HUMAN COGNITIVE PROCESSING 43

Corpus Methods for Semantics

Quantitative studies in polysemy and synonymy

Edited by
Dylan Glynn
Justyna A. Robinson

John Benjamins Publishing Company

Corpus Methods for Semantics

Quantitative studies in polysemy and synonymy

Edited by

Dylan Glynn
University of Paris VIII

Justyna A. Robinson
University of Sussex

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Corpus Methods for Semantics : Quantitative studies in polysemy and synonymy / Edited
by Dylan Glynn and Justyna A. Robinson.

p. cm. (Human Cognitive Processing, ISSN 1387-6724 ; v. 43)

Includes bibliographical references and index.

1. Semantics. 2. Cognitive grammar. 3. Computational linguistics. 4. Polysemy.
5. Corpora (Linguistics) I. Glynn, Dylan. II. Robinson, Justyna A.

P325.C595 2014
401'.43--dc23 2014004752

ISBN 978 90 272 2397 5 (Hb ; alk. paper)
ISBN 978 90 272 7033 7 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Corpus Methods for Semantics

Human Cognitive Processing (HCP)

Cognitive Foundations of Language Structure and Use

This book series is a forum for interdisciplinary research on the grammatical structure, semantic organization, and communicative function of language(s), and their anchoring in human cognitive faculties.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/hcp

Volume 43

Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy
Edited by Dylan Glynn and Justyna A. Robinson

Contributors

Timothy Colleman
Ghent University
timothy.colleman@UGent.be

Hubert Cuyckens
University of Leuven
hubert.cuyckens@arts.kuleuven.be

Sandra Deshors
New Mexico State University
sadeshors@nmsu.edu

Martine Delorge
Ghent University
martine.delorge@UGent.be

Guillaume Desagulier
Université Paris 8
Vincennes Staint Denis
Université Paris Ouest
Nanterre La Défense
UMR 7114 MoDyCo
gdesagulier@univ-paris8.fr

Dagmar Divjak
University of Sheffield
divjak@sheffield.ac.uk

Małgorzata Fabiszak
Adam Mickiewicz University, Poznań
fagosia@ifa.amu.edu.pl

Nick Fieller
University of Sheffield
nick.fieller@sheffield.ac.uk

Dirk Geeraerts
University of Leuven
dirk.geeraerts@arts.kuleuven.be

Dylan Glynn
University of Paris VIII
dglynn@univ-paris8.fr

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Anna Hebda
Adam Mickiewicz University, Poznań
ahebda@ifa.amu.edu.pl

Martin Hilpert
University of Neuchâtel
martin.hilpert@unine.ch

Jane Klavan
University of Tartu
klavan@ut.ee

Iwona Kokorniak
Adam Mickiewicz University, Poznań
kokorniak@ifa.amu.edu.pl

Karolina Krawczak
Adam Mickiewicz University, Poznań
karolina@ifa.amu.edu.pl

Natlia Levshina
Université catholique de Louvain
natalia.levshina@uclouvain.be

Florent Perek
University of Freiburg
florent.perek@frias.uni-freiburg.de

Koen Plevoets
Ghent University
koen.plevoets@UGent.be

Justyna A. Robinson
University of Sussex
justyna.robinson@sussex.ac.uk

Christopher Shank
Bangor University
c.shank@bangor.ac.uk

Dirk Speelman
University of Leuven
dirk.speelman@arts.ku.leuven.be

Joost van de Weijer
Lund University
vdweijer@ling.lu.se

Table of contents

Section 2. Statistical techniques

Outline

1. Aim of the volume

It could be argued that Cognitive Linguistics is undergoing a paradigm shift. Originally, the field sought to show the inadequacies of earlier models of language and the theories of linguistic structure based upon them. Today, the emphasis has changed to testing the various theories about how language works (Geeraerts 2006; Gries and Stefanowitsch 2006; Stefanowitsch and Gries 2006; Gonzalez-Marquez et al. 2008; Glynn and Fischer 2010). This has brought analytical methods, based on observable and quantifiable data, to the fore. In the light of these developments, this volume systematises, reviews, and promotes a range of research techniques and theoretical perspectives that currently inform work across the field of linguistics, with a particular focus on Cognitive Semantics. More precisely, the aim of this book is twofold:

i. Didactic: To broaden the understanding and application of the state-of-the-art corpus linguistic techniques for the study of conceptual structure in Cognitive Semantics.
ii. Scientific: To advance the state-of-the-art of those techniques through a collection of studies applied to the description of the conceptual structures of polysemy and synonymy.

This publication grew out of the belief that there exists a strong desire in the research community to understand and learn how quantitative corpus methods work and how to apply them to research questions that are basic to the cognitive project. Instead of a rift between linguists using corpus data and those using traditional introspective analysis, constructive communication between the methodologies should be encouraged. Both the descriptive research and the explanations of the statistical techniques included in this book seek to promote such communication. The chapters that describe the statistical techniques are written to help linguists using traditional methods both understand how these new methods work and how to apply them. The research chapters, in turn, showcase the methods described. Their aim is not only to advance corpus-driven quantitative research in Cognitive Semantics, but also to promote the possibilities that these methodologies offer. Observational data and quantitative corpus-driven methods cannot inform all research questions. However, it is hoped that this volume will advance the current state-of-the-art in their use as well as promote their application in the broader linguistic research community.

2. Structure and summary

The book divides into two sections. The first section begins with eleven chapters, arranged according to their object of study. These chapters begin with an overview of the field in "Polysemy and synonymy: Corpus method and cognitive theory" (Glynn). This chapter includes the analytical justification for approaching both lexis and morpho-syntax in terms of polysemy and synonymy as well as a justification of extending the traditional uses of the terms to cover any variation or similarity in use. The analytical chapters begin with morpho-syntactic polysemy, move to lexical polysemy, then on to lexical synonymy, and finally turn to morpho-syntactic synonymy.

Beginning with research on the polysemy of morpho-syntactic semantics, the first descriptive chapter, "Competing 'transfer' constructions in Dutch" (Delorge, Plevoets, and Colleman), considers the polysemy of a morpheme-based construction. The *ont-* prefix in Dutch combines with a range of verbs to express dispossession. Using correspondence analysis, the study seeks to capture the lexical semantic morpho-syntactic interplay associated with the construction. The next chapter, "Rethinking constructional polysemy" (Perek), also examines a grammatical construction. The syntactically encoded conative construction in English combines with a range of lexemes. Through the application of collostructional analysis, the author attempts the task of teasing out and identifying the semantic variation associated with the construction.

Turning to lexical semantics, "Quantifying polysemy in cognitive sociolinguistics" (Robinson) examines the usage of polysemous adjectives in a community of speakers from South Yorkshire, UK. The study applies cluster analysis, logistic regression, and decision tree analysis in order to examine the extent to which individual conceptualisations are non-random and can be related to the socio-demographic characteristics of the speaker. "The many uses of *run*" (Glynn) is a repeat analysis of Gries' (2006) study. Employing a combination of cluster analysis, correspondence analysis and logistic regression, it confirms Gries' findings but argues that sociolinguistic dimensions should be included in the study of polysemy.

Remaining with lexical semantics, but focusing on near-synonymy, "Visualizing distances in a set of near-synonyms" (Desagulier) examines *rather*, *quite*, *fairly*, and *pretty* in English, combining collostructional analysis and multivariate statistics such as correspondence analysis and cluster analysis. "The uses of *may* and *can* in French-English interlanguage" (Deshors and Gries) treats a lexical alternation in first language and second language use. With the use of cluster analysis and logistic regression, the authors seek to identify not only the relationship in use between *may* and *can*, but to compare this with French native speakers using English.

Moving towards morpho-syntactic semantics, "Dutch causative constructions" (Levshina, Geeraerts, and Speelman) examines a lexeme-based grammatical construction alternation. Focusing on the expression of causation in Dutch, the study employs logistic regression analysis to determine both the semantic and extralinguistic factors that determine the choice and difference in conceptualisation between the two constructions. The next study, "The semasiological structure of Polish *myśleć* 'to think'" (Fabiszak, Hebda, Kokorniak, and Krawczak) continues to move from lexical to syntactic semantics with an analysis of the near-synonymy of a set of prefix-verb combinations. The study combines introspective methods and usage-feature analysis, examined with cluster analysis, correspondence analysis and logistic regression.

"A multifactorial analysis of grammatical synonymy" (Klavan) studies a lexical-morphological alternation. Employing logistic regression, the study attempts to determine the conceptual differences that motivate speakers' choice of a preposition over a grammatical case to express the spatial relation of *on* in Estonian. "A diachronic corpus-based multivariate analysis of 'I think *that*' vs. 'I think zero'" (Shank, Plevoets, and Cuyckens) is an analysis of well-known complementiser alternation in English. Logistic regression is used to test a wide range of language factors proposed in the literature to motivate the omission of the complementiser.

The second section consists of seven chapters that explain some of the tools and methods for quantitative corpus-driven research. It is designed to introduce the application and interpretation of the statistical methods used in the first section for researchers completely new to the field. It also serves as a 'cookbook', or is a quick reference, for intermediate users of the statistical techniques and the programming environment R. The first chapter, "Techniques and tools: Corpus methods and statistics for semantics" (Glynn), is an overview of the field. It examines two corpus methods that are commonly used in Cognitive Semantics and summarises many of statistical techniques currently used in the field.

The second chapter introduces the statistical environment of R, used throughout the book (van de Weijer and Glynn). Readers with no experience in R will find this chapter useful when applying the techniques described in the previous chapters to data analysis. The following chapter, "Frequency tables: Tests, effect sizes, and explorations" (Gries), covers many of the essential and basic analytical concepts and how to apply them in R. Building on the statistical basics, the next chapter, "Collostructional analysis: Measuring associations between constructions and lexical elements" (Hilpert), explains the application of collostructional analysis, one of the most popular quantitative techniques in Cognitive Linguistics. This family of techniques are used to quantify the degree of association between linguistic forms.

The next three chapters each consider a different multivariate statistical technique. The three techniques in question have proven popular in recent Cognitive Linguistic research. The chapter "Cluster analysis: Finding structure in linguistic data" (Divjak and Fieller), focuses on a method for sorting a given set of phenomena, such as lexemes, constructions, or senses, into categories of similar and dissimilar, relative to some other set(s) of linguistic phenomena such as meanings, argument types, case marking, and so forth. This is followed by the chapter "Correspondence analysis: Exploring data and identifying patterns" (Glynn), which considers a technique similar to cluster analysis, but one that looks for correlations between different phenomena rather than categorising them. It is useful for identifying structure in complex multidimensional data. The third chapter on multivariate techniques, "Logistic regression: A confirmatory technique for variant comparison in corpus linguistics" (Speelman), considers an advanced form of statistical modelling. Logistic regression is a powerful and popular tool in the social sciences, including Cognitive Linguistics. As a confirmatory technique, regression analysis represents a level of statistical analysis that is more complex than the previous techniques covered. This chapter charts the basics of its application, interpretation and verification.

Where the first section represents a broad, yet coherent, picture of the cutting edge in the application of these techniques, the second section seeks to offer an introduction to the different statistical techniques employed in corpus-driven semantics. Focusing on the study of polysemy and synonymy of both lexical morpho-syntactic forms, these empirical analyses represent the vanguard of corpus-driven Cognitive Linguistics.

The editors

References

Geeraerts, D. (2006). Methodology in Cognitive Linguistics. In G. Kristiansen, M. Achard, R. Dirven, & F. J. Ruiz de Mendoza Ibañez (Eds.), *Cognitive Linguistics: Current applications and future perspectives* (pp. 21–50). Berlin & New York: Mouton de Gruyter.

Glynn, D., & Fischer, K. (Eds.). (2010). *Quantitative Cognitive Semantics: Corpus-driven approaches*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226423

Gonzalez-Marquez, M., Mittelberg, I., Coulson, S., & Michael S. (Eds.). (2008). *Methods in Cognitive Linguistics*. Amsterdam & Philedelphia: John Benjmains.

Gries, St. Th. (2006). Corpus-based methods and Cognitive Semantics: The many senses of *to run*. In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis* (pp. 57–99). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197709

Gries, St. Th., & Stefanowitsch, A. (Eds.). (2006). *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis*. Berlin & New York: Mouton de Gruyter.
DOI: 10.1515/9783110197709

Stefanowitsch, A., & Gries, St. Th. (Eds.). (2006). *Corpus-based approaches to metaphor and metonymy*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110199895

SECTION 1

Polysemy and synonymy

Polysemy and synonymy

Cognitive theory and corpus method

Dylan Glynn
University of Paris VIII

This chapter introduces the field of polysemy and synonymy studies from a Cognitive Linguistic perspective. Firstly, the discussion explains and defines the object of research, showing that the study of semantic relations, traditionally restricted to the description of lexical semantics, needs to be extended to include all formal structures, including morpho-syntax. Secondly, given the theoretical assumptions of Cognitive Linguistics, it is argued that quantitative corpus-driven methods are essential for the description of semantic structures. Lastly, the chapter charts the development of Cognitive Semantic research in polysemy and synonymy and demonstrates how the current corpus-driven research in the field is inherently linked to the traditions of radial network analysis and prototype semantics. It is argued that instead of an empirical revolution (as has been suggested in recent commentaries), the current trends in the use of observational data are a natural extension of the Cognitive Semantic research tradition.

Keywords: Cognitive Linguistics, corpus linguistics, polysemy, prototype semantics, quantification, radial network analysis, synonymy

1. Introduction: Theory and method

The idea of 'corpus semantics', just like the possibility of 'quantifying meaning', is not self-evident. This introduction to the field of corpus-driven Cognitive Semantics attempts to explain how semantic analysis can, and indeed, should, turn to corpus methods. It also explains why quantitative techniques are needed in this endeavour.

Assuming the Usage-Based Model (Hopper 1987; Langacker 1987), how can we identify and explain the semantic structuring of language empirically? Post-Generativist and post-Structuralist approaches to language avoid positing, *a priori*, analytical constructs to explain the structuring of language, rather treating it holistically as a dynamic and varied result of use. However, without a structurally independent *langue* or an 'ideal' speaker's competence, against what predictive model can we test our hypotheses or attempt to falsify our claims about language structure? Without constructs such as *langue* and ideal competence, linguistic research, whether Functional or Cognitive, must adopt an inductive, sample-based, methodology. To these ends, experimental techniques for the analysis of semantics have been developed, yet corpus methods remain poorly represented within the field.

Moreover, the theory of Cognitive Linguistics recognises no internal language modules, such as syntax, lexis, semantics or pragmatics. From a non-modular perspective, the study of meaning must account for the integration of all these components of language structure and do so simultaneously in a functionally and conceptually plausible manner. Corpus-driven methods, and multivariate statistics more specifically, are perfectly suited for such a task.

A usage-based semantics, therefore, must take two fundamental steps. Firstly, it must adopt inductive research methods. Whether elicited through experimentation, extracted from electronic corpora, or collated from questionnaires and field research, generalisations based on samples of data present the only possibility for hypothesis testing. Importantly, acknowledging this fact entails validating sample-based results through statistical confirmation. Secondly, it must develop corpus-driven semantic analysis. If we are to account holistically for the integrated complexity of the various dimensions of language structure, it is essential that we examine natural contextualised language production. Samples of natural language large enough to permit inductively valid claims are what we term corpora. Here again, statistics comes to the fore, though for a different reason. If we are to identify structure, sensitive to its usage context, multivariate statistics is a powerful, if not essential, tool due to the sheer complexity of the data.

The aim of this book is both to introduce quantitative corpus-driven semantic methodology to the broader research community and to advance the state-of-the-art. The methods in focus are those that are especially applicable to the lexical and constructional semantic relations of similarity and difference. Linguistic forms are used in different ways, and capturing this semasiological variation is what we term the study of polysemy. Likewise, speakers choose between different linguistic forms to express similar concepts. Explaining this onomasiological variation is what we term the study of synonymy.

The reader should be aware that the term polysemy is not restricted to 'true' polysemy, where distinct referents are indicated by a single form. Instead, meaning is understood from a usage-based perspective, where any systematic variation in use represents semasiological structure. In the same vein, synonymy is not restricted to absolute similarity since, from a Cognitive Linguistic perspective, one assumes that any variation in form is motivated by some variation in use and that 'true' synonymy is rare, if it exists at all. Lastly, it must be added that the term semantics is used to indicate encyclopaedic semantics and pragmatics, as opposed to linguistic semantics in its narrow sense.