

DE GRUYTER

LINGUISTIC CORPORA AND BIG DATA IN SPANISH AND PORTUGUESE

*Edited by Miguel Calderón Campos and
Gael Vaamonde*

DIGITAL HUMANITIES AND BIG DATA
IN IBERO-AMERICA

DE
G

Linguistic Corpora and Big Data in Spanish and Portuguese

Humanidades Digitales y Big Data en Iberoamérica

Digital Humanities and Big Data in Ibero-America

Editado por / Edited by
Ana Gallego Cuiñas y / and Azucena González Blanco

Consejo científico / Advisory board

Franco Moretti (Stanford University)

Anthony Cascardi (UC Berkeley)

Carolina Gainza (Universidad Diego Portales)

Juana Licerias (Universidad de Ottawa)

Virginia Bertoloti (Universidad de la República)

José Antonio Pérez Tapias (Universidad de Granada)

Mayte García Godoy (Universidad de Granada)

Miguel Calderón Campos (Universidad de Granada)

Cristóbal Lozano (Universidad de Granada)

Daniel Torres Salinas (Universidad de Granada)

Volumen / Volume 4

Linguistic Corpora and Big Data in Spanish and Portuguese



Edited by
Miguel Calderón Campos and Gael Vaamonde

DE GRUYTER

Esta publicación es resultado de la Unidad Científica de Excelencia “Iber-Lab. Crítica, Lenguas y Culturas en Iberoamérica” (Ref. UCE2018-04) de la Universidad de Granada

ISBN 978-3-11-078145-8

e-ISBN (PDF) 978-3-11-078146-5

e-ISBN (EPUB) 978-3-11-078152-6

DOI <https://doi.org/10.1515/9783110781465>



This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. For details go to <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

Library of Congress Control Number: 2024942975

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at <http://dnb.dnb.de>.

© 2024 the author(s), editing © 2024 Miguel Calderón Campos and Gael Vaamonde, published by Walter de Gruyter GmbH, Berlin/Boston
The book is published open access at www.degruyter.com.

Cover image: as creative atelier/DigitalVision Vectors/Getty Images

Typesetting: Integra Software Services Pvt. Ltd.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Miguel Calderón Campos & Gael Vaamonde

Introduction. Corpus Linguistics in the Era of Big and Rich Data: Methodological Perspectives on Spanish and Portuguese — 1

I Small, Tidy and Rich Diachronic Corpora: The PS-ES and the ODE Corpora

Gael Vaamonde

Not so Big Data: Assessing Two Small Specialized Corpora for the Study of Historical Variation in Spanish — 11

Inmaculada González Sopeña

Language Corpora and Lexical Arabisms in the Digital Age — 37

Miguel Calderón Campos

Corpus Size and Tagging: Methodological Strategies for Research on the History of Diminutives *-ito*, *-illo*, and *-ico* — 59

II The COSER Corpus and Newspaper Digital Libraries as Alternative Data Sources for Research on Rural and Informal Varieties

Miriam Bouzouita, Johnatan E. Bonilla & Rosa Lilia Segundo Díaz

Gaming for Dialects: Creating an Annotated and Parsed Corpus of European Spanish Dialects through GWAPs — 87

María Teresa García-Godoy

Big Data and Lexical History: Digital Newspaper Libraries in Spanish Diachronic Research — 113

III Exploiting Portuguese Reference Corpora: The CdP and the CRPC Corpora

Amália Mendes

The Reference Corpus of Contemporary Portuguese: Corpus Design and Case Study on Discourse Markers — 145

Anton Granvik

On the Origins of the Shell Noun Construction in Portuguese — 179

Katharina Gerhalter

Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio. A Corpus-Based Approach to Topicalized Infinitives in Portuguese — 207

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

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 microvariation 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 sub-corpora, 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 áudio 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/>>.
- Engiels, Renata, Clara Vanderschueren and Miriam Bouzouita (2015): “Panorama de los corpus y textos del español peninsular contemporáneo”, in Maria Iliescu and Eugene 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): *Oralía 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): “AnCorà: 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.

**I Small, Tidy and Rich Diachronic Corpora:
The PS-ES and the ODE Corpora**

Gael Vaamonde

Not so Big Data: Assessing Two Small Specialized Corpora for the Study of Historical Variation in Spanish

1 Introduction. Big Versus Rich Data?

It is a widely acknowledged fact that modern corpus linguistics has evolved hand in hand with computing. The release of the *Brown Corpus*, the first machine-readable corpus, in the 1960s is the starting point of a growing discipline that developed quickly due to advancements in new technologies and the highly improved processing and storage capacity of computers over the last 50 years.¹ The ability to process and store increasingly larger datasets can be used as a criterion for the identification of stages in the evolution of the compilation and availability of computerized corpora. By 1991, Leech considered at least three generations of corpora for English, mainly according to size:

At a basic level, the resurgence of corpus linguistics can be measured in terms of the increasing power of computers and of the exponentially increasing size of corpora, viewed simplistically as large bodies of computer-readable text. The Brown Corpus [. . .] can be thought of as a ‘first-generation’ corpus; its million-word bulk seemed vast by the standards of the earlier generation of corpus linguistics. But this size was massively surpassed by a ‘second generation’ of the 1980s represented by John Sinclair’s Birmingham Collection of English Text and the Longman/Lancaster English Language Corpus, which benefited from the newer technology of the KDEM optical character recognition device [. . .]. And perhaps the title ‘third generation’ may be given to those corpora, measured in hundreds of millions of words, almost all in commercial hands, exploiting the technologies of computer text processing (Leech 1991: 9–10).

Not long after Leech’s quotation, the growing tendency toward expanding corpora led to a substantially larger corpus: The World Wide Web made possible to

¹ See Francis (1992) and Kennedy (1998: 13–19) for a review of corpus linguistics before electronic corpora.

Note: This contribution has been realized in the framework of Grant PID2022-136256NB-I00, funded by MICIU/AEI/ 10.13039/501100011033 and by “ERDF/EU”. Also, it has been carried out within the framework of Grant C-HUM-038-UGR23 funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021–2027.

rapidly compile inexpensive databases of millions of words.² Online access to large datasets and the subsequent compilation of megacorpora certainly means another stage in the history of the quick development of corpus linguistics (Renouf 2007: 28; Tognini Bonelli 2010: 17).

At present, at least three types of corpora can be considered according to size. The first type is reference or general corpora, between one hundred and one thousand million words in size. They are intended to supply the most representative and well-balanced diachronic or synchronic sample of a language. The second type is megacorpora, amounting to several thousand million words in size, compiled by automatic download of online text. The third type is specialized or domain-specific corpora, built with texts of a specific genre, register, variety, or period. They are purpose-oriented and usually range between half a million and two million words.

The advantages and disadvantages of each type are easy to identify. Megacorpora are especially well-suited for rare phenomena or for overviews of the lexical and grammatical properties of contemporary Spanish. As they are based on data retrieved from the web, they are less representative, not so well-balanced, and less informative as regards metadata than reference corpora. By contrast, specialized corpora offer fine-grained, often manually checked annotation, and allow much more specific queries than reference corpora; the downside is that they do not have much text type variety and may not offer enough evidence of certain questions, especially if they are rare or not frequent. Additionally, they are costlier than larger corpora.

Corpora are thus intended to reach a compromise between two inversely proportional variables: size vs. granularity or, in other words, scope vs. depth:

The research or construction of a resource can grow in either breadth or depth. Given the usual limitations in time and effort, it becomes necessary to prioritize one over the other, inevitably resulting in the neglect of the other. Both are important; however, their greater or lesser relevance is more the result of a specific situation or need that arises at a certain moment than a general issue (Rojo 2010a: 1158).³

² This, in addition to the possibility of using the web as corpus (Kilgarrif & Grefenstette 2003).

³ My translation. The original reads: “La investigación o la construcción de un recurso puede crecer en extensión o en profundidad. Dadas las limitaciones habituales en tiempo y esfuerzo, es forzoso optar por privilegiar una de ellas, lo cual se convierte irremediabilmente en la preterición de la otra. Las dos son importantes, sin embargo, y su mayor o menor relevancia es más el resultado de una situación o necesidad concreta, que se presenta en un cierto momento, que de una cuestión general”.

Mair's (2006) separation *big and messy* vs. *small and tidy* corpora summarizes the —not necessarily incompatible— main features of each approach or method in this opposition when it comes to their actual use: for qualitative or for quantitative research. The same contrast is encapsulated in corpus linguistics by the opposition *big data* vs. *rich data*, a distinction borrowed from data science (Hiltunen, McVeigh & Säily 2017; Nevalainen, Suhr & Taavitsainen 2019). Reference corpora lie between these two extremes and stand out for their representative, well-balanced contents (Rojo 2016: 212–214, 2021: 77–81).

Research on contemporary Spanish relies on the three types of corpora mentioned above. Two examples of megacorpora are *esTenTen18* (Kilgarriff & Renau 2013) and *Corpus del Español NOW* (Davies 2018–). The former, available from *Sketch Engine*, contains ca. 17,000 million words; the latter, led by Mark Davies, is a 7,600 million-word corpus of online newspaper texts written between 2012 and 2019. Two obvious examples of reference corpora are the Spanish Royal Academy's *Corpus del Español Actual* (CREA) and *Corpus del Español del Siglo XXI* (CORPES XXI). At 160 and 400 million words respectively, they are freely available from the Academy's website. Finally, two examples of *small and tidy* corpora are the 1.5 million-word database *ADESSE* (Vaamonde *et al.* 2010) and the 500,000-word corpus *AnCora-ES* (Taulé *et al.* 2008). Both show several levels of syntactic, semantic and lexical tagging, and are highly suitable for research on the structure of verbal arguments in contemporary Spanish.

The above applies to historical linguistics with two major qualifications. First, no diachronic megacorpora, i.e. corpora of several thousand million words of past texts, are available, as far as we know. This is largely because past sources are not as easily available and accessible as present-day data, especially in the case of pre-19th c. data (Claridge 2008: 245–246). The above typology therefore becomes an opposition between reference and small corpora, when it comes to diachronic databases.

The second qualification is that the data available for historical corpora are, in general, of lower quality than present-day data. A number of references describe the many limitations of data-based diachronic research (Rissanen 1992; Kohnen 2007; Claridge 2008; Hernández-Campoy & Schilling 2012; Schneider 2013), namely random document availability, lacking evidence of spoken language, uncertain source authenticity, uncertain text dating, and limited availability of social and contextual metadata, among others. These well-known research constraints have been summarized under the term “bad-data problem” as follows:

Historical documents survive by chance, not by design, and the selection that is available is the product of an unpredictable series of historical accidents. The linguistic forms in such documents are often distinct from the vernacular of the writers, and instead reflect efforts

to capture a normative dialect that never was any speaker's native language. As a result, many documents are riddled with the effects of hypercorrection, dialect mixture, and scribal error. [. . .] Historical linguistics can then be thought of as the art of making the best use of bad data. The art is a highly developed one, but there are some limitations of the data that cannot be compensated for. Except for very recent times, no phonetic records are available for instrumental measurements. We usually know very little about the social position of the writers, and not much more about the social structure of the community (Labov 1994: 11).

This chapter explains how these challenges can be overcome successfully only if the corpus is compiled meticulously through every stage of data building so the data can be used to their full potential: from careful data selection to detailed tagging of textual and metatextual features and thorough philological editing. As all these are highly demanding in terms of time and effort, they are feasible only in smaller corpora, i.e. by priming *rich data* over *big data*.

Thus, it is argued here that a major effect of the *bad-data problem* is that *small and tidy* corpora become particularly relevant for historical linguistics, more so than in other linguistic disciplines. Further evidence in support of this claim is that most of the small Spanish corpora compiled in the past ten or fifteen years are historical corpora (Rojo 2021: 78).

2 Large Historical Corpora: CORDE, CdEhist and CDH

Three large diachronic Spanish corpora are available at present, namely the so-called general or reference corpora: *Corpus Diacrónico del Español* (CORDE), the diachronic component of *Corpus del Español* (CdEhist), and *Corpus del Diccionario Histórico de la Lengua española* (CDH). All three are freely available online and contain a range of texts across various genres, periods and geographical varieties spanning from the earliest Spanish texts to the mid- or late 20th c.

CORDE, the first diachronic reference corpus of Spanish, was first released in 1998 and, at ca. 250 million words,⁴ is now a static corpus. CdEhist, slightly over

4 The wordcount varies according to source: the RAE's website cites 250 million records («registros» (<https://www.rae.es/banco-de-datos/corde> [last accessed, 13/06/23]), while the CORDE manuals mention 125 million words (https://corpus.rae.es/ayuda_c.htm [last accessed, 13/06/23]). The contrast also shows in different pieces of research: Parodi (2008: 112) cites "180 million forms", Rojo (2012: 438) estimates "slightly over 260 million forms", and other studies raise the figure to 300 million words (Pascual & Domínguez 2009: 79; Rojo 2016: 201). Based on the frequency of certain highly frequent grammatical words, Davies (2009: 140) estimates between 220 and 240 million words. Our

100 million words, was released shortly afterwards, i.e. in 2001, and then revised in 2007. Finally, CDH was released in 2013. The latter consists of three main searchable subcorpora: The core CDH and two additional subcorpora, one for CORDE texts produced between the 12th c. and 1975, and one for CREA texts of the period 1975–2000. At 350 million words altogether, the CDH stands today as the largest machine-readable diachronic corpus of Spanish.

Table 1 shows the wordcounts and century coverage of the above corpora. The information of the CdEhist is from the corpus website.⁵ The information of the CDH results from *Nómina* queries available through its interface. CORDE only allows estimates, as its data source is not listed by century.⁶ The wordcount estimate is based on the average ratio between the occurrence of the most frequent forms, *de* and *que*, in CORDE and in CdEhist. Thus, *de* and *que* between 1701 and 1800 amount to 916,442 and 565,922 words in CORDE, respectively, and 594,470 and 404,205 in

Table 1: Wordcount estimate by century in CORDE, CdEhist and CDH.

Century	CORDE	CdEhist	CDH		
			Core CDH	12th c. –1975	1975–2000
<i>before 1300</i>	9,583,496	7,079,164	4,825,912	4,834,824	–
<i>1301–1400</i>	9,130,136	2,667,810	1,983,963	6,476,744	–
<i>1401–1500</i>	27,113,825	8,747,963	7,762,351	17,376,195	–
<i>1501–1600</i>	52,894,849	17,774,762	7,043,456	48,955,689	–
<i>1601–1700</i>	40,115,287	13,355,483	4,493,135	35,532,646	–
<i>1701–1800</i>	15,185,529	10,324,328	5,752,754	11,825,693	–
<i>1801–1900</i>	45,190,898	20,822,142	6,986,492	41,256,304	–
<i>1901–2000</i>	60,878,009	20,540,030	21,841,811	53,841,413	62,017,302
<i>after 2001</i>	–	–	1,341,392,	–	–
TOTAL	260,092,029	101,311,682	62,031,266	220,099,508	62,017,302

Source: CORDE, CdEhist and CDH (author's data)

estimate, based on the frequency of *de* and *que*, raises the figure to 260 million words (see Table 1). A recent publication (Sánchez Lancis 2022: 36) mentions 244 million words. The corpus size must have been revised several times between the first release and completion, but more precise information on the final wordcount is missing from the official source.

5 Data available from <https://www.corpusdelespanol.org/hist-gen/help/texts.asp> [last accessed, 13/06/23].

6 The CORDE manual only cites percentage distribution by period: Middle Ages (28%), the Golden Age (28%), and Contemporary Period (51%). A similar distribution, also compared with CdEhist, is available from Rojo (2021: 83). Molina Salinas & Sierra Martínez (2015: 313) cite CORDE and CREA relative frequencies for wordcounts and for sample texts by century, but they do not give absolute frequencies.

CdEhist. As the CdEhist wordcount for that period is 10,324,328 words, the CORDE size for the 18th c. must range between the two resulting ratios, i.e. 15,916,106 and 14,454,953. Their average ratio, namely 15,185,529, is shown in Table 1.

A reminder of the value of these corpora for diachronic research of Spanish is in order, especially in a paper on the value of specialized diachronic corpora and on the need for other courses of actions in corpus compilation. CORDE, probably the most well-known of the above corpora, stands out for the spectacular results obtained from remarkable efforts. The availability of a 100-million word diachronic corpus in the late 20th c. is “a fact of particular relevance to the research and a positive feature for the institutions that have funded its construction” (Rojo 2012: 439).⁷ It is worth noting that the diachronic corpora of English did not reach two million words at the time, even if English boasted the richest collection of electronic resources at the time, and still probably does.⁸ Nearly a decade after release, Sánchez Sánchez & Domínguez Cintas (2007: 138) described both CORDE and CREA together as “the most important resource for research on [Spanish]”.⁹ The above makes it easier to understand the influence of CORDE on our field’s growth, as it enabled and made available research on the diachronic evolution of Spanish morphology, syntax and lexis (Octavio de Toledo y Huerta 2019).

The release of CDH in 2013 improved on some of the least positive and worst rated features of CORDE, like the lack of linguistic annotation, the obsolete interface, or the distorted dates of mediaeval texts (Garachana & Artigas 2012; Rodríguez Molina & Octavio de Toledo y Huerta 2017). Finally, CdEhist is a befitting addition to the former two diachronic academic corpora. While tagging is poorer in CdEhist, the query interface allows more complex frequency data retrieval at several descriptive levels (cf. Davies 2009; Rojo 2010b). All in all, CORDE, CdEhist and CDH gather an extremely rich set of diachronic data —ca. 700 million words with some overlap— and are thus essential research tools for the history of Spanish.¹⁰

7 My translation. The original reads: “un hecho de especial relevancia para la investigación y un rasgo positivo para las instituciones que han financiado su construcción”.

8 Two well-known diachronic English corpora are the Helsinki Corpus and the ARCHER Corpus. Released in the 1990s, their sizes were 1.5 and 1.8 million words, respectively.

9 My translation. The original reads: “el recurso más importante del que se haya podido disponer jamás para el estudio de esta lengua”.

10 Rojo (2021: 127–266) presents a wide range of reference corpus exploitation possibilities for grammatical and lexical research of Spanish, both diachronic and present-day.

3 Small Specialized Historical Corpora

Both reference and specialized diachronic Spanish corpora are available at present. The latter have become more widely available and thus fortunately grant access to a range of electronic resources for diachronic research. With no pretence of exhaustiveness, the following corpora can be listed, all of them freely accessible on the web: *Biblia Medieval*, comprising Hebrew or Latin versions of the Bible and their translations into mediaeval Spanish; *Corpus de Documentos Españoles Anteriores a 1900* (CODEA+ 2022), archive fonds from before the 20th c.; *Corpus Léxico de Inventarios* (CorLexIn), goods inventories according to notarial documents of the 17th c.; *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM), American texts written between the 16th and the 19th centuries; *Corpus Documental de las Islas Canarias* (CORDICan), documents of the Canary islands produced between the 16th and the 19th centuries; *Corpus Hispánico y Americano en la Red: Textos Antiguos* (CHARTA), consisting in samples of various specialized diachronic corpora; *Old Spanish Textual Archive* (OSTA), mediaeval texts transcribed semipalaeographically by members of the *Hispanis Seminary of Medieval Studies*. Two more corpora can be added: The Spanish component of the corpus *Post Scriptum* (PS-ES), and the corpus *Oralia Diacrónica del Español*. Table 2 summarizes the main features of all the above corpora in descending order by wordcount at present:

Table 2: Main features of some specialized diachronic corpora of Spanish.

Corpus	Wordcount ¹¹	Timeframe	Corpus specialisation
OSTA	35,000,000	11th c.–16th c.	mediaeval texts
CORDIAM	14,452,855	16th c.–19th c.	American texts
<i>Biblia Medieval</i>	4,811,169	13th c.–15th c.	Spanish translations of the Bible
CODEA+ 2022	2,189,129	12th c.–19th c.	pre-20th c. texts
CorLexIn	1,599,464	17th c.	goods inventories
CHARTA	1,346,094	12th c.–19th c.	CHARTA edition guidelines
ODE	1,016,866	16th c.–19th c.	conceptual orality
PS-ES	987,390	16th c.–19th c.	personal letters
CORDICan	564,264	16th c.–19th c.	texts from the Canary Islands

As can be seen from Table 2, specialized diachronic corpora typically range between ca. 1 million and 5 million words. CORDIAM and OSTA are exceptions: A

¹¹ The wordcount is according to the information available from each corpus website by June 2023 or, when not available, according to the wordcount obtained from the query in question.

good part of the former consists in newspaper texts collected from online repositories or from literary editions, so the most demanding building stage, namely transcription, is substantially cut down; in the latter, the corpus contents are entirely transcripts of mediaeval texts. In this regard, the task undertaken by the *Hispanic Seminary of Medieval Studies* in the 1970s is worth of praise. OSTA relies heavily on their database of digital transcripts of manuscripts and incunabula produced between 1000 and 1600.

In any case, specialized corpora offer a substantially lower amount of data than reference corpora. The two corpora discussed here—the Spanish component of *Post Scriptum* (PS-ES) and the corpus *Oralia Diacrónica del Español* (ODE)— are slightly under one million words each and cover a shorter time span than the reference corpora CORDE, CdEhist and CDH (see Table 3). The contrast is thus ca. two million words vs. ca. 380 million words contained in the three diachronic reference corpora of the 16th to the 19th centuries (see Table 1). The difference is further illustrated by the fact that the CDH wordcount produced in the Philippines alone amounts as much as the PS-ES and the ODE’s wordcounts together. This means that the PS-ES and the ODE corpora amount to 0.5% of the data available from CORDE, CdEhist and CDH for the same period.

Table 3: Author’s wordcount of PS-ES and ODE by century.

Century	PS-ES	ODE
1501–1600	153,065	180,762
1601–1700	281,551	279,889
1701–1800	407,140	450,259
1801–1900	145,634	105,956
TOTAL	987,390	1,016,866

The question thus arises, what is the need for such small specialized diachronic corpora, especially as they are highly demanding projects, both timewise and otherwise. Van der Wal (2022) recently posed this question for English, a question that applies to Spanish equally well:

As historical sociolinguists we have to decide what our data are and determine from which sources data can be collected. Both large historical corpora and more specialised corpora are available for English and the compilation of digital corpora for other languages has increased in recent decades. Thus researchers may have a choice between either using existing digital corpora or compiling new specialized ones. Moreover, in a world of big data optimism, the idea may arise that the time-consuming compilation of specialised corpora is no longer needed (Van der Wal 2022: 335–336).

The answer obviously comes from the potential of *small and tidy* corpora in historical linguistics by contrast with the potential that reference corpora lack by nature. In this regard, a major point is the corpus builder's philological work, i.e. the possibility of editing the sample texts of the corpus. Small corpora (typically) do their own philologically detailed editing based on unique transcription protocols, always according to the original text's spellings and furnished with the manuscript's image. All the corpora listed in Table 2 abide by these in various degrees.

By contrast, reference corpora rely on the editions available at the time of their compilation. This practice, which Dollinger (2004) calls "philological outsourcing", builds from other authors' source editions and is, in this sense, "second level" work (Rojo 2010a):

A reference corpus is, by nature, a work that we can consider to be of a secondary level, meaning it's a resource that depends on the work previously done in various aspects. Those who construct a historical corpus, for example, depend on texts that have been edited previously, both in terms of the texts themselves and the adopted graphic conventions. In other words, one cannot expect someone intending to produce a general corpus, composed of thousands of texts and hundreds of millions of words, to also undertake the transcription of texts already published, edit those that have not been, solve textual problems, resolve gaps regarding authorship, composition dates, etc. (Rojo 2010a: 1157).¹²

The philological processing of text editions is one of several distinctive qualities of smaller diachronic corpora. Their advantages can be summarized as: i) a selection of quality sources supporting the corpus *reliability*; ii) text editing to address *faithfulness*; iii) easy access and verifiable data to prime *transparency*; and iv) encoding according to maximized *searchability*. Section 4 discusses six features of PS-ES and ODE arising from these four properties.

¹² My translation. The original reads: "Un corpus de referencia es, por naturaleza, un trabajo que podemos considerar de segundo nivel, esto es, un recurso que depende de la labor realizada previamente en muy diversos aspectos. Quien construye un corpus histórico, por ejemplo, depende de los textos que hayan sido editados con anterioridad, tanto en lo que se refiere a los propios textos como a las convenciones gráficas adoptadas. En otras palabras, no se puede pedir a quien pretende producir un corpus general, formado por miles de textos y cientos de millones de palabras, que acometa también la redacción de textos ya publicados, edite los que no lo han sido, solucione problemas textuales, resuelva incógnitas de autoría, fecha de composición, etc".

4 Advantages of Small Historical Corpora: PS-ES and ODE

The *Post Scriptum* corpus resulted from two research projects hosted by the Center of Linguistics of the University of Lisbon: *CARDS. Cartas Desconhecidas*, funded by the Foundation for Science and Technology (*Fundação para a Ciência e a Tecnologia*, FCT) from 2007 to 2009, and *P.S. Post Scriptum: A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain)*, funded by the European Research Council (ERC) from 2012 to 2017. The *Post Scriptum* corpus consists in two subcorpora slightly under 1 million words each, one of Spanish and one of Portuguese. The contents are ca. 2,500 personal letters each, written between the 16th c. and the first third of the 19th c. by authors of a range of social classes (see Vaamonde 2018).

The ODE corpus is the result of several research projects too, funded by national and regional agencies and hosted at the University of Granada from 2010 onwards. At ca. one million words, it contains slightly under 900 handwritten documents produced across Spain between the 16th and the 19th centuries. The documents can be classified as goods inventories, witness testimonies and, less significantly, surgery medical reports (see Vaamonde 2024).

These two corpora differ in kind in their contents, but they are discussed together here, because they share several features as regards their research purposes and their compilation methods. Thus, they share a selection criterion for samples close to orality, the coverage of about the same timeframe, an adequate profile for potential use in diachronic dialectology, similar editorial and linguistic annotation criteria, the same encoding language and standards (XML, TEI, EAGLES), the same online platform for dissemination and management (TEITOK) and, as a result, the same data retrieval system (CQP). Overall, they are compatible, complementary, and can be used jointly for research on certain linguistic questions.

4.1 Careful Selection of Speech-Related Sources

It is a well-known fact that historical linguistics, and especially historical pragmatics and sociolinguistics, needs access to past spoken data even if they must rely on written records as their data source. This contrast between what is desirable and what is available justifies the description of historical linguists' research as "paradoxical" (Oesterreicher 2004: 735), "frustrating" (Rissanen 1999: 188), or even "schizophrenic" (Cano Aguilar 1996: 376):

This situation has led Historical Linguistics to a behavior that we could describe as ‘schizophrenic’: The only *corpus* it can access, at least up to the current moment, is the written one [. . .], but linguists often disown it and try to find within the written material something that refers them beyond, to a communicative world that is hidden from them and of which writing barely reveals faint hints (Cano Aguilar 1996: 376).¹³

This disadvantage can be best minimized by focusing on past texts that are closest to the properties of spoken language, even if with due reservations and with the necessary qualifications. Diachronic linguists have thus turned since the 1990s to written documents where the signs of orality are best represented, e.g. personal letters —preferably by semiliterate letter-writers—, diaries, soldier stories, trial testimonies —preferably in direct speech—, and other poorly elaborated documents.¹⁴ The search and analysis of these and other text types has become a major subject of interest for the most recent research on Spanish:

This task, which endows the historian of language with an aura of discoverer or archaeologist, has attracted the attention of numerous Hispanists in recent decades, to the extent that the analysis of strategies of conceptional orality in ancient texts has become one of the most fruitful lines of research in modern philology (Del Rey Quesada 2020: 42).¹⁵

Still, these documents in varying degrees of closeness to communicative immediacy are rarely available in diachronic reference corpora of Spanish. Based on CORDE metadata, Rodríguez Puente (2018) lists at least four categories that may contain texts representative of spoken language in the corpus: i) letters and accounts of various kinds; ii) memories and diaries; iii) short and long prose drama; and iv) fictional prose, i.e. dialogues and miscellaneous modes. According to this author, the four categories amount to ca. 14 million words, i.e. 6% of the corpus contents (Rodríguez Puente 2018: 100–101). However, the view of these categories as representative of oral language is not without problems. As Rodríguez Puente notes, the CORDE query interface does not allow the use of filters in these catego-

13 My translation. The original reads: “Tal situación ha llevado a la Lingüística histórica a un comportamiento que podríamos calificar de ‘esquizofrénico’: El único *corpus* que puede manejar directamente, al menos hasta el momento actual, es el escrito [. . .], pero el lingüista suele renegar de él e intenta hallar en lo escrito algo que le remita más allá, a un mundo comunicativo que se le oculta y del que la escritura apenas le revela leves indicios”.

14 Classifications of text types close to orality are available in Oesterreicher (2004: 747–756), Culpeper & Kytö (2010: 18), or Schneider (2013: 61).

15 My translation. The original reads: “Esta tarea, que dota al investigador de la historia de la lengua de un halo de descubridor o de arqueólogo, ha ocupado la atención de numerosos hispanistas en las últimas décadas, de tal manera que el análisis de las estrategias de la oralidad conceptual en textos antiguos se ha convertido en una de las líneas de investigación más rentables de la moderna filología”.

ries, so it is not known, e.g. how many letters are personal correspondence or how many are not by literate writers. Also, personal diaries, in principle relevant as regards orality traits, are classified alongside memories and travel diaries and cannot be retrieved separately, if they are in the corpus at all. Finally, with respect to the literary representation of orality, it remains a feigned orality constructed by a single author, which limits its value as a testimony of authentic orality (Eberenz & De la Torre 2003: 22). The same limitation on oral language representation can be noticed in the CDH corpus, which is largely based on CORDE, and in CdEhist, where pre-20th c. samples “are still mainly formal texts” (Blas Arroyo 2019: 29).¹⁶

PS-ES and ODE fill this gap in part. *Post Scriptum* consists entirely in personal letters, mainly unpublished material by writers of various social backgrounds that has reached us as evidence filed in court records. Many are also by semiliterate writers or “mãos inábeis” (“unskilled hands”) (Marquilhas 2000: 235), who write unelaborately and relatively spontaneously about everyday events. The ODE corpus consists in three notarial text types where spoken language about ordinary events is also used: i) witness testimonies in court; ii) goods inventories for various purposes; and iii) detailed medical reports of injuries caused by attacks under trial. All three types of documents are written records of spoken testimonies by witnesses, surveyors, and doctors, and they often show informants’ or scribes’ vernacular language traits.

On the whole, the communicative immediacy of these corpora relatively often shows substandard dialectal or social variations that are rare in written, e.g. the use of *de que* for *que* (1a), undue use of *lo* (1b), undue use of *la* (1c), use of grammatical forms not proper of educated language (1d, 1e), dialectal words (1f), dialectal diminutive forms (1g, 1h), or /θ/ spelt as *s* (1i), and vice versa (1j):

- (1) a. crei **de que** me hescribirias (PSCR6538)
- b. Dios **los** abra los ojos, que vien lo an menester (PS6155)
- c. A la Monica **la** di la carta (PS7020)
- d. esta solo sirve para decirte **ande** as ido (PS6229)
- e. desde que sali del calabozo no e ablado con **naide** (PS8104)
- f. unas tenazas, una **rasera** que no tiene pala, un **bail** (GR1829I2005)
- g. Remito el **paniño** que me pide (PS6040)
- h. Dos **potesinos** pequeños con una poca de aceite (BA1601I7042)
- i. unos **sapatos** de **beserro** morado en **quinse** rreales (SE1780I7048)
- j. no **quizieron pazar** por el **zitio** de la quimera (MA1759D9053)

¹⁶ My translation. The original reads: “[en las centurias previas al siglo XX] continúan abundando los textos de carácter más formal”.

The collection of these highly valuable handwritten texts demands major search and selection efforts. Both the *Post Scriptum* correspondence and the ODE texts come from various historical fonds in the Iberian Peninsula that are generally not in digital format. Corpus compilers thus have to check the sources wherever they are stored: 27 and 20 archival institutions were reached for the samples collected for the PS-ES and for the ODE, respectively.

Written texts close to orality are hard to find, and their location is largely subject to chance, as in personal letters or witness testimonies and direct speech: The data recorded during compilation of these corpora reveal very low success rates, as only 8% of the court records searched for PS-ES contained personal letters (with an even lower percentage by semi-literate writers), and only 4% of the court records searched in the ODE corpus showed direct speech in witness testimonies. As historical sources attesting communicative immediacy are difficult to come by, they are also costly in time for specialized corpus building, and are missing in diachronic reference corpora.

4.2 Systematic Encoding of Geographic Metadata

Corpus sample encoding is another major point in the discussion of the difference between these corpora, i.e. how the information about each sample and their authors is encoded and structured for efficient data retrieval according to specific criteria: date, location, text type, etc. A good part of these data of historical sources is missing or inaccurate and, as a result, the metadata of diachronic corpora are, in general, rather lacking in detail (Nevalainen, Suhr & Taavitsainen 2019: 10). Regarding Spanish reference corpora, CdEhist shows the most serious limitations: At present, the interface only allows queries by text type or by period and, of the latter, only by century so no specific time intervals can be searched for. Finally, the two academic corpora rely on more detailed encoding than CdEhist, and they allow queries by geographical area, by subject topic, and also by year, not just by century. However, their potential for dialectal research is very low, as many of the corpus samples are of unknown origin, or are classified as too general or imprecise types, among other limitations:

Regarding dialectal diversity, neither the biblical corpus nor large corpora such as CORDE or the Corpus del español are suitable for studying diatopic variation, as the exact dialectal origin of many of the texts they contain is unknown. Moreover, these corpora also include

texts preserved in later testimonies in which a mixture of dialects may have occurred as a result of the intermediation of scribes in textual transmission (Enrique Arias 2009: 274).¹⁷

By contrast, ODE and PS-ES are encoded such that they outperform reference corpora: Their sources typically attest a precise and, therefore, quite reliable production date and location record. *Post Scriptum* also details the letter writers' provenance, for this and other biographical data are often explicit in the correspondence used as court evidence. Specifically, out of 2,447 letters by 1,222 writers, the provenance of 514 was identified and, therefore, encoded.

As to the ODE corpus, testimonies, goods inventories, and reports have reached us in similar conditions: The informant's oral statements—or the surveyor's oral reports—are recorded in written by a scrivener. While no biographical data are available, the informants', surveyors', or scribes' provenance must be close to the place of text production, as shown by research on the dialectal features of this type of documents (Calderón Campos 2019). The approximately 900 documents contained in the corpus display features of 25 Spanish provinces, especially from the central and southern parts of the country.

The encoding used for PS-ES and ODE thus offers unique possibilities for research on diachronic dialectology. Also, whereas the CORDE and CDH contents are classified by country, these specialized corpora are classified by province or even by more specific locations. Intrapeninsular diatopic variation in past- and present-day Spanish can thus be researched based on these resources. As evidence of this, consider the map shown in Figure 1, which is based on PS-ES data.¹⁸ This map shows the geographical distribution of *laísmo*: the undue use of clitic pronoun *la* or (less commonly) *las* in the role of indirect object, as in (1c) above. The points on the map represent the places of origin of “*laísta*” authors in this corpus (see Vaamonde 2015: 71–76).

17 My translation. The original reads: “Respecto a la diversidad dialectal, ni el corpus bíblico ni los grandes corpus como el CORDE o el Corpus del español son apropiados para el estudio de la variación diatópica, pues para muchos de los textos que contienen nos es desconocida la exacta procedencia dialectal; estos corpus incluyen además textos conservados en testimonios tardíos en los que puede haberse dado mezcla de dialectos como resultado de las [sic] intermediación de los copistas en la transmisión textual”.

18 See Calderón Campos & Vaamonde (2024) and Vaamonde (2024) for dialect maps based on ODE data.

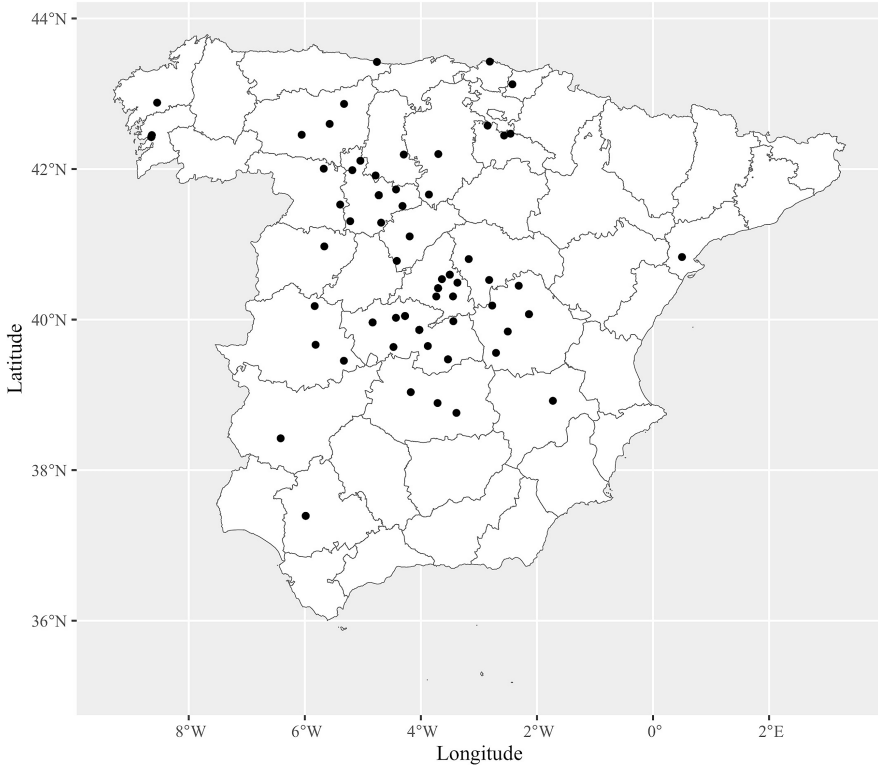


Figure 1: Geographical distribution of peninsular authors who use *laísmo* in PS-ES.

4.3 Faithful Transcription of Textual Content

The third difference concerns transcription criteria. As mentioned above, large corpora like CORDE, CdeHist, or CDH rely by necessity on the editions available at the time when they are compiled, so their control on editorial quality reaches, at best, sample selection by ecdotic quality. In practice, this means modern editions of older texts —whether digital or in print— have to be used, even if they were not designed for language research.

The samples thus selected and the corpora built therefrom entail varying degrees of editorial influence, hence a number of undesirable issues in historical research that are described elsewhere (Lass 2004; Dollinger 2004; Grund 2006; Claridge 2008: 250–251; Honkapohja, Kaislaniemi & Marttila 2009: 456–460). These issues are major limitations on the potential of diachronic corpora, e.g. the amount of text variation resulting from the guidelines used in each edition, and

the deficiency of the corpus as an experimental basis for spelling-based research, as there is no certainty that the original spelling has been preserved.

ODE and PS-ES consist in unpublished handwritten documents transcribed according to specific guidelines for confirmed corpus homogeneity. The transcripts used for these corpora stand out for their accuracy with respect to the original spelling, so they are ideal data sources for research on various language levels, including spelling and phonetics (see section 4.5).

4.4 Digital Representation of Manuscript Reality

The smaller size of these two diachronic corpora allowed thorough philological data preparation and, as a result, high quality text editions. The accuracy with respect to the original source thus involves not just spelling, but other aspects of the content of the samples too. Specifically, both ODE and PS-ES encode structural features, excerpts with particularly relevant contents, or words and phrases that may disclose information on the sample production, e.g. beginning of new page, beginning of new line, abbreviations, additions, deletions, omissions, or typographical mistakes, among others (see Table 4).¹⁹

Table 4: Some TEI elements used for ODE and *Post Scriptum* encoding.

TEI element	Description	TEI element	Description
<pb/>	page beginning	<supplied>	restored text
<lb/>	line beginning	<gap/>	omitted text
<p>	paragraph	<hi>	highlighted text
<add>	text addition	<surplus>	redundant text
	deletion	<sic>	incorrect text
<subst>	substitution	<foreign>	text in a different language
<abbr>	abbreviation	<quote>	direct speech

Both corpora encoded this information according to the *Text Encoding Initiative* (TEI) guidelines. TEI offers a widely acknowledged XML-based language for text representation in electronic format in the Digital Humanities. As a result, the end-users gain access to a language corpus, but also to a number of *scholarly digital*

¹⁹ Out of the three diachronic reference Spanish corpora, only CDH is encoded with similar information, even if from a narrower range of data than the one shown in Table 4.

editions for accurate queries according to various parameters, whether textual, discursive, or palaeographic.

The ODE query interface allows easy retrieval of all the instances of direct speech in witness testimonies, which is the closest text type to communicative immediacy in the corpus. PS-ES queries can easily be adjusted according to the discursive passages of letters, like opening and closing lines, or postscripts. The revisions encoded —additions, deletions, substitutions— may prove particularly useful as they usually evidence relevant variations at several language levels. Thus, the substitution shown in Figure 2 reveals uncertainty between the graphemes *y* and *ll*, and this may, in turn, be proof of pronunciation as /j/ (*yeísmo*) or as /ʎ/ (*lleísmo*). Similarly, deletion of the conjunction *q(ue)* in Figure 3 evidences use of *que* (*queísmo*), of which the writer is aware and, ultimately, syntactic hesitation in the formation of subordinate completive clauses. Encoding of these features as TEI-compliant XML-mark-up language allows easy retrieval by use of XPath query language.

As the facsimile images of the documents sampled in the corpus are available too, further check of the digital edition against the original manuscript is always possible.

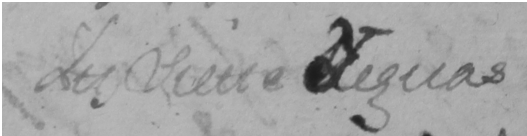


Figure 2: A snapshot of a goods inventory dated 1736 handwritten in Huelva. Source: ODE (HU1736I3518). Transcript: *Ytt siete Huegas*.

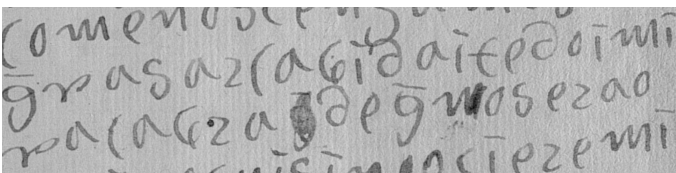


Figure 3: A snapshot of a letter dated 1628 written by a Seville-born writer. Source: PS-ES (PSCR7376). Transcript: *[. . .] i te doi mi palabra q de q no será [. . .]*.

4.5 Fully Normalized Spelling Variation

The last two points under discussion here are about the levels of annotation for more accurate data retrieval in the specialized corpora ODE and PS-ES, compared

with reference corpora. Spelling normalization, i.e. spelling according to present-day standard convention, is not available in reference corpora, even if it proves extremely useful «to enhance the searchability of historical texts» (Kytö 2011: 440).

The most obvious improvement of this level of annotation is that all spelling variants of one form can be retrieved after spelling normalization of the corpus, as shown in Table 5: The most frequent spelling of *vecino* ‘neighbor’ in ODE, of *diligencia* ‘proceedings’ in PS-ES, and of *muselina* ‘muslin’ in ODE are shown in three-column sets.

Historical texts are well-known for their high spelling instability. Table 5 is a small token of the many spellings that one and the same word may be retrieved by. These data are useful for research on spelling, e.g. in abbreviations (*vezo*, *vo*, *vzo*, *vso*, *uo* for *vecino*) or in the representation of one phoneme (*diligencia*, *dilijencia*, *dilixencia*), but also for research on phonetic traits based on spelling (see Calderón Campos & Vaamonde 2024), e.g. for spelling of /θ/ as *s* (*vesino*, *diligensia*) as shown also in example (1i) above, for instable spelling of unstressed vowels (*deligencia*, *musolina*, *musalina*, *mosulina*), or for the use of epenthetic consonants (*murcelina*, *munsolina*).

Table 5: Spelling instability of *vecino* ‘neighbor’ (ODE), *diligencia* ‘proceedings’ (PS-ES), and *muselina* ‘muslin’ (ODE).

<i>vecino</i> (ODE data)			<i>diligencia</i> (PS-ES data)			<i>muselina</i> (ODE data)		
spelling	no.	%	spelling	no.	%	spelling	no.	%
<i>vezino</i>	218	24.04	<i>diligencia</i>	47	25.41	<i>muselina</i>	51	30.54
<i>vezo</i>	161	17.75	<i>diligencia</i>	22	11.89	<i>musolina</i>	25	14.97
<i>vo</i>	107	11.80	<i>diliga</i>	19	10.27	<i>musulina</i>	21	12.57
<i>vzo</i>	102	11.25	<i>diligencia</i>	12	6.49	<i>mosolina</i>	19	11.38
<i>vecino</i>	97	10.69	<i>dilixa</i>	9	4.86	<i>murselina</i>	14	8.38
<i>veçino</i>	32	3.53	<i>diligensia</i>	9	4.86	<i>muzelina</i>	6	3.59
<i>vezno</i>	21	2.32	<i>dilijencia</i>	7	3.78	<i>mozolina</i>	6	3.59
<i>vesino</i>	19	2.09	<i>dilijençia</i>	5	2.70	<i>murzelina</i>	4	2.4
<i>uezino</i>	18	1.98	<i>deligencia</i>	5	2.70	<i>moselina</i>	4	2.4
<i>vso</i>	16	1.76	<i>dilijenzia</i>	4	2.16	<i>murcelina</i>	4	2.4
<i>bezino</i>	15	1.65	<i>delixencia</i>	3	1.62	<i>morselina</i>	3	1.8
<i>vno</i>	15	1.65	<i>dilixencia</i>	3	1.62	<i>mosulina</i>	2	1.2
<i>uo</i>	8	0.88	<i>dilixencia</i>	3	1.62	<i>morsolina</i>	1	0.6
<i>vez</i>	8	0.88	<i>deligencia</i>	3	1.62	<i>munsolina</i>	1	0.6
<i>vz</i>	7	0.77	<i>deligensia</i>	3	1.62	<i>musalina</i>	1	0.6

Three modes of visualization are thus available for each document in ODE and PS-ES: the conservative transcript, the facsimile edition, and the normalized transcript. The latter contributes not just all the spelling forms for each word, but it also provides researchers of digital manuscripts with a more legible text, linguists gain access to a formally more consistent database to search, and corpus compilers in charge of linguistic annotation may use automatic tools more effectively.

The argument for normalisation is twofold. First, that it helps to improve the accuracy of automated computational linguistic (natural language processing) techniques such as part-of-speech tagging and second, that it improves the stability and robustness of corpus linguistic methods such as keyword analysis, thereby allowing existing software tools of both types to be used unmodified (Archer *et al.* 2015: 6).

4.6 Manually Revised Linguistic Annotation

The issue of linguistic annotation follows from the previous section. Whether diachronic or synchronic, a corpus with lexical and grammatical annotation can be queried for more—and more abstract—data than a raw corpus. Corpus annotation thus plays a crucial role in the exploitation and utility of corpora as research tool. The main diachronic corpora of Spanish differ greatly in this regard, but not any stands out as a high-quality annotation standard.

Out of the three large diachronic corpora of Spanish, CORDE lacks linguistic annotation. While the query interface allows some wildcards and searches can thus be refined to some extent, it relies too heavily on specific lexical forms. As a result, it is fairly limited for morphosyntactic research, especially as regards the more abstract questions in this field (Davies 2009; Nieuwenhuijsen 2009; Garachana & Artigas 2012; Arellano 2021).

CDH and CdEhist are lemmatized and grammatically tagged. In general, tagging is performed automatically with a high degree of accuracy, but manual disambiguation is necessary at a post-editing stage and is, thus, not manageable in large corpora. Queries in CdEhist and CDH often produce false positives (wrong results), particularly when homonymy is involved. The CDH guidebook reports that «it is possible to reduce ambiguity in lemma and form queries by selecting the part of speech»,²⁰ but this approach is not consistently applied. For example, the word

²⁰ My translation. The original reads: “[e]s posible reducir la ambigüedad en las consultas de lemas y formas seleccionando la clase de palabra”.

aviso '[I] warn / warning' tagged as a verb may list homonymous nouns as results, as in (2) and, vice versa, the query for the noun may throw back verbs, as in (3).²¹

- (2) a. y aunque el **aviso** de Ferramosca era mentira tomólo el Virrey por verdad (CDH, 1550)
 b. mejor es en lo escondido del monte dejarle atado, porque no lleve el **aviso** (CDH, 1639)
 c. Nos mantuvimos en dicho acampamento, esperando el **aviso** de dicho cacique (CDH 1770)
 d. Ella y el tono en que estaba puesta eran ó un **aviso** ó un insulto (CDH, 1824)
- (3) a. Y os **aviso** que Dios no quiere ni quiso que biváis (CdEhist, 1512)
 b. Esto os **aviso**, para que en darle gusto, determinéis lo que os conviene (CdEhist, 1616)
 c. Lo único que encontré, y **aviso** a V. E. por si le gustase, es paño de seda negro (CdEhist, 1778)
 d. el 7 de junio estaré en Graus; el 4 en Estadilla, si no **aviso** de lo contrario (CdEhist, 1842)

Of course, tagging errors increase in more frequent homonyms, e.g. as in *la* as a pronoun or as an article, or in *que* as a conjunction or as a pronoun.

The automatically tagged ODE and PS-ES corpus contents were manually post-edited by linguists for higher accuracy in data retrieval. Both corpora are lemmatized and tagged morphosyntactically according to the EAGLES standard for morphosyntactic annotation of European languages (Leech & Wilson 1996). As they are small corpora, they can be annotated further, often according to the corpus compilers' research interests, as they add their fully tested analyses. Thus, PS-ES shows annotation for clitic variance (use of *la*, *le*, or *lo*) and expletive use of *de* in completive clauses (resulting in the use of *de que* sequences) besides lemmatization and morphosyntactic tags, and ODE allows retrieval of phonetic processes (/θ/ spelt as *s* and vice versa, alteration of the quality of liquid consonants, vowel variation) and lexical processes (medical terminology). Annotation granularity and accuracy are ultimately differentiating factors in specialized diachronic corpora.

²¹ An unknown number of false negatives (results that should be listed but are not) for want of (correct) tagging must be taken into consideration too.

5 Concluding Remarks

This is the era of *big data*. The rise of data availability means an academic revolution, also for language research: widespread access to digital megacorpora of several thousand million words has been described as “the modern equivalent of the 17th century invention of the telescope and microscope” (Lieberman 2014). Corpus linguists have compiled and researched increasingly larger datasets for decades now, and the right corpus size is an old, controversial question where opposite standpoints can be easily found (Valenzuela 2022: 248).

In diachronic linguistics, small corpora offer major advantages compared with larger corpora in virtually every stage of corpus building: careful data selection from sources that are hard to find and of linguistic value, thorough palaeographic text edition alongside facsimile images, highly detailed metadata encoding, or accurate and well-informed linguistic annotation. These qualities give corpora an added value in several regards (*reliability, faithfulness, transparency, searchability*), so they become more useful as data sources. Unquestionably, they are also time-demanding and call for the joint effort of a large research team. *Post Scriptum* was developed by some thirty specialists (linguists, historians, computing technicians, palaeographers) for over five years. ODE has been under construction since 2010 through several consecutive research projects. The permanent corpus staff amounts to some ten specialists, but the number of participants, including students and occasional collaborators, is much higher. This chapter highlights these two digital resources and corpora, and argues for an approach that is not always unquestioned in corpus linguistics, namely prioritizing data quality over quantity:

As understood here, “small is beautiful” looks to the future rather than the past. We shall see that the close scrutiny of carefully selected small sets of examples will remain important in the age of mega-corpora, because it is the appropriate method to tackle a few of the most challenging issues in the study of language variation and change (Mair 2013: 182)

As an active participant in the PS-ES and ODE ventures, I am confident that the time and effort invested are worth it, and that the corpus data will significantly advance research on variation and language change in Spanish historical linguistics.

Bibliography

- Archer, Dawn, Merja Kytö, Alistair Baron and Paul Rayson (2015): “Guidelines for normalizing Early Modern English corpora: Decisions and justifications”, in *ICAME Journal*, 39, pp. 5–24.
- Arellano, Nicolás (2021): “Diseño de corpus específicos para el estudio histórico gramatical: el caso de las construcciones con clítico femenino”, in *Revista de estudos da linguagem*, 29(2), pp. 711–737.
- Blas Arroyo, José Luis (dir.) (2019): *Sociolingüística histórica del español. Tras las huellas de la variación y el cambio lingüístico a través de textos de inmediatez comunicativa*. Madrid: Iberoamericana/ Frankfurt: Vervuert.
- Calderón Campos, Miguel (2019): “La configuración de la variedad meridional en el reino de Granada”, in Eugenio Bustos Gisbert and Juan Pedro Sánchez Méndez (eds.). *La configuración histórica de las normas del castellano*. Valencia: Tirant Humanidades, pp. 109–134.
- Calderón Campos, Miguel and Gael Vaamonde (2024): “Anotación y explotación de variantes gráficas de base fonética en el corpus *Oralia Diacrónica del Español*”, in *Philologia Hispalensis*, pp. 287–309.
- Cano Aguilar, Rafael (1996): “Lenguaje ‘espontáneo’ y retórica epistolar en cartas de emigrantes españoles a Indias”, in Thomas Kotschi, Wulf Oesterreicher and Klaus Zimmermann (eds.). *El español hablado y la cultura en España e Hispanoamérica*. Madrid: Iberoamericana/Frankfurt: Vervuert, pp. 375–404.
- CdEhist = Corpus del Español (Género/Histórico). Dir. Mark Davies, <www.corpusdelespanol.org/histgen/>.
- CDH = Real Academia Española. *Corpus del Nuevo diccionario histórico del español*. <<https://www.rae.es/banco-de-datos/cdh>>.
- Claridge, Claudia (2008): “Historical corpora”, in Anke Lüdeling and Merja Kytö (eds.). Berlin/ New York: Walter de Gruyter, pp. 242–259.
- CORDE = Real Academia Española. *Corpus diacrónico del español*. <<https://www.rae.es/banco-de-datos/corde>>.
- Davies, Mark (2009): “Creating Useful Historical Corpora: A Comparison of *CORDE*, the *Corpus del español*, and the *Corpus do português*”, in Andrés Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana/ Frankfurt: Vervuert, pp. 137–166.
- Davies, Mark (2018-): *Corpus del Español News on the Web (NOW)*. <<https://www.corpusdelespanol.org/now/>>.
- Del Rey Quesada, Santiago (2020): “Hacia una diacronía de la oralidad: el inicio de turno y la inmediatez comunicativa en un corpus de traducciones de Plauto y Terencio (ss. XVI y XIX)”, in *Lexis*, 44(1). <<http://dx.doi.org/10.18800/lexis.202001.002>>.
- Dollinger, Stefan (2004): “*Philological computing* vs. *philological outsourcing* and the compilation of historical corpora: a Late Modern English test case”, in *Vienna English Working Papers (VIEWS)*, 13(2), pp. 3–23.
- Eberenz, Rolf and Mariela de la Torre (2003): *Conversaciones estrechamente vigiladas: interacción coloquial y español oral en las actas inquisitoriales de los siglos XV a XVII*. Zaragoza: Pórtico.
- Enrique-Arias, Andrés (2009): “Ventajas e inconvenientes del uso de *Biblia Medieval* (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español”, in Andrés Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana/Frankfurt: Vervuert, pp. 269–283.

- Francis, W. Nelson (1992): “Language Corpora B.C.”, in Jan Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 17–32.
- Garachana, Mar and Esther Artigas (2012): “Corpus digitalizados y palabras gramaticales”, in *Scriptum Digital*, 1, pp. 37–65.
- Grund, Peter (2006): “Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights”, in *Journal of English Linguistics*, 34, pp. 105–125.
- Hernández-Campoy, Juan Manuel and Natalie Schilling (2012): “The Application of the Quantitative Paradigm to Historical Sociolinguistics: Problems with the Generalizability Principle”, in Juan Manuel Hernández-Campoy and Juan Camilo Conde-Silvestre (eds.). *The Handbook of Historical Sociolinguistics*. Oxford: Wiley-Blackwell, pp. 63–79.
- Hiltunen, Turo, Joe McVeigh and Tanja Säily (2017): “How to turn linguistic data into evidence”, in *Studies in Variation, Contacts and Change in English. Volume 19: Big and Rich Data. English Corpus Linguistics: Methods and Explorations*. <<https://varieng.helsinki.fi/series/volumes/19/introduction.html>> (30-06-2023).
- Honkaphoja, Alpo, Samuli Kaislaniemi and Ville Marttila (2009): “Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora”, in Andreas H. Jucker, Daniel Schreier and Marianne Hundt (eds.). *Corpora: Pragmatics and Discourse*. Amsterdam/New York: Rodopi, pp. 451–475.
- Kytö, Merja (2011): “Corpora and historical linguistics”, in *Revista Brasileira de Lingüística Aplicada*, 11(2), pp. 417–457.
- Kennedy, Graeme (1998): *An introduction to corpus linguistics*. London/New York: Longman.
- Kilgarriff, Adam and Gregory Grefenstette (2003): “Introduction to the Special Issue on the Web as Corpus”, in *Computational linguistics*, 29(3), pp. 333–347.
- Kilgarriff, Adam and Irene Renou (2013): “esTenTen, a Vast Web Corpus of Peninsular and American Spanish”, in *Procedia – Social and Behavioral Sciences*, 95, pp. 12–19.
- Kohonen, Thomas (2007): “From Helsinki through the centuries: the design and development of English diachronic corpora”, in Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen and Jukka Tyrkkö (eds.). *Studies in Variation, Contacts and Change in English. Volume 2: Towards Multimedia in Corpus Studies*. <<https://varieng.helsinki.fi/series/volumes/02/kohonen/>>.
- Labov, William (1994): *Principles of Linguistic Change. Vol. 1: Internal Factors*. Oxford: Blackwell.
- Lass, Roger (2004): “Ut custodian litteras: Editions, Corpora and Witnesshood”, in Marina Dossena and Roger Lass (eds.). *Methods and Data in English Historical Dialectology*. Bern: Peter Lang, pp. 21–48.
- Leech, Geoffrey (1991): “The state of the art in corpus linguistics”, in Karin Aijmer and Bengt Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London/New York: Longman, pp. 8–29.
- Leech, Geoffrey and Andrew Wilson (1996): *EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora*. <<http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>>.
- Lieberman, Mark (2014): “How big data is changing how we study languages”, in *The Guardian*, 7 April 2014. <<https://www.theguardian.com/education/2014/may/07/what-big-data-tells-about-language>>.
- Mair, Christian (2006): “Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora”, in Antoinette Renouf and Andrew Kehoe (eds.). *The Changing Face of Corpus Linguistics*. Amsterdam/New York: Rodopi, pp. 355–376.

- Mair, Christian (2013): “Using ‘small’ corpora to document ongoing grammatical change”, in Manfred Krug and Julia Schülter (eds.). *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press, pp. 181–194.
- Marquilha, Rita (2000): *A faculdade das letras. Leitura e escrita em Portugal no século XVII*. Lisboa: Imprensa Nacional-Casa da Moeda.
- Molina Salinas, Claudio and Gerardo E. Sierra Martínez (2015): “Hacia una normalización de la frecuencia de los corpus CREA y CORDE”, in *Revista Signos. Estudios de Lingüística*, 48(89), pp. 307–331.
- Nevalainen, Terttu, Carla Suhr and Irma Taavitsainen (2019): “Corpus Linguistics as Digital Scholarship: Big Data, Rich Data and Uncharted Data”, in Carla Suhr, Terttu Nevalainen and Irma Taavitsainen (eds.). *From data to evidence in English Language Research*. Leiden/Boston: Brill, pp. 1–26.
- Nieuwenhuijsen, Dorien (2009): “El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales”, in Andrés Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana/Frankfurt: Vervuert, pp. 365–384.
- Octavio de Toledo y Huerta, Álvaro (2019): “Large Corpora and Historical Syntax: Consequences for the Study of Morphosyntactic Diffusion in the History of Spanish”, in *Frontiers in Psychology*, Vol. 10, Art. 780.
- Oesterreicher, Wulf (2004): “Textos entre inmediatez y distancia comunicativas: el problema de lo hablado en lo escrito en el Siglo de Oro”, in Rafael Cano Aguilar (coord.). *Historia de la lengua española*. Barcelona: Ariel, pp. 729–770.
- Parodi, Giovanni (2008): “Lingüística de corpus: una introducción al ámbito”, in *Revista de Lingüística Teórica y Aplicada*, 46(1), pp. 93–119.
- Pascual, José Antonio and Carlos Domínguez (2009): “Un corpus para un Nuevo diccionario histórico del español”, in Andrés Enrique-Arias (ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana/Frankfurt: Vervuert, pp. 79–93.
- Rissanen, Matti (1992): “The diachronic corpus as a window to the history of English”, in Jan Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 185–205.
- Rissanen, Matti (1999): “Syntax”, in Roger Lass (ed.). *The Cambridge history of the English language*. Cambridge: Cambridge University Press, pp. 187–331.
- Renouf, Antoinette (2007): “Corpus development 25 years on: from super-corpus to cyber-corpus”, in Roberta Facchinetti (ed.). *Corpus linguistics 25 years on*. Amsterdam/New York: Rodopi, pp. 27–49.
- Rodríguez Molina, Javier and Álvaro Octavio de Toledo y Huerta (2017): “La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística”, in *Scriptum digital*, 6, pp. 5–68.
- Rodríguez Puente, Paula (2018): “En busca de lo hablado en lo escrito en los corpus diacrónicos del español: una comparativa con los corpus anglosajones”, in *E-Scripta Romanica*, 5, pp. 89–127.
- Rojo, Guillermo (2010a): “Aguja de navegar corpus”, in Víctor M. Castel and Liliana Cubo de Severino (eds.). *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística*. Mendoza: Editorial FFyL, UNCuyo, pp. 1151–1163.
- Rojo, Guillermo (2010b): “Sobre codificación y explotación de corpus textuales: otra comparación del Corpus del español con el CORDE y el CREA”, in *Lingüística*, 24, pp. 11–50.
- Rojo, Guillermo (2012): “El papel de los corpus en el estudio de la historia del español”, in Emilio Montero Cartelle and Carmen Manzano Rovira (coords.). *Actas del VIII Congreso Internacional de*

- Historia de la Lengua Española*. Meubook: Asociación de Historia de la Lengua Española, pp. 433–444.
- Rojo, Guillermo (2016): “Los corpus textuales del español”, in Javier Gutiérrez-Rexach (ed.). *Enciclopedia de Lingüística Hispánica*. London/New York: Routledge, pp. 285–296.
- Rojo, Guillermo (2021): *Introducción a la lingüística de corpus en español*. London/New York: Routledge.
- Sánchez Lancis, Carlos (2022): “Corpus diacrónicos del español de España”, in Giovanni Parodi, Pascual Cantos-Gómez and Chad Howe (eds.). *The Routledge Handbook of Spanish Corpus Linguistics*. London/New York: Routledge, pp. 33–44.
- Sánchez Sánchez, Mercedes and Carlos Domínguez Cintas (2007): “El banco de datos de la Real Academia Española: CREA y CORDE”, in *Per Abbat: boletín filológico de actualización académica y didáctica*, 2, pp. 137–148.
- Schneider, Edgar W. (2013): “Investigating Historical Variation and Change in Written Documents: New Perspectives”, in J. K. Chambers and Natalie Schilling (eds.). *The Handbook of Language Variation and Change*. Oxford: Wiley-Blackwell, pp. 57–81.
- 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.
- Tognini Bonelli, Elena (2010): “Theoretical overview of the evolution of corpus linguistics”, in Anne O’Keeffe and Michael J. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge, pp. 14–27.
- Vaamonde, Gael (2015): “Distribución de leísmo, laísmo y loísmo en un corpus diacrónico epistolar”, in *Res Diachronicae*, 13, pp. 58–79.
- Vaamonde, Gael (2018): “La multidisciplinariedad en la creación de corpus históricos: El caso de *Post Scriptum*”, in *Artnodes*, 22, pp. 118–127.
- Vaamonde, Gael (2024): “Diseño y explotación de un corpus histórico de textos oralizantes para el estudio del español clásico y moderno”, in *Revista de Humanidades Digitales*, 9, pp. 41–70. <<https://doi.org/10.5944/rhd.vol.9.2024.39834>>.
- Vaamonde, Gael, Fita González and José María García-Miguel (2010): “ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'2010)*. Valletta (Malta), pp. 1903–1910.
- Valenzuela, Javier (2022): “El *big data* en los estudios del lenguaje”, in *Estudios de Lingüística del Español*, 45, pp. 241–260.
- Van der Wal, Marijke (2022): “The challenge of historical data: from sources and corpora to answering research questions”, in *Slovo a slovesnost*, 83, pp. 335–350.

Inmaculada González Sopeña

Language Corpora and Lexical Arabisms in the Digital Age

1 Introduction. The Era of Digital Humanities

The so-called Digital Era has meant a paradigm shift in scientific research. The industrial era has undergone a transition into a digital era since the advent of computers and similar devices, as well as since access to Internet became widespread for most of the world population since the late 20th c. and early 21st c. (Jódar Marín 2010: 3). The Internet has become a new social forum where a wealth of data of all kinds are released. A range of technological tools has thus become quickly available for research in all fields to further human knowledge and has made it accessible online. In a word, the digital revolution means a paradigm shift in research practice.¹

Within the Humanities, the use of such resources, methods and tools has sparked an intense debate, to the extent that its role and the very view of the discipline have been revised regarding how to use and transfer this progress to society (Romero Frías 2014: 19). The concept of *Digital Humanities*² and its very notion as a specific field arise at this point. Roberto Busa's project on concordance design for specific search of Thomas Aquinas's works, assisted by IBM on the computational side of the project,³ has been cited as the starting point of this process (Allés Torrent 2019, Romero Frías 2014, Spence 2014b). The project brought to light the need for analog-to-digital text conversion to enable computer-assisted data search for concepts, language structures, etc. Corpus linguistics and Philology thus became pioneers in the use of these new methods. At present, the Digital Humanities aim at "the design and use of applications and models for new teaching

1 Cf., thus, the use of the term *cyberscience* or *e-Science* (Romero Frías 2014:22).

2 This label dates back to the label *Humanities Computing* of the 1950s (Spence 2014b). The publication of *A Companion to Digital Humanities* in 2004 renamed the discipline for a wider scope and full coverage of all the changes brought by the digital revolution to such a multifaceted field as the Humanities (Romero Frías 2014, Allés Torrent 2019).

3 Namely, *Index Thomisticum*.

Note: This contribution has been supported by Grant PID2022-136256NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by the European Union's ERDF/EU. Also, it has been carried out within the framework of Grant C-HUM-038-UGR23 funded by *Consejería de Universidad, Investigación e Innovación* and by ERDF Andalusia Program 2021–2027.

and research modes” (Allés Torrent 2019: 6). To this end, they include various processes regarding data processing (creation, management, storage, reuse). In the digital era, data, which may be of various kinds, are viewed as “an items’, attribute’s or empirical value’s smallest digital expression and symbol representation that is formalized digitally as a binary notation of 0 or 1” (Allés Torrent 2019: 12). Five major data types can be identified, according to their digital encoding: Text, numbers, images, video, and audio.

Accordingly, data encoding has given rise to “virtually unlimited research avenues, ranging from *Big Data* to *Small Data*” (Romero Frías 2014: 20). This is thanks to the new technological resources and tools available, both in general and specifically for research in the Humanities too. The *big data* boom, reportedly starting in 2011, has been defined variously, even if “most experts and academics agree that it is some kind of combination of algorithms, technologies and strategies capable of collecting and analysing large amounts of data” as fast as possible (Sánchez González 2014: 136). The aim is “to bring together, store and process as much information as possible” (Allés Torrent 2019: 19), with a view to subsequent quantitative analysis for patterns and dynamics, and for the construction of predictive models. In the Humanities, this amounts to the information that a researcher cannot read in a given period of time and whose interpretation requires computational tools (Graham, Milligan & Weingart 2015). Schöch (2013) put forward a closely related term, *smart data*: Unlike *big data*, smart data are semi-structured, explicit, enriched, clean, interconnected, and need human processing, e.g. XML-marked-up texts for digital editions.

As for Philology, the Digital Humanities have contributed a large number of new tools and resources for language research. Corpus linguistics and computational linguistics were the first disciplines to develop specific tools for retrieval, analysis, and interpretation of text data (Martínez-Gamboa 2016). The Digital Humanities therefore stand out for the design and compilation of language corpora as experimental bases for research on relevant questions in Spanish and in other languages.

This chapter is intended to briefly review the contribution of corpus linguistics to research on Spanish up to present-day. This includes a short account of corpus linguistics and its objectives (with a focus on text encoding), and a discussion of its use for research on Spanish. The chapter therefore reviews the properties of the main types of Spanish language corpora: Reference corpora (CORDE, CREA, CORPES, etc.) and specialized corpora (*CorLexIn*, ODE, CORDIAM), among others. It also reviews research on Arabisms in Spanish, and assesses how this specific field has been influenced by the digital revolution. Therefore, a bibliographical review is also included of the main research avenues on Arabisms and of the difficulties for their identification and retrieval in corpora.

In this regard, the chapter spans from the classics and from most traditional studies to the most recent corpus-based references. The resulting review shows the advantage in researching Arabisms, their variants, their text sources, and their chronological attestation based on the new methods and the new technological tools of corpora. Indeed, the new technologies have helped improve the lexi-co-semantic classification of Arabisms, and also revise specific statements about the periods when Arabisms were added most productively and when they were more frequently replaced by other words. These and other improvements are some of the contributions to the study of Arabisms in Spanish by TEI-compliant, XML-marked-up transcripts stored within the TEITOK platform. TEITOK (Janseen 2016) was originally designed as a platform for corpus storage and exploitation, as is the case of the most relevant corpus for this chapter, namely *Oralia diacrónica del español* (ODE). The chapter closes with short conclusions drawn from the preceding sections.

2 Corpus Linguistics and Spanish Language Corpora

Corpus linguistics, whether a discipline or a method,⁴ is rather new, as it can hardly be conceived without computer technology. The first language corpus designed for computerized use, the *Brown University Standard Corpus of Present-Day American English*, dates back to 1964 (Rojo 2016: 286). A *language corpus* is:

[. . .] un conjunto de (fragmentos de) textos, orales o escritos, producidos en condiciones naturales, conjuntamente representativos de una lengua o variedad de lengua, que se almacenan en formato electrónico y se codifican con la intención de que puedan ser analizados científicamente.⁵ (Rojo 2021:1)

The need for corpora to be built and stored “in electronic format” makes computer technology a turning point for the capacity to manage text in various formats, to store data by the thousand million, and to efficiently retrieve language

⁴ The debate on whether corpus linguistics is a new discipline, a new approach or a new method (Leech 1992, Gries 2006, Parodi 2010) has reached the generally acknowledged conclusion that it is an approach, such that the language data are researched empirically using new tools for quantitative and qualitative analysis (Rojo 2021: 49–50).

⁵ “[. . .] a collection of (passages of) texts, spoken or written, produced in natural conditions, representative of a language or language variety, stored in electronic format and encoded for scientific research”. (Rojo 2021:1).

data with new applications. Corpora are classified according to their text samples (originally, from communication not intended for research), representativeness, size, and also to the balance between chronological periods and text types⁶ (Rojo 2021: 63–69). Regardless, electronic format remains a basic, essential requirement for digital encoding, i.e. the language data of a text require conversion into machine-readable language (Rojo 2020: 94).

Corpus encoding relies on corpus design, which, in turn, is according to the corpus purpose. Encoding has developed rapidly as a result of computational progress.⁷ *Encoding* is a broad concept that encompasses information of various levels. Text transcription protocols, including decisions like what textual and paratextual information is annotated (e.g. paragraphs, side notes, comments), are examples of first level encoding. Addition of metadata of each corpus sample text (e.g. date, country, author) make a different level. Linguistic annotation, e.g. morphosyntactic tagging or lemmatization, are an additional level too.

In general, encoding starts with conversion of text samples into machine-readable data, after due arrangement and classification (Allés Torrent 2019: 10). Document conversion into machine-readable mark-up languages like XML, HTML, XHTML, or LaTeX, is more and more relevant (Rojo 2021: 74). This process is in accordance with international annotation standards specifically designed for language research, e.g. the *TEI* initiative (*Text Encoding Initiative*) (2016).

Another major point in corpus design is which metadata are to be added to each sample document, i.e. the information on the author, the text, the year of production, the country of origin, the text type, and the register, among others.⁸ The functionalities of a corpus become wider, if morphosyntactic tagging and lemmatization are added (Rojo 2016: 286, 2021: 2–3), as this makes available all kinds of grammatical and lexical information of each word and, thus, any grammatical question can be researched in depth. Computational linguistics has sup-

6 The use of the *Web* as a corpus has become a relevant issue over the past years. Such scientific use of the *Web* still suffers from shortcomings, like the dependence on appropriate search engines, difficulties for the use of regular expressions for retrieval of specific data, or the very fact that the Internet is, by nature, an ever-changing body of data, so any results may change virtually by the day (Rojo 2021: 71).

7 Note that the earlier techniques for analog-to-digital text conversion were based on OCR scanning and the subsequent creation of a .pdf file capable of sustaining unlimited searches (Rojo 2021: 88). Use of plain text for specific queries or frequency data followed afterwards. At present, mark-up language capable of telling documents from their metadata, like SGML, TEI, or CES, is used (Rojo 2021: 92).

8 In addition to data insofar as digital representations, metadata are essential for data classification (Allés Torrent 2019: 13). The possibilities are immense here, and may vary sharply according to text type.

plied a number of taggers and lemmatizers, like FreeLing, Peen Treebank, or the general guidelines of the EAGLES Consortium. Even so, all have limitations and rely on strictly linguistic decisions for the description as a tag of a given word in a given language: Which morphological elements are to be tagged in nouns or verbs? How can they be converted into machine-readable codes? How should Spanish agglutinative forms like *haberlas* be processed? Other issues must be considered too, like syntactically and semantically different homographs, e.g. *vino* ('come.PAST' vs. 'wine') or *la* ('the' vs. 'her.ACC'). Similar questions arise during lemmatization of each orthographic form, especially when it comes to subsuming a whole pronominal paradigm under a lemma, or regarding the lemmatization of locutions in Spanish. The answers to these questions are according to the purpose of each corpus.

2.1 Corpus Types in Spanish

The first Spanish corpora were built in the 1990s, and several remarks are in order prior to their description. First, corpora can be general (reference), specialized, learner corpora, and technical language corpora.⁹ These may have been designed as data sources of the evolution of a language over time (diachronic), of the most recent state of language (synchronic), or of diatopic, diastratic, or diaphasic variation.

Back to the abovementioned question of size, there are Spanish diachronic and synchronic corpora of millions of words. The main difference between these and smaller corpora is that the former contain thousands of text samples, unprovided with an editing policy because they consist in previously edited text, and unprovided with a computational structure (in many cases, without encoding and lemmatization). By contrast, small, specialized corpora have been edited much more exhaustively and rely on a text conversion and language annotation policy. Smaller corpora of the diachronic kind also allow several text presentation modes, e.g. palaeographic, critical, and even facsimile editions (Rojo 2021: 181).

The Spanish Royal Academy (hereafter, RAE) has fostered two major large corpus projects of diachronic and contemporary Spanish since the 1990s: CREA (*Corpus de Referencia del Español Actual*) and CORDE (*Corpus Diacrónico del Español*). CORDE is a 250-million word corpus encompassing from the origins of Spanish up to 1974. CREA is a 160-million word corpus produced between 1975 and 2004. At the turn of the century, these corpora were supplemented with addi-

⁹ For a review, cf. Rojo (2021: 72–75).

tional projects: The multi-layered, 300-million-word CDH (*Corpus del Nuevo Diccionario Histórico de la Lengua Española*), and CORPES XXI (*Corpus del Español del Siglo XXI*), whose latest version amounts to over 381 million words of spoken and written sources produced in the Spanish-speaking world.

By contrast, Mark Davies' *Corpus del español* hosts several subcorpora that are not entirely well-balanced: The historical subcorpus is considerably smaller (ca. 100 million words) than the dialects subcorpus (2,000 million words collected from websites¹⁰). While, in general, all are large corpora and offer overviews of specific topics, they rely on technological resources that do not allow accurate data retrieval.

The diachronic corpora cited above rely mainly on literary and formal register texts. A need for other text types has been arguably considered necessary in the past few years, and a wealth of corpora of specific periods, geographical areas, and text types have mushroomed as a result. Thus, both large and specialized medium-sized and small diachronic corpora of various Spanish text types are available at present. Following the typology by Torruella & Kabatek (2018), there are corpora of European Spanish (CODEA, *Corpus Mallorca*, CODEMA, *CorLexIn*, ODE), of American Spanish (CORDIAM, COREECOM), and of both varieties, like CHARTA's network's.¹¹

All in all, the above offer a range of language samples of various periods, Spanish-speaking geographical areas, and texts from outside the literary realm (correspondence, last wills, goods inventories, ordinances, diaries, legal agreements, chronicles, newspapers). All these corpora contain significant information for the accurate revision and for an improved account of the history of Spanish at all levels.

2.2 Technological Issues in the Spanish Diachronic Corpora, and New Methodological Proposals

Following the above review of Spanish language corpora, this chapter focuses on diachronic corpora, where digital progress has been slower and, often, unsatisfactory. Despite the large number of language corpus projects, many initiatives do not apply the potential of the new digital tools and applications to the design and

¹⁰ Since 2018, this corpus has included the subcorpus *NOW*, with over 7,000 million words produced between 2012 and 2018 (www.corpuesdelespañol.org).

¹¹ Besides these references, it is worth mentioning all the language corpora built in the 1990s cited in specific volumes like Sánchez Prieto (1995), Fontanella (1993), or Rojas (2008). A comprehensive review of Spanish corpora is available in Calderón Campos (2015), and in Rojo (2016).

exploitation of language resources (Díaz Bravo 2018: 565). This is as a result of the fairly frequent design flaws and deficient selective criteria used for diachronic corpora of old documents of all stages of the history of Spanish.

CORDE, CDH, and the historical subcorpus of *Corpus del Español* (CdE) have similar limitations as regards exploitation. Thus, CORDE is not lemmatized, is not tagged, and does not allow graphical representation of queries.¹² The CDH corpus, based on newer technology, is partially lemmatized and tagged, and sustains queries by genre and subject topic. Still the classification used is so detailed that accurate data for research on diatopic and diaphasic variation are rather difficult to retrieve. Queries by absolute and relative frequency also have limitations as regards graphical representation. Additionally, these corpora contain many documents which are according to their own edition guidelines, so a range of editorial policies coexist. The historical subcorpus of CdE uses an extremely deficient classification and description, and is also flawed by frequent errors in the links supplied for retrieval of the source samples.¹³

Some shortcomings of Spanish diachronic corpora are summarized below (Díaz Bravo 2018: 580):

- A lack of balance across centuries and genres, such that medieval texts are comparatively underrepresented in the CORDE corpus and in the CDH corpus.
- A lack of unified editorial criteria across corpus projects.
- Frequently inaccurate document metadata, often lacking in uniformity across corpus projects.
- Despite the many lemmatizers and taggers available, a bias towards contemporary Spanish that demands much-needed improvement in pre-20th c. data.

Specific methods have been made available to overcome some of the abovementioned difficulties based on digital resources and tools, and which allow consistent and accurate analysis of a number of topics by use of diachronic corpora. The historical corpora *Post Scriptum* and *Oralia Diacrónica del Español* (hereafter, ODE) set several prime examples, as their research procedures and technology rely on

¹² Albeit rather poorly, regular expressions are possible, with items like * or ?, as well as with logical operators like AND, OR, or NOT.

¹³ *Sketch Engine* is a completely different case, in that it is a digital tool for language research comprehensive of a range of subcorpora of various languages amounting to millions of words and relying on newer technology capable of sustaining multiple research options (concordances, collocations, frequency lists, etc.). Based on its more complex tagging and lemmatization, it allows queries by morphological tag, lemma, and text type, among others, as well as by other encoding procedures (<https://www.sketchengine.eu/#blue>).

the widely acknowledged TEI¹⁴ standard for XML (*eXtensible Markup Language*) mark-up text encoding and organization. Additionally, TEITOK, specifically designed by Janssen (2016) as a corpus storage platform and based on TEI-compliant, XML text conversion, is gaining ground as a widespread reference of use.¹⁵

XML exceeds .doc and other files in many respects, e.g. by not needing specific hardware or software and thus being interoperational, by being reusable as various formats, by separating form from contents, and by being extensible as a result of not using a closed tagset. It also allows text data organization and modelling by use of accurately defined, machine-readable marks, whether the data is of the structure of the document, editorial, of a semantic nature, of abbreviations, of images, etc. The marks are formalized as tags to describe a text segment. In XML, marks are arranged as angles in customizable modules according to various purposes.

XML text formatting and the TEI *Guidelines* allow text encoding comprehensive of all the textual and metatextual properties desired in a given corpus. The basic structure in TEI-compliant XML text conversion consists of: i) a header (<tei-Header>); and ii) a body (<body>). The header includes any metadata considered relevant. In old texts, they cite the title, and data on its creation, like the source archive identification and the reference signature of the bundle and page, the text type, and, if available, information of where, when, and by whom it was produced. The following example (Figure 1) is for a last will signed in Badajoz in the 17th c., of which some metadata have been recorded in the ODE corpus:

The above data selection must have been made early in the stage of corpus design, also because it is according to the type of corpus and to the samples used.¹⁶ Metadata encoding allows efficient data retrieval for various research projects.

Otherwise, the body of an XML, TEI-compliant document is the text transcript. Text transcription entails the use of a range of tags according to the TEI *Guidelines*, for identification of textual and paratextual information: Beginning of new page, beginning of new line, font change, side notes, crossed out material, overwriting, signatures, as well as strictly linguistic data (e.g. instances of /θ/ spelt as s, unstable use of liquid consonants, borrowings). Figure 2 shows the most basic structure of a page taken from a late 17th c. dowry letter:

14 The *Text Encoding Initiative* is an international standard available since 1987 specifically designed for the Humanities as regards digital editing of documents and electronic encoding (Díaz Bravo 2018). Still, its use for old documents is rather limited (Calderón Campos 2019).

15 A list of corpus projects thus built is available at <<http://www.teitok.org/index.php?action=projects>>.

16 These parameters are substantially different in spoken and written samples, and also if other parameters are added too, e.g. 'informal register', the speaker, their age, their sex, etc.

```

<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns:ff="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Testamento de Francisco Hernández Sevillano</title>
        <editor id="CORTENEX">CORTENEX: Corpus de textos notariales extremeños (siglos XVI y XVII)</editor>
        <funder/>
        <respStmt>
          <resp id="transcription">Inmaculada González Sopeña</resp>
        </respStmt>
        <respStmt>
          <resp id="standardization">Inmaculada González Sopeña</resp>
        </respStmt>
        <respStmt>
          <resp id="annotation">Inmaculada González Sopeña</resp>
        </respStmt>
        <respStmt>
          <resp id="revision"/>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <publisher>UGR, Universidad de Granada</publisher>
        <pubPlace>Granada</pubPlace>
        <distributor>HUM-278. Grupo de Investigaciones Histórico-Lingüísticas y Dialectales. UGR-Junta de Andalucía</distributor>
      </publicationStmt>
      <sourceDesc>
        <msDesc>
          <msIdentifier>
            <country>España</country>
            <settlement>Badajoz</settlement>
            <institution>Archivo Histórico Provincial de Badajoz</institution>
            <repository>Protocolos Notariales</repository>
            <idno>AHPB_PV/289</idno>
          </msIdentifier>
          <msContents>
            <summary>NA</summary>
            <msItem class="original">
              <p><locus>59r-60v</locus></p>
            </msItem>
          </msContents>
        </msDesc>
      </sourceDesc>
    </fileDesc>
  </teiHeader>

```

Figure 1: A cropped screenshot of the XML header of a last will (ODE).

```

<pb n="402v" facs="BA_IMG_2791.jpg"/>
<p>
  <lb/> <gap reason="illegible"/> sarçillos con çinco pendientes de perlas
  <lb/> y dos surtijas, la una grande de piedras verdes,
  <lb/> y la otra con diez con esmaltes vllancos y verdes
  <lb/> y una cruz pequeña con siete piedras verdes,
  <lb/> todo ello de oro que peso treçe rs||reales de a ocho y m<add place="above">
  <lb/> y dos rs||reales de plata, tasado en duzientos y qua<lb/>renta y siete rs||
  <lb/> Una artesa con su banca de madera, tasada
  <lb/> en quatro ducados.
  <lb/> Un medio arcaz de madera tasado en q<add place="above">tro</add>||quatro rs
  <lb/> Un escritorio llano de madera, tasado en
  <lb/> dozienttos rreales.
  <lb/> Quatro silla de vaqueta negras, tasadas
  <lb/> a quatro ducados cada una.
  <lb/> Mas un baul de baqueta negro usado tasado
  <lb/> en sesonta rs||reales.
  <lb/> Vn arca pequeña tasada en veynte rs||reales.
  <lb/> Un catre de madera forrado en treinta y tres r<add place="above">s</add>||r
  <lb/> Vna mesa de madera tasada en treinta y
  <lb/> tres rs||reales.
  <lb/> Un bufete pequeño con su cajon tasado en onze rs||reales.

```

Figure 2: A cropped screenshot of the XML body of a document (ODE).

Thus, specific tags are used for beginning of a new page (<pb/>), for beginning of new paragraphs (<p>), or for beginning of new lines (<lb/>). Figure 2 also shows a mark for the position of a letter above or below a line (<add>). Many tags are also enlarged with attributes and values to enable highly specific queries and meticulous detail about the information contained in a sample, e.g. <gap/> signals omission, which can be enlarged in turn with the attribute ‘reason’ for greater detail. In Figure 2, the latter is “illegible” (the original may have been torn or damaged). By contrast, other tags, e.g. “editorial”, are used for deliberate deletion of an irrelevant excerpt.

Further, various editions (palaeographic, critical) of one and the same sample may be available, and a fully tagged and lemmatized corpus may be stored within the platform TEITOK. Thus, each text is tokenized, i.e. each word is marked up with a token or numbered angled mark detailing the data of its normalized, expanded form, its morphosyntactic tag and its lemma. In old texts, each tokenized sample is processed for orthographic normalization prior to further editing. TEITOK uses a semiautomatic tool to supply a normalized form for each word¹⁷:

(1) <tok id="w-76" nform="había">auia</tok>

Example (1) illustrates the tokenization and normalization of *auia*: Tokenization first converts every form into a token, and orthographic normalization then adds a new attribute (*nform*) to the XML file with a specific value, in this case *había*. The application *Neotag* (Janssen 2012), a tagger that follows the model proposed by EAGLES (*Expert Advisory Group on Language Engineering Standards*), finally tags and lemmatizes each of the resulting forms within the platform.¹⁸ Morphosyntactic tagging translates each morpheme of the grammatical categories of Spanish into an alphanumeric code:

(2) <tok id="w-76" nform="había" lemma="haber" pos="VAII3S0">auia</tok>

¹⁷ Manual post-editing follows to revise any forms that the tool may not have normalized correctly. This is because the process starts from a training corpus large enough as to allocate orthographically standard forms during the first stages of use within the TEITOK platform. The most frequent wrong normalization mistakes in Spanish occur for homographs of the type *se* vs. *sé*, or *mi* vs. *mí*.

¹⁸ Manual post-editing of *Neotag*'s output follows for revision according to what may and may not occur in each word-class. *Neotag* relies on a probabilistic algorithm that often tags wrongly orthographically identical forms and multifunctional words, e.g. *que* (as a conjunction vs. as a relative pronoun).

Example (2) shows *Neotag*'s output of (1), whereby each token is enlarged with two new attributes for specific values: i) the attribute *lemma* (in the example, the infinitive form, because it is a verb); and ii) the attribute *pos*, where the morpho-syntactic tag VAII3S0 stands for “verb”, “auxiliary”, “indicative”, “imperfect”, “3rd person”, “singular”. The segment 0, which stands for the lack of any other morphological contents regardless of the lemma's word-class, must be included on account of the inflection for “gender” in other verbal forms, e.g. in participles.¹⁹

In the end, queries for any linguistic issue can be designed for accurate data retrieval: By transcribed forms, by standardized forms, by lemma, by tag, etc. TEI-TOK also supports CQL (*Corpus Query Language*) for queries combining many of the parameters recorded. The last part of this chapter reviews these technological possibilities in the specific field of research on Arabisms in Spanish.

3 The Research on Arabisms in Spanish

The research on Arabisms in Spanish has attracted the interest of a large number of philologists, partly for the nearly eight centuries of close contact between the varieties of Arabic and of Romance languages spoken in the Iberian peninsula since the Muslim invasion in 711 (Corriente 1977). The number and the range of studies on Arabisms is thus extremely wide and, as a result, difficult to summarize succinctly.

The lexical component of a language is a particularly challenging research field, among other reasons, because it is an ever-changing level that readily echoes diachronically external changes, whether social, cultural, political, etc. Aside from a language's morphological processes for word-formation (derivation, composition), the lexical component relies on borrowings, i.e. on elements that a language takes from another language (Gómez Capuz 2004).²⁰

Borrowings presume contact between languages or language varieties, a key component in the history of Spanish. Language contact is measured in terms of *language influence* and the *intensity of the contact*, and both are evident in the concurrence of linguistic and extralinguistic factors of Arabisms: On the one

¹⁹ The sequence of letters and numbers may vary according to the word-class, e.g. the arrangement of grammatical inflection is not the same in nouns as in pronouns.

²⁰ Borrowings have been approached from a range of theoretical positions (structuralist, formalist, functionalist), each contributing specialized knowledge for their research, e.g. *cultural borrowings*, *integral borrowing*, *lexical* and *semantic calques*, *foreign words*, or *adapted borrowings*, among others.

hand, a large number of Arabisms were borrowed into Spanish, even if there was hardly any morphosyntactic influence for the structural difference between the languages involved; on the other, the contact went through various stages at which a range of differences and sociocultural roles between the Christian and the Muslim populations can be noticed. The latter brings to the fore the relevance of *prestige* for borrowing as a process (Giménez-Eguíbar 2016). Based on the above, the following types of studies on Arabisms can be considered:

- Studies on the periods during which most Arabisms are borrowed, and into which lexico-semantic fields they were borrowed;
- Studies on the decreased borrowing of Arabisms and on their competition with other words, or their replacement by other words (lexical replacement); and
- Studies on the phonetic processes involved in the borrowing of Arabisms, and on the great spelling diversity of Arabisms since the Middle Ages.

The three types cover tens of papers, monographs and handbooks on the periods of the history of Spanish. A good part of the research on Arabisms focuses on the Middle Ages, as this is the period of closest contact and, therefore, when the largest number of Arabisms were borrowed (Oliver Pérez 2004), considering the social constraints imposed by the Christian reconquest. The subject topics thus range from Arabisms in King Alfonso X's works and in mediaeval treaties and chronicles (García González 1998, Maillo Salgado 1998, Neuvonen 1941, Pocklington 1984) to the role of the Mozarabic community in the integration of Arabisms into Spanish (García González 2007). The research on Arabisms reaches as late as the Modern era, mainly with a focus on loss or obsolescence, for the cultural disregard of the Muslim world that was concomitant with Humanism and the Greek and Latin revival (Walsh 1967), and for the competition or replacement by words of non-Arabic stock (Giménez-Eguíbar 2015, 2016). The research on Arabisms during the Spanish Golden Age examines words with low attestation records, or words that occur or remain in highly specific lexico-semantic fields (Calderón Campos 2010, González Sopeña 2017, Morala Rodríguez 2012a). A large amount of research on Arabisms is also available in specific Spanish-speaking geographical areas or on dialects, both synchronic (Garulo Muñoz 1983) and diachronic (Torres Montes 1996).

The 18th c. bears witness to a decrease in research on Arabisms, for the rise of French borrowings during the Age of Enlightenment. Still, lexicography has produced dictionaries and glossaries of Arabisms and word stock of Eastern origin since the 19th c., e.g. by Dozy & Engelmann (1869), or by Eguílaz & Yanguas (1886). The knowledge on Arabisms is finally broadened by the lexicographic information available in general and in etymological dictionaries of Spanish (e.g. the RAE

dictionaries, or DCECH, respectively), and in dictionaries of Arabisms (Corriente 1999).²¹ General descriptions of Arabisms available in handbooks of the history of Spanish can also be found, both classic (Lapesa 1981 [1942]) and newer publications (Giménez-Eguíbar 2023).²²

Indeed, the diachronic relevance of Arabisms lies behind the substantial number of titles on the topic. Even so, a significant loss of specific information on Arabisms is lost in previous databases. This is either for the technical limitations of the corpora reviewed above which supplied the experimental evidence for many of the references cited, or for being based on non-digital sources. The paradigm shift described earlier in this chapter no longer sustains the latter approach, as the scientific community demands research data to be available online and as open access.

4 Research on Arabisms and Language Corpora: The Case of *Oralia diacrónica del español* (ODE)

In the above, this chapter has described the complexities of lexical research in general. Many of them arise because the lexical level is the most superficial, the most variable, and the most difficult to systematize in a language. Lexical research must rely on sources where all the linguistic properties of a term can be attested so they can be added to exhaustive historical dictionaries²³: Lexical senses, semantic widening and reduction, lexical replacement, earliest attestations, etymological information, actual examples of a term, etc. Therefore, corpus-based diachronic lexical research is one of the most successful methods available at present. The abovementioned sources enlarge and improve the word stock, and record forms that are beyond what can be considered the usually acknowledged most standard, academic vocabulary (Morala 2012b: 200).

²¹ Federico Corriente stands out in the research on Arabisms for his dictionaries, but also for the vast number of papers and monographs he produced on the subject. Arnold Steiger (1932) is another major name, in this case for his research on the phonetic adaptation of Arabisms.

²² Some handbooks on the history of Spanish lexis also describe the words of Arabic origin in Spanish (Dworkin 2012, Colón Doménech 2002).

²³ In the case of Spanish, this task has been frequently interrupted for various reasons. At present, two historical dictionaries are pending completion. CDH, which lays the foundations for the NDHDL / *Nuevo diccionario histórico de la lengua española*, currently underway, has resumed the task of a historical dictionary of Spanish.

Specific issues arise in the case of Arabic lexis, like the great variation in spelling according to their written records, and the subsequent difficult systematization of all the possible variants of one Arabism. The description of such words in other documents was not even possible until recently, as many of the existing corpora rely on highly formal sources.

The following presents some of the advantages of new methods in the Digital Humanities. They have improved research on Arabisms in Spanish by use of the technology that is behind the corpus *Oralia diacrónica del español* (hereafter, ODE).

The methodological principles underlying the ODE corpus are part and parcel of technical resources designed specifically for corpora. The ODE corpus, a medium-sized specialized corpus, consists of three text types that go beyond the typical formal register sources used in corpora: i) goods inventories; ii) witness testimonies; and iii) incident reports (Calderón Campos 2019). Geographically, most of the text samples are from Andalusia, even if control subcorpora of other provinces in the country are also available for the period between 1492 and 1833. It is worth remarking at this stage that the ODE corpus builds on the CORDERE-GRA (*Corpus diacrónico del español del reino de Granada 1492–1833*), even if marked differences arise from the two, e.g. the geographical area covered, the use of XML-marked-up, TEI-compliant samples, or access within the TEITOK platform (Janssen 2016).

The latter two features help overcome obstacles such as the ones reviewed in Section 2 above, e.g. use of queries for transcribed text alone (i.e. excluding metadata), visualization as various types of edition (semipalaeographic, standardized), retrieval of every possible spelling variant, and availability of a corpus tagged and lemmatized according to consistent philological criteria, and of the linguistic information contained therein by use of accurate queries (Calderón Campos & Vaamonde 2020: 177).

As a result of the implementation of all levels of textual and metatextual encoding reviewed above, each XML-marked-up, TEI-compliant sample of the ODE corpus may display a header with information such as the year of production, the author, the text type, the archive, the title, the geographical area it comes from, etc. Each transcript contains tags with paratextual and with linguistic information. As all samples are morphosyntactically tagged and lemmatized, CQL queries allow exhaustive, accurate data retrieval. Figure 3 shows various query options based on the encoding described.

Data retrieval may be as transcribed, expanded, or normalized text, and by lemma or by tag. A range of parameter combinations can be used too as CQL queries (e.g. determiner + noun). Lemmatization allows retrieval of all possible variants of a term, even the least likely ones. The text type is a key point here, as the least formal texts, like those included in the ODE corpus, often contain terms

Oralia Diacrónica del Español
EN IES

Búsqueda en COL: constructor de consultas | visualizar | opciones

Constructor de consultas

Búsqueda del texto

Forma transcrita

Forma expandida

Forma normalizada

Etiqueta POS constructor de etiquetas

Lema

Búsqueda del documento

Título

Año

Lugar

Provincia [seleccionar]

Tipo textual [seleccionar]

Siglo [seleccionar]

Archivo [seleccionar]

Buscar en: Texto

Más tipos de búsqueda

- Utilice la Búsqueda comparada para realizar dos o más búsquedas de forma simultánea y poder así comparar resultados.
- Utilice la Búsqueda en el mapa para visualizar el resultado de una o más búsquedas en un mapa.
- Utilice la Búsqueda genérica para buscar cualquier palabra en los documentos XML (cabecera y texto).
- Utilice la Búsqueda con XPath para buscar en la estructura Jerárquica de los documentos XML mediante lenguaje XPath.

Lista de documentos

Powered by «TEI, 5.0.0»
Maarten Janssen, 2014

UNIVERSIDAD DE GRANADA
D. I. E. S.

Figure 3: The ODE corpus query interface.

written as the scrivener would best understand in the communicative situation in question. Thus, a query for the Arabism *tahalí* ‘sheath’ (Figure 4) retrieves unexpected forms like *taali* or *taxali*:

context	de lana blanca bordados y	taali	. Vna vasquiña y vn	1661	España, Badajoz, Badajoz
context	larga y vn frasco y	taali	en ochenta y ocho rs	1701	España, Jaén, Villacarrillo
context	olan de cristal y un	tahali	de vezero vordado de plata	1666	España, Badajoz, Badajoz
context	en quatroçientos rs. Un	tahali	de cordouan con fluecos negros	1666	España, Badajoz, Badajoz
context	calzetas nuevas, mas dos	tahalies	, vno de vaca y	1677	España, Cáceres, Cáceres
context	de a bara. Vn	talai	Vna bandola de vaqueta.	1704	España, Murcia, Lorca
context	con su mangas; un	taxali	de tela parada plateada;	1661	España, Badajoz, Badajoz

Figure 4: *Tahalí* ‘sheath’ in the ODE corpus. Query by lemma.

A similar case can be cited for *jáquima* ‘cord lead’ (Figure 5), and the Granadan variant *xaquyma*:

Retrieval of Arabisms allows, for words like *alhaja* ‘jewel’ (Figure 6), not just the concordances and variants, but also the frequency of each spelling variant recorded, as in Figure 4:

Figure 7 shows a different visualization option for Arabisms with multiple spelling variants, both in the singular and in the plural number:

In the example of *gadamecí* ‘garnished leather’, the information retrieved includes spelling variants, but also instances of /θ/ spelt as s, vowel alternation, or the addition of a liquid consonant at the end of the word. This information, com-

context	. Un caveson y una jaquima . Dos pleytas grandes.	1752	España, Sevilla, Osuna
context	seis. Un caveson y jaquimas en dos reales. Dos	1752	España, Sevilla, Osuna
context	cuchillos de coçina. Una xaquima nueva. Abriosse otro	1663	España, Cáceres, Cáceres
context	rastrillo. Dos cadenas de xaquima de mula. Vna guarniçion	1564	España, Cáceres, Cáceres
context	de la brida. Vna xaquima vieja de cañamo.	1576	España, Cáceres, Cáceres
context	de cavallo viejo. Tres xaquimas nuevas de cañamo. Dos	1564	España, Cáceres, Cáceres
context	tres o [...]. Yten tres xaquimas de cañamo guarneçidas de	1580	España, Sevilla, Sevilla
context	muerto naturalmente, desatada la xaquyma , que se abia sacado	1578	España, Granada, Iznalloz

Figure 5: *Jáquima* ‘cord lead’ in the ODE corpus. Query by lemma.

Corpus Distribution

Search Query	Lemma = alhaja		
Group query	Expanded form		
Total	129		
Reference size	1210347		

Graph: | Count: | Download:

Group	Count	WPM	Percent
alajas	46	38.01	35.66
alaxas	33	27.26	25.58
alhajas	23	19	17.83
Alajas	8	6.61	6.2
alhaxas	6	4.96	4.65
Alaxas	3	2.48	2.33
Alhajas	3	2.48	2.33
alagas	2	1.65	1.55
âlaxas	1	0.83	0.78
alaxittas	1	0.83	0.78
alajitas	1	0.83	0.78
Alhaxas	1	0.83	0.78
alfajas	1	0.83	0.78

Figure 6: The frequency of the variant forms of the lemma *alhaja* ‘jewel’ in the ODE corpus.

bined with metadata of geographical location, helps track down the evolution of such processes over time.

Yet, research on diachronic and diatopic lexical variation, especially on Arabisms, requires multilayer encoding: XML-mark-up, TEI-compliance, TEITOK access. Specifically, corpus lemmatization is a key requirement, even more in the

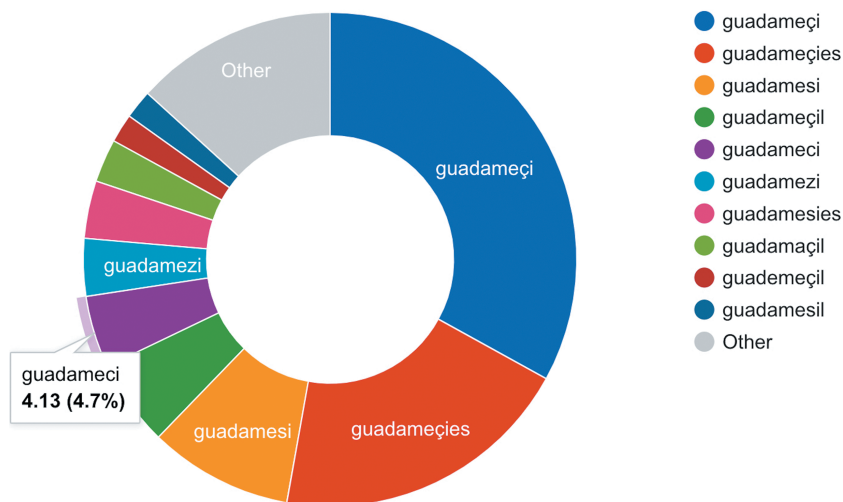


Figure 7: The frequency of the variants of *guadamecí* ‘garnished leather’ in the ODE corpus.

ODE corpus, because it contains samples that are particularly close to spoken language.

Despite the progress in the research of Arabisms based on a corpus built as described above, there is still room for improvement, e.g. by use of TEI tags (<etym>) for dictionary-making, even if this entails manual tagging of each Arabism during text conversion. Semantic annotation by lexical field²⁴ may mean a major step forward in the classification of such a versatile component as a language’s lexis, but identification of Arabisms requires human intervention too. Finally, XML’s versatility allows use of specific, customizable scripts in TEITOK corpora and, again, retrieval of Arabisms relies on the availability of a list of all the cases contained in the corpus and on their subsequent manual retrieval.

5 Conclusions

Diachronic Spanish corpora are still to face the challenge of the implementation of the technical and digital methods and resources available for language research today. This chapter reviews some of the advantages of textual and meta-

²⁴ Consider proposals like USAS (UCREL Semantic Analysis System) <<https://ucrel.lancs.ac.uk/usas/>>.

textual digital encoding of a corpus samples, as well as issues arising from the use of various taggers and lemmatizers, which are based on present-day data and neglect the language before the 20th c.

Retrieval of normalized or lemmatized forms of Arabic words alongside all their spelling variants is a breakthrough in the research of Arabisms. Other retrieval options by use of specific tags during text conversion still need improving, after careful planning of the corpus design and purpose.

While the methods described for diachronic corpora based on XML mark-up, on the use of the TEI *Guidelines*, and on use within the TEITOK platform bring obvious advantages, other factors must be taken into consideration. Thus, the above specifications are ideal for medium-sized, specialized corpora, but they demand heavy investment in terms of time, human resources, and careful work in the case of the reference corpora of millions of words cited at the beginning of the chapter. XML mark-up and the implementation of the TEI *Guidelines* are demanding too, require previous training, and may eventually result in substantial differences across projects, because they are not rigid directives.

Finally, the use of the methods presented for the ODE corpus and other corpora may help improve many dictionaries: They may carry exhaustive definitions, but they also find difficulties to cite consistent examples from a variety of text types for illustration of diatopic and diastratic variations in Spanish lexis.

Bibliography

- Allés Torrent, Susana (2019): “Sobre la complejidad de los datos en Humanidades o cómo traducir las ideas a datos”, in *Revista de Humanidades Digitales*, 4, pp. 1–28.
- Calderón Campos, Miguel (2010): “Aspectos de la vida social granadina a través de diez arabismos de las actas del ayuntamiento y de las ordenanzas municipales (1492–1552)”, in *Études romanes de Brno*, 2, pp. 179–192.
- Calderón Campos, Miguel: (2015): *El español del reino de Granada en sus documentos (1492–1833). Oralidad y escritura*. Bern: Peter Lang.
- Calderón Campos, Miguel (2019): “La edición de corpus históricos en la plataforma TEITOK. El caso de *Oralia diacrónica del español*”, in *Chimera*, 6, pp. 21–36.
- Calderón Campos, Miguel and Gael Vaamonde (2020): “*Oralia diacrónica del español*. Un nuevo corpus de la Edad Moderna”, in *Scriptum digital*, 9, pp. 167–189.
- CdE = Mark Davies (dir.): *Corpus del español*. <www.corpusdelespañol.org>
- CDH = Real Academia Española: *Corpus del Diccionario histórico de la lengua española*. <<https://www.rae.es/banco-de-datos/cdh>>.
- CHARTA = Belén Almeida Cabrejas (coord.): *Corpus hispánico y americano en la red: textos antiguos*. <<https://www.corpuscharta.es/consultas.html>>.
- CODEA = *Corpus de documentos españoles anteriores a 1900*. GITHE, Universidad de Alcalá. <<https://www.corpuscodea.es/>>.

- CODEMA = Carrasco Cantos, Inés (dir.): *Corpus diacrónico de documentación malagueña*, <<http://www.arinta.uma.es>>.
- Colón Domènech, Germán (2002): *Para la historia del léxico español*. Madrid: Arco/Libros.
- CORDE = Real Academia Española: *Corpus diacrónico del español*. <<https://www.rae.es/banco-de-datos/corde>>.
- CORDIAM = Company Company, Concepción and Virginia Bertolotti (dirs.): *Corpus diacrónico y diatópico del español de América*, Academia Mexicana de la Lengua. <<https://www.cordiam.org/>>.
- COREECOM = Arias Álvarez, Beatriz (coord.): *Corpus electrónico del español colonial mexicano*. <<https://www.iifilologicas.unam.mx/coreecom/>>.
- CorLexIn = *Corpus Léxico de Inventarios*. Universidad de León. <<https://corlexin.unileon.es/el-corpus/>>.
- CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <<https://www.rae.es/banco-de-datos/corpes-xxi>>.
- Corriente, Federico (1977): *A gramatical sketch of the Spanish-Arabic dialect bundle*. Madrid: Instituto Hispano-Árabe de Cultura.
- Corriente, Federico (1999): *Diccionario de arabismos y voces afines en iberorromance*. Madrid: Gredos.
- CREA = Real Academia Española: *Corpus de referencia del español actual*. <<https://www.rae.es/banco-de-datos/crea>>.
- DCECH = Corominas, Joan and José Antonio Pascual (1980–1991): *Diccionario crítico-etimológico castellano e hispánico*. Madrid: Gredos.
- Díaz Bravo, Rocío (2018): “Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias”, in Lidia Bocanegra and Esteban Romero Frías (eds.). *Humanidades Digitales aplicadas*. Granada/New York: University of Granada/Downhill Publishing, pp. 562–586.
- Dozy, Reinhart Pieter Anne and Willen Herman Engelmann (1869): *Glossaire des mots espagnols et portugais dérivés de l'arabe*. Leiden: Brill.
- Dworkin, Steven (2012): *A History of the Spanish Lexicon. A linguistic Perspective*. Oxford: Oxford University Press.
- Eguilaz y Yanguas, Leopoldo (1886 [1974]): *Glosario etimológico de las palabras españolas de origen oriental*. Granada: La Lealtad.
- Fontanella de Weinberg, María Beatriz (1993): *Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII*, Anejo 53. Madrid: Boletín de la Real Academia Española.
- García González, Javier (1998): “Clases de arabismos en los textos alfonsíes”, in Claudio García Turza et al. (eds.). *Actas del IV Congreso Internacional de Historia de la Lengua Española*, vol. 2. La Rioja: Universidad de La Rioja, pp. 127–136.
- García González, Javier (2007): “Una perspectiva sociolingüística de los arabismos en el español de la alta Edad Media (711–1300)”, in Inmaculada Delgado Cobos and Alicia Puigvert Ocal (eds.). *Ex admiratione et amicitia: homenaje a Ramón Santiago*, vol. 1. Madrid: Ediciones del Orto, pp. 523–548.
- Garulo Muñoz, Teresa (1983): *Los arabismos en el léxico andaluz (según los datos del Atlas lingüístico y etnográfico de Andalucía)*. Madrid: Instituto hispanoárabe de cultura.
- Giménez-Eguilbar, Patricia (2015): “Dos casos de sustituciones léxicas: los arabismos alfayate y alfajeme”, in Francisco Javier de Cos Ruiz and Mariano Franco Figueroa (coords.). *Actas del IX Congreso Internacional de Historia de la Lengua Española*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 1413–1427.
- Giménez-Eguilbar, Patricia (2016): “Attitudes toward Lexical Arabisms in 16th Century Spanish Texts”, in Sandro Sessarego and Fernando Tejero-Herrero (eds.). *Spanish Language and Sociolinguistics Analysis*. Amsterdam/Philadelphia: John Benjamins, pp. 363–380.

- Giménez-Eguíbar, Patricia (2023): “La contribución del árabe al hispanorromance”, in Steven Dworkin *et al.* (eds.). *Lingüística histórica del español. The Routledge Handbook of Spanish Historical Linguistics*. Londres: Routledge, pp. 362–371.
- Gómez Capuz, Juan (2004): *Los préstamos del español*. Madrid: Arco/Libros.
- González Sopeña, Inmaculada (2017): “Arabismos y fiscalidad”, in *Dicenda*, 35, pp. 109–130.
- Graham, Shawn, Ian Milligan and Scott Weingart (2015): *Exploring Big Historical Data. The Historian’s Macroscope*. London: Imperial College Press.
- Gries, Stefan Th. (2009): “What is corpus linguistics”, in *Language and Linguistic Compass*, 3, pp. 1–17.
- Janssen, Maarten (2012): “Neotag: a POS tagger for grammatical neologism detection”, in *Proceedings of the tenth international conference on Language Resources and Evaluation*. Istanbul, Turkey, pp. 1–7.
- Janssen, Maarten (2016): “TEITOK: Text-Faithful Annotated Corpora”, in *Proceedings of the Language Resources and Evaluation Conference*. Portoroz, Slovenia, pp. 4037–4043.
- Jódar Marín, Juan Ángel (2010): “La era digital: nuevos medios, nuevos usuarios y nuevos profesionales”, in *Razón y Palabra*, 71, pp. 1–12.
- Lapesa, Rafael (1981): *Historia de la lengua española*. Madrid: Gredos.
- Leech, Geoffrey (1992): “Corpora and theories of linguistics performance”, in Jan Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, pp. 105–122.
- Maillo Salgado, Felipe (1998): *Los arabismos del castellano en la Baja Edad Media*. Salamanca: Universidad de Salamanca.
- Martínez-Gambia, Ricardo (2016): “Big data en humanidades digitales: de la escritura digital a la lectura distante”, in *Revista chilena de literatura*, 94, pp. 39–58.
- Morala Rodríguez, José Ramón (2012a): “Arabismos en textos del siglo XVII escasamente documentados”, in *Revista de Investigación Lingüística*, 15, pp. 77–102.
- Morala Rodríguez, José Ramón (2012b): “Léxico e inventarios de bienes en los Siglos de Oro”, in Gloria Clavería Nadal, Margarita Freixas, Marta Prat Sabater and Joan Torruella Casañas (coords.). *Historia del léxico: perspectivas de investigación*, pp. 199–218.
- Neuvonen, Eero Kalervo (1941): *Los arabismos del español en el siglo XIII*. Helsinki: Finnish Literature Society.
- ODE = Calderón Campos, Miguel and María Teresa García-Godoy (2019-present) (dirs.): *Oralia diacrónica del español*. DiLEs, Universidad de Granada <<http://corpora.ugr.es/ode>>.
- Oliver Pérez, Dolores (2004): “Los arabismos dentro de la historia del español: estudio diacrónico de su incorporación”, in Manuel Cecilio Díaz *et al.* (eds.). *Estudios dedicados a José María Fernández Catón*, vol. 2. León: Centro de Estudios e Investigación San Isidoro, pp. 1073–1095.
- Parodi, Giovanni (2010): *Lingüística de corpus: de la teoría a la empiria*. Madrid: Iberoamericana.
- Pocklington, Robert. (1984): “Nuevos arabismos en los textos alfonsíes murcianos”, *Miscelánea medieval murciana*, 11, pp. 261–295.
- Real Academia Española (2014): *Diccionario de la lengua española*. Madrid: Espasa/Calpe.
- Rojas, Elena (2008): *Documentos para la historia lingüística de Hispanoamérica (siglos XVI a XVIII)*, Anejo 61. Madrid: Boletín de la Real Academia Española.
- Rojo, Guillermo (2016): “Corpus textuales del español”, in Javier Gutiérrez Rexach (ed.). *Enciclopedia de lingüística hispánica*, Vol. 2, pp. 285–296.
- Rojo, Guillermo (2021): *Introducción a la lingüística de corpus*. Londres: Routledge.
- Romero Frías, Esteban (2014): “Ciencias sociales y Humanidades Digitales: una visión introductoria”, in Esteban Romero Frías and María Sánchez González (eds.). *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, pp. 19–50.

- Sánchez González, María (2014): “El Big Data como herramienta para la *e-Research* en entornos infosaturados y complejos”, in Esteban Romero Frías and María Sánchez González (eds.), *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, 131–161.
- Sánchez-Prieto Borja, Pedro (1995): *Textos para la historia del español II. Archivo municipal de Guadalajara*. Alcalá de Henares: Universidad de Alcalá.
- Schöch, Christof (2013): “Big? Smart? Clean? Messy? Data in the Humanities”, in *Journal of Digital Humanities*, 2(3), pp. 2–13.
- Spence, Paul (2014a): “Prólogo”, in Esteban Romero Frías and María Sánchez González (eds.), *Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e-research e investigación en colaboración*. La Laguna: Sociedad Latina de Comunicación Social, pp. 9–12.
- Spence, Paul (2014b): “Centros y fronteras: el panorama internacional de las humanidades digitales”, in Sagrario López Poza and Nieves Pena Sueiro (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*. A Coruña: SIELAE-Janus, pp. 37–61.
- Steiger, Arnold (1932): *Contribución a la fonética del hispanoárabe y los arabismos en el iberorrománico y siciliano*. Madrid: CSIC.
- TEI Consortium (ed.) (2016): *TEI P5: guidelines for electronic text encoding and interchange. Text Encoding Initiative Consortium*. <<https://tei-c.org/guidelines/p5/>>.
- Torres Montes, Francisco (1996): “Nombres de medidas agrarias en la costa del antiguo Reino de Granada”, in Juan de Dios Luque Durán (ed.), *Segundas jornadas sobre el estudio y enseñanza del léxico*. Granada, pp. 265–282.
- Torruella, Joan and Johannes Kabatek (2018): *Portal de corpus históricos iberorrománicos (CORHIBER)*. <<http://www.corhiber.org/>>.
- Walsh, John (1967): *The Loss of Arabisms in the Spanish Lexicon*. Unpublished doctoral thesis. Virginia: University of Virginia.

Miguel Calderón Campos

Corpus Size and Tagging: Methodological Strategies for Research on the History of Diminutives *-ito*, *-illo*, and *-ico*

1 Introduction

In a paper published nearly two decades ago, Mair (2006) advocated for the integration of two traditions in corpus linguistics, which have historically operated independently and, at times, in opposition: the *small-and-tidy* approach, which stresses in-depth philological examination of carefully curated corpora; and the *big-and-messy* approach, which highlights the benefits of computer-assisted techniques for the analysis of massive amounts of imperfect data.

In the case of Spanish, this contrast shows in three corpus categories. The first group is small-and-tidy corpora (CHARTA; CODEA +2022; CorLexIn; Corpus Mallorca, etc.), characterized by topic specialization, smaller size, high text quality, and fine-grained tagging. Two such specialized corpora have been used for this paper: *Post Scriptum* (hereafter, PS) (Vaamonde 2015, 2018a, b) and *Oralia diacrónica del español* (hereafter, ODE) (Calderón Campos 2019a, b; Calderón Campos & Díaz Bravo 2021; Calderón Campos & García-Godoy 2023; Calderón Campos & Vaamonde 2020, 2023; González Sopeña 2023). Both corpora were built within the TEI-TOK platform (Janssen 2016; Arrabal Rodríguez 2022) and amount to ca. one million word each of unpublished 16th c. to 19th c. manuscripts: personal correspondence (PS), and goods inventories and witness testimonies in trials (ODE). Lemmatization and morphosyntactic tagging were manually checked in the two corpora.

The second group is big-and-messy corpora, typically of the type of *Sketch Engine's esTenTen18*. This corpus contains 17,000 million words of lemmatized, morphosyntactically tagged texts. It relies on an extremely flexible concordancer interface that supports CQL and regular expressions, but it also evidences some neglect of geographical metadata and text quality. This paper uses the ca. 2,000-million-word corpus *European Spanish Web 2011 (eseuTenTen11)*.

The third group is that of the Royal Spanish Academy's (hereafter, RAE) so-called reference corpora, intended to combine the best of the former two groups.

Note: This contribution has been realized in the framework of Grant PID2022-136256NB-I00, funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU. Also, it has been carried out within the framework of Grant C-HUM-038-UGR23 funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021–2027.

This paper uses the two main reference corpora of Spanish: *Corpus del español del siglo XXI* (hereafter, CORPES XXI) and *Corpus del diccionario histórico de la lengua española* (hereafter, CDH). Each of these amounts to ca. 400 million words. CORPES XXI overviews contemporary Spanish based on representative data of a range of text types of every Spanish-speaking country. The CDH corpus contains documents spanning from the 12th c. to 2005 based on the texts of the first RAE corpora, namely the *Corpus de referencia del español actual* (hereafter, CREA), and the *Corpus diacrónico del español* (hereafter, CORDE). Both CDH and CORPES XXI are lemmatized and have basic morphosyntactic annotation. Their concordancer is reportedly not particularly flexible (Calderón Campos 2024).

This chapter addresses four questions on the history of *-ito*, *-illo* and *-ico* diminutives in Spanish. Corpora of these three groups are used for each question, so their different sizes, tagging and concordancer interfaces result in a complementary methodological approach. In what follows, this chapter is arranged as four sections, one for each of the questions listed below:

1. How can present-day preference for *-ito* diminutives be proved quantitatively? The aim is to bring to light potential limitations in the morphosyntactic tagging of diminutives in large corpora. This is based on evidence retrieved from CORPES XXI and from *Sketch Engine's European Spanish Web 2011 (eseuTenTen11)*.
2. Which was the most frequent diminutive between the 16th and the 19th centuries? This is based on the specialized corpus PS, a ca. one-million-word corpus of 2,447 personal letters in Spanish written between 1510 and 1833.
3. What differences can a letter corpus such as PS and a general corpus such as CDH bring to light on the usage of diminutives in Spanish? The 18th c. data are researched based on CDH-XVIII, an ad-hoc subcorpus of the general corpus CDH, and also on the PS corpus. This section explores how far evidence from different corpora (letter vs. general) can be considered to apply beyond the limits of their corpora.
4. What was the usage of *-ito*, *-illo* and *-ico* in 18th c. Andalusia? Specifically, what was the frequency of each diminutive in two geographically distant provinces, a western one (Cádiz) and an eastern one (Granada)? This is based on the goods inventories of the corpus ODE.

These questions are intended not just to further the knowledge of the history of diminutives in Spanish, but also to put forward methodological strategies or a methodological proposal in which five corpora are used in combination, namely PS, ODE, CDH, CORPES XXI, and *eseuTenTen11*.

A note on how accurately diminutives are tagged is in order here, as it is a crucial factor for effective data retrieval. In *esTenTen*, texts are tagged according to the Spanish *FreeLing* part-of-speech tagset. Thus, a diminutive like *pisito* ‘small

apartment’ is tagged NCMS00V, standing, in this order, for noun.common.masculine.singular.*evaluative*, and is lemmatized under *piso*. Similarly, the adjective *pequeñito* is lemmatized under *pequeño* and is tagged as AQVMS00, i.e. adjective.qualifying.*evaluative*.masculine.singular. In both cases, diminutives carry the mark V for *evaluative*, last in order in nouns, and third in adjectives.

The PS corpus and the ODE corpus use a similar tagging system, except that D (for *diminutivo* ‘diminutive’) is used instead of V for marking diminutives (see Table 1).

CDH and CORPES XXI use a much more basic tagging scheme: It does not go beyond word-class annotation (in the examples, noun or adjective). CORPES XXI has tags for categories like ‘common noun’, ‘masculine’, or ‘singular’, but these categories cannot be used for queries.

Table 1: Diminutives lemmatized and tagged according to various tagging schemes.

Corpus	form	lemma	tag
PS-ODE	<i>tinajita</i> jar-DIM ‘small jar’	<i>tinaja</i> ‘jar’	NCFS00D
	<i>chiquitos</i> small-DIM.M-PL ‘tiny’	<i>chico</i> ‘small’	AQDMS00
TenTen	<i>pisito</i> flat-DIM ‘small flat’	<i>piso</i> ‘flat’	NCMS00V
	<i>pequeñito</i> small-DIM.M ‘tiny’	<i>pequeño</i> ‘small’	AQVMS00
CDH	<i>pisito</i> flat-DIM ‘small flat’	<i>piso</i> ‘flat’	Noun
	<i>pequeñito</i> small-DIM.M ‘tiny’	<i>pequeño</i> ‘small’	noun-adjective
CORPES XXI	<i>pisito</i> flat-DIM ‘small flat’	<i>piso</i> ‘flat’	noun, common, masculine, singular
	<i>pequeñito</i> small-DIM.M ‘tiny’	<i>pequeño</i> ‘small’	adjective, masculine, singular, positive

2 The Diminutives *-ito*, *-illo* and *-ico* in Contemporary Spanish

Wrong or incomplete tagging can be a serious obstacle for corpus research on diminutives. This point can be illustrated with the first question addressed in the paper: Which is the most frequently used diminutive in contemporary Spanish, and how much more frequently is it used than the other two? González-Espresati's doctoral thesis (2015) gives a quantitative answer to this question based on a dataset of 500 examples with diminutives, half from the *Corpus de conversaciones coloquiales* (Val.Es.Co, 2002), and half from various types of journalistic texts. In that dataset, *-ito* amounts to 78% of the occurrences, whereas *-illo* amounts to 10%, and *-ico* amounts to 1%.¹

González-Espresati's data can be furtherchecked with CORPES XXI. A query for nouns in *-ito* in the spoken subcorpus of Spain does yield some diminutives (*poquito* 'little bit', *librito* 'small book', *carrito* 'small trolley', etc.), but the bulk of the results is formal matches that do not meet the query profile otherwise (*ámbito* 'domain', *propósito* 'purpose', *meteorite* 'meteorite', *aerolito* 'aerolite', *hito* 'milestone', etc.). The query cannot be refined in this concordancer.

A similar query on the 25-million-word news subcorpus of the newspaper *El Mundo* within the *European Spanish Web 2011* (*eseuTenTen11*) allows more refined queries thanks to the finer tagging and the use of CQL, e.g.:

```
[word = ".*it(o|a)s? & tag="N.+V|AQV.+"]2
[word = ".*ill(o|a)s? & tag="N.+V|AQV.+"]
[word = ".*ic(o|a)s? & tag="N.+V|AQV.+"]
```

Quality results can thus obtained for diminutives in *-ito*, but not for the other two diminutives, where an extremely high number of false positives is returned (see Table 2). Thus, the first true diminutive in *-ico* retrieved (*ratico* 'short while') ranks 60th on *Sketch Engine's* frequency. This is due to merely formal matches that the software wrongly takes for diminutives: *Nico* (short for first names, e.g. *Nicolás*) being the most frequent form), and then others like *médicos* 'doctors', *periódico* 'newspaper', *económico* 'economic(al)', etc. The picture becomes even blurrier in the case of *-illo*, for the highly frequent, lexicalized diminutives in *-illo* / *-illa*, like *vistilla* 'legal proceeding for minor issues', *entradilla* 'article lead', or *gorrilla* 'illegal parking helper' (examples (1) to (3) from *El Mundo*, *eseuTenTen11*):

¹ Other diminutive suffixes are *-ín* (6%), *-ete* (5%), or *-uelo* (below 1%) (González-Espresati 2015: 330).

² Thus, any lemma tagged as a diminutive noun or as a diminutive adjective ending in *-ito*, *-ita*, *-itos*, or *-itas*.

- (1) La *vistilla* ha quedado suspendida.
‘The proceeding has been cancelled’.
- (2) Titular y *entradilla* de la noticia.
‘Headlines and lead of the article’.
- (3) Los aparcamientos llenos hasta la bandera y la calle repleta de *gorrillas*, que o reciben su dinerito, o a saber qué te hacen en el coche.
‘The parking lots full and the street crowded with illegal helpers, who receive their tips, or else who knows what they do to your car’.

Only after manual screening can the automatic tagger’s false positives be discarded, and the definitive results of Table 2 be obtained: At present, the preference for *-ito* (86%), then for *-illo* (13%), and finally for *-ico* (1%).³

Table 2: The percentage of occurrences of *-ito*, *-illo*, and *-ico* in the subcorpus of *El Mundo* (*eseuTenTen11*).

	n° of concordances	false positives	diminutives	%
ito	6623	600	6023	86
illo	1665	762	903	13
ico	1343	1290	53	1

2.1 Lexicalized (opaque) Diminutives vs. Non-lexicalized, Transparent Diminutives

The data retrieved from CORPES XXI and *eseuTenTen11* presented above are evidence of such limitations in corpus tagging to the extent that it may become an obstacle for automatic retrieval of diminutives. It is also evidence of the need for the methodological separation of two types of diminutives: non-lexicalized or transparent diminutives vs. lexicalized or opaque diminutives.

This paper deals with the former, namely non-lexicalized diminutives (also called *transparent* or *common*, NGLÉ: § 9.3), i.e. diminutives with compositional meaning which do not need dictionary information to fully understand their

³ These results are quite close to González-Espresati’s (2015), even if the list of diminutives of that thesis is not limited to *-ito*, *-illo*, and *-ico*.

meaning, e.g. *pisito* ‘small flat’, *poquillo* ‘a little bit’, or *ratico* ‘short while’ in (4) to (6) from the subcorpus of *El Mundo* within *eseuTenTen11*:

- (4) muestra orgullosa su *pisito* de dos habitaciones.
‘proudly shows her *small*, two-bedroom *flat*’.
- (5) Yo sé que es un *poquillo* especial.
‘I know it’s a *little bit* special’.
- (6) En cuanto tenga un *ratico*, entro y lo envío.
‘When I have a *short while*, I’ll go in and I’ll send it’.

By contrast, lexicalized diminutives (also called *non-transparent* and *opaque*) are the diminutives that do not convey compositional meaning, e.g. *vistilla* ‘legal proceeding for minor issues’, *entradilla* ‘article lead’, or *gorrilla* ‘illegal parking helper’, as in (1) to (3).

In most cases, distinguishing between one type of diminutive and another is not too difficult. However, from a historical point of view, it is sometimes very problematic to decide whether we are dealing with lexicalization or a common diminutive (Arrabal Rodríguez 2023b). The ODE corpus contains a considerable number of occurrences of *bufetillo* in alternation with *bufete* (‘portable table for various purposes’), *bufetico*, and *bufetito*. The *Diccionario de Autoridades* (the RAE’s Dictionary of Authorities, 1726–1739) allows the two interpretations, even if with a bias in favor of a case of lexicalization:

bufetillo: diminutivo de *bufete*. En lo literal significa el que es pequeño; pero de ordinario se suele tomar por el que sirve para el tocador de las mugeres, o para adorno de los estrados (*Diccionario de Autoridades*, 1726).⁴

The analysis becomes particularly difficult here, as the example cited by the Dictionary of Authorities is a clear case of lexicalization (“un *bufetillo* pequeño de tocador” ‘a small portable dressing table’), but the real examples do not specify which type of “*bufetillo*” they refer to, making the decision extremely complex.

The main difficulty of questions 2 to 4 above lies precisely in the need for a closed, quantifiable list of non-lexicalized diminutives of each period or geographical area under study, also in view of the issues in automatic tagging and of the ensuing disambiguation needed for cases like *bufetillo*.

⁴ *Bufetillo*: diminutive form of *bufete*. The literal meaning is for small size, but it usually refers to a dressing table for women or for decoration of stages (*Diccionario de Autoridades*, 1726).

3 Non-lexicalized Diminutives in Spanish from the 16th c. to the 19th c.

Few quantitative studies of the preference for *-ito*, *-illo* or *-ico* between the 16th c. and the 19th c. are available in the literature. Lapesa (1981, § 96, 4), based partly on a summary of research by Nández Fernández (1973), declares the following for the 16th and 17th centuries:

El sufijo diminutivo preferido era *-illo*: [. . .] Autores de las dos Castillas usan *-ico* hasta la época de Calderón, sin la limitación geográfica [que después lo ha] hecho exclusivo de Aragón, reino de Murcia y Andalucía oriental. La pujanza de *-ito* se revela en [. . .] Santa Teresa y [. . .] Calderón: en ambos ocupa el segundo lugar de frecuencia [. . .] siguiendo a *-illo*, al que no había de sobrepasar hasta el siglo XIX.⁵

Thus, according to Lapesa, the chronological distribution of diminutives in the Modern era can be summarized as follows:

1. *-illo* prevails until the 19th c.
2. *-ito* ranks the second most frequently diminutive.
3. *-ico* is used without any dialectal bias in the 16th c. and 17th c.

Fontanella (1987: 74–78) gives the following percentages of use in Buenos Aires Spanish:

Table 3: Percentage of occurrences of diminutive suffixes in Buenos Aires Spanish.

	16th c.-17th c.	18th c.
<i>-illo</i>	52%	10%
<i>-uelo</i>	44%	3%
<i>-ito</i>	4%	86%
<i>-ico</i>	–	1%

Fontanella's dataset for the 16th and 17th centuries was limited, and the subsequent conclusions about the lack of occurrences of *-ico* and the low frequency of *-ito* should be taken cautiously. A better dataset was available for the 18th c., and it at-

⁵ The preferred diminutive was *-illo*: [. . .] *-ico* was used in the two Castiles until Calderón, unconstrained by the geographical boundaries [that later made it] exclusive to Aragón, Murcia, and eastern Andalusia. The strength of *-ito* shows in [. . .] Santa Teresa's and [. . .] Calderón's texts: It stands as the second most frequently used [. . .] after *-illo*, which remained unsurpassed until the 19th c.

tests a marked preference for *-ito* in Argentinian Spanish compared with *-illo* and *-ico*. Indeed, Fontanella's dataset for the 18th c. is very similar to that of contemporary Spanish (cf. Section 2), so the modern pattern of use of diminutives may have become established in the 18th c. instead of in the 19th c., as Lapesa claimed.

Paredes García (2023) recently addressed this question again, and researched the use of *-illo*, *-ito*, *-ico*, *-ino*, *-ejo*, *-uelo*, and *-ete* from the 13th c. to the late 18th c. based on data collected from CODEA+ 2015. Table 4 shows a substantial increase in the use of *-ito* (increased frequency by 2.68 times) from the first to the second half of the century, and a sharp decrease of *-illo* in the same period (the frequency dropped nearly by half). Thus, according to CODEA+ 2015, the tendency changes in the 18th c.

Table 4: Frequency of use of diminutives in the 18th c. (CODEA+ 2015; adapted from Paredes García 2023: 124).

CODEA+ 2015	1700–1749	1750–1799
<i>-illo</i>	68%	31%
<i>-ito</i>	22%	59%
<i>-ico</i>	5%	8%
<i>-ete</i>	5%	2%

The PS corpus allows research of the evolution of diminutives in the Iberian peninsula during the Modern era for two reasons: i) it contains correspondence between various points of the peninsula (see Figure 1); and ii) morphosyntactic tagging was revised manually, which is crucial for research on diminutives, as discussed in Section 1.

Automatic search with Regular Expression 1 can be carried out by substitution of `".*ic(o|a)s?"` or `".*ill(o|a)s?"` for each century:

```
[nform = ".*it(o|a)s?" & pos="(AQD.+|N....D)"] :: match.text_lang = "ES" & int
(match.text_year) >= 1701 & int(match.text_year) <= 1800 within text6
```

A more generic regular expression (Regular Expression 2, below) was used to avoid overlook of false negatives, i.e. to avoid overlook of diminutives that are not tagged as such in the corpus. This regular expression retrieved some diminutives (*poquito* 'little bit', *gusanillo* 'little worm', *malicas* 'sick' etc.) and, especially, a large number of first names ending in *-ico*, *-ito* and *-illo* (*Anica*, *Pepito*, *Manuelillo*, etc.) that are not tagged as diminutives.

⁶ Regular expression 1 for retrieval of diminutives from the PS corpus.

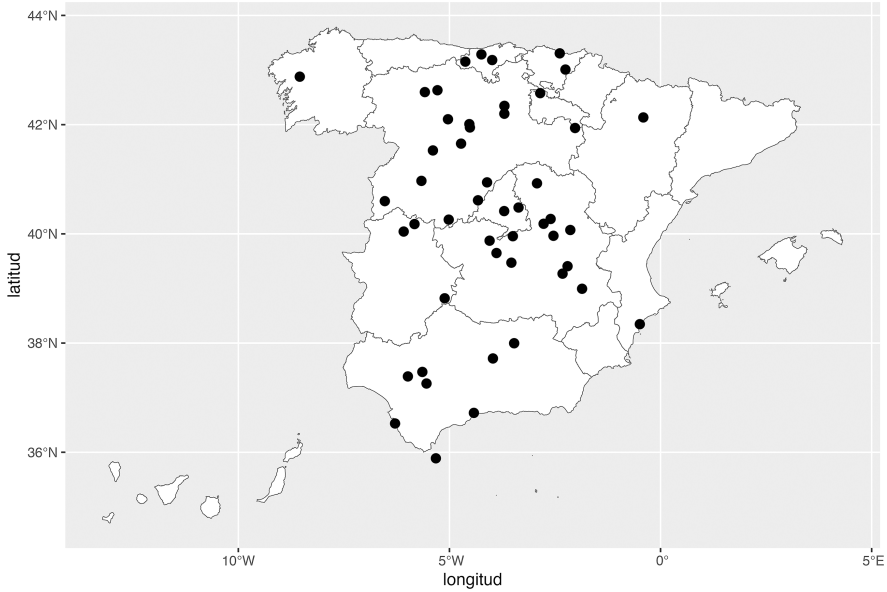


Figure 1: Origin of PS letters containing diminutives.

[nform = ".*it(o|a)s?"] :: match.text_lang = "ES" & int(match.text_year) >= 1701 & int(match.text_year) <= 1800 within text⁷

Table 5 shows the absolute frequency of non-lexicalized diminutives in the PS corpus, including first names. Figure 2 shows the percentage of each diminutive in columns by century.

Table 5: Absolute frequency of diminutives in the PS corpus, including first names.

PS	<i>-ito</i>	<i>-illo</i>	<i>-ico</i>
16th c.	10	21	26
17th c.	51	42	39
18th c.	177	85	26
19th c.	81	20	0

⁷ Regular Expression 2 for normalized variants ending in *-ito*, *-ita*, *-itos* or *-itas* in the PS corpus.

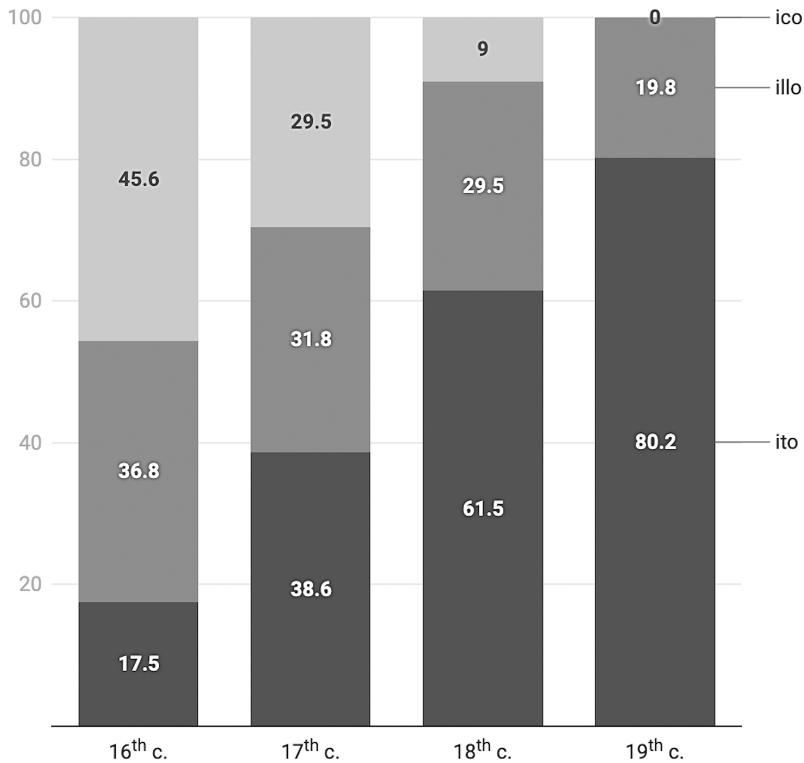


Figure 2: The evolution of *-ito*, *-illo*, and *-ico* in the PS corpus, including first names.

Figure 2 also shows the steady increase of *-ito* since the 16th c.: This form clearly outnumbers the other two suffixes in the 18th century, in accordance with the data from Fontanella and Paredes García.⁸

The strength of *-ico* in the 16th c. waned rapidly: By the 17th c. it was the least frequently used form, and this tendency was reinforced in the 18th c., so it probably became limited to emotional and dialectal connotations.

Finally, *-illo* displays the most stability throughout the period: It ranks higher in frequency than *-ito* in the 16th c., and it becomes the second most frequently used diminutive from the 18th c. onwards.

All in all, the most relevant findings occur in two centuries: i) the 16th c., when *-ico* prevails over *-illo* and *-ito* (in this order of frequency); and ii) the 18th

⁸ The observed proportions of *-ico*, *-ito*, and *-illo* for the 17th century do not allow for significant conclusions.

c., when the frequency of *-ito* is twice as high as the frequency of *-illo* and nearly seven times as high as the frequency of *-ico*.

The following two sections go deeper into the usage of diminutives in the 18th c., first based on CDH data for general tendencies in Spanish beyond what a letter corpus may show, and then with ODE data for the dialectal use of diminutives.

4 Non-lexicalized Diminutives in a General Subcorpus of the 18th c.

The data obtained from the PS corpus for the 18th c. illustrate the usage of diminutives in personal correspondence within the Iberian peninsula. These data need to be compared with those of a reference corpus, where source variety is wider, in order to see whether they are representative of usage in Spanish in general. To this end, this paper relies on a 12.4 million-word subcorpus for 18th c. Spanish extracted from the CDH corpus.⁹

Tagging is very similar in CDH and in CORPES XXI (see Section 2), so diminutives are not easy to retrieve automatically. Thus, a query for nouns ending in *-ico* (**ico*) retrieves mainly formal matches that are not diminutives (*médico* ‘doctor’, *pico* ‘peak’, *eclesiástico* (*sic*) ‘clergyman’, *músico* ‘musician’, *rico* ‘rich man’, etc.), and have to be screened manually.

A ca. one-million-word .txt corpus was built with the concordances obtained from the CDH corpus for more efficient retrieval. Table 6 shows a set of eight CDH queries used to ensure the subcorpus contains enough instances of diminutives.

The first query was intended to retrieve all instances ending in *-ito*, *-ico*, or *-illo* (**ito*, **ico*, **illo*) for the period 1701–1710 in the subcorpus “Spain”. As the CDH corpus allows to export up to 1,000 concordances at a time, the result was downloaded as a .txt file. The file was then screened to delete bibliographical references and other unnecessary metatextual matter, so only text remained. The process was repeated for the queries shown in Table 6.

The corpus thus obtained was researched with *Sketch Engine*, as it allows to obtain a table with frequencies automatically (Figure 3). The resulting table can then be easily used for counts, e.g. of transparent diminutives (*Juanito* ‘Johnny’, *boquita* ‘tiny mouth’, *arbolitos* ‘small trees’, *manchitas* ‘tiny spots’), and for discard of false positives.

⁹ As the PS corpus consists mainly in correspondence of the Iberian peninsula, this paper does not take account of 18th c. American texts, which amount to 5.6 million words.

Table 6: Queries used for retrieval of items ending in *-ito*, *-ico*, *-illo* and their inflected forms from *CDH-Spain*.

Period	Suffixes
1701–1710	<i>-ito, -ico, -illo</i>
1711–1720	<i>-ita, -ica, -illa</i>
1721–1730	<i>-itos, -icos, -illos</i>
1731–1740	<i>-itas, -icas, -illas</i>
1761–1770	<i>-ito, -ico, illo</i>
1771–1780	<i>-ita, -ica, illa</i>
1781–1790	<i>-itos, -icos, -illos</i>
1791–1800	<i>-itas, -icas, -illas</i>

207	<input type="checkbox"/>	Juanito	7
208	<input type="checkbox"/>	Primogenito	7
209	<input type="checkbox"/>	lícitos	7
210	<input type="checkbox"/>	boquita	7
211	<input type="checkbox"/>	bonitas	7
212	<input type="checkbox"/>	arbolitos	7
213	<input type="checkbox"/>	señoritas	7
214	<input type="checkbox"/>	manchitas	7

Figure 3: The frequency of words ending in *-ito* [word=".*it(o|a)s?"] according to *Sketch Engine*.

Table 7 shows the data obtained from the CDH-Spain-18th c. subcorpus of diminutives and from the manual count of the relevant cases contained therein. The table contrasts the absolute frequency of the two 18th c. corpora, namely PS-18th c. and CDH-Spain-18th c.

Table 7: Absolute frequency and percentages of diminutives in two corpora of the 18th c. Iberian peninsula.

	<i>-ico</i>	<i>-ito</i>	<i>-illo</i>
Corpus			
CDH-Spain-18 th c.	38 (2.13%)	1037 (58.3%)	703 (39.5%)
PS-18 th c.	26 (9%)	177 (61.4%)	85 (29.5%)

Pearson's Chi-squared test of the absolute frequencies in Table 7 gives 44.93 (2), i.e. an extremely low p -value (< 0.001) and, therefore, a statistically significant difference in the use of diminutives between the two corpora. By contrast, Cramer's V (a measure of the effect size) is as low as 0.147 and therefore indicates a modest overall effect. Thus, the two variables (corpus type and diminutive suffix) are related, but the relation is weak, and the difference in use may be for several reasons. In such cases, the similarities and dissimilarities in the table must be analysed carefully.

Figure 4 compares the frequency of use of diminutives in PS-18th c. vs. CDH-Spain-18th c., and confirms two main points: i) diminutives rank identically in Spain in the 18th c., such that *-ito* ranks highest, and then *-illo* and *-ico* in this order; and ii) the percentage of use of *-ito* is the same regardless of the corpus (ca. 60%), which is evidence of the strength of this diminutive in this century.

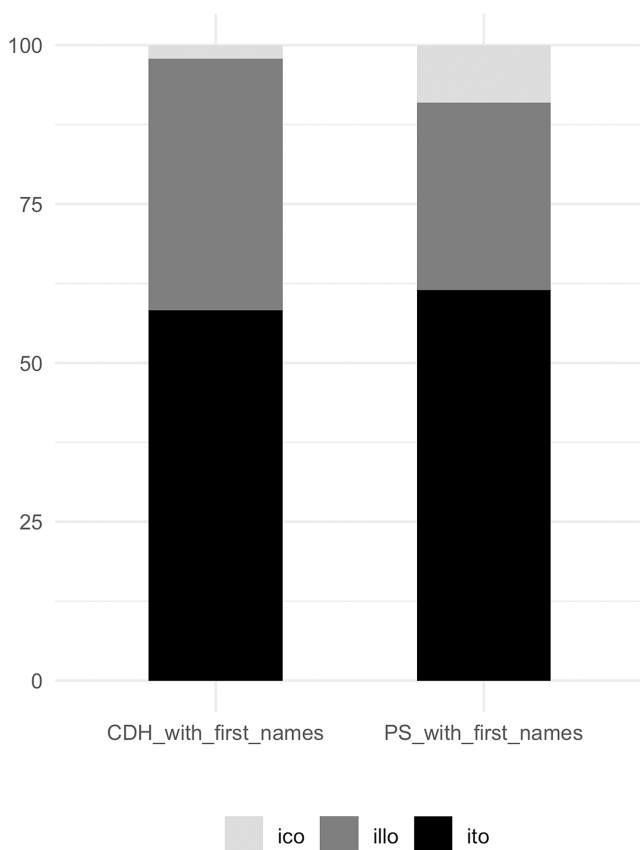


Figure 4: A comparison of the use of diminutives in PS-18th c. and in CDH-Spain-18th c.

Figure 4 also lends itself to a discussion of the different percentages found for *-illo* and *-ico*, and which may be associated with the style of the contents of PS: A much higher percentage of first names with a diminutive suffix can be expected from personal correspondence than from the contents of a general corpus, and this may explain the divergence between the two corpora.

Table 8: Percentage of first names with diminutives in CDH vs. PS for the 18th c.

		First names with diminutives	No. of di- minutives	% of first names with diminutives
Corpus				
<i>-ico</i>	CDH-Spain-18 th c.	6	38	15.7%
	PS-18 th c.	20	26	77%
<i>-ito</i>	CDH-Spain-18 th c.	139	1037	1.3%
	PS-18 th c.	103	177	75%
<i>-illo</i>	CDH-Spain-18 th c.	83	703	11.8%
	PS-18 th c.	21	85	24.7%

Diminutives are overrepresented in the PS corpus as a result of the occurrence of first names (see Table 8): For *-ico* and *-ito*, first name diminutives amount to over 75%. Discard of first names brings CDH and PS in line, as shown in Table 9 and in its graphical representation in Figure 5.

Table 9: Diminutives in CDH and PS after discard of first names.

	<i>-ico</i>	<i>-ito</i>	<i>-illo</i>
Corpus			
CDH-Spain-18 th c. (without first names)	32 (2%)	898 (58%)	620 (40%)
PS-18 th c. (without first names)	6 (4%)	74 (51%)	64 (44.4%)

Statistical analysis of the contingency table (Table 9) gives a *p*-value above 0.05 and a negligible overall effect. This suggests that there is no strong evidence of a difference in the use of suffixes (*-ico*, *-ito*, and *-illo*) between the two corpora.

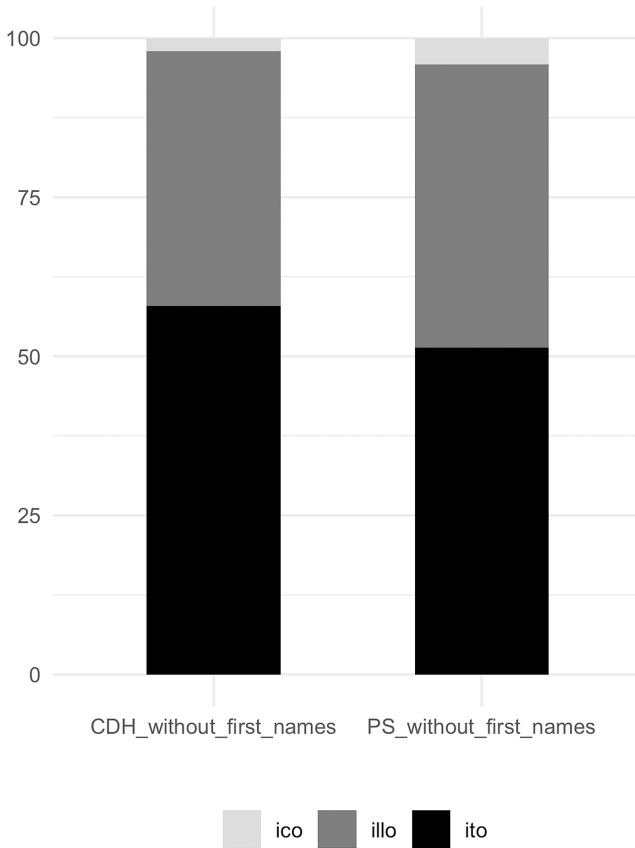


Figure 5: A comparison of the use of diminutives in CDH-Spain-18th c. and 18th c. after discard of first names.

Pearson's Chi-squared test: 4.22 (2), $p = 0.1209$

Cramer's V = 0.05, 95% CI [0, 0.093].

Two conclusions can be drawn from the comparison of CDH-Spain-18th c. with PS-18th c.:

1. The two corpora show more similarities than dissimilarities. They show the same frequency rank of diminutives for the 18th c., whether with or without account of first names, namely *-ito* > *-illo* > *-ico*.
2. The contrast between CDH and PS is related to the latter's oversized first names diminutives. Only minor differences remain between the two corpora after discard of first names.

5 Non-lexicalized Diminutives in 18th c. Goods Inventories of Cádiz and Granada

This section addresses a much more specific question than the former: Whether there is any contrast in the usage of *-ico*, *-ito*, and *-illo* diminutives between western and eastern Andalusia, each represented by Cádiz and Granada, respectively (cf. Arrabal Rodríguez 2023a: 313–361). This point is discussed based on evidence retrieved from the ODE corpus. Like the PS corpus, the ODE corpus was built within the TEITOK platform. At present, it contains a subcorpus of goods inventories amounting to 670,000 tokens. The Granada subcorpus consists of 32 documents and 37,299 tokens, whereas the Cádiz subcorpus consists of 59 documents and 69,844 tokens.

Goods inventories are a particularly adequate text type for research on diminutives, as house surveyors often use evaluative suffixes to describe the quality of the goods assessed (see examples (7) to (9)). Suffixation may describe size (*cajetitas* ‘small boxes’, *botoncillos* ‘small buttons’, *perolico* ‘small pot’), or a subjective view of the state of repair of the item, e.g. *chupilla vieja* ‘old well-worn vest’ in (8),¹⁰ where the surveyor refers to a well-worn garment, not to a small one.

- (7) Dos *cajetitas* ordinarias de plata (ODE, Cádiz, 1703).¹¹
‘Two ordinary *small* silver boxes’.
- (8) Una *chupilla vieja* de lienzo naranjado con *botoncillos* de plata (ODE, Cádiz, 1708).
‘An old, *well-worn* orange cloth *vest* with small silver buttons’.
- (9) Un *perolico chiquito* de cobre con dos asas de hierro en tres reales (ODE, Granada, 1707).
‘A *small*, tiny copper *pot* with two iron handles worth three reales [old currency unit]’.

¹⁰ In the 18th c., *chupa* denoted the vest worn by men over their shirts and under the jacket. Thus, *casaca, chupa y calzón* ‘jacket, wrap and trousers’ (a frequent combination in the ODE corpus), referred to the so-called *vestido a la francesa* ‘French fashion’ that was typical from the last third of the 17th c. until the late 18th c.

¹¹ The ODE examples used are the normalized version; the palaeographic version is available on the corpus’ website.

As in former sections, the comparison between Cádiz and Granada data does not take into consideration *-ico*, *-ito*, and *-illo* in lexicalized derivatives with non-transparent meaning, as in *calzoncillos* ‘underpants’, *pajecito* ‘lamp table’, and *caserillo* ‘home-made cloth’ in (10) to (12).

- (10) veinte y tres sábanas de *caserillo* apreciadas en doscientos y treinta reales (ODE, Cádiz, 1765).
 ‘Twenty-three blankets of *home-made cloth* worth two hundred and thirty reales’.
- (11) Un *pajecito* de el velón en cuatro (ODE, Cádiz, 1710).
 ‘A *candle table* worth four [pesos, currency unit]’.
- (12) Ítem, dos *calzoncillos* blancos (ODE, Cádiz, 1733).
 ‘Idem, two white *underpants*’.

In the 18th c., *calzones* was an outer garment, as in the definition by the *Diccionario de Autoridades* ‘the garment covering the body from the waist to the knee joint’ and in the following example from Seville: “Un vestido que se compone de casaca y calzones de paño fino, color de ámbar, forrada la casaca en tafetán doble” (‘An outfit consisting of a jacket and trousers, both of fine amber-colored cloth and the jacket with a double taffeta lining’) (ODE, Seville, 1720).

Calzoncillos ‘underpants’ are called thus “no porque sean menores” (‘not because they are smaller’), but because “se trahen debaxo de los otros calzones” (‘they are worn under other pants’) (*Diccionario de Autoridades*). The dictionary authors’ former remark fully illustrates the difference between lexicalization and evaluative suffixation, as discussed in Section 2.

Further examples of lexicalization are *caserillo* (‘a home-made cloth for everyday use by contrast with more expensive, imported cloth’) in (10), and *pajecito* (‘[in Andalusia] a small table for lamps and candles’) (11) (*Diccionario de Autoridades*).

The theoretical separation between lexicalized and non-lexicalized diminutives has a significant impact on ODE’s morphosyntactic tagging and lemmatization. The lemma of transparent diminutives is the base of the derivative (*caja* ‘box’, *chupa* ‘vest’, *botón* ‘button’, *perol* ‘pot’, *chico* ‘boy’), whereas the lemma of lexicalized diminutives is the base plus the suffix (*caserillo* ‘home-made cloth’, *pajecito* ‘lamp table’, *calzoncillo* ‘underpants’). The tag for morphosyntactic diminutives includes letter D both in nouns (NCMS00D) and in adjectives (AQDMS0).

The diminutives in the 18th c. goods inventories of Cádiz and Granada are thus easily counted automatically, after manual check of the tags.

ODE data were thus extracted with Regular Expressions 3 and 4:

```
[nform = ".*it(o|a)s?" & pos="(AQD.+|N.....D)"] :: match.text_province = "Granada"
& match.text_cat = "inv" & match.text_century = "XVIII"12
```

```
[nform = ".*it(o|a)s?" & pos="(AQD.+|N.....D)"] :: match.text_province = "Cádiz" &
match.text_cat = "inv" & match.text_century = "XVIII"13
```

The linguistic profile of the query is given between square brackets: [nform = ".*it(o|a)s?" & pos="(AQD.+|N.....D)"]. In this case, the query searches for every normalized instance ending in *-ito*, *-ita*, *-itos* or *-itas* (signalled by the attribute *nform* in ODE) tagged as a diminutive qualifying adjective (AQD.+) or as a diminutive noun (N.....D).

The extralinguistic profile to be combined with the linguistic profile is for diminutives whose province metadata match Granada or Cádiz (match.text_province = "Granada" or "Cádiz"), of the text type "goods inventory" (match.text_cat = "inv") and of the 18th c. (match.text_century = "XVIII"). The forms ".*it(o|a)s?", ".*ill(o|a)s?" and ".*ic(o|a)s?" are substituted so every occurrence is retrieved.

Table 10 shows the frequency of occurrence as absolute values and as percentages. For easier reading, the linguistic variants (diminutive suffixes) are listed as columns, and the categories of explanatory variables or predictors (provinces) are listed as rows (Brezina 2018: 108).

Table 10: The frequency of *-ico*, *-ito*, and *-illo* as absolute values and as percentages for each province (ODE, 18th c.)

Corpus	<i>-ico</i>	<i>-ito</i>	<i>-illo</i>	total
Cádiz	1 (0,7%)	99 (70.2%)	41 (29%)	141
Granada	67 (61.5%)	34 (31.1%)	8 (7.3%)	109

The following statistical results are obtained from Table 10:

Pearson's Chi-squared test: 115.85 (2)

p < 0.001

Cramer's V = 0.681 and 95% Confidence Interval [0.554, 0.802],

Effect size: LARGE EFFECT.¹⁴

¹² Regular Expression 3 for 18th c. Granada diminutives in *-ito*, *-ita*, *-itos*, *-itas*.

¹³ Regular Expression 4 for 18th c. Cádiz diminutives in *-ito*, *-ita*, *-itos*, *-itas*.

¹⁴ Lancaster Stats Tools Online, <http://corpora.lancs.ac.uk/stats/toolbox.php>.

The values of the Chi-squared test and the p -value show a significant association between province (Cádiz vs. Granada) and the diminutive suffix used. The overall effect is large. This description is based on the value of Cramer's V (0.681, 95% CI [.554, .802]), and indicates that the association between the two categorical variables is not only statistically significant, but also practically meaningful. A “large effect” means that the relationship between the variables is substantial and important to consider.¹⁵

A mosaic plot or Marimekko chart (Figure 6) is an effective graphical representation of the data in a cross-tab table. Unlike the contingency table, the mosaic plot typically plots the categories of the explanatory variable on the x axis, and the categories of the linguistic variable on the y axis (Brezina 2018: 110).

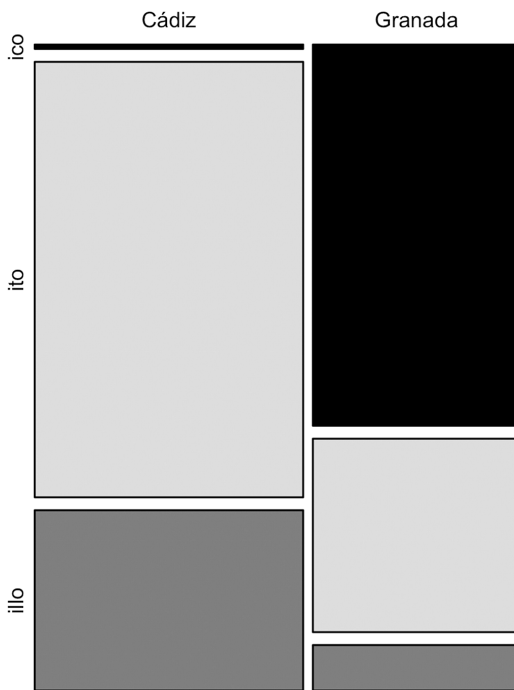


Figure 6: Mosaic plot showing the association between provinces and ODE diminutives in the 18th c.

¹⁵ Cramer's V is a measure of the strength of association between the two categorical variables, ranging from 0 (no association) to 1 (perfect association). The value obtained is 0.681, which suggests a strong association. The 95% Confidence Interval (CI) for Cramer's V: [0.554, 0.802] is the range within which we are 95% confident that the true population value of Cramer's V lies. The wider the interval, the less precise the estimate. In this case, the interval is relatively narrow, so it suggests a reasonable degree of precision.

The y axis of the mosaic plot shows the proportion to which *-ico*, *-ito*, and *-illo* are used in each province. Thus, *-ito* (70.2%, light grey) is preferred in Cádiz, very much unlike *-ico* (lower than 1%); by contrast, *-ico* prevails in Granada (slightly higher than 60%).

The x axis of the mosaic plot compares the usage in Cádiz with that of Granada, i.e. it shows how much more often a given diminutive is used in one province compared with the other. The figure clearly shows a much higher number of occurrences of *-ico* in Granada than in Cádiz, and a much higher number of occurrences of *-ito* and *-illo* in Cádiz than in Granada.

The probability to record a given diminutive in one or the other province is calculated with a measure of the effect size known as *Probability Ratio* (PR), as follows:

PR = probability of outcome of interest in province 1 / probability of outcome of interest in Province 2 (adapted from Brezina 2018: 115).

The probability for *-ico*, *-ito* and *-illo* in each province of Table 10 yields the following results after use of the above PR formula:

PR of *-ico* in Cádiz vs. Granada = $0.7 / 61.46 = 0.011$

PR of *-ico* in Granada vs. Cádiz = $61.46 / 0.7 = 86.66$

PR of *-ito* in Cádiz vs. Granada = $70.21 / 31.19 = 2.25$

PR of *-ito* in Granada vs. Cádiz = $31.19 / 70.21 = 0.44$

PR of *-illo* in Cádiz vs. Granada = $29.07 / 7.33 = 3.96$

PR of *-illo* in Granada vs. Cádiz = $7.33 / 29.07 = 0.25$

Thus, for example, the *-ico* suffix is 86.66 times more likely to occur in Granada than in Cádiz. Similarly, the probability to find *-ito* and *-illo* in the goods inventories of Cádiz is, 2.25 and 3.96 times higher than in Granada, respectively.¹⁶

Finally, Pearson residuals¹⁷ help identify deviations between the observed values and the expected values for each table cell. This signals the cells with particularly large deviations, and supplies additional information on the nature of the association between variables (Cramer's V only shows the strength of the association).

¹⁶ The PR values can be interpreted as follows: A PR value 1 signals no difference between provinces. A PR value lower than 1 signals a lower probability for the variant form in question to occur in Province 1 than in Province 2 (cf. the PR value of *-ico* for Cádiz vs. Granada, and the PR of *-ito* and *-illo* for Granada vs. Cádiz). A PR value higher than 1 signals a higher probability for the variant form in question to occur in Province 1 than in Province 2 (cf. the PR value of *-ico* for Granada vs. Cadiz, and the PR value of *-ito* and *-illo* for Cádiz vs. Granada).

¹⁷ Pearson residuals are calculated by taking the difference between the observed count (O) and the expected count (E), and dividing by the square root of the expected count: $(O - E) / \sqrt{E}$.

Table 11: Expected frequencies vs. observed frequencies (in brackets) (ODE, 18th c.).

Corpus	-ico	-ito	-illo
Cádiz	38.35 (1)	75.01 (99)	27.63 (41)
Granada	29.64 (67)	57.98 (34)	21.36 (8)

Table 11 shows the expected frequency and the observed frequency (in brackets) for each cell. Based on these data, Pearson residuals show positive values to signal a higher observed frequency than expected, and negative values to signal the opposite (cf. Table 12 and Figure 7). The higher the deviation from 0, the larger the effect is. Values higher than 2 are considered significant (+2, -2) (Desagulier 2017:183–185).

Table 12: Pearson residuals (ODE, 18th c.).

Corpus	-ico	-ito	-illo
Cádiz	-6.03	2.76	2.54
Granada	6.85	-3.15	-2.89

Pearson residuals suggest extreme behaviours for the combined categories "Cádiz-*ico*" and "Granada-*ico*". The figure obtained for *-ico* is much lower than expected in Cádiz (negative residual -6.031) and much higher than expected in Granada (positive residual 6.859). A significant association between *-illo* and *-ito* in the two provinces can be identified, as the values are higher than absolute value 2.¹⁸

Figure 7 is a graphical representation to show how the residuals confirm significant differences between Cádiz and Granada in each cell, i.e. two opposite patterns of use of diminutives in the 18th c.

The picture emerging from Figures 6 and 7 thus shows a sharp contrast between diminutive usage in Granada and in Cádiz. The values obtained for Granada resemble the values obtained from the PS corpus for the 16th c. (see Figure 2). By contrast, Cádiz develops in line with the general standard of the PS and the CDH data presented above:

Granada: *-ico* > *-illo* > *-ito*

Cádiz: *-ito* > *-illo* > *-ico*

¹⁸ An absolute value 2 in Pearson residuals is not a clearcut threshold, but it is commonly used for the identification of potentially significant deviations.

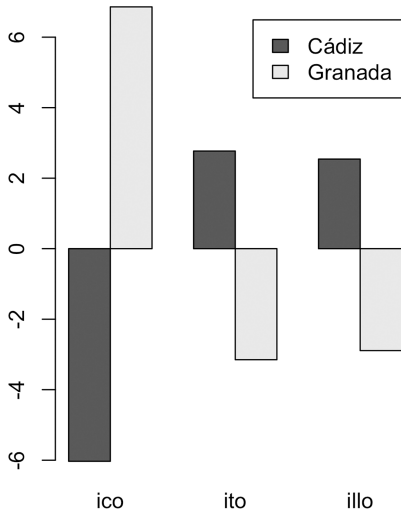


Figure 7: Association plot with Pearson residuals (ODE, 18th c.).

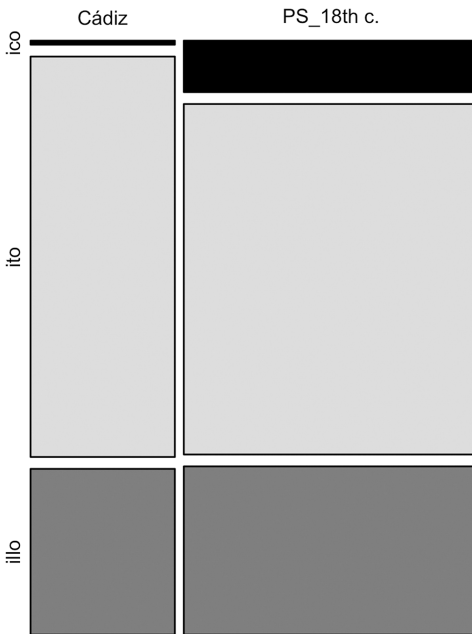


Figure 8: Converging diminutive usage in Cádiz vs. PS-18th c.
 Pearson's Chi-squared test: 11.54 (2), $p = 0.00311$, Cramer's V = 0.164
 95% CI [0.057, 0.253] SMALL EFFECT.

Figures 8 and 9 show mosaic plots of the 18th c. data for Granada and Cádiz, compared with the PS data for the same century. The figure's overview supports the above claim: unlike Granada's dialectal divergence, Cádiz develops in line with the general tendency of 18th c. Spanish. The statistical data also confirm a small effect for Cádiz vs. PS-XVIII and a large effect for Granada vs. PS-XVIII.

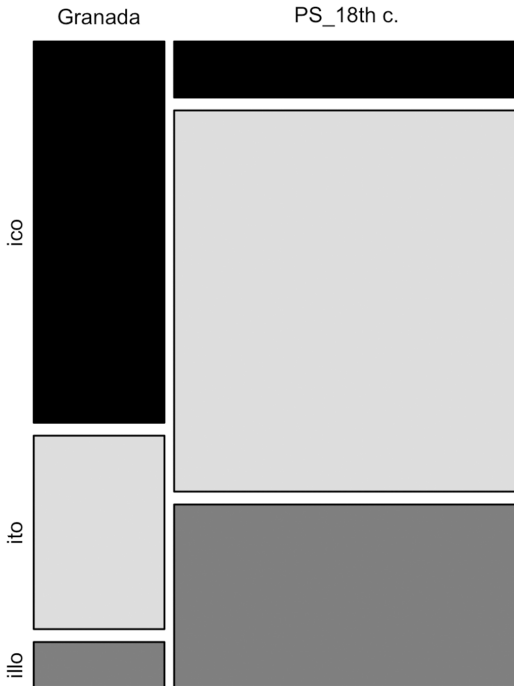


Figure 9: Diverging diminutive usage in Granada vs. PS-18th c.
 Pearson's Chi-squared test: 123.05 (2), $p < 0.001$, Cramer's V = 0.557
 95% CI [0.456, 0.653] LARGE EFFECT.

6 Conclusions

This paper discloses objective difficulties an optimal corpus for research on diminutives must deal with. The main difficulty lies in the inadequacy of automatic tagging, for the formal identity between diminutives and other words ending in *-ito*, *-illo*, or *-ico* (*ámbito* 'domain' vs. *pisito* 'small flat', *médico* 'doctor' vs. *ratico* 'short while', etc.), and also for the extremely subtle difference between lexicalized and non-lexicalized diminutives, especially from a diachronic point of view

(*caserillo* ‘home-made cloth’, *pajecito* ‘lamp table’, *chupilla* ‘well-worn vest’, *calzoncillos* ‘underpants’, etc.).

Diminutives can be analysed automatically only after the corpus’ tagging has been checked manually. As this is viable only in small corpora (ODE, PS), issues about text representativeness arise.

The limitations due to insufficiently accurate tagging can be circumvented by use of versatile concordancers, specifically of the CQL type allowing data collection as frequency lists that can be manually screened for true diminutives.

Despite the above difficulties, the combined use of the five corpora cited in this chapter (PS, ODE, CDH, CORPES XXI, and *eseuTenTen11*) brings to light the following findings on the development of diminutives in Spanish:

1. The 18th c. is the turning point when the current tendency toward *-ito*, then *-illo* and, finally, *-ico* became established. This tendency was reinforced in the 19th c.
2. The data obtained from the PS corpus (a corpus of personal correspondence) and from the CDH corpus (a general corpus) confirm the above rank as regards diminutive use in peninsular Spanish in the 18th c., despite the different size and genre of each corpus.
3. The ODE subcorpus of 18th c. goods inventories reveals a sharp contrast between western and eastern Andalusia (Cádiz and Granada, respectively) as regards the use of *-ito*, *-illo*, and *-ico*.
4. The 16th c. deviates from the other centuries under study in that *-ico* is preferred over *-illo* and *-ito*. This point deserves further attention, especially by comparison of PS data with bigger general corpora.

Bibliography

- Arrabal Rodríguez, Pilar (2022): “TEITOK, a visual solution for XML/TEI encoding: editing, annotating and hosting linguistic corpora”, in RIDE, 15. <<https://ride.i-d-e.de/issues/issue-15/teitok/>>.
- Arrabal Rodríguez, Pilar (2023a): *Variación morfológica y corpus lingüístico: los diminutivos -ico, -ito, -illo en la provincia de Almería (siglos XVIII y XIX)*. Doctoral dissertation supervised by Miguel Calderón Campos. University of Granada.
- Arrabal Rodríguez, Pilar (2023b): “Los sufijos -ico, -ito, -illo en inventarios de bienes de los siglos XVIII y XIX: significado transparente y lexicalizado en el diminutivo”, in Miguel Calderón Campos and Inmaculada González Sopeña (eds.), *Scripta manent. Historia del español, documentación archivística y humanidades digitales*. Lausanne: Peter Lang, pp. 461–485.
- Brezina, Vaclav (2018): *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge: Cambridge University Press.
- Calderón Campos, Miguel (2019a): “Los corpus del español clásico y moderno. Entre la filología y la lingüística computacional”, in *Revista de lingüística teórica y aplicada*, 57(2), pp. 41–64.

- Calderón Campos, Miguel (2019b): “La edición de corpus históricos en la plataforma TEITOK: el caso de *Oralia diacrónica del español* (ODE)”, in *CHIMERA: Romance Corpus and Linguistic Studies*, 6, pp. 21–36.
- Calderón Campos, Miguel (2024): “Spanish Corpora: Big (Quality) Data”, in Gallego Cuiñas, Ana and Daniel Torres Salinas (eds.). *Humanities and Big Data in Ibero-America. Methodological Issues and Practical Applications*. Berlin/Boston: De Gruyter, pp. 109–127.
- Calderón Campos, Miguel and Rocío Díaz Bravo (2021): “An online corpus for the study of historical dialectology: *Oralia diacrónica del español*”, in *Digital Scholarship in the Humanities*, 36(2), pp. 30–48.
- Calderón Campos, Miguel and M^a. Teresa García-Godoy (2023): “ALEA-XVIII: un corpus lingüístico para cartografiar la Andalucía del Setecientos”, in *Études Romanes de Brno*, 44(2), pp. 153–175.
- Calderón Campos, Miguel and Gael Vaamonde (2020): “*Oralia diacrónica del español*: un nuevo corpus de la Edad Moderna”, in *Scriptum digital*, 9, pp. 167–189.
- Calderón Campos, Miguel and Gael Vaamonde (2024): “Anotación y explotación de variantes gráficas de base fonética en el corpus *Oralia diacrónica del español*”, in *Philologia Hispalensis*, 38(1), pp. 301–323.
- CDH = Real Academia Española: *Corpus del Diccionario histórico de la lengua española*. <<https://www.rae.es/banco-de-datos/cdh>>.
- CHARTA = Belén Almeida Cabrejas (coord.): *Corpus hispánico y americano en la red: textos antiguos*. <<https://www.corpuscharta.es/consultas.html>>.
- CODEA +2022 = *Corpus de documentos españoles anteriores a 1900*. GITHE, Universidad de Alcalá. <<https://www.corpuscodea.es/>>.
- CORDE = Real Academia Española: *Corpus diacrónico del español*. <<https://www.rae.es/banco-de-datos/corde>>.
- CorLexIn = *Corpus Léxico de Inventarios*. Universidad de León. <<https://corlexin.unileon.es/el-corpus/>>.
- CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <<https://www.rae.es/banco-de-datos/corpes-xxi>>.
- Corpus Mallorca = *Corpus Mallorca. Documentos castellanos en archivos de las Islas Baleares*. <<https://www.corpusmallorca.es/>>.
- CREA = Real Academia Española: *Corpus de referencia del español actual*. <<https://www.rae.es/banco-de-datos/crea>>.
- Desagulier, Guillaume (2017): *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Cham: Springer.
- Fontanella de Weinberg, Beatriz, (1987): *El español bonaerense. Cuatro siglos de evolución lingüística (1580–1980)*. Buenos Aires: Hachette.
- González Sopena, Inmaculada (2023): “Confusión de sibilantes y neutralización *-r/-l* en los inventarios de bienes andaluces del siglo XVIII”, in *Études romanes de Brno*, 44(1), pp. 141–162.
- González-Espresati García-Medall, Carlos (2015): *Problemas de morfopragmática del diminutivo en español*. Doctoral dissertation supervised by Emilio Ridruejo Alonso y Joaquín A. García Medall Villanueva. University of Valladolid.
- Janssen, Maarten (2016): “TEITOK: Text-Faithful Annotated Corpora”, in *10th conference on International Language Resources and Evaluation (LREC-16)*, pp. 4037–4043. <<https://aclanthology.org/L16-1000>>.
- Lapesa, Rafael (1981): *Historia de la lengua española*. Madrid: Gredos.
- Mair, Christian (2006): “Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora”, in A. Renouf and A. Kehoe (eds.). *The Changing Face of Corpus Linguistics*. Amsterdam/New York: Rodopi, pp. 355–376.

- Náñez Fernández, Emilio (1973): *El diminutivo: historia y funciones en el español clásico y moderno*. Madrid: Gredos.
- NGLE = Real Academia Española y Asociación de Academias de la Lengua Española (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.
- ODE = Calderón Campos, Miguel and M^a. Teresa García-Godoy (2019 -present) (dirs.): *Oralia diacrónica del español*. DiLEs, Universidad de Granada. <<http://corpora.ugr.es/ode/>>.
- Paredes García, Florentino (2023): “Evolución de las formas de diminutivo en español. La historia del cambio -illo > -ito”, in Belén Almeida Cabrejas y Pedro Sánchez-Prieto Borja (coords.). *Varia lección de la lengua española. Estudios sobre el corpus CODEA*. Valencia: Tirant lo Blanch, pp. 97–135.
- PS = P. S. *Post Scriptum*. *Archivo digital de escritura cotidiana en Portugal y España en la Edad Moderna*. CLUL, Universidad de Lisboa. <<http://teitok.clul.ul.pt/postscriptum/>>.
- Vaamonde, Gael (2015): “P. S. Post Scriptum. Dos corpus diacrónicos de escritura cotidiana”, in *Procesamiento del lenguaje natural*, 55, pp. 57–64.
- Vaamonde, Gael (2018a): “Escritura epistolar, edición digital y anotación de corpus”, in *Cuadernos del Instituto Historia de la Lengua*, 11, pp. 139–164.
- Vaamonde, Gael (2018b): “La multidisciplinariedad en la creación de corpus históricos. El caso de *Post Scriptum*”, in *Artnodes: revista de arte, ciencia y tecnología*, 22, pp. 120–129.

II The COSER Corpus and Newspaper Digital Libraries as Alternative Data Sources for Research on Rural and Informal Varieties

Miriam Bouzouita, Johnatan E. Bonilla & Rosa Lilia Segundo Díaz

Gaming for Dialects: Creating an Annotated and Parsed Corpus of European Spanish Dialects through GWAPs

1 Introduction

Despite the recent increase in interest in dialectal grammars by both dialectologists (e.g., Fernández-Ordóñez & Pato 2020) and generative linguists (e.g., Castillo *et al.* 2020), it is well-known that, for Spanish, diatopic microvariation in morpho-syntax has been much less explored in comparison to lexical and phonetic matters. Though there exist sociolinguistic corpora that focus on spoken Spanish, such as those that are part of the *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA, “Project for the Sociolinguistic Study of Spanish from Spain and America”; Moreno Fernández 2005) which encompasses more than 40 research groups gathering oral data from various cities in the Hispanic-speaking world, or the *Corpus Oral de Lenguaje Adolescente* (COLA, “Oral Corpus of Adolescent Language”; Jørgensen *et al.* 2002–2017; Jørgensen & Eguía Padilla 2015) which contains samples of youth speech from Madrid, Buenos Aires, Santiago de Chile and Managua, few of these corpora are morpho-syntactically annotated and/or parsed. There are some notable exceptions though, such as the Spanish part of the *Integrated Reference Corpora for Spoken Romance Languages* (C-ORAL-ROM; Moreno-Sandoval *et al.* 2005; Moreno-Sandoval & Guirao 2006), the *Corpus Oral de Español como Lengua Extranjera* (COR-ELE, “Oral Corpus of Spanish as a Foreign Language”; Campillos Llanos 2016), and the *Corpus del Habla de Baja California* (CHBC, “Corpus of Baja California Speech”; Rico-Sulayes *et al.* 2017).

However, none of these corpora focus on the European Spanish diatopic varieties. In recent years, there have also been some initiatives that center on spontaneous web speech, such as the *Latin American Spanish Discussion Forum Treebank* (LAS-DisFo; Taulé *et al.* 2015), which is not an open access tool and is lodged behind a pay wall. Although this morpho-syntactically annotated corpus also contains non-standard fragments and thus shares the problem with oral corpora that Natural Language Processing (NLP) tools trained on standard written texts perform badly when applied to this type of data (see also section 5.2.3), it should not be forgotten that LAS-DisFo’s focus is on written spontaneous language, which has its own idiosyncrasies, not necessarily shared with its oral counterpart (e.g., typos, emoticons, see Taulé *et al.* 2015).

In sum, up until now there are no morpho-syntactically annotated and parsed corpora available for spoken Spanish dialects. In what follows, we will present an interdisciplinary crowd-sourced project that aims to fill this gap in order to stimulate and enhance more fine-grained dialectal morpho-syntactic research.¹

More specifically, the general aim of this project consists in creating a morpho-syntactically annotated and parsed corpus of the European Spanish dialects, the so-called *Corpus Oral y Sonoro del Español Rural – Anotado y Parseado* (COSER-AP, “Annotated and Parsed Audible Corpus of Spoken Rural Spanish”) and later renamed COSER-UD (COSER-Universal Dependencies, Bonilla *et al.* 2022, 2023). As the name of this new annotated and parsed corpus indicates, its basis is the *Corpus Oral y Sonoro del Español Rural* (COSER, “Audible Corpus of Spoken Rural Spanish”, Fernández-Ordóñez 2005-present), currently the largest collection of spoken Spanish data. Recently, similar initiatives to create annotated and parsed corpora for spoken dialects have been undertaken for other languages too, such as the PARLARS corpus for Valencian Catalan (Esplà-Gomis & Sentí in prep.; Montserrat & Segura 2020) and the *Gesproken Corpus van de zuidelijk-Nederlandse Dialecten* for Southern Dutch (GCND, “Spoken Corpus of Southern Dutch Dialects”; Farasyn *et al.* 2022; Breitbarth *et al.* 2020; Ghyselen *et al.* 2020), among others.

As regards the objectives of this paper, we will present the project design and the challenges that have been encountered in the first two project phases (see section 3) while constructing the COSER-UD through the interdisciplinary approach that was used, in which the fields of Dialectology, NLP and Human-Computer Interaction (HCI) intertwine. As will become clear, this project also employs a citizen science methodology given that linguistic confirmations and corrections are obtained from the general public through various online Games With A Purpose (GWAPs), collectively referred to as *Juegos del español* (Bouzouita *et al.* 2022).

As concerns the structure of this paper, first, we will introduce the COSER corpus on which this interdisciplinary crowd-sourced project is based (section 2). Subsequently, we will present the different project stages and detail the various tasks involved in each stage (section 3). In section 4, we briefly introduce the various GWAPs of *Juegos del español* that have been created for the confirmation and correction of the automatically generated morpho-syntactic tags, while section 5 deals with various tasks of Phase I and II carried out by the Linguistics team, such as the pre-processing of the transcriptions of the COSER corpus (section 5.1), morpho-syntactic annotation of the COSER corpus (section 5.2), which provides

¹ This international research project, titled “A (Respeaking and) Collaborative Game-Based Approach to Building a Parsed Corpus of European Spanish Dialects” (I000418N; PI: M. Bouzouita), has been financed by the Flemish Research Fund (*Fonds voor wetenschappelijk onderzoek*, FWO; 2018–2023).

details on the framework used (section 5.2.1), the creation of the COSER-PoS (COSER-Parts-of-Speech; section 5.2.2), on the results of a study evaluating the tagging accuracy of one of the automatic taggers that have been tested (section 5.2.3) and of the human annotators (section 5.2.4), and on the post-tagging knowledge transfer of the data obtained by the players of *Juegos del español* (section 5.2.5). In the final part, we will draw some conclusions (section 6).

2 The COSER Corpus: Contents and Transcription Protocol

As mentioned before, the COSER corpus (Fernández-Ordóñez 2005-present) is the beating heart of the COSER-UD. In view of this, this section will provide more details on this corpus, its contents, and goals. The primary objective of the COSER corpus is to document diatopic variation, especially morpho-syntactic one, in rural areas of Spain. This database of spoken Spanish is constructed using transcriptions of semi-directed sociolinguistic interviews with elderly men and women living in rural parts of Spain, who have little to no formal education and who enjoyed limited mobility throughout their lives. In other words, the COSER informants coincide largely with those used in traditional dialectology, i.e., the so-called NORM (Non-mobile Older Rural Men, Chambers & Trudgill 1998), though women's speech has also been included in the COSER corpus. As of December 2022, 2,961 informants have been interviewed, 1,415 men and 1,546 women to be precise, for 1,415 locations in 55 provinces and islands. As the COSER focuses on the speech of elderly, the average age of its informants is quite high, 74.2 years to be exact: 75 years for the men and 73.6 for the women. In total, 1,772 interviews have been conducted, for which 1,910 hours of spoken Spanish have recorded, and of which 218 interviews have been transcribed, corresponding to 295 hours and 48 minutes. In other words, at the moment, a mere 12.3% of the total interviews has been transcribed (or 15.4% of the total recorded hours), though we are currently exploring ways to improve the transcription rate significantly using newly developed automatic tools that use large-scale weak supervision, such as Whisper (Radford *et al.* 2023).

As regards the COSER's transcriptions, although that the conventions have changed a few times due to newly acquired insights gained during the corpus construction process, currently, the transcription protocol mixes orthographic and non-standard considerations. More precisely, they try to reflect the pronunciation of mostly phonological (and not phonetic) phenomena that can be found in spoken rural Spanish, while also adopting commonly used spelling rules. This decision distinguishes COSER from other spoken Spanish corpora, such as PRESEEA

and COLA, that have adopted orthographic transcription protocols (see Ghyselen *et al.* 2020 for the advantages and disadvantages of the various types of transcription protocols). The two main phonological changes that have been included in the COSER transcriptions are the omission and addition of phonological segments (de Benito Moreno *et al.* 2016: 79). Let us consider the transcription in example (1), in which the speech of the informant (I1) contains both types of phonological changes (unlike the one of the interviewer, E1):

- (1) I1: *me pagaban el rebaje y ganaba ochocientas cincuenta pesetas **to** los meses.*
 E1: *¿El rebaje qué es?*
 I1: *La **comía**. Lo que cuesta la **comía** en el cuartel.*
 [...] *y yo, ¿qué gastaba? Si diba al bare o al aquello que salía con las chicas, le compraba una peseta **e** golosinas o tal, que una peseta daban... buah. Y, o echaba un café en el bar o cosas de esas. **Demás**, no tenía que gastar. ¿**Pa** qué? **Comía** tenía, pues... Y, ya digo, traje el doble dinero del que llevé.*
 I1: ‘They paid me the rebate and I earned eight hundred and fifty pesetas every month.’
 E1: ‘What is the rebate?’
 I1: ‘The food. What the food costs in the barracks.’
 ‘[...] and I, what did I spend? If I went to the bar or to there where I went out with girls, I would buy her a peseta of candy or something, and they would give a peseta... buah. And, or I would have a coffee at the bar or things like that. Besides, I didn’t have to spend. For what? Food I had, well... And, I’m telling you, I brought twice as much money back as I took.’
 (COSER-5506-01: M, 88 years, El Remo, Los Llanos de Aridane, in La Palma)
- (2) I1: “*Coño, [NP], ¿qué **t’ha** pasao?*”. Digo: “*Pues, tonto este, que **s’ha** dejao el jefe la escopeta [...]*”
 “*Damn, [Name], what happened to you?*” I say: “*Well, this fool, the boss left the shotgun [...]*”
 (COSER-0222-01: M, 82 years, Povedilla, in Albacete)

As can be seen, the COSER transcription presented in example (1) respects, on the one hand, (i) the suppression of phonological segments, represented in bold, as in *to* for *todos* ‘all’, *comía* for *comida* ‘food’, *e* for *de* ‘of’, *demás* for *además* ‘besides’, *pa qué* for *para qué* ‘what for’, and, on the other, (ii) the increase in phonological segments in the underlined items, as in *diba* for *iba* ‘I went’ and *bare* for *bar* ‘bar’, the latter of which is a typical case of paragogic *-e*, commonly found among elderly speakers in the Canary Islands (see Castillo Lluch *et al.* 2022 for more de-

tails). Similarly, non-standard stress changes and the concatenation of sounds are represented in the COSER transcriptions, as for instance in the pronunciation of *pájaro* [paˈxaro] instead of the standard *pájaro* [ˈpaxaro] ‘bird’, or *t’ha* and *s’ha* for *te ha* and *se ha* respectively, as illustrated in example (2) (de Benito Moreno *et al.* 2016: 79, Bonilla *et al.* 2022).

Features that are typical of spontaneous conversations are also represented in the COSER transcriptions, such as overlaps, interruptions, and self-corrections. Concretely, overlaps are inserted within the transcription of the speech of the first speaker at the point where the overlapping fragment starts (though no indication is given on where it ends), and are marked with square brackets followed by HS, which stands for *Habla simultánea* ‘simultaneous speech’, and a specification of who the second interlocutor is, as in [HS:E1] (see the transcription section on the COSER website for more information on other types of overlaps and their transcription). Interruptions are signalled by a hyphen (-), while self-corrections by a vertical bar (|) that indicates that the interruption is followed by a sequence that does not repeat the interrupted fragment but instead is self-corrected (de Benito Moreno *et al.* 2016: 80).

Although these transcription decisions were taken with the aim to represent the original pronunciation as closely as possible, they create additional challenges for the construction of a morpho-syntactically annotated and parsed corpus, such as the tokenization and lemmatization process. To ease this burden, the COSER also adapted a special transcription rule, the so-called disambiguation convention, whereby phonological reductions are restored to eliminate ambiguous interpretations of the item in question. To exemplify, as in spontaneous speech the deletion of the final /-r/ of the infinitive of verbs of the 1st conjugation class (i.e., verbs ending in *-ar*) can give rise to a form that can also be interpreted as the feminine past participle, the speech is disambiguated between brackets using the equals sign (=), as in *cant(á=ar)* ‘to sing’ or *cant(á=ada)* ‘sung’ (de Benito Moreno *et al.* 2016: 80–81; Fernández-Ordóñez & Pato 2020: 76). As the COSER is still under development, the transcription team continues to transcribe the recorded audio and video material in addition to work on the disambiguations.

3 The Creation of COSER-UD: Project Phases and Tasks

Now that the base corpus and its transcription protocol have been introduced, we will briefly discuss the different phases of the current research infrastructure project. The goal of this project consists in creating a morpho-syntactically anno-

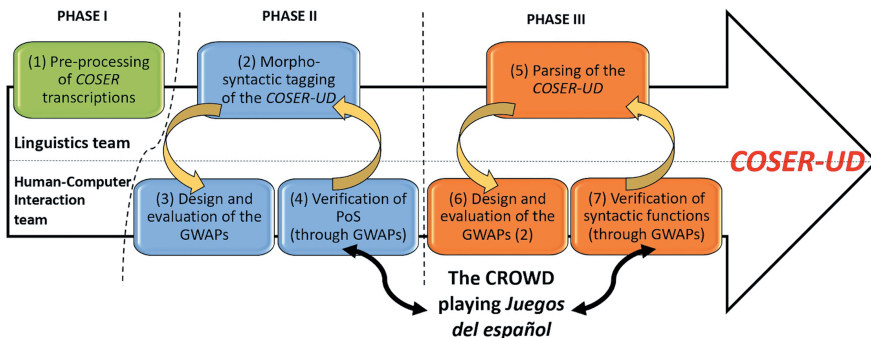


Figure 1: The Creation Process of the COSER-UD.

tated and parsed version of the COSER corpus, originally called COSER-AP (Bonilla *et al.* 2022) and later renamed COSER-UD, due to the Universal Dependencies (UD; Nivre *et al.* 2016, 2020) framework employed for the current research (see section 5.2.1). The resulting COSER-UD, which is presented in a treebank format following the UD guidelines (Bonilla 2022), is the first treebank for oral Spanish.

As illustrated in Figure 1 (adapted from Bonilla *et al.* 2022: 81), we can identify three different project phases, represented by the different colors in the workflow: namely, (i) the COSER pre-processing phase (Phase I), shown in the green rectangle; (ii) the morpho-syntactic annotation stage (Phase II) in light blue; and (iii) the parsing phase (Phase III) in orange. Each task within the different phases is described in the corresponding rectangle, with a numerical sequence indicating the ordering of the tasks. To create the COSER-UD treebank, two specialist teams have been working in parallel, as can be seen by the different levels in the workflow. More specifically, the upper level details the tasks of the Linguistics team, which is responsible for the NLP and dialectology matters of the project, that comprise (1) the pre-processing of the COSER transcriptions, (2) the manual and automatic morpho-syntactic annotation, and (5) the parsing of the corpus, needed to build as an output the COSER-UD treebank (Bonilla 2024a, submitted). Morpho-syntactic annotation or PoS tagging is the procedure whereby a word or a token is assigned a label, which either indicates its grammatical category (e.g., noun, adjective) or its status as a punctuation mark, symbol, or incomplete word. A reference corpus, the so-called Gold Standard (GS) dataset, also termed the COSER-PoS (Bonilla 2024a, 2024b) is created in Phase II. Parsing refers to the process whereby the syntactic function (e.g., subject, direct object, etc.) of a word is identified.

This project also includes an HCI team, given that a collaborative game-based approach has been adopted, whereby the crowd, i.e., (non-expert) members of the public, helps to review the automatic morpho-syntactic annotation and pars-

ing (Bonilla *et al.* 2022; see also Segundo Díaz *et al.* 2023a, 2023b, 2024). As can be seen, the HCI team focuses on the design and evaluation of the various sets of GWAPs (tasks (3) and (6)), that have been specifically developed to verify the automatically annotated PoS and the syntactic functions (tasks (4) and (7)).

These various GWAPs have been collectively referred to as *Juegos del español* (Bouzouita *et al.* 2022). The collaborative aspect of the project is represented in Figure 1 by the two bidirectional black arrows that point to these tasks and the crowd that plays one of the GWAPs included in *Juegos del español*. In other words, the public confirms or corrects the grammatical categories or functions that have been assigned automatically to the words while playing the GWAPs of *Juegos del español*. Importantly, both teams need to collaborate closely and exchange different types of results, as illustrated by the yellow arrows between the two project teams. For example, the results of the morpho-syntactic annotation carried out by taggers and the manual expert validation of the tags are incorporated into the design of the first series of GWAPs. In turn, the verifications (i.e., both corrections and confirmations) by the players of this automatic annotation can, in theory, serve to retrain the language model to improve the accuracy of the automatic PoS tagging. Likewise, the results produced by the parsers can form the basis for the second series of GWAPs, whose goal is to confirm and correct the syntactic functions that have been generated automatically by the parsers. These crowd-sourced verifications can then, in turn, enhance the accuracy of automatic parsing.

It is important to keep in mind that the workflow of the creation process of the COSER-UD presented in Figure 1 is an abstract representation of the various phases and tasks involved in this project and that various of these tasks are complex ones and thus comprise sub-tasks, this is the case for the tasks of both teams. The pre-processing phase of the COSER, for instance, includes the selection of transcriptions based on geographical distribution criteria, the ridding of these transcriptions of marks typical of the COSER transcription protocol, as well as the sentence extraction from the selected texts (see section 5.1).

Similarly, the morpho-syntactic annotation of the COSER-UD, i.e., task (2) (see also section 5.2), can be subdivided into tasks (2.1) the automatic pre-annotation by the best performing tagger, whose pipeline includes the sentence segmentation, tokenization, lemmatization, PoS tagging, and parsing of the data, (2.2) the manual and semi-automatic tag verification by members of the Linguistics team for the creation of the reference model or GS corpus, COSER-PoS, (2.3) the evaluation of the automatic tagging using the developed reference model, and (2.4), the fine-tuning of the language model of the best performing tagger, which, in theory, can be done on the basis of the confirmative and corrective feedback by the members of the general public once obtained and processed.

Likewise, task (3), which concerns the design and evaluation of the GWAPs of *Juegos del español*, carried out by the HCI team, contains various sub-tasks, such as (3.1) the conceptual design of low-fidelity prototypes of the GWAPs (version 1.0), (3.2) the development of high-fidelity prototypes of the GWAPs, (3.3) the evaluation of GWAPs in terms of Player Enjoyment (PE) and the Game Design Elements (GDEs) integrated in the games to study their influence on the PE, (3.4) the improvement of the GWAPs (version 2.0) based on the results of the previous evaluation, (3.5) the implementation of a mechanism to assess the inter-annotator agreement, which automatically accepts confirmations and corrections of tags provided by the players, and (3.6) the implementation of the crowdsourcing environment, in which the GWAPs are launched to the crowd (for more details, see Segundo Díaz 2024: chapters 5–8 and 10).

Finally, it should be pointed out that although Figure 1 represents the various project stages as conceived originally in the project proposal, not all phases and tasks have been concluded during the funded period (until April 2023). This is due to several reasons, such as lesser obtained funding, which resulted in hiring fewer project members, as well as delays due to the physical and mental consequences of the pandemic outbreak on the project members. Although the Linguistics team was able to evaluate the PoS-taggers and fine-tune the language model for the PoS tagging (Bonilla 2024a: chapter 5, 2024b), this was not done using data resulting from the verified tokens provided by the crowd and the HCI team, but based on expert manual validation. Similarly, as regards the parsing phase, though task (5) has very recently been completed by the Linguistics team (see Bonilla 2024a: chapters 7–9, submitted), no new GWAPs have been designed nor evaluated for the verification of the syntactic functions (tasks (6) and (7)) by the HCI team.

In view of this, not all tasks of the workflow will be discussed. In this contribution we will focus mainly on the tasks carried out by the Linguistics team, such as the pre-processing of the COSER transcriptions (task (1); section 5.1) and the morpho-syntactic annotation tasks (tasks (2) and (5); section 5.2). Though the various created game concepts will be introduced in section 4, we refer to Segundo Díaz *et al.* (2022, 2023a, 2023b, 2024) for more details on the results obtained by the HCI team. For more details on the various GWAPs, see Segundo Díaz *et al.* (2023a, 2024).

4 HCI Team’s Tasks: Design and Evaluation of the GWAPs for PoS Verification

While the Linguistics team worked on the pre-processing of the COSER transcriptions (task (1), see Figure 1), the HCI team started designing and evaluating the three GWAPs of *Juegos del español* through which the crowd could help confirm and correct the automatically generated PoS tags (task (3) in Figure 1). As explained in section 3, the design and evaluation of the GWAPs involves various stages, going from creating low-fidelity prototypes of the GWAPs, then high-fidelity ones to the creation of the GWAPs version 2.0 (Segundo Díaz 2024: chapters 5–8). To build engaging games, the HCI team carried out several studies on the GDEs and examined their correlation to PE (see also Segundo Díaz *et al.* 2022). Later, the HCI team also researched the correlations between GDEs, PE and the Personality Traits of the players (PT; for more details, see Segundo Díaz *et al.* 2023b, 2024, Segundo Díaz 2024: chapter 7). Once the Linguistics team completed the automatic tagging process in Phase II (see sections 5.1 to 5.2.2), the relevant data was passed to the HCI team to integrate into the three GWAPs of *Juegos del español*. The HCI team then moved to the fine-tuning, testing, and eventually launching of the GWAPs to the crowd (for further details on the various iterations, see Segundo Díaz 2024: chapters 5–8). In what follows, we will briefly outline the various games included in *Juegos del español*.



Figure 2: The GWAPs in *Juegos del español*: *Agentes*, *Tesoros* and *Anotatlón*.

Three different game concepts have been designed to contrastively examine the effect of various GDEs on PE. As shown in Figure 2 on the left, *Agentes* centers its narrative around the topic of secret agents. This GWAP is a clicker game in which various PoS tags are presented around a sentence with a highlighted word (in this case the determiner *la*) and in which players need to confirm or correct the PoS tag by either clicking on the appropriate one or dragging the word to it.

The second game is called *Tesoros*, in which players need to gather coins and win treasure chests by building a path for an avatar named Gummy, who needs to walk or jump along the constructed path (Figure 2 on the upper right). The path is built every time the player identifies the PoS of a highlighted word.

The third GWAP is a racing game called *Anotatlón*, in which players drive a car to avoid obstacles and reach the finish line. At the finish line, the player must select the appropriate PoS tag for a highlighted word, as illustrated in Figure 2 on the second line.

The three games of *Juegos del español* contain two distinct session types, to wit a training and a playing mode. In the training mode, the highlighted word and its corresponding PoS tag are displayed in the same color to help players become acquainted with the various tags, as shown for *Agentes* and *Tesoros* in Figure 2. Some adjustments to the PoS tags have been introduced though. To illustrate, the *ADP* tag is used in the UD framework for labelling both prepositions and postpositions (see Table 1 in section 5.2.1). Nonetheless, as Spanish does not have the latter, it was decided to present players with a preposition PoS tag (which internally corresponds to the *ADP* tag), given that this is also the denomination used in the Spanish-speaking educational systems.

A mechanism to disambiguate certain tags was also implemented, for example, when a token with a *SCONJ* tag appears in the game, a *CCONJ* and *PRON* tag are also be added (for more details, see Segundo Díaz 2024: chapter 8). Note that the training mode also offers a definition and examples of each PoS to assist players in recognizing the words and their respective PoS tags more effectively (see the examples on the first line of Figure 2).

Besides familiarizing players with the PoS tags and the game mechanics of each GWAP, the training session also serves to establish the confidence score of each player. This score is subsequently used in assessing the inter-annotator agreement, which is needed for automatically accepting confirmations and corrections of the PoS tagging and the extrapolation of the verified data in the morpho-syntactic annotation phase (sections 5.2.3 and 5.2.5).

Once the training session has been completed and the player passes to the playing mode, color cues and PoS definitions are no longer provided, thus challenging players to identify the grammatical categories without help, as shown by the screenshot of *Anotatlón* on the second line in Figure 2, in which no color cue with corre-

sponding definition appears. Finally, note that the games show sentences from the geographic variety that the players have selected when they register to play (Segundo Díaz 2024: chapter 8). This feature was introduced to investigate whether the geographic origin of the players influences the annotation quality, as demonstrated in Bonilla *et al.* (2023).

5 Linguistics Team’s Tasks

5.1 Phase I: Pre-Processing of the COSER Transcriptions

As concerns the pre-processing of the COSER transcriptions, firstly, they have been classified into different regional zones, based on the administrative-political division of Spain into autonomous communities or, in some cases, on the grouping of some of these autonomous communities, as has been done for the Principality of Asturias and the Community of Castile and León due to their shared linguistic heritage (cf. Menéndez Pidal 1906; Tuten *et al.* 2016; for recent work on varieties from this region, see d’Andrés Díaz *et al.* 2017). This regional classification is of utmost importance when aiming to construct a geographically balanced reference model, the COSER-PoS, and ultimately the treebank, COSER-UD, that is representative for the European Spanish rural varieties. Secondly, per region between 500 to 600 conversational turns have been randomly extracted, their transcriptions have been altered to remove XML tags and features that are typical of the COSER transcription protocol, such as, for instance, the abbreviations and punctuation marks to indicate overlapping speech, interruptions and self-corrections (see section 2). Concatenations, represented by the apostrophe, have been detached to ensure the success of the subsequent tokenization of each lexical item, a task belonging to the morpho-syntactic annotation phase (see section 5.2). To illustrate, *s’ha* and *t’ha* in example (2) in section 2 have been divided each into three parts, to wit *s* and *t* respectively, the apostrophe, followed by *ha* (Bonilla *et al.* 2022: 79; for further details, see Bonilla 2024a: chapter 5, 2024b).

5.2 Phase II: Morpho-Syntactic Annotation

Now that the pre-processing of the COSER interviews, the first phase of the project, has been introduced, we will discuss the morpho-syntactic annotation process. Before discussing the details of the PoS tagging, it should be pointed out that this project is not the first to morphologically annotate the COSER corpus. Indeed, as

described in de Benito Moreno *et al.* (2016: 81–82), FreeLing (Carreras *et al.* 2004), an open-source tool, has been used for this in an earlier project. In total, around 180h of transcribed material have been lemmatized and annotated using this tool. Unfortunately, no information exists on the accuracy of the tagging of the COSER corpus with FreeLing, which prevents us from comparing results.

5.2.1 The UD Project

In order to select the most accurate PoS tagger, the morpho-syntactic annotation phase of the project started with a study evaluating the tagging accuracy of three different state-of-the-art open-source taggers, which are based on neural network architectures: to wit, spaCy (Honnibal *et al.* 2020), Stanza NLP (Qi *et al.* 2020), and UDPipe (Straka *et al.* 2016). These taggers have been trained on UD treebanks (Nivre *et al.* 2020), created by the open UD community, which aims to create a cross-linguistically consistent treebank annotation system (see the UD website: <https://universaldependencies.org/>, 15-10-2023).

This immensely successful project currently comprises around 200 treebanks in over 100 languages. Various levels of representation exist within the morpho-syntactic UD annotation scheme: apart from a lemma representing the base form of the word, tokens can get assigned a coarse-grained PoS tag that indicates the word's grammatical category (as in Table 1) and (ii) a more fine-grained label, which describes lexical and grammatical properties associated with the form in question (Nivre *et al.* 2020: 4035; de Marneffe *et al.* 2021).² To exemplify, the PoS tag assigned to the Spanish word *mujer* 'woman' is *NOUN*, whereas the fine-grained features (FEATS) common noun, gender, and number specify a more specific subcategory of the PoS *NOUN* and that the word in question is feminine and singular (for a more elaborate example in treebank format, see Segundo Díaz *et al.* 2023a: 139; Bonilla 2024a chapter 5).

Due to the UD project's general aim to achieve cross-linguistic consistency, the annotation system proposed by the UD project is a universal one. As such, there is a list of universal PoS tags which includes 17 different grammatical categories question (Nivre *et al.* 2020: 4036; de Marneffe *et al.* 2021: 261). Note that, although languages are not required to use all categories, this list cannot be ex-

² The UD project follows traditional grammar by considering words as the primary units, which are interconnected by dependency relations (de Marneffe *et al.* 2021: 259). Observe, however, that this morpho-syntactic notion does not always coincide with the phonological nor orthographic one, as is the case for clitics, which cannot appear without a phonological host, such as those in *t'ha* and *s'ha* in example (2) in section 2.

tended with language-specific PoS tags (cf. section 4 for a discussion on *ADP*). Indeed, as can be seen in Table 1, the morpho-syntactic annotation of Spanish data requires 16 tags out of the full set of 17.³

Table 1: Set of Spanish PoS Tags.

PoS Tag	Grammatical Category
<i>ADJ</i>	Adjective
<i>ADP</i>	Adposition
<i>ADV</i>	Adverb
<i>AUX</i>	Auxiliary
<i>CCONJ</i>	Coordinate conjunction
<i>DET</i>	Determinant
<i>INTJ</i>	Interjection
<i>NOUN</i>	Noun
<i>NUM</i>	Number
<i>PRON</i>	Pronoun
<i>PROPN</i>	Proper noun
<i>PUNCT</i>	Punctuation sign
<i>SCONJ</i>	Subordinate conjunction
<i>SYM</i>	Symbol
<i>VERB</i>	Verb
<i>X</i>	Other

While most of the PoS tags in Table 1 are self-explanatory, some are less so. For instance, *X* is used for incomplete words, as well as unanalyzed lexical items from other languages, which, unsurprisingly, can be found more frequently in the bilingual regions of Spain (e.g., Basque Country, Galicia, Catalonia, Balearic Islands, etc.). The PoS tag *AUX* is used in UD more widely than is usual (among linguists and the general public), given that this tag not only encompasses auxiliary verbs, such as *haber* ‘to have’ in perfect tenses and *ser* and *estar* ‘to be’ in passive and progressive constructions, but also modal verbs, such as *poder* ‘to can’ and *deber* ‘to must’, *soler* ‘to tend to’, as well as *ser* and *estar* ‘to be’ that are used as copulas (see also Bonilla *et al.* 2022: 89–90; Bonilla 2024a: chapter 5). All instances of *ser* (and *soler*) are always classified as auxiliary verbs, irrespective of

³ The COSER-UD only uses 16 of these tags, unlike the UD Spanish AnCorá treebank, which also uses the particle tag (*PART*), as for instance for tagging *no* ‘no(t)’ in *no obstante* ‘notwithstanding’. In COSER-UD, *no* is classified as an *ADV* when modifying a verb and as an *INTJ* when appearing alone. *No obstante* is in COSER-UD classified as a Multi-Word Expression (MWE), which is tagged as a *CCONJ* (https://universaldependencies.org/treebanks/es_ancora/es_ancora-pos-PART.html, 15-03-2024; Bonilla 2022).

their actual status as a verb that supports another one, as for instance in *soy profesora* ‘I am a teacher’. As de Marneffe *et al.* (2021: 273) indicate, the distinction between copula and auxiliary verbs is restored at the parsing level, given that auxiliary verbs are treated as dependents of the main verb through the *aux* relation, whereas copula as dependents of a non-verbal predicate, such as an adjective or a noun, through the *cop* relation. The variants of existential *haber*, in contrast to the use of *haber* in perfect tenses, are classified as *VERB*.

5.2.2 Creation of the COSER-PoS: Automatic Tagging, Data Review and Creation of a Reference Model

As concerns Spanish treebanks, the most widely used one is AnCora_Es (ANnotated CORpora; Taulé *et al.* 2008), which has been developed using Spanish newspapers articles (from the Spanish EFE news agency and the newspaper *El Periódico*) and material from the *Léxico Informatizado del Español* corpus (LexEsp, “Computerized Lexicon of Spanish”, Sebastián Gallés *et al.* 2000).

The LexEsp corpus is a balanced corpus of 6 million words which includes various literary genres, news articles, scientific texts, etc., all written in European Spanish. For the coarse-grained tagging, an accuracy rate above 98% has been reported for the taggers trained with the AnCora_Es treebank (e.g., <https://spacy.io/models/es> and <https://stanfordnlp.github.io/stanza/performance.html>). Nevertheless, their performance with data from other types of language varieties, such as spoken speech, has not been evaluated yet. It is expected though that the accuracy rate with spoken speech will be lower, considering that these taggers have been trained using written language, predominantly from the journalistic domain.

One of the challenges is indeed that there is insufficient labelled and publicly accessible data from other domains. More representative language models are thus needed. In view of this, we aim in this project to fill this gap by evaluating the taggers through the construction of a reference corpus, COSER-PoS, which is a 200.000-word sub-corpus of European rural spoken Spanish, based on a geographically balanced sample from the COSER corpus (Bonilla 2022, 2024a: chapter 5, 2024b). The creation of this reference corpus, the GS, will allow us to measure the accuracy of the current taggers, trained on written varieties close(r) to the standard variety, when dealing with spoken data, as well as to calculate the players’ confidence and resulting inter-annotator agreement scores in the training mode of the GWAPs (see section 4).

The evaluation of these taggers can in turn help determine which features are at the basis for the flaws of the current models when tagging non-standard oral speech. For the creation of the COSER-PoS, three taggers based on neural network

architectures and trained with AnCora_Es (Taulé *et al.* 2008), have been selected: namely, spaCy (Honnibal *et al.* 2020), Stanza NLP (Qi *et al.* 2020), and UDPipe (Straka *et al.* 2016). A geographically balanced sample of around 200.000 tokens has been selected for the construction of the COSER-PoS (see section 5.1), given that the (transcribed part of the) COSER corpus contains more than 4 million words, and that the manual revision of the morpho-syntactic annotation would be too labor-intensive. To ensure that this reference model can be reused and replicated, the annotation criteria and labels from the UD project have been followed (section 5.2.1). This reference corpus has then been used for the accuracy evaluation tasks of the tagging and for the subsequent retraining of the taggers (see section 5.2.3).

Once the geographically balanced sample has been selected, adequately delimited and the changes in the transcriptions discussed in section 5.1 have been implemented (task (1) in Figure 1), various automatic procedures have been carried out using the spaCy NLP library, to wit, sentence segmentation, tokenization, lemmatization, and morpho-syntactic tagging, as will be discussed now.

As regards sentence segmentation, during this process the conversational turns extracted during the pre-processing of the sample (see 5.1) have been separated into sentences, which have then been given a unique identifier, composed of the first four letters of the name of the region and an incremental integer (e.g., extr=Extremadura). The use of this type of identifier makes geographical classification (and thus searching) of the data possible. Note that, even though a similar number of conversational turns has been included per region, the total number of sentences can still vary due to differences in the length of the conversational turns (see Bonilla *et al.* 2022: 85 for specific details per region; Bonilla 2024a: chapter 5, table 5.2).

Once the different sentences have been separated, the words (or tokens) contained in these sentences have been extracted through tokenization. Subsequently, all tokens have been lemmatized, a procedure whereby the inflectional complexity of words is reduced to a common base form, i.e., the lemma, and then morpho-syntactically tagged. The results of all this have been adapted to the CoNLL-X format (Buchholz & Marsi 2006; Computational Natural Language Learning), which is an UD adaptation in which CoNLL-U format properties are assigned to a document, its sentences, and tokens, using the spaCy_conll library (version 3.0, Vanroy 2021).

The spaCy library offers three Spanish models that have been created using convolutional neural networks and which vary in size (small *sm*, medium *md*, and large *lg*) and one model that uses Transformer (*trf*) architecture, in this case the Spanish version of BERT (Cañete *et al.* 2020). According to data published on spaCy's website, the models *sm*, *md*, and *lg* (version 3.0) have an PoS tagging accuracy of 0.98, while *trf* achieves 0.99. All these models have used the AnCora_Es treebank for their training. In this project, we used the large model (*es_core_*-

news_lg) given that the *trf* model had not been released yet when the Linguistics team started the morpho-syntactic tagging task.

After having transformed the geographically balanced COSER sample into the CONLL-U format, the lemmas, coarse-grained PoS tags, and more fine-grained FEATS have been validated both manually and semi-automatically (using regular expressions). This corrected dataset then served to establish the COSER-PoS, a reference corpus or GS, which is freely available for consultation on GitHub (Bonilla 2022).⁴

5.2.3 Automatic Taggers' Accuracy Evaluation

Once the data has been reviewed and the COSER-PoS has been created as a GS, the accuracy evaluation of the various taggers took place. For this sub-task, each of the sentences has been tokenized and tagged using the various versions of the spaCy tagger (*sm*, *md*, *lg*, and *trf*), Stanza NLP, and UDPipe, which all use neural network architecture and have been trained with the AnCora_Es corpus. The results of these processes have then been verified with those of the corrected reference corpus to assess the accuracy rate of each tagger. For the accuracy calculations, the *scikit-learn* library has been used to evaluate the models in terms of precision, recall, F1-score, and accuracy (Pedregosa *et al.* 2011). These reference statistics are important to determine whether the domain adaptation of these taggers, trained on written standard language, to oral-dialectal data, as found in the COSER corpus, improves the tagging accuracy. The comparison of the different accuracy rates of the various taggers reveals that the differences are not significant. Nonetheless, the spaCy's *trf* model outperforms the others with an accuracy rate of 0.927, while the other models' rates range between 0.90 (UDPipe) to 0.920 (Stanza NLP), with the *sm*, *md* and *lg* models of spaCy occupying an intermediate position with an 0.913 accuracy rate (Bonilla *et al.* 2022).

Interestingly, minor differences in the tagger performance can be observed depending on the geographic origin of the text that is tagged, as observed by Bonilla *et al.* (2022: 87). For instance, UDPipe obtained the least accuracy rate for Andalusian and Murcian Spanish (0.89) and the best for Balearic Spanish (0.92). Though the spaCy *trf* model achieved consistently higher results than the other models, for some varieties, such as Castilian, Basque Country and Aragonese

⁴ Note that the size of the COSER-PoS has evolved over time due to various data cleaning phases (cf. Bonilla *et al.* 2022: 13.402 sentences, 204.899 tokens vs. Bonilla 2024a: chapter 5, 13.219 sentences, 196.372 tokens).

Spanish, Stanza NLP performed equally well. Indeed, the difference between these two models is never greater than 1% for the various Spanish regions, whereas the maximum difference between UDPipe and spaCy *trf* can reach 3%, as is the case for Andalusian Spanish. Given the small differences between the geographical areas, it is not possible to draw conclusions on which dialectal zones are closer or more removed from the journalistic language on which the various models were trained.

As the spaCy *trf* model consistently achieved the highest accuracy rates, the next part of the analysis only used this Transformer model, whereby the performance of this tagger is reviewed for each PoS (see also Table 2 in section 5.2.4). Summarizing the most important observations made by Bonilla *et al.* (2022: 88–89), the lowest F1-scores are found in the tagging of incomplete words (*X*: 0) and interjections (*INTJ*: 0.53). These findings are in line with those of Moreno-Sandoval and Guirao (2006: 201–206) for the C-ORAL-ROM corpus. In this project, however, there is no need to increase the size of the lexicon nor to implement a grammar to disambiguate certain linguistic aspects, as machine learning techniques were used for the training of the tagger, whereby it learns from the data input without having to rely on predefined lexicon or rules.

The highest F1-scores, in contrast, are obtained for numbers (*NUM*: 0.96), punctuation signs (*PUNCT*), coordinate conjunctions (*CCONJ*) and prepositions (*ADP*), the latter three of which received an F1-score of 0.99. These results are expected given that, on the one hand, incomplete words and interjections are typical of colloquial speech and, as such, are not found in written journalistic genres, and thus unknown features for the tagger, and that, on the other, the categories with the highest F1-scores are all invariable ones, and thus do not present a challenge for the tagger. Note further that the low F1-score for the interjections can be in part attributed to the UD guidelines (Bonilla 2024a: 76, 2024b), given that words used in exclamations obtain the original PoS. Consequently, *Dios* ‘God’ in *Dios mío* ‘my God’ will be classified as *NOUN*, while the whole construct is also tagged as a Multi-Word Expression (MWE).

Conversely, grammatical categories that exhibit morphological variation associated with oral and/or dialectal idiosyncrasies, such as adjectives (*ADJ*: 0.84), verbs (*VERB*: 0.91), and auxiliaries (*AUX*: 0.75), present the tagger with more difficulties and thus obtain lower F1-scores than the invariable ones. For instance, adjectives with diminutives, deverbal adjectives that coincide with past particles (e.g., *los calderos todos ahumaos* ‘lit. the cauldrons all smoked’), and verbs that reflect dialectal pronunciation and are transcribed differently from the standard form (e.g., *trabajábanos* = *trabajábamos* ‘we worked’) are real hurdles for the tagger as they present unknown morphological characteristics.

5.2.4 Human Annotators' Accuracy Evaluation

Apart from verifying the accuracy of the automatic tagging process, the accuracy rate of the human annotators, i.e., the players of the GWAPs included in *Juegos del español*, was also examined. As already mentioned in section 4, the players' performance in the training mode is used to assign a confidence score for each participant, whereby the accuracy of PoS tagging is compared with the GS. This score serves to calculate the inter-annotator agreement to automatically accept PoS verifications. For a specific token to be assigned an inter-annotator agreement score, at least three players need to have verified the token in question and the score for the tag should have a coefficient of at least 0.75 (Bonilla *et al.* 2023). Once a token receives an inter-annotation agreement score, annotation stops. In what follows, we will report on a study that examines the accuracy of human annotators.

The overall human accuracy rate is 0.80 in the study reported on in Bonilla *et al.* (2023), in which 121 participants took part, who managed to verify 5.976 tokens. This is considerably lower than the accuracy rates of the automatic taggers (spaCy's *trf* model: 0.927; Stanza NLP: 0.920; spaCy *sm*, *md* and *lg* models: 0.913; UDPipe: 0.90; see section 5.2.3; Bonilla *et al.* 2022). However, this difference in general accuracy rate should not lead to an abandonment of the citizen science approach implemented in this research infrastructure project. On the contrary, the comparative analysis of the F1-scores per tagged PoS, shown in Table 2,⁵ in which the bold items indicate the highest F1-score per PoS, reveals that for certain grammatical categories the human tagging accuracy score is higher than for the automatic tagger spaCy's *trf*, thus indisputably demonstrating that human input is not obsolete and, more generally, that citizen science projects can positively contribute to advancing knowledge and technological progress.⁶

Indeed, humans outperform the Transformer tagger for the interjections (*INTJ*: 0.86 vs 0.53) and proper nouns (*PROPN*: 0.93 vs 0.69). These results are expected given that interjections tend to be absent in certain genres, such as journalist texts, on which the automatic taggers were trained (see section 5.2.2). Similarly, humans tend to be better at inferencing that a given token in a certain linguistic context is a proper noun, such as a name or place name, without necessity of previously having

5 For additional details, such as the precision and recall scores for the spaCy *trf* model and for the human annotators, we refer the interested reader to Bonilla *et al.* (2022) and Bonilla *et al.* (2023), respectively. For a study on the variables that significantly influence the human annotators' tagging accuracy, such as educational level, the field of study, and geographic upbringing, see Bonilla *et al.* (2023).

6 This said, we do not claim that taggers cannot be trained to achieve as high accuracy rates as human annotators or even higher. Though, human input will be needed for this task too.

Table 2: Comparison of Human and Automatic Tagging Accuracy (F1-scores per PoS).

PoS	F1-scores	
	Human annotators	SpaCy <i>trf</i> model
<i>ADJ</i>	0.76	0.84
<i>ADP</i>	0.83	0.99
<i>ADV</i>	0.86	0.91
<i>AUX</i>	0.61	0.75
<i>CCONJ</i>	0.83	0.99
<i>DET</i>	0.84	0.94
<i>INTJ</i>	0.86	0.53
<i>NOUN</i>	0.94	0.94
<i>NUM</i>	0.89	0.96
<i>PRON</i>	0.69	0.91
<i>PROPN</i>	0.93	0.69
<i>PUNCT</i>	/	0.99
<i>SCONJ</i>	0.46	0.87
<i>SYM</i>	/	0.91
<i>VERB</i>	0.69	0.91
<i>X</i>	0.79	0

heard or being familiar with the proper noun in question. The F1-score of the automatic tagger for incomplete words (*X*) is 0 given that this PoS is absent in written language models, whereas human annotators are completely used to it due to its high frequency in oral speech. As such, humans performed significantly better (0.79).

As can be seen, the differences between the human and automatic tagging accuracy scores are for some PoS very great. For the PoS *NOUN*, in contrast, the F1-scores for the tagging accuracy are the same for the *Juegos del español* participants and the spaCy *trf* model: to wit, 0.94, which is the highest F1-score for human tagging when comparing all the PoS scores. The lowest human F1-score is obtained by the *SCONJ* (0.46), which was in almost half of the cases confused with its coordinate counterpart, *CCONJ*, probably due insufficient technical linguistic knowledge by the participants (Bonilla *et al.* 2023). Equally, the *AUX* category received a low human F1-score (0.61), which is quite likely due to the classification used in the UD guidelines (see section 5.2.1), which diverges from the one taught in schools and appears to lead to confusion among the players. Concretely, *ser* is always regarded as an *AUX* (e.g., in its copular and passive function) regardless of its status as a supporting verb to another one, whereas *estar* ‘to be’ receives the *AUX* tag when being a copula (e.g., *Lili está enferma* ‘Lili is ill’) and part of the progressive construction (e.g., *Johnatan está estudiando* ‘Johnatan is studying’), but can also function as a

VERB, as in *Miriam está en el despacho* ‘Miriam is in the office’ (Bonilla 2022). However, the same cannot be said verbal periphrases, whereby *voy* in *voy a cantar* ‘I am going to sing’ [*ir* ‘to go’+ preposition *a* + infinitive] and in *voy cantando* ‘I’m going (while) singing’ [*ir* ‘to go’ + gerund], or *ando* in *ando cantando* ‘I walk around singing’ [*andar* ‘to walk’ + gerund] are regarded as a *VERB*, despite the clear parallelism between these cases with the *estar*-constructions with a gerund.

5.2.5 Post-Tagging Knowledge Transfer

As we will see in this section, it is not necessary to verify the PoS tag for each token in the corpus given that the knowledge gained from the players of *Juegos del español*, who confirmed or corrected PoS tags, can be extrapolated to unverified PoS tags on condition that certain requirements are met. To exemplify, the preposition *de* ‘of/from’ can only ever be a preposition (*ADP* in UD). In view of this, the human annotators’ confirmation of the accuracy of this PoS tag can be extrapolated to all 4.699 cases in COSER-UD, thus tremendously upscaling the crowd-sourced input. In contrast, the PoS tag verifications of *bueno*, which can function either as an *ADJ* ‘good’ or as an *INTJ* ‘well’, but which both have the same lemma, namely *bueno*, cannot be extrapolated because of its lemma’s inherent polyfunctionality. Notwithstanding this, homographs that have been assigned different lemmas can be extrapolated, as is the case for instance for *la* which can be the feminine article ‘the’ (*DET*) but also a feminine object pronoun ‘her’ (*PRON*), the latter of which has been attributed the lemma *él* ‘he’, while the former *el* ‘the’.

The knowledge transfer of verified PoS tags has been implemented in a semi-automated manner. Initially, all text underwent conversion to lowercase to ensure uniformity across the dataset, crucial for consistent text analysis, while eliminating duplicate entries. Next, PoS corrections suggested by human annotators have been reviewed by experts, who adjusted the PoS tags when needed. For instance, all cases of *ser* ‘to be’ and *ir* ‘to go’ are considered *AUX* and *VERB* respectively in UD (see also section 5.2.4). Additionally, erroneous corrections, such *umbilical* ‘umbilical’ or *última* ‘last’, which were tagged as *NOUN* instead of *ADJ*, have been rectified. Furthermore, the lemmas of the tokens have been automatically extracted and then manually confirmed. This step is needed as the original database included only IDs, which complicates the large-scale knowledge transfer of verified tags. The final stage of the post-tagging knowledge transfer consists in an extrapolation process of the verified PoS tags. First, the set [token + lemma + PoS tag] of the verified data is matched with their counterparts in the automatically tagged data of the COSER-UD treebank. As they coincide, the latter cases can thus be regarded as veri-

fied by extension. Table 3 provides the number of matches of this extrapolation operation per PoS tag, counting MWEs as single entities.

Table 3: Post-Tagging Knowledge Transfer per PoS tag.

PoS tag	Extrapolated cases
<i>ADJ</i>	180 (0.25%)
<i>ADP</i>	15.499 (21.44%)
<i>ADV</i>	8.812 (12.19%)
<i>AUX</i>	1.465 (2.03%)
<i>CCONJ</i>	10.329 (14.29%)
<i>DET</i>	14.413 (19.93%)
<i>INTJ</i>	1.052 (1.45%)
<i>NOUN</i>	4.657 (6.44%)
<i>NUM</i>	633 (0.88%)
<i>PRON</i>	8.806 (12.18%)
<i>PROPN</i>	178 (0.25%)
<i>SCONJ</i>	4.120 (5.7%)
<i>SYM</i>	0
<i>VERB</i>	2.139 (2.96%)
<i>X</i>	21 (0.03%)
Total	72.304

Leaving punctuation signs aside, the knowledge transfer of the 467 verified tokens by players of *Juegos del español* yielded the confirmation of a total of 72.304 tokens. In other words, while human annotators confirmed or corrected a mere 0.29% of COSER-UD’s tokens, the knowledge transfer upscaled these results to 45.1% (72.304/160.321) of the treebank, which is not an insignificant feat.⁷ However, lemma-wise the results of this extrapolation process are a lot less impressive as it only affects 3.34% (189/5.662) of all lemmas.

As can be seen in Table 3, the extrapolation of prepositions (*ADP*: 21.44%) is responsible for more than a fifth of all cases, closely followed by the determiners (*DET*: 19.93%), and then the coordinate conjunctions (*CCONJ*: 14.29%), adverbs (*ADV*: 12.19%) and pronouns (*PRON*: 12.18%). These results indicate that, though maintaining players’ interest in GWAPs is not an easy task (Chamberlain *et al.*

⁷ In summary, 147 players confirmed and/or corrected 8.215 PoS annotations, which resulted in the full verification of 467 tokens (using the inter-annotator agreement score; see section 5.2.4), while 3.428 tokens have been verified partially (up until the 23rd of February 2024). It should not be forgotten that these PoS annotations only include those obtained in the playing mode (see section 4). In the training session, players provided 10.576 annotations.

2013; Poesio *et al.* 2017), upscaling verification results by knowledge transfer can advance the PoS verification process considerably, leaving more time for the human annotators to handle the polyfunctional, more difficult cases.

6 Conclusions

In summary, this contribution outlines an interdisciplinary research infrastructure project integrating the fields of Dialectology, NLP, and HCI aimed at developing a morpho-syntactically annotated and parsed corpus of the diatopic varieties of European Spanish, known as COSER-UD. Employing a citizen science approach, this project engages the public through various GWAPs, collectively termed as *Juegos del español*, to verify the automatic PoS tags. In addition to delineating the three games comprising *Juegos del español*, this discussion detailed the tasks undertaken by the Linguistics team during the initial two phases of the project, which encompass the transcription pre-processing and morpho-syntactic annotation of the COSER. Regarding the latter, we elaborated on the creation of COSER-PoS as a reference model and the automated tagging process. Subsequently, we provided a comparative analysis of tagging accuracy rates between the spaCy Transformer model and human annotators. While human annotators generally exhibit lower tagging accuracy scores, they notably surpass the Transformer model for specific PoS categories, such as interjections (*INTJ*), proper names (*PROPN*), and incomplete words (*X*). This underscores the effectiveness of the collaborative approach employed in this corpus creation endeavor, wherein human expertise and NLP algorithms complement each other. The final phase of morpho-syntactic annotation involves the post-tagging transfer of verified data, resulting in a significant expansion of the verification rate (from 0.29% to 45.1% of the treebank). Prepositions (*ADP*) and determiners (*DET*) emerge as the primary PoS categories driving this extrapolative procedure, collectively constituting more than 40% of the instances therein.

Regarding future tasks pertaining to the morpho-syntactic annotation of the COSER-UD, approximately half of the PoS tags of the treebank are yet to be verified. Given the difficulty of sustaining player engagement in gaming, alternative data verification methods are currently under exploration, as well as variations on the post-tagging knowledge transfer.

Bibliography

- Bonilla, Johnatan E. (2022): COSER-UD <https://github.com/johnatanebonilla/UD_Spanish-COSER> (21-10-2023).
- Bonilla, Johnatan E. (2024a): *Universal Dependencies for Spoken Spanish*. Doctoral Dissertation. Ghent University/Humboldt University of Berlin.
- Bonilla, Johnatan E. (2024b): “Spoken Spanish PoS Tagging: Gold Standard Dataset”, in Language Resources and Evaluation. <<https://doi.org/10.1007/s10579-024-09751-x>>.
- Bonilla, Johnatan E. (submitted): “Development of the First Spoken Spanish Treebank within the Universal Dependencies Framework”.
- Bonilla, Johnatan E., Miriam Bouzouita and Rosa Lilia Segundo Díaz (2022): “La construcción del *Corpus Oral y Sonoro del Español Rural – Anotado y Parseado* (COSER-AP): avances en el etiquetado de partes del discurso”, in *Revista Internacional de Lingüística Iberoamericana*, 22(40), pp. 77–96.
- Bonilla, Johnatan E., Rosa Lilia Segundo Díaz and Miriam Bouzouita (2023): “Using GWAPs for Verifying PoS Tagging of Spoken Dialectal Spanish”, in *10th International Conference on Behavioural and Social Computing (BESC)*. Institute of Electrical and Electronics Engineers, pp. 1–7. <<https://doi.org/10.1109/BESC59560.2023.10386542>>.
- Breitbarth, Anne, Melissa Farasyn, Anne-Sophie Ghyssele and Jacques Van Keymeulen (2020): “Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (GCND)”, in *Handelingen (KZM)*, LXXII, pp. 23–38. <<https://doi.org/10.21825/kzm.v72i0.17914>>.
- Buchholz, Sabine and Erwin Marsi (2006): “CoNLL-X Shared Task on Multilingual Dependency Parsing”, in Lluís Màrquez and Dan Klein (eds). *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City: Association for Computational Linguistics, pp. 149–164. <<https://aclanthology.org/W06-2920.pdf>> (21-10-2023).
- Campillos Llanos, Leonardo (2016): “PoS-Tagging a Spanish Oral Learner Corpus”, in Margarita Alonso-Ramos (ed.). *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, pp. 78–89.
- Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang and Jorge Pérez (2020): “Spanish Pre-trained BERT Model and Evaluation Data”, in *Practical Machine Learning for Developing Countries at the International Conference on Learning Representations 2020*, pp. 1–10.
- Carreras, Xavier, Isaac Chao, Lluís Padró and Muntsa Padró (2004): “FreeLing: An Open-Source Suite of Language Analyzers”, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association (ELRA), pp. 239–242.
- Castillo, Lorena, M.^a Pilar Colomina and Irene Fernández (2020): “El *Atlas Sintáctico del Español* (ASinEs): una herramienta para codificar la variación”, in Ángel J. Gallego and Francesc Roca Urgell (eds.). *Dialectología digital del español*. Santiago de Compostela: Verba: Anuario Galego de Filoloxía, Anexo 80, pp. 47–69. <<https://dx.doi.org/10.15304/9788418445316>>.
- Castillo Lluch, Mónica, Cristina Peña Ruedo and Michiel de Vaan (2022): “¿‘Pronunciar’ o ‘pronunciare’? Esa es la cuestión”, in Ana Estrada, Beatriz Martín and Carlota de Benito (eds.). *Como dicen en mi pueblo: el habla de los pueblos españoles*. Madrid: Pie de Página, pp. 63–75.
- Chamberlain, Jon, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade and Massimo Poesio (2013): “Using Games to Create Language Resources: Successes and Limitations of the Approach”, in Iryna

- Gurevych and Jungi Kim (eds.). *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Berlin, Heidelberg: Springer, pp. 3–44. <https://doi.org/10.1007/978-3-642-35085-6_1>.
- Chambers, Jack and Perter Trudgill (1998): *Dialectology*. Cambridge: Cambridge University Press.
- COSER = Fernández-Ordóñez, Inés (dir.) (2005–present): “Corpus Oral y Sonoro del Español Rural”. <www.corpusrural.es> (11-08-2024).
- d'Andrés Díaz, Ramón, Fernando Álvarez-Balbuena García, Xulio Miguel Suárez Fernández, and Miguel Rodríguez Monteavaro (2017): *Estudiu de la transición llingüística na zona Eo-Navia, Asturias (ETLEN). Atlas llingüísticu dialectográficu - horiométricu - dialectométricu*. Trabe: Universidá d'Uviéu.
- de Benito Moreno, Carlota, Javier Pueyo and Inés Fernández-Ordóñez (2016): “Creating and Designing a Corpus of Rural Spanish”, in Stefanie Dipper, Friedrich Neubarth and Heike Zinsmeister (eds.). *Proceedings of the 13th Conference on Natural Language Processing KONVENS 2016*. Bochum: Bochumer Linguistische Arbeitsberichte, pp. 78–83.
- de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman (2021): “Universal Dependencies”, in *Computational Linguistics*, 47(2), pp. 255–308.
- Esplà-Gomis, Miquel and Andreu Sentí (in preparation): “An Annotated Informal and Spoken Corpus for Dialectal Conversations: The Parlars Corpus for Valencian Catalan”.
- Farasyn, Melissa, Anne-Sophie Ghyselen, Jacques Van Keymeulen and Anne Breitbarth (2022): “Challenges in Tagging and Parsing Spoken Dialects of Dutch”, in *Journal of Historical Syntax*, 6, pp. 4–11. <<https://doi.org/10.18148/hs/2022.v6i4-11.92>>.
- Fernández-Ordóñez, Inés and Enrique Pato (2020): “El Corpus Oral y Sonoro del Español Rural (COSER) y su contribución al estudio de la variación gramatical del español”, in Ángel J. Gallego and Francesc Roca Urgell (eds.). *Dialectología digital del español*. Santiago de Compostela: Verba: Anuario Galego de Filoloxía, Anexo 80, pp. 71–100. <<https://dx.doi.org/10.15304/9788418445316>>.
- Gelbukh, Alexander, Sulema Torres and Hiram Calvo (2005): “Transforming a Constituency Treebank into a Dependency Treebank”, in *Procesamiento del Lenguaje Natural*, 35, pp. 145–152.
- Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen and Arian van Hessen (2020): “Clearing the Transcription Hurdle in Dialect Corpus Building: The Corpus of Southern Dutch Dialects as Case Study”, in *Frontiers in Artificial Intelligence*, 3. <<https://doi.org/10.3389/frai.2020.00010>>.
- Honnibal, Matthew, Ines Montani, Sofie van Landeghem and Adriane Boyd (2020): “spaCy: Industrial-strength Natural Language Processing in Python”. <<https://github.com/explosion/spaCy>> (11-08-2024).
- Juegos del español* = Bouzouita, Miriam, Johnatan E. Bonilla, Rosa Lilia Segundo Díaz, Véronique Hoste, Karin Coninx and Gustavo Roveló Ruiz (2022): “Project *Juegos del español*”. <www.juegosdelespanol.com> (11-08-2024).
- Jørgensen, Annette Myre and Esperanza Eguía Padilla (2015): “Presentación de COLA, un corpus oral de lenguaje adolescente en línea”, in Sigrún A. Eriksdottir (ed.). *Actes du XIX^{ème} Congrès des romanistes scandinaves*. Reykjavik: Institute of Foreign Languages. <<https://conference.hi.is/rom14/rom-lectures/>> (11-08-2024).
- Jørgensen, Annette Myre, Esperanza Eguía Padilla, Anna-Brita Stenstrom, Juan Antonio Martínez López, Eli Marie Drange Danbolt, Mariano Reyes Tejedor, Anna Acevedo, Giovanna Angela Mura, Stine Huseby, Lise Holmvik, Solfrid Hernes, Evert Jakobsen, Kristine Eide and Marie Espeland (2004–2017): “Proyecto COLA. Corpus Oral de Lenguaje Adolescente”. <<https://blogg.hiof.no/colam-esp/>> (21-09-2021).

- Martínez Alonso, Héctor and Daniel Zeman (2016): “Universal Dependencies for the AnCorra Treebanks”, in *Procesamiento del Lenguaje Natural*, 57, pp. 91–98.
- Menéndez Pidal, Ramón (1906): “El dialecto asturleonés”, in *Revista de Archivos, Bibliotecas y Museos*, 2–3, pp. 128–172 and pp. 294–311.
- Montserrat, Sandra and Carles Segura (2020): “Un corpus col-loquial i dialectal del valencià: PARLARS”, in *Zeitschrift für Katalanistik*, 33, pp. 9–44. <<https://doi.org/10.46586/Zfk.2020.9-44>>.
- Moreno Fernández, Francisco (2005): “Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA”, in *Oralia: Análisis del discurso oral*, 8, pp. 123–140.
- Moreno-Sandoval, Antonio, Guillermo de la Madrid, Manuel Alcántara, Ana González, José M. Guirao and Raúl de la Torre (2005): “The Spanish Corpus”, in Emanuela Cresti and Massimo Moneglia (eds.). *C-ORAL-ROM: Integrated Reference Corpus for Spoken Romance Languages*. Amsterdam: John Benjamins, pp. 135–161.
- Moreno-Sandoval, Antonio and José M. Guirao (2006): “Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation”, in Yuji Kawaguchi, Susumu Zaima and Toshihiro Takagaki (eds.). *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: John Benjamins, pp. 199–218.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman (2016): “Universal Dependencies v1: A Multilingual Treebank Collection”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 1659–1666.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers and Daniel Zeman (2020): “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association (ELRA)*. Marseille: European Language Resources Association, pp. 4034–4043.
- Predgosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay (2011): “Scikit-learn: Machine Learning in Python”, in *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Poesio, Massimo, Jon Chamberlain and Udo Kruschwitz (2017): “Crowdsourcing”, in Nancy Ide and James Pustejovsky (eds.). *Handbook of Linguistic Annotation*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-024-0881-2_10>.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher Manning (2020): “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. <<https://arxiv.org/abs/2003.07082>> (11-08-2024).
- Radford, Alec, Jing Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey and Ilya Sutskever (2023): “Robust Speech Recognition via Large-Scale Weak Supervision”, in *Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 202, pp. 28492–28518. <<https://proceedings.mlr.press/v202/radford23a/radford23a.pdf>> (11-08-2024).
- Rico-Sulayes, Antonio, Rafael Saldívar-Arreola and Álvaro Rábago-Tánori (2017): “Part-of-Speech Tagging with Maximum Entropy and Distributional Similarity Features in a Subregional Corpus of Spanish”, in *Ingeniería y Competitividad*, 19(2), pp. 55–67.

- Sebastián Gallés, Núria, M. Antònia Martí Antonín, Manuel Francisco Carreiras Valiña and Fernando Cuetos Vega (2000): *LEXESP: Léxico informatizado del español*. Barcelona: Edicions Universitat de Barcelona.
- Segundo Díaz, Rosa Lilia (2024): *Juegos del español – Iterative Design, Evaluation and Implementation of Games with a Purpose to Enhance Parts-of-Speech Tagging in a Corpus of European Spanish Dialects*. Doctoral Dissertation. Hasselt University/Ghent University.
- Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita and Karin Coninx (2022): “Building Blocks for Creating Enjoyable Games – A Systematic Literature Review”, in *International Journal of Human – Computer Studies*, 159. <<https://doi.org/10.1016/j.ijhcs.2021.102758>>.
- Segundo Díaz, Rosa Lilia, Johnatan E. Bonilla, Miriam Bouzouita and Gustavo Rovelo Ruiz (2023a): “Juegos con propósito para la anotación del *Corpus Oral Sonoro del Español Rural*”, in *Dialectologia et Geolinguistica*, 31, pp. 135–164. <<https://doi.org/10.1515/dialect-2023-0007>>.
- Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita, Véronique Hoste and Karin Coninx (2023b): “The Influence of Personality Traits and Game Design Elements on Player Enjoyment: A Demo on GWAPs for Part-of-Speech Tagging”, in Mads Haahr, Alberto Rojas-Salazar and Stefan Göbel (eds.). *Serious Games, 9th Joint International Conference, JCSG 2023, Lecture Notes in Computer Science (LNCS, volume 14309)*. Cham: Springer, pp. 353–361. <https://doi.org/10.1007/978-3-031-44751-8_28>.
- Segundo Díaz, Rosa Lilia, Gustavo Rovelo, Miriam Bouzouita, Véronique Hoste and Karin Coninx (2024): “The Influence of Personality Traits and Game Design Elements on Player Enjoyment: An Empirical Study in GWAP for Linguistics”, in Pierpaolo Dondio *et al.* (eds.). *Games and Learning Alliance, 12th International Conference, GALA 2023, November 29-December 1, 2023, Lecture Notes in Computer Science (LNCS, volume 14475)*. Cham: Springer, pp. 1–10. <https://doi.org/10.1007/978-3-031-49065-1_20>.
- Straka, Milan, Jan Hajič and Jana Straková (2016): “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing”, in Nicoletta Calzolari *et al.* (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož: European Language Resources Association (ELRA), pp. 4290–4297.
- 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 the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech: European Language Resources Association (ELRA), pp. 96–101.
- Taulé, Mariona, M Antònia Martí, Ann Bies, Montserrat Nofre, Aina Garí, Zhiyi Song, Stephani Strassel and Joe Ellis (2015): “Spanish Treebank Annotation of Informal Non-standard Web Text”, in Florian Daniel and Oscar Diaz (eds.). *Current Trends in Web Engineering, ICWE 2015, Lecture Notes in Computer Science*. Cham: Springer, pp. 15–27. <https://doi.org/10.1007/978-3-319-24800-4_2>.
- Tuten, Donald N., Enrique Pato and Ora R. Schwarzwald (2016): “Spanish, Astur-Leonese, Navarro-Aragonese, Judaeo-Spanish”, in Adam Ledgeway and Martin Maiden (eds.). *The Oxford Guide to the Romance Languages*. Oxford: Oxford University Press, pp. 382–410.
- Vanroy, Bram (2021): spaCy_conll 3.0. <https://github.com/BramVanroy/spaCy_conll> (11-08-2024).

María Teresa García-Godoy

Big Data and Lexical History: Digital Newspaper Libraries in Spanish Diachronic Research

1 Introduction

The advent of the Royal Spanish Academy's (hereafter, RAE) first diachronic database in the late 20th c. has allowed free access to ten centuries of the history of Spanish texts in the form of a large annotated corpus ever since. A 250-million-word corpus covering from the earliest records to 1974, CORDE trailed the blaze for the big data resources that later became available for research on the history of Spanish and, thus, for the new paradigm of experimentally-based research that was to come. This is so much so that, all the RAE's databases, both of the earliest (CORDE, CREA) and of the latest generation (CDH and CORPES XXI) are considered reference corpora still today, and are essential for research on Spanish, whether present-day (CREA and CORPES XXI) or of the past (CORDE, CDH). Still, the RAE's databases are not enough for research on the history of diaphasic and diatopic variation, for two reasons: i) the genres where older colloquial forms are most likely to occur are underrepresented in these corpora; and ii) the geolocation of specific uses is not specific enough for research of intradialectal variation. The geographical classification of CORDE and CDH samples by country deny the possibility of research on regional or local forms. Thus, while the RAE diachronic corpora were the first big data resources available for diachronic research on Spanish, they are, paradoxically, also fairly limited as regards colloquial and dialectal diversity.

All the RAE corpora are based on a range of text types (fiction, notarial documents, chronicles, historiography, scientific treatises, etc.). Yet, not all the discourse genres are equally represented in the diachronic corpora, and this paper shows how the literary genres are primed over the non-literary ones. The resulting data unavailability can be noticed especially in documents of communicative immediacy, more likely to represent the spoken language in written (García-Godoy 2015). A lack of journalistic texts can also be felt in modern Spanish (18th and 19th centuries): Neither CORDE nor CDH cover a substantial amount of his-

Note: This contribution has been realized in the framework of Grant PID2022-136256NB-I00, funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU. Also, it has been carried out within the framework of Grant C-HUM-038-UGR23 funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021–2027.

torical press.¹ This neglect is particularly serious for the period 1750–1900, when the journalistic genre arose and grew in Spanish.

A lack of balance is also noticeable in the RAE diachronic databases as regards diatopic representativeness. Not all geolects are equally well represented in the CORDE and CDH corpora, e.g. American Spanish is represented by fewer documents than European Spanish and, in the latter, the variety of the *Meseta* prevails over other Spanish geolects. As the geographical classification in the diachronic corpora is by country, it is not possible to track down, e.g. the development of neologisms from their geographical origin. Language creativity and its geographical dissemination are easier to research from 1700 onwards, when the journalistic genre appeared. This is because, unlike other text types, journalistic texts, especially their sections of local news, always carry specific data on their day, month, year of publication, and on location, i.e. on key data for the geolocation of language variants.

Historical press is widely acknowledged as a major data source for research on diversity in modern Spanish, actually as one that is unparalleled by other text types in this regard. While the Spanish journalistic genre dates back to the 17th c., sparked by the spirit of the Enlightenment, journalism as we know it today developed from the first decree of press freedom issued by the Parliament of Cádiz (*Cortes de Cádiz*) in the 19th c. Since then, the Spanish press has proved a melting pot of texts of various types, topics and styles over the country. The early Spanish press discloses the inter(national), regional and local life of a specific time and thus depicts modern Spanish (i.e. of the 18th and 19th centuries) in ways that are not so noticeable in other genres. Current research thus relies on the digital newspaper libraries' search engines for retrieval of data unavailable from RAE corpora that may complete the picture of the language of the time, especially as regards diversity and the origin and development of colloquial forms in modern Spanish. The lack of experimental data of pre-20th c. spoken Spanish constrains diachronic research on diaphasic variation to the spoken language attested in written. As this chapter shows, journalistic texts from the 18th c. onwards have become data sources for colloquial forms and their geographical distribution which are hard to find outside newspaper libraries and, thus, largely outside of the genres recorded by reference diachronic corpora.

The relevance of texts for the account of language change is worth remarking. Their importance for research on the origins and development of specific cases has

¹ The CORDE website lists the text types of the corpus. Fiction samples (poetry, prose, drama) amount to 40%, whereas press and commercial genres are only 8%. The CDH corpus, partly based on CORDE, classifies samples by topic instead of by genre, so the actual number of literary and press samples is hard to measure.

been underlined in the past decade by a number of authors (García-Godoy 2021). Epistemologically, journalism –and its range of topics– appears as a major genre where language change in the 18th and 19th centuries can be tracked. Due to the CORDE and CDH's neglect of newspaper texts, researchers turn to newspaper libraries of historical press for access to a wealth of evidence of pan-Hispanic language diversity. The digital newspaper section of the National Library (hereafter, HD) has become the most frequently used resource in this regard. It stores 2165 newspapers published nationally, regionally and locally in the Spanish speaking world of the period between the 18th and the 21st centuries (see section 3). Admittedly, a newspaper library's search engine cannot be compared with a computerized language corpus concordancer and the potential for data retrieval of its lemmatization, morphosyntactic annotation, or data size management, among others. At present, no newspaper library allows automatic data screening or analysis, so research on the evolution of Spanish based on digitalized press requires manual analysis of the data collected (see section 4.2.).

This chapter is intended to prove the relevance of the journalistic genre in the history of certain colloquial forms of modern Spanish (18th and 19th centuries), a crucial period for language standardization. Evidence of colloquial forms, specifically of the adverb *cabalito*, is compared between newspaper vs. other data sources, based on the HD and the CDH, respectively.

The chapter develops from this *Introduction* by way of four sections. First, a review of the state of the art on digital newspaper libraries for diachronic research on Spanish. Section 3 discusses corpora and methods. The history of *cabalito* then follows in Section 4 according to corpus evidence with regard to its diachronic, textual and diatopic properties, after its morphosyntactic description according to the grammars of reference. Section 5 deals with the formalization of *cabalito* in Spanish lexicographic practice over time, and then leads to the chapter's conclusions.

2 Digital Newspaper Libraries and Modern Spanish. State of the Art

In the past decade, large data bodies of historical press have raised the interest in diachronic research on Spanish, especially on modern Spanish. This is largely due to the development of online newspaper libraries, as they offer more and more evidence of the history of Spanish journalistic texts, so rare in Spanish corpora. Still, this resource is used for research on the evolution of Spanish in barely ten recent publications, mainly on lexical semantics and, less frequently, on the mor-

phosyntax of Spanish. The results obtained from digital newspaper libraries are worthy of special note for the following three processes of language variation and change: i) neology; ii) standardization; and iii) dialectalization.

In neology, the research on lexical and morphosyntactic change under the influence of the early journalistic genre stands out, especially in four sections of historical press: i) scientific dissemination; ii) crime and accidents; iii) advertisements; and iv) society. Campos Souto (2018) has proved that, for the former of these, HD is crucial for the compilation of the historical dictionary of Spanish, especially for the so-called specialty languages. Campos Souto argues that the digital contents of HD, whether generic or specialty, attest the earliest records of a large number of technical terms in the fields of medicine, music, or fashion between the 17th and the 20th centuries: Over 60 entries of these domains listed in the *Diccionario histórico del español* are first attested in the official press of the 18th c. (*gripe* ‘flu’, *guitarra* ‘guitar’, *violin* ‘violin’, *oboe* ‘oboe’, etc.), of technical and scientific journalism of the 19th c. (*gripal* ‘flu-related’, *saramponioso* ‘infected with measles’, *corsetería* ‘corsetry’, *clubrecorsé* ‘camisole’, . . .) and, particularly, of the press in general of the 20th c. (*griposo* ‘infected with flu’, *lepra*, ‘leprosy’, *violonchelista* ‘cello player’). García-Godoy’s (2015, 2017) lexical research also proves that lexical innovations currently in use date back to the political press between 1813 and 1823, e.g. *retaliación* ‘retaliation’, *retaliar* ‘retaliate’, *complotado* ‘conspirator’, *sufragar* ‘vote’, and neological senses of *departamento* ‘department’, *cantón* ‘canton’, *cantonal* ‘cantonal’, or *prefectura* ‘prefecture’ for administrative divisions. HD is also used for attestation of new morphosyntactic formations, e.g. *madama* ‘Mrs.’ / *madamita* ‘Miss’ under the influence of French in 18th c. society news (García-Godoy 2021), and the prepositional locution *a por* ‘towards’ with motion verbs in advertisements and in crime and accidents sections of the 19th c. (Company & Flores Dávila 2017, 2018).

In the references cited above, HD is used not only for the identification of lexical and morphosyntactic neologisms in journalistic texts, but also to assess their diatopic spread prior to standardization. The precise chronological dating of each neological use in journalistic texts by nature makes digital newspaper libraries a key data source for research on the convergence and divergence between European and American Spanish from the 18th c. onwards. Regarding European Spanish, Company & Flores Dávila (2017, 2018) and García-Godoy (2021) claim that *a por* ‘towards’ and *madama* ‘Mrs.’ / *madamita* ‘miss’ start out in the press of Madrid, wherefrom they quickly reached the rest of the country by imitation of the new formations of the Spanish of the court as a standard of prestige.

Regarding American Spanish, García-Godoy (2015, 2017) shows how the earliest attestation of public lexical divergence between the colonies and the metropolis can be found in the pro-independence press of the American emancipation period

across South America (Colombia, Mexico, Argentina, Venezuela). These divergences are today viewed as specifically American, century-old formations. Diatopically, these first instances of research based on HD evidence focus on European vs. American Spanish as two geolects separated by the Atlantic as an isogloss. Thus, geolocation of HD data help identify large-scale continental dialects taking the Atlantic as a boundary line, even if further research at a smaller scale is needed, and where HD may also supply earliest attestation data. In the latter respect, and despite the large amount of regional and local records available from the HD, the so-called local terms are in need of research with regard to the journalistic genre. Octavio de Toledo's (2016) "Sin CORDE pero con red: *algotras* fuentes de datos" pioneered this field by defining boundaries on the *terra incognita* of the historical map of dialects in the Iberian peninsula, based on online sources (e.g. specific subject field websites, data retrieval from *Google Books* by specific search engines), whenever the RAE corpora failed to supply the necessary data. This author does not use digital newspaper libraries for his research on *algotro* (and its morphological variants), but his methods are very similar to those reviewed in this section in that language traits unsupported from data available from reference corpora and metalinguistically marked in diatopically vague and contradictory ways in historical dictionaries are researched based on online resources searchable with specific engines. Against traditional dialectal lexicography, Octavio de Toledo thus proves that the quantifier *algotro* comes historically from Extremadura instead of Andalusia, and that it eventually was also used therefrom in American Spanish.²

3 Corpora and Experimental Methods

This chapter is based on three complementary databases: i) HD for the history of the journalistic genre; ii) CDH as a control corpus: as it covers a range of genres (except for press texts), it is used here to tell whether the journalistic genre made wider use of *cabalito* or not compared with HD; and iii) the *Nuevo Tesoro Lexicográfico de la Lengua Española* (hereafter, NTLLE) for the occurrence of *cabalito* in the history of Spanish dictionaries.

Our base corpus, the HD, focuses on press language and covers digitally available newspapers published between the 18th and the 21st centuries. The chronological frame is defined by HD's earliest and latest attestation of *cabalito* (1790–2022). The 2022 actualization of the virtual HD relies on 2,165 titles and over

² Calderón Campos (2023: 119–121) uses the EsTenTen18 macrocorpus (Sketch Engine) and the social network X to analyze the current stigmatization of *algotro* in America.

seven million pages in digital format, with European press prevailing over American press, as mentioned above. This big data source of historical press offers the largest body of data on the history of *cabalito* currently available, namely 327 attestations between the 18th and the 21st centuries compared with 27 occurrences in the 3.5-million-word CDH corpus (specifically, 355,740,238 words) used for the RAE's historical dictionary.

The 322 years' span is divided into 30-year sections for identification of any neological stage and its subsequent evolution stages. Regarding the contents, the corpus is a scale model of the HD in that it accounts for the range represented in the HD too: Politics, satire, humour, science, religion, cultivated, leisure, sports, arts, literature, etc. This connection with the HD also shows in the bias towards Spanish press to the detriment of American press, and in the availability of regional and local press only for European Spanish.

As for data collection, the HD application allows a range of queries and access to the digital version of the original texts. Data collection is far from perfect, as it lets in undesired cases in the search for *cabalito*, e.g. *caballero* 'gentleman', *caballito* 'little horse', *cabellera* 'mane', etc., to be discarded manually. This limitation, largely inherent in newspaper libraries, precluded automatic compilation of the base corpus and required further data selection for identification and separation of so-called false positives. Even so, HD proved essential for the experimental data attestation used in this chapter.

On the other hand, the history of Spanish texts parallels in the main the RAE's diachronic corpora, namely CORDE and CDH. The latter shows major advantages by virtue of its three layers of documents, and they justify its use as a control corpus: a) a higher philological control of the sources; b) lemmatization; and c) morphosyntactic annotation.

The HD evidence is contrasted with CDH evidence from outside newspaper sources. Section 4 presents data analysis according to text type, diachrony, and diatopy. In the latter, geolocation data attested in historical press are more precise and homogeneous than in the rest of text types. This is because only newspapers data can count on exact production dates (day, month, year) and location.³ The third corpus, of a lexicographic kind, covers Spanish dictionaries over time and relies on the NTLLE for the attestation of *cabalito* in (non-)academic lexicographic practice.

³ Early newspaper practice often published one and the same text on various dates and in various locations. For more accurate chronological attestation, the earliest date is recorded here for neologisms, even if the qualitative analysis takes account of all the published instances of the same use.

4 The History of *cabalito* ‘exactly’ in Corpora. Grammatical Status and Use Attestation

Some diminutive adjectives can be used as adverbs in present-day Spanish (*justito* ‘just.DIM’, *rapidito* ‘fast.DIM’). In this context, the case of *cabalito* is a bit of a mystery for the following reasons: i) grammatically, this adverb has been neglected by both synchronic and diachronic research; and ii) the reference corpora do not attest examples for the 21st c., and even past evidence is scarce. The following sections show the sharp contrast between this lack of evidence of *cabalito* and the frequent use in historical press.

The following is intended to cast light based on the analysis of the evolution of *cabalito* in Spanish by identification of its morphosyntactic, diachronic and diatopic profiles. To that end, the state of the art is first reviewed and the data collected for *cabalito* from HD as the base corpus and from CDH as the control corpus are then contrasted and discussed.

4.1 Diminutive Adverbs in Spanish Grammar

According to the reference grammar (NGLE: § 9.2), the adjectives that can be used as adverbs often do so in their diminutive forms too. This has been described synchronically, and has been reported to be more frequent in American Spanish than in European Spanish. Table 1 shows the contrast according to dialect, here

Table 1: Diminutive adverbs in NGLE. Cross-dialectal examples.

American Spanish	American and European Spanish
ahicito	cerquita
ahorita	despacito
alrededorcito	poquito
allacito, allicito	prontito
apenitas	
aquicito	
antesito	
despuesito	
detrasito	
enantito	
nomasito	

illustrated with a dozen cases available only in American Spanish vs. three cases available in both American and European Spanish.

Gerhalter (2020: 190–194) confirms this dialectal contrast in her account of the paradigm of focusing adverbs, *cabalito* among them. In this paradigm, seven adjectives/adverbs are listed, and three of them have a diminutive form: *preciso* ‘precise[ly]’, *exacto* ‘exact(ly)/*exactito*’ exact(ly)-DIM’, *justo* ‘just’/*justito* ‘just-DIM’ and *cabal* exact(ly)/*cabalito* ‘exact(ly)-DIM’. Gerhalter illustrates the use of the three diminutive adverbs (*exactito*, *justito*, and *cabalito*) with examples by South American literary authors recorded in the RAE corpora. Still, while the first two are illustrated with 20th c. examples, *cabalito* is illustrated with evidence of the 19th c. The suffix *-ito* in the adverbs *exactito*, *justito*, and *cabalito* adds an affective and emphatic nuance of meaning, but only in *cabalito* does Gerhalter underline the following difference: It is a statement marker in 19th c. use (dated 1875), as in examples by the Peruvian author Ricardo Palma (1).

- (1) El padre Arce quedó un minuto pensativo; y luego, pegándose una palmada en la frente, como quien ha dado en el quid de intrincado asunto, exclamó: “¡*Cabalito!* ¡Eso es!” (Ricardo Palma, *Tradiciones peruanas*, third series, 1875. CDH, cited by Gerhalter 2020: 191).

All in all, the little evidence available of the adverb *cabalito* suggests that only this adverb developed specific pragmatic values, as attested in the American variety at least since the late 19th c. The following questions, unanswered so far, can then be asked to give shape to the following section on data analysis: Where, when and how did this functional peculiarity of *cabalito* arise? Is this grammatical profile of *cabalito* limited to the 19th c.? Is *cabalito* historically one of the few diminutive adverbs available both in European and in American Spanish, or is it only in American Spanish?

4.2 *Cabalito* in Corpora

This is a contrastive analysis of the diachronic use of *cabalito* in HD (corpus base) and CDH (control corpus), according to four variables: i) evidence of use and morphosyntactic profile; ii) chronological attestation; iii) text types; and iv) diatopy.

4.2.1 Evidence of Use and Morphosyntactic Status

Table 2 shows a much stronger attestation of *cabalito* in the base corpus (317 occurrences) than in the control corpus (27 occurrences). The former, 317 occurrences in HD, are dated between 1790 (example 2) and 2022 (example 3), whereas the latter, 27 occurrences in CDH, are dated between 1771 (example 4) and 1977 (example 5).

HD attests four early examples of *cabalito* in the 18th c. (1790–1799) by contrast with one example in CDH (1771). The latter is, incidentally, the earliest attestation available. Thus, the 18th c. examples of *cabalito* in the base corpus are four times as many as the ones in the control corpus. This lack of balance grows exponentially in the 19th and 20th centuries. The two sources also differ widely as regards current use: *cabalito* is attested for the 21st c. in HD, but it is not in the RAE corpora.

The contrast in the number and in the chronological distribution of examples suggests that HD and CDH are two entirely different accounts of *cabalito*, as will be shown below.

Table 2: Evidence of *cabalito* in the base corpus vs. the control corpus (18th–21th centuries).

	BASE CORPUS (HD)	CONTROL CORPUS (CDH +CORPES XXI)
18th c.	4	1
19th c.	160	19
20th c.	127	7
21th c.	26	0
	317 occurrences in 302 documents	CDH: 27 occurrences in 14 documents

- (2) 1790. Madrid. Señor Editor: he visto en el n. 403 que el Caballero A.C. escribe quejándose del Señor *Quiqondam* y de mí. ¡Válgate Dios, que nunca hemos de poder contentar a todos! [. . .] Unos lloran de lo que otros ríen *cabalito*: eso es el mundo (*Correo de Madrid o de los ciegos*, Letter to the editor, 27/10/1790, page 7).
- (3) 2022. Toledo. No tengo memoria de mi río. Nací, justo, *cabalito*, aquel año que lo robaron (*ABC*, Toledo edition, 19/06/2022, editorial article on local topics, page 65).

- (4) 1771. Madrid. En el Lavapiesillo, / por el verano, / de aquesta forma cantan / majas y majos: “Que si ronda mi calle / Paco el herrero, / no le importa a ninguno, *cabalito* / (ea, ea, ea, ea, ea, ea), / segurito / (ea, ea, ea, ea, ea, ea). No le importa a ninguno / y yo requiero” (Anonymous, “El juego del burro. Tonadilla a tres” in *Tonadillas teatrales*. Madrid: Tipografía de Archivos, 1932. In CDH: Fiction/poetry).
- (5) 1977. Perú. “Los amores de un bebé y una anciana que además es algo así como su tía” –me dijo una noche la tía Julia, mientras cruzábamos el Parque Central–. “*Cabalito* para un radioteatro de Pedro Camacho” (Vargas Llosa, Mario, *La tía Julia y el escribidor*, Barcelona: Seix Barral. In CDH: Fiction, novel).

As for the grammatical status of *cabalito*, the earliest and latest attestations in the two sources are evidence of adverbial use alone, as this is the prevailing use in the two corpora. The quantitative analysis of all the occurrences shows that the adjectival use of *cabalito* is not attested in CDH and is rare in HD. Actually, only three out of the 317 examples of the base corpus are used as adjectives (see 6–8 below), and the rest are instances of adverbial use. Overall, the diminutive adverb *cabalito* is, typically, an adverb used during the course of ordinary spoken interaction. In these interactions, *cabalito* starts turn-taking and is often used for the expression of agreement (‘language accuracy’), but it may also be used for disagreement in contexts like (4), ironical and jocular, where ordinary interaction takes place by the less privileged social strata in the area known as *Lavapiés*, in Madrid. Such syntactic and semantic specialization of *cabalito* becomes its main sign of identity and is a regularity of diachronic use attested in the two corpora. Besides language accuracy, and much less prominently, the adverb *cabalito* may also denote numerical or mathematical accuracy in the measurement of time, weight, etc. This semantic nuance of meaning, rarely attested in the corpora, is illustrated with example (3), where *justo* and *cabalito* are used as synonymous adverbs.

- (6) 1884. Burgo de Osma. *Cosas y casos*. [. . .] Caballeros, les advierto a ustedes que soy ilustre [. . .] y aun cuando juzgan que soy el loco de la casa, no dejo de tener mi juicio muy *cabalito* (*La Propaganda: revista quincenal de intereses materiales, ciencias y literatura*, local news, 14/11/1884).
- (7) 1966. Madrid. [Informaciones de espectáculos]. Tercera de la Feria granadina. Falleció esta madrugada don Ricardo Calvo, dos toreros y un toro. Granada. El conde de la Corte envió un encierro *cabalito* de peso. Hubo uno de 435 kilos y otro de 437. Seguramente dos perdieron 100 gramos en el

viaje y no pudieron pasar. ¡Estos toritos que ponen tan poco de su parte! (*Informaciones*, bullfight news, 13/06/1966, page 13).

- (8) 2008. Sevilla. Si Pepín Liria se despide de El Puerto con un indulto, el público *cabalito* que conservó sus entradas, aunque no cayó nada bien la sustitución que anunció la empresa el mismo día (*Diario de Sevilla*, bullfighting, culture and leisure, 16/08/2008,⁴ page 50).

4.2.2 Diachronic Evidence: Stages in the Evolution of *cabalito*

The lifespan of *cabalito* is over two centuries according to the base corpus, specifically 232 years from the earliest attestation, in 1790, to the latest in 2022. As described above, this period has been divided into 30-year segments of time, and each segment counts at least on ten attestations. By contrast, ten occurrences per segment is the highest record in the control corpus, and it is available only in one of the eight chronological segments (1821–1851). Even more, no attestations of *cabalito* are available from the earliest and latest chronological segments of CDH, namely 1790–1820 and 2007–2022, respectively.

Table 3: HD and CDH diachronic evidence of *cabalito*: Chronological distribution.

	Base corpus (HD)	Control corpus (CDH)
1771	0	1
1790–1820	10	0
1821–1851	35	9
1852–1882	68	9
1883–1913	95	2
1914–1944	62	1
1945–1975	12	4
1976–2006	18	1
2007–2022	17	0

⁴ This bullfight piece of news was published the same day in the *Diario de Cádiz* (page 42) and in the *Diario de Jerez* (page 39).

Table 3 shows four landmarks in the evolution of the use of *cabalito* according to the base corpus: i) the earliest records (1790–1830); ii) the earliest evidence of stabilization (1831–1851); iii) standardization (1852–1952); and iv) obsolescence from the mid-20th c. up to present-day. These stages are described in detail below.

Cabalito is recorded as a neologism in the base corpus between 1790 and 1830, as this is the segment where the earliest ten examples are attested in ten sources. As mentioned above, the earliest records date back to the last decade of the 18th c. During this period, this use started to seep into written Spanish, even if it must have been in the spoken language at least since the mid-18th c.

The number of attestations in the base corpus increased sharply in the second stage (1831–1851), namely more than three times as many compared with the former segment of time: From 10 to 35 instances. This record shows that *cabalito* is no longer a neologism and has become stable in the history of Spanish texts by the mid-19th c.

The third stage spans a century of general change (1852–1952), during which the use at issue becomes standardized. This period at the turn of the 20th c. attests the highest number of examples of *cabalito*: Compared with the first segment above, the number of occurrences is between 6 and 9 times as high: Nearly 100 of the 317 records of the base corpus date from this period and, as a result, standardization can be said to peak during this period in the evolution of *cabalito*.

The last segment covers from the mid-20th c. up to present-day. The figures decrease markedly and fall below 20 instances, i.e. between 4 and 8 times less than in the former stage.

The picture obtained from the control corpus is quite otherwise. Unlike the base corpus, where the case under study is attested by over 300 authors, the control corpus gives evidence of only 11 occurrences. According to the CDH, only one case is attested in the 18th c. and another in the first half of the 19th c., specifically in 1832, so the use of *cabalito* in these two periods is quite marginal. Still, the second half of the 19th c. attests 11 occurrences. Thus, while the control corpus suggests a marked decrease in the use of *cabalito* at the turn of the 20th c., the base corpus, remarkably, suggests the opposite and gives evidence of the widest spread recorded. No attestation is available from the RAE corpora (CDH and CORPES XXI) for the period 1971–2023, so *cabalito* appears to be falling out of use, even if the base corpus confirms the sustained occurrence from the 18th c. up to present day.

Finally, the morphosyntactic status of *cabalito* develops differently in the two corpora too. All the CDH occurrences are adverbs, whereas the base corpus shows that *cabalito* was used also as an adjective from the third stage onwards, which is when it became most widely used: Four HD records dated from 1884 onwards evidence an occasional use as an adjective, even if the use as an adverb prevails and is attested in all the four stages considered here. All in all, CDH data, scarce and

chronologically sparse, attest only the adverbial use, and a rare, peripheral use as an adjective. As shown in Table 3, the base corpus shows the opposite.

It should also be underlined that the data of *cabalito* in the base corpus do not allow, strictly speaking, the identification of evolutionary cycles according to relative frequency. The technological resources of newspaper libraries do not offer statistical accounts of *cabalito* over time based on accurate quantitative methods. The results presented here are based on manual counts of absolute frequencies after due data selection. Despite the difficulties inherent in the use of HD as a source of unencoded data, the historical press reveals a longer lifespan and a wider distribution of *cabalito* than the RAE corpora suggest. Note that the morphosyntactic tagging used in the latter corpora does not allow quantification of diachronic tendencies: Each occurrence of *cabalito* must be marked as adjectival or adverbial manually, because the CDH tagger marked each occurrence of *cabalito* both as an adjective and as an adverb. As a result, only after manual analysis can it be ascertained that all the 27 occurrences of *cabalito* in the CDH verify the adverbial use, and that no evidence of the adjectival use is available.

4.2.3 Textual Evidence

Table 4 shows the journalistic genre's major role in the development of *cabalito*, as it is attested historically in two text types: i) HD newspaper texts; and ii) literary texts, extensively covered by CDH. Most of the occurrences of *cabalito* in CDH are recorded in fiction texts, whether poetry or prose. Out of 27 instances, only 4 are from non-literary sources: They are from a volume on bullfight news written in 1970 by Díaz-Cañabate. As the news had already been published in specialized

Table 4: The occurrence of *cabalito* in two text types.

	Newspaper	Literary texts	Total
1771	0	1	1
1790–1820	10	0	10
1821–1851	35	9	44
1852–1882	68	9	77
1883–1913	95	2	97
1914–1944	62	1	63
1945–1975	16	0	16
1976–2006	18	1	19
2007–2022	17	0	17
Total	321	23	343

forums (sports publications), these four instances are listed under the journalistic genre (Table 4, row 7), whose quantitative relevance compared with literary texts appears at the foot of the table: 321 occurrences in newspapers vs. 23 in fiction texts, i.e. nearly 14 times as many.

Diachronically, the journalistic genre influences the four stages of the evolution of *cabalito* above especially heavily during the stages of innovation (1790–1830) and standardization (1852–1952). Most of the occurrences of *cabalito* in the two text types under consideration occur between 1852 and 1944: 225 instances are from the press, whereas CDH attests 12 occurrences by six authors (Ascasubi, Gaspar Enrique, José María de Pereda, Ricardo Palma, Jacinto Octavio Picón, and Eduardo Blanco). While the latter CDH data may suggest that *cabalito* was specific of literary texts, the former evidence proves that the press was the true catalyst for the standardization of *cabalito*. The journalistic genre also comprehends a number of text subtypes, so *cabalito* is recorded in news on a range of topics: Parliamentary news (9), opinion articles (10), letters to the editor (13), editor's releases (11), and even bullfight news (14).

- (9) 1838. Madrid. [Intervención acalorada del Sr. Sancho, diputado por Valencia] El Sr. SANCHO: “Pues ahora contesto que ese cálculo es monstruosamente exagerado [. . .] El Sr. Martínez de la Rosa dice, que si no damos el diezmo, el clero se queda sin comer; pues yo digo lo contrario, si damos el diezmo el clero se queda sin comer. *Cabalito*. [. . .] Para mí es inconcebible” (*El Correo nacional*, 31/5/1838, page 3).
- (10) 1845. Madrid. [Artículo de opinión] Conociendo el poco valor de semejante testimonio, se apresuró a pronosticar que los que habíamos dicho que la carta anterior era falsa, diríamos que también lo era la nueva. *Cabalito*: la misma fuerza nos hace la una que la otra (*La Esperanza*, Madrid, 20/11/1845, page 1).
- (11) 1882. Burgos. [Nota del editor en respuesta a la de dos lugareños que se denominan “cándidos”] si ustedes son cándidos hay que meter en la cárcel a los innumerables mártires de Zaragoza, *cabalito* (*El Papa Moscas: periódico satírico*, Burgos, 04/06/1882, page 2).
- (12) 1891. Madrid. [Correspondencia particular] Si son pocos los números que le faltan, tal vez podamos remitírselos. Y eso sería lo mejor. *Cabalito* (*Madrid Cómico*. 21/02/1891, page 7).

- (13) 1894. Soria. [Carta al director] ¿Construcciones o ruinas? [. . .] Ya veo que se me va a hacer una pregunta suelta: ¿Y la Sociedad de socorros mutuos de Soria qué piensa de construcciones? *Cabalito*, ciudadanos. Formada esa sociedad obrera de pobres y ricos; confundidos en ella los de chaqueta con los de levita; hermanadas las ideas y los propósitos, deberíamos todos tomar un nuevo rumbo. [Firma el Pobrete de la clase] (*El Noticiero de Soria*, 03/03/1894, page 2).
- (14) 1952. Logroño. [Taurinas] En Madrid sale en hombros un logroñés. Fenómeno habemus. . . Ya tiene La Rioja su torero. . . De pocos días data el sensacional descubrimiento. . . pero el hecho es cierto, *cabalito*, sin lugar a la más ligera duda (*La Rioja. Diario político*, Logroño. 23/11/1952, page 3).

Still, most of the occurrences of *cabalito* in newspaper texts come from the section on society news about various events. This highly popular section started in the 19th c. to cover local news, gossip, tales, humour and all kind of reviews about current issues, all of which were key to a newspaper's commercial viability. From 1840 onwards, the section also covered serial fiction, with novels in serial form. In the base corpus, *cabalito* is attested in these within excerpts of simulated orality, similarly to others in CDH (see examples under 1, 4 and 5 above). From the mid-19th c. onwards, the highly successful serial fiction takes over the contents of the society news, so the news on current issues are diverted to smaller sections of variety, gossip, and the like. Thus, the data on newspaper texts from the mid-19th c. onwards of Table 4 may also include excerpts of novels. This section contains 12% of the occurrences of *cabalito* in press texts, remarkably both literary and most of the non-literary too. The authors of this section in the 19th c. are typically fond of lexical fashions and of the most colloquial registers. Whether literary or not, the contents of the society news display features of less elaborate, rushed writing than in other newspaper sections, they forerun yellow press, and they also use local vocabulary that is eschewed in other text types. Thus, the society section discloses, from the mid-19th c. onwards, a more representative gamut of colloquial forms of the time than the literary canon does.

Overall, only HD offers evidence of *cabalito* both in fiction and non-fiction. For over two centuries, the attestation of this diminutive has been limited to the production by a small group of 11 authors, but the newspaper corpus proves that *cabalito* was not just a stylistic device of the poets and novelists of the literary canon since the 18th c. The journalistic genre also has an added value in that it gives evidence of the language proper to press language, but also of fiction texts intended for the masses. As will be shown below, the authors of such sections write informally, for a wide readership, often under a pen name, and become

both major participants and unique informants as regards language diversity, especially of the diaphasic and diatopic kind.

4.2.4 Diatopic Evidence

Diminutive adverbs are described above as more frequent in American than in European Spanish. Their geolectal nature is often underlined in the literature, and specific forms are recorded only for American Spanish (Table 1). Still, the history of the adverb *cabalito* seems to show that specifically European forms may have been available too. The diatopic indicators of the base corpus suggest that *cabalito* was a divergence in European Spanish since its earliest occurrences in the 18th c. until its standardization at the turn of the 20th c., as shown in Table 5.

Table 5: Diachronic interdialectal distribution of *cabalito*, where the first figure is the number of American examples and the second figure is the number of European examples.

	Base corpus (HD)	Control corpus (CDH)
1771	0/0	0/1
1790–1820	0/10	0/0
1821–1851	0/35	2/7
1852–1882	2/66	6/3
1883–1913	0/95	1/1
1914–1944	2/60	0/1
1945–1975	0/12	0/4
1976–2006	0/18	1/0
2007–2022	0/17	0/0

Pan-hispanically, the geolocation of the HD data refer virtually all the occurrences of *cabalito* to European Spanish (313 out of 317 occurrences), and only four to American Spanish (Cuban and Argentinian Spanish as in 15–17). At this point, it is in order to underline that *cabalito* is not recorded in the *Corpus Diacrónico y Diatópico del Español Americano* (CORDIAM). As this corpus covers a wide range of genres (letters, legal and administrative texts, literature, press) dated between 1494 and 1905, *cabalito* then becomes more strongly associated with European Spanish, as suggested by the base corpus data. Table 5 shows that the HD data suggest an extensive use in European Spanish press and quite the opposite in American Spanish press. Both CORDIAM and HD data suggest that *cabalito* has

little history American Spanish: The four occurrences in American Spanish attested in the base corpus are dated between 1860 and 1928. *Cabalito* must have been used rarely outside Spain during this period, with occasional instances in the literary section of South American press. The examples 15–17 show that the fiction dialogues of some American newspapers have used *cabalito* occasionally since 1860, i.e. 70 years after the earliest records of the European Spanish press.

Unlike CORDIAM and HD, CDH data reveal a well-balanced distribution of *cabalito* between American and European literary authors in the control corpus, so a more detailed interdialectal analysis is in order. All the instances contained in CDH are by eleven authors, five American (Ricardo Palma and Vargas Llosa from Peru, Manuel Eduardo de Gorostiza from Mexico, Eduardo Blanco from Venezuela, and Hilario Ascasubi from Argentina) and six European (Mariano José de Larra, Ayguals de Izco, Antonio Díaz Cañabate, Enrique Gaspar, Jacinto Octavio Picón, and José María de Pereda). Thus, out of the 27 CDH occurrences, 10 are American and 17 are European. Of these, the former also suggest the opposite evolution to the one obtained from the base corpus data: The latest attestation is recorded in Peru in 1977, and even for the period 1852–1882 the RAE corpus lists more occurrences of American Spanish (6 occurrences) than of European Spanish (3 occurrences). By contrast, the base corpus points in the opposite direction based on six times as many examples than in CDH: Only 2 out of 68 occurrences of *cabalito* are from American Spanish. The 66 records of European Spanish are thirty times as many as those of American Spanish, and they grow steadily in the base corpus too. The conflict between such opposite interdialectal data may be as a result of CDH's design, whereby fiction by canonical authors is primed and the journalistic genre, which is where the form under study is more likely to occur, is largely overlooked. The balance between American and European authors in CDH may be due to the literary connections between educated authors across the Atlantic, such that their exchange of texts may have been an elite channel for the specific dissemination of lexical fashions at various points of the Spanish speaking countries.

- (15) 1860. Cuba. *Memorias de una viuda. Mi segundo marido* (continuación) [. . .] Ah! Boca de serafín. . . Dios te guarde, pico de oro. . . acertaste. . . *cabalito, cabalito*. Estuve en la gloria con aquel bribonzuelo (*El Moro Muza*, La Habana, 08/01/1860, page 3).
- (16) 1869. Cuba. ¡*Un artículo de punta!* [editorial] No sé si lo de *punta* querrá decir artículo agudo, ó tal vez *punta* que hiera. [. . .] *D. Pacuato* quiere hacer milicianos nacionales, se insurreccionan combatiendo la nacionalidad

española, para lo mismo que en Alcolea ¡*cabalito!* . . . para dar la libertad al país (*El Moro Muza*, La Habana, 09/05/1869, page 1).

- (17) 1928. Argentina. *Los regionales. Pava y Varita*, por Fausto Burgos. [. . .] En clase, a hurto del catedrático, sacaba yo los billetes para contarlos y recontarlos. “Te ha dado justo, ¿che?”, preguntaba Chumbo. “*Cabalito*”. No veíamos la hora de llegar a la calle, de llegar a nuestra casa de huéspedes [. . .] “¿Te dieron de más?”. “¡*Cabalito!*”. Y echábamos cuenta: 200 x 400 = 80.000 pesos (*Caras y Caretas*, Buenos Aires, 01/12/1928, pages 150–151).

As the for the intradialectal properties of *cabalito* in European Spanish, the questions arise: When did this innovative use start, and how widespread did it become in European Spanish? As noted above, only the base corpus can supply geolocation by country, province, and specific location, in addition to the date and place of publication. The analysis of the diachronic diatopic data shows that the earliest records date from the last decade of the 18th c. and refer to publications in Madrid. Certainly, the highest number of attestations are from Madrid throughout the eight segments of time under consideration here. Newspapers from other places can be added at the turn of the 19th c. as early records, i.e. as records from before 1815: (18) from Zaragoza, (19) from Cádiz and (20) from Seville. From 1815 onwards, the number of publication places increases steadily until widespread occurrence over the Spanish press in the first decade of the 20th c. According to the base corpus, the dissemination amounted to newspapers published in up to 43 locations. The diatopic distribution of the occurrences of the base corpus is shown by region in the tables of Appendix 1. Their data are graphically represented in the following Map 1.

- (18) 1798. Zaragoza. *Carta publicada en los Diarios de Madrid números 199 y 200 del miércoles y jueves, 18 y 19 de julio de 1788* [. . .] Por qué el adverbio *souvent* le (sic) traduce ahí con frecuencia y vele (sic) aquí, quatro líneas más abaxo, de *quando en quando*? “*Cabalito*”, dixo él: “souvent lo mismo es que frecüentemente” (*Semanario de Zaragoza*, Zaragoza, 13/08/1798, page 3).
- (19) 1813. Cádiz. ¡Qué poco calcula el que tal cosa propone! Peor lo habíamos de pasar: *cabalito*; peor (*El Duende de los cafées*, Cádiz, 08/08/1813, page 6).
- (20) 1814. Sevilla. [. . .] ¡Hombre necio! Si la casa se está quemando! Piensa en socorrerla, y luego en adornarla. *Castaña*. ¡*Cabalito!* Pero agregue usted a eso las otras dos disposiciones (*La Tía Norica*, Sevilla, 1814, no. 20, page 4).



Map 1: *Cabalito* in Spanish newspapers (HD, 1790–2022). Intradialectal diatopic distribution.

Attestations of *cabalito* can be found in the local press during the standardization stage, e.g. in newspapers of Jerez (province of Cádiz), Orihuela (province of Alicante), Burgo de Osma (province of Soria), Gandía (province of Valencia) and many others, both in reprints of news of Madrid and also in local news. These records bear witness to the wide spread of this colloquial term in Spain until the mid-20th c. From this point onwards, its distribution in the press becomes much more limited as *cabalito* becomes gradually obsolescent. By the 21st c., the attestations in the base corpus are occasional and mainly limited to the region of Castilla-La Mancha. Further research may reveal whether *cabalito* remained only in this gelect and fell out of use elsewhere.

The final picture emerging from the base corpus presents *cabalito* as an European singularity arisen in Madrid in the last decade of the 18th c. and used in the rest of the country one century later after dissemination between 1850 and 1950. According to the HD, *cabalito* receded mainly in the 21st c., as it seems to have become a dialectal form used in Toledo and nearby areas. This adverb therefore evolved from a catchy colloquial form of Madrid in 1790 to a frequent term in the press of the country for two centuries and, finally, probably a mere diatopic form of La Mancha in the 21st c.

5 Encoded Use: *Cabalito* in the History of Spanish Dictionaries

An innovative use's relevance for metalinguistic analysis and its full acceptance by a language's reference codes are clear evidence of standardization. This section researches the encoding stage of *cabalito* in the history of Spanish dictionaries. The corpus of dictionaries used here covers all the academic and non-academic references contained in the NTLLE, i.e. over 90 titles published between the 15th and the 20th centuries. The 21st c. is here accounted for by academic and non-academic references too, namely the electronic editions of the RAE Dictionary (DLE 2023) and of the *Diccionario del español actual* (DEA 2023). The latter's diatopic coverage is limited to European Spanish.

The aim is to find out when and how the term *cabalito* was recorded in the history of Spanish dictionaries according to the gloss type, grammatical classification, use marks and examples. Table 6 shows in chronological order (1853–2023) how *cabalito* was recorded in academic and non-academic dictionaries, and classifies the four variables listed above as columns, where the leftmost column shows the models recorded by the dictionaries as in examples 21 through 23.

Table 6: *Cabalito* in the history of Spanish dictionaries.

Year, author	Gloss	Grammatical category	Usage	Occurrences
1853. Domínguez, s.v. <i>cabalito</i> . (idem Domínguez 1869)	Ironical reply to express negation, mockery, confirmation. Similar uses are also reported	Diminutive adjective for <i>cabal</i>	Unmarked	3 use patterns (examples under 21)
1917. Alemany, s.v. <i>cabalito</i>	Cabalmente	Diminutive for <i>cabal</i> . Masculine adverb	Informal	No examples
1936. RAE, s.v. <i>cabalito</i>	Cabal. Cabalmente	Masculine adverb	Informal	4 use patterns (examples under 22)
2023. DEA, s.v. <i>cabal</i> . (idem DEA 1999 and DEA 2007).	Exactamente. Dicho para asentir a lo que acaba de oírse 'expressing agreement'	Adverb	Colloquial	1 use pattern (example under 23)

- (21) 1853 Domínguez (ídem in Domínguez 1869). [Para la negación] “*Me hará vd. gusto*”. “*¡Cabalito!*”; [para expresar burla] “*Cabalito, que me gusta mucho*”; [para expresar la decisión] “*¿Irás a verle?*”. “*Cabalito*”.
- (22) 1936. RAE (*Diccionario histórico*) “Empiece usted por su casa / a corregir el exceso” / “*¿Por mi casa?*”. “*Cabalito*” (Ramón de la Cruz, 1731–1824); Que sois mi suegro, / *cabalito*, en dos palabras (Leandro Fernández de Moratín, 1760–1828); “*¿Le amáis por fe?*”. “*Cabalito*” (Juan Eugenio Hartzenbusch, 1806–1880); *Cabalito*. Eso quiero, que gastes de lo tuyo (Jacinto Octavio Picón, 1852–1923. *La honrada*, ed. 1924).
- (23) 2023. DEA (ídem en DEA 1999 y en DEA 2007). “Esa es la fuente de vino que dicen ¿no?”, le preguntaron. “*Cabalito*” (Ángel María de Lera. 1912 Baides-Castilla la Mancha-1984, Madrid; *La boda* [1959], in *Novelas*. 1966).

As can be seen from the above, *cabalito* is recorded in few dictionaries, current or past. Two major points from Table 6 can be underlined: i) the RAE has not recorded *cabalito* in its official dictionary of Spanish in three centuries, but it did in the first diachronic dictionary (RAE 1936); and ii) the earliest record is by non-academic lexicography (Domínguez in 1853), and it also records this form in present-day (DEA 2023 is the only 21st c. dictionary to cite this diminutive in current Spanish use). Table 6 is inspected in further detail below.

Cabalito is recorded only in four dictionaries during the period 1853–2023: i) the two editions of *Diccionario nacional* by Domínguez (1853, 1869); ii) Alemany’s (1917) dictionary; iii) the RAE’s first historical dictionary (RAE 1936); and iv) the three editions of the dictionary by Manuel Seco, Olimpia Andrés and Gabino Ramos (DEA 1999, DEA 2007, DEA 2023). In all cases, the same information is given across editions of the same dictionary. Column 1 shows that this diminutive is not recorded the same in all dictionaries. Domínguez was first to record *cabalito* and also to list it as a separate entry in his 1853 dictionary. Alemany (1917) and the RAE (1936) followed this convention in the 20th c. The relevance of *cabalito* in the 1853 and 1917 general dictionaries may stem from the relevance of the term at the time, which is also the period of highest frequency in the base corpus. At the turn of the 21st c., the DEA differs from the above and, following more orthodox criteria, lists *cabalito* as a variant form of *cabal*.

Table 6 evidences disagreement in the glosses. While Alemany (1917) and the RAE (1936) merely point out the semantic equivalence of *cabalito* with *cabalmente* and *cabal* ‘upright’ within the same word family, Domínguez (1853 and 1869) and the DEA (1999, 2007 and 2023) describe its use as a colloquial form. Notably, only the earliest and latest lexicographic records highlight the pragmatic function of

cabalito in conversation: Domínguez describes it as an ironical reply to express negation, mockery, confirmation, and the like (“estribillo irónico que manifiesta: una negación, la burla, la decisión; tiene otros usos del mismo tenor”), and the DEA dictionary lists *cabalito* (as a variant of *cabal* ‘upright’) to express agreement (“para asentir a lo que acaba de oírse”) meaning ‘precisely’ (“exactamente”). The semantic range is wider according to Domínguez than to the DEA dictionary, maybe for the more frequent use of *cabalito* in the mid-19th c. than at present, as shown by the newspaper corpus used here.

As for grammatical status, *cabalito* is first described as an adverb in the 20th c. dictionaries. Domínguez just mentions the morphological form as the diminutive adjective from *cabal* (“adjetivo diminutivo de cabal”), and Alemany (1917) and the RAE dictionary (1936) refer to it as a masculine adverb (“adverbio masculino”). As mentioned above, the DEA (1999–2023) equates *cabalito* and *cabal* when the latter occurs as an adverb.

Table 6 also shows that the 20th c. dictionaries mark the term *cabalito* as proper of *informal* use (Alemany 1917, RAE 1936) and later as *colloquial* (DEA 1999, 2007, and 2023). The *Diccionario nacional* (1853) does not add an explicit categorization, but Domínguez links up *cabalito* with colloquial register and, thus, describes it as a deeply ironic feature of spoken language. In either case, dictionaries explicitly or implicitly associate *cabalito* with the most informal registers. Colloquial forms more proper of spoken than of written language are typically not recorded in dictionaries, even if they have become colloquial pet phrases, as Domínguez suggests for *cabalito*. Actually, as mentioned above, this diminutive, first attested in 1760–1770 and widely used in the 19th c. press according to corpus evidence, was never recorded by the RAE’s general dictionary.

Examples 21 through 23 illustrate *cabalito* respectively as in Domínguez (1853), the first historical dictionary (RAE 1936), and the DEA 2023. Except for Domínguez (1853), who inserted his own examples in the gloss, all the examples are from literary texts: The RAE dictionary (1936) gives examples by four 18th and 19th century authors (Ramón de la Cruz, Moratín, Hartzbusch, and Picón), and the DEA 2023 cites *La Boda*, a 1966⁵ novel written by Lera. Whether fiction or non-fiction, all the examples are direct speech, and thus underline this diminutive’s colloquial nature.

Finally, no dictionary record gives diatopic information of use, whether inter- or intra-dialectal. The list of literary authors cited may be used, though, as indirect

5 This is the only attestation in the three editions of the DEA: A 1966 occurrence in literary language. The wider chronological coverage of the third edition (1950–2023, according to the foreword) compared with the first two (European Spanish of the second half of the 20th c.) does not add any attestation of *cabalito*.

evidence of diatopic information, if their provenance is researched (see Table 7): Thus, all the examples cited in the historical dictionary (RAE 1936) are by four authors from Madrid of the 18th and 19th century, one of whom is also listed in the CDH corpus (Jacinto Octavio Picón). Remarkably, the 17 European occurrences of *cabalito* in the RAE are by five authors born in Madrid, and one in Cantabria. Overall, the entire list of examples is by ten authors, nine of whom are from Madrid. Even the 1771 anonymous record is referred to Madrid. As a result, *cabalito* may have been historically associated with the Court of the 18th to the 20th centuries. However, this dialectal bias is not supported by HD evidence, as shown in the previous section. Diatopically, *cabalito* arose in Madrid in the second half of the 18th c., spread over European Spanish in the 19th c., and seems to have retreated to Castilla-La Mancha in the 21st c. This is also the origin of the 1966 example of *cabalito* by the novelist cited in the DEA (1999, 2007 and 2023).

The RAE data may suggest that *cabalito* arose and disappeared as a form specific of the Spanish spoken in Madrid, even if the evidence of newspaper data shows that it reached the dialectal status of a typical European form in the Spanish-speaking countries.

Table 7: The diachronic use of *cabalito* according to RAE 1936 and CDH. Birthplace of Spanish authors.

Author /Title	Date	Country	Author's provenance	No. of occurrences
Anonymous	1771	Spain	Madrid	1 (CDH)
Ramón de la Cruz (Madrid 1731-Madrid 1824), <i>Obras</i>	18th c.	Spain	Madrid	1 (RAE 1936)
Leandro Fernández de Moratín (Madrid 1760-París 1828) <i>Obras</i>	18th c.	Spain	Madrid	1 (RAE 1936)
Mariano José de Larra (Madrid 1809-Madrid 1837) <i>Traducción de Roberto Dillón</i>	1832	Spain	Madrid	1 (CDH)
Juan Eugenio Harzzenbusch (Madrid 1806-Madrid 1880), <i>Obras</i>	19th c.	Spain	Madrid	1 (RAE 1936)
Gorostiza <i>Contigo pan y cebolla. . .</i>	1833	Mexico	Madrid	2 (CDH)
Ayguals de Izco <i>La Bruja de Madrid. . .</i>	1850	Spain	Madrid	7 (CDH)

Table 7 (continued)

Author /Title	Date	Country	Author's provenance	No. of occurrences
Ascasubi	1852,	Argentina		1 (CDH)
<i>Paulino Lucero</i> (1852)	1872			3 (CDH)
<i>Aniceto el Gallo</i> (1872)				
Gaspar Enrique (Madrid 1852-Francia 1902)	1868	Spain	Madrid	1 (CDH)
<i>La Chismosa</i>				
José María de Pereda (Polanco 1833, Santander 1906)	1871	Spain	Cantabria	2 (CDH)
<i>Tipos y Paisajes</i>				
Ricardo Palma	1875	Peru		1 (CDH)
<i>Tradiciones peruanas</i> , 3ª serie, (1875)	1877			
<i>Tradiciones peruanas</i> 4ª serie, (1877)				1 (CDH)
Jacinto Octavio Picón (Madrid 1853-Madrid 1923)	1890	Spain	Madrid	1 (RAE 1936)
<i>La honrada</i>				ídem (CDH)
Eduardo Blanco	1912	Venezuela		1 (CDH)
<i>Tradiciones épicas y cuentos viejos</i>				
Antonio Díaz-Cañabate (Madrid 1987-Madrid 1980)	1970	Spain	Madrid	4 (CDH)
<i>Paseíllo por el planeta de los toros, crónicas taurinas</i>				
Vargas Llosa	1977	Peru		1 (CDH)
<i>La tía Julia . . .</i>				

6 Conclusions

This paper underlines the need for a combined use of diachronic corpora (linguistically annotated, big data) and digital newspaper libraries (linguistically non-annotated, big data) in the pan-Hispanic research on certain issues of the evolution of colloquial Spanish, e.g. the adverb *cabalito* for the low frequency in diachronic reference corpora. This paper relies on HD as the base corpus and on CDH as a control corpus, and discloses to a large extent the evolution of this adverb on ac-

count of its diachrony evidence, its occurrence across text types, its diatopic evidence, and its grammatical status.

In the former respect, *cabalito* is an innovative form of the mid-18th c. It became widespread and reached full standardization in the late 19th c. In the 20th and 21st centuries it fell out of use. These radical changes can be found only in the base corpus, for two reasons: i) *cabalito* is attested for over two centuries (1790–2022) only in the base corpus, whereas it is attested in CDH only continually with major chronological gaps; and ii) the number of occurrences in the base corpus is 12 times as high as in the CDH corpus (317 vs. 27 occurrences, respectively).

Regarding text types, the development of *cabalito* was heavily influenced by the journalistic genre. The term is attested more frequently in rushed writing and in conditions of communicative immediacy, and is much more open to informal language: *Variety news* (local events, gossip) and *serial publications* (literary and non-literary). In these subtypes, *cabalito* is always recorded in direct speech and starts a turn. This is evidence of this diminutive's association with the spoken mode.

Pan-Hispanically, *cabalito* is specific of European Spanish. Interdialectal distribution shows that the term was attested in Spain but not in America for over two centuries. Interdialectally, it started in Madrid and spread from the geolect of the *Meseta* to the rest of diatopic varieties.

Morphosyntactically, the adverb *cabalito* expresses agreement with a previous statement, or disagreement, if used ironically. Under this morphosyntactic and pragmatic specialization, it is highly frequent, even if it is rarely recorded in Spanish dictionaries.

Appendix

Spanish Historical Press, with Attestations of Cabalito by Region

Andalusia. Journalistic examples of *cabalito* (1809–1944)

Place	Newspaper. Date, Page
Almería	<i>El Radical</i> . 16/07/1906, p. 4
Almería	<i>Yugo</i> . 09/07/1944, p. 7
Almería/Vélez Blanco	<i>La Opinión</i> . 14/11/1895, p. 1
Almería/Vélez Rubio	<i>El Loro</i> . 01/12/1913, pág. 1
Cádiz	<i>El Duende de los cafés</i> . 08/08/1813, p. 6

(continued)

Place	Newspaper. Date, Page
Cádiz	<i>El correo de Cádiz</i> . 04/08/1920, p. 1
Cádiz/Jerez	<i>El Progreso</i> . 21/04/1870, p. 2
Cádiz/Jerez	<i>El Guadalete</i> . 12/08/1920, p. 1
Córdoba	<i>El amigo católico</i> . 17/06/1875
Córdoba	<i>La voz</i> . 17/05/1924, p. 16
Córdoba/Pozoblanco	<i>El Cronista del Valle</i> . 28/05/1960, p. 2
Granada	<i>El Defensor de Granada</i> . 03/04/1890, p. 2
Granada	<i>El Defensor de Granada</i> . 23/12/1924, p. 2
Málaga	<i>Atalaya patriótico de Málaga</i> . 01/04/1809, p. 9
Málaga	<i>El Folletín</i> . 28/02/1875, p. 3
Sevilla	<i>La Tía Norica</i> . 1814, nº 20, p. 4
Sevilla	<i>ABC-Sevilla</i> . 22/04/1942, p. 9

Aragon. Journalistic examples of *cabalito* (1798–1927)

Place	Newspaper. Date, Page
Teruel	<i>Guía del magisterio</i> . 30/01/1881, p. 5
Teruel	<i>El Mercantil</i> . 09/01/1914, p. 3
Zaragoza	<i>Semanario de Zaragoza</i> . 13/08/1798, p. 3
Zaragoza	<i>Heraldo de Aragón</i> . 14/01/1927, p. 8

Asturias. Journalistic examples of *cabalito* (1886–1926)

Place	Newspaper. Date, Page
Gijón	<i>Gijón cómico</i> . 14/09/1889, n. p.
Gijón	<i>Páginas escolares</i> . 01/11/1926, p. 2
Oviedo	<i>La cruz de la victoria</i> . 25/11/1886, p. 2

Balearic Islands. Journalistic examples of *cabalito* (1871–1914)

Place	Newspaper. Date, Page
Mallorca	<i>El genio de la libertad</i> . 24/11/1839, p. 1
Mallorca	<i>El juez de paz</i> . 21/09/1871, p. 4
Mahón	<i>El bien público</i> . 8/03/1913, p. 2
Mahón	<i>La Alquitara</i> . 21/03/1914, p. 5

Canary Islands. Journalistic examples of *cabalito* (1908–1911)

Place	Newspaper. Date, Page
Tenerife/Sta. Cruz	<i>La gaceta de Tenerife</i> . 16/07/1910, p. 3
Tenerife/Sta. Cruz	<i>El Tiempo</i> . 05/12/1908, p. 2
Tenerife/Sta. Cruz	<i>Gaceta de Tenerife</i> . 19/08/1910, p. 4
Tenerife/Sta. Cruz	<i>El Progreso</i> . 22/02/1911, p. 3

Cantabria. Journalistic examples of *cabalito* (1844–1922)

Place	Newspaper. Date, Page
Santander	<i>La Verdad, diario de la mañana</i> . 08/01/1844, p. 2
Santander	<i>La Abeja montañesa: periódico de intereses locales</i> . 02/11/1865, p. 3
Santander	<i>La Abeja montañesa</i> . 1865, p. 3
Santander	<i>El Atlántico</i> . 28/08/1889, p. 2
Santander	<i>El Avisador</i> . 14/03/1897, p. 2
Santander	<i>La Atalaya</i> . 27/03/1905, p. 3
Santander	<i>El pueblo cántabro</i> . 22/07/1922, p. 8

Castile and León. Journalistic examples of *cabalito* (1882–1932)

Place	Newspaper. Date, Page
Burgos	<i>El Papa-Moscas: periódico satírico</i> . 04/06/1882, p. 2
Burgos	<i>El Papa-Moscas: periódico satírico</i> . 22/02/1885, p. 1
León	<i>Diario de León</i> . 05/05/1908, p. 3
León	<i>Diario de León</i> . 02/07/1932, p. 4
Salamanca	<i>La Legalidad: revista de asuntos administrativos</i> . 30/01/1982, p. 3
Salamanca	<i>El Salmantino</i> . 05/02/1910, p. 2
Salamanca/Béjar	<i>La Victoria: semanario de Béjar</i> . 26/07/1902, p. 1
Salamanca/Béjar	<i>La Victoria: semanario de Béjar</i> . 24/02/1906, p. 2
Soria/Burgo de Osma	<i>La Propaganda</i> . 14/11/1884, p. 2
Soria/Burgo de Osma	<i>La Propaganda</i> . 05/08/1891, p. 3
Soria	<i>El noticiero de Soria</i> . 03/03/1894, p. 2
Soria	<i>Ideal numantino</i> . 21/04/1911 p. 2

Castile-La Mancha. Journalistic examples of *cabalito* (1898–1930)

Place	Newspaper. Date, Page
Ciudad Real	<i>El Pueblo manchego</i> . 21/03/1911, p. 2
Cuenca	<i>El Catequista</i> . 31/01/1907, p. 3
Toledo	<i>La Aurora</i> . 18/10/1898, p. 4
Toledo	<i>El Heraldo toledano</i> . 29/06/1930 p. 10

Catalonia. Journalistic examples of *cabalito* (1881–1915)

Place	Newspaper. Date, Page
Barcelona	<i>El Mundo ilustrado. Biblioteca de las familias.</i> 1881–1883. N° 152, p. 26
Barcelona	<i>Iris.</i> 18/4/1903. p. 9
Gerona	<i>La nueva lucha: diario de Gerona.</i> 190/02/1888, p. 1
Gerona/Olot	<i>El Eco de la montaña.</i> 16/05/1897, p. 2
Tarragona	<i>Diario del comercio.</i> 06/08/1898, p. 3
Tarragona	<i>La Cruz.</i> 10/03/1915, p. 1
Tarragona/Tortosa	<i>El estandarte católico.</i> 11/07/1899, p. 2
Tarragona/Tortosa	<i>El Restaurador.</i> 30/10/1908, pág.3

Extremadura. Journalistic examples of *cabalito* (1885–1908)

Place	Newspaper. Date, Page
Badajoz	<i>El Avisador de Badajoz.</i> 19/03/1885, p. 3
Badajoz	<i>Noticiero extremeño.</i> 17/06/1906, p. 3
Cáceres	<i>Revista de Extremadura. Ciencia y arte.</i> 01/09/1903, p. 42
Cáceres	<i>El bloque.</i> 14/06/1907, p. 3

Galicia. Journalistic examples of *cabalito* (1889–1927)

Place	Newspaper. Date, Page
Lugo	<i>El Lucense.</i> 20/04/1889, p. 1
Lugo	<i>El norte de Galicia.</i> 02/07/1906, p. 3
Lugo	<i>Acción social.</i> 15/02/1921, p. 8
Lugo	<i>El Progreso: diario liberal.</i> 08/02/1927, p. 3

La Rioja and the Basque Country. Journalistic examples of *cabalito* (1893–1933)

Place	Newspaper. Date, Page
Logroño	<i>La Rioja: diario político.</i> 06/09/1893, p.2
Logroño	<i>La Rioja: diario político.</i> 28/03/1933, p. 4
Álava	<i>Heraldo Alavés: Diario independiente.</i> 15/04/1902, p. 1

Murcia. Journalistic examples of *cabalito* (1900–1903)

Place	Newspaper. Date, Page
Murcia	<i>La Juventud literario</i> . 10/04/1900, p. 3
Murcia	<i>Heraldo de Murcia</i> . 10/4/1902, p. 3
Murcia	<i>El Liberal</i> . 31/01/1903, p. 31

Valencia. Journalistic examples of *cabalito* (1881–1910)

Place	Newspaper. Date, Page
Alicante	<i>El Graduador</i> . 18-08-1898, p. 1
Alicante	<i>La Voz de Alicante</i> . 28/04/1906, p. 1
Alicante/Orihuela	<i>La lectura popular</i> . 15/09/1884, p. 2
Alicante/Orihuela	<i>El nuevo alicantino</i> . 10-09-1897, p. 2
Alicante/Alcoy	<i>Heraldo de Alcoy</i> . 04/09/1902, p. 4
Alicante/Gandía	<i>Revista de Gandía</i> . 30/05/1908, p. 1
Valencia	<i>Periódico monárquico</i> . 21/12/1881, pág. 2
Valencia	<i>El Pueblo: diario republicano de Valencia</i> . 09/08/1910, p. 2

Bibliography

- Alemay y Bolufert, José (1917): *Diccionario de la lengua española*. Barcelona: Ramón Sopena.
Available in NTLLE.
- Calderón Campos, Miguel (2024): “Spanish Corpora: Big (Quality) Data?”, in Ana Gallego Cuiñas and Daniel Torres-Salinas (eds.). *Humanities and Big Data in Ibero-America. Methodological Issues and Practical Applications*. Berlin/Boston: De Gruyter, pp. 109–127.
- Campos Souto, Mar (2018): “Bibliotecas y hemerotecas digitales en el NDHE”, in *Cuadernos del Instituto Historia de la Lengua*, 11, pp. 237–255.
- CDH = Real Academia Española: *Corpus del diccionario histórico de la lengua española*. <<https://www.rae.es/banco-de-datos/cdh>>.
- Company Company, Concepción and Rodrigo Flores Dávila (2017): “Género textual, diacronía y valoración de un cambio sintáctico”, in BRAE XCVII-CCCXV, pp. 203–239.
- Company Company, Concepción and Rodrigo Flores Dávila (2018): “El contraste *a por vs. por* con verbos de movimiento”, in *Revista de Filología Española*, 98(2), pp. 281–318.
- CORDE = Real Academia Española: *Corpus diacrónico del español*. <<https://www.rae.es/banco-de-datos/corde>>.
- CORDIAM = Company Company, Concepción and Virginia Bertolotti (dirs.): *Corpus diacrónico y diatópico del español de América*, Academia Mexicana de la Lengua. <<https://www.cordiam.org/>>.
- CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <<https://www.rae.es/banco-de-datos/corpes-xxi>>.
- CREA = Real Academia Española: *Corpus de referencia del español actual*. <<https://www.rae.es/banco-de-datos/crea>>.

- DLE = Real Academia Española (2023): *Diccionario de la Lengua Española*. <<https://dle.rae.es/>>.
- Domínguez, Ramón Joaquín (1846–1847/²1853): *Diccionario nacional o gran diccionario clásico de la lengua española*. Madrid-Paris. Available in NTLLE.
- García-Godoy, María Teresa (2015): “Political and Lexical Emancipation in Spanish America. The Nineteenth Century in the History of Americanisms”, in *Nineteenth-Century Context*, 37(4), pp. 321–339.
- García-Godoy, María Teresa (2017): “La diferenciación léxica del español de América. Anglicismos jurídicos e institucionales en la Colonia tardía”, in *Hispania*, 100(1), pp. 65–78.
- García-Godoy, María Teresa (2021): “De *madamas* y *madamitas*. Un tratamiento galicado en la historia del español moderno”, in *Rilce*, 37(1), pp. 46–72.
- Gerhalter, Katharina (2020): *Paradigmas y polifuncionalidad. Estudio diacrónico de preciso/precisamente, justo/justamente, exacto/exactamente y cabal/cabalmente*. Berlin/Boston: De Gruyter.
- HD = Biblioteca Nacional de España: *Hemeroteca Digital* <<https://hemerotecadigital.bne.es/hd/es/advanced>>.
- NGLE = Real Academia Española y Asociación de Academias de la Lengua Española (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.
- NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española*. <<https://apps.rae.es/ntlle/SrvltGUILoginNtlle>>.
- Octavio de Toledo y Huerta, Álvaro S. (2016): “Sin CORDE pero con red: algotras fuentes de datos”, in *Revista Internacional de Lingüística Iberoamericana (RILI)* 28, pp. 19–48.
- Seco, Manuel, Olimpia Andrés and Gabino Ramos (1999/2007): *Diccionario del español actual*. Madrid: Espasa Calpe. 2 vols.

III Exploiting Portuguese Reference Corpora: The CdP and the CRPC Corpora

Amália Mendes

The Reference Corpus of Contemporary Portuguese: Corpus Design and Case Study on Discourse Markers

1 Introduction

The compilation of the *Corpus de Referência do Português Contemporâneo (CRPC)* / Reference Corpus of Contemporary Portuguese began at the Centro de Linguística da Universidade de Lisboa, in 1988, under the coordination of Maria Fernanda Bacelar do Nascimento (Bacelar do Nascimento 2000; Bacelar do Nascimento *et al.* 2014). The objective was to apply a certain number of criteria that distinguish a reference corpus from other types of text compilations. First, a reference corpus is intended to be able to serve certain analysis purposes, as Sinclair (1996) mentions:

A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.

The goal was not to put together texts indiscriminately, but to carefully select texts that would be references for the elaboration of lexical and lexicographic materials, and other studies. The CRPC began with this objective in mind and included a diverse selection of literary texts by renowned authors, school manuals from different years and disciplines, didactic and scientific texts, among others. This option involved a long process of identifying, locating and preparing texts that were obtained in paper format.

The CRPC also intended to be representative at various levels: on the one hand, to cover the varieties of Portuguese in the world, although European Portuguese is necessarily the most represented variety; on the other hand, to include a great diversity of types of texts; and also to include written texts and oral recordings/transcriptions. To this end, the corpus was large (at least for the time) and today exceeds 300 million words. The type of material selected affects the way the corpus is made accessible: the written subcorpus is only available for context queries, without access to the full text, due to copyright limitations. The annota-

Acknowledgment: This work was developed with the support of Fundação para a Ciência e a Tecnologia, in the scope of the project UIDP/00214/2020.

tion and lemmatization of the corpus has enabled a wider set of query options, which are essential for many studies on the lexicon and on syntax.

Our goal is two-fold: first, we wish to present the written and spoken subparts of the CRPC and the query systems in use. Some of this information has already been made available, but the current access to the spoken subcorpus on TEITOK has not been previously discussed, nor is there a publication with a general overview of the spoken and written subcorpora in their current access. Second, we plan to use the CRPC for the study of the words that co-occur with the discourse marker (DM) *claro* ‘of course’ (literally: clear), specifically taking into account the identification of specific properties linked to the written and spoken modality and to the different text types.

In section 2, we will present the CRPC and discuss its internal composition: the written subpart in 2.1 and the spoken subpart in 2.2. The PoS annotation and lemmatization are discussed in 2.3, and the availability of the corpus in 2.4. Section 3 is devoted to the case study on the co-occurring words with the DM *claro*: the different values of the DM pointed out in the literature in 3.1, the significant co-occurring words found in the spoken and written subcorpora in 3.2, and an analysis of their functions in 3.3. We conclude in section 4.

2 The Reference Corpus of Contemporary Portuguese – CRPC

The constitution of a corpus is based on the principle that the data will be representative of a language or some aspect of the variation of a language (Leech 1991). This is also true for a reference corpus, but it is even more challenging for this type of corpora, as it is intended to be representative of the standard usage of the language in most of its diversity. One can address this concern by including a large array of text types and of verbal interactions, but still it is impossible to achieve a corpus that is fully representative. Another way to address the issue is to look for balance between text types. In fact, the number of texts and the proportions of the corpus sections will vary greatly and depend on the project’s goals but also on practical aspects, such as the duration of the project and the available funding.

Representativeness and balance are evaluated in terms of the total size of the corpus. A reference corpus will need to achieve a certain dimension to include the variation of the language. Also, studies show that lexical diversity is difficult to achieve: for instance, an analysis of a 1-million-word corpus showed that a new word appears approximately every 30 words (Kennedy 1998: 100), and some

meanings of highly polysemous words may be rare. Concomitantly, the presence of the internet has enabled the availability of large numbers of texts in digital format. These factors explain the ever-growing dimension of corpora, as the projects try to provide as much data as possible for linguistic analysis. One such tendency is to use texts that are available on the internet to create a linguistic corpus, and the approach known as “web as corpus” has gained importance (Kilgarriff & Grefenstette 2003). On the one hand, this implies that there is less control over the texts included in the corpus compared to reference corpora, so less balance, but, on the other hand, it certainly provides a high number of contexts for analysis. The contrast is illustrated by the British National Corpus, a corpus with 100 million words, designed and planned to be a reference corpus, and the Bank of English, with over half a billion words (McEnery & Hardie 2012) and continuing to be enlarged (the Bank of English is defined as a “monitor” corpus, meaning that it serves to monitor a language, its diversity and changes).

Before presenting in more detail the CRPC, let us consider other corpora of large dimensions for European Portuguese. The corpus CETEMPúblico (Linguateca) contains 190 million words extracted from editions of the daily newspaper Público between 1991–1998. The news articles having been subdivided into excerpts of a few sentences for copyright reasons (Rocha & Santos 2000). The corpus is annotated with several levels of information: PoS, verbal and nominal inflection, lemmatization and syntactic constituents with the software PALAVRAS (Bick 1999). The corpus can be queried online, and the IMS CQP system allows concordance queries that make use of the full set of annotation, as well as information on the distribution of the hits in the corpus. The fact that the texts come from a single source, and that copyrights have been secured, have made it possible to make it fully available upon request, making this corpus an extremely valuable resource, not only as a source of data for linguistic analysis, but also for the development of tools for the automatic processing of Portuguese. The unique source of the data is however a drawback, as the corpus lacks representativeness of more diverse text types.

Another example is the Corpus do Português, with 1.1 billion words, that includes written texts and oral transcriptions for European Portuguese and Brazilian Portuguese, from the 19th century to the XXth century (Davies 2014). The corpus was tagged with PoS and lemma information. The diversity of the corpus, and the consequent copyright issues, is also the reason why the corpus is only available for queries, using a platform that allows searches by century, with information on the Portuguese variety and the genre. An overview of corpora for Portuguese is presented in Vanderschueren & Mendes (2015), and Mendes (2016).

As mentioned in the introduction, the CRPC has now over 300 million words, more specifically 309 million written words and 1.6 million spoken words, repre-

senting the varieties of Portuguese, and compiled from 356,208 documents (Table 1). The corpus covers the twentieth century, although fiction texts from 1850 forward, from classic authors, are also included, as well as some texts from the beginning of the XXI century. The decades of the twentieth century are not comparable in size because the corpus includes mostly texts from the chronological period between 1970 to 2008 (due to the life time of the project). Special attention was given to the Portuguese variety of the text, to assure that only texts from native speakers of each variety of Portuguese, with no external linguistic influences, were included. The metadata include the country of edition of the written text, but most significantly for our purposes, it includes a field with the national variety of Portuguese that the author represents (the two fields may not be equivalent, for instance, when a Brazilian author is published in Portugal).

Table 1: Composition of the CRPC (written and spoken modality) per variety.

Composition of the CRPC	Words
Portugal	291,311,212
Angola	10,801,990
Brazil	3,562,947
Macau	2,093,538
Cape Verde	1,474,682
Mozambique	1,152,465
São Tomé and Príncipe	562,887
Guinea-Bissau	389,437
East-Timor	125,984
Goa	1,840
Total	311,476,982

2.1 Written Corpus

A large section of the written subcorpus of the CRPC was carefully planned to include texts that are considered a reference for the variety that they represent. This is especially the case for the European variety, that was the main focus of the corpus compilation. Fiction work from important Portuguese authors from the second half of the XIX century and from the XX century were included. Scientific and technical texts were selected from a diverse set of areas of knowledge, as well as schoolbooks for all courses and levels of primary and secondary education. Those texts were mostly compiled from paper books, and involved a time-consuming four-step process: sampling of sections from the beginning, intermedi-

ate and final parts of each text, digitalization with OCR, manual correction and final revision by a different team member. Concomitantly, all texts compiled from paper books were described with a detailed set of metadata. As the compilation progressed, the internet became an increasing source of data. Also, some texts were obtained directly in digital format from their owners, such as the parliamentary debates. The corpus design became closer to the notion of a monitor corpus, by including all texts that were made available, but that were still identified as reference texts for Portuguese. When working on a specific project that required a more balanced composition of the corpus, a subcorpus would be designed specifically for the task at hand; this led, for instance, to the design of smaller corpora for the identification of collocations (COMBINA-PT project),¹ and for the creation of a frequency lexicon.² When preparing the corpus to be made available online for queries, although we excluded some of the data, we decided to make as much of this material available as possible, although providing means to restrict the query to specific text types (cf. section 2.4).

The written subpart of the CRPC is organized in eight broad text types, that reflect the source of the data, the communicative setting, and internal properties of the texts (Table 2). The largest subpart is labelled “politics” and includes the published diaries of the National Parliament, plus additional documentation from the Parliament. These are mostly transcriptions of the Parliament sessions (that suffer some normalization and use written conventions such as commas and full stops to represent the discourse of the deputies). This is prepared and formal spoken data, transcribed according to the norms of written texts. Although labelled as written data in the CRPC, it shows marks of interaction between the deputies during the sessions. The second most frequent text type are newspaper articles, that cover different sections of a newspaper, and a large array of daily and weekly publications. The section Book is subdivided in Fiction, Scientific/Technical and Didactic (schoolbooks), and the section Magazines has a Scientific/Technical subset. Law includes juridical texts, such as rulings from the Supreme Court. There is a small subset of Correspondence and Brochures. Finally, any text that is not integrated in one of these subsets is included as *Varia*.

The set of documents retrieved from the internet had to be automatically cleaned of metatags and some non-relevant lexical content. We used the tool NCleaner that automatically segments the text into short textual units, mainly paragraphs. NCleaner requires the creation of an annotated corpus to learn to distinguish “relevant” from “not relevant” segments. We consider irrelevant all

1 <https://clul.ulisboa.pt/projeto/comбина-pt-combinatorias-lexicais-do-portugues>

2 <https://clul.ulisboa.pt/projeto/lexico-multifuncional-computorizado-do-portugues-contemporaneo>

Table 2: Text types distribution in the CRPC-written.

Text type	% of texts	Words
Politics (política)	52.70%	163,267,089
Newspaper (jornal)	35.67%	110,503,376
Book (livro)	6.63%	20,557,296
Magazine (revista)	2.47%	7,581,850
Varia	1.55%	4,806,176
Law (direito)	0.94%	2,927,953
Correspondence (correspondência)	0.02%	88,370
Brochure (folheto)	0.02%	80,833
Total	100%	309,812,943

information that is not part of the text, that do not represent a typical use of language within a collection of texts of a specific genre and on a defined subject, and distort the analysis of language that human experts, but especially NLP tools, could produce.

2.2 Spoken Corpus

The spoken subset of the CRPC was created within the scope of two projects that gave rise to specific and independent corpora for European Portuguese: the Fundamental Portuguese, that started in the 1970s, and the corpus C-ORAL-ROM. Fundamental Portuguese includes 1800 recordings of spontaneous conversations by speakers of different ages, levels of education and professions (Bacelar do Nascimento *et al.* 1987a,b), with 1400 of these recordings being transcribed (700,000 words). The transcripts were recently updated to XML format, aligned with the EXMARaLDA program and annotated with PoS information. This new version is freely available for research in the ELRA catalogue. The corpus C-ORAL-ROM was compiled more recently and constitutes a set of comparable spoken corpora for four Romance languages (Portuguese, Spanish, French and Italian) (Bacelar do Nascimento *et al.* 2005). The transcription of the Portuguese part contains 300,000 words, and has been revised according to the CHAT guidelines, text-to-sound aligned with the EXMARaLDA software (Schmidt 2012), and automatically lemmatized and annotated with PoS information. This updated version of the C-ORAL-ROM Corpus is available for research purposes in the ELRA catalogue. Other projects that provided spoken material to the CRPC include Português Falado - Spoken Portuguese (Bacelar do Nascimento 2001), with recordings and transcriptions of the varieties of Portuguese in the world.

Recently, the set of audio files and transcriptions have been revised to eliminate repetitions between projects and to identify the set of files with sound-to-text alignment. The result is the CRPC-Oral corpus, available for queries on the TEITOK platform.³ This corpus is composed of 490,903 tokens, of which 88% represent European Portuguese. The composition of the corpus in terms of register and communication setting is presented in Table 3. Informal register include conversations between a group of people, dialogues, and some monologues. The corpus includes telephone conversations between friends and acquaintances, recorded without prior knowledge (informed consent was provided but no specific date of recording was given to the participants). The remaining data is of formal register. It includes broadcasted programs on radio and television, on different topics. Finally different situational contexts of formal registers are included, such as teaching, political debate, and preaching.

Table 3: Composition of the CRPC-oral on TEITOK.

Recording situation	Number of tokens	% of tokens
Informal, private		
Face to face (conversation, dialogues, monologues)	314,946	64.16
Telephone	24,479	4.99
Formal		
Broadcasting/Talk Show	20,679	4.21
Broadcasting/Interview	18,413	3.75
Broadcasting/Reportage	10,984	2.24
Broadcasting/Scientific Press	10,021	2.04
Broadcasting/Sport	5,917	1.21
Broadcasting/News	3,984	0.81
Broadcasting/Meteorology	2,094	0.43
Business	10,487	2.14
Conference	10,846	2.21
Teaching	9,882	2.01
Political Speech	9,149	1.86
Political Debate	8,992	1.83
Law	6,682	1.36
Preaching	6,591	1.34
Professional Explanation	6,588	1.34
Unspecified	10,169	2.07
Total	490,903	100

³ teitok.clul.ul.pt/crpcoral

2.3 Annotation

The written subset of the CRPC is tokenized, i.e. it is segmented into tokens or linguistically significant units, using an adapted version of the LX-tokenizer (Branco and Silva 2003) specifically prepared to deal with properties of the Portuguese language. The handling of clitics by the LX-tokenizer was kept as such in the tokenization of the CRPC, i.e., clitics are separated from the verb form and tagged independently. However, the contractions of prepositions and articles was adapted to our purposes: instead of separating contractions into two units (e.g., *pelo* in two tokens “por_ o”), we keep them as a graphic unit (*pelo*). The reason for this decision is that the CRPC is mostly used by linguists that seek lexical items or constructions and want to be able to retrieve examples for their studies that have little manipulation. However, this tokenization decision has implications for POS tagging, as we will discuss below.

For POS-tagging we used MBT (Daelemans *et al.* 1996), a memory-based tagger. The written CINTIL corpus (annotated for POS and manually verified) was used for training and evaluation. The tagset uses a large set of categories, that enable to make fine grained distinctions. For instance, past participles are divided into the tag PPT, for participles found in compound tenses, and PPA, for the remaining cases. And auxiliary verbs are divided into 3 subtags: *_VAUX* if the auxiliary verb is in a finite tense (e.g., *tinha feito*), *_INFAUX* if the auxiliary verb is in the infinitive form (*ter feito*), *_GERAUX* if the auxiliary verb is in the gerund form (*tendo feito*). The CRPC is also annotated with inflection tags that cover, for non-verbal categories, gender, number, person, and degree, and for verbal categories, tense, mood, person, number and, in the case of the PPA tag, gender.

For lemmatization, we used MBLEM, that combines a dictionary lookup with a machine learning algorithm to produce lemmas. The classifier is a memory-based learning algorithm (Daelemans and van den Bosch 2005). As basis for the dictionary we used a list of wordform - POS-tag combinations mapped to lemmas. This list was produced in-house. The dictionary used in MBLEM contains 102,196 word forms combined with 27,860 lemmas, leading to 120,768 word-form-lemma combinations (for more details on the PoS annotation and lemmatization of the CRPC, cf. Génèreux *et al.* 2012). A list of the tagset used for tagging the CRPC is provided in the corpus user manual (Mendes *et al.* 2023).

2.4 Availability

The CRPC is available for queries in two platforms: the written subpart uses the CQPweb interface (Hardie 2012),⁴ running over the IMS CQP - Corpus Query Processor (Evert & Hardie 2011), while the spoken subpart runs on TEITOK⁵ (Janssen 2016). The spoken subcorpora are also freely available for academic purposes through the ELRA catalogue and both spoken and written subparts are found in the repository of the PORTULAN CLARIN infrastructure.

CQPweb provides a robust system for large corpora, making full use of query options and providing additional features. Queries can be limited to specific text types and/or varieties of Portuguese by using the “restricted query” option. The node of a query can be a word, a part of a word, a continuous or discontinuous sequence of words, and regular expressions. All of them can be combined with the annotation layers: PoS category, inflection tags and lemma. Queries can be made by using the “simple query” mode, that is more intuitive and doesn’t require a formal structure of the query, or by using the “CQP syntax” mode. The result of a query are concordances, i.e., the list of the contexts of the query node in the corpus, presented in Key Word In Context (KWIC) format, that highlights the node. Concordances can be sorted, and then downloaded to be later imported to an excel environment, for instance. The user can obtain the metadata for each context. The CQPweb allows users to restrict the number of hits by applying the option “Thin”, either specifying a percentage of the results or a specific number of hits. This is especially useful to obtain a manageable number of results while keeping a random selection throughout the corpus. Other options are the possibility to obtain the distribution of the query node in the corpus, in this case, the option provides the distribution per variety and per text type. Concordances are one type of information that can be obtained in CQPweb. Another is frequency information, either indirectly (provided when looking for concordances) or directly, when choosing the option “word lookup”: specifying the query as word-forms starting with “livr*” will provide a list of all word-forms starting with these characters found in the corpus (with a link to their concordances hits), and the number of their occurrences. The option “frequency list” will return the full list of word-forms in the corpus, sorted by the number of occurrences. Selecting two different frequency lists in the option “Keywords” will return a list of word-forms that show different frequency patterns in the two corpora, or a list of word-forms that occur in one but not the other corpus. Finally, another option is to obtain the

⁴ gamma.clul.ul.pt/CQPweb.

⁵ teitok.clul.ul.pt/crpcoral.

collocations of a query node. Several statistical measures are available, such as Mutual Information or T-score. Users need to register to use the full CRPC, and the advantage is that the query history of the user is saved, and that specific queries may be saved by the user.

Figure 1: Homepage of the CRPC-written on CQPweb.

The homepage of the CRPC corpus on CQPweb is illustrated in Figure 1. The left menu provides access to the query options (Standard query, Restricted query, Word Lookup, Frequency lists, Keywords and Analyze corpus); to the query history of the user; to Corpus Info (View corpus metadata provides the list of the metadata fields, and Corpus documentation opens a new page where the manual of the CRPC can be accessed). Figure 1 provides an image of a Simple query of word forms starting with “livr”, restricted to European Portuguese (Portugal) and to fiction books (*livro_literário*).

The first lines of the concordances results of the query in Figure 1 are provided in Figure 2. The query retrieved hits for the word forms “livre, livralhada, livres, livros”, among others. The top window provides additional options after obtaining the concordances, such as Thin, Distribution, Sort, Download, Save current set of hits (save query).

Figure 3 shows the collocations of the concordances of the lemma “livro” in fiction books of European Portuguese, processed with the statistical measure Mutual information. The settings of the concordance window mean that each collo-

Your query "livr*", restricted to texts meeting criteria "Fonte: livro_literario; Pais do autor: Portugal", returned 4,142 matches in 359 different texts (in 8,056,234 words [417 texts]; frequency: 514,14 instances per million words)

Line View | Show Page: 1 | Show in random order | Page 1 / 83

No	Filename	Text
1	L0111	à cara os punhos fechados . Mas , o tio Carlos , livre delas , muito esquecido , ia -se embo
2	L0111	visitou e lhe disse : Há maneira de Vossa Excelência se ver livre do Bravo . Diga ! Vossa Excelência c
3	L0111	em vez de o enfraquecer , fazia -o homem . O ar livre , respirado de costas direitas e coraçõ
4	L0308	professor , pois sabia de coisas de roubos muito mais que a livralhada . Que ainda havia de fazer uma lei no
5	L0091	gente bastante para se opor à nossa passagem . O caminho continuava livre diante das armas cristãs . Em vez de marchar em direcção a
6	L0091	organizar um corpo de tropas que cortasse as comunicações , até af livres , do Infante com as praças da fronteira e que , na
7	L0524	género de vida que levava em Coimbra , a influência de certos livros aliás nem sempre bem entendidos , a de vários companheiros , a

Thin... | New query | Frequency breakdown | Distribution | Sort | Collocations... | Download... | Categories... | Save current set of hits...

Figure 2: Examples of concordances on CQPweb and further options.

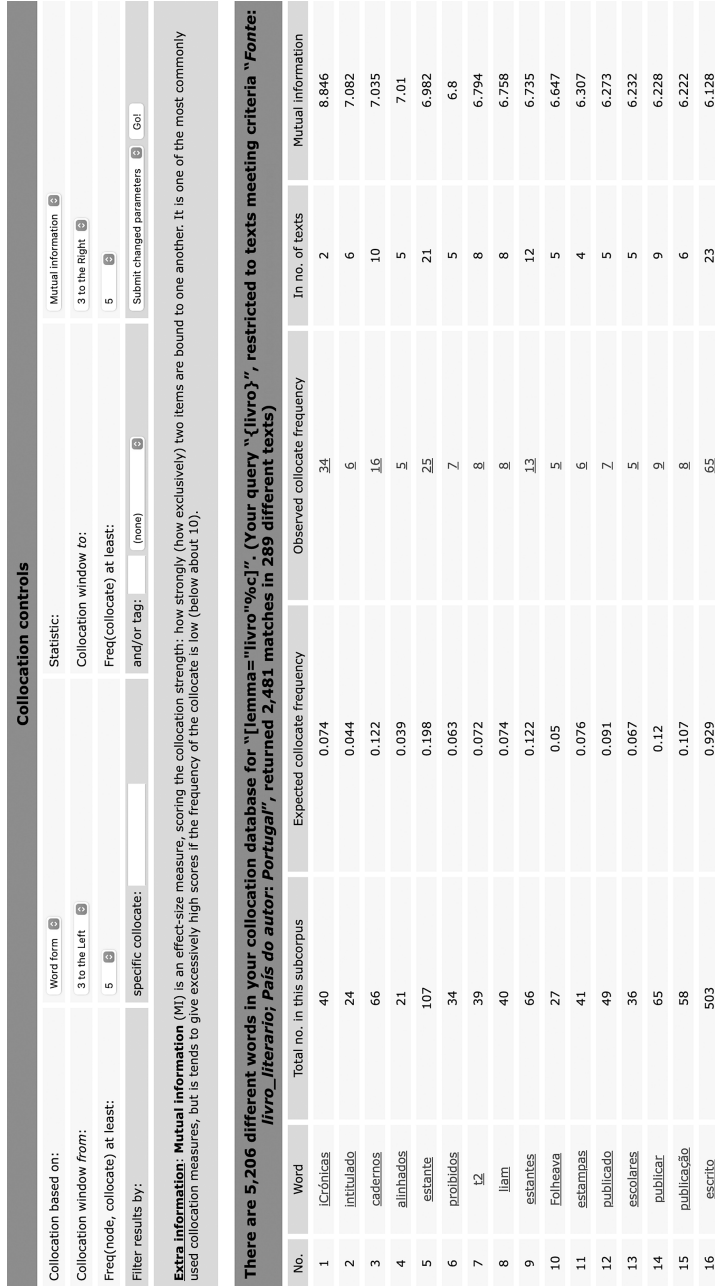


Figure 3: Collocations of *livro* in a subset of the CRPC on CQPweb.

cate listed may occur up to 3 positions to the left and to the right of the lemma “livro”.

The spoken corpora *Português Fundamental* (Fundamental Portuguese) and *C-ORAL-ROM* have been converted to the EXMARaLDA software (Schmidt 2012) and text-to-sound aligned. They can be obtained and queried locally through the Exakt module of EXMARaLDA. Exakt allows concordances to be extracted, and metadata to be used to restrict the query. Each line of the concordance hits is linked to the audio file and can be listened to. Alternatively, the full set of the transcribed and aligned spoken data may be queried online, using the TEITOK platform, that allows to visualize the transcription, make queries and listen to the contexts (Janssen 2016). TEITOK includes a query builder function to create a query without making use of the CQP query format, as illustrated in Figure 4.

Figure 4 shows the options of the query builder. In this case, the tag builder was used to select the category “main verb” (the system will automatically translate to the correct PoS tag). In Document Search, the country and the text type (Channel) have been used to restrict the results: concordances hits will provide an overview of the main verbs used in preaching settings in European Portuguese. The frequency distribution of the hits per PoS tag can be automatically retrieved and gives an overview of the use of, for instance, subjunctive, imperative, and the use of second person verb forms.

When accessing the full transcription, the view options include the “Audio” setting, that enables the user to hear the speech turn, as illustrated in Figure 5. In the “Transcription” view option, all the tokens produced by the speaker are visible. Some are marked with strikethrough, meaning that they are repetitions or hesitations. In the “Dialogue Form” view, these are not visible. Figure 5 shows the Transcription View on top, and below the Dialogue Form view of the same 3 lines: the repetitions in the second line (“um pro um”) are eliminated in the Dialogue Form view.

This presentation of the CRPC corpus and its design and compilation show how it is the result of a 20 years process (more than 30 if we include making it available in its current platforms). This period shows dramatic changes with the advent of the internet, and a new approach to corpus compilation. The CRPC is a testimony of this change and how it affects corpus design and compilation. The initial goal of creating a reference corpus (representative and balanced) was slightly adapted to include large batches of texts that were made available through official sources. This affected the balance of the corpus and stirred it in the direction of a monitor corpus. The initial design work and the diversity of texts included in the corpus still account for the representativity that was intended, at least for European Portuguese and the XXth century.

Corpus Search

Search query_builder | visualize | options

SQL Query:

Tag Builder: COPLE2 tags

Main POS:

tense

person

number

gender

infinitive definiteness

imperative type

Query Builder

Text Search

Dialogue form	<input type="text" value="matches"/> <input type="button" value="x"/>
Regularized form	<input type="text" value="matches"/> <input type="button" value="x"/>
COPLE2 tags	<input type="text" value="tag_builder"/>
CINTIL tags	<input type="text" value="matches"/> <input type="button" value="x"/>
Lemma	<input type="text" value="matches"/> <input type="button" value="x"/>

Document Search

Year	<input type="text" value=""/>
Country	<input type="text" value="Portugal"/> <input type="button" value="x"/>
Project	<input type="text" value="[select]"/> <input type="button" value="x"/>
Channel	<input type="text" value="Formal/Preaching"/> <input type="button" value="x"/>
Title	<input type="text" value=""/>

CRPC-ORAL

Main Menu

- Home
- Search
- Team
- Tagset
- Login

Powered by <TEI>:ITOK<
Maarten Janssen, 2014-

R&D Unit funded by
FACT
Funded by the European Union
View project page

Figure 4: Query builder on TEITOK.


CRPC-ORAL

Main Menu

- Home
- Search
- Team
- Tagset
- Login

Powered by <TEI:TOK>
Maarten Janssen, 2014-

R&D Unit funded by



pfammn19

pfammn19.txt

Title pfammn19.txt

Project C-ORAL-ROM

Country Portugal

City Lisboa

Date 20-11-2001

View options

Text: [Transcription](#) | [Dialogue form](#) | - Show: [Colors](#) | [Audio](#) | - Tags: [POS tag](#) | [CINTIL pos](#) | [Lemma](#)

▶ |
-12:54 >>

LSC - [🔊](#) agora / o que eu tenho impressão / é que / o instrutor também / hhh / ch +

LSC - [🔊](#) enfim / imagina ~~um-pre~~ um programa / muito / leve //

LSC - [🔊](#) um / dado / dizer-lhe que tinha problemas de coluna / e mais isto / e mais aquilo / e mais o outro //

View options

Text: [Transcription](#) | [Dialogue form](#) | - Show: [Colors](#) | [Audio](#) | - Tags: [POS tag](#) | [CINTIL pos](#) | [Lemma](#)

▶ |
-12:54 >>

LSC - [🔊](#) agora / o que eu tenho impressão / é que / o instrutor também / hhh / ch +

LSC - [🔊](#) enfim / imagina um programa / muito / leve //

LSC - [🔊](#) um / dado / dizer-lhe que tinha problemas de coluna / e mais isto / e mais aquilo / e mais o outro //

Figure 5: View options of the CRPC-spoken on TEITOK.

3 Discourse Studies: a Case Study of *claro* Using the CRPC

Discourse studies focus either on the analysis of full texts (written or transcriptions of spoken language), or on the analysis of concordances of specific words or structures. The written subpart of the CRPC gives access to concordances (and to a larger context of the hit word) but not to full texts, due to copyright issues. The spoken subpart, free from copyright issues, enables such an integral approach to texts. One topic in discourse studies is the study of discourse markers, i.e. units with propositional, structural and modal/interpersonal functions in discourse (see, among others, Pons Bordería 1998; Halliday & Matthiessen 2004; Cuenca & Marín 2009; Cribble & Degand 2019). An overview of different perspectives on the study of DMs in Portuguese is presented in Mendes & Lejeune (forth.), and a compilation of recent studies on Portuguese DMs in a contrastive approach with other languages is pre-

sented in Duarte & Ponce de León (2020). Discourse markers can perform functions at the propositional level, linking two propositions together and specifying the semantic nature of this connection, such as a causal, contrastive or conditional value. On the other hand, they can perform functions at the level of textual organization, signaling that between two textual segments there is a relation of change of topic or of summarization, for instance. Finally, discourse markers can have pragmatic functions, associated with interpersonal values and modal values, related to the interaction between speakers. A single DM may be polyfunctional, as it may perform several functions, in different contexts (Crible & Degand 2019).

The CRPC has been widely used for studies on Portuguese discourse markers, as it provides both spoken and written modalities and a diversity of text types. In this section, we plan to analyze one specific discourse marker, *claro*, and to observe its use in the spoken and in the written subparts of the CRPC in terms of the words that tend to co-occur with it in the corpus. This DM has been previously studied for Portuguese in Lopes (2021), and also in a contrastive perspective with French in Mendes & Lejeune (2023). The Catalan and Spanish equivalents of *claro* have been the topic of several studies (Fuentes Rodríguez 1993; Cuenca & Marín 2012; Pons Bordería 2003, 2012). Our goal is to explore a specific functionality of concordancers, namely the possibility of extracting information on patterns of co-occurrence of *claro* with other tokens. We will explore collocations of *claro*, i.e., co-occurrences between two or more words that tend to be more frequent than expected based on the frequency of each element in a corpus (Sinclair 1991; Mendes & Antunes 2016). The study of the collocates of a node word in a corpus provides information on the meaning of the word in context and on the different uses of the word in different text types, as significantly expressed by Firth: “You shall know a word by the company it keeps” (Firth 1957:11). The study of corpus data revealed a tendency for words to co-occur, even when the meaning of the sequence was still compositional, without idiomatic interpretation, and drew the attention to the fact that language was composed of “prefabricated chunks” that were still semantically and syntactically analysable (Sinclair 1991). Our goal is to explore the possibility that most concordancers provide nowadays of automatically extracting information on collocations of a node word and to examine how this information may be valuable to identify patterns of use and functions of the DM *claro* in different contexts.

We will first present in 3.1 a summary of studies of functional equivalents of *claro* in Spanish and Catalan, and also proposals addressing the Portuguese *claro*. We will then present the collocations of *claro* in the spoken subpart of the CRPC in 3.2, and analyze their functions in 3.3.

3.1 Previous Studies on *claro* and on its Functional Equivalents in Spanish and Catalan

The Spanish discourse marker *claro*, and the Catalan (*és*) *clar*, have been described as having a modal function of signalling agreement, while denoting that the propositional content is considered shared knowledge (Pons 2012; Cuenca 2013). These markers have also been described as having a value of contrast or concession, and a structural function, marking the beginning of a speech turn and signalling that it is aligned with the previous intervention of the interlocutor (Cuenca 2013). The Portuguese DM *claro* was analyzed in Lopes (2021), which distinguishes uses that constitute dialogues from other uses. In uses in dialogues, *claro* is described as configuring an emphatic affirmative response to speech acts of question or request, in which the use of *claro* is different from the use of *sim* ‘yes’, as is also the case in Castilian and Catalan (Pons 2012; Cuenca 2013), because *claro* denotes that the propositional content over which it has scope is shared knowledge. In dialogues, Lopes (2021) also point to a value of agreement, and a value that signals an attitude of cooperative attention in the interaction. In the remaining uses, an epistemic modalizing value and a concessive value are also encountered. Another recent study of *claro* presents a contrastive approach with functional equivalents in French (Mendes & Lejeune 2023) and also proceeds by distinguishing the use of *claro* as a response in contexts of dialogue from other contexts. The authors point out that in most contexts where *claro* is used, a possible equivalent in French is *bien sûr*, and that the adjective *clair*, ethymologically close to the Portuguese word form *claro*, is only possible in predicative contexts (*c’est clair* ‘it is clear’, *il est clair que* ‘it is clear that’). In contexts of dialogue, *claro* expresses agreement, eliminates possible alternatives and denotes that the information provided by its interlocutor is shared knowledge. In certain contexts, it may not denote that the content is shared knowledge but rather factual, inevitable under the circumstances (and possible French functional equivalents in comparable corpora are *évidemment* ‘obviously’, *de fait* ‘indeed’, *c’est un fait* ‘it’s a fact’) (1). Another value is a negative evaluation, and exclusion, of another stance in the discourse (2).

- (1) *PED: [<] <e> depois <a polícia foi-se embora durante a noite> //
 *SAN: [<] <hhh>\$
 *NUN: <hhh> //
 *AMA: [<] <claro> // (CRPC-Oral) (Mendes & Lejeune, forth., ex (29))
 ‘- PED: and then the police went away during the night - SAN: hh - NUN: hhh
 - AMA: of course’ (our translation)

- (2) Exprime as suas exigências pela boca do presidente da CIP, advoga a tranquilidade dos espíritos e a ordem nas ruas e nas empresas, a « sua tranquilidade » e a « sua ordem », claro está. (CRPC-escrito) (Mendes & Lejeune, forth., ex (26))
 ‘He expresses his demands through the mouth of the president of the CIP, he advocates the tranquility of spirits and order in the streets and in companies, “his tranquility” and “his order”, of course.’ (our translation)

In monologic contexts, *claro* may qualify a previous utterance, explicit or virtual, as inevitable (determined by circumstances), involving shared knowledge between speaker, addressee and the discourse community of the speaker. The authors conclude that *claro* is used in contexts of dialogue or contexts of interdiscursive dialogism.

While in dialogue contexts, *claro* may evaluate negatively another stance, in monologic contexts, it may establish the factuality of p and be followed by a relation of opposition, explicitly marked or left implicit (3).

- (3) Os actos racistas e xenófobos são completamente inaceitáveis na nossa Comunidade, seja qual for o local em que ocorram. São contrários aos princípios que estão na base da fundação da União Europeia, como o disse ontem o presidente Havel: os princípios da liberdade, da democracia e do respeito pelos direitos humanos. Claro que nestes últimos anos se fizeram progressos, mas temos de continuar a esforçar-nos juntos por criar um clima de tolerância (. . .) (Europarl 16487) (Mendes & Lejeune, forth., ex (38a))
 ‘Racist and xenophobic acts are completely unacceptable in our Community, wherever they occur. They are contrary to the principles that underlie the founding of the European Union, as President Havel said yesterday: the principles of freedom, democracy and respect for human rights. Of course, progress has been made in recent years, but we must continue to work together to create a climate of tolerance’ (our translation)

3.2 Collocations of *claro* in the CRPC

After presenting a review of previous analyses of the values of *claro*, our objective in the next sections is to observe how the collocations of *claro* may provide a contribution to the analysis of this DM. We will first observe collocations of *claro* in the spoken subpart of the CRPC. For our analysis, we restricted the corpus to the CORAL-ROM corpus, the largest and more recent part of the spoken data. A search of

the query node *claro* in the C-ORAL-ROM corpus on TEITOK retrieves 739 hits. When analyzing each context, we find out that only a few cases are in fact the adjectival word form *claro* (for instance, a single instance of *claro* modifying a noun, *um recado claro* ‘a clear message’, and 8 occurrences in predicative structures with the verbs *ser*, *estar*, *ficar*, *tornar* (e.g., *acho que ficou muito claro* ‘I think it’s very clear’).

The remaining contexts correspond to the discourse marker usage of *claro*, although in different syntactic patterns, revealed by sorting the concordances by the left and the right context of the node. First of all, *claro* as the sole answer to a question is only found in 19 hits, 12 of them in tag-questions with *não é* ‘right’.

The option of identifying collocates of a query node is not available on the TEITOK, the platform used for the spoken subpart of the CRPC. However, the different spoken subcorpora that are included in the CRPC may be queried locally with available concordancers. For this analysis, we used the software AntConc (Anthony 2019) over a non-annotated version of the C-ORAL-ROM corpus. AntConc provides two options to analyze the patterns of co-occurrence of a query node: the option “clusters (n-grams)” retrieves sequences of 2 or more words (2-grams, 3-grams, etc.) that include the query node, and sorts the results per frequency or per probability; the option “collocates” automatically extracts words that co-occur with the query node using several options for the window size, the frequency and the statistical measures. While in TEITOK the name of the speaker is not included as a token, in AntConc this is not the case, and one has to filter manually the collocates that are abbreviations of speakers. We left out of the discussion contexts where *claro* occurs with verbs *ser*, *ficar* and *estar*. Table 4 provides information

Table 4: Significant 2-grams including the DM *claro* in the C-ORAL-ROM.

freq.	search term left	freq.	search term right
48	claro que	19	pois claro
6	claro / que	15	pois // claro
2	claro // que	10	pois / claro
18	claro // claro	6	e claro
6	claro / claro	2	e / claro
5	claro // e	5	sim / claro
2	claro / e	3	sim // claro
5	claro // mas	1	sim claro
3	claro / depois	3	mas claro
2	claro // eu	1	mas // claro
1	claro / pois	1	agora / claro
1	claro // exacto	1	claro claro
1	claro // evidentemente	1	depois claro
		1	depois / claro

on 2-grams where the search term is on the left and on the right window. Short pauses are marked with “/” and long pauses with “//”.

For the extraction of collocations, we established the limit of 3 words on the left and 3 words on the right of the query node (Table 5). The collocates were sorted by frequency and also by applying the statistical measure log-likelihood (other measures are available on AntConc). Again, due to the fact that speakers’ names are treated as tokens, some collocates were excluded from our analysis. The 5 columns of Table 5 provide the following information: the collocate, the total frequency of the collocation; frequency with collocate on the left of the node; frequency with collocate on the right of the node; statistical measure. For instance, in row 2, the collocate *pois* occurs 58 times on the left of *claro* (*pois claro*) and 10 times on the right of *claro* (*claro / pois*). By clicking on the collocate, in AntConc, the user visualizes the concordances of the collocation, i.e., all the contexts where the node and the collocate occur together in the corpus. This is extremely useful to quickly grasp the type of context of the collocation and exclude cases where there is no relation between both units.

Table 5: Most significant collocates of the DM *claro* in the CRPC-spoken.

collocate	total freq	freq (collocate left)	freq (collocate right)	statistical measure (log-likelihood)
claro	82	41	41	710.53088
pois	68	58	10	382.84855
que	101	13	88	238.39841
e	89	24	65	214.69497
mas	34	8	26	112.97331
depois	18	6	12	62.21670
sim	17	13	4	58.40884
porque	15	4	11	38.66077
eu	16	4	12	24.61289

The CQPweb software, used for the queries of the written subpart of the CRPC, enables the automatic extraction of collocations. We restricted our search to European Portuguese, by selecting the corpus “Portugal” on CQPweb.⁶ As the corpus was automatically annotated for POS, we searched the item *claro* tagged with the PoS category Adverb. Although the results still include adjectival uses of *claro*, due to errors in the automatic annotation, the use of the PoS category is useful to filter unwanted contexts. The query returned 20,180 hits. We extracted collocations with CQPweb and selected Log-likelihood as the statistical measure, and a

⁶ gamma.clul.ul.pt/CQPweb/portugal

window of 3 left and right, and minimum frequency of 3. CQPweb includes all the tokens of the written corpus when retrieving collocates, and tokens include punctuation marks. The relevant collocates for the DM use of *claro* were selected from the top 50 collocates, and are listed in Table 6.

Table 6: Most significant collocates of the DM *claro* in the CRPC-written restricted to Portugal.

collocate	freq of word in corpus	expected collocate freq	observed collocate freq	Log- likelihood value
que	7437019	3199.164	9017	7339.414
!	539436	232.048	2098	5541.429
pois	139837	60.153	439	989.603
?	416056	178.974	621	663.152
sim	61212	26.331	219	543.388
mas	782660	336.675	739	358.835
claro	45914	19.751	131	273.577

A quick comparison of both lists shows that punctuation marks are found as collocates in the written corpus, but not the spoken one. The remaining collocates in the written corpus are also found in the spoken corpus. There are some collocates that achieve high statistical measure in the spoken corpus but not in the written corpus: the conjunction *e* has negative statistical values (-50.728) and *depois* receives a value close to zero (0.224), probably due to the high overall frequency of the collocate. The other two collocates have low statistical values: *porque* (36.463) and *eu* (28.87).

3.3 Functions of the Collocations

The collocations in the written and spoken subcorpora point to modal and structural domains (Cuenca & Marín 2009) in which *claro* occurs: a modal domain with a function of marking agreement; a structural/modal domain with a function of interaction and text structuring; a structural/modal domain with a function of interaction and contrast. We discuss these values below.

3.3.1 Modal Domain

The repetition of *claro*, in the written and spoken subsets, with or without a weak or strong pause between the two elements, brings a reinforcement of the value of agreement and of shared knowledge, as exemplified in (4).

- (4) SRA - é mesmo um dar mais aos outros do que só para receber
 NUN - claro claro (pfammn08.txt)
 ‘SRA - it’s really more about giving to others than receiving. NUN – of course of course’

The reinforcement of the agreement is also found with the collocates *exacto* and *evidentemente* (*claro // exacto*; *claro // evidentemente*) on the right position of the node, with a strong pause between the two (the fact that there is only one context of each collocate suggests that the spoken corpus is relatively small for the study of collocations). In these two cases, the two collocates reinforce the value of shared knowledge and inevitability of *claro*.

The collocates *pois* and *sim* also occur with *claro* in contexts of response, where the speaker agrees with the content of the previous speech turn. Both collocates are placed in left position of the node in the written and spoken modalities. When there is a weak or strong pause between the two elements (*pois / claro*; *sim / claro*), the collocate signals agreement and *claro* reinforces the agreement value and adds that the content was already known by the speaker (5). The collocate *pois* differs from *sim* in two aspects: first, it may just signal attention of the speaker to the speech turn of his addressee, rather than agreement, and the collocation would suggest a rising level of agreement; second, contrary to *sim*, the DM *pois* occurs more frequently in immediate adjacency to the node *claro*, without any pause between the two units, as in (6).

- (5) MAR - e tem estado / muito / há coisa de quase um ano / seis meses / oito meses / que está / bastante fraco //
 ZEB - isso normalmente / essas coisas têm altos e baixos / não é ?
 MAR - sim / claro // (C-ORAL-ROM, pfamdl16.txt)
 ‘MAR – and he has been / very / about one year or so / six months / eight months / that he is / quite weak ZEB - usually / those things have highs and lows, right? MAR – yes / of course’
- (6) PBF - então / e eu estava contigo nessa altura //
 BAP - pois sim senhor
 PBF - depois foi cada um para seu lado /

BAP - pois claro // (C-ORAL-ROM, pmedin03.txt)

‘PBF – then / and I was with you at the time BAP – yes indeed PBF – then each one went his one way BAP – indeed’

In the written subcorpus, the collocate *pois* is also mostly found in left adjacency to *claro* without any punctuation between node and collocate. Figure 6 shows the distribution of the collocate per text type in the corpus: *pois* is more frequent (per million words) in subsets of the corpus that either transcribe dialogues (e.g., the transcripts of the Portuguese Parliament sessions (*politica*) and newspapers (*jornal*)) or include fictional representations of dialogues as in fiction works (*livro_literario*). On the contrary, values are low in technical/scientific books (*livro_tecnico*) and technical/scientific journals (*revista_tecnica*). The collocate *sim* occurs in the left or in the right window (*sim, claro / claro que sim*) and is also more frequent (per million words) in subsets of the written corpus that include the representation of a dialogue (see Figure 7). However, the collocate *sim* has a lower frequency per million words in the transcripts of the sessions of the Portuguese parliaments than the DM *pois*. This might be related to the value of negative evaluation of *pois claro* over the utterance of the previous speaker (7) – a property frequently found in contexts of antagonist response, frequent in the transcripts of the Parliament sessions –, while *claro que sim* or *sim, claro* are frequently used in monologic contexts where the speakers asks a question and answers it himself (8) (Mendes *et al.* 2020).

- (7) A aplicação de um regime de transição para a tributação de rendimentos resultantes da actividade pecuária intensiva em IRS (considerando 40 %, 60% e 80% desses rendimentos, respectivamente em 1989, 1990 e 1991) e em IRC (taxando em 20%, 25% e, 31%, respectivamente, 1989, 1990 e 1991), é, pois, uma medida que o Grupo Parlamentar do PSD vê com agrado.

O Sr. José Magalhães (PCP): - Pois, claro! (A129504, sessions of the Parliament)
 ‘The application of a transition regime for the taxation of income resulting from intensive livestock activity in IRS (considering 40%, 60% and 80% of these income, respectively in 1989, 1990 and 1991) and in IRC (taxing at 20%, 25 % and 31%, respectively, 1989, 1990 and 1991), is, therefore, a measure that the PSD Parliamentary Group welcomes.

Mr. José Magalhães (PCP): - Yes, surely!’

- (8) As juventudes partidárias fazem sentido? Claro que sim. (J39155, newspaper)
 ‘Do political party youth associations make sense? Sure they do.’

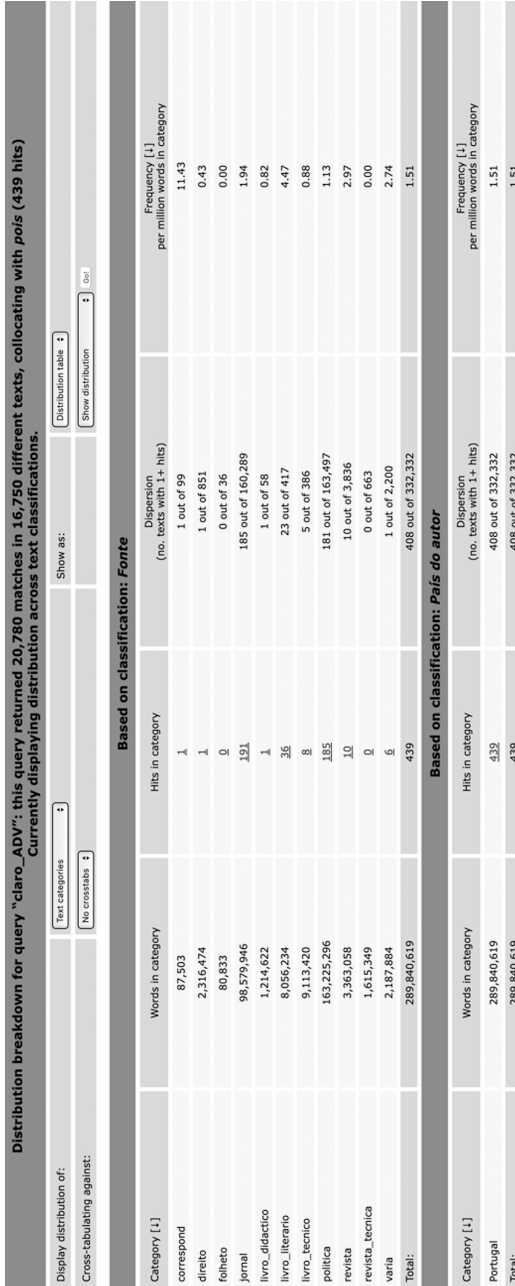


Figure 6: Distribution of the collocate *pois* with *claro* in the corpus CRPC-written restricted to Portugal.

Distribution breakdown for query "claro_ADV": this query returned 20,780 matches in 16,750 different texts, collocating with *sim* (219 hits)
 Currently displaying distribution across text classifications.

Display distribution of: Show as:

Cross-tabulating against:

Based on classification: Fonte

Category [1]	Words in category	Hits in category	Dispersion (no. texts with 1+ hits)	Frequency [1] per million words in category
correspond	87,503	0	0 out of 99	0.00
direito	2,316,474	0	0 out of 851	0.00
folheto	80,833	0	0 out of 36	0.00
jornal	98,579,946	121	112 out of 160,289	1.23
livro_didactico	1,214,622	1	1 out of 58	0.82
livro_literario	8,056,234	23	19 out of 417	2.85
livro_tecnico	9,113,420	3	2 out of 386	0.33
politica	163,225,296	55	55 out of 163,497	0.34
revista	3,363,058	10	10 out of 3,836	2.97
revista_tecnica	1,615,349	3	3 out of 663	1.86
varia	2,187,884	2	2 out of 2,200	0.91
Total:	289,840,619	219	204 out of 332,332	0.76

Based on classification: Pais do autor

Category [1]	Words in category	Hits in category	Dispersion (no. texts with 1+ hits)	Frequency [1] per million words in category
Portugal	289,840,619	219	204 out of 332,332	0.76
Total:	289,840,619	219	204 out of 332,332	0.76

Figure 7: Distribution of the collocate *sim* with *claro* in the corpus CRPC-written restricted to Portugal.

3.3.2 Structural/Modal Domain

A significant collocate of *claro* in the spoken corpus is the causal conjunction *porque*: *claro* signals agreement and shared knowledge and *porque* adds information on the reasons that lead to the situation presented in the speech of the addressee. In that sense, although the speaker agrees with the addressee, *porque* introduces a textual segment that ensures textual progression and identifies the speaker as holding more knowledge than the addressee.

- (9) GRA – é dos tais / que não se altera / porque é realmente bem educado / mas que chega a qualquer sítio / e ao fim de cinco minutos / está a falar sobre / a guerra de África // e até se ir embora / fala sobre a guerra de África // eu só tenho um termo em / francês / para definir um tipo destes // é um emmerdeur (. . .) quer dizer / é um emmerdeur

LGM – claro / porque não resolve a vida dele e nem a das pessoas também (C-ORAL-ROM, pfamcv03.txt)

‘GRA – he’s one of those / who keeps his head / because he’s really well educated / but he arrives anywhere / and after five minutes / he’s talking about / the war in Africa // and until he leaves / he talks about the war in Africa // I only have one term in / French / to define a type of this // it’s an emmerdeur (. . .) I mean / it’s an emmerdeur

LGM – of course / because he doesn’t solve his life or anyone else’s either’

In the written corpus, we find few cases of the collocation *claro, porque* as a response, and they involve a negative evaluation of the content of the speech of the addressee (the context preceding *claro*), as in (10). In monologic contexts, the speaker uses the collocation to comment on his own speech as in (11):

- (10) O Sr. Carlos Encarnação (PSD): - Sr. Presidente, Sr. Deputado José Magalhães, como tenho saudades de si na oposição!

Risos do PSD.

O Sr. José Magalhães (PS): - Claro, porque estava no Governo!

Risos gerais. (noCOD_1001370 – politica)

‘Mr. Carlos Encarnação (PSD): - Mr. President, Mr. Deputy José Magalhães, how I miss you in the opposition!

Laughter from the PSD.

Mr. José Magalhães (PS): - Of course, because you were then in the Government! General laughter.’

- (11) Eu respondi-lhe que o racismo sempre existira só que não era tão evidente nem tão incómodo. Para os brancos, claro, porque os negros nunca viveram bem com isso. (J77816, newspaper)
 ‘I told him that racism had always existed, but it wasn't as obvious or as uncomfortable. For white people, of course, because black people have never had a good time with it.’

Two punctuation marks are found as collocates in the written corpus: the exclamation and the interrogation marks. Exclamation mark co-occurs in the right window of *claro* (e.g., *claro!* / *claro está!* / *claro que é!*) in responses and in comments to what was previously said, as an interjection that emphasizes the fact that the previous content is known and obvious. The interrogation mark occurs on the left window of *claro* and marks the ending of a question to which the speaker responds with *claro*. The distribution of both punctuation marks is very similar to what was found with the collocate *pois*, namely, they tend to occur in contexts of verbal interaction or reproducing a dialogue.

3.3.3 Structural/Modal Domain: Elaboration

In the spoken subcorpus, the conjunction *e* ‘and’ receives a high statistical measure as collocate of *claro*: when it occurs in the right position, it is always preceded by a pause (*claro / e*). *Claro* closes the preceding segment and evaluates the situation it denotes as inevitable, and the conjunction *e* introduces a new textual sequence. The conjunction *e* also occurs in left position, frequently without a pause: the DM *claro* has scope over the right context marking it as inevitable. In (12), despite the lack of punctuation between node and collocate, each DM seems to have an individual function. However, in other contexts, it is difficult to clearly identify the function of each DM separately, as both seem to contribute to the structural domain, by providing further elaboration on a topic (13).

- (12) MJC – partiu um pé e claro ao princípio andou com gesso e andou com gesso durante bastante tempo (pfammn23.txt)
 ‘[he/she] broke [his/her] foot and of course at first [he/she] was wearing a cast and was wearing a cast for a long time’
- (13) MJC - e o que é certo / é que eles morreram todos // e eu vi-os morrer todos // e claro / aquela gente toda / e tanta gente que eu conhecia / foi morta assim // nem todos conseguiram fugir (pfammn23.txt)

‘and what is certain / is that they all died // and I saw them all die // and of course / all those people / and so many people I knew / were killed like that // not everyone managed to escape’

3.3.4 Structural/Modal Domain: Contrast

The conjunction *mas* ‘but’ can occur on the right position of *claro*, separated from the node by a pause (*claro / mas*). An inspection of the corpus shows that all but one context involve in fact two different speakers, one uttering *claro* and the other *mas* (this is in fact an issue when extracting concordances of spoken data). The exception is (14), where *claro* marks agreement with the previous speaker and *mas* introduces a sequence that reframes the context of the agreement. In all other contexts, *mas* is located in the left position, typically without pauses (15), and expresses contrast, while *claro* adds information on the fact that the situation is inevitable.

- (14) HLR - foi mau / e às vezes é doloroso / é difícil / é essas coisas todas //
 mas isso é
 JOS - mas é uma aprendizagem sempre //
 HLR - claro / mas / isso é num quadro mais vasto (pfamdl21)
- (15) MAI - e lá me fizeram um sapato novo // hhh / mas claro / demorou oito dias / a fazer o sapato (pfammn05)
 ‘and there they made me a new shoe // hhh / but of course / it took eight days / to make the shoe’

In the written corpus, the collocate *mas* occurs either on the left or the right of the node, separated by a comma, or on the left position with no comma (*mas claro*). In (16), the concessive value is marked by “é certo . . . mas claro” (‘it is true . . . but of course’).

- (16) E, é certo, eu acredito na arte popular, mas claro, vejo também que a arte, que a literatura são casos de cultura e tem de haver uma ampla informação para se conseguir realizar. (L0941, livro_literário)
 ‘And, certainly, I believe in popular art, but of course, I also see that art, literature are cases of culture and there must be extensive information to be able to achieve it.’

The collocate *depois* ‘after’ is also found in the results provided by AntConc, both on the left and the right context in the spoken corpus. The adverb *depois* contributes with a temporal meaning, and *claro* marks an inevitable situation. However, the contexts frequently involve a contrast between two temporal situations, as in (17).

- (17) LAL – mas deve ser ao princípio / para fazer clientela / depois claro que começam a pagar (ptelpv03)
 ‘but it must be at first / to build clientele / then of course they start paying’

Finally, the personal pronoun *eu* ‘I’ is ranked as relatively significant as a collocate of *claro*. These are contexts where the marking of agreement is followed by a personal stance regarding the previous utterance. In (18), the speaker’s opinion (*eu acho*) matches and reinforces the opinion of the addressee. Example (19), from the spoken corpus Spoken Portuguese (Português Falado) illustrates contexts where *claro* has a face saving function: the position of the speaker JOS is challenged by the addressee (NAZ), and the personal pronoun *eu* introduces a segment that reframes the opinion of the speaker JOS.

- (18) ZEB - porque eu acho que é uma prenda útil / não é ?
 GAB - hum hum / claro / eu acho que é ótimo / até // (ptelpv14)
- (19) NAZ - eu acho que não // perde a virilidade porquê ? acho que isso não faz /
 JOS - não
 NAZ - sentido
 JOS - claro // eu não digo que perca a virilidade
 NAZ - exacto
 JOS – por isso // o que eu digo é que alguns podem pensar que perdem (Português Falado, Bom Senso e Bom Rosto.txt)

The analysis of these collocations has showed in that many contexts both units contribute with their individual meaning. This is most obvious when a pause (in the spoken corpus) or punctuation (in the written corpus) occurs between the node and the collocate. However, there are several cases where the collocation may occur with no pause or punctuation, or even cases where the absence of pause is more frequent in the corpus. The collocates are, for the most part, conjunctions, pragmatic markers and adverbs. The lack of pause or punctuation could be a sign of some integration of the two units. These co-occurrences of DMs has been discussed and categorized for Catalan and Spanish (Cuenca & Marín 2009), French (Crible 2018) and revised, based on English data, in Cuenca & Crible (2019). The au-

thors distinguish between different categories of integration of the DMs. One category is *juxtaposition*, that applies when the DMs (typically conjunctions) take scope over different units (e.g., *and if*, *and because*, *because when*). The second category is *combination*, subdivided in two levels of integration: *addition* and *composition*. Addition applies “when the markers take scope over the same discourse unit and they exhibit distinct but compatible meanings. Specifically, the second marker either narrows down or reinforces the meaning of the first one, which is more general and underspecified.” (Cuenca & Crible 2019: 178). Examples of addition are *but in fact* (the contrastive meaning of *but* is reinforced by *in fact*), and the cluster *and secondly* (*secondly* specifies the position of the following segment in an enumeration introduced by *and*). In the case of a relation of addition, the second DM may be replaced by a synonym or near-synonym (for instance, *but in fact* could be rephrased as *but actually*). Finally, two “co-occurring DMs form a compound DM when they have the same scope and jointly express a single meaning, that is, the contribution of the individual markers can no longer be disentangled.” (Cuenca & Crible 2019: 179). The authors are aware of ambiguous cases where the same cluster of DMs may correspond to different levels of integration in different contexts. This is the case of *and then*, that, depending on the context, shows properties of juxtaposition, addition or composition.

In the case of *claro*, we discussed in section 3.2 the following co-occurrences in the corpus:

claro claro
pois claro
sim claro
e claro
mas claro
depois claro

The first three cases are responses or comments to a previous utterance (also comments of the speakers' own utterance in some contexts). *Claro claro* is a repetition of the same DM: both DMs have the same meaning and scope over the same unit. The co-occurrence of the two DMs reinforces the marking of agreement, as in (4). According to these criteria, the co-occurrence *claro claro* falls into the addition category. *Pois claro* and *sim claro* are both cases where the two DMs and *claro* have the same scope, and a related meaning of agreement, and *claro* adds a value of shared knowledge. This also falls into the suggested category of addition. However, some cases of *pois claro* are less clear-cut, as we discussed regarding example (6). The exact nature of the contribution of each DM in this example seems difficult to pinpoint, and the context is closer to a single contribution of

meaning in terms of strong agreement. This collocation might be closer to composition in some contexts, but an analysis of more data would be required.

The conjunction *e* and *claro*, when co-occurring, have scope over the same unit, and seem to contribute with separate meanings in contexts such as (12). The conjunction introduces a new situation and *claro* qualifies it as known and obvious, and this relation between the two DM falls into the properties of the category addition. However, in (13), it is difficult to clearly separate the contribution of each DM because the conjunction doesn't introduce new information, but rather reformulates the previous segment and makes the text progress. The DM *claro* seems redundant as a marker of known information, since the speaker has just mentioned the death of all the people. In this sense, example (13) could point to a relation between addition and composition.

In contexts of contrast with *mas*, as in (15) and (16), each DM contributes individually to the meaning of the collocation: *mas* establishes a contrast or concession value, and *claro* qualifies the contrastive segment as known and inevitable. The DM *claro* establishes a relation of addition with the DM *mas*.

4 Conclusion

The Reference Corpus of Contemporary Portuguese project started in 1988, so it's far from being a new resource for Portuguese. The different stages of its compilation have already been described, but recent changes, mostly regarding access to its spoken subpart had not yet been addressed. We take this opportunity to describe the written and spoken subcorpora of the CRPC under a global approach and then to focus on the two query platforms that give access to the data. The written subcorpus is available on CQPweb, a robust platform, able to deal with a large corpus of written language. The spoken subcorpus, a collection (without repetitions) of all the data collected for different projects, can be queried on TEITOK. This platform has the advantage of including the text-to-sound alignment, so that the full recording can be listened to, or individual utterances of the recording, linked to the orthographic transcriptions. The different annotation levels on each platform are also discussed, as well as the query options, adapted to each modality.

In the second part of the paper, we analyze significant co-occurrences of the DM *claro* in both corpora. We discuss the extraction of these preferred co-occurrences, referred to as collocations of the node, and the difficulties that are encountered. The collocations of *claro* highlight the intrinsic interactional dialogism of the DM, as a response or comment to a previous utterance, or a comment over an utterance, but also the interaction with structural DMs that ensure the textual progression as a con-

junction of events, a temporal ordering of events, or contrast between events. In most contexts, each DM contributes with its individual meaning, in a relation that Cuenca & Crible (2019) define as a relation of addition. However, in certain contexts of co-occurrences with *e* and *pois*, the individual meaning of each DM is difficult to establish and the collocations *e claro* and *pois claro* seem ambiguous between an addition and a composition relation.

These findings need to be explored in more detail, through the analysis of more contexts, in both the spoken and the written subcorpora. Two lines of research are enabled by the corpora and could provide interesting results: a detailed analysis of the distribution of the collocations in the corpus, per text type (some results are provided for two DMs in the paper and provide evidence for internal properties of each text type in terms of level of interaction); an analysis of the distribution of the collocations per variety of Portuguese.

Bibliography

- Anthony, Laurence (2019): *AntConc (Version 3.5.8)* [Computer Software]. Tokyo, Japan: Waseda University. Available from <<https://www.laurenceanthony.net/software>>.
- Bacelar do Nascimento, Maria Fernanda (2000): *Corpus de Référence du Portugais Contemporain*, in Mireille Bilger (ed.). *Corpus, Méthodologie et Applications Linguistiques*. Paris: Champion/Presses Universitaires de Perpignan, pp. 25–30.
- 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 [CD-ROM].
- 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 Ferreira (eds.). *Working with Portuguese Corpora*. London: Bloomsbury.
- Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Rita Veloso, Sandra Antunes, Florbela Barreto and Raquel Amaro (2005): “The Portuguese Corpus”, in Emanuela Cresti and Massimo Moneglia (eds.). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphia, Benjamins, pp. 163–207.
- Bacelar do Nascimento, Maria Fernanda, Maria Lúcia Garcia Marques and Maria Luísa Segura da Cruz (1987 a): *Português Fundamental, Métodos e Documentos, tomo 1: Inquérito de Frequência*. Lisboa: INIC, CLUL.
- Bacelar do Nascimento, Maria Fernanda, Paul Rivenc and Maria Luísa Segura da Cruz (1987b): *Português Fundamental, Métodos e Documentos, tomo 2: Inquérito de Disponibilidade*. Lisboa: INIC, CLUL.
- Bick, Eckhard (1999): *The parsing system PALAVRAS*. Aarhus University Press.
- Branco, António and João Silva (2003): “Contractions: breaking the tokenization-tagging circularity”, vol. 2721, chap. *Lecture Notes in Artificial Intelligence*. Springer, pp. 167–170.
- Crible, Ludivine (2018): “Discourse Markers and (Dis)fluency. Forms and Functions across Languages and Registers”, in *Pragmatics & Beyond New Series*. Amsterdam: John Benjamins.

- Crible, Ludivine and Liesbeth Degand (2019): “Domains and Functions: A two-dimensional account of discourse markers”, in *Discours*, 24.
- Cuenca, Maria Josep and Maria Josep Marín (2009): “Co-occurrence of discourse markers in Catalan and Spanish oral narrative”, in *Journal of Pragmatics*, 41(5), pp. 899–914.
- Cuenca, Maria Josep and Maria Josep Marín (2012): “Discourse markers and Modality in Spoken Catalan: The Case of (*és clar*)”, in *Journal of Pragmatics*, 44(15), pp. 2211–2225.
- Cuenca, Maria Josep and Ludivine Crible (2019): “Co-occurrence of discourse markers in English: From juxtaposition to composition”, in *Journal of Pragmatics*, 140, pp. 171–184. <<https://doi.org/10.1016/j.pragma.2018.12.001>>.
- Cuenca, Maria Josep (2013): “The fuzzy boundaries between discourse marking and modal marking”, in Liesbeth Degand, Bert Cornillie and Paola Pietrandrea (eds.). *Discourse markers and modal particles. Categorization and description*. Amsterdam: John Benjamins, pp. 181–216.
- Daelemans, Walter and Antal Van den Bosch (2005): *Memory-Based Language Processing*. Cambridge: Cambridge University Press.
- Daelemans, Walter, Jakub Zavrel, Peter Berck and Steven Gillis (1996): “Mbt: A memory-based part of speech tagger generator”, in *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, pp. 14–27.
- Davies, Mark (2014): *Creating and using the Corpus do Português and the Frequency Dictionary of Portuguese*, in Tony Berber Sardinha and Telma Ferreira (eds.). *Working with Portuguese Corpora*. London: Continuum, pp. 89–110.
- Duarte, Isabel and Rogélio Ponce de León (eds.) (2020): *Marcadores Discursivos. O português como referência contrastiva*. Frankfurt/Wien: Peter Lang.
- Evert, Stefan and Andrew Hardie (2011): *Twenty-first century Corpus Workbench: Updating a query architecture for the new Millennium*. Paper presented at Corpus Linguistics 2011, University of Birmingham, UK. <http://cwb.sourceforge.net/files/EvertHardie_CL2011_paper.pdf>.
- Firth, John Rupert (1957): “A Synopsis of Linguistic Theory, 1930–1955”, in *Studies in Linguistic Analysis*.
- Fuentes Rodríguez, Catalina (1993): “Claro: modalización y conexión”, in *Sociolingüística Andaluza*, 8, pp. 99–126.
- Généreux, Michel, Iris Hendrickx and Amália Mendes (2012): “A Large Portuguese Corpus On-Line: Cleaning and Preprocessing”, in Helena Caseli, Aline Villavicencio, António Teixeira and Fernando Perdigão (eds.). *Computational Processing of the Portuguese Language PROPOR 2012*, Lecture Notes in Computer Science, vol. 7243. Berlin, Heidelberg: Springer, pp. 113–120. <https://doi.org/10.1007/978-3-642-28885-2_13>.
- Halliday, Michael A. K., and Christian M. Matthiessen (2004): *An Introduction to Functional Grammar* (3^a ed.). London: Arnold.
- Hardie, Andrew (2012): “CQPweb – combining power, flexibility and usability in a corpus analysis tool”, in *International Journal of Corpus Linguistics*, 17(3), pp. 380–409.
- Janssen, Maarten (2016): “TEITOK: Text-Faithful Annotated Corpora”, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- Kennedy, Graeme (1998): *An Introduction to Corpus Linguistics*. London/New York: Longman.
- Kilgarriff, Adam and Gregory Grefenstette (2003): “Introduction to the Special Issue on Web as Corpus”, in *Computational Linguistics*, 29(3), pp. 333–347.
- Leech, Geoffrey (1991): “The state of the art in corpus linguistics”, in Karin Aijmer and Bengt Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London/New York: Longman, pp. 8–29.
- Lopes, Ana Cristina Macário (2021): “Contributos para o estudo do marcador discursivo 'claro' em português europeu”, in *Revista Galega de Filoloxía*, 14, pp. 71–83.

- McEnery, Anthony and Andrew Hardie (2012): *Corpus Linguistics*. Cambridge: Cambridge University Press.
- Mendes, Amália (2016): “Linguística de Corpus e outros usos do corpus em Linguística” in Ana Maria Martins and Ernestina Carrilho (eds.). *Manual de Linguística Portuguesa*. Berlin/Boston: De Gruyter, pp. 224–251.
- Mendes, Amália and Pierre Lejeune (2023): “Marcadores discursivos com funções modais: uma análise contrastiva de *claro* e seus equivalentes funcionais em francês”, in *Revista da Associação Portuguesa de Linguística*.
- Mendes, Amália and Pierre Lejeune (forth.): “Discourse markers in Portuguese”, in Maj-Britt Mosegaard Hansen and Jacqueline Visconti (eds.). *Manual of Discourse Markers in Romance*. Berlin/Boston: De Gruyter.
- Mendes, Amália and Sandra Antunes (2016): “Collocations in Portuguese: A corpus-based approach to lexical patterns”, in Begoña Sanromán (ed.). *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*. Helsinki: Société Néophilologique, pp. 141–166.
- Mendes, Amália, Pierre Lejeune and Carolina Nunes (2020): “Perguntas-respostas em textos escritos: uma análise no âmbito das relações discursivas”, in *Revista da Associação Portuguesa de Linguística*, 7, pp. 226–241. DOI: <<https://doi.org/10.26334/2183-9077/rapln7ano2020a14>>.
- Mendes, Amália, Michel Généréux, Iris Hendrickx and Sandra Antunes (2023): “Manual for the CRPC on the CQPweb interface”, v. 1.6. Centro de Linguística da Universidade de Lisboa. <http://gamma.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_6_en.pdf>.
- Pons Bordería, Salvador (1998): “Conexión y conectores: estudio de su relación en el registro informal de la lengua”, in *Cuadernos de Filología*, Anexo XXVII. Valencia: Universitat de Valencia.
- Pons Bordería, Salvador (2003): “From Agreement to Stressing and Hedging: Spanish *Bueno* and *Claro*”, in Gundrun Held (ed.). *Partikeln und Höflichkeit*. Frankfurt: Peter Lang.
- Pons Bordería, Salvador (2012): “Claro. Una palabra sobre los apellidos de la sintaxis”, in José Jesús de Bustos Tovar et al. (eds.). *Sintaxis y análisis del discurso hablado en español. Homenaje a Antonio Narbona*. Sevilla: Servicio de Publicaciones de la Universidad de Sevilla.
- Rocha, Paulo and Diana Santos (2000): “CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa”, in Maria das Graças Volpe Nunes (ed.). *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, Atibaia, São Paulo, Brasil, pp. 131–140.
- Schmidt, Thomas (2012): “EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language”, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC2012)*, Istanbul, Turkey, pp. 236–240.
- Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, John (1996): *EAGLES Preliminary recommendations on Corpus Typology*. EAG-TCWG-CTYP/P. <<https://www.ilc.cnr.it/EAGLES96/corpusstyp/corpusstyp.html>>.
- Vanderschueren, Clara and Amália Mendes (2015): “Panorama de los corpus y textos del portugués europeo”, in Maria Iliescu and Eugen Roegiest (eds.). *Manuel des Anthologies, Corpus et Textes Romans*. Berlin/Boston: De Gruyter, pp. 58–80.

Anton Granvik

On the Origins of the Shell Noun Construction in Portuguese

1 Introduction

Shell nouns (cf. Schmid 2000, 2018) are abstract nouns which serve the important function of encapsulating, or classifying, other, often more complex elements present in the discourse as simple concepts such as *idea*, *conclusion*, *fact*. Flowerdew & Forest (2015: 46) describe them as “open-class vocabulary items with discourse signalling functions, items which [. . .] exhibit a requirement for specification in context and which [. . .] act as discourse signposts.” Sanches Duran & Volpe Nunes (2019: §1) add that “In encapsulation there is an encapsulating word (pronoun or noun phrase) that is used in place of an entire sentence encapsulated by it.”¹

As shown in (1), *certeza* and *facto* make reference to the following passages, *que . . . a parte social iria continuar comigo* and *eu ser uma pessoa preocupada*, respectively. Importantly, the information presented in these passages can be considered a *certainty* and a *fact*; and, without the following passages we do not know what the certainty and the fact are.

- (1) Tenho a **certeza** de que se deixasse de trabalhar em televisão, ou se me reformasse da televisão, a parte social iria continuar comigo, é indissociável do **facto** de eu ser uma pessoa preocupada com as pegadas que deixo aqui quando me for embora; de ser mulher, portanto confrontada com as desigualdades que ainda existem cá e lá fora. (CP-NOW, <https://online.sapo.pt/641931>)

‘I’m sure that if I stopped working in television, or if I retired from television, the social part would remain with me, it’s inseparable from the fact that I’m a person who is concerned about the footprints I leave here when I leave; that I’m a woman, and therefore confronted with the inequalities that still exist here and there.’

In present-day Portuguese, shell nouns typically appear in three constructional formats, each with two variants, as shown in (2)–(4):

¹ My translation. The original reads: “No encapsulamento há uma palavra encapsuladora (pronome ou sintagma nominal) que é usada no lugar de toda uma oração por ela encapsulada.”

- (2) a. **(det) N (adj) *ser que* + oração**
A idéia apresentada é que as pessoas dediquem um dia por mês para trabalhos voluntários.²
 ‘The idea is for people to dedicate one day a month to voluntary work.’
- b. **(det) N (adj) *ser* + (de) infinitive**
 A minha parte eu vou pagar e posso assegurar que **a idéia** de todos é **fazer** o mesmo
 ‘I’ll pay my share and I can assure you that everyone’s idea is to do the same’
- (3) a. **(det) N (adj) *de* + infinitive**
 Com[o] surgiu **a idéia de fazer** esse filme de guerra?
 ‘How did the idea of making this war film come about?’
- b. **(det) N (adj) *de* + infinitivo pessoal**
 –Ah! meu pai! **essa idéia de irmos** para tão longe, sem esperança de um dia podermos voltar...
 ‘Oh, my father, this idea of going so far away, with no hope of ever coming back. . .’
- (4) a. **(det) N (adj) *de que* + oração**
 –O senhor fala que o PDT quer espaço político, mas fica sempre **a idéia de que** espaço significa cargo.
 ‘You say that the PDT wants political space, but there’s always the idea that space means office.’
- b. **(det) N (adj) *que* + oração**
 Por isso se criou **a idéia que** uma economia saudável é aquela que cria megasuperávits, o que não é correto...
 ‘That’s why the idea has been created that a healthy economy is one that creates mega surpluses, which is not correct...’

In the medieval period, however, the available constructions were fewer, including only those in (2a), (3a), (3b) and (4b), as in (2c) to (4c):

- (2) c. **Ca *não é razão que*** aquele que recebe a entrega mingue nada do seu.
 ‘Because it is not right [reason] that the one who receives the gift shall be without all that is his’

² If not marked otherwise, all the examples are taken from Davies & Ferreira’s (2006-) *Corpus de Português* (CP).

- (3) c. Ca a Sennor que o atan ben dá non á ome **razon de lle furtar nen de roubar-ll'** o seu nen llo filar...‘Because a Lord who gives so well a man does not have the right [reason] to rob him or to take it from him. . .’
- (4) c. E don Vaasco fernandez Maestre do Tempre e os ffreyres dessa Ordim da outra **per Razom que** o dicto procurador delRey dizia que o dicto Maestre e ffreyres tragiã ascondudas...‘And Don Vaasco Fernandez, Master of the Templars, and the freedmen of that Order of the other, for the reason that the said procurator of the King said that the said Master and freedmen were carrying X hidden. . .’

Given these differences in construction format between the medieval period and the present-day language, in this paper I propose to investigate and describe

- A. how the shell-noun construction in Portuguese arises and evolves over the centuries;
- B. how the use of the constructional formats develops over time (*N ser que* + oração, *N de* + infinitive, *N (de) que* + clause);
- C. what the relationship is between the function (shell-noun use or not?), the form (construction) and usage context of these constructions.

To reach these goals the paper focuses on nine shell-nouns which are representative of the above constructions in three different periods: the 13th and 14th centuries, the 16th and 17th centuries and the 19th and 20th centuries. The analysis of the nine nouns, i.e. *caso*, *facto*, *ideia*, *mercê*, *questão*, *razão*, *sinal*, *temor* and *vontade*, is based on data retrieved from *Corpus do Português* (Davies & Ferreira 2006-), a large diachronic reference corpus of Portuguese. In line with the interests of Digital Humanities, the analysis is based on different quantitative analyses, such as collostructional analysis (Stefanowitsch & Gries 2003) and logistic regression (Gries 2013, Winter 2020).

The interest of investigating the history and early development of the shell-nouns in Portuguese is further motivated by the fact that previous studies have mainly focused on the present-day language. Secondly, observations regarding these nouns in other languages suggest that their use is connected to the rise of informative and journalistic texts in the 17th century (Borreguero 2018, Borreguero Zuloaga & Octavio de Toledo 2007). Previous studies on Spanish, however, have shown that some nouns are used as shell nouns as early as the 13th century

(Granvik 2019).³ It is thus interesting to see whether this is the case in Portuguese as well.

The paper is structured as follows. Section 2 introduces the concept of shell nouns and their encapsulating function. What I call the shell-noun construction is also presented and defined. The corpus and methods used in the analysis are presented in Section 3. Section 4 focuses on analyzing the data first from a diachronic perspective, and then from the perspective of determining which criteria best explain the difference between encapsulating and non-encapsulating uses. Section 5 concludes.

2 Theoretical Background

The notion of *shell noun* was coined by Schmid in the late 20th century (Schmid 2018: 110–111) and is used to refer to what he calls a functionally defined category: “A noun is turned into a shell noun when a speaker decides to use it in a shell-content complex in the service of certain aims. Shell-nounhood is thus a functional property” (Schmid 2000: 13). Shell nouns are thus not a category a noun either belongs to or not, but rather many, often abstract nouns are turned into, or function as, shell nouns in certain uses, i.e. when they are used to encapsulate, classify, or signal another discourse element as what the designate, such as a *fact*, an *idea*, a *conclusion*, etc.

Schmid’s (2000) monograph on shell nouns approaches this group of nouns from three perspectives, highlighting their semantic, cognitive, and textual function. According to Schmid (2000: 14), the semantic function corresponds to “characterizing and perspectivizing complex chunks of information”, the cognitive function to “temporary concept-formation (encapsulate these complex chunks . . . in temporary nominal concepts)” and the textual function to “linking these nominal concepts with clauses.”

The text or discourse function of shell nouns had, of course, been noticed way before Schmid’s (1998, 1999, 2000) work. Within the discourse tradition these referring, signalling, linking, encapsulating, and characterizing functions (cf. Schmid 2018: 111) all refer to the fact that the nouns “act as labels, also known as encapsulators, [and] being compactors, they are responsible for changing or linking topics

³ See also Schmid & Mantlik (2015) for some observations regarding early shell-noun uses in English, dating from, at least, the 14th century.

and also contribute to preserving textual continuity by introducing new information within old information” (Paredes Silva & Bezerra 2013: 219–220).⁴

The essence of these nouns is the concept of encapsulation which, in the words of Guimarães (2011:30) “suggests a process of linguistic compacting of events around a lexical item ... This means that a word class summarizes content to be enunciated.” For Andrade (2019: 180), encapsulating nouns imply “a choice capable of summarising information present in the context and, at the same time, transforming it into a new object of discourse.”⁵ Flowerdew & Forest (2015: 48) also take encapsulation, i.e. “equative encapsulation with lexical specifics provided elsewhere in the text”, as the defining “criterion for SN membership”. Schmid (2000: 21) acknowledges that the contextual identification of what the shell noun encapsulates is central, i.e. what he calls *experiential identity*, the notion that the shell noun and the shell content express ideas about the same thing.

This is what happens in (5), where there are two nouns acting as shells, i.e. *conceção* and *ideia*. *Conceção* makes anaphoric reference to the immediately preceding textual segment —*Adquirir novos conhecimentos* ‘acquiring new knowledge’— whereas *ideia* encapsulates (by cataphoric reference) the following completive sentence —*o aprender é entendido como um meio para saber novos conteúdos* ‘learning is understood as a means of acquiring. . .’.

- (5) Como podemos observar, a *conceção* mais referida diz respeito ao Adquirir novos conhecimentos, sendo enunciada por 49,1% dos participantes. Esta **conceção** refere-se à *ideia* de que o aprender é entendido como um meio para saber novos conteúdos. (Carvalho e Ribeiro 2017: 210)

‘As we can see, the most mentioned conception relates to Acquiring new knowledge, which was mentioned by 49.1 per cent of the participants. This conception refers to the idea that learning is understood as a means of acquiring new content.’

Note that, what the presence of an encapsulator as *ideia* adds to the sentences in (5), is a characterisation of the proposition “learning is understood as a means of acquiring new content” as an ‘idea’, i.e. as a kind of mental entity. But we could

4 My translation. The original reads: “Os SNs que atuam como rótulos, também chamados encapsuladores, sendo compactadores, são responsáveis por mudar ou ligar os tópicos e contribuir, também, na preservação da continuidade textual ao introduzir as informações novas dentro das velhas” (Paredes Silva & Bezerra 2013: 219–220).

5 My translation. The original reads: “O encapsulador é um processo referencial que merece destaque, pois se apresenta como uma escolha capaz de sumarizar informações presentes no contexto e, ao mesmo tempo, transformá-las em novo objeto de discurso.”

substitute *ideia* by a noun such as *facto* ‘fact’, and the overall meaning would not be very much altered, except that we would now be talking about something conceived of as a fact instead of as an idea. The function of *ideia* in (5) corresponds perfectly to Abad Serna’s (2015: 230) observation that “nouns used as encapsulators have the singularity of synthesising the reference of the textual segment to which they refer”.⁶

According to Schmid (2000: 21), the encapsulating function can be considered a prototypical category, the core of which is the experiential identity relation (Schmid 2000: 21; cf. López Samaniego 2011: 447; Abad Serna 2015), i.e. “the notion that the shell noun and the shell content express ideas about the same thing”. Returning to example (5), above, this experiential identity can be verified by noting that the clause “learning is understood as a means of acquiring new content” is an *idea*. However, if the core of the category of shell nouns is the notion of experiential identity, and the category is prototypical in nature, it is to be expected that there are less typical and even marginal cases around it. Thus, for Schmid (2000: 25–26), nouns used as part of “expanded predicates”, in combination with functional verbs of the type *to have* and *to do*, “cannot be considered good examples of encapsulating nouns”. Likewise, nouns of temporal or locative meaning, such as *time*, *place*... “are treated as marginal” (Schmid 2000: 26). Examples (6) to (8) illustrate different degrees of typicality, where the notion of “experiential identity” becomes increasingly difficult to maintain:

- (6) Por exemplo, **no caso de serem vistas de longe**, deveria imprimir força e monumentalidade. (CP, 19Ac:Br:Enc)
 ‘For example, in case they were to be seen from a distance, they should be strong and monumental.’
- (7) Que se **fizesse**, porém, grande **caso de juntar com a pregação os exercícios de humildade**. (CP, 15:Lucena:SFXavier)
 ‘However, it would be a good idea to combine preaching with exercises in humility.’
- (8) Também não **faça caso de que lhe faltam fôrças** para a batalha porque isto é soberba e não humildade (CP, 16:Chagas:Cartas)
 ‘Don’t pretend that they lack the strength for battle either, because that’s pride, not humility’

⁶ My translation. The original reads: “los sustantivos empleados como encapsuladores tienen la singularidad de sintetizar la referencia del segmento textual al que remiten.”

In (6), although *no caso de* is a lexicalized expression, it is transparent enough that it is clear that ‘to be seen from a distance’ is what constitutes the *case*, so there is clearly identification and encapsulation. In (7), on the other hand, *caso* is combined with the light-verb *fazer* ‘to do, make’, and the compound predicate *fazer caso* is combined with *de* which appears to maintain some of its sense of ‘aboutness’ (see Granvik 2014). It seems possible to interpret that the *case* is *to combine preaching with exercises*. . ., but this interpretation is clearly less salient than in (6). In (8) the experiential identity relation is even more peripheral, since the same compound predicate *fazer caso de* is used in another sense, that is, ‘to make an issue about’ or, as the translation suggest, ‘to pretend’. The complement clause in (8) thus corresponds more to a direct object than to an appositional clause (see Leonetti 1993, 1999).⁷

When determining whether an instance of an abstract noun constitutes a case of the shell noun use or not, it seems natural to take the notion of experiential identity as a defining criterion. That is, a noun is considered to have an encapsulating function if and only if there is “experiential identity” between the noun and another element present in the immediate context (typically, the complement clause following it). Indeed, this is how shell nouns, and indexical nouns, are often defined (see above; Flowerdew & Francis 2015). So, if the answer to the question “Is N equal to the (complement) clause? (or, is the clause (an) N?) is affirmative, it would count as a shell use.

However, as examples (6) to (9) indicate, the distinction between experiential identity and its absence is not always clear (let alone objective). There is, for example, no biunivocal relation between one or more syntactic constructions (e.g. *N de que* or *N é que*) and the relation of identity. As Samaniego’s (2011: 446) discussion of Leonetti’s (1999) examples reveals, the same noun can be combined with both argumentative (9) and appositive complement clauses (10) in the same syntactic construction:

- (9) La explicación [de que suspendas siempre] es que no te preocupas por entender bien la asignatura,
 ‘The explanation [that you always fail] is that you don’t bother to understand the subject well’

7 According to López Samaniego (2011: 446–447), “the discursive actualisation of the noun only appears in the complement clause when it maintains an appositive relation of identity with the noun”. The alternative is that the complement clause has a predicate function, constituting an argument of the nominal nucleus, reflecting Leonetti’s (1993, 1999) distinction between argumentative and appositive complement clauses.

- (10) La explicación [de que han aumentado los gastos] no es muy convincente.
‘The explanation [that expenses have increased] is not very convincing.’

In (9) there is no identification of the noun *explanation* as “you always fail”, whereas in (10) “that expenses have increases” is the explanation.

On the other hand, consider examples (11) and (12), involving the noun *temor* ‘fear’. In (11), it is difficult, if not impossible, to determine whether the complement clause “as minorias sérvias . . . seriam perseguidas” is what constitutes the ‘fear’ (*temor*), or whether it is the object of the fear, i.e. what the fear is about. Clearly, this is a question of degree, where the semantics of the noun *temor*, which parallels the modal verb *temer* ‘to fear’, is compatible with both an identifying and an object interpretation. In the first interpretation, we would say that the fear is that the Serbian minorities are being persecuted; in the second, we would say that this is what is feared. Compare this to the use of *temor* in the equative construction with *ser* in (12), where the identification of *que os Brasil suspende as exceções . . .* as a *fear* is made explicit.

- (11) Com a onda de independências a partir de 90, os nacionalistas sérvios disseminaram **o temor de que** as minorias sérvias em as outras repúblicas seriam perseguidas como ocorreu quando sérvios (CP, 19N:Br:Folha)
‘With the wave of independence from the 1990s onwards, Serbian nationalists spread the fear that Serbian minorities in the other republics would be persecuted as happened when Serbs’
- (12) **O grande temor** dos empresários argentinos **é que**, em outubro, o Brasil suspende as exceções às restrições às importações. (CP, 19Or:Br:Intrv:ISP)
‘The great fear of Argentine businesspeople is that in October Brazil will lift the exceptions to import restrictions’

Since the relation of experiential identity between the noun and the complement clause is often a question of degree and may thus be difficult to establish objectively, Granvik (2019) introduced an alternative —formal and therefore objective— approach to the analysis of shell nouns. Starting from Schmid’s (2018: 11) observation that “a limited number of lexico-grammatical patterns are particularly suitable for encoding the link of experiential identity between shell noun and shell content”, Granvik considered all combinations of N with a complement clause (i.e. the constructions included in points b) and c), below), a case of the encapsulating construction. The four grammatical constructions that, according to Schmid greatly facilitate that a noun is used with an encapsulating function are listed in the following:

- a) demonstrative + noun (e.g. *this dilemma...*)
- b) noun followed by modifying clause (*the dilemma + that/infinitive...*)
- c) noun as the subject of an identifying clause with the verb *ser* (*the dilemma is that...*)
- d) neuter demonstrative or neuter personal pronoun as subject and encapsulating noun as predicate (*this is a dilemma*) (following Rodríguez Espiñeira 2015: 658; the examples are mine)

Focusing on the constructions in b) and c) finds support in Schmid's (2018: 115) consideration that these two lexico-syntactic patterns (N *be* + *that*-clause and N + *that*-clause) "stand out from the others by providing the kind of environment in which shell nouns thrive".

I propose, then, that there is an encapsulating construction in the schematic sense given by Construction Grammar (cf. Croft & Cruse 2004; González-García 2012; Traugott & Trousdale 2013; Hilpert 2019). This shell noun construction has four variants, which all involve the presence of a complement clause in either infinitive or finite form: N *que* + clause, N *de que* + clause, N *de* + infinitive and N *ser* + clause (finite or infinitive) (see Schmid 2018: 124–125 for a discussion of the construction status of the encapsulating construction(s)).⁸ As will become evident below, one of the focal points of the analysis is to verify to what degree the formal characteristics of the shell noun construction corresponds to the encapsulating function of the nouns.

3 Data and Methods

The data analyzed in this paper was retrieved from Davies & Ferreira's (2006–) reference corpus of Portuguese called *Corpus do Português* (CP). The diachronic part of this corpus includes roughly 45 million words, ranging from the 13th to the 20th century.

The nouns were selected using a series of criteria. As a starting point I use Granvik's (2017b: 39) century-wise collocation analysis (Stefanowitsch & Gries 2003) of the most frequently attested nouns in the sequences N (*de*) *que*, N *ser que* and N *de* + infinitive.⁹ Second, among the most highly ranked nouns in each cen-

⁸ I thereby exclude the formats involving demonstratives, types a) and d) from consideration in this paper.

⁹ Note that the collocation analysis presented in Granvik (2017b) was performed on data ranging from the 1500s to the 1900s (due to scarcity of data in previous centuries) and focused on the sequences N *que*, N *de que* and N *de* + infinitive. For the 1200s to 1400s, I manually inspected

tury, I selected nouns that fall into different types of nouns, following Schmid's (2000) classification of shell nouns into factual, mental, modal, linguistic, and circumstantial ones. Third, inspecting the nouns and their usage frequencies across the construction formats and the centuries permitted selecting a total of nine nouns, three for each period as shown in the table below.

1200–1399 *mercê* (modal), *razão* (factual), *vontade* (mental)
 1500–1699 *caso* (factual), *sinal* (factual), *temor* (mental)
 1800–1999 *facto* (factual), *ideia* (mental), *questão* (mental)

Having limited the study to these nine nouns, the following step was to extract all instances of them in the following sequences:

N *de* N *de* + infinitive, N *de que* + clause
 N *que* N *que* + clause, *é* N *que* + clause
 N [dist=9] *é que* N *é que* + clause

This provided me with an initial dataset of a little over 9350 instances. As the figures in Table 1 show, the nine nouns appear in these constructions in a step by step fashion. In the 13th century only *mercê*, *razão*, *sinal* and *caso* are documented, with *vontade* and *temor* appearing in the 14th. Table 1 also reveals that the usage frequencies of the attested nouns go up significantly and the six nouns are used in all four constructions from the 15th century on. In the 18th century the first uses of *ideia* and *questão* appear, and in the 19th and 20th centuries, all nine nouns are attested. The usage frequencies continue to increase for all except *mercê*, which appears to lose its status as a salient shell noun in the last two centuries. Finally, as the percentages shown on the last row of Table 1 indicate, the relative frequency of the use of these nouns in the different formats of the shell-noun construction increases steadily, moving from around 2 per cent in the medieval period, to approximately 5 per cent from the 1400s to the 1700s only to surpass the ten per cent line in the last centuries.

Out of the 9300 concordances extracted from the CP I created a subsample by arbitrarily selecting up to 100 cases of each noun and period (1200s–1300s, 1400–1700s, and 1800s–1900s). Compared with the figures in Table 1, then, the sample which was used for more detailed analysis included all 58 cases from the

all nominal collocates of the same sequences as above as well as N *é* + clause and selected three nouns which are indeed used with an encapsulating function. See Granvik (2015: § 5) and Schmid & Küchenhoff (2013) for discussion of the details of the collostructional analysis. The partition of the data into centuries is due to architecture of the CP.

Table 1: Distribution of the nine nouns across the 13th to the 20th century in the five constructions.

Nouns	1200s	1300s	1400s	1500s	1600s	1700s	1800s	1900s	Sum
<i>mercê</i>	3	8	50	45	73	11	7	1	198
<i>razão</i>	8	17	66	106	105	45	169	201	717
<i>vontade</i>	0	3	95	65	26	41	407	936	1573
<i>sinal</i>	6	2	34	43	72	33	159	169	518
<i>caso</i>	1	7	55	111	103	39	342	273	931
<i>temor</i>	0	3	31	47	16	18	18	44	177
<i>facto</i>	0	0	0	0	0	0	204	2841	3045
<i>ideia</i>	0	0	0	0	0	8	674	1057	1739
<i>questão</i>	0	0	0	0	0	4	41	419	464
Sum	18	40	331	417	395	199	2021	5941	9362
Nro. de N	990	1590	5575	9487	6825	4561	21256	40693	90977
%	1.8%	2.5%	5.9%	4.4%	5.8%	4.4%	9.5%	14.6%	

13th and 14th centuries; all 112 cases of *temor*, all 8 cases of *ideia* and all 4 cases of *questão* as well as 100 arbitrarily selected cases of *mercê*, *razão*, *vontade*, *sinal* and *caso* in the period running from 1400 to 1799. Finally, all 62 cases of *temor*, all eight of *mercê* as well as 100 arbitrarily selected cases of the remaining seven nouns from 1800 to 1999. The total number of cases included in the qualitative analysis is 1446 (see Table 4 in Section 4, below).

This subsample was annotated in detail by means of the variables presented in Table 2. As the table shows, for each concordance line extracted from the corpus I registered the construction involved (N *de* + infinitive, N *que* + clause, N *de que* + clause or N *é que* + clause), the noun (*caso*, *facto*, *ideia* . . .), the century, the immediate usage context of the noun (involving three structural properties, see Table 2), a formal characterisation of the usage as typical, less typical or untypical (based on the three subparts of the usage context variable), and, finally, whether it functions as a shell noun in the strict sense of there being an identifying relation between it and another element in the (immediate) usage context. The aim of this annotation was to capture the essential characteristics of the usage context of the nouns in order to determine to what degree they affect the way the nouns are used.

The key variable of the annotation is Identity relation, which distinguishes between the encapsulating uses of shell nouns, and the use of the nouns in other functions, particularly as part of complex predicates in combination with light-verbs, such as *fazer* ‘to do, make’, *dar* ‘to give’, *haver* ‘to be, exist’, *ser* ‘to be’, etc. The distinction between the two functions, shell vs. non-shell use, was made with the help of two questions:

Table 2: Variables and variable values used in the analysis.

Variable	Values
Construction (cxn)	inf(itive), <i>que, de que, é que</i>
N(oun)	<i>mercê, razão, vontade, caso, sinal, temor, facto, ideia, questão</i>
Century	1200s to 1900s
Usage context	a) Determiner or not; b) syntactic function (object, prepositional complement, predicative, subject); c) the head element of the noun (lexical verb, noun, etc; light verb, preposition, no head)
Formal typicality	typical, less typical, untypical ¹⁰
Identity relation	0 (no relational identity), 1 (identification with previous element or with a following complement clause)

- 1) Is the noun used in a verb-like fashion? Does the subordinate clause function as if it were a direct object?
- 2) Are the paraphrases “the noun is the (subordinate) clause content” or “the clause content is a N” acceptable?

If the answer to the first question is “Yes”, the function is considered non-encapsulating, i.e. “predicate”; if the answer to the second question is “Yes”, then the noun is considered to have a shell function (cf. Schmid 2000: 30). This means that the shell function of a noun is equated to there being a relation of identity or identification between the noun and the subordinate clause, as is the case in (13) as compared to (14), where there is no identification.

- (13) Parece um contra-senso, mas não é, **pela** simples **razão de que**, no meio da anarquia, a lógica não sobrevive! (CP, 19N:Pt:Jornal)
‘It sounds like nonsense, but it isn’t, for the simple reason that in the midst of anarchy, logic can’t survive!’
- (14) A Rosa **tem** toda **a razão de** acender uma vela aos pés de santo Antônio. (CP, 19:Fic:Br:Resende:Braco)
‘Rosa is absolutely right to light a candle at the feet of St Anthony.’

¹⁰ See the discussion of Table 3, below, for more details.

The formal traits making up the Usage context variable, determiner, syntactic function and head element, were used by Granvik (2019) to capture the formal typicality of the use of the nouns. The degree of Formal typicality was determined numerically by applying the value system presented in Table 3. Each example was assigned a number, and sum values from 0–1 were considered typical usage contexts, values 2–3 were considered less typical, whereas the highest values, 4 and 5, constitute untypical usage contexts. These correspond to usage contexts where the nouns bear no determiner, are used as direct objects of highly frequent verbs, often light verbs, such as *dar*, *fazer*, *haver*, *ter* or as prepositional complements of grammatical prepositions such as *com*, *sem*, *a*, *em*, *por*.

Table 3: Values assigned to formal properties to determine the degree of typicality of the usage context of each noun.

Criterion / level	0	1	2
Determiner	definite article demonstrative pronoun possessive determiner	indefinite article or pronoun	no determiner
Syntactic function	absolute, subject	(in)direct object, predicative, prepositional complement, unknown	
Head	lexical verb, non-grammatical prepositions no head	verb <i>ser</i> grammatical prepositions (<i>a</i> , <i>de</i> , <i>em</i> , <i>para</i> , <i>por</i>)	light verb (<i>haver</i> , <i>ter</i> , <i>fazer</i> , <i>dar</i>), prepositions <i>com</i> and <i>sem</i>

In comparison to Granvik (2019), where the formal characterisation was used directly to capture the differences in use of the course of the centuries, in this paper, an additional goal is to verify to what degree the formal characterisation proposed correlates with differences in function, i.e. Identity relation or not. The hypothesis is that formally typical uses will correlate with uses in which the nouns have an encapsulating function, involving an identity relation and establishing an anaphoric or cataphoric relationship to another contextual element.

4 Analysis of the Results

4.1 Diachronic Considerations

From a historical perspective, the main question is, naturally, how the nouns inserted into the shell noun construction are used in the medieval period and how the construction evolves over time. I will address these questions from the perspective of first the nouns and then the construction formats as a reflex of the encapsulating function. As was already seen in Table 1, above, the nine nouns are introduced subsequently into the different construction formats, and their usage frequency increases as we approach the present-day language. This is, of course, totally expected given that the nouns were selected based on their representativity (as measured by usage frequency) of different time periods.

As the figures in Table 4 show, there is a clear difference between the nouns with regard to their use as shell nouns or not. Thus, of the nouns which are attested already in the medieval period, *caso*, *mercê*, and *sinal* are mainly used as shell nouns, whereas *razão*, *temor* and *vontade* are more often used with a non-encapsulating function (*haver/ter razão de* + infinitive ‘have reason to’, *haver/ter/com/sem/por temor de* + infinitive ‘have fear of/that’ and *haver/ter/com vontade de* + infinitive ‘have the will to’). In the following centuries, 1400s to the 1700s, *temor* and *vontade* continue to be used mainly in a non-encapsulating function, *razão* and *sinal* are used in both functions, whereas *caso* and *mercê* are mostly used as shell nouns. In the 19th and 20th centuries, *facto*, *ideia* and *questão* are introduced. While *facto* and *ideia* are almost exclusively used as shell nouns, *questão* is used with both functions. In the 1900s *questão* appears mainly in combinations where no encapsulation is involved (*fazer questão de* + inf). The above considerations, then, allow dividing the nouns into three groups:

- A. Typical shell nouns: *caso*, *mercê*, *ideia* and *facto*
- B. Multi-use nouns: *questão*, *razão* and *sinal*
- C. Non-encapsulating nouns: *temor* and *vontade*

Table 4: Use of the nine nouns with an encapsulating or non-encapsulating function across the centuries.

Nouns / identity relation	1200	1300	1400	1500	1600	1700	1800	1900	Grand total
caso	1	7	17	37	33	12	43	57	207
identity	1	7	17	34	32	12	43	57	203
no				3	1				4
facto							3	97	100
identity							3	97	100

Table 4 (continued)

Nouns / identity relation	1200	1300	1400	1500	1600	1700	1800	1900	Grand total
ideia						8	32	67	107
identity						8	32	66	106
no								1	1
mercê	3	8	23	27	47	3	7	1	119
identity	2	8	22	26	46	3	6	1	114
no	1		1	1	1		1		5
questão						3	9	90	102
identity						3	5	30	38
no							4	60	64
razão	7	17	22	26	36	16	49	51	224
identity	2	6	3	15	17	2	23	19	87
no	5	11	19	11	19	14	26	32	137
sinal	6	2	18	24	41	16	51	49	207
identity	3	2	11	18	22	11	8	32	107
no	3		7	6	19	5	43	17	100
temor		3	30	47	16	17	18	44	175
identity					1	1	2	17	21
no		3	30	47	15	16	16	27	154
vontade		3	51	23	13	15	30	70	205
identity			5	3		1	2		11
no		3	46	20	13	14	28	70	194
identity (sum)	8	23	58	96	118	41	124	319	787
no (sum)	9	17	103	88	68	49	118	207	659
Grand total	17	40	161	184	186	90	242	526	1446

Examples (15) to (25), below, illustrate typical uses of each noun in the century in which its use is the most salient as compared both to the other nouns and to other centuries. In examples (16), (18), (22) and (25), *mercê*, *caso*, *ideia* and *facto* all function as shell nouns, encapsulating the information of the underlined passages. The three multi-use nouns appear once in each usage type, as shells in (19), (20), (24) and not shells in (15), (23), (26), in the century when this use is most salient. In examples (17) and (21), finally, *vontade* and *temor* are used in compound predicate-like expressions where the underlined complement clauses are comparable to direct objects.

- (15) Ca a Sennor que o atan ben dá non á ome **razon de lle furta**r nen de roubar-ll' o seu (CP, Mettman:CantigasSM3)
 'From the Lord who gives it so well a man has no reason to steal it or take it from him.'
- (16) Senhor, **seja vossa mercee que mandees entregar** o seu aos que son com ho Cide. (CP, 13:CIPM:CGEsp)
 'Sire, may it be your honour that you order theirs to be delivered to those who are with the Cide.'
- (17) Danubre, que bem vio que morreria e que **havia vontade de vingar sa morte**. . . (CP, 14:CIPM:Demanda)
 'Danubre, who knew that he would die and wanted to avenge his death...'
- (18) & ser lhe ha tido em segredo,**em caso que elle não queria ser accusador**, para que mais liurementemente o possa descobrir: (CP, 15:Liao:Leis)
 '& it shall be kept secret from him, in case he didn't want to be the accuser, so that he could find out more easily'
- (19) cuidando que se podiam salvar na mesquita, acabaram nela, e assi **era razão que** no lugar onde tinham perdido as almas, dessem sepultura aos corpos. (CP, 15:Barros:Asia2)
 'Thinking that they could save themselves in the mosque, they ended up there, and so it was right that in the place where they had lost their souls, they should bury their bodies.'
- (20) e não se sabe até agora mais que haverem-se ouvido tiros pela madrugada, **sinal de que** foram sentidos. (CP, 16:Vieira:Cartas)
 'and so far no more is known than that shots were heard in the early hours of the morning, a sign that they were felt.'
- (21) se casaria de boa vontade com huma destas damas **sem temor de ser alguma noite degolado** (CP, 17:Macedo:Antidoto)
 'he would willingly marry one of these ladies without fear of being beheaded one night'
- (22) Laura estremeceu de pudor **com a idéia de que em breve estaria totalmente despida** e descomposta (CP, 18:Azevedo:Demônios)
 'Laura shuddered at the thought that she would soon be completely naked and uncomposed.'

- (23) Ela meneou a cabeça afirmativamente, e ele **fez-lhe sinal** de que o esperasse por detrás do cortiço, no capinzal dos fundos. (CP, 18:Azevedo:Cortiço)
‘She nodded, and he made a sign to her to wait for him behind the tene-ment, in the back garden.’
- (24) Venham donde vierem as ideias. A origem pouco importa, **a questão é que elas sejam boas.** (CP, 18:Dinis:Fidalgos)
‘Wherever the ideas come from. It doesn’t matter where they come from, the point is that they should be good.’
- (25) O ascendente que o PC teve no Alentejo deveu-se fundamentalmente **ao facto de o Partido Comunista ter conseguido uma certa mobilização** das massas trabalhadoras (CP, 19Or:Pt:Intrv:Pub)
‘The rise of the CP in the Alentejo was fundamentally due to the fact that the Communist Party managed to mobilise the working masses to a certain extent’
- (26) Esta é a convicção do treinador-adjunto vitoriano, Romeu, que, todavia, **fez questão de salvar a importância** com que todos os jogos são encarados (CP, 19N:Pt:Jornal)
‘This is the conviction of Vitoria’s assistant coach, Romeu, who nevertheless made a point of emphasising the importance of every match.’

Apart from the individual nouns, the formal typicality of the usage context and the construction format also experience changes over time. Consider, first, the figures in Table 5 and the graphs illustrating different facets of these figures in Figures 1a and 1b. As, the figures in Table 5 show, the typical uses show a steady increase over time, reaching approximately a third of all the uses in the 19th and 20th centuries. The less typical uses show comparatively stable frequencies, whereas the frequency of the untypical uses goes down from over 70 per cent until 1699 to approximately a third of the cases in from 1700 on. These figures are directly reflected in the graphs of Figure 1a.

Table 5: Distribution of the three degrees of typicality of the usage context, 1200s to 1900s.

Typicality	1200	1300	1400	1500	1600	1700	1800	1900	Total
typical	0	6/15%	11/7%	11/6%	19/10%	17/20%	74/31%	183/35%	323
less typical	4/22%	5/13%	44/28%	36/20%	36/19%	35/38%	68/28%	173/33%	405
untypical	13/78%	29/73%	106/64%	137/74%	131/70%	38/42%	100/41%	170/32%	725
Total	17	40	161	185	186	90	242	526	1446

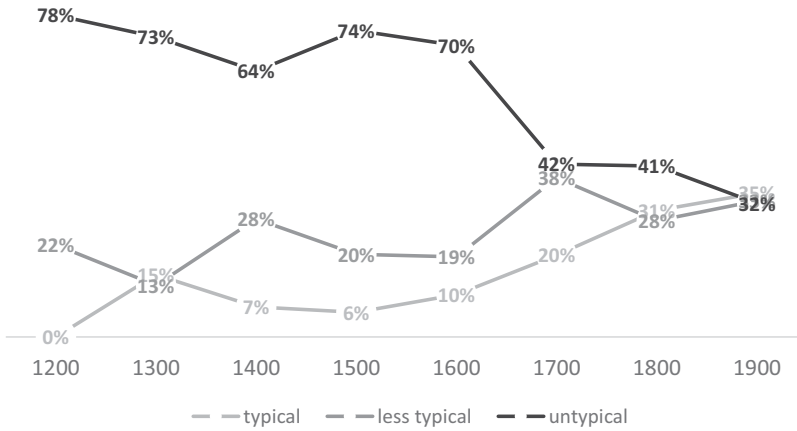


Figure 1a: Distribution of the three degrees of typicality of the usage context, 1200s to 1900s.

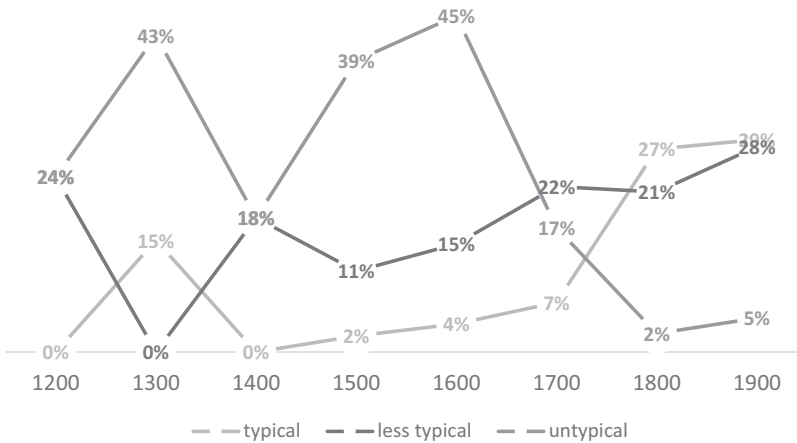


Figure 1b: Proportion of the three degrees of typicality of the usage context of the shell noun uses (involving an identity relation), 1200s to 1900s.

The graphs in Figure 1b, on the other hand, reflect the same tendencies as above, but considering only shell noun uses (with identity relation), illustrating the proportion of the three usage contexts of the cases with an identity relation of all uses (hence the percentages do not add up to 100%). In Figure 1b, the formally typical uses are shown to be of minor importance for the shell noun function all up to the end of the 18th century (0–15 % of all uses); but in the 19th and 20th centuries, the shell noun uses in typical contexts stand for about 30 per cent of all

uses. The line for the untypical uses shows the opposite evolution, with an average of around 40 per cent until 1699 and less than 5 percent from 1800 on. The correlation coefficient between the typical usage context and the shell uses is a convincing 0,95, and for the relationship between the untypical contexts and the non-shell uses it also reaches 0.85. This indicates that the formal usage context is a good predictor of the encapsulating function: when the nouns appear without determiners, in combination with light verbs and as direct objects or prepositional complements they are not used as shell nouns.

Table 6 and Figure 2a portray the distribution of the four construction formats over the centuries, with Figure 2b adding the impact of the construction format on the shell noun uses only. As Table 6 and Figure 2a show, both the infinitive (N *de* + infinitive) and the finite complement clause (N *de que*) construction show an increasing usage frequency over time. The equative construction (N *é que*) seems to maintain its frequency (with the exception of the 20th century), while the finite N *que* construction steadily loses the importance it had in the medieval period. The loss of usage frequency of the N *que* construction format is of course to be expected on the basis of previous studies (see Bogard & Company 1989, Pountain 2014 for Spanish, and Granvik 2015, 2017a, 2017b for Spanish and Portuguese) and must be put in relation to the development of the N *de que* format, which also involves finite complement clauses.

Table 6: Distribution of the four different construction formats, 1200s to 1900s.

Construction	1200	1300	1400	1500	1600	1700	1800	1900	Total
N <i>é que</i>	2/12%	6/15%	17/11%	32/17%	27/15%	1/1%	40/17%	39/7%	163
N <i>de inf</i>	7/41%	23/58%	103/64%	97/53%	92/49%	66/73%	135/56%	360/68%	883
N <i>de que</i>	0	0	0	2/1%	15/8%	6/7%	27/11%	94/18%	144
N <i>que</i>	8/47%	11/28%	41/25%	53/29%	52/28%	17/19%	40/17%	33/6%	255
Total	17	40	161	184	186	90	242	526	1446

When the four construction formats are observed from the perspective of the encapsulating function of the nouns, Figure 2b shows the same trends as above but with some important differences. First, it shows that the equative construction (N *é que*) is, overall, a format where almost all uses are encapsulating ones, with frequencies ranging from 100 to 83 per cent. Overall, 154 out of 163 cases (94%) involve encapsulation; and only *mercê*, *questão* and *razão* are used in this construction format without there being an identity relation. Second, the proportions of N *de que* and N *de* + infinitive construction formats of all shell noun uses also increase over time in Figure 2b. For N *de que* format, the rise starting in the 1600s is much more pronounced as compared to Figure 2a. Third, the trendlines also reveal that N *de* +

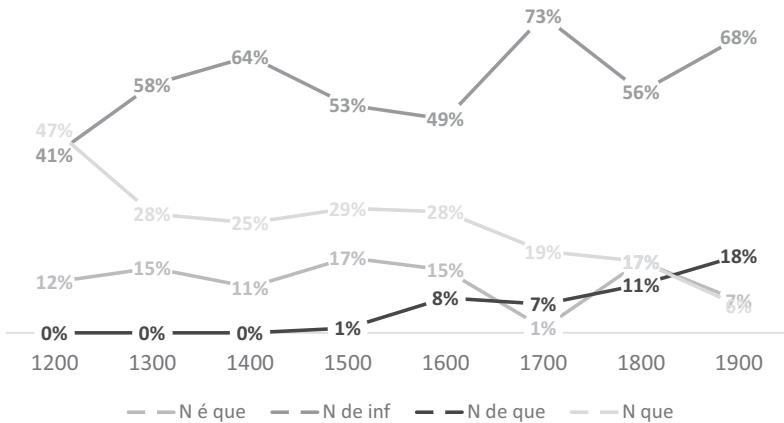


Figure 2a: Distribution of the four different construction formats, 1200s to 1900s.

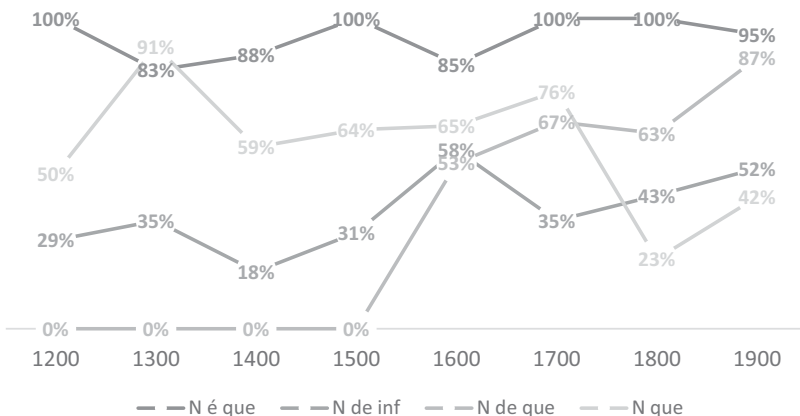


Figure 2b: Proportion of shell uses (involving an identity relation) of the four different construction formats, 1200s to 1900s.

infinitive and the *N que* construction formats run almost parallel until the 17th century; in the 18th century, the *N de + infinitive* format loses some of its importance, only to rise again in the last two centuries, while the descent of *N que* starts in the 19th century.

Figure 2b clearly shows that, overall, *N é que* is a construction format where encapsulation is almost guaranteed. It is clearly the least frequent of the four construction formats, but also the one most specialized on the encapsulating function. This is, of course, a very natural consequence of the construction itself,

which literally equates the noun with its complement clause given the structure *N é que* + clause. The *N de que* format, on the other hand, also shows a growing tendency to involve encapsulating uses of the nouns which culminates in the 20th century. This is due to two things: first, the typical shell nouns *facto*, *ideia* and the multi-use *senal*, often have an encapsulating function when followed by a *de que* clause. Second, almost two thirds of all uses of the *N de que* format are concentrated in the 19th and 20th centuries (see Table 6), when the overall frequency of the encapsulating function reaches its maximum.

The *N de* + infinitive and the *N que* + clause formats, on the other hand, are the least prone to encapsulation. For the *N de* + infinitive format the average of shell noun uses lies below 50 per cent in the whole period analyzed, and 504 out of the 659 cases (74%) of non-encapsulating uses in the data set are in the *N de* + infinitive format. The slight increase that can be observed between the 18th and the 20th centuries (in Figure 2b) is thus best explained by the apparition of the nouns *facto* and *ideia* in the data set, since these nouns are almost uniquely used in an encapsulating function, even when followed by infinitives. The *N que* construction format, finally, is the most frequent of the four construction formats until the 1500s, and from 1600–1700 it is the second most frequent after *N de* + infinitive, but it starts to lose ground with the rise of the other finite construction format, *N de que*, in the 19th century. In the first six centuries, *N que* is more frequent with encapsulating uses than non-encapsulating, but both uses are always attested, and from the 19th century on *N que* is more often used when the nouns do not have an encapsulating function.

4.2 Factors Determining the Encapsulating Function

In this subsection, I present the results of a mixed-effects logistic regression analysis performed to verify which criteria best help determine whether a noun is used in an encapsulating function or not. As we will see, the results align nicely with the tendencies presented in Section 4.1. The dependent variable is the identity relation (see Table 2, above), distinguishing between identification and no identification. The independent variables whereby the model tries to predict shell vs. non-shell uses, are Construction (CXN: *N que*, *N de que*, *N de* + infinitive, *N é* + clause) and Typicality (typical, less typical, untypical, based on the formal values). The Centuries (1200–1900) and the nouns (LEX) are considered as random effects, by which the importance of the other two variables is allowed to vary (see Winter 2020: Ch. 14).

The analysis was performed in R, using the **glmer** function. The numerical results of the regression model are presented in Table 7.¹¹ As the figures in bold face in the rightmost column of Table 7 indicate, the model as a whole is statistically significant (Intercept p value of 0,004) and four of the five independent variables values also stand out as significant. The Odds Ratio values in the second column indicate the direction of the impact, with values below 1 indicating a tendency towards no identification and values above 1 a tendency towards an identity relation. The positive Odds Ratio value (24.85) of the Intercept means that the two variable values which are not shown in the table, namely the Typical usage context and the *N de que* construction format, strongly favor the Identity relation. With an OR value of 14.27, the *N é que* construction format is strongly associated with Identity, i.e. encapsulation. The untypical uses as well as the *N de* + infinitive and the *N que* construction formats, on the other hand, do not favor there being an identity relation. Figure 3 illustrates these tendencies graphically, with the variable values marked in red favoring the absence of an identity relation and the blue ones, especially the one corresponding to the *N é que* construction format, showing predictors for the encapsulating function.

Table 7: Summary of the mixed-effect logistic regression model.

Predictors	Effects on Identity relation		
	Odds Ratios	CI	p
(Intercept)	24.85	2.74 – 225.09	0.004
Less typical	1.13	0.61 – 2.08	0.701
Untypical	0.15	0.08 – 0.27	<0.001
<i>N é que</i>	14.27	5.27 – 38.61	<0.001
<i>N de</i> infinitive	0.24	0.12 – 0.46	<0.001
<i>N que</i>	0.31	0.16 – 0.61	0.001

¹¹ The formula used to create the model was: `ID_REL_NUM ~ TYPICAL_NUM + CXN3 + (1 | LEX) + (1 | SEC)`. The AIC value of the model is 785.3, which can be considered reasonable. The full R script and dataset used to create the tables and graphs presented here can be obtained from the author upon request. Table 7 was created by using the `tab_model()` function and Figures 3 and 4a +b by the `plot_model()` function found in the `sjPlot` package (see http://www.strengejacke.de/sjPlot/reference/plot_model.html).

Table 7 (continued)

Predictors	Effects on Identity relation		
	Odds Ratios	CI	p
Random Effects			
σ^2			3.29
τ_{00} LEX			9.82
ICC			0.75
N_{LEX}			9
Observations			1446
Marginal R^2 / Conditional R^2			0.191 / 0.797

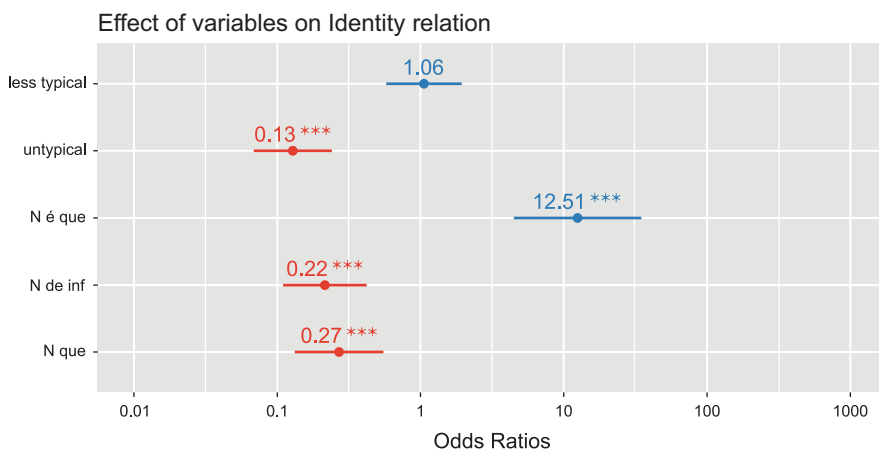


Figure 3: Odds Ratio values indicating the effect of the variables on the Identity relation, i.e. the encapsulating function.

Turning to the random effects, Lexemes and Century, Figures 4a and 4b illustrate similar plots of the effects of the individual lexemes and the centuries on the Identity relation. As Figure 4a shows, there are clear differences between two groups of nouns: those that favor encapsulation, *mercê*, *ideia*, *facto* and *caso*, and those that don't: *vontade*, *temor*, *sinal*, *razão* and *questão*. In agreement with the three-way distinction presented above (§ 4.1), the multi-use nouns, *questão*, *razão* and *sinal* stand out as presenting much higher OR values as compared to non-encapsulating *vontade* and *temor* (i.e. they are much less strongly associated with the non-encapsulating function).

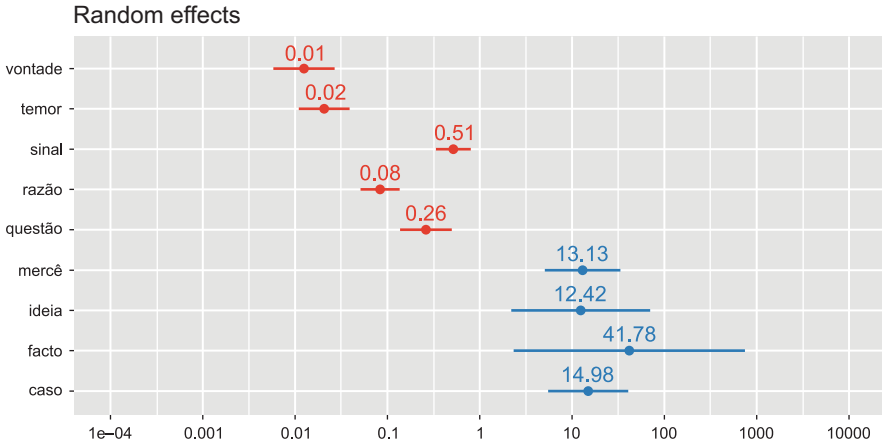


Figure 4a: Odds Ratio values indicating the effect of the different nouns on the Identity relation, i.e. the encapsulating function.

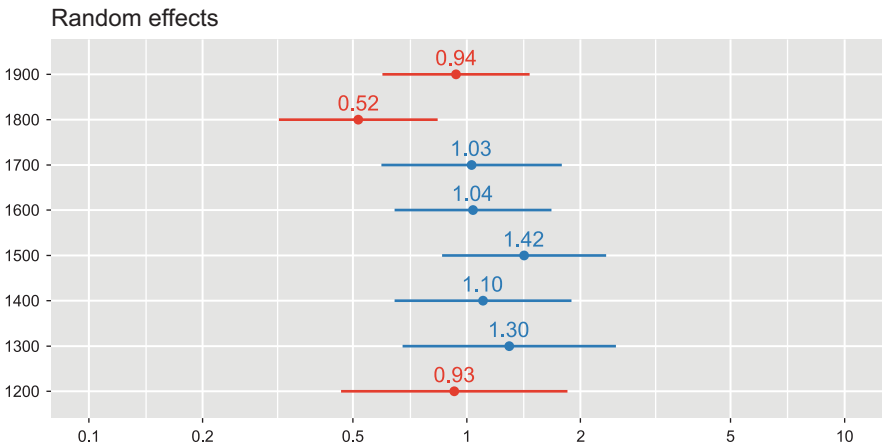


Figure 4b: Odds Ratio values indicating the effect of the different centuries on the Identity relation, i.e. the encapsulating function.

Figure 4b, finally, shows small differences between the centuries and the ranges of the OR values overlap to such a degree that Century as a variable does not seem to be an important predictor for the encapsulating function. This may be explained by the fact that the nouns which are strong predictors are found across the centuries: *caso*, for example, is a strong predictor for the Identity relation and it is used as a shell noun in all centuries. Although *ideia* and *facto* are also strongly associated to uses with an Identity relation, and are nouns which are

only used in the 19th and 20th centuries, the third present-day noun *questão* is strongly associated with non-encapsulating uses. This highlights the fact that, overall, the individual nouns are stronger predictors of the encapsulating uses as compared to the diachronic dimension, at least from the perspective of the nine nouns analysed in this paper.

5 Concluding Remarks

The results of the analysis allow me to answer the research questions in the following fashion. Starting with the last one of them, we have seen that there is a clear correlation between the formal characterisation of the usage context and the shell uses of the nouns: In typical (and less typical) contexts, the nouns are more often used with encapsulating function as compared to untypical contexts. I thus conclude that, for Portuguese, the formal characterisation proposed by Granvik (2019) seems to be a quite reliable proxy for identifying shell-noun uses, understood as uses in which there is an identity relation between the noun and some other element. There is also a (statistically significant and reliable) relationship between some of the constructional formats and the encapsulating function, so that the [N *é* + clause] and the [N *de que* + clause] constructional formats are associated with shell-noun uses. In contrast, when the nouns are inserted in the other two formats, N *que* + clause and N *de* + infinitive the encapsulating function is not much more frequent than a predicative-like function. From this point of view, then, the main encapsulating constructions in Portuguese seem to be [N *é* + clause] and [N *de que* + clause].

Interestingly, although the analysis shows that the shell function becomes more frequent over time (see Table 4, above) the diachronic dimension (centuries) does not stand out as a significant predictor in the regression analysis. Instead, the nouns and the usage context are more important, and the main reason for the increase in frequency of the shell-uses in the 20th century are *facto* and *ideia*, which are almost exclusively used as shells. In fact, as part of the analysis, the nouns were divided into three groups: typical shell nouns (*caso*, *mercê*, *ideia* and *facto*), multi-use nouns: *questão*, *razão* and *sinal*, and the non-encapsulating nouns *temor* and *vontade*. However, it is worth remembering that even when an untypical shell noun such as *questão* is used in a context classified as non-encapsulating (as in (27)), there is often a nuance of encapsulation, marginal though it may be, which reminds us that the construction formats included in this analysis are, really, part of a broader shell-noun construction.

- (27) Ela **fizera questão de** manter sempre trancada a porta que separava os dois cômodos (CP, 19:Fic:Br:Penna:Menina)
 ‘She made a point of always keeping the door that separated the two rooms locked’

In (27), then, the relationship between *questão* and the complement clause is ambiguous. On the one hand, the *question* or *point* is ‘to keep the door locked’. On the other, keeping it locked is what is aimed at, it functions as a sort of direct object, the objective, of *his or her* action, where *fazer questão de* corresponds to ‘making a point’ of achieving something.

From a historical perspective, it is important to notice that the encapsulating uses appear already in the 13th and 14th centuries, with *mercê*, *caso*, *razão*, *sinal*. This is thus not a modern innovation which can be associated with informatically dense texts, as Borreguero (2018) argues. But future work will need to address in more detail which nouns are indeed used as encapsulators in the medieval period and in which text types. A broader understanding of the origins of the shell-noun construction in Portuguese would also need to consider other shell-noun constructions, i.e. demonstrative pronoun + N and (Schmid 2000; Rodríguez Espiñeira 2025), which were not considered here, as well as other salient nouns in order to complete the picture.

Bibliography

- Abad Serna, Silvia (2015): *Estudio contrastivo del funcionamiento semántico de los encapsuladores nominales en la prensa española y alemana: de la anáfora a la catáfora conceptual*. Doctoral dissertation. Universidad Autónoma de Madrid. <<https://repositorio.uam.es/handle/10486/669678>>.
- Bogard, Sergio and Concepción Company Company (1989): “Estructura y evolución de las oraciones completivas de sustantivo en el español”, in *Romance Philology*, XLIII (2), pp. 258–273.
- Borreguero Zuloaga, Margarita (2018): “Los encapsuladores anafóricos: una propuesta de clasificación”, in *Caplletra*, 64, pp. 179–203. DOI: 10.7203/Caplletra.64.11380.
- Borreguero Zuloaga, Margarita and Álvaro S. Octavio de Toledo y Huerta (2007): “Presencia y función de los encapsuladores en las crónicas periodísticas del s. XVII”, in *Philologia Hispalensis*, 21, pp. 125–159.
- Croft, William and Alan D. Cruse (2004): *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Davies, Mark and Michael Ferreira (2006-): *Corpus do Português: 45 million words, 1300s-1900s*. Online at <<https://www.corpusdoportugues.org/hist-gen/>>.
- Flowerdew, John and Richard W. Forest (2015): *Signalling nouns in English. A Corpus-Based Discourse Approach*. Cambridge: Cambridge University Press. <<https://doi.org/10.1017/CBO9781139135405>>.

- González-García, Francisco (2012): “La(s) gramática(s) de construcciones”, in Iraide Ibarretxe-Antuñano and Javier Valenzuela (dirs.). *Lingsüística cognitiva*. Barcelona: Anthropos, pp. 249–280.
- Granvik, Anton (2014): “Hablando *de, sobre y acerca de* la gramaticalización y la lexicalización: panorama diacrónico de las relaciones entre preposiciones y locuciones prepositivas dentro del campo semántico de tema/asunto”, in José Luis Girón Alconchel and Daniel M. Sáez de Rivera (eds.). *Procesos de gramaticalización en la historia del español*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 77–117. <<https://doi.org/10.31819/9783954871988-006>>.
- Granvik, Anton (2015): “Oraciones completivas de sustantivo: un análisis contrastivo entre español y portugués”, in *Verba*, 42, pp. 347–401.
- Granvik, Anton (2017a): “Oraciones completivas de sustantivo en español y portugués: ¿infinitivo y oración finita?”, in *Cuadernos de Lingüística de El Colegio de México*, 4 (1), pp. 103–180. DOI: <<http://dx.doi.org/10.24201/clecm.v4i1.54>>.
- Granvik, Anton (2017b): “Análisis histórico-comparativo de las oraciones completivas de sustantivo en español y portugués: nacimiento y evolución de una alternancia sintáctica”, in *Neuphilologische Mitteilungen*, I, CXVIII (2017), pp. 31–63.
- Granvik, Anton (2019): “Sobre los orígenes de la construcción encapsuladora en español”, in Marta Blanco, Helia Olbertz and Victoria Vázquez Rozas (eds.). *Corpus y construcciones. Perspectivas hispánicas*. Anexo 79 de *Verba. Anuario galego de filoloxía*, 2019, pp. 41–79. DOI: <<https://dx.doi.org/10.15304/9788417595876>>.
- Gries, Stefan Th. (2013): *Statistics for Linguistics with R. A practical introduction*. Berlin/New York: De Gruyter Mouton.
- Guimarães, Igor Caixeta Trindade (2011): *A formação nominal em português: um estudo sintáticosemântico de bases enunciativas*. MA thesis (Dissertação de mestría), UFMG. Belo Horizonte.
- Hilpert, Martin (2019): *Construction Grammar and its application to English*. 2nd ed. Edinburgh: Edinburgh University Press.
- Leonetti, Manuel (1993): “Dos tipos de completivas en sintagmas nominales”, in *Lingüística (ALFAL)*, 5, pp. 5–40.
- Leonetti, Manuel (1999): “La subordinación sustantiva: las subordinadas enunciativas en los complementos nominales”, in Ignacio Bosque and Violeta Demonte (coords.). *Gramática descriptiva de la lengua española*. 3 vols. Madrid: Espasa Calpe, pp. 2083–2104.
- López Samaniego, Anna (2011): *La categorización de entidades del discurso en la escritura profesional: las etiquetas discursivas como mecanismo de cohesión léxica*. Doctoral dissertation. Universidad de Barcelona. <<http://hdl.handle.net/10803/48757>>.
- Paredes Pereida da Silva, Vera Lúcia and Gabrieli Pereira Bezerra (2013): “SNS como rótulos em livros didáticos de história do Brasil: simples ou complexos?”, in *Revista diacrítica*, 27 (1), pp. 217–237.
- Pountain, Christopher (2014): “Preposición + que en español”, in *Cuadernos de Lingüística del Colegio de México*, 2. Estudios de cambio y variación, pp. 9–54.
- Ribeiro, Dayhane Alves Escobar (2010): “Levantamento lexical dos encapsuladores utilizados nas redações de alunos do pré-vestibular”, in *Cadernos do CNLF*, XIV (4, tomo 3), pp. 2781–2808.
- Rodríguez Espiñeira, María José (2015): “El sustantivo *hecho* como ejemplar de nombre encapsulador factual”, in *Studium grammaticae: homenaje al profesor José A. Martínez*. Oviedo: Universidad de Oviedo, pp. 655–674.
- Sanches Duran, Magali and Maria das Graças Volpe Nunes (2023): “Aposições anafóricas e catafóricas no português e sua anotação no esquema Universal Dependencies”, in *Anais do*

- Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. DOI: 10.5753/stil.2023.233699.
- Schmid, Hans-Jörg (1997): “Constant and ephemeral hypostatization: *thing*, *problem* and other “shell nouns”, in Bernhard Caron (ed.). *Proceedings of the 16th International Conference of Linguistics* (Paris, July 20–25 1997). Elsevier.
- Schmid, Hans-Jörg (1999): “Cognitive effect of shell nouns”, in Karen Van Hoek, Andrej A. Kibrik and Leo Noordman (eds.). *Discourse Studies in Cognitive Linguistics: Selected papers from the 5th International Cognitive Linguistics Conference, Amsterdam, July 1997*. Amsterdam/Philadelphia: John Benjamins, pp. 111–132. <<https://doi.org/10.1075/cilt.176.09sch>>.
- Schmid, Hans-Jörg (2000): *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin/ New York: Mouton de Gruyter. <<https://doi.org/10.1515/9783110808704>>.
- Schmid, Hans-Jörg (2018): “Shell nouns in English – a personal round-up”, in *Caplletra*, 64, pp. 109–128. DOI: 10.7203/Caplletra.64.11368.
- Schmid, Hans-Jörg and Annette Mantlik (2015): “Entrenchment in Historical Corpora? Reconstructing Dead Authors’ Minds from their Usage Profiles”, in *Anglia*, 133(4), pp. 583–623. DOI 10.1515/anglia-2015-0056.
- Schmid, Hans-Jörg and Helmut Küchenhoff (2013): “Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings”, in *Cognitive Linguistics*, 24(3), pp. 531–577.
- Stefanowitsch, Anatol and Stefan Gries (2003): “Collostructions: Investigating the interactions of words and constructions”, in *International Journal of Corpus Linguistics*, 8:2, pp. 209–243.
- Traugott, Elisabeth Closs and Graeme Trousdale (2013): *Constructionalization and constructional changes*. Oxford: Oxford University Press. <<https://doi.org/10.1093/acprof:oso/9780199679898.001.0001>>.
- Winter, Bodo (2020): *Statistics for linguists. An introduction using R*. New York/London: Routledge.

Katharina Gerhalter

Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio. A Corpus-Based Approach to Topicalized Infinitives in Portuguese

1 Introduction

This paper addresses the topicalized infinitive construction in Portuguese. This construction comprises a fronted infinitive that announces the topic of the sentence, followed by an inflected verb form of the same lemma in the comment (cf. Bastos 2001; Bastos-Gee 2009; Reich 2011; Hein 2020; Andrade 2020: 110–112). The following example contains two instances of topicalized infinitives (=TI), namely *escrever* ('as for writing') and *ler* ('as for reading'):

- (1) Se não fosse a intuição não falava línguas como falo sem nunca ter aprendido nada. Inglês, francês, espanhol, italiano. *Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio.* (Portugal, CdP, 19:N:Pt:Expr)

In contrast to most previous studies on topicalized infinitives, which primarily relied on isolated sentences obtained from anecdotal evidence and introspection, our study is based on corpus data. The examples were compiled from a large corpus: the 20th-century section of the *Corpus do Português: Genre / Historical* (=CdP). We will discuss mainly quantitative data that aims to answer the following questions: In which of the four genres established in CdP (fiction, newspaper, academic texts, and spoken language) is the TI-construction most prevalent? Does the data indicate any differences between Brazilian Portuguese (= BP) and European Portuguese (= EP)? Furthermore, corpus data allows to examine the question of whether the TI-construction can take any verb, i.e., whether it is a productive pattern.

As observed by Andrade (2020: 101), the contextual properties of topic constructions have not yet been discussed in depth on the basis of real spoken data so far. This study represents a first attempt to fill this gap. Our objective is to address the following question based on the context in which TI-constructions are produced: Do TI reply to a verb that is already present in the previous context, or can they also introduce completely new verbs into the discourse? In order to analyze the function of this construction in discourse, we will analyze TI as explicit instantiations of the immediate *question under discussion* (= QUD), as suggested by Muñoz Pérez & Verdecchia (2022).

2 Theoretical Background: Topicalized Infinitives

The topic-comment partition encodes the information structure of a sentence. The topic announces what the comment is about and defines the frame within which the information uttered in the comment is valid and true (cf. Duarte 2013; Leonetti 2014; Leonetti & Escandell Vidal 2021: 77). Syntactically marked topics are placed in a detached position, such as hanging topics, left and right dislocations, and topicalizations of verbal projections (Andrade 2020: 101), which we refer to as TI. The majority of topics are nominals, referring to stable entities. In contrast, predicates or bare infinitives are much less frequent as topics and, as such, refer to a state of affairs that can be activated and recovered from the context (Maslova & Bernini 2006: 80–81).¹

The pattern under study has been variously designated in the literature as *topicalization of verbal projections* (Bastos-Gee 2009; Andrade 2020: 110–112), *topicalization of verbal constituents* (Bastos 2001), *predicate doubling* (Verdecchia 2021; Muñoz Pérez & Verdecchia 2022), *predicates as topics* or *infinitival topic expressions* (Maslova & Bernini 2006: 81–83), *fronted infinitives* (Reich 2011), *bare infinitive and predicate clefts* (Vicente 2007), *topic constructions* in which the topic is a sentence (Leitão Vasco 2006), and *vP-topicalization* (Saab 2017).

In this paper, we adopt a constructionist approach, in line with the definition of constructions as linguistic patterns that consist of conventionalized form and meaning pairings (cf. Goldberg 2006: 3–5; Hoffmann & Trousdale 2013). In this sense, the TI-construction has a conventionalized form: the same verb appears in both the topic (infinitive) and as part of the comment (inflected verb form).² This is represented by the following structure:

[_{TOPIC} infinitive_i] + [_{COMMENT} inflected verb form_i]

The infinitive may be combined with its inner objects (therefore some authors speak of topicalization of *predicates*, see above). As illustrated by example (1)

¹ Verb doubling-patterns are documented in numerous languages worldwide and do not always indicate a topic-reading. For an overview of the possible patterns observed in 45 languages, see Hein (2020); for examples of predicates as topics in several European languages, see Maslova & Bernini (2006: 80–83).

² For a formal explication of the requirement of lexical identity between the TI and the inflected verb form in the comment, we refer to the analysis proposed by Bastos (2001), Vicente (2007), and Muñoz Pérez & Verdecchia (2022). In essence, the comment must align with the previously announced topic. In order to express an assertion of the previously announced topic, the inflected verb form must be identical to the TI.

cited in the introduction, the second topic contains not only the TI *ler* ‘to read’ but also its direct object:

[_{TOPIC} Escrever] [_{COMMENT} não escrevo], mas [_{TOPIC} ler um livro, ou um jornal, uns versos], [_{COMMENT} leio].

The comment that follows a TI typically conveys a simple assertion or negation of the state of affairs described in the topic (Maslova & Bernini 2006: 81; Reich 2011). Therefore, the speaker emphatically asserts the veracity of the proposition (Vicente 2007: 64). In the aforementioned example, indeed, the first comment is merely a negation (*não escrevo* ‘I do not write’), whereas the second comment is a simple assertion (*leio* ‘I read’).

Verdecchia (2021) and Muñoz Pérez & Verdecchia (2022) analyze TI as contrastive topics in the sense of Büring (2003). Contrastive topics break down the implicit or explicit *question under discussion* (= QUD) at a given point in the conversation into alternative subquestions (see also Frascarelli 2017: 473–474). More precisely, TI introduce the answer to an immediate QUD, which is mostly a polar question. At the same time, they point to implicit alternative QUDs (Verdecchia 2021; Muñoz Pérez & Verdecchia 2022). In this sense, TI are contrastive topics that introduce an answer that only partially satisfies the contextual question; they indicate that the speaker employs a strategy to answer only part of a more ample and complex question (Andrade 2020: 111; see also Silva 2019).

This “missing part” is frequently addressed in a subsequent adversative sentence that expresses a contradiction or objection to the proposition affirmed in the TI-construction.³ This construction thus triggers a so-called *continuation effect* (Verdecchia 2021: 8). As illustrated in example (1), both sentences containing a TI-construction are linked and contrasted via the adversative conjunction *mas* ‘but’. In the absence of an explicit adversative sentence following the TI-construction, it can be inferred from the context. In this case, the TI-construction triggers adversativity as a conversational implicature (cf. Bastos 2001: chapter 2.3; Vicente 2007: 65–68; Reich 2011).

In order to fully account for the function of TI-constructions, it is necessary to analyze such sentences in their discourse context since they are most often used as replies or objections to a previous statement (Bechara 2009: 639–640; see

³ In a specific pattern analyzed by Valenzuela, Hilferty, & Garachana-Camarero (2005), in Spanish, this adversative continuation is even considered an integral part of the construction. As we will briefly discuss in section 5, not all corpus examples show such a continuation. Therefore, we consider only the pattern [_{TOPIC} infinitive_i] + [_{COMMENT} inflected verb form_i] a conventionalized construction.

also Narbona 2015: 57, 168–169 on Spanish). This paper will determine whether the TI tends to repeat a verb that has been mentioned previously or whether it can also introduce a new lexical unit into the discourse (see section 4.3).

Finally, TI-constructions seem more natural and frequent in spoken language, particularly in colloquial conversations (for Spanish, see Vicente 2007: 3, 68 and Narbona 2015: 57). It can therefore be reasonably assumed that the TI-construction is more common in more informal subcorpora, particularly in the oral subcorpora of CdP (see section 4.1).

3 Corpus and Methodology

The search for TI-constructions in any corpus presents two principal challenges:

- (i) **Low frequency:** the majority of topicalizations concern nominal elements, such as noun phrases, nouns, adjectives, or prepositional phrases (e.g., Duarte 2013 only discusses nominal topics). In comparison to nominal categories, topicalized infinitives or predicates are relatively uncommon. For instance, the study on spoken BP by Leitão Vasco (2006) reveals that of a total of 1,292 instances of topicalizations, only 14 instances (1.08%) contain an infinitive followed by the same verb. Additionally, 10 instances (0.77%) contain a verb in the topic, but without repetition in the comment (Leitão Vasco 2006: 142–143). Moreover, in the spoken Spanish corpora studied by Hidalgo Downing (1999), only 0.6% to 2.4% of all topicalizations contain an infinitive, whereas the vast majority of topicalizations consist of nominal elements (Hidalgo Downing 2003: 199).
- (ii) **Schematicity:** The form of a construction can be either entirely schematic or partially lexically filled (Goldberg 2006: 3–5; Hoffmann & Trousdale 2013). TI-constructions are completely schematic and, presumably, all verbs can fill in this pattern (as we will discuss in section 4.2). In between the topicalized infinitive and the inflected verb form of the same lemma, there may be several other words. This makes it difficult to define a search string that precisely addresses these patterns.

To counter the first challenge, namely the documentation of infrequent phenomena, a large corpus is needed. Concurrently, to meet the second challenge, this corpus must be PoS-tagged and lemmatized in order to enable the creation of search strings that differentiate between inflected verbs and infinitives.

The CdP (section Historical/Genre) is a fairly large corpus, comprising a total of 45 million words, including historical data. This paper focuses on the data from the

20th century, which still offers a considerable size of approximately 20 million words (ca. 10 million for BP and ca. 10 million for EP). The 20th-century data is further divided into four genres: there is a more or less equal number of words from academic texts, fiction texts, and newspapers, compared to a smaller dataset from spoken language (see Table 1).

Table 1: Distribution of sources in CdP Historical/Genre (only 1900s).

Country	Number of words	Genre
Portugal	3,087,052	Academic
	3,271,328	Newspaper
	3,048,020	Fiction
	1,100,303	Spoken
Brazil	2,816,802	Academic
	3,346,988	Newspaper
	3,028,646	Fiction
	1,078,586	Spoken
total	20,777,725	

It should be noted that the classification into genres was done for each text as a whole; there is no differentiation between different genres within a single source. This approach may result in some erroneous categorizations. For instance, one example in our sample was classified as “academic”, yet it is a transcript of an interview with a patient within a scientific article on medicine. Therefore, in our sample, this example was included in the “spoken” section.

CdP compiled also sources from other corpora (Mendes 2016: 232, 234). Of particular relevance to our purposes are the spoken subcorpora, as the authors did not transcribe their own corpus of spoken language but selected previously published corpora. The spoken subcorpora comprise transcripts from radio and television shows, representing approximately half of the spoken data for both countries. The other half of the data was taken from conversational corpora. For Portugal, these are the *Corpus Dialetal para o Estudo da Sintaxe* (= CORDIAL-SIN) and the *Corpus de Referência do Português Contemporâneo* (= CRPC). For Brazil, these corpora are *A linguagem falada culta na cidade de São Paulo* and *A linguagem falada culta na cidade de Recife*.

Most importantly for our study, the majority of examples of TI-construction in the spoken CdP-subcorpora originate from CORDIAL-SIN, as well as from the spoken parts of CRPC, two POS-tagged and lemmatized corpora based on the transcription of spontaneous and semi-directed speech collected in fieldwork. In our dataset, we double-checked all these examples in the original corpora. This re-

sulted in the exclusion of certain examples from CORDIAL-SIN, as the transcription of infinitives in CdP was occasionally found to be misleading. CdP presents a single linear transcription, which fails to accurately reflect the additional layer of normalized transcriptions present in CORDIAL-SIN. This includes instances of truncations, repetitions, and hesitation phenomena, which are marked as such in the original corpus. While a large corpus such as CdP offers a substantial quantity of data, this data is not as meticulously tagged as that of smaller corpora such as CORDIAL-SIN, with manually revised annotations and more detailed, multilayered tagging. Consequently, the combination of both corpora balances quantity and quality (for further discussion on different Portuguese corpora, refer to Mendes 2016; for further discussions on the reliability of large corpora, refer to Calderón Campos 2024).

The search queries were based on the list of the 200 most frequent verbs in CdP. For each verb, six different patterns were searched, distinguishing between inflected verbs that immediately follow the infinitive and inflected verbs that occur up to five words after the infinitive. This is exemplified by the verb *chorar* ‘to cry’ in Table 2.

Table 2: Search queries for TI-construction in CdP.⁴

Search string	Possible examples
CHORAR_vr* CHORAR	<i>Chorar chorei muito</i>
CHORAR_vr* * CHORAR	<i>Lá chorar, ele chorou</i>
CHORAR_vr* * * CHORAR	<i>Mas chorar ela não chorava nunca</i>
CHORAR_vr* * * * CHORAR	<i>Chorar, não gosto de chorar</i>
CHORAR_vr* * * * * CHORAR	etc.
CHORAR_vr* * * * * * CHORAR	etc.

Nevertheless, the search queries yielded a considerable number of irrelevant examples, necessitating the manual filtering of the results. Some examples contained two instances of TI, as evidenced by example (1) cited above (*escrever* ‘as for writing’ and *ler* ‘as for reading’). In such instances, each TI-construction was counted separately in the sample.

In total, our compilation of TI-constructions in the 20th-century section of CdP results in a dataset of 60 occurrences.⁵ This is a rather modest result when consid-

⁴ Capitalized letters are used to search for lemmas, the asterisk substitutes any word, and *_vr** is a PoS-marking that specifies that the preceding lemma should be a verb in its infinitive form.

⁵ The search query was conducted in the whole diachronic corpus, leading to a huge dataset of more than 130 examples. The examples from earlier centuries will be analyzed in a separate paper on the diachrony of the TI-construction.

ering the size of CdP (approx. 20 million words), which lends support to the relatively low usage rate of TI (relative frequency: 2.89 occurrences per one million words). It is important to note that our search results are constrained by the limitations of the search query, which could not retrieve all potential instances of TI-constructions. If the list of searched verbs were expanded to include more than 200 lemmas and if the distance between the infinitive and the inflected verb form were increased to more than five words, additional examples would be identified. Indeed, in the second TI-construction in example (1), *ler um livro, ou um jornal, uns versos, leio*, there are seven words between the infinitive and the inflected verb form. This example was only discovered by chance, as it occurred after the TI-construction *escrever não escrevo*, found via the search string `ESCREVER_vr* * ES-CREVER`. Thus, our results should be regarded as a compromise between exhaustiveness and the feasibility of manual filtering.

4 Results

4.1 Genre and Country

The results across the four different genres corroborate the hypothesis that the TI-construction is more frequent in spoken language (see Figure 1).

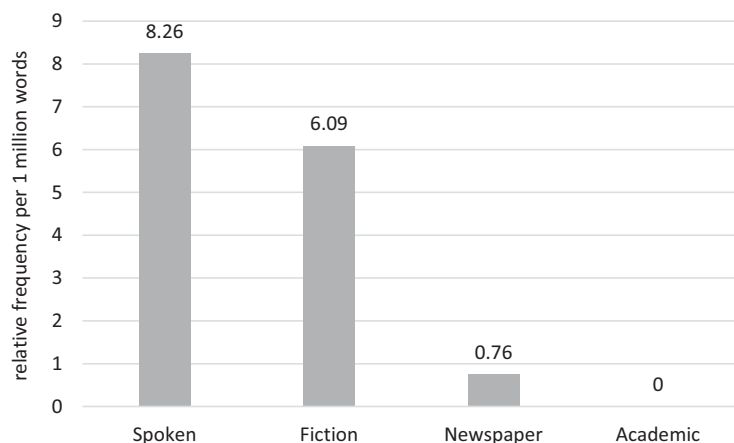


Figure 1: Relative frequency of TI-constructions by genre (CdP, 1900s).

The TI-construction is most prevalent in the spoken sub-corpus, which is presumed to represent the most informal and spontaneous registers. In the second place, they appear in fiction, which encompasses literary texts such as theater plays or novels that may include direct speech imitating conversations. In contrast, the TI-construction is notably less frequent in more formal genres, such as newspapers, and is entirely absent in academic texts.

The division of the CdP into four genres does not differentiate between different text parts within the same source. In order to ascertain the extent to which TI-constructions are present in dialogues, our sample was further manually categorized. Of the 60 examples identified, 33 (55%) were found in dialogues, including transcripts of spoken conversations, interviews, and reported speech on the radio or newspaper, as well as fictitious dialogues in novels or theatre plays (see, e.g., examples 3 and 4 in the next section). Twelve examples (20%) are found in interior monologues, i.e., in self-conversations (see, e.g., examples 1 and 2). Finally, fifteen examples (25%) are found in narrative, descriptive passages in novels and newspapers (see, e.g., example 11 in section 4.4). This demonstrates the versatility of the TI-construction, which is not solely employed as a reply in conversations (its primary function) but can also be used to mark the topic of a sentence in narrative descriptions.

Unexpectedly, the TI-construction is documented twice as frequently in Portugal (3.81 occurrences per one million words) than in Brazil (1.95 occurrences per one million words). There are three different explanations for that:

- i. First, this discrepancy may reflect a diatopic variation that has not been previously documented in the literature.⁶
- ii. Second, it is possible that our methodology (the search string for filtering relevant examples) is insufficient and produces random results. This would mean that the 60 examples found in our extensive search in a corpus of 10 million words for EP and 10 million words for BP were not sufficient to be representative.
- iii. Third, this remarkable difference might also be due to the corpus design of CdP itself. This would imply that the sources selected to represent EP and BP

⁶ The corpus study by Leitão Vasco (2006) on spoken BP includes a comparison with EP regarding different types of topicalizations. For example, he notes a higher frequency of dislocated subject topics in BP due to the progressive loss of the null subject parameter (Leitão Vasco 2006: 201). However, the frequency of TI-constructions is not contrasted. Furthermore, Andrade (2020: 117–122) contrasts the acceptability of certain topic-constructions that are acceptable in EP but not in BP, but TI are not considered there. Finally, the work by Bastos (2001) on TI is dedicated explicitly to BP rather than EP. A comparison between these two varieties addresses the acceptability of certain complex sentences within TI-constructions (*Visitar os amigos, que a Maria visita é evidente*), which are deemed acceptable in EP but not in BP (Bastos 2001: 77). As her work is not based on corpus data, it is not possible to ascertain the frequency of these constructions.

may be quantitatively equivalent in terms of total words, but not qualitatively comparable. For instance, a smaller proportion of spontaneous colloquial sources from Brazil compared to Portugal within the spoken subcorpora could have distorted our results.

Now, if we split the data into both genres and countries, there are, indeed, striking differences between Portugal and Brazil regarding the spoken data (see Figure 2 and Table 3).

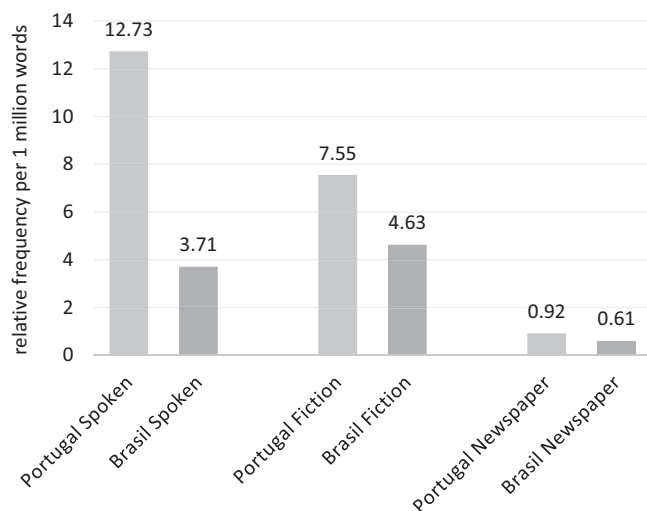


Figure 2: Relative frequency of TI-construction per genre and per country (CdP, 1900s).

Table 3: Distribution of examples (country and genre), absolute and relative frequency (CdP, 1900s).

Country	Genre	Absolute frequency	Relative frequency (per 1 million words)
Portugal	Academic	0	0
	Newspaper	3	0.92
	Fiction	23	7.55
	Spoken	14	12.73
Brazil	Academic	0	0
	Newspaper	2	0.61
	Fiction	14	4.63
	Spoken	4	3.71

In spoken EP, the relative frequency of TI-constructions is as high as 12.73 occurrences per one million words, whereas in the spoken BP corpus, it is considerably lower (3.71). In the Brazilian corpus, TI are mostly documented in the genre “fiction” (4.63), but within this subcorpus, the majority of examples are from dialogues (reported speech) and monologues. This observation lends support to the third explanation previously proposed: the collection of sources for spoken EP and spoken BP in CdP, without a meticulous selection of not only quantitatively but also qualitatively comparable sources, may result in a corpus that is not entirely representative of diatopic variation in spontaneous spoken Portuguese. In the case of rather infrequent phenomena such as TI-constructions, the search results in these subcorpora may, to a certain extent, be random. This discrepancy is somewhat unexpected, given that the total word count of the oral subcorpora of CdP is equivalent for Brazil and Portugal: for both countries, approximately half of the data comes from television and radio, and half from spoken, conversational corpora (see section 3).

The 14 examples of the subcorpus Portugal/Spoken (see Table 3) are distributed as follows: 11 examples from CORDIAL-SIN, 2 examples from CRPC, and only one example from radio or television programs. This further confirms that the TI-construction is more frequent in spontaneous, unplanned conversations. In contrast, radio and television programs represent a public and therefore more formal, more planned, and more distant orality. These sources are oral in terms of the channel or medium, but conceptually some radio and television shows are more “written” in the sense of Koch & Oesterreicher (2011).

In the Brazilian spoken subcorpus, however, all four spoken examples (see Table 3) originate from radio and television shows or published interviews. Our search query did not yield any results from the conversational corpora of São Paulo and Recife (i.e., the sources that, according to the data from Portugal, are expected to contain the majority of spoken examples).

A closer look at the conversational sources included in CdP may offer an explanation: the EP conversational subcorpus (from CORDIAL-SIN and CRPC) includes transcriptions from 1108 different sources/conversations (and therefore different speakers/informants), whereas the BP conversational subcorpus contains longer transcriptions from only 51 different sources (i.e., data from much fewer speakers/informants from São Paulo and Recife). In summary, the data retrieved from spoken corpora for EP and BP are comparable in terms of quantity (total count of words), but less so in terms of quality. The spoken data from Portugal is more balanced and therefore probably more representative than the spoken data from Brazil.

Focusing on the three written subcorpora, the data in Figure 2 and Table 3 still confirm that TI-constructions are slightly more frequent in EP than in BP.

However, in the case of the “fiction” subcorpora, the EP subcorpus is, again, more balanced than the BP subcorpus, since it contains about twice as many different sources (i.e., novels, plays, etc.).

In light of the above, the use of TI-constructions may be regarded as an idiosyncratic trait, with some speakers employing it on occasion, while others do not use it at all. To give an example of idiosyncratic style, the 37 examples of TI-construction in “fiction” (EP and BP) were produced by 26 different authors. Six of these authors produced two instances of TI-constructions, whereas the Portuguese writer Alfonso Ribeiro produced even six instances of TI-constructions. This individual author’s relative frequency is as high as 156.92 TI-constructions per one million words.

Consequently, in a more diverse and balanced corpus, TI-constructions are more likely to be documented and quantitative distortions due to idiosyncratic style are better compensated. Given the relatively small data set (60 examples), we believe that a possible diatopic variation –as indicated by our data– needs to be further confirmed in comparable, balanced corpora of EP and BP colloquial speech.

4.2 Productivity

The 60 examples (or tokens) of TI of our compiled sample correspond to as many as 44 different verbs, i.e., 44 different types. Only a few verbs occur several times, e.g., *falar* ‘to speak’ (5 tokens), *poder* ‘to be able/possible’ (5 tokens), *trabalhar* ‘to work’ (3 tokens), and *dormir* ‘to sleep’ (3 tokens). In contrast, most types –37 verbs– occur only once, e.g., *ajoelhar* ‘to kneel’, *destruir* ‘to destroy’, *garantir* ‘to guarantee’, *mamar* ‘to suck’, *merecer* ‘to deserve’, *perceber* ‘to realize’, or *tentar* ‘to try’ (one token each). This means that 84% of the types (37 out of 44 verbs) and 62% of the tokens (37 out of 60 examples) in our dataset are so-called *hapax legomena*, which is a relatively high rate indicating a relatively high level of productivity.⁷

⁷ The hapax/type ratio is considered an indicator for measuring the productivity of a construction: the higher this ratio, the more productive a construction is compared to other constructions (see Baayen 2008; van Wetteere 2021: 405; van Wetteere 2022: 169). This ratio is quite stable in large corpora (van Wetteere 2022). In this study, we have no data to compare the hapax/type ratio of the TI-construction (0.84) with that of other marked topicalization constructions in the same corpus; we can only state that 0.84 is a relatively high value. To give just an example, the hapax/type ratio of the French semi-copula-construction “*tomber* ‘to fall’ + adjective” is 0.74, a value that van Wetteere (2021: 410) considers to be relatively high.

This shows that the TI-construction, despite being a rather rare phenomenon, is perfectly productive in the sense that it tends to be filled by a wide range of different types (verbs) rather than by a small set of highly frequent types. In our sample, there is no single type/verb (or a small group) that stands out as being extremely frequent (this would be an indication of conventionalization rather than productivity, see, e.g., van Wetteere 2021: 398).

Some lexical-semantic restrictions have been discussed in the literature, though. According to Bastos (2001: 40–43) copula verbs such as *ser*, *estar* or *parecer* are not acceptable or highly questionable in the TI-construction, at least in BP (e.g., **Ser, o João é inteligente* ‘As for being, John is intelligent.’). Such sentences are judged unacceptable because they do not refer to an event. On the contrary, Vicente (2007: 62–63) judges that in Spanish there are no semantic restrictions on the type of predicate that can be topicalized: even verbs such as the Sp. copula verb *ser* and the Sp. existential verb *haber* can be topicalized because they have “enough lexical content to constitute a well-formed topic” (Vicente 2007: 73). The only exception are auxiliary verbs since they lack a lexical content and therefore do not make felicitous topics.

As our sample shows, copula verbs like *ser* ‘to be’ can indeed be topicalized in Portuguese, too. Note that example (2) does not come from the 20th-century dataset analyzed in this paper, but from the historical dataset of BP. It is a monolog of a theater play:

- (2) “Artigo duzentos. Toda pessoa real que, esquecendo o decoro que deve a si própria e ao povo, der escândalo público, será julgada por um Conselho composto de quatro Ministros de estado, e, averiguado o delito, condenada a pena última”. Se se pudesse sofismar este maldito artigo duzentos! Vejamos por partes: [. . .] “que esquecendo o decoro que deve a si própria e ao povo. . .” Disto se esqueceu ela. Comeu queijo! “der escândalo público. . .” Escândalo foi! Lá *ser, foi*. É o diabo! Não há meio de sofismar! E o Conselho não pode estar à espera! (Brazil, CdP, 18:Azevedo:Princesa)

The king is reflecting on the main QUD of whether his daughter, the princess, has infringed Article 200 and is to be prosecuted. He divides the law into several parts, asking himself if each of them is indeed applicable. The implicit immediate QUD addressed by the TI-construction is: *Was it a scandal or not?* The answer is affirmative (Pt. *Escândalo foi! Lá ser, foi!* ‘It was a scandal! As for being, it was’). As shown in this example, there must be a specific context that allows the use of *ser* ‘to be’ as a TI. In this case, the copula appears in the previous statement *Escândalo foi* ‘It was a scandal’. This indicates that the copula verb *ser* ‘to be’ can in-

deed be topicalized if this verb has already been mentioned previously and can therefore function as an immediate QUD.

In a later work, Bastos-Gee (2009: 181–182) states that certain patterns such as **Ir, o João foi para a Bahia* ‘As for going, John went to Bahia’ and **Ser, o João era inteligente* ‘As for being, John was intelligent’ are unacceptable because the verbs *ir* and *ser* have suppletive stems (*fo-* and *e-* in the past tense).⁸ However, as shown in example (2), the pattern *ser, foi* is documented in BP. In another CdP-corpus, Web/Dialects, there are also attestations of topicalized *ir* combined with a suppletive stem (*ir, vou*) in EP:

- (3) – Não, falaram todos, mas eu tenho a impressão que ainda fiquei mais confuso. Se calhar não percebi nada e não devia era ter ouvido o debate. . .
 – Olha o melhor é ires aos comícios todos cá do concelho para ver se ficas menos confuso.– Eu *ir vou*, mas se o meu voto é sempre igual ou ao da minha prima [. . .], ou ao do meu primo [. . .] ou ao do meu cunhado [. . .], se calhar o melhor é não ir a nenhum. (Portugal, CdP Web/Dialects, Blog)

Such lexical-morphological restriction can therefore be ruled out by pragmatics, i.e., by certain conversational contexts that license rather unusual topicalizations.

Also, the impersonal, existential verb *haver* ‘there is/there are’ is documented as an TI in our corpus, see example (4) for EP:

- (4) INQ1 Isso era uma seara de trigo. E depois aparecem essas.
 INF Pois. [vocalização] É, é [vocalização]. Às vezes, apareciam muitas ervitas assim. A gente não lhe sabia, às vezes. . .
 INQ1 Eram assim: davam uma folha, uma flor encarnada. . .
 INF Pois, pois. Pois é.
 INQ1 Se calhar aqui não há.
 INF [vocalização] *Haver*, pode *haver*. Mas o que é é que a gente, às vezes, não lhe lembra o. . . [pausa]
 INQ2 Pois. Mas. . . Papoila ou pimpoila não lhe diz nada?
 INF Não me Não me lembra o nome. Não me lembra o nome dela. (Portugal, CORDIAL-SIN, Assanhas, Figueiró da Serra, excerto 5)⁹

⁸ On the syntactic level, the suppletive stems supposedly block a copy of the verb stem and its subsequent movement to the topic position (Bastos 2009: 181). Contrarily, Vicente (2007: 74) considers that Spanish examples such as *Ir, Juan va or Ir, Juan fue* are acceptable.

⁹ We quote the original source, CORDIAL-SIN, because we cite the transcript as it is provided there. Example (4) is an example of spontaneous, unplanned conversation as recorded in COR-

Again, it is the specific context that allows the verb *haver* to be topicalized: the utterance *Haver, pode haver* ‘As for being, there can be’ is a reaction to the previous utterance *Se calhar aqui não há* ‘Maybe there aren’t any [of these flowers] here’. There is, thus, an antecedent of *haver* that licenses the topicalization of the existential verb.

Furthermore, Bastos (2001: 83) considers that the verb *poder* ‘to be able’, when it expresses the ability to do something (and not a permission), is unacceptable in this construction in BP: **Poder chover, pode. . .* ‘As for the possibility of raining, it is possible. . .’. In our corpus, however, the verb *poder* is also topicalized when it expresses possibility. For example, in (5), *poder* refers to the possibility that something will happen in the future:

- (5) – [. . .] Existem muitos shows gravados pela EMI? Como eles foram gravados? Me parece que existe até mesmo um show gravado em Manaus. Existe alguma apresentação em especial que te chamou a atenção? Será que desses shows poderemos ter algumas novidades boas para os fãs? Um novo disco ao vivo pode ser um lançamento futuro?
 – *Poder, pode* sim, até porque os dois shows do Parque Antártica (SP, 1990) foram gravados em 24 canais pela EMI e na época aprontado para lançamento. Esse disco ao vivo quase saiu, mas foi abortado em favor do “Música p/ Acampamentos” em 92. (Brazil, CdP, 19:Or:Br:Intrv:Web)

In light of our data, the TI-construction is a fully schematic construction that is apparently extendable to any verb. Restrictions that apply to isolated sentences disappear in specific contexts where such sentences are no longer unnatural. A corpus-based approach is thus helpful in assessing the extensibility of the TI-construction. Our data shows that this construction is fully productive and can probably occur with any verb if the discourse context allows it.

4.3 Discourse Context

As mentioned in the section above, TI tend to appear in reactions (replies or responses) to a previous statement. There must be a specific context that favors this construction. In our sample, the TI-construction is never used to open a new con-

DIAL-SIN (there are truncations, correctios, uncomplete sentences, etc.). On the contrary, example (5) is also classified as ‘oral’, but it is clearly a much more planned and formal radio interview, i.e., example (4) is conceptually oral, and example (5) is conceptually written (in the sense of Koch & Oesterreicher 2011), see our discussion in 4.1.

versation or to introduce a completely new topic out of the blue. As already shown in examples (2) to (5) in the previous section, there is an antecedent in the immediately preceding context for the TIs *ser*, *haver*, and *poder*, i.e., the TI refers to a discourse-old topic, which is at issue at this specific point of the discourse. More specifically, the TI makes explicit the immediate QUD of the current discourse (Muñoz Pérez & Verdecchia 2022).

This leads to the question of whether an explicit antecedent (or ‘anchor’, Andrade 2020: 103) is necessarily obligatory. In our sample, we analyzed the previous context of each example to determine whether the verbs in the TI-constructions have been mentioned before or not. The results are shown in Figure 3.

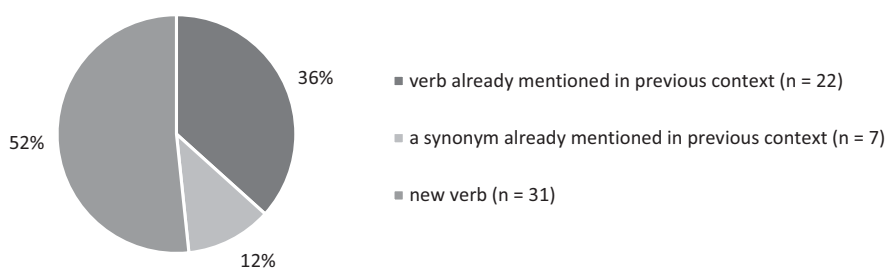


Figure 3: Previous discourse context of TI-constructions (n = 60).

4.3.1 TI with Antecedents

Surprisingly, in our sample only 22 examples (36%) take a verb that has already been mentioned explicitly in the previous context. In most of these cases, the antecedent is to be found in the immediately preceding statement of the interlocutor (see also examples 2, 3, and 4 in the previous section). In 7 examples (12%) it is not the same verb but a synonym that appears in the previous context as an antecedent.

For example, in (6) the TI *crer* ‘as for believing’ is a synonym of the antecedent *acreditar*:

- (6) – Reza, Ananias?
 – É o costume, meu santo Nhô Ambrósio cura pelo rastro.
 – E você *acredita* nessas cousas?
 – Bem, *crer* não *creio* muito não. Mas se não tem outro remédio? (Brazil Cdp, 19:Fic:Br:Carvalho:Somos)

The (polar) QUD *Do you believe in these things?* is answered in the negative ('as for believing, I don't believe much'), but at the same time, the speaker signals that this answer is not sufficient to fully explain his behavior. This objection is also introduced by the discourse maker *bem* 'well', that signals acceptance with reservations. In point of fact, the sentence that follows the TI-construction is introduced by *mas* 'but' and relativizes the answer given in the TI-construction by posing an alternative QUD (*What if praying is the only thing one can do?*).

Also in example (7), the TI responds to a question. In this example, a lady is persistently questioning a servant of her daughter, who is living with a man and not telling her mother all the details:

- (7) – E ouve lá, Maria José, eles nunca *falam em casar*? – Maria José detestava D. Matilde e o seu feitio mexeriqueiro; tinha-lhe medo e às intrigas que deixava como rastro da sua passagem cada vez que vinha a Lisboa.
 – Saiba V. Ex.a que a minha Senhora. . . – e deixava tombar uma resposta tola, propositadamente tola.
 – Mas não foi isso que eu te perguntei, rapariga; não te faças parva. Eles *falam em casar*?
 – Lá *falar* não *falam*, pelo menos à mesa, que é quando eu os oiço. Mas isso não quer dizer nada. (Portugal, CDP, 19:Fic:Pt:D'Arcos:Tons)

The QUD *Eles falam em casar?* 'Are they talking about getting married?' is answered in the negative (*não falam* 'they are not talking'), but this negative is not enough to fully answer the question and all its implicatures regarding the probability of these two getting married. By explicitly marking the topic (*falar* 'as for speaking'), the servant signals that she can only partially answer to the QUD because it needs to be split up into several sub-QUDs:

- sub-QUD1: *Do they talk about getting married in front of you, at the table?*
- Answer: *No.*
- sub-QUD2: *Do they talk about getting married when you are not listening?*
- Answer: . . .
- sub-QUD3: *Do they secretly think about getting married, without even talking about it?*
- Answer: . . .

The servant evades her commitment to the negative answer she gives by evoking alternative sub-QUDs and not answering them, i.e., by leaving open the option of positive answers to these alternative QUDs. These may give a more satisfactory and conclusive answer to the 'big' question. In example (7), the TI-construction is thus used as part of a mitigation-strategy.

In certain instances, the inflected verb in the comment may remain implicit: *Cozinhar, eu. . . quase que nada* (Leitão Vasco 2006: 143). Such examples were not systematically addressed in our search query since we explicitly searched for inflected verb forms following the TI. However, a few instances were found by coincidence. They demonstrate, in our view, that the TI-construction is conventionalized: elliptic comments are possible, as the verb form in the comment can be easily recovered from the context and by knowing that the TI is followed by an inflected verb form of the same lexeme. This is the case of the TI *ajoelhar* ‘as for kneeling’ in example (8). In this example, both TIs (*andar* ‘to walk’ and *ajoelhar* ‘to kneel’) have antecedents. The antecedent *eu nem posso me ajoelhar* ‘I can’t even kneel’ probably facilitates the ellipsis of the inflected verb form in the second TI-construction: the comment *não* ‘not’ can easily be reconstructed as *não me ajoelho* or *não posso me ajoelhar*:

- (8) INF E eu bordava, bordava, bordava, o bocadinho que me ficava, bordava. Depois a minha filha, a mais velha, foi crescendo e eu fui para a fábrica da conserva, ainda trabalhei dez anos lá, [pausa] mas depois acabou. Trabalhei em casa duma senhora, [pausa] a dias. Ia lá [vocalização] umas horazinhas, a minha filha ficava em casa. Depois também eu fiquei, das pernas, que estou muito doente das minhas pernas, eu nem posso me *ajoelhar*. *Ando*. *Andar, ando*; mas me *ajoelhar* e a coisa, não.
INQ1 Pois, pois.
INF Sempre me tratando. O Senhor Doutor já me quis operar dos joelhos mas eu não quis. (Portugal, CORDIAL-SIN, Camacha, excerto 10)

In this conversation, the informant reports that she stays at home and cannot work anymore because of her sick legs. The implicit QUD that arises after this statement would be something like *How sick are the legs?* This QUD can be answered by responding to several polar sub-QUDs such as *Are you able to walk?* and *Are you able to bend your knees?* Each one is addressed by a TI. The first one is answered in the affirmative, and the second in the negative.

In most examples discussed so far, the TI-construction is followed by an adversative sentence, i.e., there is a continuation starting with *mas* ‘but’ that introduces an opposite answer to an alternative QUD. However, this is not always the case, as in example (5), cited in the previous section:

(see example 5 in section 4.1)

Será que desses shows poderemos ter algumas novidades boas para os fãs? Um novo disco ao vivo pode ser um lançamento futuro?

– *Poder, pode* sim, até porque os dois shows do Parque Antártica (SP, 1990) foram gravados em 24 canais pela EMI e na época aprontado para lançamento. Esse disco ao vivo quase saiu, mas foi abortado em favor do “Música p/ Acampamentos” em 92. (Brasil, CdP)

The QUD regarding the possibility is answered in the affirmative (*pode sim* ‘yes, it is possible’), the possibility of releasing a live album is emphatically asserted (see Vicente 2007: 64). An objection in the form of an alternative QUD –probably regarding the realistic probability of such a disc– is not explicitly stated, but a contrasting negative answer to such an alternative QUD may be inferred: e.g., *mas não é muito provável* ‘but it’s not very realistic’ or *mas não sei se vai sair* ‘but I don’t know’ (in fact, the following utterance provides an example of a live album that almost was released but that was abandoned). The absence of an adversative sentence following the TI-construction *poder, pode sim* may trigger such a pragmatic inference, the ‘but-effect’ as stated by Bastos (2001: chapter 2.3). This corroborates that TI-constructions are highly conventionalized as a strategy that responds to a polar QUD while simultaneously evoking an alternative polar QUD that must be answered in the opposite way (though this effect is not fully automatic: for exceptions, see section 5).

4.3.2 TI without Antecedents

In 31 examples (52%) the verb that appears in the TI-construction is a new verb that was not mentioned in the previous discourse. For example, in (9) the TI *comer* ‘to eat’ is not a reaction to a previous utterance containing this verb. In this conversation, the informant talks about the experience of visiting her son in Angola:

- (9) e ele queria que eu fosse para lá // mas eu / fui lá / estive lá seis meses / mas eu lá também + eu gosto muito de lá / que Luanda é muito lindo e Golungo Alto é muito lindo // mas eu não me dou lá com o calor // o calor era muito // eh / eu estava sempre a suar // só queria beber / só queria beber // *comer* não *comia* / só queria beber // só bebia quick. (Portugal, CRPC-ORAL, Casto Daire / Viseu, 1970, pf0894pu)

The discourse topic is the following: the son of the informant has been living in Africa for 20 years and has a family there. She is sad about that because they don’t see each other very often. He’s her only child and she’s a widow, so the question is, why doesn’t she just move there like her son suggested? She tried that, but she could not stay in Angola for more than six months. In example (9), she lists the different arguments: she liked the cities, but it was very hot; because of the heat, she only wanted to drink, but she couldn’t eat. The topic *comer* ‘as for eating’ does not address a QUD that was explicitly asked before, but it is easily

recoverable from the previous context, in contrast to drinking. The TI addresses a sub-QUD that answers part of the bigger QUD:

- QUD: *Why is the heat a problem for moving to Angola?*
 - sub-QUD1: *Do you sweat a lot?*
 - Answer: *yes* (Pt. *eu estava sempre a suar* ‘I was always sweating’)
 - sub-QUD2: *Do you drink a lot?*
 - Answer: *yes* (Pt. *só queria beber* ‘I only wanted to drink’)
 - sub-QUD3: *Do you eat a lot?*
 - Answer: *no* (Pt. *comer não comia* ‘as for eating, I did not eat’)

Each response to a sub-QUD triggers a new sub-QUD. In this context, *comer* is an accessible topic according to the common ground knowledge about the effects of heat on the human body. It is therefore a topic that can be expected in this discourse; it is not discourse-old but hearer-old. This allows the TI-construction to appear without an explicit antecedent. Furthermore, the TI-construction marks a contrast to the previous two topics: to sweat and to drink a lot, but –on the contrary– to not eat at all. In order to mark this contrast, the topic (*comer*) is explicitly announced and thus this statement is set apart from the previous two statements.

Also, in example (10) there is no antecedent of the TI *garantir* ‘as for guaranteeing’:

- (10) Ele manda na política de educação? De saúde? Participa da elaboração das altas estratégias do governo? Então, que mando é esse? Para o presidente, aparentemente superada a crise da semana passada, o governo anda em “céu de brigadeiro”. Coincidência ou não, a opinião do ACM é parecida: ele fala em “mar de almirante”. Será? Pode ser, mas *garantir*, ninguém *garante*. Porque a reeleição já foi mais sólida do que está. (Brazil, CdP, 19:N:Br:Cur)

The QUD in this example is whether it is true that the government is actually in a calm coalition at this moment, after passing a political crisis (*Será?* ‘Is it like that?’). The answer, given by the speaker to himself, is divided into two sub-QUDs, one answered in the positive and the other in the negative:

- Sub-QUD1: *Is it possible?*
 - Answer: *Yes*. (Pt. *Pode ser* ‘it is possible’)
- Sub-QUD2: *Is it guaranteed?*
 - Answer: *No*. (Pt. *mas garantir, ninguém garante* ‘as for guaranteeing, no one guarantees’)

Also in this case, the sub-QUD2 (*is it guaranteed?*) is triggered as a continuation of the affirmation of sub-QUD1 (*is it possible?*), again relying on common ground and world knowledge that if something is possible this does not mean that it actually happens. Note that there are similar sub-QUDs involved in example 5 (*poder, pode sim*) mentioned above.

Finally, in example (11), the TI *trabalhar* ‘to work’ has no antecedent, but names an easily recoverable topic from the common ground (workers do work):

- (11) “Anda lá, ó Gustavo! E tu, Glória! Mostrem àguêles senhores as habilidades!”
 Mas eram precisamente os olhos daqueles senhores que anquilosavam as pernas dos dançarinos. Sem consciência clara disso, sentiam que era uma degradação cantar e bailar para divertir patrões. *Trabalhar, trabalhavam*, porque o pão sobrepunha-se a tudo, mesmo às maiores vergonhas. Mas *cantar e dansar* era uma coisa que se podia escusar quando não fôsse imposta pelo coração. (Portugal, Cdp, 19:Fic:Pt:Torga:Vindima)

The implicit QUD here is what kind of physical activities the workers are willing to do for the landlords of the vineyard. This QUD is split into several polar sub-QUD:

- sub-QUD1: *Do they work?*
 - Answer: *yes* (Pt. *Trabalhar, trabalhavam*. . . ‘as for working, they work’)
- sub-QUD2: *Do they sing and dance?*
 - Answer: *no* (Pt. *Mas cantar e dansar*. . . ‘but singing and dancing. . .’)

Again, both sub-QUDs are answered in an opposite manner: the workers accept to work, but they are ashamed to sing and dance to entertain the owners of the vineyard and refuse to do so. Both sentences are introduced by infinitives as contrastive topics: *trabalhar* as the TI in a TI-construction, and *cantar e dansar* as infinitives in the subject position, i.e., in the unmarked topic position of a sentence.

5 Outlook: Adversativity

In the examples discussed so far, there is consistently a contrast (adversativity) between the response to the sub-QUD addressed by the TI and the response to its alternative sub-QUDs. In the case of a negation in the comment of the TI-construction (e.g., *não falam, não creio, não comia, ninguém garante*), the alternative sub-QUDs are either answered in the affirmative or such a possible affirmation is insinuated (conversational implicature). Conversely, if the comment in the TI-construction contains an affirmation (e.g., *trabalham, pode, foi, pode haver*), the answers to actual or

insinuated alternative sub-QUDs are negative. However, this is not the case in all examples of our sample. In this section, we provide a brief outlook on some examples that do not entail an adversativity continuation. A detailed analysis of the role of adversativity in TI-constructions is a subject for further investigation.

To give an example, in (12), the TI-construction is employed to emphasize an affirmation that relativizes or even drowns out possible alternative QUDs:

- (12) A hipoteca era o recurso supremo naquela casa. Já fora ela que garantira a minha ida para o Brasil, e quando nos últimos anos da Universidade meu tio estrangulou a mesada, convencido de que me pagara já os anos de labuta na Morro-Velho, meu Pai sossegava-me evocando-a.
 – *Acabar, acabas*. Nem que eu tenha de tirar dinheiro a juros sobre a Cortinha da Fonte. Não fora preciso então, e também não queria que o fosse agora. (Portugal, CdP, 19:Fic:Pt:Torga:Criacao)

In this example, the previous conversation is not reproduced by the narrator. Most likely, the son told his father that he did not know if he would be able to finish university because he was running out of money, i.e., there was probably an antecedent of *acabar*, or this topic could easily be recovered from the discourse context (the topic is discourse-new, but hearer-old). There are two connected QUDs, one about graduating from university and one about the money needed. *Acabar, acabas* ‘as for finishing [your studies], you will finish’ is an emphatic assertion of the main topic. Connected QUDs such as *Do we have money for university?* are to be answered negatively at this moment (e.g., *but I don’t know yet with what money you will finish*). However, the negative answer to the QUD about money is presented as if it would not interfere with the positive answer to the QUD about finishing university: in the worst case, the father would take out a mortgage. This means that whatever the answer to the QUD on money is, and even if it is negative, the answer to the QUD on finishing university must be positive.

Finally, there are also examples in which the TI-construction does not trigger alternative QUDs at all and in which there is no adversativity continuation (explicit or implicit). This is the case of example (2) already quoted in section 4.2:

(see example 2)

“Artigo duzentos. Toda pessoa real que, esquecendo o decoro que deve a si própria e ao povo, der escândalo público, será julgada por um Conselho composto de quatro Ministros de estado, e, averiguado o delito, condenada a pena última”. Se se pudesse sofismar este maldito artigo duzentos! Vejamos por partes: [. . .] “que esquecendo o decoro que deve a si própria e ao povo. . .” Disto se esqueceu ela. Comeu queijo! “der escândalo público. . .” Escândalo foi! Lá ser, foi. É o diabo! Não há meio de sofismar! E o Conselho não pode estar à espera! (Brazil, CdP, 18:Azevedo:Princesa)

The QUD at this point of the discourse is whether the wording of the law, *der escândalo público* ‘cause public scandal’, is fully and literally applicable. The answer is affirmative, and the TI-construction emphasizes this assertion by expressing that the topic ‘to be a scandal’ is absolutely accurate (*Escândalo foi! Lá ser, foi*. ‘It was a scandal! As for being, it was.’). There is no alternative, contrasting QUD that is evoked, nor an adversative continuation that relativizes this affirmation. All parts of the legal text, i.e., all QUDs, are to be answered affirmatively and the daughter is to be judged guilty.

Our preliminary conclusion is therefore that the core function of the TI-construction is emphatic affirmation or negation of the event or state of affairs presented in the topic. Implicit or explicit continuation effects, that is, adversativity to alternative QUDs, are common contextual effects, but they are not always automatically triggered by the TI-construction itself.

6 Conclusion

Due to the difficulties encountered in searching for TI-constructions in large corpora such as CdP (namely low frequency and high schematicity), the sample analyzed in this paper is relatively small (60 examples). Nevertheless, it is still sufficiently representative to indicate tendencies and allow us to draw several conclusions:

With regard to discourse genres (diaphasic variation), the corpus data confirms that TI-constructions are predominantly employed in colloquial conversations. Furthermore, the data reveals diatopic variation, with the pattern being documented at a notably higher rate in Portugal (EP) than in Brazil (BP). However, this finding may be somewhat influenced by the corpus design. The Brazilian and Portuguese subcorpora of CdP are comparable in terms of quantity (number of words), but less so in terms of quality when it comes to oral data. Specifically, the spoken data for EP is much more balanced than the data for spoken BP. Therefore, to confirm that IT-constructions are more frequent in EP, further studies based on comparable corpora of colloquial speech are necessary. Such corpora should comprise a similar large number of different informants/speakers to ensure that idiosyncratic results do not distort the whole picture.

One of the advantages of large lemmatized and PoS-tagged corpora is that they enable the measurement of the productivity of a pattern. The corpus data demonstrated that, despite being relatively infrequent, IT-constructions are highly productive because they are documented with a multitude of different verbs. We even argue that there are no lexical restrictions at all: any verb can be topicalized when the specific discourse context allows it.

The verb that is topicalized can be either explicitly addressed in a previous utterance (antecedent) or recovered from the common ground. Presumably, verbs that are less amenable to topicalization, such as copula *ser* and existential *haver*, require a specific context with an antecedent (i.e., an occurrence of *ser* and *haver* in a previous utterance). Conversely, other verbs that refer to specific actions, such as *falar*, *crer*, *comer*, or *trabalhar*, do not necessarily require an antecedent. They can also be topicalized if they are discourse-new, but easily recoverable from the common ground (i.e., hearer-old).

Context has demonstrated to be crucial. The TI-construction responds to an immediate polar QUD, which can be either explicitly formulated by the interlocutor or implicitly arise in the previous discourse. A recurrent pattern can be observed: a complex QUD must be split into several sub-QUDs in order to answer some of them negatively and others positively. The TI addresses one of these sub-QUDs and the action or state of affairs addressed by the TI is asserted or negated in the following comment. Frequently, as a contextual effect, a contrast to an alternative polar QUD that must be answered in the opposite way arises, i.e., TI-constructions may trigger adversativity as a continuation effect or as a conversational implicature.

In certain examples, such alternative sub-QUDs are even more decisive and important for answering the main question. In such cases, the TI-construction is used to mitigate the assertion or negation expressed by evoking or introducing alternative QUDs. On the contrary, in other corpus examples, the TI-construction emphatically responds to the most important QUD, overruling potential alternative QUD. In any case, the continuation effect is not fully conventionalized as part of the TI-construction, as there are also corpus examples without any explicit or implicit adversative continuation. In future studies, the role of contrast and adversativity and the different context-dependent pragmatic functions of TI-constructions will be analyzed in greater detail.

Bibliography

- Andrade, Aroldo Leal de (2020): “Construções de tópico marcado no português brasileiro”, in *Cuadernos de la ALFAL* 12/2, pp. 100–125.
- Baayen, R. Harald (2008): “Corpus Linguistics in Morphology: Morphological Productivity”, in Anke Lüdeling and Merja Kytö (eds.). *Corpus linguistics. An International Handbook*, Vol. 2. Berlin: De Gruyter, pp. 899–919.
- Bastos, Ana Cláudia Pinto (2001): *‘Fazer, eu faço!’ Topicalização de constituintes verbais em português brasileiro*. Master’s Thesis, Universidade Estadual de Campinas.

- Bastos-Gee, Ana Claudia (2009): “Topicalization of Verbal Projections in Brazilian Portuguese”, in Jairo Nunes (ed.). *Minimalist Essays on Brazilian Portuguese Syntax*. Amsterdam: Benjamins, pp. 161–189.
- Bechara, Evanildo (2009): *Moderna Gramática Portuguesa*. [37th ed.]. Rio de Janeiro: Nova Fronteira.
- Büring, Daniel (2003): “On D-Trees, Beans, And B-Accents”, in *Linguistics and Philosophy* 26/5, pp. 511–545.
- Calderón Campos, Miguel (2024): “Spanish Corpora: Big (Quality) Data?”, in Ana Gallego Cuiñas and Daniel Torres-Salinas (eds.). *Humanities and Big Data in Ibero-America. Theory, methodology and practical applications*. Berlin/Boston: De Gruyter, pp. 109–128.
- CdP= Mark Davies and Michael Ferreira (2006): *Corpus do Português* <<http://www.corpusdoporlugues.org/>>.
- CORDIAL-SIN = Martins, Ana Maria (1999–2022): *CORDIAL-SIN: Corpus Dialetoal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects*. <<https://cordialsin.wordpress.com>>.
- CRPC = Centro de Linguística da Universidade de Lisboa [CLUL] (1988–2012): *Corpus de Referência do Português Contemporâneo (CRPC)*. <<http://teitok.clul.ul.pt/crpcoral/index.php?action=home>>.
- Duarte, Inês (2013): “Construções de Topicalização”, in Eduardo Paiva Raposo do Nascimento, Maria Fernanda Bacelar da Mota, Maria Antónia Coelho, Luisa Segura and Amália Mendes (eds.). *Gramática do Português. Vol. I*. Lisboa: Fundação Calouste Gulbenkian, pp. 401–426.
- Frascarelli, Mara (2017): “Dislocations and framings”, in Andreas Dufter and Elisabeth Stark (eds.). *Manual of Romance Morphosyntax and Syntax*. Berlin/Boston: De Gruyter, pp. 472–501.
- Goldberg, Adele E. (2006): *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Hein, Johannes (2020): *Verb Doubling and Dummy Verb. Gap Avoidance Strategies in Verbal Fronting*. Berlin/Boston: De Gruyter.
- Hidalgo Downing, Raquel (2003): *La tematización en el Español hablado. Estudio discursivo sobre el Español peninsular*. Madrid: Gredos.
- Hoffmann, Thomas and Graeme Trousdale (2013): “Construction Grammar: Introduction”, in Thomas Hoffmann and Graeme Trousdale (eds.). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, pp. 1–12.
- Koch, Peter and Wulf Oesterreicher (2011): *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. [2nd ed.]. Berlin: De Gruyter.
- Leitão Vasco, Sergio (2006): *Construções de tópico na fala popular*. Ph.D. thesis, Universidade Federal do Rio de Janeiro.
- Leonetti, Manuel (2014): “Gramática y pragmática en el orden de palabras”, in *LinRed*, 12, pp. 1–25.
- Leonetti, Manuel and María Victoria Escandell Vidal (2021): “La estructura informativa. Preguntas frecuentes”, in Manuel Leonetti and María Victoria Escandell Vidal (eds.). *La estructura informativa*. Madrid: Visor Libros, pp. 15–181.
- Maslova, Elena and Giuliano Bernini (2006): “Sentence topics in the languages of Europe and beyond”, in Giuliano Bernini and Marcia L. Schwartz (eds.). *Pragmatic Organization of Discourse in the Languages of Europe*. Berlin/New York: De Gruyter, pp. 67–120.
- Mendes, Amália (2016): “Linguística de corpus e outros usos dos corpora em linguística”, in Ana Maria Martins and Ernestina Carrilho (eds.). *Manual de linguística portuguesa*. Berlin/Boston: De Gruyter, pp. 224–251.
- Muñoz Pérez, Carlos and Matías Verdecchia (2022): “Predicate doubling in Spanish: On how discourse may mimic syntactic movement”, in *Natural Language & Linguistic Theory*, 40, pp. 1159–1200.

- Narbona Jiménez, Antonio (2015): *Sintaxis del español coloquial*. Sevilla: Editorial Universidad de Sevilla.
- Reich, Uli (2011): “Frontalizaciones de la semántica verbal en español y portugués”, *Communication at XVIII Deutscher Hispanistentag*, Passau.
- Saab, Andrés (2017): “Varieties of verbal doubling in Romance”, in *Isogloss. A journal on variation of Romance and Iberian languages*, 3/1, pp. 1–43.
- Silva, Fernanda Rosa (2019): “Deslocamento de tópico contrastivo no português brasileiro: uma proposta semântico-pragmática”, in *Revista de Estudos da Linguagem*, 27/2, pp. 771–809.
- Valenzuela, Javier, Joseph Hilferty and Mar Garachana-Camarero (2005): “On the reality of constructions: The Spanish reduplicative-topic construction”, in *Annual Review of Cognitive Linguistics*, 3, pp. 201–215.
- van Wette, Niek (2021): “Productivity of French and Dutch (semi-)copular constructions and the adverse impact of high token frequency”, in *International Journal of Corpus Linguistics*, 26/3, pp. 396–428.
- van Wette, Niek (2022): “The hapax / type ratio. An indicator of minimally required sample size in productivity studies?”, in *International Journal of Corpus Linguistics*, 27/2, pp. 166–190.
- Verdecchia, Matías (2021): “Impossible Presuppositions. On factivity, focus, and triviality”, in *Glossa*, 6/1, pp. 1–29.
- Vicente, Luis (2007): *The Syntax of Heads and Phrases. A Study of Verb (Phrase) Fronting*. Ph.D. dissertation, Leiden University.

