Collostructional Methods

STEFAN T. GRIES

Introduction

One of the most central assumptions in all of corpus linguistics is the so-called distributional hypothesis, which is best captured in Harris (1970, p. 785f):

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

The distributional hypothesis leads directly to one of the most central corpus-linguistic methods, the study of association phenomena, that is, the question of which linguistic elements—morphemes, words, syntactic constructions, and so forth—"like" to co-occur with which other linguistic elements (i.e., there is **attraction**) or "dislike" to co-occur with which other linguistic elements (i.e., there is **repulsion**), and what that reveals. With some simplification, two main domains of association studies can be distinguished: the study of lexical co-occurrence (**collocation**) and the study of lexico-syntactic, or lexico-constructional, co-occurrence (**collostructional analysis**, a blend of *collocation* and *construction*; often, this is also referred to as **colligation**). This overview is concerned with the latter approach. I will first introduce the three main methods of the "collostructional family," before discussing some applications and then turning to recent developments and future desiderata.

Collostructional Analysis

The family of methods of **collostructional analysis** includes three main applications:

- **collexeme analysis**, which quantifies the degree of attraction/repulsion of words to a syntactically defined slot in a construction (cf. Stefanowitsch & Gries, 2003), for example, how much does each verb occurring in the ditransitive (C) like to do so?
- **distinctive collexeme analysis**, which quantifies which words are attracted to/repelled by one of several constructions (cf. Gries & Stefanowitsch, 2004a), for example, how much does each verb occurring at least once in the ditransitive or the prepositional dative like to do so?
- **covarying collexeme analysis**, which identifies (dis)preferred pairs in two slots of one construction (cf. Gries & Stefanowitsch, 2004b), for example, how much does the lemma *fool* like to occur with *thinking* (as in, e.g., *He fooled*$_{\text{verb1}}$ *her into thinking*$_{\text{verb2}}$ *much better of him*)?

Just like nearly all corpus-linguistic association measures, this quantification is usually based on 2 × 2 co-occurrence tables such as Table 1, which schematically represents the frequencies of occurrence of one element ($a + b$) and one construction ($a + c$) as well as the frequency of their co-occurrence (a) and the corpus size (N, usually in constructions), from which the remaining cells (b, c, and d) can be computed.

Many indices can be computed from such tables, and most of them are based on comparing the observed frequencies in Table 1 to those that would be expected under the null hypothesis of there

The Encyclopedia of Applied Linguistics. Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal0258.pub3

Table 1 Frequencies of an element *E* within and outside of a construction *C*

	Construction C	*Other constructions*	Sum
Element *E*	*a*	*b*	$a + b$
Other elements	*c*	*d*	$c + d$
Sum	$a + c$	$b + d$	$a + b + c + d = N$

Table 2 Frequencies of *give* in ditransitives and other constructions in the ICE-GB

	Ditransitive	*Other constructions*	Sum
Give	461	699	1,160
Other verbs	574	136,930	137,504
Sum	1,035	137,629	138,664

being no relation whatsoever between *E* and *C*. The most frequently used measure in collostructional analysis is $-\log_{10}$ of the $p_{\text{one-tailed}}$-value of a Fisher–Yates exact test, but the log-likelihood score $G^2 \rightarrow$ and ΔP have also been used. The following three sections exemplify each collostructional method, one with each statistic.

Collexeme Analysis

The first method, collexeme analysis, requires exactly the kind of input represented schematically in Table 1 and exemplified more concretely in Table 2, which shows how the lemma *give* is distributed across ditransitives and elsewhere in the British Component of the International Corpus of English.

Using the open-source programming language and environment R, it is easy to compute **the log-likelihood score G^2** for *give* in the ditransitive, which amounts to 3,206.235 (because *give* occurs 461 times in the ditransitive, while a random distribution would have it occur there approximately nine times).

```
Table.2 <- matrix(c(461, 574, 699, 136930), ncol=2)¶
2*sum(Table.2 * log(Table.2/chisq.test(Table.2, correct=F)$exp))¶
[1] 3206.235
```

Analogous tests would be done for all verb/lemma types occurring at least once in the ditransitive, and then these verb lemmas can be ranked according to their **collexeme strength**, that is, the strength of their attraction to the ditransitive. With G^2, *give* is the verb most strongly attracted to the ditransitive (followed by *tell*, *send*, *offer*, and *show*), which provides some support for analyses that stipulated that the function of the ditransitive has to do with "transfer" (see also below).

Distinctive Collexeme Analysis

This method contrasts two (or more) constructions with each other regarding which words they prefer to occur with. Table 3 helps determine which construction of the dative alternation (*he gave him a book* vs. *he gave a book to him*) *give* is more attracted to.

Since *give*'s observed frequency in the ditransitive (461) is greater than the one expected by chance ($607 \times 1{,}035/2{,}954 \approx 213$), one can compute the negative $\log_{10}$ of the *p*-value of a **Fisher–Yates exact test** as follows (using the hypergeometric distribution). Again, we obtain a very high value, indicating that *give* strongly prefers the ditransitive over the prepositional dative.

```
-log10(sum(dhyper(461:607, 1035, 1919, 607)))¶
[1] 119.7361
```

Table 3 Frequencies of *give* in ditransitive and prepositional datives in the ICE-GB

	Ditransitive	*Prepositional dative*	*Sum*
Give	461	146	607
Other verbs	574	1,773	2,347
Sum	1,035	1,919	2,954

Table 4 Frequencies of *fool* and *thinking* in the into-causative in the BNC

	Verb 1: fool	*Other verbs in slot 1*	*Sum*
Verb 2: *thinking*	46	31	77
Other verbs in slot 2	101	1,408	1,509
Sum	147	1,439	1,586

Analogous computations for all verbs occurring at least once in one of the two constructions identify the verbs most strongly attracted to each construction, which are listed in (1) and (2) in decreasing strength of association strength:

1. ditransitive: *give*, *tell*, *show*, *offer*, *cost*, *teach*, *wish*, *ask*, *promise*, *deny*, and so forth.
2. prepositional dative: *bring*, *play*, *take*, *pass*, *make*, *sell*, *do*, *supply*, *read*, *hand*, and so forth.

Such results, too, provide strong support for analyses attributing different senses to the two constructions. For example, the ditransitive has been argued to involve constructional senses of transfer, enablement of transfer, communication as transfer, and so forth. The results are also compatible with the constructions' acquisition patterns (where, for example, *give* is a path-breaking verb for the acquisition of the ditransitive).

Covarying Collexeme Analysis

This last method studies the interrelations of words in two slots of one construction, as in the *into*-causative mentioned above and exemplified here in Table 4.

A completely different association measure to compute on such tables is ΔP, a unidirectional/asymmetric association measure that can quantify separately how much

- the element in the first row attracts/repels the element in the first column;
- the element in the first column attracts/repels the element in the first row.

ΔP is just the difference of row/column percentages, which shows here that, while *fool into thinking* seems strongly mutually attracted (given the high G^2 of >134), the ΔP-values reveal that the attraction in one direction is much stronger than in the other:

```
(46/147) - (31/1439) # how much fool attracts thinking¶
[1] 0.2913824
(46/77) - (101/1509) # how much thinking attracts fool¶
[1] 0.5304709
```

Gries & Stefanowitsch (2004a) and Wulff et al. (2007) discuss several case studies revealing patterns explainable with recourse to cognitive, cultural, and stereotypical factors.

Applications

The above three methods have been applied in a wide variety of contexts, registers, and languages (e.g., English, German, Dutch, Italian, Standard Arabic, Mandarin Chinese, and many more). Much work has focused on argument structure constructions, but many other less-semantically loaded constructions have exhibited similar verb-specific effects; examples include *will*-future versus *going to* V (cf. (3)), particle placement (cf. (4)), or *to* versus *ing*-complementation (cf. (5)).

(3) a. He will mess it up.
    b. He is going to mess it up.

(4) a. He will mess up the whole talk.
    b. He will mess the whole talk up.

(5) a. He tried to mess up everything.
    b. He tried messing up everything.

In addition, a growing number of studies use collostructional analysis in

- psycholinguistic contexts, for example, as a measure of verb-construction preferences that will help predict syntactic priming; see Gries (2005);
- second/foreign-language learning, for example, to determine whether learners are acquiring, or have acquired, native-like verb-construction preferences; see especially Ellis et al. (2016);
- language change; see Hilpert (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, Gries et al. (2010) show on the basis of a sentence-completion task and self-paced reading data that the behavior of native speakers of English can sometimes be predicted better on the basis of collexeme strengths ($-\log10\, p_{\text{Fisher–Yates exact}}$) than on the basis of raw frequencies or conditional probabilities alone; see also Flach (2020b). As for the latter, Gries and Wulff (2009) show how advanced German learners of English exhibit collostructional preferences similar to those of native speakers for both the dative alternation and the two complementation patterns exemplified in (5) in both sentence completion tasks and acceptability ratings; see also Flach (2020a).

Current Developments and Desiderata

While collostructional methods have now been widely used for many years, the field is still undergoing active development, both by its main developers and many users of the method. This section outlines some of the recent and current debates, developments, and proposals.

One important matter of debate has concerned the choice of association measure. Most studies have relied on $p_{\text{Fisher–Yates}}$ because of (a) its mathematical properties (as an exact test, it has no distributional requirements) and (b) the fact that this test is a good heuristic in how it combines both frequency of (co-)occurrence and association in one simple, sortable dimension (see Gries, 2015), which might be sufficient for many applied purposes. At the same time, the conflation of these dimensions also loses valuable information, which is why Gries (2019) suggested to have collostructional methods not return a single value per word merging many dimensions (unrecognizably) but a tuple of orthogonal measures, in particular (a) frequency, (b) association measures that do not include frequency, and (c) textual dispersion of co-occurrences. This allows the analyst to tease apart

what much work has conflated, which is necessary for, say, psycholinguistic analyses interested in frequency (entrenchment), but also contingency (association), and exposure/priming (dispersion).

Another important issue is how most association studies in corpus linguistics do not consider the variability of their results, which threatens to undermine any studies that involve ranking the top associations. Gries (2022) proposed to add bootstrapped uncertainty ellipses to collostructional studies.

More linguistically, one of the most important improvements was discussed by Bernolet and Colleman (2016), namely, the need to not just explore correlations between constructions and words but between constructions and word-sense pairings. Since different senses of words can be attracted to different constructions (differently much), the purely form-based tradition of much of collostructional analysis can miss important details/patterns; recent developments in distributional semantics might ultimately help addressing this issue in a more manageable nonmanual way (see Perek and Hilpert, 2017).

In sum, while the basic logic of collostructional methods has stood the test of time, the method, like every other one, would benefit from continuous refinement to make its findings more precise, robust, and instructive.

SEE ALSO: Testing Independent Relationships

References

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

Ellis, N. C., Römer, U., & Brook O'Donnell, M. (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)). John Wiley.

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

Flach, S. (2020b). Reduction hypothesis revisited: Frequency or association? In C. Sanchez-Stockhammer, F. Günther, & H.-J. Schmid (Eds.), *Language in mind and brain* (pp. 16–22). LMU Open Access.

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

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

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., Hampe, B., & Schönefeld, D. (2010). Converging evidence II: More on the association of verbs and constructions. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 59–72). CSLI.

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

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

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*, 164–187.

Harris, Zelig S. (1970). *Papers in Structural and Transformational Linguistics*. Reidel.

Hilpert, M. (2008). *Germanic future constructions: A usage-based approach to language change*. John Benjamins.

Perek, F., & Hilpert, M. (2017). A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics*, *22*(4), 490–520.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Wulff, S., Stefanowitsch, A., & Gries, S. T. (2007). Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In G. Radden, K.-M. Köpcke, T. Berg, & P. Siemund (Eds.), *Aspects of meaning construction* (pp. 265–281). John Benjamins.

Suggested Readings

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction* (2nd ed.). Routledge, Taylor & Francis Group.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.