This is a contribution from *Corpus-based Approaches to Construction Grammar*.
Edited by Jiyoung Yoon and Stefan Th. Gries.
© 2016. John Benjamins Publishing Company

CHAPTER 1

Corpus-based approaches to Construction Grammar

Introduction*

Jiyoung Yoon and Stefan Th. Gries
University of North Texas / University of California, Santa Barbara

For a long period of time, generative approaches to grammar dominated the field of theoretical linguistics, and that theoretical dominance was coupled with a similar dominance of the 'method' of judgments of acceptability of (typically) decontextualized sentences to 'determine' whether a particular sentence was formed in accordance with the postulated rules of the grammar. However, during the 1980s, the field of theoretical linguistics began to change with the advent of cognitive / usage-based linguistics, and the concomitant cognitive commitment towards "providing a characterization of general principles for language that accords with what is known about the mind and brain from other disciplines" (Lakoff, 1990: 40). While early work in cognitive linguistics and Construction Grammar was characterized by methods quite similar to those of the generative approach, a first difference consisted in the fact that even some of the earliest Construction Grammar studies were already more based on observational data (even if those were often collected somewhat eclectically). Over time, the increase of usage-based linguistics on the one hand and discussions of the limits of cognitive-linguistic theorizing on the other hand led to a slow but steady increase of the (range of) methods that are being employed: Usage-based linguistics in general and Construction Grammar in particular are now brimming with experimental and observational studies, and the number of studies that also use statistical methods has been increasing to the point that there are now publications discussing the quantitative turn in cognitive linguistics (Janda, 2013).

* Part of this research has been supported by the University of North Texas Research and Creativity Enhancement (RCE) Grant and by the Spanish Ministry of Economy and Competitiveness, grant no. FFI2013-43593-P. We would like to thank anonymous reviewers and the series editors of Constructional Approaches to Language, Jan-Ola Östman and Kyoko Ohara, for their invaluable feedback on the earlier versions of this edited volume. We would also like to thank Francisco Ruiz de Mendoza who helped us at the initial stage of this project which was first planned for the 44th SLE conference in Rioja, Spain.

doi 10.1075/cal.19.01yoo

The methodological approach that has been growing most quickly in cognitive linguistics and Construction Grammar involves, arguably, the use of corpus data. The notion of *corpus* is a prototype category and the prototype of a corpus is a collection of files that contain text and/or transcribed speech that is supposed to be representative and balanced for a certain language, variety, register, or dialect; often, the files contain not just the language that has been written or spoken (often in UTF-8 encoding to cover different orthographies), but also extra annotation (often part-of-speech information, but also morphological, semantic, or other information in the form of XML annotation). Corpora differ most importantly in size (from a few narratives narrated in an underdocumented or even already extinct language to corpora of many billions of words) and in the degree of naturalness of the data they contain (from completely spontaneous dialog between two speakers to experimental highly-constrained situations).

In a recent survey of different kinds of data in Construction Grammar, Gries (2013) discusses four different kinds of quantitative uses of corpus data, which can be grouped into three different categories (in ascending order of statistical complexity):

- absolute frequencies and (conditional probabilities) of occurrence of constructions; this category would include as a limiting case studies based on the observation whether something is attested (and how) or not, but can also include cases where frequencies or probabilities are used to rank-order words in constructions etc.
- measures of association strength that quantify the degree of attraction or repulsion of two kinds of linguistic construction: if the two linguistic constructions are both words, corpus linguists have referred to these co-occurrences as *collocations*; if the two linguistic constructions are words and some other kind of construction (often, argument structure constructions or other constructions with lexically-unspecified slots), they are usually referred as *colligations* or *collostructions*;
- detailed co-occurrence data based on the annotation of many aspects of constructional uses which are then analyzed statistically using multifactorial or multivariate methods (such as regression methods/classifiers or exploratory tools such as cluster analysis/multidimensional scaling and others).

While the three kinds of quantitative uses of corpus data can be situated on a cline of statistical complexity, this does not mean that all studies should always be aiming for the highest level of complexity in the analysis: while more detailed analyses can be, all else being equal, potentially more revealing, different linguistic questions require different levels of analytical granularity (see Arppe et al., 2010). This is nicely exemplified by the studies in the present volume, which exemplify insightfully all three levels of statistical resolution in how they tackle different kinds of constructions. However, the present volume also provides another, from our point of view, welcome differentiation from much existing work: Just like in linguistics or cognitive linguistics as a whole, corpus linguistics has for a long time been dominated by studies of synchronic native-speaker English, a tendency which has been reinforced by the widespread availability of many English-language corpora and the much more slowly growing availability of both general and more specific corpora in other languages. None of the case studies in this volume is on synchronic native-speaker English – rather, they study various aspects in Dutch, Spanish, Italian (both synchronic native speaker data and L1-acquisition data), and in diachronic (Old) English on the basis of a much wider range of corpora than are typically found. In what follows, we provide a brief overview of the contributions in this volume.

The first part of this volume comprises three studies which are examples of the first kind of quantitative corpus method, namely Beliën's study of Dutch post-positions or particles, Quochi's analysis of light-*fare* ('do') verb constructions in Italian, and Vázquez Rozas & Miglio's study of subject and object experiencers in Spanish and Italian. Specifically, the main question in the study contributed by Beliën is the long-standing constituency issue of the Dutch particle constructions. Beliën points out that traditional syntactic constituency tests such as topicalization, passivization, and pronominalization did not satisfactorily provide an answer about the constituency of the Dutch particle construction. In her contribution, she employs a cognitive-grammar analysis based on Langacker (1997) in order to determine 'conceptual' constituency which is one type of constituency distinguished in a cognitive-grammar theory. A semantic analysis of the construction under study based on actually attested, rather than invented, examples suggests that Dutch particle constructions are analyzed similar to (transitive) separable complex verb constructions rather than (intransitive) preposition constructions.

Quochi's article explores the development and representation of light-*fare* verb constructions in Italian. On the basis of an analysis of language acquisition data from the CHILDES database (i.e., the Italian collection of longitudinal transcriptions of interactive sessions with eleven Italian-speaking children), she studies type-token ratios in children's and adult data as well as relative frequencies of co-occurrence data and proposes that light verb constructions in Italian can be viewed as a family of constructions or a radial category (Goldberg, 1995), which includes the central construction labelled as the Perform Intransitive Action Construction (e.g., *fare una passeggiata* 'take a walk'), the Perform/Emit Sound Construction (e.g., *fare chiasso* 'make noise'), the Perform Transitive Action Construction (e.g., *fare un colpo* 'hit'), and the Cause Emotion Construction (e.g., *fare rabbia* '(lit) do anger to someone'; 'make someone angry'). This family of constructions accounts for both the specificity of each construction and its proximity to the more general transitive construction. The findings of the study suggest that a light-*fare* 'do' (pivot) schema (that accounts for both conventional expressions and for new productive formations) may really exist, and may be taken to support the idea that light verbs act as facilitators in the learning of (argument structure) constructions (Ninio, 1999).

In the final chapter of this first part, Vázquez Rozas and Miglio provide a comparative study on constructions with subject versus object experiencers in Spanish and Italian. This study explores 'Experiencer-as-Subject' constructions (ESC) and 'Experiencer-as-Object' constructions (EOC) in Spanish and Italian using the ARTHUS corpus and the BDS/ADESSE database for Spanish, and BADIP, C-ORAL, and *La Repubblica* as corpora for Italian. In both languages, verbs that denote feeling or emotion involve two participants – an experiencer and a stimulus – but the puzzling fact is that some of these clauses construe the experiencer as a subject and the stimulus as object (e.g., *Amo esta ciudad* 'I love this city'), while others have experiencers coded as dative or accusative objects and stimuli as subjects (e.g., *Me gusta la música* 'Me-dative likes the music [I like music]'). In order to gain insight into how both constructions are used by speakers, the authors analyze the relative frequencies and distributions of a number of discourse-related properties of the arguments, such as animacy, person, and syntactic category, as well as textual genres. The results indicate that the distribution of the discourse-related properties of the arguments is not random when comparing the ESCs with the EOCs in the verbs of feeling and emotion: for instance, the use of the 1st person is more frequently found in EOCs than in ESCs, and EOCs are associated more with oral discourse than with written discourse.

The second part of this volume contains studies that are concerned with the co-occurrence of words and constructions on the basis of the family of methods that has come to be known as collostructional analysis (see Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004; Gries, 2015); this method quantifies the degree to which words and constructions are mutually attracted to, or repelled by, each other and what such attractions/repulsions reveal about the functional characteristics of constructions.

In the first chapter of this part, the fourth overall, on Spanish constructions of directed motion, Pedersen provides a language-specific view of Construction Grammar. Pedersen analyzes Spanish telic motion constructions with the constructional environment [V *a* 'to' NP] (e.g., **caminar a la biblioteca* 'to walk to the library') in order to revisit the Talmian typological distinction between satellite-framed languages (in which the verb encodes the manner as in English) and verb-framed languages (in which the verb encodes the path as in Spanish). To this end, the author applies collexeme analyses (see Stefanowitsch & Gries, 2003) to data extracted from the Corpus del Español, which confirm the basic encoding pattern of the Talmian typology: the verbal encoding of the path component, with the verb meaning 'path of motion leading to an end point'. Pedersen proposes that, from a Construction Grammar point of view, the constraining role of the verb is essential in Spanish while the role of the schematic construction is not as predominant as in Germanic languages such as English. In other words, the encoding of the Spanish argument structure is basically verb-driven (as opposed to construction-driven), but he cautions that 'verb-driven' is not the same as categorizing Spanish as a verb-framed language as defined in the Talmian tradition.

The second chapter of this part tackles the alternation of a complementation pattern, in this case the alternation between infinitival and sentential complement constructions in Spanish. In their contribution on infinitival and sentential complement constructions in Spanish, Yoon & Wulff analyze 561 instances of infinitival complements and 795 instances of sentential complements retrieved from a corpus of journalistic prose. Through the application of a distinctive collexeme analysis (Gries & Stefanowitsch, 2004), the authors identify the verbs most distinctively associated with either type of complementation. The results indicate that the two complementation patterns are in fact distinct constructions in the constructionist sense of the term (Goldberg, 1995, 2006): the infinitival complementation construction attracts verbs that denote 'desire' (e.g., *querer* 'want,' *intentar* 'try,' *preferir* 'prefer') whereas the sentential complementation construction is distinctively associated with verbs of 'communication' (e.g., *decir* 'say,' *explicar* 'explain,' *anunciar* 'announce') and 'mental activity' (e.g., *creer* 'believe,' *recordar* 'remember,' *reconocer* 'recognize'). At the same time, Yoon & Wulff stress the importance of the usage-based constructionist approaches in the sense that verbs do not fall into two mutually exclusive classes with each class licensing either type of complementation only, but are rather distributed probabilistically.

The final chapter of this second part is Bernolet & Colleman's contribution on sense-based and lexeme-based alternation biases in the Dutch dative alternation. This study raises the issue of whether the subcategorization probabilities of Dutch verbs partaking in the dative alternation are biased by the verbal lexeme or by the verb senses. In order to answer this question, the authors run a sense-based distinctive collexeme analysis (Gries & Stefanowitsch, 2004) on corpus data supplemented by a syntactic priming experiment. A total of 15 polysemous ditransitive verbs with two senses (i.e., sense 1: 'concrete', sense 2: 'figurative') were selected and analyzed for their association strengths with the double object (DO) construction and the prepositional dative (PD) construction with *aan*. The authors find that the distinct senses of the same verb display markedly different alternation biases toward either DO or PD constructions, showing that sense-based data in a collostructional analysis, and also in other kinds of analyses, provide a more precise picture of the Dutch dative alternation than the standard lexeme-based analysis. The additional psycholinguistic experiment involving the participation of twenty-five native speakers of Dutch, on the other hand, shows no effect of the lexeme-based or sense-based biases of prime verbs, but the results still support the position that language users are sensitive to sense-based verb biases and that they store such information in memory.

The final part of this volume contains two studies that involve very detailed case-by-case annotation of concordance results and, consequently, more advanced statistical methods. In the first chapter of this part, Shank, Plevoets, & Van Bogaert provide a multifactorial analysis of *that*/zero alternation and discuss the diachronic development of the zero complementizer with *think*, *guess* and *understand*. This study uses stepwise logistic regression analysis in order to evaluate the effects of eleven structural features such as length of the complement clause subject, presence versus absence of additional material in the matrix clause, matrix clause tense, etc. on complementizer realization with three verbs of cognition: *think*, *guess*, and *understand*. After analyzing a total of nearly 19,000 tokens from both spoken and written corpora from 1560–2012, the authors challenge the long-standing assumption of a diachronic trend towards a preference of the zero complementizer. Their finding indicates that *guess* is the only verb exhibiting such a diachronic increase. At the same time, the authors suggest that among many other factors, the lack of matrix internal elements and also the written or spoken mode are good conditioning factors for the presence or the absence of complementizers.

Last but not least, Levshina's contribution investigates a geometric exemplar-based model of semantic structure in her analysis of the Dutch causative construction with *laten* 'let'. Her innovative approach questions the commonly assumed notion of 'prototypical senses' of a construction in Construction Grammar, and presents a corpus-based bottom-up approach that can be used to model semantic structures. A sample of 731 occurrences of the causative *laten* randomly selected from the Corpus of Spoken Dutch as well as newspaper register is analyzed by visualizing semantic similarities between the exemplars of a construction in a semantic map computed using Multidimensional Scaling. This semantic map makes it possible to see the main semantic dimensions and senses of the Dutch causative construction with the auxiliary *laten*: In the map, the more features two exemplars share, the smaller the distance between them. The result suggests that the constructional semantics is organized as a doughnut, with an empty center and extensive periphery, which means that there is not necessarily a central sense, or prototype. Levshina concludes that the exemplars of *laten* are related in a family-resemblance fashion, with the main senses not being discrete, but representing a continuum.

"These are exciting times to be a …" is a construction and a cliché, but here it is true. Cognitive linguistics is following the same trend that has been visible in linguistics at large: an evolution towards a more rigorously empirical and quantitative discipline, and a discipline that looks more and more outside of synchronic and L1 English. The articles in this volume, which reflect these trends, also leave a mark on Construction Grammar, with data from other languages, from a wide variety of corpora, and with very different quantitative approaches (for very different questions). While the evolution of (cognitive) linguistics in general and Construction Grammar in particular is certainly not coming to an end, the increased diversity of languages studied, questions explored, and methods/techniques used is a promising sign that Construction Grammar is maturing – may its journey/evolution continue along those lines …

References

Arppe, Antti, Gilquin, Gaëtanelle, Glynn, Dylan, Hilpert, Martin & Zeschel, Arne. (2010). Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora*, 5(1), 1–27. doi:10.3366/cor.2010.0001

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

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

Gries, Stefan Th. (2013). Data in construction grammar. In Graham Trousdale & Thomas Hoffmann (Eds.), *The Oxford handbook of construction grammar* (pp. 93–108). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780195396683.013.0006

Gries, Stefan Th. (2015). More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff (2013). *Cognitive Linguistics*, 26(3), 505–536. doi:10.1515/cog-2014-0092

Gries, Stefan Th. & Stefanowitsch, Anatol. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9, 97–129. doi:10.1075/ijcl.9.1.06gri

Janda, Laura A. (Ed.). (2013). *Cognitive linguistics – The quantitative turn*. Berlin & Boston: De Gruyter Mouton. doi:10.1515/9783110335255

Langacker, Ronald W. (1997). Constituency, dependency, and conceptual grouping. *Cognitive Linguistics*, 8(1), 1–32. doi:10.1515/cogl.1997.8.1.1

Lakoff, George. (1990). The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics*, 1(1), 39–74. doi:10.1515/cogl.1990.1.1.39

Ninio, Anat. (1999). Pathbreaking Verbs in Syntactic Development and the Question of Prototypical Transitivity. *Journal of Child Language* 26, 619–653.

Stefanowitsch, Anatol & Gries, Stefan Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243. doi:10.1075/ijcl.8.2.03ste