DE GRUYTER MOUTON
Folia Linguistica 2016; 50(2): 345–353

Introduction

Małgorzata Fabiszak, Martin Hilpert and Karolina Krawczak*

Usage-based cognitive-functional linguistics: From theory to method and back again

DOI 10.1515/flin-2016-0013

This special issue grew out of a theme session organized at the *Societas Linguistica Europaea* conference in Poznań in 2014. The focus of this session was on the interface between linguistic theories and methods and, more specifically, on how current empirical methods, especially those employed in usage-based linguistics, should inform and, ultimately, advance linguistic theory. With this underlying question, the present collection brings together corpus-based research in areas such as construction grammar (Hennemann; Krawczak et al.; Pijpops and Van de Velde), historical linguistics (Barteld et al.), semantics (Glynn), and typology (Levshina). It also includes a survey article addressing the cognitive plausibility of statistical classification modeling applied to observational and experimental data (Klavan and Divjak) and arguing for the need to combine the two in order to refine explanatory models.

The empirical turn in linguistics is well documented (Geeraerts 2006; Stefanowitsch 2008; Fischer 2010; Glynn 2014b). Within Cognitive-Functional linguistics alone, there has been a substantial increase in the use of corpus-based quantitative methods, covering multiple domains. Morpho-syntactic phenomena have been addressed by Wulff (2003), Gries et al. (2005), Goldberg (2006), Hilpert (2008a, 2008b, 2013), Levshina (2012), Krawczak and Glynn (2015), and many others. Quantitative research in semantics can be exemplified by the work of Geeraerts et al. (1994), Schmid (2000), Gries (2006), Divjak and Gries (2006), Glynn (2009, 2014a), Divjak (2010), Fabiszak et al. (2014), and Krawczak (2014). Among sociolinguistic explorations in Cognitive Linguistics that employ quantitative methods, we could mention Szmrecsanyi (2005), Heylen et al. (2008), and

***Corresponding author: Karolina Krawczak,** Faculty of English, Adam Mickiewicz University, Poznań al. Niepodległości 4 61-874 Poznań, Poland, E-mail: karolina@wa.amu.edu.pl
Małgorzata Fabiszak, Faculty of English, Adam Mickiewicz University, Poznań al. Niepodległości 4 61-874 Poznań, Poland, E-mail: fagosia@wa.amu.edu.pl
Martin Hilpert, Institut de langue et littérature anglaises, Université de Neuchâtel, Espace Louis-Agassiz 1, CH-2000 Neuchâtel, Switzerland, E-mail: martin.hilpert@unine.ch

Peirsman et al. (2010). There have also been a number of well-received edited volumes over the past few years promoting quantitative corpus-based methods, including Gries and Stefanowitsch (2006), Glynn and Fischer (2010), or Glynn and Robinson (2014).

Despite the general sense of optimism that is inspired by these developments in empirical research, it is probably justified to pause for a moment of (self-)critical reflection and to consider difficulties that still lie ahead. In this context, one important problem concerns the theoretical pluralism in linguistics, in which many different approaches and theoretical models coexist. It is an open question whether linguistics can "turn its current theoretical chaos (...) into a situation of cumulative development" (Geeraerts 2006: 21). In other words, can linguistics test and compare the existing theoretical models to demonstrate in a systematic manner which of them best explain the empirical phenomena under investigation? For such a goal to be attainable, theory and method must be integrated so as to be mutually informative. In this theoretical-analytical cycle, theory should serve as the basis for operationalizing research questions and hypotheses, while methods should not only provide descriptive results but should ultimately serve to test theoretical proposals (see Tummers et al. 2005; Geeraerts 2010; Glynn 2010). The present volume focuses on corpus linguistics, as practiced in cognitive-functional linguistics, paying particular attention to how corpus-based methods are employed for the purposes of testing hypotheses and refining theoretical claims.

In the most general terms, quantitative methods applied in corpus linguistics fall into three main categories, together forming the "distributional" approach (after Heylen et al. 2015). The first type, dating back to Church and Hanks (1990) or Sinclair (1991), focuses on surface associations, combining automatic identification of contextual clues (collocations and colligations) with largely subjective interpretation. This method has been further developed through the work of Gries and Stefanowitsch (e. g., 2006) or Hilpert (e. g., 2008a). The second group of quantitative corpus methods comprises what is known as the profile-based or multifactorial usage-feature approach. This method integrates qualitative analysis of contextual clues, which is manual for the most part, with multivariate modeling (e. g., Geeraerts et al. 1994; Gries 2006; Glynn 2009; Divjak 2010), thus revealing frequency-based multifactorial profiles of language use. Finally, the last type, which has emerged most recently, could be considered a fusion of the other two. Similarly to the first approach, it determines the relevant contextual parameters automatically, based on directly observable features. Similarly to the second approach, it employs quantitative modeling to reveal usage patterns (e. g., Turney and Pantel 2010; Heylen and Ruette 2013; Levshina and Heylen 2014; Heylen et al. 2015).

All three approaches represent important methodologies and each of them has its assets. For example, both the first and third are designed for analyzing large data samples that would not be amenable to manual annotation. The profile-based approach, on the other hand, permits meticulous examination of linguistic features that are hard to observe directly, pertaining to such areas as semantics or pragmatics. Another crucial advantage offered by all three methods is that they produce falsifiable results. Importantly, irrespective of which method one implements, what should receive special attention is how these state-of-the-art methods, in addition to providing descriptive insights, can improve our understanding of language and communication. How can they afford answers to questions that offer theoretical value?

There is clearly no need today to convince anyone within cognitive-functional linguistics of the crucial role that empiricism plays in research, but there still seems to be some need for emphasizing how empiricism should link back to theory. This need to bring theory and method together is what lies at the heart of all the contributions to this volume. Except for the survey article by Klavan and Divjak, all of the studies in this special issue can be taken to be representative of the profile-based approach. We will now discuss each of the articles in some more detail.

Jane Klavan and Dagmar Divjak present a survey article in which they discuss the advantages of testing statistical models of corpus data against language users' behavior in experimental settings. The authors stress that neither procedure should be considered superior to the other, as each has its weaknesses, which can only be overcome when they are both used in combination. As pointed out by Mitchell (2012), experimental studies often suffer from limited external validity. They have two underlying weaknesses. Firstly, insufficient attention is given to the possible interaction between factors in the natural setting, which may result in an ad hoc choice of variables for analysis. The second potential problem is artificial stimuli. Both of these issues can be overcome by relying on corpus data and using multivariate analysis. The article focuses on how modeling multivariate corpus data may benefit from juxtaposition with experimental results. The meta-analysis of four studies in four different languages (Arabic, English, Estonian, and Russian) shows that the models based on the analysis of corpus data either performed at a similar level of accuracy or outperformed native speakers completing a forced choice task. In the former case, such results can be taken to confirm the cognitive validity of the model built on observational data. In the latter case, i. e., if the model outperforms native speakers, this may be viewed as an indication that the explanatory factors should be revised and the model fine-tuned: not all regularities that can be discovered in the data are exploited by speakers.

Fabian Barteld, Stefan Hartmann and Renata Szczepaniak investigate the development of the sentence-internal capitalization of nouns in Early New High German. While capitalization of nouns is a general orthographic convention of written Present-day German, diachronic corpora show that there used to be variation involving the factors animacy, frequency, and complexity. Specifically, nouns with animate referents spearheaded the development and settled earlier into fully capitalized usage. With regard to frequency, Barteld, Hartmann and Szczepaniak observe that low-frequency nouns show a greater extent of variation, while conversely, high-frequency nouns are quickest to adopt a stable capitalized pattern. Capitalization appears to be promoted by the complexity of a noun as well. Additionally, the factors of animacy and frequency are shown to interact, so that for example the effect of high frequency is not as pronounced for nouns denoting female human beings, which have a very high likelihood of capitalization to begin with. A general result of the study is that also graphemic language use reflects cognition and therefore should not be neglected.

In his paper on cognitive semantics, Dylan Glynn addresses a fundamental, yet often overlooked, methodological consequence of Prototype Set Theory – the difficulty of result falsification. First, he revisits the notion of prototype semantics, focusing, in particular, on the fluidity of polysemy and its continuous nature. The author shows that recent empirical, and thus falsifiable, studies describing polysemy start from a list of senses (cf. Gries 2006), which, despite being cognitive in orientation, does not conform to the theoretical tenets of prototype theory or the usage-based model. Turning to the methodological consequence of these tenets, the author notes that in early cognitive studies, e. g., on the polysemy of prepositions (Brugman 1983; Lakoff 1987), result falsification was not possible, for, unlike in formal semantics, such descriptions did not contain rules that could be falsified with counter-examples. It is argued that such early research on the prototype structuring of meaning is better understood as hypotheses about semantic structure, rather than actual case studies. In order to obtain an empirical and, therefore, falsifiable method which still adheres to the theoretical tenets of Cognitive Linguistics, Glynn suggests employing the Usage Feature Analysis (Glynn 2008, 2009, 2010), also known as the Behavioral Profile Analysis (Divjak 2006; Gries 2006; Gries and Divjak 2009). In his case study of *annoy*, he follows a step-by-step Usage Feature Analysis, demonstrating how to arrive at sense aggregates, rather than discrete senses, in an entirely bottom-up manner. Unlike an introspective approach, the quantitative analysis of observational data allows the author to test the predictive power and descriptive accuracy of the results obtained.

Anja Hennemann in her study of two Spanish constructions *creo* and *creo yo* faces the problem of scarcity of oral interactional data in Spanish corpora. Under the circumstances, she makes a strong point for the value of qualitative analysis as a theory building method. Her analysis elaborates on earlier descriptions of the functions of *creo*/*creo yo* by showing that *creo yo* can be treated as an intersubjectivity marker, as it is interpreted by the interlocutors as an invitation to express their opinion. This novel observation will be further investigated quantitatively in future research drawing on data from discussion forums and other forms of online interaction.

The study by Karolina Krawczak, Małgorzata Fabiszak and Martin Hilpert is a corpus-based quantitative account of complement alternation observed for a set of mental verbs in English, German, and Polish. The constructional alternation investigated involves the choice between nominal and clausal complementation. The authors employ statistical methods to test two specific hypotheses, informed by relevant prior research in one of the languages examined. The hypotheses draw on the well-established distinction between descriptivity/objectivity and performativity/subjectivity (see Benveniste 1971; Nuyts 2001; Verhagen 2005), which is here operationalized in terms of boundedness and picturability. The proposed operationalization was methodically implemented through manual annotation of the data. It was expected that third-person (descriptive) occurrences of the predicates would more readily correlate with bounded and picturable objects, while the contrary pattern was hypothesized for the first-person (performative) uses of such verbs. To test the accuracy of this hypothesized tendency and to determine what usage properties motivate the choice between the two complementation patterns, the obtained metadata were submitted to bivariate and multivariate modeling, as appropriate. The results not only offer valuable, if unexpected, insights into the theoretical distinction investigated in the study, but also present methodological value for future research on the topic.

Natalia Levshina adopts a corpus-based quantitative perspective to examine the onomasiological variation in lexical and analytic causative constructions across fifteen European languages. The data for this study were extracted from a multilingual parallel corpus of film subtitles compiled by the author. In her analysis, she tests the explanatory accuracy of a range of variables that have been discussed in the typological literature on cross-linguistic variation of causatives. More precisely, the author seeks to establish whether the choice between the two types of causative constructions is determined by a single semantic dimension, as prior research seems to suggest, or rather by a combination of semantic and syntactic variables. To answer this question, she employs the statistical methods of correspondence analysis and conditional random forests. Importantly, the results thus obtained are informative not only descriptively, but also theoretically.

Overall, the findings confirm the iconicity-based accounts of variation in causative constructions, but, at the same time, the study clearly shows that the variation is structured along a number of semantic and formal parameters, which cannot be explained in terms of iconicity alone. Hence, it becomes evident that a multivariate analysis is needed to provide a comprehensive explanation of the investigated phenomenon. Moreover, by integrating qualitative analysis of corpus data for the sampled languages with quantitative modeling, the author demonstrates the feasibility of token-based statistical analyses in typological research and makes a plea for more such work in the field. This constitutes an important contribution to typological research, where quantitative corpus-based studies do not represent the norm yet.

In their contribution, Dirk Pijpops and Freek Van de Velde develop a new theoretical notion that they call "constructional contamination". What is meant by this term is that the usage of one construction may be influenced by another construction that is superficially similar. To illustrate this idea, the authors present a case study of the Dutch partitive genitive construction, which exhibits morphological variation. In this construction, an adjective in the genitive may be marked with an *-s* suffix, as in *iets verkeerds* 'something wrong', or the adjective may be bare, as in *iets verkeerd*. Pijpops and Van de Velde observe that the *s*-less variant is strongly preferred in cases where the quantifier and the *s*-less version of the adjective co-occur frequently in other constructions. A frequent *s*-less collocation may thus be seen as a contamination that affects the partitive genitive construction. Pijpops and Van de Velde operationalize the concept of constructional contamination in four different ways, analyze the predictive power of each operationalization, and present detailed critiques of each one, concluding that constructional contamination works through both formal and semantic resemblance. On the whole, the analysis supports the idea that speakers engage in shallow parsing (Ferreira and Patson 2007) and use exemplar-based representations of syntactic structures (Dąbrowska 2014).

In our view, all the contributions in this special issue usefully connect theory and empirical research in order to test currently held assumptions, thus advancing our understanding of the investigated language phenomena. The authors in this volume have turned to corpus data to achieve this, which, of course, represents only one line of empirical research. However, irrespective of the type of data employed, the general argument that we propose here remains unchanged: we should always seek to complete the research "cycle" by demonstrating how our results bear upon relevant theoretical questions. In other words, while describing language behavior, we should never lose sight of that which we are seeking to explain, i. e., language structure.

References

Benveniste, Emile. 1971. Subjectivity in language. In Emile Benveniste (ed.), *Problems in general linguistics*, 223–230. Coral Gables: University of Miami Press.

Brugman, Claudia. 1983. *The story of* over: *Polysemy, semantics, and the structure of the lexicon*. Trier: LAUT.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1). 22–29.

Dąbrowska, Ewa. 2014. Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics* 25(4). 617–653.

Divjak, Dagmar. 2006. Ways of intending: A corpus-based Cognitive Linguistic approach to near-synonyms in Russian. In Stefen Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 19–56. Berlin: Mouton de Gruyter.

Divjak, Dagmar. 2010. *Structuring the lexicon: A clustered model for near-synonymy*. Berlin: De Gruyter Mouton.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2. 23–60.

Fabiszak, Małgorzata, Anna Hebda, Iwona Kokorniak & Karolina Krawczak. 2014. The semasiological structure of Polish *myśleć* 'to think': A study in verb-prefix semantics. In Dylan Glynn & Justyna Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 223–252. Berlin: De Gruyter Mouton.

Ferreira, Fernanda & Nikole Patson. 2007. The "good enough" approach to language comprehension. *Language and Linguistics Compass* 1. 71–83.

Fischer, Kerstin. 2010. Quantitative methods in cognitive semantics. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative cognitive semantics: Corpus-driven approaches*, 43–61. Berlin: De Gruyter Mouton.

Geeraerts, Dirk. 2006. A rough guide to Cognitive Linguistics. In Dirk Geeraerts (ed.), *Cognitive Linguistics: Basic readings*, 1–28. Berlin: Mouton de Gruyter.

Geeraerts, Dirk. 2010. The doctor and the semantician. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative cognitive semantics: Corpus-driven approaches*, 63–78. Berlin: De Gruyter Mouton.

Geeraerts, Dirk, Stefan Grondelaers, & Peter Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2008. Lexical fields, grammatical constructions and synonymy. A study in usage-based cognitive semantics. In Hans-Jörg Schmid & Susanne Handl (eds.), *Cognitive foundations of linguistic usage-patterns: Empirical studies*, 89–118. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2009. Polysemy, syntax, and variation: A usage-based method for Cognitive Semantics. In Vyvyan Evans & Stéphanie Pourcel (eds.), *New directions in cognitive linguistics*, 77–106. Amsterdam: John Benjamins.

Glynn, Dylan. 2010. Testing the hypothesis: Objectivity and verification in usage-based cognitive semantics. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative cognitive semantics: Corpus-driven approaches*, 239–270. Berlin: De Gruyter Mouton.

Glynn, Dylan. 2014a. The many uses of *run*: Corpus methods and socio-cognitive semantics. In Dylan Glynn & Justyna Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 117–144. Berlin: De Gruyter Mouton.

Glynn, Dylan. 2014b. Polysemy and synonymy: Cognitive theory and corpus method. In Dylan Glynn & Justyna Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 7–38. Berlin: De Gruyter Mouton.

Glynn, Dylan & Kerstin Fischer (eds.). 2010. *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, Berlin: De Gruyter Mouton.

Glynn, Dylan & Justyna Robinson (eds.). 2014. *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins

Goldberg, Adele E. 2006. *Constructions at work: On the nature of generalization in language*. Oxford: Oxford University Press.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57–99. Berlin: Mouton de Gruyter.

Gries, Stefan Th. & Dagmar Divjak. 2009. Behavioral profiles: A corpus-based approach towards cognitive semantic analysis. In Vyvyan Evans & Stéphanie Pourcel (eds.), *New directions in cognitive linguistics*, 57–75. Amsterdam: John Benjamins.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16. 635–676.

Gries, Stefan Th. & Anatol Stefanowitsch (eds.). 2006. *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexicon*. Berlin: Mouton de Gruyter.

Heylen, Kris & Tom Ruette. 2013. Degrees of semantic control in measuring aggregated lexical distances. In Lars Borin & Anju Saxena (eds.), *Approaches to measuring linguistic differences*, 253–374. Berlin: De Gruyter Mouton.

Heylen, Kris, José Tummers & Dirk Geeraerts. 2008. Methodological issues in corpus-based cognitive linguistics. In Gitte Kristiansen & René Dirven (eds.), *Cognitive sociolinguistics: Language variation, cultural models, social systems*, 91–128. Berlin: Mouton de Gruyter.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman & Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172.

Hilpert, Martin. 2008a. *Germanic future constructions: A usage-based approach to language change*. Amsterdam: John Benjamins.

Hilpert, Martin. 2008b. New evidence against the modularity of grammar: Constructions, collocations, and speech perception. *Cognitive Linguistics* 19(3). 391–411.

Hilpert, Martin. 2013. *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.

Krawczak, Karolina. 2014. Shame, embarrassment and guilt: Corpus evidence for the cross-cultural structure of social emotions. *Poznań Studies in Contemporary Linguistics* 50. 441–475.

Krawczak, Karolina & Dylan Glynn. 2015. Operationalizing mirativity: A usage-based quantitative study of constructional construal in English. *Review of Cognitive Linguistics* 13(2). 253–282.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Levshina, Natalia. 2012. Comparing constructions: A usage-based analysis of the causative construction with *doen* in Netherlandic and Belgian Dutch. *Constructions and Frames* 4(1). 76–101.

Levshina, Natalia & Kris Heylen. 2014. A radically data-driven construction grammar: Experiments with Dutch causative constructions. In Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.), *Extending the scope of construction grammar*, 17–46. Berlin: De Gruyter Mouton.

Mitchell, Gregory. 2012. Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science* 7(2). 109–117.

Nuyts, Jan. 2001. *Epistemic modality, language, and conceptualization*. Amsterdam: John Benjamins.

Peirsman, Yves, Kris Heylen & Dirk Geeraerts. 2010. Applying word space models to sociolinguistics. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in cognitive sociolinguistics*, 111–138. Berlin: De Gruyter Mouton.

Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: Mouton de Gruyter.

Sinclair, John M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stefanowitsch, Anatol. 2008. Negative entrenchmen: A usage-based approach to negative evidence. *Cognitive Linguistics* 19. 513–531.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1. 113–149.

Tummers, José, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1. 225–261.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188.

Verhagen, Arie. 2005. *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford: Oxford University Press.

Wulff, Stephanie. 2003. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics* 8. 245–282.