Collostructional Analysis

Gaëtanelle Gilquin[a], **and Stefan Th Gries**[b], [a] UCLouvain, Louvain, Belgium and [b] UC Santa Barbara, United States & JLU Giessen, Giessen, Germany

Abstract

This article deals with the methods of collostructional analysis, which make it possible to investigate the words that typically occur in certain slots of one or more constructions (in the Construction Grammar sense). It describes the origin of collostructional analysis, its status as a usage-based approach, and the different methods it covers. Computation aspects are discussed, as well as techniques to analyze the output of the computation semantically, stylistically, or according to other features. It is also shown how collostructional analysis has evolved over the years, being applied to an increasing range of constructions, languages, and fields, and undergoing conceptual and methodological developments.

Key Points

- Collostructional analysis was developed in the early 2000s as a family of methods to measure the degree of interaction between words and constructions;
- The study of the collexemes attracted to a construction provides insights into the semantic, stylistic, functional, cultural, etc. features of the construction;
- Since its advent, collostructional analysis has been applied to a growing number of constructions, languages, and fields, and several changes and extensions have been proposed.

Introduction

In a trilogy of articles published in the early 2000s, Stefan Th. Gries and Anatol Stefanowitsch introduced a new method, or rather family of methods, known as collostructional analysis (**Gries & Stefanowitsch, 2004a**,**b**; **Stefanowitsch & Gries, 2003**). Collostructional analysis was presented as an "extension of collocational analysis" (**Stefanowitsch & Gries, 2003**, p. 209). While collocational analysis examines the interaction between words (e.g., how the verb SUGGEST interacts with other words to the left and/or to the right such as *I*, *would*, *that*, or *strongly*), collostructional analysis considers the interaction between words and constructions. For example, it can investigate the words that typically occur in the verb slot of the [*SUGGEST* V_{ing}] construction (e.g., *going*, *leaving*, *using*). Like collocational analysis, it relies on authentic language data from corpora and on statistics to identify these words—called "collexemes" in collostructional analysis.

The notion of construction as it is used in collostructional analysis corresponds to a pairing of form and meaning/function, as defined in **Goldberg's (1995**, **2006)** Construction Grammar. Collostructional analysis can therefore apply to any type of construction recognized in Construction Grammar, including partially filled constructions such as [*SUGGEST* V_{ing}] and fully abstract constructions such as [Subj V Particle Obj], as well as morphological constructions like the *-se* or *-ra* endings for Spanish verbs in the imperfect subjunctive (see **Guzmán Naranjo, 2017**). It has also been used to distinguish between alternating constructions, such as the dative or active-passive alternations (**Gries & Stefanowitsch, 2004a**), which, in keeping with Construction Grammar, are considered distinct constructions with their own meanings/functions.

By building on collocational analysis and relying on the notion of construction, collostructional analysis has brought together the frameworks of Corpus Linguistics and Construction Grammar, as indicated by the term "collostruction" itself, a blend of "collocation" and "construction". It is thus arguably a usage-based approach in the two senses of the word: (i) in the original sense of an approach assuming that language systems gradually emerge from usage, as advocated in Construction Grammar (see **Diessel, 2015**), and (ii) in the recent reinterpretation of the term describing an approach based on authentic language data.

The Methods of Collostructional Analysis

Traditional collostructional analysis comprises three main methods: collexeme analysis, (multiple) distinctive collexeme analysis, and covarying collexeme analysis. Collexeme analysis investigates one slot in a particular construction, as illustrated by the [*SUGGEST* V_{ing}] example or by **Hartmann & Ungerer, 2024** analysis of [*the mother of all* X]. Distinctive collexeme analysis makes it possible to compare the collexemes of two constructions that fulfill similar functions. **Wulff (2007)**, for instance, uses distinctive collexeme analysis to compare the verb slot of [*GO-and*-V] and [*GO*-V], two constructions that are functionally similar but turn out to attract quite different verbs. When more than two constructions are compared, a multiple distinctive collexeme analysis can be used, as is the case in **Gilquin's (2006)** comparison of the non-finite verb slots of ten periphrastic causative constructions in English ([X *CAUSE* Y $V_{to\text{-inf}}$], [X *GET* Y V_{ing}], [X *GET* Y $V_{to\text{-inf}}$], etc.). Covarying collexeme analysis, finally, considers the association between two slots within one construction, such as the comparative filler phrases in the comparative correlative construction (e.g., *the sooner, the better*) in **Hoffmann et al. (2019)**.

The computation of collostructional analysis implies that we know the frequency of the construction(s) under study and the frequency—inside and outside the construction(s)—of the words occurring in the selected slot(s). Because of the more syntactic nature of many of the constructions that have been studied using collostructional analysis, parsed corpora are a useful resource because they include information about syntactic structure, functions, and relationships that can help with the automatic retrieval of constructions. Gries and Stefanowitsch's early studies (**Gries & Stefanowitsch, 2004a**; **Stefanowitsch & Gries, 2003**), for example, exploited the parsed version of ICE-GB, the British component of the International Corpus of English. With more manual work (and possibly lower recall), however, part-of-speech tagged corpora can also serve as a basis for the retrieval of constructions through the search for sequences of words and/or parts of speech.

The quantitative information derived from corpora makes it possible to compare observed and expected frequencies for each combination of construction and word, using some distributional statistic. Words that are more frequent than expected in a construction are said to be attracted to the construction, whereas those that are less frequent than expected are repelled. Scripts are available that compute these statistics (**Gries's (2004/2022)** Coll.analysis 3/4 script and **Flach's (2021)** collostruction package) and provide different statistical measures of the strength of association or repulsion between a word and a construction. So far, the focus of collostructional analysis has mainly been on words that display a significant degree of attraction to a construction, but repelled words have been shown to be worthy of attention too (**Stefanowitsch, 2008**).

Typically, the output of a collostructional analysis is analyzed by grouping the collexemes around identifiable semantic classes (although other kinds of classes—functional, stylistic, cultural, or register-based ones—have also been used, see below). Thus, **Wulff et al. (2007)** compare the collexemes in the verb slots of the *into*-causative construction ([Subj V1 Obj *into* $V2_{ing}$]) in British and American English. They find that, in the V1 slot, the collexemes in British English tend to express physical force (e.g., *push*) whereas those in American English more often involve verbal persuasion (e.g., *talk*). While their semantic analysis of the collexemes is based on manual classification, attempts have been made to automate the classification process, for example through cluster analysis (**Gries & Stefanowitsch, 2010**) or network analysis (**Gries & Ellis, 2015**). A related important question is how to take polysemy into consideration, i.e. the fact that, e.g., verbs in constructional slots often have different senses. Studies that have specifically addressed this issue include **Gilquin (2010**, **2013)**, **Colleman and Bernolet (2012)**, **Bernolet and Colleman (2016)**, and **Gries (2023)**.

Applications of Collostructional Analysis

As already mentioned, collostructional analysis can be—and has been—applied to different types of constructions, including alternating constructions. By focusing on the collexemes of these constructions, it has underlined their phraseological nature, since constructions often show strong preferences for certain words in a given slot. Even more importantly, collostructional analysis has relied on collexemes to improve our understanding of the meanings of constructions and to emphasize semantic differences between constructions often considered equivalent. In **Wulff's (2007)** study referred to above, for example, the collexemes indicate that the [*GO-and*-V] construction refers to the whole action, from beginning to end, whereas the [*GO*-V] construction "only denotes the initiation of an action and is inherently atelic" (**Wulff, 2007**, p. 121). Collexemes can also offer useful insights into the stylistic features of constructions. **Stefanowitsch and Gries (2008)** thus show that the passive construction is associated with more formal collexemes than the active construction, both in writing and speech. Sometimes, differences in collexemes can be related to cultural specificities. In **Wulff et al.'s (2007)** study of the *into*-causative construction in British and American English, the collexemes appear to include movement-initializing verbs (e.g., *stir*) in British English as opposed to movement-restricting verbs in American English (e.g., *rope*). This difference is said to be related to "the strong and explicit emphasis on mobility as an essential condition for a happy and free life as we find it in American culture" (**Wulff et al., 2007**, p. 279). Interestingly, collostructional analysis can also help discover constructions, as demonstrated by **Hampe (2011)**. Her investigation of the collexemes of the [Subj V Obj_{Dir} $Compl_{Pred}$(NP)] argument-structure construction reveals the presence of many denominative verbs (e.g., *call*, *name*, *term*), which leads her to postulate what she calls the "Denominative Construction", described semantically as [X SAY [Y BE Z]].

While the first collostructional analyses were carried out on English language data, including different varieties of it (e.g., Hong Kong, Indian, and Singapore English in **Mukherjee & Gries, 2009**), further studies have tackled other languages. **Hilpert (2006)**, for example, performs collexeme analyses to study Swedish future constructions (with *ska*, *skall*, *skulle*, *komma*, and *tänka*), while **Jensen (2012)** uses the same method to study the [V *ihjel*] ("V *to death*") construction in Danish. Distinctive collexeme analysis is applied by **Colleman (2009)** to the Dutch dative alternation and by **Verroens, 2024** to the French ingressive constructions [*se mettre à* V_{inf}] and [*commencer à* V_{inf}]. Some scholars have also compared constructions in several languages. **Noël and Colleman (2010)** investigate the "accusative and infinitive" construction (of the type [X *BELIEVE* Y $V_{to\text{-}inf}$]) and the "nominative and infinitive" construction (of the type [Y *BE* believed $V_{to\text{-}inf}$]) in English and Dutch. The crosslinguistic comparison is drawn by means of a distinctive collexeme analysis in English and another one in Dutch. In **Gilquin (2015)**, the infinitive verb slots of the English causative construction [X *MAKE* Y V_{inf}] and its French equivalent [X *FAIRE* V_{inf} Y] are compared through a single distinctive collexeme analysis, on the basis of a translation of the French verbs into English.

Contrastive linguistics is not the only field that has benefited from collostructional analysis. The collostructional approach has also been used in diachronic linguistics, as a way to trace the historical evolution of constructions (see, e.g., **Hilpert, 2012**). **Noël and Colleman's (2010)** study mentioned above investigates the collexemes of the "accusative and infinitive" and "nominative and infinitive" constructions in English and Dutch over a period of almost three centuries (1640–1920). The findings suggest that the "nominative

and infinitive" construction in English, and probably in Dutch too, is not just a passive equivalent of the "accusative and infinitive" construction but has always had a distinct symbolic value. Second language acquisition is another field of linguistics to which collostructional analysis has made a valuable contribution. Comparing the collexemes of a construction in native and learner language highlights possible differences between the two varieties. The distinctive collexeme analysis of [X *MAKE* Y V_{inf}] in **Gilquin (2012)**, applied to the comparison of native English and learner English, shows, for example, that while both learners and native speakers use copular verbs in the construction, learners favor *be* and *become*, whereas native speakers prefer more specific copulas such as *seem*, *appear*, *sound*, or *look*. This type of information can then be used to develop targeted pedagogical activities. Among other fields of linguistics that have adopted collostructional analysis, contact linguistics can also be mentioned. In **Mukherjee and Gries (2009)**, the collexemes of the intransitive, monotransitive, and ditransitive constructions are compared across Hong Kong, Indian, and Singapore English, three varieties of New Englishes that correspond to different evolutionary stages. The analysis reveals that the more advanced in its evolution a New English variety is, the more its collostructional preferences differ from those in British English, the historical input variety, pointing to an increasing degree of nativization in collexemes.

New Developments

Over the years, changes to and extensions of the three traditional methods have been proposed. Some of these are conceptual in nature, such as when distinctive collexeme analyses do not explore two constructions but one construction in two or more different settings. Examples include **Wulff et al.'s (2007)** or **Mukherjee and Gries's (2009)** studies comparing constructions across varieties of English and **Gilquin's (2012)** comparison of constructions in native versus learner English.

Other changes/extensions (also) involve methodological questions. One recurrent question is what distributional statistic(s) to use for quantifying attraction/repulsion. Most studies have used the (negative of the logged) *p*-value of a Fisher-Yates exact test because the way it heuristically conflates frequency and association into a single sortable value and avoids distributional assumptions was very appealing for exploratory and applied purposes. However, it has also been argued (**Gries, 2019**, **2023**) that, for theoretical, cognitive, or psycholinguistic work, it would be better to measure and express collostructional attraction/repulsion:

- using statistics that measure frequency and both directions of association (e.g., verb-to-construction and construction-to-verb) as independently from each other as possible (so that the contributions of each statistical dimension of information to learning, processing, priming, etc. can be properly distinguished);
- maybe adding more dimensions to the mix such as dispersion, i.e. the degree to which collostructional co-occurrences are distributed evenly or unevenly in corpora (so that frequency-based statistics are not undermined by underdispersion in corpora, as when *fold* and *process* are attested often in the imperative, but only in one of 500 corpus files);
- avoiding conflating dimensions into a single value but instead reporting a tuple of collostructional statistics.

Gries (2019) demonstrates how this kind of approach can shed new light (in the form of "more precision") on old results.

A related improvement that has been discussed involves the fact that collostructional methods usually only provide point estimates for collostructional strengths, but they do not provide any indication of their robustness, or uncertainty, and this even though confidence intervals and similar statistics are very widespread in other areas of corpus linguistics. **Gries (2022**, **2023)** reports on first applications of bootstrapped confidence intervals, which suggest that the clear-cut rankings of many previous analyses may be much fuzzier than has been assumed so far.

A final set of current developments include increasing the dimensionality of the co-occurrence data studied. The vast majority of studies are two-dimensional in nature in how they involve one lexical item (or not) in one construction (or one or more others) or in one variety or another, etc. Yet, including more dimensions of co-occurrence information is ultimately indispensable, especially from a usage-based perspective. **Stefanowitsch and Gries (2005)** proposed a first three-dimensional extension of co-varying collexeme analysis, but this was not taken up much. More recently, however, there has been a renewed interest in developing a more fine-grained picture of collostructional patterns of attraction. For example, **Stefanowitsch and Flach (2020)** propose a "distinctive co-varying collexeme analysis" to explore associations between two slots in two related constructions, looking at how certain adjectives combine with different verbs in the [*too* Adj *to* V] and [Adj *enough to* V] constructions. Much more broadly, **Hampe and Gries (2018)** explore eight mini constructions on the basis of four additional dimensions (binary predictors specifying verb and clause types as well as the presence of polarity markers and perfects) and a multinomial regression to, among other things, determine constructional prototypes. Using even more dimensions, **Olguín Martínez & Gries, 2024** specifically propose a type of multivariate collexeme analysis and exemplify it with a five-dimensional study of counterfactual constructions.

In sum, while collostructional methods are, at their core, quite simple and have been around for a while, there are several new developments that promise to increase their precision, the rigor, and the comprehensiveness.

Conclusion

This article has provided an overview of collostructional analysis, showing how its methods can be used to investigate the interaction between words and constructions with a view to improving our understanding of constructions: their phraseological nature, their semantic or

stylistic preferences, their cultural specificities, their variation according to registers or varieties, etc. The wide range of applications of collostructional analysis has also been emphasized, with many constructions, languages, and fields having benefited from these methods, and new developments likely to increase the scope and significance of collostructional analysis.

As evidence for collostructional preferences accumulates, it has become clear that collexemes are central to the description of constructions—so much so that **Herbst (2018**, p. 18) has suggested transforming **Goldberg's (2006**, p. 18) credo "it's constructions all the way down" into "it's collexemes (or items) all the way down". **Herbst (2018)** has also argued that the description of constructions should include a "collo-profile", with examples of items typically occurring in the open slots of the constructions, and that this information should be part of a "collostructicon", as a new kind of reference work supplanting dictionaries and grammar books. More generally, collostructional analysis has gradually established itself as one of the most popular approaches in Construction Grammar. With its reliance on corpus data, it can be said to have contributed to the empirical turn of Construction Grammar. This, in turn, has helped enhance the value of the Construction Grammar framework and its validity as a usage-based model, for the benefit of all its users.

References

Bernolet, S., & Colleman, T. (2016). Sense-based and lexeme-based alternation biases in the Dutch dative alternation. In Yoon, J., & Gries, S.T. (Eds.), Corpus-based approaches to construction grammar (pp. 165–198). Amsterdam: John Benjamins.

Colleman, T. (2009). Verb disposition in argument structure alternations: A corpus study of the dative alternation in Dutch. Language Sciences, 31(5), 593–611.

Colleman, T., & Bernolet, S. (2012). Alternation biases in corpora vs. picture description experiments: DO-biased and PD-biased verbs in the Dutch dative alternation. In Divjak, D., & Gries, S.T. (Eds.), Frequency effects in language representation (pp. 87–125). Berlin: de Gruyter.

Diessel, H. (2015). Usage-based construction grammar. In Dąbrowska, E., & Divjak, D. (Eds.), Handbook of cognitive linguistics (pp. 296–322). Berlin: de Gruyter.

Flach, S. (2021). Collostructions: An R implementation for the family of collostructional methods. Package version v.0.2.0 https://sfla.ch/collostructions/.

Gilquin, G. (2006). The verb slot in causative constructions. Finding the best fit. Constructions, SV1. https://constructions.journals.hhu.de/article/view/445.

Gilquin, G. (2010). Corpus, cognition and causative constructions. Amsterdam: John Benjamins.

Gilquin, G. (2012). Lexical infelicity in English causative constructions. Comparing native and learner collostructions. In Leino, J., & von Waldenfels, R. (Eds.), Analytical causatives. From "give" and "come" to "let" and "make" (pp. 41–63). München: Lincom Europa.

Gilquin, G. (2013). Making sense of collostructional analysis. On the interplay between verb senses and constructions. Constructions and Frames, 5(2), 119–142.

Gilquin, G. (2015). Contrastive collostructional analysis: Causative constructions in English and French. Zeitschrift für Anglistik und Amerikanistik, 63(3), 253–272.

Goldberg, A.E. (1995). Constructions. A construction grammar approach to argument structure. Chicago: The University of Chicago Press.

Goldberg, A.E. (2006). Constructions at work. The nature of generalization in language. Oxford: Oxford University Press.

Gries, S.T. (2004/2022). Coll.analysis 3/4. A program for R.

Gries, S.T. (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, S.T. (2022). Towards more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more. Research Methods in Applied Linguistics, 1(1), 100002.

Gries, S.T. (2023). Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. Cognitive Semantics, 9(3), 351–386.

Gries, S.T., & Ellis, N.C. (2015). Statistical measures for usage-based linguistics. Language Learning, 65(S1), 228–255.

Gries, S.T., & Stefanowitsch, A. (2004a). Extending collostructional analysis. A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics, 9(1), 97–129.

Gries, S.T., & Stefanowitsch, A. (2004b). Covarying collexemes in the *into*-causative. In Achard, M., & Kemmer, S. (Eds.), *Language, culture, and mind*. Stanford: CSLI (pp. 225–236).

Gries, S.T., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In Rice, S., & Newman, J. (Eds.), Empirical and experimental methods in cognitive/functional research (pp. 73–90). Stanford: CSLI.

Guzmán Naranjo, M. (2017). The se-ra alternation in Spanish subjunctive. Corpus Linguistics and Linguistic Theory, 13(1), 97–134.

Hampe, B. (2011). Discovering constructions by means of collostruction analysis: The English Denominative Construction. Cognitive Linguistics, 22(2), 211–245.

Hampe, B., & Gries, S.T. (2018). Syntax from and for discourse II: More on complex-sentences as meso-constructions. In Hampe, B., & Flach, S. (Eds.), Yearbook of the German cognitive linguistics association (pp. 115–142). Boston: de Gruyter.

Hartmann S. Ungerer T. 2024Attack of the snowclones: A corpus-based analysis of extravagant formulaic patternsJournal of Linguistics603599634 .

Herbst, T. (2018). Is language a collostructicon? A proposal for looking at collocations, valency, argument structure and other constructions. In Cantos-Gómez, P., & Almela-Sánchez, M. (Eds.), Lexical collocation analysis: Advances and applications (pp. 1–22). Cham: Springer.

Hilpert, M. (2006). A synchronic perspective on the grammaticalization of Swedish future constructions. Nordic Journal of Linguistics, 29(2), 151–172.

Hilpert, M. (2012). Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In Allan, K., & Robinson, J. (Eds.), Current methods in historical semantics (pp. 133–160). Berlin: de Gruyter.

Hoffmann, T., Horsch, J., & Brunner, T. (2019). The more data, the better: A usage-based account of the English comparative correlative construction. Cognitive Linguistics, 30(1), 1–36.

Jensen, K.E. (2012). Fatal attraction: Inheritance and collostruction in the *ihjel*-construction. Skandinaviske Sprogstudier, 3(2), 1–30. https://tidsskrift.dk/sss/article/view/6250/5416.

Mukherjee, J., & Gries, S.T. (2009). Collostructional nativisation in new Englishes: Verb-construction associations in the international corpus of English. English World-Wide, 30(1), 27–51.

Noël, D., & Colleman, T. (2010). *Believe*-type raising-to-object and raising-to-subject verbs in English and Dutch: A contrastive investigation in diachronic construction grammar. International Journal of Corpus Linguistics, 15(2), 157–182.

Olguín Martínez J. Gries S.T. 2024*If not for-if it weren't/wasn't for* counterfactual constructions: A multivariate extension of collostructional analysisCognitive Semantics102158189 .

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. Cognitive Linguistics, 19(3), 513–531.
Stefanowitsch, A., & Flach, S. (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 Corpas Pastor, G., & Colson, J.-P. (Eds.), Computational phraseology (pp. 247–272). Amsterdam: John Benjamins.
Stefanowitsch, A., & Gries, S.T. (2003). Collostructions: Investigating the interaction of words and constructions. International Journal of Corpus Linguistics, 8(2), 209–243.
Stefanowitsch, A., & Gries, S.T. (2005). Covarying collexemes. Corpus Linguistics and Linguistic Theory, 1(1), 1–43.
Stefanowitsch, A., & Gries, S.T. (2008). Channel and constructional meaning: A collostructional case study. In Kristiansen, G., & Dirven, R. (Eds.), Cognitive sociolinguistics: Language variation, cultural models, social systems (pp. 129–152). Berlin: de Gruyter.
Verroens F. 2024Zooming in on the semantics of French ingressives: A collostructional analysisJournal of French Language Studies343305322 .
Wulff, S. (2007). *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In Gries, S.T., & Stefanowitsch, A. (Eds.), Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis (pp. 101–126). Berlin: de Gruyter.
Wulff, S., Stefanowitsch, A., & Gries, S.T. (2007). Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In Radden, G., Köpcke, K.-M., Berg, T., & Siemund, P. (Eds.), Aspects of meaning construction (pp. 265–282). Amsterdam: John Benjamins.