Corpora and Representativeness Corpora and Representativeness: Where to go from now?

Sophie Raineri, Camille Debras

CogniTextes
Revue de l'Association française de linguistique cognitive

Volume 19 | 2019
Corpora and Representativeness

Corpora and Representativeness: Where to go from now?

Sophie Raineri and Camille Debras

Electronic version
URL: http://journals.openedition.org/cognitextes/1311
ISSN: 1958-5322

Publisher
Association française de linguistique cognitive

Electronic reference
Sophie Raineri & Camille Debras, « Corpora and Representativeness: Where to go from now? », *CogniTextes* [Online], Volume 19 | 2019, Online since 08 June 2019, connection on 22 October 2019.
URL : http://journals.openedition.org/cognitextes/1311

Corpora and Representativeness: Where to go from now?

Sophie Raineri and Camille Debras

1 Twentieth-century structuralist and generative linguists argued that the study of the language system (langue, competence) must be separated from the study of language use (parole, performance). For Saussure or Chomsky, no generalizations about language could be made based on the observation of patterns, regularities and rules of language performance. For Saussure, "Il n'y a donc rien de collectif dans la parole ; les manifestations en sont individuelles et momentanées. Ici il n'y a rien de plus que la somme des cas particuliers selon la formule (1+1'+1''+1'''...)" (Saussure 1964 (1916): 38). For Chomsky, "any natural corpus will be skewed. Some sentences won't occur because they are too obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list" (Chomsky 1957: 159). Since the early 1980s, this view of language and language study started to be called into question by functionalist (Givón 1979; Hopper 1987; Bybee 1985; 2010) and cognitive approaches (Langacker 1987), which argued that language use is central in the structure and organization of a speaker's linguistic knowledge, and needs to be empirically investigated.

2 In the wake of Langacker's (1987; 1991) foundational work on cognitive grammar, cognitive linguistics has long been concerned with how speakers represent, process, and actually use language. Although first-generation cognitive linguistics was theory-driven (Fillmore 1982; Lakoff 1987; Talmy 2000), the usage-based approach of language has been at the core of cognitive linguists' work from the start. In parallel, corpus linguistics has developed as one source of evidence for improving descriptions of the structures and use of languages. It can be described as an approach that empirically analyzes language use in large and principled collections of authentic texts, thanks to automatic and/or computerized tools and based on a combination of quantitative and qualitative techniques (Biber, Conrad & Reppen 1998: 4). The introduction of corpus-based methods in usage-based cognitive linguistics, also described as its "empirical turn"[1], can be traced back to the 2000's (Fanego 2004; Geeraerts 2003; Tummers et al. 2005; Gries &

Stefanowitsch 2006; Gibbs 2007). Over the last few years, these methods have thrived, relying on increasingly complex statistical tools (Glynn 2010; Gries 2011; Perek 2015).

3 Collecting and analyzing corpus data in usage-based approaches to linguistic investigation relies on one major assumption, namely that the corpus is representative of the linguistic phenomenon under scrutiny. But, of course, corpus representativeness itself is a construct, both a theoretical (Halliday 2005) and a methodological one (Leech 2006; Habert 2010): language corpora are tools constructed by linguists, and their structural limitations constrain and condition the validity of linguistic findings. Data gathering from a corpus and theorizing are not separate activities: the textual instance is valued as a window on the linguistic system. This is especially true of spoken corpora: any form of transcription operates a drastic selection on the original spoken material, and as such transcription already constitutes a form of theorization (Ochs 1979).

4 Multiple theoretical, methodological and practical questions are raised by the issue of corpus representativeness, which, although by no means new, continues to draw attention in current research (see, for instance, Gray, Egbert & Biber 2017; Egbert, Biber & Gray forthcoming). A first question is whether corpus representativeness is achievable at all, and if so, how we can identify the relevant criteria to decide whether a corpus represents language use. If corpus representativeness cannot be fully gauged empirically, researchers must remain aware of the extent to which representativeness relies on intuition. Since a corpus cannot realistically be representative of all features of language use, one can wonder how bias in sampling can be addressed. A related issue concerned with corpus design is to what extent representativeness necessarily entails balance. More broadly, the question remains open as to whether the design of a corpus can be totally free from any form of theorization, i.e. whether "pure corpora" can exist.

5 Some answers to these questions have already been provided. Indeed, the representativeness of written corpora may rely on a variety of features, namely variability, sampling and balance. According to Biber (1993: 244), "[r]epresentativeness refers to the extent to which a sample includes the full range of variability in a population." Variability can be defined as the interaction between situational (e.g. format, setting, author, addressee, purposes, topics) and linguistic, distributional parameters (e.g. frequencies of word classes). Sampling is usually based on extralinguistic (sociological, demographic) criteria (Crowdy 1993). Balance, i.e. a proportion of sampled elements that reflects their frequency in the targeted language, is claimed to characterize some corpora (e.g. the Brown Corpus (Francis & Kucera 1979) and the Lancaster-Oslo-Bergen corpus (Johansson et al. 1978)), though it is not a prerequisite.

6 Although increasingly larger corpora, including monitor corpora, can be compiled from the Web (Baroni et al. 2009), large size is not necessarily a priority. As Fillmore put it, "I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way" (Fillmore 1992). "Big is beautiful" in the realm of corpora is, perhaps, a "delusion" (Svartvik 1992: 10). Large corpora are often presented as an ideal but, in practice, "small" corpora can go a long way in such domains as gesture studies (Debras 2018), English language teaching (Ghadessy, Henry & Roseberry 2001), the study of metaphors (Cameron and Deignan 2003) and dialectology (Hollmann & Siewierska 2007; Boas & Schuchard 2012), among others.

Parallel corpora, i.e. collections of original texts and their translations in one or more languages, are particularly useful in areas of research such as contrastive linguistics, translation studies and computational linguistics (Kenning 2010), but their alleged lack of representativeness has called for inventive ways of using them (Nádvorníková 2017).

7 In the area of spoken corpora, collecting data that represents the variability of the multiple dimensions of speech (phonology and phonetics, prosody, gesture) remains a challenge today. Collecting, transcribing, annotating and analyzing data, is a slow, sometimes complicated, task. Although phonological and prosodic annotations can be partially systematized (Bertrand et al. 2008), technological advances are yet to be made in the automatic recognition of speech and gesture in interactional contexts. Automatic motion capture technologies for gesture research are promising (Priesters & Mittelberg 2013: Guez et al. 2013), but so far little advanced. As part of initiatives such as the TGIR Huma-Num Multi-Com–CORLI Consortium, multimodality researchers collaborate to develop collective harmonized practices of collection, transcription and archiving of spoken corpora. Overall, even if many advances are yet to be made in the construction of representative spoken corpora, the field is making fast progress.

8 Given the multiplicity of issues raised by corpus representativeness and the limited space of a journal's special issue, the contributions assembled here are not intended to give a comprehensive overview of the topic in linguistics today. Rather, they allow to address a number of specific issues raised across various approaches and types of data and in relation to different research goals within linguistic traditions grounded in or having close theoretical and methodological links with cognitive linguistics. All the contributions stem from presentations given at the AFLiCo JET workshop "Corpora and representativeness", which was held on May 3-4, 2018 at the Université Paris Nanterre.

9 One major issue is the articulation between intuitive and corpus-based approaches to language study. Traditionally, usage-based models have sought to ground linguistic analysis in the observation of naturally occurring data, which, they argue, offers a more representative picture of language than does introspective evidence (see Lemmens's arguments on why only a usage-based model can rise to the challenge of representativeness). Interestingly, the habitual dichotomy between the two approaches is softened at several points in the thematic issue. Corpus methodologies are not necessarily meant to eliminate intuitions altogether but rather to provide support for them and, from there, for the construction of linguistic theories (see Ranger for the argument). For the corpus linguist, intuition may prove to be particularly useful in detecting corpus oddities (see Egan's conclusion on this point). Rocking the usage-based linguistics boat, Newmeyer goes as far as to claim that data drawn from big corpora yield essentially the same results as intuition-based data.

10 The fundamental rationale for relying on corpora is the unique possibility they offer of using quantitative data in linguistic analysis. In usage-based linguistics, frequencies of occurrences and co-occurrences of forms are key to elaborating theoretical models. This is not to say that qualitative analyses should be abandoned. In this special issue, linguists working in various theoretical frameworks stress the importance of integrating quantitative and qualitative approaches. While the necessary back-and-forth methodological movement between frequencies and specific contexts of use is either explicitly addressed (see in particular Egan's article) or exemplified in practice throughout the usage-based contributions, approaches not traditionally based on usage and/or data quantification – enunciative theory, generative grammar – seize the

opportunity to use quantitative tools to complement qualitative research (see Ranger's and Newmeyer's contributions, respectively).

11 A second issue addressed in the present collection of articles is corpus size. In specific avenues of research, notably the study of grammatical structures on the basis of conversational corpora, "big" can be argued to be indeed "beautiful" (see Newmeyer's argument). But a number of disadvantages can also be pointed out. First, as mentioned by Lemmens, without a fair amount of knowledge of sophisticated statistical tools, it is difficult, if not impossible, to make sense of large volumes of data, so that big or "mega" corpora in the end may turn against the researcher. In other cases, big corpora may simply be no more efficient than corpora of more reasonable size, since there is a point at which the saturation of new information is reached during corpus construction (see Parisse's article for the demonstration). Finally, a more crucial aspect than size, as it emerges from a number of articles, is sampling. A corpus may be small but more representative of a language, variety or register than larger ones if sampling is based on systematic, linguistically-motivated decisions rather than convenience or some principle of authority, as was perhaps often the case with first-generation corpora. In the debate over whether sampling should aim at representing the full range of texts produced (i.e. production-based sampling) or the diversity of high-impact texts (i.e. reception-based sampling), the authors addressing the question explicitly or implicitly advocate production, each stressing a different reason. Sampling according to production rather than reception acts as a safeguard against including in the corpus a lot of intentionally unusual material which is recognized as such by the language user and has arguably no influence on their productive linguistic system (see this point developed in Egan's article). Only a production-based sampling method enables the corpus compiler to gauge and represent the diversity of a given genre or register (see Perrez et al.'s argument). Specifically, using the production criterion ensures that the corpus does not over-represent features of individual speakers at the expense of others (see Grieve-Smith's evaluation of the French corpus Frantext regarding this issue).

12 Finally, there runs throughout this special issue the tacit agreement that complete representativeness of a language might never be achieved. As a consequence, perhaps, most contributors choose to build or carry out their analyses on corpora focused on some specific type of language production, for which corpus samples may represent the target population more exhaustively: sign language, language acquisition, political discourse, spontaneous conversation, fictional spoken language and performance texts. Even then, representing those types of language use is not without its challenges. Issues specifically related to spoken and signed language corpora are discussed in a number of contributions. Given the complexity of this type of unscripted data, one may wonder to what extent it may serve as a basis for the study of grammar, for instance (see Newmeyer's concluding remarks). In this special issue, two different ways of handling such complexity are presented. The first one is to turn it into an object of study in itself and seek solutions for how to best represent this complexity. In the field of sign languages, this means creating the most accurate tools for the transcription, visualization and searchability of data (see Boutet et al. for an understanding of the issue). The second one is to develop an adaptation strategy and rely on data drawn from fictional spoken language and performance texts as the closest representation of speech (cf. Terry's and Grieve-Smith's articles). In terms of corpus construction and use, this nonetheless involves two major challenges, as the corpus should be representative of the performance

genre itself, which in turn, should be representative of the spontaneous spoken interaction the genre is taken to reproduce, despite its inherent planned and editable nature. Arguably, although fictional spoken language and performance texts are an approximation of on-line speech production, the gap between the representation and the reality represented may be significantly reduced through adequate sampling techniques (see in particular Grieve-Smith on this point).

13 Although the many echoes between the various contributions suggested more than one way of structuring the special issue, we have decided to group them under three sections:

14 I. Representativeness in (target-specific) corpus construction

15 II. Addressing representativeness through case studies

16 III. Theorizing corpus representativeness

17 The first section, dedicated to representativeness in (target-specific) corpus construction, includes three papers written by Boutet et al., Parisse and Perrez et al. respectively. The first contribution by Dominique Boutet et al. presents the typefont Typannot, a transcription system dedicated to the annotation of sign languages. Unlike vocal languages, which are monolinear in nature, sign languages are characterized by multilinearity, since meaning is expressed simultaneously with various distinct articulators (e.g. hands, face, body). Other existing annotation systems remain limited: they either fail to provide information on form altogether, are cumbersome when describing language signs on the basis of form, or fail to account for the multilinearity of sign languages. Representing sign language in graphic form is hence a major challenge for sign language corpus developers. With Typannot, Boutet et al. take up that challenge. This modular typographic system aims to integrate three levels of information: the parameter (handshape, orientation, location, movement, facial expression), the components of the parameter, and the characteristics of each component, so as to provide a fine-grained description of the form of language signs for optimal representativeness of the data. Typannot's typefaces are created on the basis of the four underlying principles of scriptability, readability genericity, and modularity: they are meant to be easy to write and read, as well as to account for both low-level and concatenated information. Typannot stems from an interdisciplinary research involving linguists, designers and developers, and aims at being transferable to all existing sign languages.

18 Christophe Parisse's paper contributes to research on dense corpora by assessing the optimal size of a longitudinal dense corpus. As he explains, dense longitudinal corpora aim to be as representative as possible of a child's language development. They allow tracing the development of linguistic knowledge based on numerous samples of child's speech and input over relatively short periods of time. Since the compilation of dense corpora is time-consuming, identifying the optimal minimum size of this type of corpus is crucial. To do so, Parisse relies on two measures: word-metric, used to represent the development of lexical knowledge, and bigram-metric, used to represent the development of basic syntactic knowledge over time. Measuring word-metric and bigram-metric allows him to determine the smallest number of sessions to be included in a longitudinal dense corpus, in order to provide enough data to predict the use of a word or a bigram. His results show that the quality of both word and bigram coverage obtained in about 40 sessions is high enough to study language acquisition. Of course, this result does not undermine the value of very large dense corpora, but shows that high-quality

research can be conducted on smaller dense corpora. His paper therefore sets a useful practical upper limit in the area of dense corpus construction.

19 The question at the heart of Julien Perrez et al.'s article is the following: how is the category of political discourse (or should be) defined? How homogeneous is it? Linguistics studies of political discourse bear mostly on productions by political elites such as presidential or parliamentary debates, presidential addresses or public speeches, and tend to leave aside other forms such as media or citizen discourses. Their article aims to define the genre of political discourse, based on both extralinguistic and linguistic features. First, they apply the bibliometric method PRISMA (borrowed from the political sciences) on a sample of 172 scientific articles from the Scopus database, so as to map out what types of discourse have been categorized as political in linguistic research over the past twenty years, and to identify their extralinguistic features (e.g. type of actors, materials, themes, geographical origin). Second, they assess how consistent the notion of political discourse is from a linguistic point of view. Using multidimensional register analysis (Biber & Conrad 2009), the authors study the formal linguistic features of three subtypes of political discourse (parliamentary debates, televised debates and citizen corpora) so as to assess similarities between their textual registers. Results show great divergence, suggesting that political discourse is best defined as an abstract generic category for a range of different registers whose linguistic features are sensitive to the situational context of use.

20 The second section of this special issue addresses representativeness through three case studies proposed by Grieve-Smith, Terry and Ranger, respectively. In the first of these contributions, Angus Grieve-Smith, driven by the challenge of how to best represent the way people used to talk in 19th-century France, decides to turn to the language of the theater as one of the closest genres to spoken conversation. A corpus of theatrical productions can be considered to be representative of spoken language, he argues, providing adequate sampling methods are used. Grieve-Smith shows that the 'principle of authority' behind the sampling of the drama section of FRANTEXT results in the under-representation of some authors and a general bias towards formal language features. The author then presents his ongoing compilation project, the Digital Parisian Stage corpus, for which he uses as a sample frame an exhaustive list of every play that premièred in Paris in the nineteenth century. The representativeness of FRANTEXT and the Digital Parisian Stage corpus are then compared through a case study of the various syntactic realizations of negation. The results show that the Digital Parisian Stage corpus provides a more accurate picture of the language used over the period, including recent grammatical innovations.

21 Starting from the general assumption that the language of TV series is a polished, but truthful version of naturally occurring conversation, Adeline Terry's paper deals with the following research questions: are the metaphors in TV series representative of those that can be found in naturally occurring conversation? Are the main source domains used similar in naturally occurring conversation and in TV series? To answer them, she compares results drawn from a corpus of TV series and results of previous studies on metaphor in non-fictional corpora. Major differences stand out: compared to naturally occurring conversation, metaphors in the TV series corpus are more often creative and/or extended, are more used for humorous functions, and are used for characterization, dramatization, and aesthetic purposes. Terry also highlights the multifunctionality of metaphors in the TV series corpus. Differences in the use of metaphors between real-life

conversation and TV series discourse suggest that they constitute distinct genres, and hence that TV series discourse is a genre that should be studied for itself, and not as a representative of real-life conversational discourse.

22 In the last contribution of the section, Graham Ranger brings together the methodological strengths of corpus linguistics with the robust theoretical tools of the French Théorie des Opérations Prédicatives et Énonciatives in a case study of the English marker 'along'. The author starts with a critical appraisal of the status of linguistic data in the enunciative framework. Often drawn from a single genre, realistic fiction, linguistic evidence is traditionally submitted to manipulations about which intuitive acceptability judgements are made. According to Ranger, this methodology goes against the enunciative approach's commitment to study language in the diversity of its naturally occurring manifestations. The author then proposes a corpus study of 'along' using the data of the British National Corpus. Using statistical measures of occurrences and co-occurrences of terms, he identifies four contextual configurations, which are mapped with four main values – spatial, temporal, subjective and argumentative – for the marker. In each value, 'along' marks a dynamic identification between a locatum and a locator, construed as an unbounded, sequentially ordered space. The specificity of the marker is highlighted through a comparison with the compound form 'alongside', a closely related but semantically distinct marker, as evidenced by corpus data.

23 The third and final section of the special issue focuses on theorizing corpus representativeness. It includes a paper by Egan, and a tandem paper by Newmeyer and Lemmens. Thomas Egan's contribution addresses some practical and theoretical issues of representativeness in the design and use of corpora. The author discusses the ideal composition of a written corpus, which, he argues, should be based on production and avoid text types likely to misrepresent actual language use (e.g. grammar books, poetry, historical fiction). He then turns to multilingual corpora, for which he presents methods to obtain maximally representative data and avoid translation effects. Specifically, he shows the benefits of two models: the (expandable) four-text model and the three-text model. In the four-text model, the researcher makes use of bidirectional corpora, which contain original texts and translations in each of the languages represented. In the three-text model, linguistic analysis is based on the comparison of productions in two target languages, while the source language functions as tertia comparationis.

24 In the last contribution, Frederick Newmeyer and Maarten Lemmens use the original format of a tandem paper to bring into discussion generative grammar and usage-based cognitive linguistics, two traditionally contrasted theoretical frameworks, on the issue of corpus data and representativeness. The article starts with Newmeyer's demonstration of the limits of small conversational corpora when used to build theoretical models of grammar. The author goes on to reaffirm the relevance of introspective data in grammatical theory, since, he argues, they provide essentially the same insight into grammatical structure than data obtained from big corpora. Newmeyer's conclusions prompt Lemmens to enlarge the perspective and engage in a defence of usage-based linguistics over intuition-based approaches, highlighting the "complex and all-pervasive role of frequency with respect to the structure of grammar" and using the opportunity to give an updated overview of the field. The arguments proposed by the two authors in the initial position and response statements give rise to genuine discussion in Newmeyer's rejoinder and Lemmens's final acknowledgment. Closer connections than usually

assumed are suggested between the two theoretical approaches, in particular with respect to their underlying assumptions and goals.

25 We hope that the reader will find answers as well as more food for thought in relation to corpus representativeness in this special issue. On a final note, we would like to thank the anonymous reviewers for their detailed feedback and insightful suggestions. We are also grateful to the authors for the care they took in incorporating the remarks in their paper revisions. The quality of the special issue has been enhanced as a result.

BIBLIOGRAPHY

Baroni, Marco et al. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bertrand, Roxane, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, et al. 2008. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. Traitement Automatique des Langues, *ATALA*, 49(3). 105-134.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Boas, Hans C. & Sarah Schuchard. 2012. A corpus-based analysis of preterite usage in Texas German. *Proceedings of the 34th Annual Meeting of the Berkeley Linguistics Society.* Berkeley, CA: Berkeley Linguistics Society.

Bybee, Joan. 1985. *Morphology: A study on the relation between meaning and form.* Amsterdam: John Benjamins.

Bybee, Joan. 2010. *Language, usage and cognition.* Cambridge: Cambridge University Press.

Cameron, Lynne & Alice Deignan. 2003. Combining Large and Small Corpora to Investigate Tuning Devices Around Metaphor in Spoken Discourse. *Metaphor and Symbol* 18(3). 149-160.

Chomsky, Noam. 1957. *Syntactic structures.* The Hague: Mouton.

Crowdy, Steve. 1993. Spoken Corpus Design. *Literary and Linguistic Computing* 8(4). 259-265.

Debras, Camille. 2018. Petits et grands corpus en analyse linguistique des gestes. *Corpus* 18|2018, mis en ligne le 09 juillet 2018. URL : http://journals.openedition.org/corpus/3287

Egbert, Jesse, Douglas Biber & Bethany Gray. (Forthcoming). *Designing and evaluating language corpora.* Cambridge: Cambridge University Press.

Fanego, Teresa. 2004. Is Cognitive grammar a usage-based model? Towards a realistic account of English sentential complements, *Miscelánea. A Journal of English and American Studies* 29. 23-58. Special Issue on Language and Linguistics containing the plenary lectures from ESSE7 (Seventh

International Conference of the European Society for the Study of English), University of Zaragoza, 8-12 September 2004.

Fillmore, Charles J. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". In Svartvik, Jan (ed.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*, 35-60. Berlin/ New York: Mouton de Gruyter.

Francis, W. Nelson & Henry Kučera (1979). Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers. Department of Linguistics. Brown University. URL: http://www.hit.uib.no/icame/brown/bcm.html.

Geeraerts, Dirk. 2003. 'Usage-based' implies 'variational': On the inevitability of Cognitive Sociolinguistics. Paper presented at the 8th International Cognitive Linguistics Conference 2003, Logrono, Spain, July 20-15.

Geeraerts, Dirk, Gitte Kristiansen & Yves Peirsman (eds.). 2010. *Advances in Cognitive Sociolinguistics.* Berlin / New York: De Gruyter Mouton.

Ghadessy, Mohsen, Alex Henry & Robert L. Roseberry. 2001. *Small Corpus Studies and ELT.* Amsterdam: John Benjamins.

Gibbs, Raymond W. 2007. Why cognitive linguists should care more about empirical methods. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*, 2-18. Amsterdam: John Benjamins.

Givón, Talmy. 1979. *On understanding grammar.* New York: Academic Press.

Glynn, Dylan. 2010. Corpus-driven cognitive semantics. an introduction to the field. In Dylan Glynn & Kerstin Fischer (eds.), *Corpus-Driven Cognitive Semantics. Quantitative approaches*, 1-42. Berlin / New York: Mouton de Gruyter.

Gray, Bethany, Jesse Egbert & Douglas Biber. 2017. "Exploring methods for evaluating corpus representativeness." Paper presented at the Corpus Linguistics International Conference 2017. Birmingham, UK.

Gries, Stefan Th. & Anatol Stefanowitsch (eds.). 2006. Corpora in Cognitive Linguistics. *Corpus-Based Approaches to Syntax and Lexis.* Berlin / Boston: De Gruyter Mouton.

Gries, Stefan. 2011. Corpus data in usage-based linguistics. What's the right degree of granularity for the analysis of argument structure constructions? In Mario Brdar, Stefan Gries & Milena Žic Fuchs (eds.), *Cognitive Linguistics Convergence and Expansion*, 237-256. Amsterdam and Philadelphia: John Benjamins.

Halliday, Michael A. K. 2005. Computational and quantitative studies, volume 6. In *The collected works of M. A. K. Halliday*. Hong Kong: Continuum.

Hollmann, Willem B. & Anna Siewierska. 2007. A construction grammar account of possessive constructions in Lancashire dialect : some advantages and challenges. *English Language and Linguistics* 11(2). 407-424. DOI: 10.1017/S1360674307002304.

Hopper, Paul. 1987. Emergent grammar. In Jon Aske, Natasha Beery, Laura Michaelis, & Hana Filip (eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 139-157. Berkeley, CA: Berkeley Linguistics Society.

Johansson, Stig, Geoffrey Leech, & Helen Goodluck. 1978. Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Department of English. University of Oslo. URL: http://clu.uni.no/icame/manuals/LOB/INDEX.HTM.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them ?. In Anne O'Keefe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*, 487-500. London : Routledge

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Vol. I: Theoretical Prerequisites.* Stanford: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Vol. II: Descriptive application.* Stanford: Stanford University Press.

Nádvorníková, Olga. 2017. Pièges méthodologiques des corpus parallèles et comment les éviter, Corela [Online], HS-21 | 2017, Online since 20 February 2017, http://corela.revues.org/4810 ; DOI : 10.4000/corela.4810

Ochs, Elinor. 1979. Transcription as Theory. In Ochs, Elinor & Bambi Schieffelin (eds.), *Developmental Pragmatics*, 43-72. New York: Academic Press.

Perek, Florent. 2015. *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: John Benjamins.

Priesters, Matthias A. & Irene Mittelberg. 2013. Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study. *Proceedings of the Tilburg Gesture Research Meeting (TiGeR)*, June 19-21, 2013.

Saussure, Ferdinand (de). 1964 (1916). *Cours de Linguistique Générale*. Paris: Payot.

Svartvik, Jan. 1992. Corpus linguistics comes of age. In Svartvik, Jan (ed.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*, 7-16. Berlin / New York: Mouton de Gruyter.

Svartvik, Jan. 2007. Corpus linguistics 25+ years on. In Facchinetti, Roberta (ed.), *Corpus Linguistics 25 years on*, 11-25. Amsterdam: Rodopi.

Tummers, J., Kris Heylen, & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). 225-261.

NOTES

1. "Cognitive Linguistics before and after the empirical turn" was the main theme of the 2016 AFLiCo JET Workshop organized by Guillaume Desagulier at the Université Paris Nanterre.

AUTHORS

SOPHIE RAINERI

Université Paris Nanterre, CREA (EA370)/GReG

CAMILLE DEBRAS

Université Paris Nanterre, CREA (EA370)/GReG