ratios but we still need a dispersion measure that is not by definition already very highly correlated with the frequencies and we will use DP_{nofreq}. Quantifying the dispersion of verbs in the ditransitive with it is done in the following way:

1. one computes the regular *DP* measure for verbs in the ditransitive. Regular *DP* is computed as follows (see Gries, 2020, 2022c for more discussion and exemplification based on 6 corpora):
   a. one computes each corpus part's/file's size as a fraction of the overall corpus size; call this vector file.sizes.rel. For instance, the first file of the ICE-GB makes up 0.002296902 of the corpus as a whole (measured in main verbs);
   b. one computes for each verb how much of its uses in the ditransitive is in each corpus part; call this vector ditr.verbs.by.files.rel. For instance, 0.005300353 of all uses of GIVE in the ditransitive are in the first file of the ICE-GB;
   c. for each verb, take half of the sum of the absolute pairwise differences of these two vector: sum(abs(file.sizes.rel – ditr.verbs.by.files.rel))/2, that is *DP* for each verb;
2. one computes the largest possible *DP* value, i.e. the *DP*-value that would result if all *n* ditransitive uses of the verb occurred in the smallest corpus part/file; call this DP.upp; this value is the same for all verbs because it just assumes that 100% of all instances of any verb in a construction are in that smallest file;
3. for each verb, one computes the smallest possible *DP*-value, i.e. the *DP*-value that would result if the *n* ditransitive uses of the verb occurred as evenly distributed across the corpus parts/files as possible; call this DP.low;
4. for each verb, compute a min-max transformation of the *DP*, DP.upp, and DP.low, which means you set DP.low to 0, DP.upp to 1, and determine where on that continuum from 0 to 1 the actual *DP*-values falls – that is the frequency corrected value of DP_{nofreq} for that verb.

The logic may seem arbitrary/convoluted, but (i) it really only means that each verb's *DP*-value is interpreted against that background of what the possible range of *DP*-values of a verb with that frequency would be (and that is what 'partials frequency out of dispersion') and (ii) this is done with a min-max transformation that is actually common in other contexts. For instance, anyone who has ever read a paper reporting a Nagelkerke's R^2 (the most widely used R^2 reported for logistic regressions) has already encountered this logic: Nagelkerke's R^2 is essentially a min-max-transformed version of Cox and Snell's R^2 (and similar considerations come into play for other statistics (e.g. φ for frequency tables) Also, implementing this is easier than it sounds and requires maybe a dozen lines of code, no more. However, representing such results visually can be a bit more challenging for print publications because we

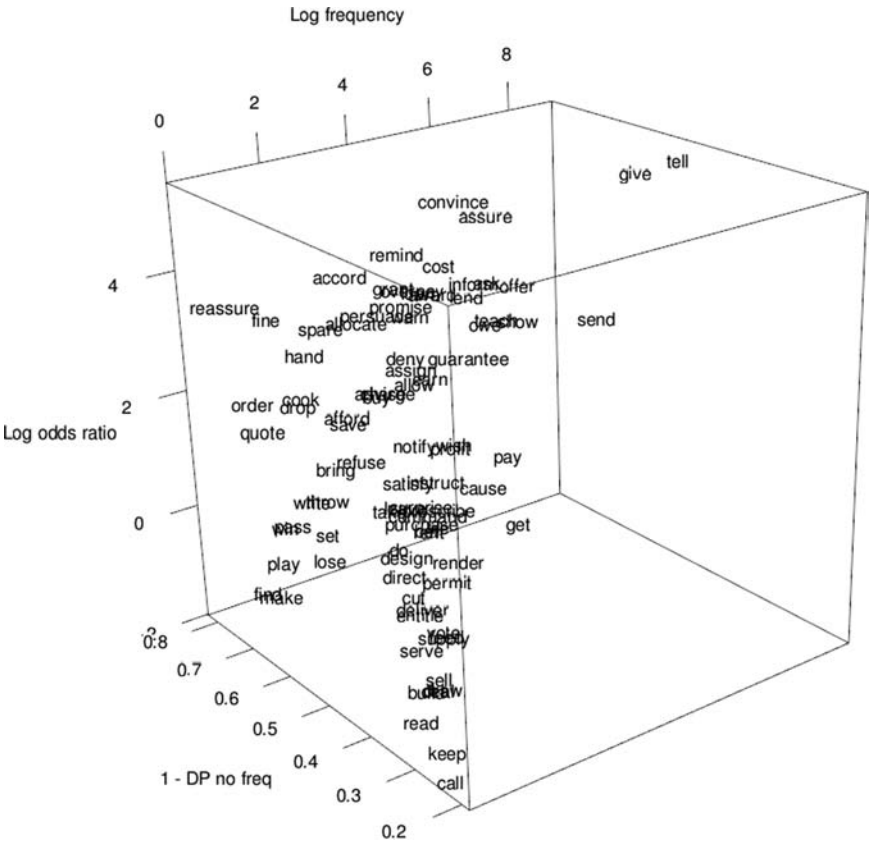

FIGURE 6 Ditransitive collexeme tuples

now have three numeric dimensions to plot – (logged) frequency, association (log odds ratios), and dispersion (1- DP_{nofreq}).[3] A 3-D version from one hopefully useful angle is shown here in Figure 6.

One of the maybe more easily interpretable ways of plotting these results is Figure 7, where

– the *x*-axis represents the association between verbs and ditransitives (measured in a way that is not already by definition determined by co-occurrence frequency);
– the *y*-axis represents the dispersion of verbs in ditransitives (measured in a way that is not already by definition determined by co-occurrence frequency);

3 We use 1-DP_{nofreq} so that low and high values mean clumpy and even distribution respectively.

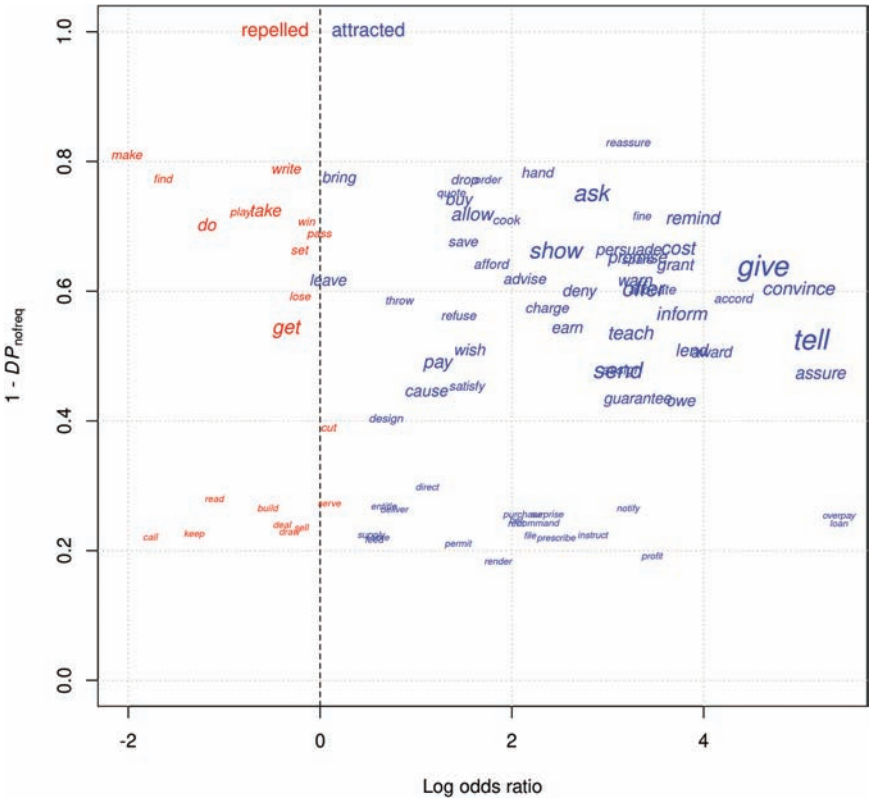

FIGURE 7 Ditransitive collexemes: frequency, association, dispersion

– the sizes of the verbs represent their logged frequencies in the ditransitives;
– the colors of the verbs indicate whether a verb occurs more often in the ditransitive than expected by chance.

Both plots make clear that *give* and *tell* are still the most 'prototypical' verbs for the ditransitive, but also why: Much of this result – not all, but much – is due to their high frequency in the ditransitive, because the two are by far the most frequent verbs in the construction while they are neither the most strongly attracted ditransitive verbs nor the most evenly dispersed ones. For instance, in terms of pure association, other communication verbs such as *convince* or *assure* score higher, but they are much much rarer. On the other hand, we can also see that verbs like *ask* or *hand*, while not as strongly attracted to the ditransitive as others, are more evenly distributed in the ditransitive than most others. Any analysis of a construction that was interested in the construction's learnability, overall distribution, or how its verb slot changes over time would benefit from the higher degree of precision that this approach offers over the more descriptive one that is still prevalent. To drive home this point, let

us briefly look at another construction, the *way*-construction (in the British National Corpus).

4.3 *The Way-Construction*

Let's apply the exact same logic to the *way*-construction (e.g., *he fought his way through the crowd* or *she made her way to the top*). For the purposes of this case study, I extracted from the British National Corpus (XML edition) all sequences of a lexical verb (a verb that is not tagged as *be*, *do*, *have*, or a modal), a possessive determiner (CLAWS tag: DPS), and the word *way*. This returned 7013 potential examples, the vast majority of which were 'proper' *way*-constructions – a small number of exceptions include, for instance, *get my way*, but the overlap of the by far most verb lemmas in the slot with previous studies is very high. The 7013 examples involved 596 different verb lemmas, for each of which I determined its frequency in the *way*-construction, its log odds ratio to the *way*-construction, and its dispersion (DP_{nofreq}) in the *way*-construction in the exact same way as before. Correspondingly, we have the same visualization options as before: a 3D-plot as in Figure 8 or the maybe more easily interpretable 2-D plot of Figure 9.

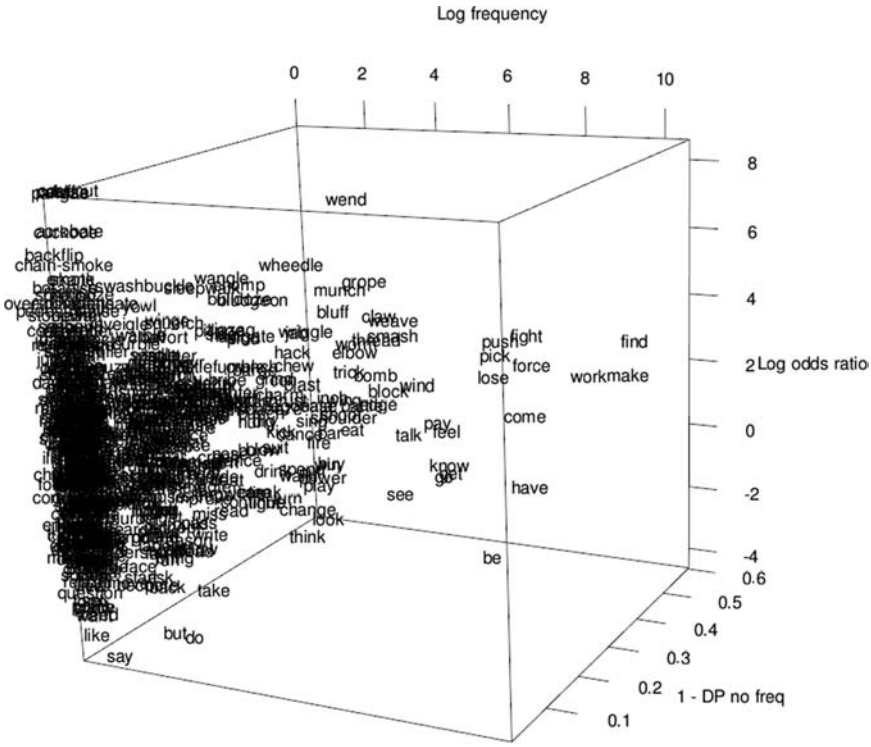

FIGURE 8 The *way*-construction's collexeme tuples

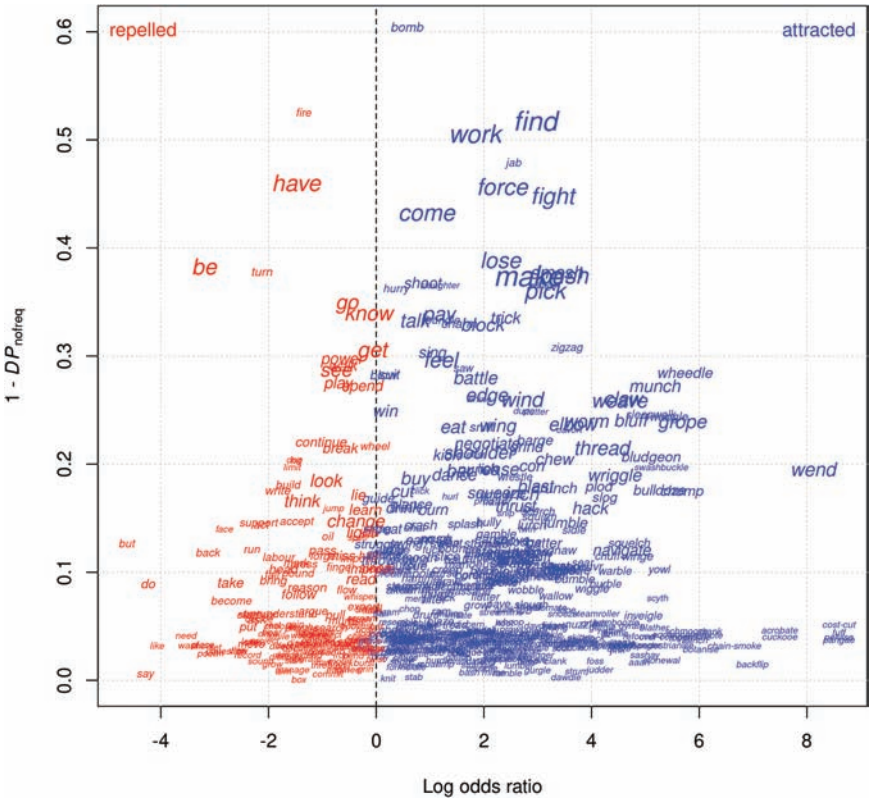

FIGURE 9 The *way*-construction's collexemes: frequency, association, dispersion

We find there is a cluster of verbs that score highly on all three dimensions; these include *find*, *make*, *work*, *fight*, and *force*, followed closely by *push* and *pick*. These are arguably the set of prototypical verbs of the *way*-construction: They occur frequently in the construction, they occur in that construction in a wide variety of situations/settings, and they are fairly strongly attracted to the construction. However, there are many verbs that have a stronger association, but many of these verbs are fairly rare in the construction and also not widely distributed in the corpus (meaning, their results might be due to few corpus files); as per Gries (2011), many of these include verbs beginning with a/w/ sound such as *wend*, *wheedle*, *wriggle*, *warble*, *wind*, *wiggle*, but also, more forceful verbs such as *bludgeon*, *hack*, *blast*, and a class of verbs including *munch*, *chomp*, *chew*, *gnaw*. Regardless of one's interpretive/theoretical goals – language acquisition? language change? – it is clear that this approach provides a much more fine-grained resolution on how exactly each verb is associated with the *way*-construction than if we just slapped one number on each

verb, and in fact a number that somehow blends a lot of frequency, a bit of association data, and another bit of dispersion.

4.4 *Adding Senses: Verb-Particle Constructions*

The above discussion approached the addition of more information and the corresponding expansion from a single simple collexeme strength value to a tuple (of, minimally, frequency, association, and dispersion) from a statistical perspective. There is, however, another way in which more information should be added and it's maybe an even more important dimension of information. This is because, in a sense (pun intended), much collostructional work (including my own) has not really stayed true enough to the theoretical foundation of the method. *Collostructional* as a term was chosen as a blend of *collocation* and *construction*, the latter being used in the above-mentioned Construction Grammar sense of the term. However, that Construction Grammar perspective then of course implies that what is of interest is the co-occurrence (patterns) of constructions, i.e. pairings of form and function/meaning. But note how often we all have strayed from that perspective. Yes, usually one of the (sets of) items featured on a collostructional study was a syntactic/linking/argument structure construction – the ditransitive, the *way*-construction, the two verb-particle constructions, etc. – but what about the elements occurring in (slots of) them? Construction Grammarians always say something to the effect 'it's construction all the way down', meaning the words occurring in the verb (and particle) slots of the above constructions are indeed also constructions, just ones that are less schematic than more syntactic/linking/argument structure constructions. But that means they, too, are pairings of form and function/meaning even if very many studies (again, including my own) has nearly exclusively focused on their form side and disregarded the function side. With some hopefully illustrative exaggeration, most collostructional studies have not measured and discussed co-occurrences of one construction (e.g. one of the two verb-particle constructions) with many other constructions (all the phrasal verbs used with it) – they measured and discussed co-occurrences of one construction (e.g. one of the two verb-particle constructions) with *letter sequences*, ignoring the different functions/meanings those could come with. Put differently, most collostructional studies essentially said 'we pretend the thing we spell as *put down* is one construction but ignore its (potentially many diverse) functions/meanings' when, I think, any Construction Grammarian would agree that what we should actually be saying is the very explicit 'there are several (often related) constructions that share their form side/ phonological pole, which we might spell as *put down*, but that have different functional sides/ semantic poles'; here are some examples with rough paraphrases:

– *put down* 'kill' as in *I had the dog put down*;
– *put down* 'place' as in *I put down the coffee cup*;
– *put down* 'write/register' as in *Oh, it's a potluck? Put me down for some pudding*;
– *put down* 'denigrate' as in *must you put him down like that? He's only starting to learn this*.

And then there are of course also idiomatic uses with fully specified lexical DOs (e.g., *put* POSS *foot down* 'insist'). And this problem increases as the versatility/generality of the meaning of the more abstract construction increases and, thus, probably its productivity (which is, after all, often measured on the basis of the type frequency or the hapax type frequency of a slot). For instance, such kinds of polysemy are not much of a problem with the N *waiting to happen* construction, which has a much more restricted set of nouns it occurs with (chiefly, *accident* and *disaster*) but they are much more of a problem with much more general constructions such as the verb-particle constructions (let alone tense, aspect, or voice or linking constructions).

Now, admittedly, this problem of 'missing meaning/function poles' has not been completely unnoticed. Gries and Stefanowitsch's (2004) analysis of the dative alternation already makes an implicit reference to verbal polysemy when they explain the at first sight rather surprising strong attraction of *play* to the *to*-dative, which they explained by pointing out that those instances are mostly from sports commentary (e.g., *he played to ball back to Messi*), and their note 5 discusses the polysemy of *have on* (*you have your bra on* vs. *they are having a bit of a sale on*) and explicitly states that "in some cases it might be more precise and rewarding to not just look at the distinctive collexemes of verbs, but of verb senses, i.e. verb-sense specific patterns […]". However, too little has come of that approach during the next 15+ years. Laudable exceptions to this relatively strong neglect of polysemy in the collexemes include Wiechmann (2008), Gilquin (2010), Colleman and Bernolet (2012), and Bernolet and Colleman (2016). In a few nutshells,

– Wiechmann found that collostructional results that distinguished senses of verbs (whether they take a nominal or a sentential complement) outperform collostructional results that did not distinguish senses when it comes to predicting reading time latencies;
– Gilquin did a multiple distinctive collexeme analysis of periphrastic causatives and showed that different senses of verbs have very different preferences for causative constructions;
– Colleman and Bernolet showed how polysemy effects can partially explain the lack of (better) agreement between corpus and experimental data on the dative alternation;

– Bernolet and Colleman illustrated how verb senses can have widely different alternation biases (in a study of the Dutch dative alternation) and how, in an experiment, priming was affected by an interaction of the strength of priming and sense-specific biases of target verbs.

As the last main point of this paper, I would therefore like to re-emphasize the point made by all these must-read studies, if only by the smallest of add-ons to one of the above case studies, the distinctive collexeme analysis of the verb-particle constructions. I took three of the most frequent transitive phrasal verbs – one with a very strong preference for one constructions (*carry out*, which prefers VPO) and two with no strong constructional preferences (*pick up* and *put down*), read all instances, and tried to assign a sense to each use. The senses for *put down* are the ones listed above, the senses for *carry out* were 'execute/perform' (e.g., *Allied Forces carried out air raids*) and one sense one might term 'propel/move' (*All he could do was bottle the burning feeling tightly inside,* [...] *and then let the thin oozings of rage carry him out on the morning trail.*), and the senses for *pick up* included

– 'register/notice/learn' as in *so don't be discouraged if you don't pick things up quite as quickly as everyone else*;
– 'take (typically with one's hand)' as in *He picked a pencil up, rocking it between a thumb and finger*;
– 'meet (typically romantically)' as in *You picked a girl up on the train*;
– 'address' as in *I think you pick up the point about how efficient the council is*, and a few others.

(Again, these labels are heuristic: as so often in semantic annotation, it's sometimes not clear when to lump and when to split.) If the phrasal verbs *carry out*, *pick up*, and *put down* are then replaced by themselves with the relevant sense tag added to them (e.g., *carry out* becomes *carry out* (*execute*) or *carry out* (*propel*)), we can redo all of the above analysis; consider Figure 10.

If one looks at each of the verbs, one finds that, as one might expect given the above discussion, the addition of the senses can make a difference. All 49 uses of *carry out* with the 'execute/perform' sense were in VPO and the one other sense was in VOP; thus, here nothing much changes so the transitive phrasal verb construction most strongly attracted to VPO is now *carry out* (*'execute'*). For *pick up*, the results are more diverse: 'register/notice/learn', 'take', and 'meet (romantically)' prefer VOP while 'pay', 'play to', and 'address' prefer VPO. For *put down*, the changes are maybe most pronounced. Without looking at senses, this verb has no strong preference but, once senses are distinguished,

– 'place', the 'insist' idiom, and 'denigrate' prefer VOP;
– 'write down', one unclassifiable instance (which could have been 'write' or 'denigrate'), and 'kill' prefer VPO.

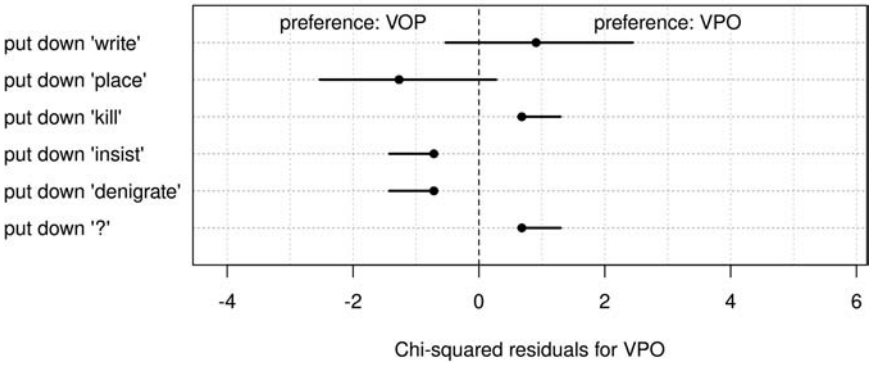

FIGURE 10 Senses of *put down* and their preferences

The differences in Figure 10 are not huge,[4] but I submit that the general message of 'senses can make notable difference' still holds: It is potentially self-contradictory to self-identify as Construction Grammarians, do collo*structions*, but take seriously only the constructionhood – a pairing of form *and* function – of one of the constructions in one's study (e.g. the ditransitive or verb-particle constructions) but reduce the other one to the form side of 'what letters are being used'. If I had it my way (…), all collostructional studies of semantically general or versatile constructions at least should now consider senses!

5 Conclusion

The main message of this paper can be visualized as a flowchart: If one's goal is really only description/exploration, then an association measure combining (a lot of) frequency and (a little bit of) association is still a good option, and in order to facilitate and accelerate analyses, I propose to use the residuals of a chi-squared test, which are extremely highly correlated with the current 'gold standards' of Fisher-Yates exact and G^2, but require literally 5% or, with big data, much much less of the computational efforts – the time and resources spent on the old tests should then rather be invested into a quick and proper bootstrapping so that our collexeme strengths come with uncertainty/confidence intervals.

4 The confidence intervals are not particularly informative here, given the low frequencies of co-occurrence. Also, it is interesting to note just on the side that these preferences are not easily explainable on the basis of a cline from literal via metaphorical to idiomatic.

However, if one's goal is theoretical/explanatory, then we all need to do better, especially if we're coming from a cognitive or psycholinguistic perspective. Instead of pretending that a single blended association measure can cover all the cognitive or psycholinguistic dimensions of information that everyone agrees are relevant, we need to measure them all and we need to measure them in ways that orthogonalize them as much as we can: Our association measure should return mostly association (like the log odds ratio would), our dispersion measures should return mostly dispersion information (like DP_{nofreq} would), all dimensions of information per, in most studies, word/lemma should be represented in a tuple of values, and the unit to which any and all measures are applied should really be constructions, not generalized letter sequences. Then, and only then, will we be able to assess all the dimensions of collostructional attraction or repulsion that are relevant to our cognitive and psycholinguistic goals.

Acknowledgements and Dedication

This article follows up on a lecture I presented online as part of the Innovation in Linguistics Talk Series in Oct. 2022. Innovation in Linguistics is an online zoom/voov meeting talk series organized by the journal Cognitive Semantics. It is a non-profit organization and attended for free. Each talk may last for from a minimum of 60 minutes to a maximum of 120 minutes. The presentation is video-recorded and is posted on the WeChat official account (titled 认知语义学, cognitive semantics in Chinese) in China and the YouTube channel (@cognitivesemantics3004) maintained by the editorial team. The video is available for future audience viewing. The present talk, titled "Overhauling Collostructional Analysis", was delivered on 11 Oct 2022 and can be viewed at https://mp.weixin.qq.com/s/I9zg6Cr-7SiSswGqZ6CyhA and https://www.youtube.com/watch?v=Wu4uTGBJn8I&t=6047s.

This paper is dedicated to Thomas Berg, Professor of English Linguistics in the Department of English of the University of Hamburg, on the occasion of his 65th birthday and as a small "thanks!" for the positive influence he's had on me during the early stages of 'my academic coming of age'.

References

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

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

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 naïve discriminative learning. *Brazilian Journal of Applied Linguistics* 11(2): 295–328.

Backus, Ad and Maria Mos. 2011. Islands of productivity in corpus data and acceptability judgments: contrasting two potentiality constructions in Dutch. In Doris Schönefeld (ed.), *Converging Evidence: Methodological and Theoretical issues for Linguistic Research*, 165–192. Amsterdam and Philadelphia: John Benjamins.

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

Brysbaert, Marc and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4): 977–990.

Bybee, Joan. 2010. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.

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

Church, Kenneth W. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/_2$ than p^2. In Proceedings of the COLING 2000 (The 18th international conference on computational linguistics), np.

Colleman, Timothy and Sarah Bernolet. 2012. Alternation biases in corpora vs. picture description experiments: DO-biased and PD-biased verbs in the Dutch dative alternation. In Dagmar S. Divjak and Stefan Th. Gries (eds.), *Frequency Effects in Language Representation*, 87–125. Berlinand New York: Mouton de Gruyter.

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

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1): 1–24.

Ellis, Nick C. 2002a. Frequency effects in language processing and acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2): 143–188.

Ellis, Nick C. 2002b. Reflections on frequency effects in language acquisition: A response to commentaries. *Studies in Second Language Acquisition* 24(2): 297–339.

Ellis, Nick C. and Fernando Ferreira-Junior. 2009a. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3): 370–385.

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

Ellis, Nick C., Ute Römer, and 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.

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

Flach, Susanne. 2020a. Schemas and the frequency/acceptability mismatch: Corpus distribution predicts sentence judgments. *Cognitive Linguistics* 31(4): 609–645.

Flach, Susanne. 2020b. Reduction Hypothesis revisited: Frequency or association? In Christina Sanchez-Stockhammer, Franziska Günther, and Hans-Jörg Schmid (eds.), *Language in Mind and Brain: Multimedial Proceedings of the Workshop Held at lmu Munich, December 10–11, 2018*, 16–22. München: LMU Open Access.

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Amsterdam and Philadelphia: John Benjamins.

Goldberg, Adele, E. 2006. Constructions at Work: The Nature of Generalization in Language. Oxford: OUP.

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

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

Gries, Stefan Th. 2011. Phonological similarity in multi-word symbolic units. *Cognitive Linguistics* 22(3): 491–510.

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. 2015. More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3): 505–536.

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

Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 99–118. Berlin and New York: Springer Press.

Gries, Stefan Th. 2022a. Towards more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics* 1(1): 100002.

Gries, Stefan Th. 2022b. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(1): 1–33.

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

Gries, Stefan Th., Beate Hampe, and 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, and Doris Schönefeld. 2010. Converging evidence II: more on the association of verbs and constructions. In Sally Rice and John Newman (eds.), *Empirical and Experimental Methods in Cognitive/functional Research*, 59–72. Stanford, CA: CSLI.

Gries, Stefan Th. and Anatol Stefanowitsch. 2004a. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129.

Gries, Stefan Th. and Anatol Stefanowitsch. 2004b. Co-varying collexemes in the into-causative. In Michel Achard and Suzanne Kemmer (eds.), *Language, Culture, and Mind*, 225–236. Stanford, CA: CSLI.

Gries, Stefan Th. and 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. and Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7. 163–186.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2): 243–257.

Hilpert, Martin. 2008. *Germanic Future Constructions: A Usage-based Approach to Language Change*. Amsterdam and Philadelphia: John Benjamins.

Jaeger, T. F. and Snider, N. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In Bradley C. Love, Ken McRae, and Vladimir M. Sloutsky (eds), *Proceedings of the Cognitive Science Society Conference*, 1061–1066. Washington, DC: Cognitive Science Society.

Maechler, Martin. 2022. Rmpfr: R MPFR – Multiple Precision Floating-Point Reliable. R package version 0.8–9, https://CRAN.R-project.org/package=Rmpfr.

Schmid, Hans-Joerg. 2010. Entrenchment, salience, and basic levels. In Dirk Geeraerts and Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, 117–138. Oxford: Oxford University.

Schmid, Hans-Joerg and 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 and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

Stefanowitsch, Anatol and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1): 1–43.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English*. Berlin and New York: De Gruyter Mouton.

Wiechmann, Daniel. 2008. Initial parsing decisions and lexical bias: Corpus evidence from local NP/S-ambiguities. *Cognitive Linguistics* 19(3): 439–455.