Miguel Calderón Campos & Gael Vaamonde

Introduction. Corpus Linguistics in the Era of Big and Rich Data: Methodological Perspectives on Spanish and Portuguese

The current state of play in language corpora is one where the range of resources is wide, many of them available online for free. The resulting scenario is unique, in that language research can nowadays be undertaken from manifold vantage points, and evidence of a given case can be easily and efficiently retrieved from various languages, language varieties, genres, periods, or registers. The progress in computing, with an ever-growing data storage and processing capacity, resulted in increasingly larger corpora, so corpus size, from one to thousands million forms, becomes one more choice opportunity in language corpus research.

At least three broad types of language resources can be distinguished: small corpora, general or reference corpora, and mega corpora or web corpora, each with its own advantages and disadvantages (Rojo 2021: 77–81). The former type usually amounts to some million words and is designed for specific research, or for the analysis of specific language cases. Small in size, these corpora have richer and more fine-grained encoding and annotation compared to general corpora, so they are well-suited for highly accurate data retrieval; however, their limited size may weaken their representativeness and their value as a basis for generalization. By contrast, general or reference corpora are intended to be as representative as possible for a language or for a language variety, so they contain evidence of various text types and geographical language varieties. They typically amount to several hundred million forms, are highly versatile and, thus, can support a wide range of language research projects; on the downside, they may not prove sufficient for research on specific areas, or for highly specialized projects. Finally, mega corpora rise to thousands of million words, usually built on the immense amount of data uploaded to the World Wide Web. These resources are highly powerful for large-scale quantitative analysis, but they are often impaired by design flaws, especially as regards balance and representativeness, and their annotation is often defective, so retrieval of geographical or typological data may be inaccurate.

Corpus selection is according to the specific research objectives and to the end-user's methodological and analytical needs (Hunston 2008: 166). Small corpora are exploited qualitatively, as research relies on rich data, i.e. on texts furnished with a highly detailed (extra-)linguistic metadata array. By contrast, mega corpora are exploited quantitatively, as research uses big data for statistical analysis of massive datasets or for retrieval of unprecedented or rare cases that are

https://doi.org/10.1515/9783110781465-001

unavailable from other corpora. General corpora lie in between, and stand out for their balance and representativeness. Both Spanish and Portuguese can be researched with a range of resources that run the gamut from small and tidy corpora to big and messy ones (for an overview, see Davies 2008; Vanderschueren & Mendes 2015; Enghels, Vanderschueren & Bouzouita 2015).

The *TenTen* corpora, available via *Sketch Engine* (Kilgarriff *et al.* 2003), stand out among the latter type of corpora, especially the ca. 20,000-million-word *Spanish Web corpus* (esTenTen), and the *Portuguese Web Corpus* (ptTenTen), with over 12 thousand million words (Kilgarriff & Renau 2013; Kilgarriff *et al.* 2014). Some corpora of these two languages hosted on the corpus management platform developed by Mark Davies at Brigham Young University (BYU) are within this group too, e.g. the 7.6-billion-word *NOW* (News on the Web) *Corpus of Spanish* (Davies 2018a), and the ca. 2-billion-word *Web / Dialects Corpus of Spanish* (Davies 2016a). The Portuguese counterparts are in the range of one billion words for each of the two corpora (Davies 2018b, 2016b). Language research can also be based on big data, by use of numberless digital repositories available, e.g. *Google Books* (via *Google Books Ngram Viewer*), and of digital libraries, newspaper and otherwise, like *Biblioteca Digital Hispánica*, *Biblioteca Digital de la Real Academia Española*, or *Biblioteca Virtual Miguel de Cervantes* for Spanish, and *Biblioteca Nacional Digital de Portugal*, *Biblioteca Nacional Digital de Brasil*, or *Biblioteca Digital Camões* for Portuguese.

Several reference corpora have been compiled by the Spanish Royal Academy, synchronic and diachronic, like *Corpus de Referencia del Español Actual* (CREA) and *Corpus del Español del Siglo XXI* (CORPES XXI) among the former, and *Corpus Diacrónico del Español* (CORDE) and *Corpus del actual Diccionario Histórico* (CDH) among the latter. Their sizes range from 150 to 350 million forms. For Portuguese, the *Corpus de Referência do Português Contemporâneo* (CRPC), with over 300 million words, and the *Carolina* corpus (*Corpus Geral do Português Brasileiro Contemporâneo*), with over 800 million words, are worth citing (Bacelar do Nascimento *et al.* 2014; Sturzeneker et al. 2022). Other reference corpora are the ca. 100-million-word *Corpus del Español Genre / Historical*, and the 45-million-word *Corpus do Português Genre / Historical* (Davies 2002; Davies & Ferreira 2006).

Small corpora arise from specific research needs. As a result, the scope covered by small corpora is hard to summarize as a short list. Time-consuming building stages are frequent here, so historical corpora (especially if based on strict selection criteria and/or thorough philological editions), spoken corpora (where data processing requires the transcription of metalinguistic properties with varying degrees of detail), and parsed corpora (where manual postediting is needed according to the detail and comprehensiveness of the syntactic annotation) are small corpora. A list of the above for illustration purposes and with no pretence

to comprehensiveness could mention Spanish or Portuguese corpora like *Corpus Hispánico y Americano en la Red: Textos Antiguos* (CHARTA), *Oralia Diacrónica del Español* (Calderón Campos y García-Godoy 2019-), *Post Scriptum* (CLUL 2014), *Corpus de Textos Antigos* (CTA), and *Corpus Histórico do Português Tycho Brahe* (Galves *et al* 2017), among historical corpora, *Corpus Oral y Sonoro del Español Rural (COSER)* and *Português Falado* (Bacelar do Nascimento 2001) among spoken corpora, and *Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (ADESSE), *AnCora-ES* (Taulé *et al.*), and the CORDIAL-SIN/Synapse projects (Martins 1999–2022; Magro & Vaamonde 2022), among parsed corpora, to mention only some.

This volume brings together research on Spanish and Portuguese based on corpora of various sizes and designs. The volume thus presents various approaches to corpus linguistics and discusses the advantages and disadvantages of rich data and big data for linguistic research. The volume has three parts according to the corpus used for each research project: Part I is about small, *enriched* diachronic corpora, Part II is about *alternative* corpora for research on diatopic or diaphasic microvariation, and Part III is about canonical reference corpora.

Part I *Small, Tidy and Rich Diachronic Corpora: The PS-ES and the ODE Corpora* consists of three papers based on the specialized corpora *Post Scriptum* (PS-ES) and *Oralia Diacrónica del Español* (ODE). The three papers lay emphasis on the need for small, specialized corpora enriched with text data and (extra-)linguistic annotation in addition to the annotation available from large reference corpora (CDH, CORDE, CdEhist). Part II *The COSER Corpus and Newspaper Digital Libraries as Alternative Data Sources for Research on Rural and Informal Varieties* discusses further the limitations of reference corpora and the search for alternative sources of dialectal and of informal language data. The former case turns to Project COSER-UD, an annotation scheme of *Corpus Oral y Sonoro del Español Rural* (COSER), and the latter turns to data from the *macrocorpus Hemeroteca Digital de la Biblioteca Nacional de España* (HD). Part III *Exploiting Portuguese Reference Corpora: The CdP and the CRPC Corpora* presents three papers based on two of the main reference corpora of Portuguese, and thus prove the relevance of large, well-balanced, annotated corpora.

In Linguistics, but especially in Historical Linguistics and in Dialectology, "representativeness may be more important than the sheer size of the corpus" (Brezina 2018: 221). Large corpora often offer little variation in terms of text types (few semiformal or informal samples) and in terms of speaker social or regional provenance (few samples by semi-educated speakers or by speakers of geographical provenance other than of standard language). Large corpora also rely on text editions produced according to various scientific criteria and, sometimes, according to dissimilar quality standards and rigour. The first two parts of the volume thus focus on data sources representative of variation and change, where the

quality of language data and metadata is given priority over the quantitative approach of big data, as in personal correspondence (PS-ES), goods inventories (ODE), regional press (HD), or rural questionnaires (COSER).

The volume opens with **Vaamonde**'s "Not so Big Data: Assessing Two Small Specialized Corpora for the Study of Historical Variation in Spanish", a crucial chapter for the question of the value of small corpora (one to two million words) in the era of Big Data. The answer is illustrated with the use of two corpora of Modern Spanish (16th–19th centuries) available within TEITOK, namely the PS-ES corpus and the ODE corpus. The former is a corpus of personal correspondence, and the second is a corpus of goods inventories and witness testimonies in trials. For Vaamonde, these are useful corpora, if their low size is supplemented with descriptive contents for language description, with quality editions, or with digital resources unavailable from larger corpora, e.g. selection of text samples close to orality, accurate chronological and geographical metadata, or accurate transcripts and representation of the original source. TEITOK has the added value of allowing corpus access in various modes: palaeographical diplomatic edition, modern edition according to current standards, and facsimile edition. The use of a CQL-supported search engine allows quality retrieval of linguistic data, which is further enhanced with manual postediting of morphosyntactic tagging and lemmatization.

The chapter by **Inmaculada González Sopeña** "Language Corpora and Lexical Arabisms in the Digital Age" examines the technical profile of the ODE corpus and discusses its properties with regard to historical research on Arabisms in Spanish. The focus is on how TEITOK manages exhaustive retrieval of formal variants of Arabisms (e.g. *guadameçi*, *guadameçil*, *guadamezi*, *guadamesil*, *guadameçiles*, etc.) starting out from the lemma or modern version (*guadamecí* 'garnished leather').

In the last chapter of Part I, "Corpus size and tagging: Methodological Strategies for Research on the History of Diminutives *-ito*, *-illo* and *-ico*", **Miguel Calderón Campos** proves how various corpus types may be necessary for the thorough account of language change. The current use of *-ito*, *-illo* and *-ico* is examined according to evidence obtained from big data sources (EsEuTenTen11), specifically PS-ES for the analysis of the evolution of diminutives in informal Spanish between the 16th and the 19th centuries, and PS-ES, CDH and ODE for a description of standard and dialectal usage in the 18th c.

Part II consists of two chapters about the need for further resources than just reference corpora in language research. The chapter by **Miriam Bouzouita**, **Johnatah E. Bonilla** & **Rosa Lilia Segundo Díaz** "Gaming for Dialects: Creating an Annotated and Parsed Corpus of European Spanish Dialects through GWAPs" presents project COSER-UD, intended to spawn a morpho-syntactically annotated and parsed version of COSER (*Corpus Oral y Sonoro del Español Rural*). Led by

Inés Fernández-Ordóñez, the COSER corpus, a collection of transcripts of spoken Spanish from the mainland and from the Balearic and the Canary Islands was started in the 1990s to meet the need for means for research on diatopic microvation in the morphology and syntax of European Spanish. COSER-UD is intended to furnish the base corpus with a new layer of curated morphosyntactic tagging for enhanced data retrieval accuracy and efficiency. Three Games with a Purpose (GWAPs) were designed for tagging enhancement, whereby users can revise automatic tagging.

The second chapter of Part II "Big Data and Lexical History: Digital Newspaper Libraries in Spanish Diachronic Research" by **María Teresa García-Godoy**, surveys the history of the diminutive adverb *cabalito* 'exactamente', a colloquial form rarely attested in the CDH. The data retrieved from the digital newspaper section of the National Library amount to 317 attestations, by contrast with 27 attestations in the CDH. The paper underlines the relevance of press language for attestation of educated neologisms and for research on 18th c. and 19th c. informal lexical innovations. Ultimately, the paper argues for more journalistic samples in reference corpora, and exposes methodological issues of unencoded and non-annotated mega corpora, when it comes to the identification of the origin and spread of linguistic innovations that may become dialectal particularities of European Spanish.

Part III consists of three chapters about the use of reference corpora for research on Portuguese. The first chapter, "The Reference Corpus of Contemporary Portuguese: Corpus Design and Case Study on Discourse Markers", is by **Amália Mendes**, and discusses the *Corpus de Referência do Português Contemporâneo* (CRPC), an annotated corpus that is currently over 300 million words. This chapter serves a two-fold purpose: i) it overviews this digital resource with useful, up-to-date information about its design, contents, linguistic annotation, and online access, both for the written and for the spoken component (available via CQPWeb and TEITOK, respectively); and ii) presents the case study of the discourse marker *claro* 'naturally, of course', where the most frequent collocations are researched based on 739 and 20,180 occurrences retrieved from the spoken and written subcorpora, respectively. The results obtained show various properties of the form *claro* in each mode and thus contribute towards the standing knowledge of this discourse marker in contemporary European Portuguese.

Anton Granvik's "On the Origins of the Shell Noun Construction in Portuguese" researches the diachronic evolution of the so-called shell noun construction in Portuguese from the 13th c. to the 20th c. Based on previous research, Granvik studies nine shell nouns (*mercê* 'mercy', *razão* 'reason', *vontade* 'will', *sinal* 'sign', *caso* 'case', *temor* 'fear', *facto* 'fact', *ideia* 'idea', and *questão* 'question'). The initial 9,362 attestations within a range of syntactic structures available from *Corpus do Português* (CdP) are filtered to a well-balanced sample of 1,446

cases. After manual annotation, statistical analysis shows that the shell noun construction is attested in Portuguese as early as the 13th c. and 14th c., that it is used increasingly frequently over time, and that its evolution reveals major change patterns, syntactically and lexically. The results of a mixed-effects logistic regression analysis also shows a statistically significant relation between the shell nouns' encapsulation and the syntactic structure, the type of noun, and the context.

Finally, the chapter by **Katharina Gerhalter** "*Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio*. A Corpus-Based Approach to Topicalized Infinitives in Portuguese" examines the so-called topicalized infinitive construction in contemporary Portuguese. Despite the difficulties in data retrieval of this type of examples (for their low frequency and for the varying distance between the infinitive and the inflected verb form), the author gathers 60 occurrences from the 20th c. section of the *Corpus do Português* (CdP). The dataset obtained shows that the topicalized infinitive construction is highly productive in Portuguese, in that a wide range of verbs may be used and that it is associated with informal spoken language. The data also show that the construction is more frequent in European than in Brazilian Portuguese, even though this may admittedly be as a result of a bias in the corpus design. The quantitative analysis is followed by a qualitative analysis of the contextual and discursive properties of this construction according to the 'question under discussion' (or 'QUD') framework.

We would also like to express our gratitude to the reviewers for their valuable suggestions that helped improve the chapters in this volume.

Bibliography

ADESSE = García-Miguel, José María (dir.). *ADESSE: Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*. <http://adesse.uvigo.es/>.

Bacelar do Nascimento, Maria Fernanda (coord.) (2001): *Português falado, documentos autênticos, gravações audio com transcrições alinhadas*. Lisboa, Centro de Linguística da Universidade de Lisboa e Instituto Camões [cederrón]. <https://catalog.elra.info/en-us/repository/browse/ELRA-S0345/>.

Bacelar do Nascimento, Maria Fernanda, Amália Mendes, Sandra Antunes and Luísa Pereira (2014): "The Reference Corpus of Contemporary Portuguese and related resources", in Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira (eds.). *Working with Portuguese Corpora*. London: Bloomsbury, pp. 237–256.

Brezina, Vaclav (2018): *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

CDH = Real Academia Española (2013): *Corpus del Diccionario histórico de la lengua española (CDH)*. <https://apps.rae.es/CNDHE>.

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <https://www.corpuscharta.es/>.

CLUL (ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt>.

CORDE = Real Academia Española: Banco de datos (CORDE). *Corpus diacrónico del español.* <http://www.rae.es>.

CORPES XXI = Real Academia Española: Banco de datos (CORPES XXI). *Corpus del Español del Siglo XXI (CORPES)*. <http://www.rae.es>.

COSER = Fernández-Ordóñez, Inés (dir.): *Corpus Oral y Sonoro del Español Rural*. <http://www.corpusrural.es/>.

CREA = Real Academia Española: Banco de datos (CREA). *Corpus de referencia del español actual.* <http://www.rae.es>.

CTA = Sobral, Cristina (coord.): *Corpus de Textos Antigos em português até 1525*. <http://teitok.clul.ul.pt/teitok/cta/>.

Davies, Mark (2002): *Corpus del Español: Historical/Genres*. <http://www.corpusdelespanol.org/hist-gen/>.

Davies, Mark (2008): "New directions in Spanish and Portuguese corpus linguistics", in *Studies in Hispanic and Lusophone linguistics*, 1(1), pp. 149–186.

Davies, Mark (2016a): *Corpus del Español: Web/Dialects*. <http://www.corpusdelespanol.org/web-dial/>.

Davies, Mark (2016b): *Corpus do Português: Web/Dialects*. <http://www.corpusdoportugues.org/web-dial/>.

Davies, Mark (2018a): *Corpus del Español: NOW*. <http://www.corpusdelespanol.org/now>.

Davies, Mark (2018b): *Corpus do Português: NOW*. <http://www.corpusdoportugues.org/now>

Davies, Mark and Michael Ferreira (2006): *Corpus do Português: Historical Genres*. <http://www.corpusdoportugues.org/hist-gen/>.

Enghels, Renata, Clara Vanderschueren and Miriam Bouzouita (2015): "Panorama de los corpus y textos del español peninsular contemporáneo", in Maria Iliescu and Eugeen Roegiest (eds.). *Manuel des antologies, corpus et textes romans*. Berlin/Boston: De Gruyter, pp. 147–170.

Galves, Charlotte, Aroldo Leal de Andrade and Pablo Faria (2017): *Tycho Brahe Parsed Corpus of Historical Portuguese*. <https://www.tycho.iel.unicamp.br/corpus/>.

Hunston, Susan. (2008): "Collection strategies and design decisions", in Anke Lüdeling and Merja Kytö (ed.). *Corpus linguistics: An international handbook*. *Vol. 1*. Berlin: De Gruyter, pp. 154–168.

Kilgarriff, Adam and Irene Renau (2013): "esTenTen, a vast web corpus of Peninsular and American Spanish", in *Procedia-Social and Behavioral Sciences*, 95, pp. 12–19.

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

Kilgarriff, Adam, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha and Pete Whitelock (2014): "PtTenTen: a corpus for Portuguese lexicography", in Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira (eds.). *Working with Portuguese Corpora*. London: Bloomsbury, pp. 111–130.

Magro, Catarina and Gael Vaamonde (coords.) (2022): *SynAPse – The Syntactic Atlas of European Portuguese*. Lisboa: Centro de Linguística da Universidade de Lisboa. <http://corpora.ugr.es/synapse/>.

Martins, Ana Maria (coord.) (1999–2022): *CORDIAL-SIN: Corpus Dialetal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects*. Lisboa: Centro de Linguística da Universidade de Lisboa. <https://cordialsin.wordpress.com/>.

ODE = Calderón Campos, Miguel and María Teresa García-Godoy (dirs.) (2019-present): *Oralia Diacrónica del Español (ODE)*. <http://corpora.ugr.es/ode>.

Rojo, Guillermo (2021): *Introducción a la lingüística de corpus en español*. London/New York: Routledge.

Sturzeneker, Mariana Lourenço, Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte and Cristiane Namiuti (2022): "Carolina's Methodology: building a large corpus with provenance and typology information", in Cassia Trojahn, Maria José Finatto, Renata Vieira and Valéria de Paiva (eds.). *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing* (2nd DHandNLP 2022). CEUR-WS, Vol. 3128. <https://ceur-ws.org/Vol-3128/>.

Taulé, Mariona, M. Antònia Martí and Marta Recasens (2008): "AnCora: Multilevel Annotated Corpora for Catalan and Spanish", in Nicoletta Calzolari *et al*. (eds.). *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC'2008)*. Marrakesh, pp. 96–101.

Vanderschueren, Clara and Amália Mendes (2015): "Panorama de los corpus y textos del portugués europeo contemporáneo", in Maria Iliescu and Eugeen Roegiest (eds.). *Manuel des antologies, corpus et textes romans*. Berlin/Boston: De Gruyter, pp. 58–80.