Quantitative Methods in the Humanities and Social Sciences

Pascual Cantos-Gómez
Moisés Almela-Sánchez *Editors*

Lexical Collocation Analysis

Advances and Applications

Springer

Quantitative Methods in the Humanities and Social Sciences

Quantitative Methods in the Humanities and Social Sciences is a book series designed to foster research-based conversation with all parts of the university campus – from buildings of ivy-covered stone to technologically savvy walls of glass. Scholarship from international researchers and the esteemed editorial board represents the far-reaching applications of computational analysis, statistical models, computer-based programs, and other quantitative methods. Methods are integrated in a dialogue that is sensitive to the broader context of humanistic study and social science research. Scholars, including among others historians, archaeologists, new media specialists, classicists and linguists, promote this interdisciplinary approach. These texts teach new methodological approaches for contemporary research. Each volume exposes readers to a particular research method. Researchers and students then benefit from exposure to subtleties of the larger project or corpus of work in which the quantitative methods come to fruition.

More information about this series at http://www.springer.com/series/11748

Pascual Cantos-Gómez • Moisés Almela-Sánchez
Editors

Lexical Collocation Analysis

Advances and Applications

Springer

Editors

Pascual Cantos-Gómez
Department of English
University of Murcia
Murcia, Spain

Moisés Almela-Sánchez
Department of English
University of Murcia
Murcia, Spain

ISSN 2199-0956 ISSN 2199-0964 (electronic)
Quantitative Methods in the Humanities and Social Sciences
ISBN 978-3-319-92581-3 ISBN 978-3-319-92582-0 (eBook)
https://doi.org/10.1007/978-3-319-92582-0

Library of Congress Control Number: 2018950661

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introduction

Borderline phenomena are a fertile ground for scientific inquiry. They stimulate theoretical controversy and open up new opportunities for exploring innovative methodologies. The concept of collocation is illustrative of these possibilities. The special character of collocation, particularly its intermediate position between lexical and grammatical patterning, has favored an integration of perspectives of analysis that in previous stages of linguistics had belonged to separate areas of study. This integration of perspectives is proving fruitful. Six decades after the concept of collocation was introduced – it is attributed to the writings of J. Firth published in the 1950s – the range of topics explored in the literature on collocation and the sophistication of the methods proposed in this field are still far from being exhausted.

Collocational studies are, we dare say, one of the most productive areas of research over the last five decades, judging by the abundance of literature dealing with the topic and by the multiplicity of theoretical insights, methodological frameworks, and practical applications that have resulted from this field of research. The results obtained from collocational research have played a central role in the *lexicalist turn* of the last decades and in the reformulation of the boundaries between vocabulary and grammar. Concepts such as the Sinclairian *idiom principle* or Hoey's *lexical priming* are good epitomes of this tendency. So is the integration of corpus collocation studies and construction grammar, famously initiated by Gries and Stefanowitsch. The fruitfulness of collocational research is further illustrated by the diversity and the effectiveness of practical applications derived from advances in this field. Applied collocational research has produced promising results in various disciplines, including lexicography, second language teaching/learning, and computational linguistics, among others.

It is today beyond question that one of the key factors in the boosting of collocational research has been the incorporation of the new technologies into the tools of linguistic description. As Sinclair envisioned four decades ago, the use of computers and electronic corpora has facilitated the creation of ever more powerful methods of description that, in turn, have made it possible to lay bare forms of lexico-grammatical organization that had remained unnoticed to the unaided observer. This volume lays special emphasis on the coupling of collocational research and computational corpus tools. The common denominator of the papers presented here is the use of computational corpora and quantitative techniques as a means to explore aspects of language patterning that overlap the boundaries between lexis and grammar.

The book opens with a proposal for integrating both collocational and valency phenomena within the overarching theoretical framework of construction grammar. This first chapter, by Thomas Herbst, combines insights from Bybee's usage-based approach to language, from Goldberg's construction grammar, and from Gries and Stefanowitsch's collostructional analysis as a way to account for properties of both collocational patterns and valency patterns.

In Chap. 2, Violeta Seretan makes the case for integrating advances in syntactic parsing and in collocational analysis. After observing that parsing technologies and collocational research have often followed separate paths, Seretan contends that these two areas would benefit mutually from a joint approach to syntactic analysis and to collocation extraction.

Chapter 3 submits an interesting and innovative proposal for complementing corpus data and dictionaries in the identification of specific types of collocations consisting of restricted predicate-argument combinations (*collocates* and *bases*, in Hausmann's terminology). The chapter is authored by Isabel Sánchez-Berriel, Octavio Santana Suárez, Virginia Gutiérrez Rodríguez, and José Pérez Aguiar. As the authors explain, association measures face serious limitations as methods for extracting this type of collocations, which are structurally and semantically more restricted than the Sinclairian node-collocate pair. The strategy proposed by the authors of this chapter for solving this problem is to complement corpus collocational data with network analysis techniques applied to dictionary entries.

In Chap. 4, Vaclav Brezina explains the potential of collocational graphs and networks both as a visualization tool and as an analytical technique. Brezina provides three case studies showing the use of this technique in several areas of descriptive and applied linguistics, particularly in discourse analysis, language learning research, and lexicography.

In Chap. 5, Alexander Wahl and Stefan Gries propose a new, data-driven approach to the identification and extraction of multi-word expressions from corpora. The approach, termed by the acronym MERGE (Multi-word Expressions from the Recursive Grouping of Elements), is based on the selection of bigrams using log-likelihood and their successive combination into larger sequences. The results are validated via human ratings.

Finally, in Chap. 6, Peter Uhrig, Stefan Evert, and Thomas Proisl undertake a thorough analysis and evaluation of factors influencing the performance of collocation extraction methods in parsed corpora. The authors compare the impact of several factors, including parsing scheme, association measure, frequency threshold, type of corpus, and type of collocation. The results of this profound study offer valuable criteria for methodological decisions on collocation extraction.

We would like to conclude this introduction by expressing our gratitude to all the contributors to this volume for having joined us in this project and for helping to make it a reality. A word of gratitude goes also to the referees who have kindly agreed to assist us in the review process, supplying valuable feedback and advice to the authors.

Thanks are also due to Springer's staff Matthew Amboy, Editor Operations Research, for believing in this project and for his assistance and support throughout the preparation of this book, and to Faith Su, Assistant Editor, for her guidance during the production of this volume.

We are confident that this collection can contribute to the development of collocation analysis by providing an interesting illustration of the current trends in this field of research.

Universidad de Murcia, Murcia, Spain
Moisés Almela
Pascual Cantos

Contents

Chapter 1
Is Language a Collostructicon? A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions

Thomas Herbst

Abstract This chapter argues in favour of not regarding collocation and valency as strictly discrete categories but rather seeing them as near neighbours in the lexis-grammar continuum. Following Bybee's (Usage-based theory and exemplar representation of constructions. In Hoffmann T, Trousdale G (eds) The Oxford handbook of construction grammar. Oxford University Press, Oxford, pp. 49–69, 2013) analysis of the *drive me crazy* construction, a suggestion will be made for presenting both collocational and valency phenomena in terms of constructions. It will be argued that the constructicon representing speakers' linguistic knowledge contains both item-specific information and generalized information in the form of Goldbergian argument structure constructions (Goldberg 2016) and in particular that the description of valency slots should provide exemplar representations based on the principles of collostructional analysis as developed by Stefanowitsch and Gries (Inter J Coprus Lingusitics 8:209–243, 2003).

1 Why We Know So Much More About Language

1.1 Exciting Times for Linguists

We live in exciting times for linguists. After being dominated by one particular line of thinking for decades with other approaches leading a rather peripheral (!) existence, at least in theoretical linguistics, we now seem to have reached a point where linguists of many different fields who for some reason or other had not been persuaded by the generative enterprise appear to be agreeing on at least a

T. Herbst (✉)
English Linguistics and Interdisciplinary Centre for Research on Lexicography, Valency and Collocation, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
e-mail: thomas.herbst@fau.de

P. Cantos-Gómez, M. Almela-Sánchez (eds.), *Lexical Collocation Analysis*,
Quantitative Methods in the Humanities and Social Sciences,
https://doi.org/10.1007/978-3-319-92582-0_1

rough outline of a different framework, which brings together scholars working in cognitive linguistics, corpus linguistics, foreign language acquisition (including lexicography) and also historical linguistics—sailing under labels such as the usage-based approach or construction grammar (Langacker 1987, 2008; Sinclair 2004; Goldberg 1995, 2006; Dąbrowska 2015; Ellis 2003; Lieven 2014; Behrens 2009; Bybee 2010, 2015; Beckner et al. 2009). These developments are largely paralleled and caused by the enormous development in computer technology, as was pointed out by John Sinclair (1991: 1) more than 25 years ago:

> Starved of adequate data, linguistics languished—indeed it became almost introverted. It became fashionable to look inwards to the mind rather than outwards to society. Intuition was the key, and the similarity of language structure to various formal models was emphasized. The communicative role of language was hardly referred to.

Although Sinclair and many other corpus linguists could not be called cognitive linguists, many corpus linguistic insights, especially those concerning multi-word units, collocation and the idiom principle, provide important evidence supporting (and may in some cases have been instrumental in formulating) the position about the nature of language taken in constructionist approaches.

1.2 CxG

It cannot be emphasized too strongly that while not all of the descriptions provided in constructionist frameworks present new insights into the phenomena in question as such, what is worth demonstrating is that these phenomena can be described within this framework, which, after all, is seen by many as offering a more convincing approach towards a comprehensive theory of language than Chomskyan generative linguistics. As most readers will be aware, the fundamental positions proposed in these models include the following[1]:

- Constructionist approaches do not see speakers' linguistic knowledge as being based on inborn properties of the human mind, but envisage it as a network of learned form-meaning pairings called constructions (e.g. Goldberg 2006: 5).[2]
- Constructionist approaches reject a strict dividing line between grammatical and lexical knowledge and work on the assumption of a lexicogrammatical continuum (Langacker 2008: 5).
- Constructionist approaches take linguistic knowledge to be emergent (Bybee 2010: 2).
- Constructionist approaches aim at descriptive adequacy and cognitive plausibility.

[1] Cf. also, for example, Beckner et al. (2009) and Hoffmann and Trousdale (2013: 1–3)

[2] For an outline of the advantages of an approach to language that assumes that knowing a language involves only one type of knowledge, see Stefanowitsch (Stefanowitsch 2011a).

2 Generalizations and Item Specificity

From the point of view of foreign language linguistics, in which phenomena such as collocation and valency (or complementation), which take a position in the centre of the lexicogrammatical continuum assumed by both corpus linguists (Sinclair 2004) and cognitive linguists (Langacker 2008; Goldberg 2006), play a central role, constructionist approaches are an attractive framework because they allow for both item-specific and generalized knowledge to coexist (Goldberg 2006; Bybee 2010), but the role which either plays (for what) is still very much a matter of debate:

> It is as yet not known whether we simply store more and more tokens upon repeated usage, or whether we store more repeated information on a more general and abstract level when available, or whether we do both. (Behrens 2007: 209)

It would indeed be strange to assume that there was no place for generalization in language: Constructionist theories tend to see L1 learning as a process of storage of input and abstraction from it (Bybee 2010; Dąbrowska and Lieven 2005; Lieven 2014; Tomasello 2003). At the same time, it is obvious—and must become clear to language learners at some point of learning a language—that many generalizations do not necessarily apply generally, e.g.:

- Looking at *see*, *bee*, *fee,* etc., one could generalize that words that have a long /iː/ are spelt with a double <ee>; however, looking at *sea*, *tea*, *read*, etc., one could generalize that they are spelt <ea>, and further spellings occur in words such as *key*, *piece*, *be*, *police*, *quay* and *Beauchamp* (Gimson 1989: 101).
- Similarly, *toes*, *foes*, *potatoes* and *tomatoes* allow a generalization of the kind that the grapheme sequence <-oes> is pronounced /əʊz/ in English; but then there is *does* /dʌz/, which, however, makes up 73% of all word-final <-oes>-tokens in the British National Corpus.
- In the area of word formation, we are faced with very much the same sort of situation: *kind* ➔ *kindness, great* ➔ *greatness, polite* ➔ *politeness* etc., but other adjectives nominalize with {-ity} (*brevity, neutrality*), others take both (*clearness, clarity*), etc.

What this means is that we will have to account for the fact that—in a large number of cases, at least—generalizations cannot replace knowledge about the item in the sense that speakers must know which items belong to a particular generalization. What it does not mean is that generalizations are pointless, because, as a rule, knowing about the various options available for deriving a noun from an adjective will facilitate the learning process when a language learner encounters an established nominalization for the first time.

It is the purpose of this article to take up these issues in the areas of valency and collocation and to explore how a number of cases could be dealt with in a constructionist framework.

3 Argument Structure Constructions: The ITECX View

3.1 Argument Structure Constructions as a Challenge for Valency Theory

Theories of valency or complementation have tended to account for differences such as the ones exemplified by the following examples from the point of view of the verb:

(1) a *Obey the speed limit and* **avoid** *being ticketed*. COCA 2015 NEWS
b *... she always* **managed** *to get away with it*. COCA 2015 FIC
(2) a *Die Sachen* [accusative] **bearbeitet** *er allein.* DWDS DIE ZEIT 1948 (He is dealing with these matters on his own.)
b *... auch der Rechnungshof hat sich der Sache* [genitive] **angenommen**.
(... the Financial Control Authorities have also attended to the matter.)

Students of Latin were taught that certain verbs *govern* accusative objects and others dative objects; similarly, students learning German must learn whether a verb *takes* a genitive complement (*sich einer Sache annehmen*), an accusative complement (*eine Sache bearbeiten*) or a prepositional complement (*sich an etwas erinnern*) just as learners of English must learn that avoid *has* a valency slot for a V-ing-clause but not for a *to*-infinitive-clause, for example. The metaphors we use to describe the relationship between the verbs and the elements they occur with imply a dominating role of the verb, which, indeed, is the perspective taken in dependency grammars, in particular in valency theory, and, in fact, all other projectionist approaches (Jacobs 2009). The mere fact that information on valency (or complementation) is included in learners' dictionaries or special valency dictionaries shows that they are considered to be related to particular items.[3]

This item-related view was challenged by Goldberg's (1995, 2006) concept of argument structure constructions, which postulates constructions at a high level of abstraction such as the ditransitive and the caused-motion construction (Tables 1.1 and 1.2).

There are two very good reasons to claim that such general constructions exist in the minds of speakers: one is that when speakers are confronted with test sentences such as:

(3) *They meeped him something*.

the majority of speakers will assign some kind of "transfer"-meaning ("intend-CAUSE-RECEIVE") to the invented verb. The other reason is that creative uses occur with verbs used in a construction that speakers will not have experienced before (Goldberg 2006: 73):

[3] When valency or different complementation patterns are dealt with in grammars, they are usually accompanied by lists of verbs that occur in these patterns. See also the pattern grammar approach taken by Francis, Hunston and Manning (Francis et al. 1996; Francis et al. 1998).

Table 1.1 The ditransitive construction (Goldberg 2006: 20)

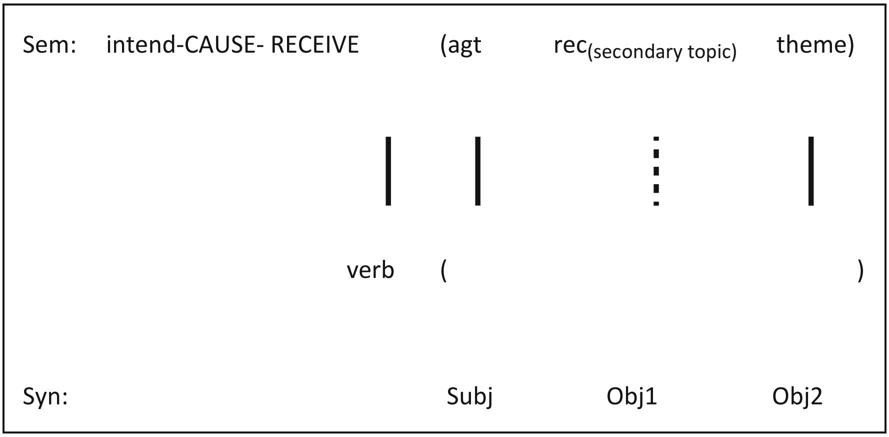
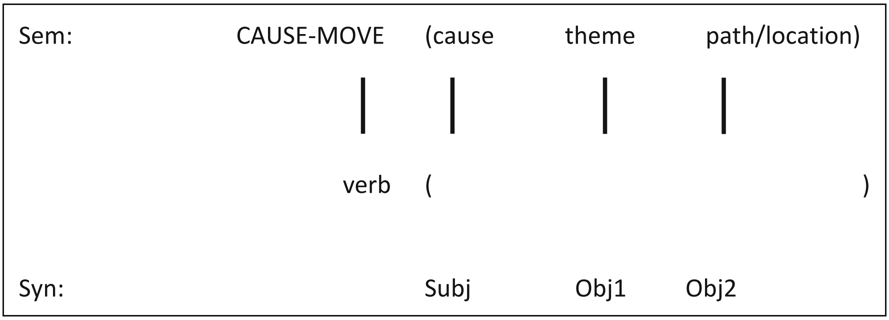


Table 1.2 The caused-motion cx Goldberg (2006: 41)



(4) *Pat sneezed the foam off the cappuccino*.

Since Goldberg's (2006) outline of argument structure constructions offers both an explanation of creative language use of the kind demonstrated in (4) and an account of the meaning of constructions, it goes far beyond traditional accounts of verb complementation such as valency theory.

3.2 Valency as a Challenge for the Theory of Argument Structure Constructions

There is thus a case for integrating elements of Goldberg's theory of argument structure constructions into e.g. valency theory (Herbst 2011a, Welke 2011; see also Engelberg et al. 2011). However, the opposite is also true, because it is difficult to explain restrictions on the use of particular verbs in particular constructions simply in terms of saying that "the more specific participant role of the verb must

Table 1.3 The English ditransitive construction

AGENT	Action	RECIPIENT	THEME
NP1	Give Tell etc.	NP2	NP3

be construable as an instance of the more general argument role"—Goldberg's (2006: 40) semantic coherence principle. A particularly prominent example of this, so prominent that Goldberg chose it as the title of a book dealing with such restrictions—*Explain Me This*—is the fact that the verb *explain* does not occur in the ditransitive construction in English (whereas *erklären* does in German):

(5) a *The starship explained the physics of resistance fields to her* ... COCA 200 FIC
b ?? *The starship explained her the physics of resistance fields...*

One way of accounting for such restrictions is to supplement the semantic coherence principle by a valency realization principle (Herbst 2011a, 2014ab) to account for the dominating role of stored valency information.[4] However, such a principle is not explicitly required if we assume the items that occur in a construction in established use to be part of the representation of the construction (Goldberg forthcoming). The representation of the ditransitive construction could then take the following form (Table 1.3).

3.3 *Collexemes and ITECXes*

How do we know which verbs are represented in a construction in the minds of speakers? The answer is: we don't. First of all, if we follow the exemplar theory advocated by Bybee (2010), according to which every new language experience changes our knowledge of our language, then the representations speakers have will depend on their individual language experiences. Secondly, we do not know enough about how repeated experience of the same type (say *Person X meets Person Y*) is processed and stored in the brain.

However, it would seem reasonable to assume that the analysis of corpora can at least provide us with some indication of which constructions speakers of a language are likely to have experienced, in what form and how often. Note that this, if applied with sufficient caution, neither ignores differences between individuals nor entails that the mental constructicon be a corpus or like a corpus. But it would be very strange if the analysis of the input would not tell us anything about the nature of the knowledge gained by the input.

[4] See also Boas (2003, 2011), Engelberg et al. (2011), Faulhaber (2011), Herbst (2009, 2010, Herbst 2011a, Herbst 2014a, b), Perek (2015) and Stefanowitsch (2011b). This is why the role of lower-level constructions has been stressed by a number of researchers in cognitive linguistics ("mini-constructions" Boas (2003), Hampe and Schönefeld (2006)).

Table 1.4 Collexemes most strongly attracted to the ditransitive construction (Stefanowitsch and Gries 2003)

Collexeme	Collostruction strength	Collexeme	Collostruction strength
Give (461)	0	Allocate (4)	2.91E-06
Tell (128)	1.6E-127	Wish (9)	3.11E-06
Send (64)	7.26E-68	Accord (3)	8.15E-06
Offer (43)	3.31E-49	Pay (13)	2.34E-05
Show (49)	2.23E-33	Hand (5)	3.01E-05
Cost (20)	1.12E-22	Guarantee (4)	4.72E-05
Teach (15)	4.32E-16	Buy (9)	6.35E-05
Award (7)	1.36E-11	Assign (3)	2.61E-04
Allow (18)	1.12E-10	Charge (4)	3.02E-04
Lend (7)	2.85E-09	Cause (8)	5.56E-04
Deny (8)	4.5E-09	Ask (12)	6.28E-04
Owe (6)	2.67E-08	Afford (4)	1.08E-03
Promise (7)	3.23E-08	Cook (3)	3.34E-03
Earn (7)	2.13E-07	Spare (2)	3.5E-03
Grant (5)	1.33E-06	Drop (3)	2.16E-02

The obvious method to measure the association between a construction and the verbs that occur in it is that of collostructional analysis developed by Stefanowitsch and Gries (2003) (Gries and Stefanowitsch 2004a, b; Stefanowitsch 2014). In their pioneering article outlining the concept of the method, Stefanowitsch and Gries (2003: 214) use the verb slot of the English ditransitive construction to demonstrate that certain lexemes "are strongly attracted or repelled by a particular slot in the construction (i.e. occur more frequently or less frequently than expected)"—lexemes attracted to a construction are called collexemes (Stefanowitsch and Gries 2003: 215). Their analysis, which is based on ICE-GB, identifies the following 30 verbs as showing the highest collostructional strength (Table 1.4).

Stefanowitsch and Gries (2003) use the Fisher-Yates exact test to calculate the probability of an item occurring in a particular construction in a corpus. As in the analysis of collocations, different association measures can be applied, whose characteristics have been discussed widely in the literature (e.g. Evert 2005, 2008, Bartsch 2004, Pecina 2010 or Proisl in preparation).

Fundamental objections to collostructional analysis come from Bybee (2010: 101), who observes[5]:

> lexemes that occur only once in a construction within a corpus are treated in two ways by Collostructional Analysis: if they are frequent throughout the corpus, then they are said to be repelled by the construction and if they are infrequent in the corpus, then they are likely to be attracted to the construction. (Bybee 2010: 101)

[5] Compare also Schmid and Küchenhoff (2013) and Gries (2015). For the influence of frequency and the relevance of different types of frequency measures, see Divjak and Caldwell-Harris (2015).