Corpus Linguistics: Quantitative Methods

STEFAN TH. GRIES

Introduction

Ever since technological development has made it possible to search large corpora in a very short time, corpus linguists have done a lot of interesting work in linguistics in general, and in applied linguistics in particular. Given both a large interest of corpus linguists in lexicographic applications and the fact that words are among the linguistic elements most easily recoverable (in the usual suspects of well-researched Indo-European languages at least), most corpus-linguistic work until now has been concerned with words and/or *n*-grams (i.e., sequences of words), their distribution with regard to other words, and their distributions across different modes, registers, genres, varieties, and so forth.

More recently, however, the situation has changed and corpus-linguistic research has begun to address many more syntactic phenomena. While this is to some extent due to the increased availability of syntactically annotated corpora, it is also due to corpus linguists' and many cognitive linguists' adoption of the assumption that syntax and lexis are not qualitatively different (see Hunston & Francis, 2000, or Hoey, 2005, in corpus linguistics and Langacker, 2000, or Goldberg, 1995, 2006, in cognitive linguistics). Only recently, however, have words and syntactic patterns, or constructions, been treated on a par not only *theoretically*, but also *empirically*. One example is the application of association measures that are usually applied to co-occurrences of words (aka collocations) to the co-occurrences of words with syntactic patterns. This approach is referred to as *collostructional analysis* (a blend of *collocation* and *construction*), and three different kinds of applications have been proposed:

- **collexeme analysis**, which quantifies the degree of attraction or repulsion of words (typically verbs) to a syntactically defined slot in a construction (see Stefanowitsch & Gries, 2003), for example: how much does *give* like to occur in the ditransitive?
- **distinctive collexeme analysis**, which quantifies which words (typically verbs) are attracted to or repelled by one of several constructions (see Gries & Stefanowitsch, 2004a), for example: how much does *give* prefer to occur in the ditransitive as opposed to the prepositional dative?
- **covarying collexeme analysis**, which identifies preferred and dispreferred pairs in two slots of one construction (see Gries & Stefanowitsch, 2004b), for example: the two verb slots in *He tricked her into marrying him*.

These methods have been applied in a variety of domains and languages including constructional senses and complementation patterns, syntactic alternations of a variety of constructions, verb-specific syntactic priming effects, and so forth. In this article, applications of distinctive collexeme analysis to data from second-language learners of English will be discussed briefly.

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

DOI: 10.1002/9781405198431.wbeal0258

Table 1 Frequencies of *give* in ditransitive and prepositional datives in the ICE-GB (from Gries and Stefanowitsch, 2004a, p. 102)

	Ditransitive	*Prepositional dative*	*Total*
give	461	146	607
Other verbs	574	1773	2,347
Total	1,035	1,919	2,954

Distinctive Collexeme Analysis

Like nearly all corpus-linguistic association measures, distinctive collexeme analysis is based on a two-by-two co-occurrence table such as Table 1, which exemplifies how the lemma *give* is distributed across ditransitive and prepositional datives in the British component of the International Corpus of English (ICE-GB).

In collostructional analysis, the association measure used most frequently to evaluate such tables is the negative log to the base of 10 of the $p_{\text{one-tailed}}$-value of a *Fisher–Yates exact test*. Using the open-source programming language and environment R (see R Development Core Team, 2010, available from http://cran.at.r-project.org/), this measure can be computed easily as follows (when the observed frequency in the upper-left cell is larger than the one expected by chance, i.e., 607 · 1035 / 2954 ≈ 213):

```
607*1035/2954 # expected frequency¶
[1] 212.6760
– log10(sum(dhyper(461:607, 1035, 1919, 607))) # – log10 p-value¶
[1] 119.7361
```

Line 3 of the above code computes the negative log to the base of 10 (– log10) of the sum (sum) of all probabilities from the hypergeometric distribution (dhyper) from the observed frequency of 461 to the theoretically possible extreme of 607, given that the data contain 1,035 ditransitives, 1,919 prepositional datives, and 607 instances of *give*. (Such computations can be performed automatically with a script available from http://tinyurl.com/collostructions.)

If the observed frequency of the cell of interest is less than the expected one (as it is here for the occurrence of *give* in the prepositional dative), this formula changes to the following, which computes the negative log to the base of 10 (– log10) of the sum (sum) of all probabilities from the hypergeometric distribution (dhyper) from the observed frequency of 146 to the theoretically possible extreme of 0, given that the data contain 1,919 prepositional datives, 1,035 ditransitives, and 607 instances of *give*:

```
– log10(sum(dhyper(0:146, 1919, 1035, 607))) # expected frequency¶
[1] 119.7361
```

Analogous tests can be done for all verb or lemma types occurring at least once in either the ditransitive or the prepositional dative, and then these verb lemmas can be ranked according to the strength of their attraction or repulsion to the two constructions (an interactive R script for this offering different measures of association strength is available from the author). The verbs that are most strongly attracted to the ditransitive and the prepositional dative are listed in (1) and (2) respectively (listed in decreasing strength of association strength).

1. *give*, *tell*, *show*, *offer*, *cost*, *teach*, *wish*, *ask*, *promise*, *deny*, *award*, *grant*, *cause*, *drop* . . .
2. *bring*, *play*, *take*, *pass*, *make*, *sell*, *do*, *supply*, *read*, *hand*, *feed*, *leave*, *keep*, *pay* . . .

Such results are interesting because they provide strong support for analyses of the two constructions that invoke different constructional senses. For example, the ditransitive has been argued to involve constructional senses of transfer, enablement of transfer, non-enablement of transfer, communication as transfer, and others. In addition, they are also compatible with what is known about the two constructions' acquisition patterns (where, for example, *give* is a path-breaking verb for the acquisition of the ditransitive).

While many analyses of this kind were targeted at argument-structure constructions, other less semantically loaded constructions have exhibited similar verb-specific effects; examples include *will*-future versus *going to* V (see [3]), particle placement (see [4]), or *to* versus *ing*-complementation (see [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.

This collostructional approach has returned interesting and new results regarding many of the above constructions and others, and there is even experimental evidence from sentence-completion and self-paced reading tasks that indicates that the behavior of native speakers of English can sometimes be predicted better on the basis of association strengths than on the basis of raw frequencies or conditional probabilities (see Gries, Hampe, & Schönefeld, 2005, 2010).

Applications

The above kind of corpus-based measurement of association strengths has many interesting implications and applications. For example, there is an increasing body of evidence that shows that children and adults are very sensitive to distributional patterns in language: infants less than a year old can notice statistical co-occurrence patterns in their ambient language; language change is strongly correlated with the frequencies of words and syntactic patterns; and linguistic representation and processing exhibit frequency and conditional-probability effects. Therefore, the computation of probabilistic associations between different linguistic elements can inform many aspects of theoretical linguistics, but also applied linguistics. The following two sections discuss how such corpus-based methods can also be correlated with experimental data and show, here for second and foreign-language learners, how the corpus-based association strengths help to reliably predict second-language learners' experimental priming responses.

Ditransitive Versus Prepositional Datives

Gries and Wulff (2005) performed a sentence-completion task in which the results of Gries and Stefanowitsch's (2004a) distinctive collexeme analysis, parts of which were listed in (1) and (2), were correlated with the results of a sentence-completion priming experiment with German learners of English (mean number of years of English instruction: 11.1 years). In that experiment, the subjects were presented with sentence fragments of two kinds in an alternating fashion: sentence fragments that suggested a particular completion (as in [6]), followed by sentence fragments that did not (as in [7]).

6. a. The racing driver showed the helpful mechanic . . . [suggests a ditransitive]
   b. The racing driver showed the torn overall . . . [suggests a prepositional dative]
7. The racing driver showed . . . [does not suggest a specific constructional completion]

The question was whether subjects' completion of a fragment of the type in (6) would prime them to complete the fragment of the type in (7) with the same construction, and the learner subjects did exhibit such a significant priming effect. More interestingly in the present connection, however, is the fact that the subjects exhibited different priming effects for different verbs: the subjects were significantly more likely to be primed for ditransitives when the sentence fragment ended in a verb that the distinctive collexeme analysis of the native English speaker identified as preferring the ditransitive, and vice versa. Even more interestingly, Gries and Wulff also showed that this significant correlation between native-speaker corpus preferences and learner experimental preferences cannot be reduced to the English verbs' translational equivalents in German.

Similar evidence was obtained by Wulff and Gries (in press) on the basis of (German and Dutch) learner corpus data from the International Corpus of English (ICLE; Granger, 1993). They found a highly significant correlation between native-speaker corpus preferences and learner corpus preferences.

In sum, for the alternation of ditransitives and prepositional datives, different studies using the collostructional approach revealed that the two constructions exhibit markedly different preferences for different verbs, which in turn correlate with cognitive-linguistic accounts of the two constructions and their sense extensions, and these preferences are robust across native speakers and learners, and across experimental and observational data.

to-Versus *ing*-Complementation

In a similar set of case studies, Gries and Wulff (2009) studied the two complementation patterns exemplified in (5). They first conducted a distinctive collexeme analysis of the two constructions in native-speaker corpus data to identify which verbs they prefer. They found that the *to*-construction and the *ing*-construction preferred the verbs listed in (8) and (9) respectively (listed in decreasing strength of association strength).

8. *try*, *wish*, *manage*, *seek*, *tend*, *intend*, *attempt*, *hope*, *fail*, *like*, *refuse*, *learn*, *plan* . . .
9. *keep*, *start*, *stop*, *avoid*, *end*, *enjoy*, *mind*, *remember*, *go*, *consider*, *envisage*, *finish* . . .

Again, many of the claims about the semantic differences between the two constructions are confirmed. For one, the verbs most distinctively associated with the infinitival construction, *try* and *wish*, both denote potentiality, while the verbs most distinctive for the gerundial construction, *keep*, *start*, and *stop*, denote actual events. Along similar lines, many of the collexemes distinctive for the infinitival construction are future-oriented (*intend*, *hope*, *learn*, and *aim* are just a few examples), while the distinctive collexemes of the gerundial construction evoke an interpretation in relation to the time of the utterance (*avoid*, *end*, *imagine*, *hate*, etc.).

As before, the question arises as to what extent learners are aware of these statistical tendencies, especially since these two patterns provide few other clues such as, for instance, the order of semantic roles they involve. Gries and Wulff therefore performed a similar sentence-completion experiment involving priming with German learners of English (mean number of years of English instruction: 11 years). (This study included several additional factors that are of no concern here.) In a logistic regression involving priming and verbs' attraction to both constructions, Gries and Wulff found that the collostructional preference of the verb in the target fragment was by far the strongest predictor of the learners' sentence completions. Also, Wulff and Gries (in press) show that the same native-speaker collostructional preferences are also highly significantly correlated with learners' preferences obtained from the German part of the ICLE.

As with the ditransitives and prepositional datives, different kinds of evidence support the collostructional approach and its implications: native speakers and learners exhibit very similar preferential patterns of construction use.

Conclusion

This article has discussed several different case studies—involving different experiments and different corpus data—all of which yield converging evidence in support of a quantitative corpus-linguistic method to explore the syntax–lexis interface, the collostructional approach. This approach yields replicable quantitative data for the general description of constructions' distributional characteristics and/or verb subcategorization preferences as well as other processing-related accounts of acquisition, learning, and priming. However, another feature of this approach that is just as attractive is that it is compatible with much recent work in usage-based cognitive linguistics and psycholinguistics that adopts an exemplar-based perspective, in which learning is based on the memorization of, and probabilistic abstraction from, thousands of exemplars. Such collostructional studies are therefore more than just a convenient quantification of co-occurrence phenomena: they also provide a motivated way for relating empirical results and contemporary linguistic and psycholinguistic theorizing.

SEE ALSO: Testing Independent Relationships

References

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

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

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 57–69). Amsterdam, Netherlands: Rodopi.

Gries, St. Th., Hampe, B., & Schönefeld, D. (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics, 16*(4), 635–76.

Gries, St. Th., Hampe, B., & Schönefeld, D. (2010). Converging evidence II: more on the association of verbs and constructions. In J. Newman & S. Rice (Eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language* (pp. 73–90). Stanford, CA: CSLI.

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

Gries, St. Th., & Stefanowitsch, A. (2004b). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–36). Stanford, CA: CSLI.

Gries, St. Th., & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics, 3*, 182–200.

Gries, St. Th., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics, 7*, 164–87.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London, England: Routledge.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Philadelphia, PA: John Benjamins.

Langacker, R. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 1–63). Stanford, CA: CSLI.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org

Stefanowitsch, A., & Gries, St. Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–43.

Wulff, S., & Gries, St. Th. (in press). Corpus-driven methods for assessing accuracy in learner production. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. Philadelphia, PA: John Benjamins.

Suggested Readings

Gries, St. Th. (2009a). *Quantitative corpus linguistics with R: A practical introduction*. London, England: Routledge.

Gries, St. Th. (2009b). *Statistics for linguistics with R: A practical introduction*. Berlin, Germany: De Gruyter.

Gries, St. Th. (in press). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *New horizons in corpus linguistics* (pp. 269–91). Frankfurt am Main, Germany: Peter Lang.