DE GRUYTER

Corpus Linguistics and Linguistic Theory

Introduction to the special issue on collostructions

Journal:	*Corpus Linguistics and Linguistic Theory*
Manuscript ID	Draft
Manuscript Type:	Article
Classifications:	
Keywords:	collostructions, collexemes, construction grammar, log-likelihood score, COCA, COHA

1 Introduction

A little more than 20 years ago, Anatol Stefanowitsch and Stefan Th. Gries published a series of four articles – Stefanowitsch & Gries (2003, 2005) and Gries & Stefanowitsch (2004a, b) – that proposed to apply the decades-old tradition of quantifying the co-occurrence of words (collocates) with node words using statistical association measures to the occurrence of words with one or more constructions (in the Construction Grammar sense of *construction*, as in, back then, Goldberg 1995). Over time, the resulting family of methods came to be known as *collostructional analysis* – a blend of *collocation* and *construction* – and included three main methods:

- **collexeme analysis**, which quantifies the degree to which different words are attracted to, or repelled by, a specific slot in one construction. For instance, which verbs are attracted to the main-verb slot of the ditransitive construction?
- **(multiple) distinctive collexeme analysis**, which quantifies the degrees to which different words are attracted to, or repelled by, comparable slots in two (or more) functionally similar constructions. For instance, which verbs are attracted to the main-verb slot of the ditransitive constructions and which verbs are attracted to the main-verb slot of the prepositional dative construction?
- **co-varying collexeme analysis**, which quantifies the degree to which words in two different slots of one construction are attracted to, or repelled by, each other. For instance, which verbs_1 are attracted to which verbs_2 in the into-causative (e.g., *He tricked*$_{\text{verb1}}$ *her into signing*$_{\text{verb2}}$ *the contract*)?

On the one hand, these methods were relatively straightforward extensions of calculations that had been used in collocation research in corpus linguistics for a long time; on the other hand, these methods also happened to become extremely successful – for both Gries and Stefanowitsch, the four articles mentioned above amount to their most cited works (with, according to Google Scholar in March 2025, a combined number of >4000 citations) because they fortuitously rode and, with all due humility, maybe co-created several 'waves' or trends co-occurring at the same time:

- the trend of linguistics in general to become more quantitative;
- the trend of linguistics to become more computational, which back then especially meant that corpora and corpus-linguistic work were becoming more widespread;
- the rise of interest in (esp. Goldbergian) Construction Grammar in cognitive linguistics, which coincided with a concomitant rise of interest in Pattern Grammar in corpus linguistics (Hunston & Francis 1998).

This success was manifested in, according to an informal bibliography compiled by Anatol Stefanowitsch, literally hundreds of collostructional papers since 2003. Because of collostructions' enduring success and because of the fact that Stefanowitsch & Gries also co-founded *Corpus Linguistics and Linguistic Theory* (*CLLT*) at exactly that time – *CLLT*'s first issue appeared in 2005 with Stefanowitsch & Gries (2005) as its lead article – Anatol Stefanowitsch had the idea of commemorating, so to speak, 20 years of collostructions and this special issue of *CLLT* is one way in which this idea was realized. The special issue brings together contributions from a variety of researchers, some of whom used collostructions early on, some of whom only became interested

in it later, but all of whom help paint a picture of the current state of collostructional methods involving both the 'traditional' approach and newer developments that aim to broaden the scope and make the method more useful as the theoretical and methodological landscapes are changing.

2 The papers in this current issue

The papers in this issue adopt approaches towards their objects of study that usually differ along a variety of dimensions, e.g. how many constructions and slots in constructions are looked at, whether they are examined at the same time (i.e. in a multivariate way) or sequentially monofactorially, what corpus data were used, what statistics are used, what follow-up analyses are pursued, etc. The following presentations are therefore selective on what they highlight to suggest 'groups of papers', and other arrangements would be equally possible.

The paper by Chen targets Degree Adverb Constructions (an English example would be *very good*) in the Academia Sinica Balanced Corpus of Mandarin Chinese, a corpus of more than 10 million words covering a wide variety of topics, genres, and styles. Using POS tags, he retrieves ≈15,000 instances of the sequence of a degree adverb, a modified head, and the associative marker *de* (after removal of hapax combinations and cases where degree adverbs were attested with fewer than 10 different head types). As for the collostructional application, he applies a co-varing collexeme analysis on the pairs of degree markers and modified heads using, like most studies historically have, -log10 $p_{\text{Fisher-Yates exact}}$ as the measure of collexeme strength (attraction or repulsion).

The collexeme pairs resulting from this analysis are then explored further with two network analysis, a collexeme-based one and a construction-based one. In the former, the nodes consist of the lexical tokens of the constructions and the strengths of links are determined by collostruction strengths and ChatGPT 3.5-based embeddings; in the latter, the nodes are collexeme pairs with links again based on embeddings, this time based on pairwise semantic similarities in a >1500 dimensional vector space. The networks are then studied with community detection methods with an eye to exploring semantically-based co-occurrences and semantic fields emerging from the communities identified.

The results show that degree adverbs form "pivot constructions" with small semantically-motivated groups and that a range of communities can be found, several of which form metaphorical coherences with horizontal relations among them giving rise to generalizations of higher-level constructional schemas or constructional families.

Liao, Gries, & Wulff is another study on Mandarin Chinese. They target the dative alternation – an alternation of five different constructions – in two corpora: (i) the Text of Recent Chinese corpus, a small (≈1m words) written corpus but one that sampled nicely comparably to the Brown corpus of American English and (ii) the CallFriend-Mainland Mandarin corpus, a small (≈273,000 words) spoken corpus. They POS-tag the corpora and then retrieve all instances of 354 verb candidates that have been identified as participating in ditransitive constructions using a comprehensive sampling strategy to strike a balance between a decent coverage of constructional occurrences, a token frequency threshold for each verb of 5 in each of the written and the spoken data, and a minimization of the effect that repeated measurements in the form of multiple occurrences of ditransitives from a single author/speaker might have. They then apply different versions of multiple distinctive collexeme analysis to the verb-by-construction resulting from the previous step, comparing the traditional binomial tests against alternatives such as Pearson

residuals (Gries 2023), multiple log odds ratios, and contributions to the Kullback-Leibler divergence *KLD* (Gries 2024).

The results are interesting on a linguistic level in how the verbs attracted to the five ditransitive constructions indicate different semantic/functional preferences of the constructions, in particular with regard to what is transferred in the ditransitive and the directionality of the transfer events. In addition, the study offers methodological advice for multiple distinctive collexeme analyses by suggesting in particular the use of contributions to the *KLD* because of the combined advantages of the ability to distinguish directions of attraction/repulsion, lower correlation with mere co-occurrence frequency, and a high speed of computation, which is attractive for computing confidence intervals for collostructional strengths, an unfortunately still underutilized method (see Gries 2023, 2024, Olguín Martínez & Gries 2025).

Daugs & Lorenz explore English negative modal constructions comparing contracted vs. non-contracted versions (e.g. *shouldn't* vs. *should not*) in the 1990-2021 part of COCA (the Corpus of Contemporary American English). They retrieved ≈200,000 trigrams, namely modal constructions with pronominal subjects (personal pronouns, existential *there*, *this*, *that*, *wh*o, and *which*). They apply a hybrid of a distinctive collexeme analysis and a co-varying collexeme analysis they call distinctive co-varying collexeme analysis (following Stefanowitsch & Flach, who essentially re-used the hierarchical configural frequency analysis approach of Stefanowitsch & Gries 2005) and as their statistical measures they use a simplified version of the log-likelihood score G^2 called simple log-likelihood G^2_{simple}. together with surprisal values computed as $-\log_2 p(\text{verb}|\text{subj mod}_{\text{neg}})$.

Their findings suggest that, even though there is of course considerable overlap in co-occurrences and even though contracted and uncontracted forms with the same subject and verb will not have different communicative functions, negative modal contractions and their uncontracted parent form still deserve to be treated separately, given their different degrees of entrenchment and conventionalization, which in turn merit different idealized associative networks for contracted forms and their uncontracted counterparts; combinations of subjects, modals, and verbs do have different preferred modal meanings.

Jensen's study also uses COCA data – specifically the 2010-2019 subset of COCA of about 250m words – and contrasts the *go (a)round* V*ing* with the *go (a)round and* V construction. He applies (i) simple collexeme analyses to the verb slots of each construction separately (with an eye to inductively identifying semantic and discourse prosodies from the results) but, more importantly, (ii) distinctive collexeme analysis to the comparison of the two constructions, where the main innovative feature is that the method is applied to not just the verbs in the constructional slots, but also to other contextual features such as semantic and discursive prosodies, colligational patterns that the constructions are used in (e.g., do support, imperative, infinitives, etc.), speech acts (statements vs. directives, questions, and commissives). The *go (a)round* V*ing* construction has distinctly negative semantic and discourse prosodies and serves as a negative stance marker, while the *go (a)round and* V construction is much rarer and exhibits more diverse/less systematic patterns; but the more important contribution is the way in which the distinctive collexemic approach is extended from the typical constructional slot (often, the verb in the construction) to other features that usage-based theories claimed should be relevant for constructional profiles but that collostructional studies often did not include (at least quantitatively).

Like the previous two studies by Daugs & Lorenz and Jensen, the next study is also on American English, but while all studies discussed so far were synchronic and highly quantitative in nature, Schönefeld is a study of smell verbs that adopts a diachronic perspective and highlights

the usefulness of collostructional methods in a more qualitative perspective. From three different time periods of COHA (the Corpus of Historical American English) – the 1820s, the 1920s, and the 2010s – she retrieves instances of the verb lemmas SMELL, STINK, REEK, and SCENT in eight structural patterns (including, but not limited to, intransitive constructions, V *of*/*like*/*with* N, particle verb constructions).

Like Jensen, Schönefeld uses simple collexeme analysis and distinctive collexeme analysis (with the log-likelihood ratio G^2 as the measure of collexeme strength) to see what types of smell descriptors were used by American English speakers in the time periods studied, how they differ in terms of prominence, and what diachronic changes can be observed and maybe explained. Her results show that most diachronic effects are lexical in nature: the words in the constructions change more than the constructions themselves and, in general at least, there is a notable increase in frequency and degree of diversification over time. However, there are also clear exceptions to these overall trends and, more interestingly even, there are diachronic trends specifically applying to 'more metaphorical' or evaluative uses of, e.g., STINK, namely when applied to case of socially stigmatized behaviors (Schönefeld's examples include condescension and illiteracy); however, the results do not support previous work's findings that smell words are primarily used figuratively.

Studies in learner corpus contexts, or applied linguistics kind of contexts, can also benefit from collostructional methods, as is demonstrated by the next two studies. The first of these is Gilquin's study of transfer of collostructions in the case of causative constructions (such as *John makes Mary laugh*); specifically, a first analysis compares verbs in the V_{inf} slot of the English construction and its French equivalent, [X FAIRE V_{inf} Y], and a second one compares verbs used in the V slot of [X MAKE Y V_{inf}] by native speakers of English, French-speaking learners of English and learners of English from other mother tongue backgrounds. Her native-speaker English data are from a 5m word sample from the academic texts of the BNC (British National Corpus, 258 causatives) while her native-speaker French data are from an equally-sized academic writing component of Scientext (2015 causatives), and she uses the log odds ratio (as again an association measure that is less strongly correlated with mere co-occurrence frequency than the default choice of p_{FYE}).

Her first analysis reveals a variety of differential preferences, but the even more interesting part is the one with the analyses of (i) contrasting native and learner English (the native vs. interlanguage comparison in the Integrated Contrastive Model she adopts as her theoretical foundation) and (ii) French vs. general learner English. Her results suggest the existence of collostructional transfer by the learners from French to English as when change of state or location verbs (or other specific verbs) are statistically preferred in the French learner data or when copular verbs other than *be* are dispreferred.

The other learner study is De Los Reyes and Römer-Barron's exploration of Japanese noun-modifying clause constructions (NMCCs), a frequent construction that has so far mostly been studied only qualitatively. Their data come from I-JAS (the International Corpus of Japanese as a Foreign Language), an 8m-words corpus containing Japanese written and spoken by more than 1000 learners and detailed metadata regarding the language users and their proficiency levels. Specifically, they focus in the dialogue task part of that corpus (≈3.2m words) and retrieve more than 4400 concordance lines with NMCCs from 850 learners and 50 native speakers and then run two simple collexeme analyses on the head nouns – one for the learners, one for the native speakers – based on the log-likelihood score G^2 (with a Bonferroni correction for multiple post-hoc tests) as their measure of collexeme strength.

Their results have relevant implications on both a theoretical/linguistic level and on an

applied/pedagogical level. This is the first study to identify POS (sub-)categories that are most frequent in the modifying clauses' predicates and the types of nouns in the constructions. For example, while both learners and native speakers of Japanese use auxiliary verbs most frequently as the clause's predicate, the exact lexical choices differ; the authors are able to relate this difference to how Japanese for Foreign Language learner textbooks describe and exemplify NMCCs and to how exercises often prompt learners to identify people and things in picture description tasks.

The final study in this special issue is by Newman, who revisits a construction that was used as an explanatory vehicle in the very first collostructional study by Stefanowitsch & Gries (2003), the N *waiting to happen* construction. The 2003 paper used the 100m word BNC and discussed the often negative overtones of the construction as revealed by *accident* and *disaster* being the strongest collexemes of the construction's noun slots, but Newman's study now uses the 1b word COCA with its eight registers and submits the total 735 instances of some noun in this construction to a simple collexeme analysis. His results return the same two strongest collexemes and the same negative connotations of the construction, but Newman then proceeds to discuss the implications of several methodological choices that, in one way or the other, underlie nearly all collostructional studies and whose consequences may not always have been sufficiently explored. These include

- the notion of tokenization, such as what counts as 'the word' in a construction slot – in the case of nouns, e.g., just head nouns or also noun compounds?
- whether or not to use lemmas (like most collostructional studies have done) or inflectional forms (which come with more precision but also lower numbers; see Rice & Newman 2005, Newman & Rice 2006, and Gries (2011) for earlier systematic comparison of forms vs lemmas);
- how much context of a (slot in a) construction needs to be used for making correct inferences regarding the semantic, functional, or connotational characteristics of a construction;
- which parts of a context – e.g. which registers and/or time slices– are utilized for a collostructional study.

While all of these issues have been discussed in many different corpus-linguistic applications, they certainly have been understudied in collostructional studies, leading to a maybe often simplistic view, or one that is very heuristic and not very granular, which means that 'meta studies' such as Newman's are important to critique, improve, and extend corpus methods like collostructional analysis.

3 Concluding remarks and where to go from here

Collostructional analysis 'has had a good run': it has been a very widely used method in especially cognitive-linguistic or usage-based linguistics, but also more generally in corpus linguistics; the two main implementations – Gries's coll.analysis R function and Flach's collostructions R package – have been used in a huge number of studies, and our understanding of many constructions and their semantic, functional, discourse-prosodic, and connotational characteristics has benefited immensely from the ease of applicability and interpretability of the results offered by

collostructional analysis. That being said, the studies from this special issue highlight that collostructional analysis should not be resting on its laurels and, thankfully, some work has already begun to expand our view. With the bias that is naturally coming with the two authors of this introduction, the main desiderata come under the (partially interrelated) headings of *increased resolution* and *multivariateness*. Increased resolution addresses the fact that, in some sense, traditional collostructional studies involve really very little information, namely only some construction and lemmas in one slot; thus, the suggestions are to

- input not just lemmas but maybe also inflectional forms;
- input not just simple words or forms, but also, e.g., compounds and especially word-sense combinations;
- include not just constructions and material specific to one (in the sense of collexeme or distinctive collexeme analysis) or two (co-varying collexeme analysis) slots but also other information 'surrounding' the construction.

These points, all of which were discussed in the papers of this special issue, would massively increase the amount of information we would get from the corpus data. However, that also means we must up our quantitative game by recognizing the multivariateness that results from the increased resolution. This can be handled in several ways, too:

- we can make sure we do not rely too much on quantitative corpus measures that conflate various dimensions of information such that
  - we should make sure that our measures of collexeme strength are interpretable and do not conflate frequency and association in irrecoverable ways;
  - we should probably distinguish directions of attraction;
  - we should incorporate dispersion (either on the time slice, register, or even file/speaker level);
- we can annotate multiple features of the constructional uses at the same time and include them in simple extensions, as when Stefanowitsch & Flach (2020), Olguín Martínez & Gries (2024), or Jensen & Gries (2025, a follow-up to Jensen, this issue) explore different ways to include more than just two things – one or two constructions and one set of things in some slots of theirs – in the analysis; in the same vein, this can lead to the recognition that more complex methods such as network analysis need to be used more often and broadly, or that more powerful follow-up methods (e.g. from the realm of predictive modeling) are integrated as well;
- we can make sure that we provide confidence intervals for our results (see Gries 2019, 2023).

Ideally, of course, all these things would happen at the same time. However insightful collostructional analysis has been over the last 20 years, it is time to move on from what was an essentially crude but insightful first and monofactorial heuristic – something that in modeling would be written as CONSTRUCTION ~ LEMMA – to improved versions that mirror how much more sophisticated quantitative corpus linguistics has become. To put it somewhat polemically: we do not need the 534th study of some niche construction in some language or niche register that otherwise does everything like it was done 15-20 years ago, we need the field to follow the current developments (and of course the current special issue's authors' lead) and move collostructions to

the next level; that's how this approach will remain meaningful and consequential in both theoretical and applied linguistic contexts.

References

Chen. Alvin. this issue. From sequentiality to schematization: A two-tier network analysis of covarying collexemes in Mandarin Degree Adverb Constructions. *Corpus Linguistics and Linguistic Theory*.

Daugs, Robert & David Lorenz. this issue. A radically usage-based, collostructional approach to assessing the differences between negative modal contractions and their parent forms. *Corpus Linguistics and Linguistic Theory*.

De Los Reyes, Nicole C. & Ute Römer-Barron. this issue. A collostructional approach to Japanese noun-modifying clause construction use and acquisition: a learner corpus study. *Corpus Linguistics and Linguistic Theory*.

Jensen, Kim Ebensgaard. this issue. *Well, maybe you shouldn't go around shaving poodles*: collostructional semantic and discursive prosody in the *go (a)round Ving* and *go (a)round and V* constructions. *Corpus Linguistics and Linguistic Theory*.

Gilquin, Gaëtanelle. this issue. Transfer of collostructions: the case of causative constructions. *Corpus Linguistics and Linguistic Theory*.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure.* Chicago: University of Chicago Press.

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. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In Mario Brdar, Stefan Th. Gries, & Milena Žic Fuchs (eds.), *Cognitive linguistics: convergence and expansion*, 237-256. Amsterdam & Philadelphia: John Benjamins.

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

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

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.

Hunston, Susan & Gill Francis. 1998. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam & Philadelphia: John Benjamins.

Jensen, Kim Ebensgaard & Stefan Th. Gries. 2025. GO *(a)round and* V vs. GO *(a)round* V*ing*: A multivariate distinctive collo-profiling analysis based on association rules. *Review of Cognitive Linguistics*.

Liao, Shengyu, Stefan Th. Gries, Stefanie Wulff. this issue. Transfer five ways: applications of multiple distinctive collexeme analysis to the dative alternation in Mandarin Chinese.

Corpus Linguistics and Linguistic Theory.
Newman, John. this issue. Revisiting N *waiting to happen*: word, construction, and corpus choices in a collostructional analysis. *Corpus Linguistics and Linguistic Theory*.
Newman, John & Sally Rice. 2006. English adjectival inflection: a radical Radical Construction Grammar approach. Paper presented at Conceptual Structure, Discourse, and Language, San Diego, CA, USA.
Olguín Martínez, Jesús Francisco & Stefan Th. Gries. 2024. *If not for-if it weren't/wasn't for* counterfactual constructions: A multivariate extension of collostructional analysis. *Cognitive Semantics* 10(2). 159-189.
Olguín Martínez, Jesús Francisco & Stefan Th. Gries. 2025. The similative-pretence alternating pair and filler-slot relations: A revised version of distinctive collexeme analysis. *Constructions and Frames*.
Rice, Sally & John Newman. 2005. Inflectional islands. Paper presented at the International Cognitive Linguistics Conference, Seoul, South Korea.
Schönefeld, Doris E. this issue. Expressing smells in (American) English. *Corpus Linguistics and Linguistic Theory*.
Stefanowitsch, Anatol and Susanne Flach. 2020. *Too big to fail* but *big enough to pay for their mistakes*: A collostructional analysis of the patterns [*too* adj *to* V] and [adj *enough to* V]. In Gloria Corpas Pastor & Jean Pierre Colson (eds.), *Computational Phraseology*, 247-272. Amsterdam & Philadelphia: John Benjamins.
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.