RiCL Research in Corpus Linguistics

Review of Blanco, Marta, Hella Olbertz and Victoria Vázquez Rozas eds. 2019. *Corpus y Construcciones: Perspectivas Hispánicas*. (*Verba: Anexo* 79). Santiago de Compostela: Universidade de Santiago de Compostela. ISBN: 978-8-417-59587-6. https://dx.doi.org/10.15304/9788417595876

Miriam Thegel
Uppsala University / Sweden & Université catholique de Louvain / Belgium

This edited volume, with a special focus on corpus-based research on grammatical constructions and corpus compilation, is the outcome of the scientific event *Corpus y Construcciones: Perspectivas Hispánicas*, which was held at the University of Santiago de Compostela in November 2018 and hosted by the research group *Gramática del Español*. Apart from an introductory chapter by the editors, Marta Blanco, Hella Olbertz and Victoria Vázquez Rozas, who present a background to the current focus of the mentioned research group as well as a summary of the chapters in the volume, the book is divided into two parts, each comprising five contributions. Whereas the first part consists of synchronic and diachronic research studies based on corpus data, with theoretical implications for studies of language change in general, the second part is mainly oriented towards practical and applied issues considering corpus design, morphosyntactic tagging and the use of corpora for didactic purposes.

The first chapter, entitled "Grammars in contact in a bilingual corpus," is written by Rena Torres Cacoullos and Catherine E. Travis. Their aim is to test the hypothesis of grammatical convergence which states that bilingualism leads to change in at least one of the languages in contact, usually the minority language. While such changes have been documented at a lexical and phonetic level, the authors state that accounts of morphosyntactic changes are scant. Therefore, Torres Cacoullos and Travis focus on the frequency of three morphosyntactic phenomena drawing on data from the *New Mexico Spanish-English Bilingual Corpus* (NMSEB). The corpus consists of recorded

sociolinguistic interviews in Spanish and English of speakers who belong to an established bilingual community in northern New Mexico. Three structures with different uses and distribution in the two languages are investigated, namely, the use of the progressive construction (*estar* + V-*ndo* vs. *be* + V-*ing*), the use of subjunctive vs. indicative in subordinated clauses and the frequency of pronominal subjects. The authors explore whether these bilingual speakers show more similarities between their two languages in the use of these constructions than their monolingual counterparts do, which would be a sign of grammatical convergence, but no such conclusion could be drawn. The results show that the bilingual speakers maintain two independent grammars in Spanish and English, which points towards linguistic continuity rather than language change.

However, although the bilingual and monolingual varieties show similar patterns in general, the findings indicate that the frequency of the subjunctive in negated sentences is much lower among the bilingual speakers. A more in-depth discussion of this difference in comparison to the other constructions studied would perhaps have been of interest to the readers. Nevertheless, Torres Cacoullos and Travis present a robust study, whose results have theoretical implications beyond the particular languages under examination. It will certainly be enlightening for scholars interested in language contact and bilingualism in general.

The second chapter, by Anton Granvik, is called "On the origins of the shell noun construction in Spanish." The shell noun construction is a schematic entity that may appear in four different syntactic patterns, namely i) N *de* + infinitive, ii) N *de que* + clause, iii) N *que* + clause and iv) N *ser* + clause, and where the abstract head noun (N) encapsulates the information expressed in the complement. The aim of the author is to explore if the shell noun construction was already used in the medieval period or is a later development and which were the first nouns with this function. Furthermore, Granvik seeks to answer how the grammatical function of the construction is related to the textual one, on the one hand, and to different nouns and time periods, on the other. His data has been extracted from the *Corpus del Nuevo Diccionario Histórico del Español* and is analyzed combining quantitative and qualitative methods.

Based on three formal features, namely, the presence of a determiner (definite or indefinite article), the syntactic function of the noun and the kind of element (verb, preposition, etc.) on which the noun depends, Granvik is able to classify the uses of nine

analyzed nouns (*causa*, *condición*, *convicción*, *esperanza*, *idea*, *noticia*, *ocasión*, *señal* and *sospecha*) as a) typical, b) less typical and c) marginal. The findings demonstrate that there are typical shell noun uses already in the medieval period and that these typical uses increase over time. Additionally, after having discussed in detail less prototypical cases of encapsulation, the author draws the conclusion that, in order to classify as a shell noun construction, there either has to be an identity relation between the shell noun and the shell complement or the shell has to function as a link between two discursive elements.

Granvik provides a thorough analysis of the shell noun construction, taking both its diachronic development and its current state into account. He highlights the heterogenic nature of the construction in his discussion of atypical cases, where he convincingly shows the strengths of combining a quantitative perspective with a detailed textual analysis.

In the third chapter, entitled "The contribution of Corpus Linguistics to the analysis of a prepositional construction with *entre* 'amid'," Belén López Meirama and Carmen Mellado Blanco analyze the constructional idiom [*entre* + $N_{plural/bodily}$] which until now has passed by practically unnoticed in Spanish grammars. Inspired by the framework of Construction Grammar (Goldberg 2006), the authors carry out a careful corpus study of the construction, based on almost 1,200 cases from the *Corpus del Español del Siglo XXI* (CORPES XXI), in which its main syntactic, semantic and pragmatic properties are analyzed. The construction functions as a second predicate, referring to an action that occurs simultaneously with the main verb, and that often can be paraphrased with a gerund (*Me lo dijo entre sollozos/sollozando* 'S/he told me this amid sobs/sobbing', p. 86).

Furthermore, it is shown that this construction is partially schematic, allowing three different patterns, namely i) *entre* + bare noun, ii) *entre* + noun + modifier and iii) *entre* + noun + coordinated clause, among which the bare noun construction is the prototypical one. The noun filling the N slot usually originates from the semantic fields of communication or bodily expression, for instance *risas* 'laughter', *lágrimas* 'tears' and *aplausos* 'aplauses', whereas the main verb tends to be related to communication or, to a lesser extent, displacement. The authors successfully apply Construction Grammar in the discussion of their findings, relating the variability of the constructional frame and the high number of different nouns occurring in it to a high degree of productivity and entrenchment. The study is a valuable example of how phraseological units can be

examined through a corpus-based approach, which can serve as an inspiration for future studies of other prepositional patterns.

In the fourth chapter, called "On the concept of *behavioral profile*," Inmaculada Mas Álvarez reviews the previous definitions of 'behavioral profile' found in the bibliography and discusses recent initiatives inspired by this concept. The idea of the behavioral profile arose in the 1990s as a result of the growing work of corpus-based linguistics. This new methodology permitted the extraction of a large number of concordances used, for instance, to establish a relationship between the different meanings of a lemma and its most frequent syntactic patterns. Mas Álvarez cites Hanks (1996: 79), among the first to elaborate on the concept, who defines his own work on the behavioral profile as "an attempt to encapsulate its established norms (patterns of usage) on the basis of analysis of a body of evidence of actual usage (a corpus)." Thereafter, she describes the pioneer project *Base de Datos Sintácticos del Español Actual* (BDS), an initiative taken to describe the syntactic usage patterns of Spanish verbs. This initiative later evolved into the project *Base de Datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (ADESSE), where semantic information was added to the syntactic patterns in the BDS, establishing a categorization of different verb classes and allowing comparison between similar lexical elements.

As an example of a recent project inspired by the concept of the behavioral profile, Mas Álvarez mentions the tool *SketchEngine* (Kilgarriff *et al*. 2014), where the user can perform lemmatized queries in several corpora at the same time, in order to find the most frequent (as well as low frequency) constructions and meanings. Even though these new tools are valuable for linguistic research, the author concludes that the automatized functions still remain insufficient in certain aspects, for instance, syntactic tagging that causes errors in the analyses provided. Therefore, she states, it is still necessary to combine new hardware, based on quantitative methods, with a detailed manual analysis.

The chapter stands out from the rest in this first section of the volume in that it does not offer an empirical study on constructions based on corpus research. Nevertheless, it brings attention to the concept of behavioral profile and gives an overview of both the possibilities and the challenges that corpus projects may entail.

The fifth chapter, with the title "Pragmatic functions in Brazilian Portuguese: A Functional Discourse Grammar account," is written by Hella Olbertz. From the perspective of Functional Discourse Grammar, the author describes the evolution of the

innovative function of topic in Portuguese spoken in Brazil, in which a personal pronoun is added immediately after the nominal subject, although being syntactically redundant, to mark a new, contrastive or general topic. The author explains this process of pragmaticalization as an effect of two recent changes in the pronominal system, leading to an overload of functions for third person singular: i) the substitution of second person plural (*tu*) by third person singular *(você*) and the introduction of the nominal phrase *a gente* to refer to first person plural, but with a verb morphology of third person singular. This, Olbertz states, has led to a generalization of the use of third person subject pronouns to contexts where they are not syntactically necessary, but where they perform the pragmatic function of topic.

Through a qualitative analysis of spoken data from the corpus *Iboruna*, the author convincingly discusses the innovative function in Brazilian Portuguese in different contexts and compares it with the functions of topic and focus in the European variants of Portuguese and Spanish based on the oral corpora *PRESEEA de Alcalá de Henares*, *C-Oral-Rom* and *Português Falado*. She concludes that whereas European Spanish can express focus by placing the subject in clause final position, it does not have a morphosyntactic way of expressing topic. In contrast, the variant of Portuguese spoken in Europe lacks both pragmatic functions, while the Brazilian Portuguese has developed the innovative function of topic, described in this chapter. Olbertz offers a well-written contribution that is highly interesting, in that it concerns a recent evolution emergent in spoken corpora, and she successfully manages to relate it to patterns in other Romance languages.

In chapter six, called “The *Reference Corpus of Present-day Galician* (CORGA): Composition, codification, POS-tagging and use,” Eva María Domínguez Noya, María Sol López Martínez and Francisco Mario Barcala Rodríguez summarize almost thirty years of work with the largest corpus of contemporary Galician. The paper begins with a background of the normativization process of Galician in the 1980s and 1990s, which eventually culminated in the compilation of the *Corpus de Referencia do Galego Actual* (CORGA). The corpus covers the period from 1975 to the present and contains roughly 40 million words, mainly from written sources, but also 25 hours of orthographic transcriptions of radio programs. Due to the inclusion of texts published before the standardization of the written language, CORGA shows a rich graphic and morphologic variation. The authors point out that this variation has caused problems in the automatic

tagging of lemmas, while being at the same time a testimony of the historical language contact between Galician and other languages spoken in the region. The process of creating CORGA is carefully described, by presenting the digitalization, codification and the POS-tagging of the corpus texts. A smaller training corpus, analyzed manually, has laid the foundation for the automatized tagging tool, which is now able to recognize many contracted forms, typical of the Galician language, although there is improvement to be made. The authors end with a section on how the online application can be used and what information there is to retrieve from the corpus. For readers interested in corpus design and corpus composition, this chapter provides a valuable contribution where all the necessary steps of creating a corpus are thoroughly presented.

The seventh chapter, by Elisa Fernández Rei and Xosé Luís Regueira and with the title "CORILIGA: A corpus for the study of variation and linguistic change in spoken Galician," also concerns the construction of a corpus of Galician, namely the *Corpus Oral Informatizado da Linga Galega* (CORILIGA), which consists of spoken data covering a period from 1965 until present. The compilation process of the corpus started in the early 2010s with the incorporation of previous collections of spoken data, mostly recordings done within traditional works of dialectology that consisted of informal spoken Galician from rural areas. In order to improve the representativity of the corpus and to facilitate the possibility to study the evolution of spoken Galician over time, efforts have been made to include more genres and registers, such as urban varieties, data from Galician youth and formal discourse. The tool *ELAN* (Wittenburg *et al.* 2006) was used to transcribe and annotate the recordings and the program *Freeling* (Padró and Stanilovsky 2012) was adapted to Galician to carry out a morphosyntactic tagging. The authors stress the fact that the corpus project was developed in close collaboration with speech technologists, resulting in an improvement of tools for automatic speech recognition in Galician.

Furthermore, Fernández Rei and Regueira provide a detailed description of the online interface of CORILIGA, where queries can be specified according to genre, topic, year of recording and kind of speaker, to name only a few criteria for selection. Although the corpus is not yet openly accessible to the public, the authors point out that research based on data from CORILIGA has been published recently, concerning political discourse and politeness strategies, among other topics. Fernández Rei and Regueira conclude that with a public access in the near future, any user will be able to continue

studying the variation and language change in Galician, both on matters related to sociolinguistics and to morphosyntax.

Chapter eight, entitled "Problems encountered in the morphosyntactic tagging of the ESLORA corpus," is written by Eva María Domínguez Noya, Raquel Rivas Cabanelas, María Paula Santalla del Río and Rebeca Villapol Baltar. The authors stress the importance of taking into consideration elements proper of spoken language when creating and tagging an oral corpus. They provide a description of the morphosyntactic annotation of the *Corpus para el Estudio del Español Oral* (ESLORA), a corpus of Spanish spoken in Galicia, highlighting the tagging problems that arise when oral language is analyzed based on models of written language. In particular, the authors emphasize the need to use a training corpus that represents the same kind of registers and genres as the larger corpus, in order to create a tagging tool with a high degree of accuracy. In order to tag ESLORA, four tools were developed, namely, the tagger XIADA,[1] based on Galician but adapted to Spanish, a catalogue with all the morphosyntactic labels used, a dictionary where lemmas were connected to these morphosyntactic labels, and a training corpus to teach the tagger how to annotate the larger corpus. The training corpus, after being automatically tagged, was manually revised to solve problems in the tagging process. Labels adapted for elements proper of spoken language, such as pauses, interrupted words, lengthened syllables and communicative noises were added. Other linguistic elements mentioned by the authors as particularly challenging for the tagging process are discourse markers, set phrases and idioms, which frequently play a different role in spoken language than written language and where previous models for tagging remain unsatisfactory. For instance, the discourse marker *hombre* has been classified as an interjection in ESLORA, whereas it is labeled as a noun in the CORPES XXI by the *Real Academia Española*. All the examples given in the chapter strengthen its main argument that, in order to build a useful corpus with a reliable tagging function, special care has to be put to the composition of the training corpus as well as the manual tagging, especially for spoken language with all its particularities.

The ninth chapter, with the title "The *Corpus de Aprendices de Español* (CAES) and its applications to the teaching/learning of Spanish as a foreign language" is written by Ignacio Palacios Martínez, Francisco Mario Barcala Rodríguez and Guillermo Rojo.

[1] http://corpus.cirp.es/xiada

The first part of the contribution focuses on the compilation process and the design of CAES, whereas the second part concerns the practical applications of the corpus for research as well as for teaching and learning Spanish as a foreign language. CAES consists of short texts written by learners of Spanish ranging from the levels A1 to C1 and with six different first languages (L1). The tagger *FreeLing* (Padró and Stanilovsky 2012) was used for the morphosyntactic labeling and lemmatization of the linguistic elements and was combined with a manual revision in which labels were added and errors generated in the automatic tagging process were detected.

The authors highlight the fact that there are few studies published on the practical applications of learner corpora and that the number of these corpora is still limited. To show the usefulness of CAES, they present several areas where the corpus could offer valuable contributions. For instance, the authors state it could be used to compare the influence that different L1s have on the learning of Spanish. Moreover, queries can be made in CAES to conduct error analysis, such as to check to which extent the learners use the verb pairs *ser*/*estar* correctly. Finally, Palacios Martínez, Barcala and Rojo show how samples from the corpus can be used in the classroom, serving as inspiration for learning activities for the students and giving them the opportunity to analyze authentic examples of interlanguage. The chapter is well-written and reveals a strong engagement for didactic matters; however, several of the issues raised by the authors are not new ideas and, therefore, fall short. For example, the possible effect that the L1 of the learner has on L2 or L3 has been studied since several decades (e.g. Ellis 1994). Furthermore, it is well known that the verb pair *ser/estar* causes problems for many learners and that this is confirmed in the present corpus too is not a truly new finding. It might thus have been preferrable to put the main focus on the new contributions of the study, for instance, to give more examples of how a learner corpus can be used in the classroom to enrich the repertoire of learning activities and to expand the final sections on possible applications for curriculum design and evaluation.

The final and tenth chapter of the volume, written by Irene Doval and Tomás Jiménez, is entitled “Multifunctionality of parallel corpora, exemplified by German-Spanish corpus PaGeS.” The aim is to describe the composition of the corpus, the process of segmentation and sentence alignment, the possibilities for searching through and visualizing the corpus data and, finally, to give an outline of the steps to be taken in the future. PaGeS offers new opportunities for research of comparative linguistics, translation

studies and the teaching and learning of foreign languages in that it consists of genres, namely narrative texts such as novels and essays, that are not represented in previous parallel corpora. Moreover, the texts in PaGeS are of high quality, being published by well-established publishing houses. The corpus consists of 25 million words from 140 titles (original language and translations combined) and covers a time period from 1960 until present. Additional material, such as texts from the European Parliament and TED talks, can also be found in the corpus. The authors refer to the process of sentence alignment as a crucial step of the corpus construction where every source text and its translation are segmented on a sentence level and aligned to each other. Problems in the alignment process may occur and, as stated by Doval and Jiménez, fiction usually supposes a major challenge compared to other genres, due to possible differences between the original version and the translation. For instance, the translator can choose to omit or add sentences or change their structure, of which clear examples are offered in the chapter. Therefore, the automatic alignment has to be complemented with a manual revision. As for the interface of PaGeS, the corpus aims to be user-friendly, fast and to offer three different search levels to satisfy the needs of various types of users. The authors end the chapter by mentioning future steps to be implemented. Word alignment is planned for, which may be particularly useful for students in translation and foreign language studies. Furthermore, two other parallel corpora are prepared by the research group, namely Spanish-Dutch and Spanish-Chinese.

This volume offers a multifaceted picture of what can be done in corpus-driven research, and it highlights the necessary steps to be taken when creating a new corpus as well as the challenges that may arise in the process, not least related to automatic tagging that in general requires a manual revision in order to achieve the desired accuracy. The editors have managed to bring together scholars working on a rich variety of topics and projects, which, although they may seem widely different at first glance, are united by the fact that they all offer new insights on corpus linguistics. Several chapters provide inspirational accounts of how corpora can be used to study issues related to language change, dealing with grammaticalization (Chapters 2 and 3) and pragmaticalization (Chapter 5), while others offer thorough descriptions of corpus tagging and discuss useful tools in the compilation process (Chapters 6–10). Other topics that unite several of the contributions are didactic applications of corpora (Chapter 9 and 10) and the particularities of oral corpora (Chapters 1, 5, 7 and 8).

In summary, this collection of texts is a testimony of the usefulness of corpora for linguistic research as well as for language teaching and training, and as such it offers a valuable contribution to a wide variety of readers, both researchers, corpus builders, teachers and students.

REFERENCES

Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Goldberg, Adele, E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1/1: 75–98.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36.

Padró, Lluís and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association, 2473–2479.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of* the *Fifth International Conference on Language Resources and Evaluation*. Genoa: Language Resources Association,1556–1559.

Reviewed by

Miriam Thegel
Uppsala University
Department of Modern Languages, Romance Languages
Engelska parken, Thunbergsvägen 3L
Box 636
751 26 Uppsala
Sweden
e-mail: miriam.thegel@moderna.uu.se