



Spanish corpora and their pedagogical uses: challenges and opportunities

Claudia Sánchez-Gutiérrez, Barbara De Cock & Nicole Tracy-Ventura

To cite this article: Claudia Sánchez-Gutiérrez, Barbara De Cock & Nicole Tracy-Ventura (2022): Spanish corpora and their pedagogical uses: challenges and opportunities, Journal of Spanish Language Teaching, DOI: [10.1080/23247797.2022.2157078](https://doi.org/10.1080/23247797.2022.2157078)

To link to this article: <https://doi.org/10.1080/23247797.2022.2157078>



Published online: 29 Dec 2022.



Submit your article to this journal [↗](#)



Article views: 234



View related articles [↗](#)



View Crossmark data [↗](#)



Spanish corpora and their pedagogical uses: challenges and opportunities¹

Corpus de español y sus usos pedagógicos: desafíos y oportunidades

Claudia Sánchez-Gutiérrez^a, Barbara De Cock^b and Nicole Tracy-Ventura^c

^aDepartment of Spanish and Portuguese, University of California, Davis, CA, USA; ^bLanguage and Communication Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium; ^cDepartment of World Languages, Literatures & Linguistics, West Virginia University, Morgantown, WV, USA

ABSTRACT

While first-language (L1) corpora have been a key tool for linguistic research in Spanish, their use for pedagogical purposes is still limited, and more corpora are needed that document the varieties of Spanish used by learners or heritage speakers of the language. In this article, we provide an overview of the research that has already been carried out with available Spanish corpora and we propose avenues for further development of the areas where there is room for growth. As an introduction to this special issue, we also summarize the highlights of each of the six articles in this volume. Finally, we conclude with a call for collaboration among Spanish corpus researchers to address the current limitations when it comes to creating corpora that include oral and longitudinal data, that are more accessible, and more useful for daily use in classrooms of Spanish as a second, foreign or heritage language.

RESUMEN

Aunque los corpus han sido un instrumento clave en el desarrollo de la investigación lingüística en español, su uso para fines pedagógicos sigue siendo limitado. Se necesitan además más corpus que documenten las variedades utilizadas por aprendices y hablantes de herencia de español. En este artículo, ofrecemos una panorámica general de la investigación que ya se ha llevado a cabo con los corpus disponibles y proponemos sugerencias de desarrollo en ciertas áreas con potencial de crecimiento. Como introducción a este monográfico, también resumimos los contenidos de los seis artículos que componen este volumen. Finalmente, concluimos llamando a la colaboración entre investigadores en corpus en español para atajar las limitaciones actuales del campo, así como la creación de más corpus orales o longitudinales, que sean más accesibles, y que resulten más útiles para las necesidades diarias en las aulas de español como lengua extranjera, segunda lengua o lengua de herencia.

ARTICLE HISTORY

Received 15 June 2022

Accepted 15 November 2022

KEYWORDS

Corpus linguistics; learner corpora; Spanish language teaching (SLT); data-driven learning; second-language acquisition

PALABRAS CLAVE

Lingüística de corpus; corpus de aprendices; enseñanza del español; enseñanza basada en datos; adquisición de segundas lenguas

1. Introduction

Corpora have been the centerpiece of the Spanish philological, linguistic and lexicographic tradition since the creation of the *Real Academia Española* (RAE). However, their use for pedagogical purposes

or in the context of research on Spanish as a second (L2), foreign (FL) or heritage (HL) language is still limited. Concretely, pedagogical and acquisitional research based on corpora in Spanish is lagging behind the well-established tradition of learner corpora and data-driven learning developed in the realm of English as a L2/FL/HL. Of the 137 learner corpora listed on the *Learner corpora around the world* website maintained by the Université catholique de Louvain, 60% focus on English, while only 8% focus on Spanish. This lack of representation of corpus-based research in the context of L2/FL/HL Spanish, as well as the limited array of studies on pedagogical uses of corpora in L2/FL/HL Spanish classrooms, do not adequately reflect the international interest in learning Spanish, with more than 21 million L2/FL Spanish learners worldwide (Instituto Cervantes 2019).

The purpose of this monograph is to highlight the many ways in which Spanish corpora are already being utilized in L2/FL/HL Spanish research, demonstrate the applicability and impact that these tools can have on language teaching, and promote the development of new corpora as well as innovative uses of current and future first language (L1) and L2/FL/HL Spanish corpora. Specifically, the monograph consists of two complementary parts, each comprising three articles:

Part 1. Research and applicability of L1 corpora to L2/FL/HL Spanish teaching and learning

Part 2. Design and use of learner and heritage speaker corpora for L2/FL/HL Spanish research and teaching

While the first part provides an overview of existing L1 Spanish corpora, their potential applications in the L2/FL/HL classroom, and the ways in which appropriate teacher training can encourage greater use of L1 corpora for pedagogical purposes, the second part focuses on currently available learner corpora, their affordances for research on L2/FL/HL Spanish development and how their design modulates the type of research questions that can be answered with them. Each article includes an in-depth and up-to-date review of the current literature as well as actionable suggestions on how findings from said literature can be utilized to increase and improve the use of corpora as research and teaching tools. With this format, which combines the presentation of research findings with specific pedagogical recommendations, we hope the articles will be useful for students and researchers specializing in L2/FL/HL Spanish but also for teachers and language program coordinators/directors who may be interested in incorporating corpora in their L2/FL/HL Spanish curricula.

2. Corpus linguistics and available Spanish corpora

2.1. Introduction to corpus linguistics

Corpus linguistics is an approach that utilizes computers to analyze large collections of naturally-occurring spoken and written data (stored in electronic format) to “investigate how speakers and writers exploit the resources of their language” (Biber et al. 1998, 1). With the help of computers, millions of words can be searched nearly instantaneously, and patterns can be uncovered that would have otherwise been impossible to find. A corpus is defined as “a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety” (McEnery, Xiao, and Tono 2006, 5). Language corpora can be general, in that they aim to represent common uses of a language across a range of registers, or specialized, when they document language uses from a specific register or specific groups of speakers. The first large electronic corpus was the Brown Corpus (1-million words), compiled in the 1960’s to include published written texts representing American English. Later in the 1970’s, the parallel London-Oslo-Bergen (LOB) corpus was compiled with written texts of British English (Biber and Reppen 2012). These corpora, and those that followed, allowed for the empirical analysis of natural language use, resulting in descriptive studies of how language varies across registers and discourse contexts. During the same time, dictionaries and reference grammars also began to be based on corpora. The influence of corpus linguistics can be found in several areas of applied linguistics, including language teaching and learning. However, to date, the majority of corpus-based studies

and reference materials focus on English, even though everyday corpora in other languages, including Spanish, are becoming more accessible.

2.2. L1 Spanish corpora

In Spanish, corpus linguistics has been long associated with the diachronic and synchronic corpora compiled by the RAE and, more recently the *Asociación de Academias de la Lengua Española* (ASALE). The corpora compiled by these institutions have a general scope and are organized chronologically, such that each corpus covers different time periods. The *Corpus Diacrónico del Español* (CORDE, *Real Academia español s.d.*) contains over 250 million tokens and aims to document the evolution of the Spanish language from its first written traces to 1974. The *Corpus de Referencia del Español Actual* (CREA, *Real Academia Española s.d.*) contains 160 million words from a variety of sources and countries which were collected from 1975 to 2004. Finally, the *Corpus del Español del Siglo XXI* (CORPES, *Real Academia Española s.d.*) aims to register the general uses of the Spanish language across the 21st century through a collection of over 300 million tokens. While the CORPES has made considerable efforts to represent a range of varieties of Spanish, with around 70% of the corpus data coming from Spanish-speaking countries other than Spain, the RAE has repeatedly been criticized for over-representing the varieties of Spanish spoken in Spain. In this context, the *Corpus del Español, Web/Dialects* (Davies 2016) provides an alternative general corpus with two billion words gathered from webpages from 21 countries across the Spanish-speaking world. All the corpora mentioned in this paragraph are freely accessible and can be consulted via dedicated, rather intuitive interfaces, which has undoubtedly contributed to their being widely used.

In terms of specialized corpora, certain corpora have focused on specific geographical areas, covering zones as broad as Latin America in the *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM; Academia Mexicana de la Lengua *s.d.*), or Mexico in the *Corpus del Español Mexicano Contemporáneo* (CEMC; Colegio de México *s.d.*). Other corpora have focused on a specific city, as is the case of the *Corpus Sociolingüístico de la Ciudad de México* (CSCM; Martín Butragueño and Lastra 2011-2015), which is itself composed of sub-corpora that contain data from interviews with speakers with three different levels of education: high, medium, and low. In addition to corpora that specialize on specific geographical varieties of Spanish, others have focused on specific registers, such as the corpus *Valencia, Español Coloquial* (Val.Es.Co; Pons Bordería, *s.d.*) which explores the characteristics of colloquial Spanish through the collection and analysis of spontaneous conversations from the Spanish city of Valencia. The project *América y España, Español Coloquial* (AMERESCO; Albelda and Estellés *s.d.*) further develops this idea by compiling additional conversational data from cities in Mexico, Colombia, Argentina, Cuba, Panamá and Chile.

L1 corpora have been traditionally used to study a wide variety of linguistic research topics, including morphosyntactic, semantic and pragmatic phenomena (see Parodi, Cantos-Gómez, and Howe 2022 for an overview). Corpora that allow for access to the broader sentential and textual context have also been used for discursive analyses and some spoken corpora have been utilized to study phonetic phenomena (e.g. Hidalgo Navarro 2019). Moreover, corpora have also allowed for the study of language variation, including geographic (e.g. López Meirama 2020), register (Biber et al. 2006) or sociolinguistic variation (e.g. Moreno Fernández 2009). Even though L1 corpora have traditionally been considered as research tools for linguists, proponents of data-driven learning have long advocated for the introduction of corpus data in the L2/FL/HL classroom (Gilquin and Granger 2010; Godwin-Jones 2017; Johns 1991). The purpose of this teaching approach is to let students become language researchers, or “detectives” (Johns 1991, 101), who search for cues about specific linguistic patterns in authentic utterances found in corpora, make informed hypotheses about them, and discuss their findings with peers and instructors. L1 corpora can thus be key pedagogical tools in the classroom, used to further students’ metalinguistic knowledge and increase learners’ autonomy (Asención-Delaney et al. 2015; Jablonkai et al. 2020; Reinhardt 2010; Yanto and Nugraha 2017).

In Spanish, for example, Benavides (2015) used the *Corpus del Español* in an advanced grammar course in a US university. Students had to search examples of different grammatical structures that were presented in class (i.e., *ser* vs. *estar* / preterite vs. imperfect) and analyze them to better understand the uses and nuances of those forms. Yao (2019) used the CREA to have his students analyze examples of specific vocabulary items in context to infer the meaning of the words. In both studies, learners reported positively on the corpus-based activities and learned more than the students in control groups that followed more traditional methods. Marcos Miguel (2020) confirms that, contrary to teachers' beliefs, corpus activities are generally well received and are considered easy to complete by students in US higher-education settings who are learning L2/FL Spanish. In addition to using corpora for the development of grammatical skills, these could also be useful tools to introduce pragmatic content in the L2 classroom (Taguchi 2015). The corpus AMERESCO shows great potential as a tool to compare turn-taking strategies during daily conversations in different varieties of Spanish, for instance. Similarly, the *Corpus del Español* or CORPES offer examples of grammatical and lexical variation across the Spanish-speaking world, which could afford L2/FL/HL Spanish students opportunities to identify similarities and differences across geographical areas (Marcos Miguel, 2022). Heritage speaker corpora could also be a resource for learners of L2 Spanish who learn the language in places where it is spoken by bilingual speakers in their community. Increasing learners' awareness of the varieties of Spanish spoken in their own environment may make them realize that the language they are learning is not as "foreign" as they may think. While, to the best of our knowledge, the latter pedagogical uses of L1 and heritage corpora have not been documented in L2 Spanish classes, this volume aims to encourage new and more systematic pedagogical uses of corpora and to open conversations on how to do that.

Even though studies that document the use of L1 and heritage corpora in L2/FL/HL classrooms show encouraging results and corpora are increasingly easy to access and consult, it is still rare to see teachers who are not corpus researchers make use of corpora as teaching tools in a systematic way (Poole 2020). While several factors may have contributed to this situation, a lot of focus has been placed in the literature on teachers' insufficient familiarity with corpora and with the tools necessary to consult them (e.g., concordancers). In this context, some solutions have been proposed to support teachers willing to introduce corpora in their classes. First, more online tools are currently available that do not require users to know any programming languages. Many Spanish L1 corpora, such as *Corpus del Español*, CREA or CORPES, include user interfaces that allow for direct searches of words, multiword expressions and grammatical structures. Additionally, free programs, such as *AntConc* (Anthony 2022) or *WordsandPhrases* (<https://www.wordandphrase.info/analyzeText.asp>) allow any user to upload texts and search specific words from the texts to obtain frequency counts or other relevant information, such as their most frequent collocates. Another way to avoid technological issues when working with corpora in class is for the teacher to extract examples from the corpus and print them so that students do not need to access the internet or interact with a new interface while still working on hypothesis testing through the analysis of selected data from L1 corpora (Boulton 2010). Finally, if data-driven learning is to become a broadly adopted teaching approach, instructors need to receive specific training to become familiar with existing L1 corpora and the tools used to consult them (Chen, Flowerdew, and Anthony 2019; Leńko-Szymańska 2017).

3. Learner corpus research in the Spanish-speaking world

3.1. Features of learner corpora

A learner corpus is a type of specialized corpus composed of language by L2 or FL learners. Learner corpora (LC) are defined as "systematic collections of authentic, continuous, and contextualized language use (spoken or written) by L2 learners stored in electronic format" (Callies and Paquot 2015, 1). Although LC can include spoken or written data, the majority tend to be written due to

the ease with which the data can be collected. Oral data first need to be transcribed, which is still rather labor intensive, with several important decisions to be made along the way (see Tracy-Ventura and Huensch 2018; Bell et al. 2021). The (semi-)naturalistic data provided by LC differ in important ways from the more controlled types of assessments traditionally used in second language acquisition (SLA) research. In LC it is critical that “learners can use their own wording rather than being prompted to produce a specific linguistic feature” (Granger 2021, 245). Learner Corpus Research (LCR) is a field that emerged from corpus linguistics, and thus follows several of its key principles including authenticity.

The first major LC was the *International Corpus of Learner English* (ICLE: Granger et al. 2002) developed by a team of researchers at the Université catholique de Louvain. ICLE was designed to complement the *International Corpus of English*, by contributing a L2/FL/HL language learner variety as one of many varieties of English included. ICLE could be purchased by researchers and, as a result, several researchers outside of the project team began conducting their own studies with the corpus. From there, LCR grew quickly to become a dynamic international community and now includes a professional organization, a biannual conference, and its own journal, the *International Journal of Learner Corpus Research*. LCR was initially quite descriptive in nature, documenting how L2 learners were using different grammatical features or lexical items. This differed from SLA research, which has primarily been theory-driven, utilizing many different types of data, some very controlled like grammaticality judgment tasks and others more open-ended like participant interviews. As Myles (2021) notes, “the data needs of different SLA theories can be very different, depending on what hypotheses need to be tested” (260). Therefore, LC have been utilized in SLA research, and increasingly so recently, but they are not the primary source of data like they are in LCR. Although SLA and LCR were initially quite separate disciplines, over the years they have been benefiting from increased collaborations, with several new recent publications focusing specifically on the interface between the two fields (e.g., Le Bruyn and Paquot 2021; Tracy-Ventura and Paquot 2021).

Current LC can reach quite large sizes, with many including over one million words (e.g., *The Trinity-Lancaster Corpus of Learner English*: Gablasova et al. 2019). Such large corpora typically include data from hundreds of learners at different proficiency levels. Other types of metadata, in addition to proficiency, may be available such as age, L1, country of origin, number of years studying the language, time spent abroad, etc. These metadata make it possible to investigate a variety of questions. While large corpora offer several research opportunities, they still may not be large enough to study rare linguistic features. In such cases, it is advisable to also include more controlled tasks that have been found to elicit the features under investigation (see Tracy-Ventura and Myles 2015). Complementing corpora with more experimental methods, like those used in SLA, is one way that researchers have been attempting to better understand learners’ knowledge and use of different aspects of their L2/FL/HL (Gilquin 2022). Although many large LC tend to be available to the research community (a very welcome endeavor), most researchers build their own small specialized LC (e.g., Czerwionka and Olson 2020).

As was the case with L1 corpora, LC have mainly been used for research purposes but can also be useful tools in the design of pedagogical materials (Götz and Mukherjee 2019). Concretely, Paquot (2018), among others, suggests that LC offer unique insights into the structures and words that learners over- or under-use, at different proficiency levels, information that can serve to establish which linguistic structures may require more, or less, attention in pedagogical materials for a given group of learners. An example of this type of effort is the INTELeNG project (Mendikoetxea, Murcia Bielsa and Rollinson 2010) where researchers collected and error-corrected texts from students of L2 English at a Spanish University. The resulting error database was the basis for the creation of pedagogical materials introduced in L2 English classes to provide students with opportunities to reflect on the issues present in erroneous utterances and to find ways to correct them. However, despite some documented instances of such pedagogical uses of LC, these are still mostly anecdotal and far from being widespread.

3.2. Spanish Learner Corpora

The first L2/FL/HL Spanish LC were composed of oral or written samples from small groups of students generally enrolled in language courses in the institution or geographical area of the researchers. For example, the *Díaz Rodríguez Corpus* (Díaz Rodríguez 2002) contains audio and interview transcripts and other oral tasks (e.g., picture description, oral elicitation task) completed by eight learners of Spanish with various L1s. Another example of a corpus that constituted the first steps of L2/FL/HL Spanish LC would be the *Corpus de Textos Escritos por Universitarios Taiwaneses Estudiantes de Español* (Lin 2005) which taps into the writing of learners in a very specific learning context.

Over the years, larger LC that aimed to represent a broad range of L2/FL/HL Spanish students (with multiple L1s) across the globe started to emerge. This is the case of the *Corpus Escrito del Español L2* (CEDEL2; Lozano 2009, 2021) or the *Corpus de Aprendices de Español* (CAES; Rojo and Palacios 2016). Some LC, such as the *Corpus of Written Spanish L2 and Heritage* (COWS-L2H; Yamada et al. 2020) or the *Spanish Corpus Proficiency Levels Training* (Koike and Witte 2016), have also recently started including data from heritage speakers of Spanish, mostly in the United States. While most of these corpora contain only written data, efforts have been made to develop more oral corpora, such as the *Spanish Learner Oral Language Corpus* (SPLLOC; Mitchell et al. 2008) or the *Corpus Oral de Peticiones en Interacciones Naturalizadas en Español* (COPINE; Marsily 2022). Additionally, some LC offer longitudinal data of students as they progress through their coursework in a language program (see COWS-L2H), or as they participate in a study abroad program (see the *Languages and Social Networks Abroad Project*, LANGSNAP: Tracy-Ventura, Mitchell and McManus 2016). This type of data is a welcome addition to the current LC landscape, as their longitudinal nature allows for more fine-grained analyses of learners' individual trajectories (Sánchez-Gutiérrez and Fernández-Mira 2022).

The diversity of Spanish learner corpora currently available has already allowed researchers to carry out investigations in a broad range of areas of linguistic development, such the aspectual differences between Spanish past tenses (Domínguez et al. 2013; Minnillo et al. 2022), gender agreement errors (Gudmestad, Edmonds, and Metzger 2021) or lexical diversity (Fernández Mira et al. 2021; Tracy-Ventura, Mitchell, and McManus 2016). Recent studies have also been published that look into pragmatic aspects of the development of a L2/FL/HL. For instance, Marsily (2018) analyzed the mitigation strategies used in requests by L1 French learners of Spanish in the COPINE corpus, while Vázquez Veiga (2016) studied discourse markers in L1 English learners of Spanish in CEDEL2 and SPLLOC.

In addition to these research findings through LCR, researchers have also made concrete proposals to use these LC for pedagogical purposes. Sampedro Mella (2021), for instance, describes a classroom activity where L2/FL/HL Spanish students had to analyze the pragmatic appropriateness of learners' utterances in CAES as part of a class devoted to learning how to make requests. Bowles (2022) used SPLLOC and CEDEL2 to select appropriate vocabulary for a placement test that aimed to differentiate L2 learners from heritage speakers. Some researchers are also interested in developing automatic error-correction tools, using LC as training sets for their models (e.g., Davidson et al. 2020). Finally, the *Spanish Corpus Proficiency Levels Training* (Koike and Witte 2016) was designed, among other things, to train Spanish language instructors in determining learners' proficiency levels based on a corpus of video-recordings. These different examples illustrate the multiple roles that Spanish LC can take when it comes to language education, beyond the traditional focus on research. Our hope is that this monograph will serve to call for more initiatives in this line of work, fostering an open attitude towards LC development and use for pedagogical purposes.

4. Articles in this special issue

In the opening article of this monograph, René Venegas, Iris Viviana Bosio and Constanza Cerda offer an overview of Spanish L1 corpora, both general ones such as CREA, CORPES XXI and *Corpus del*

español, and specialized ones such as PRESEEA, Val.Es.Co, AMERESCO. Most of the research carried out based on those corpora has focused on analyzing linguistic phenomena from a theoretical perspective, such as morphosyntactic features (e.g. verb mode and pronoun use), lexical phenomena (e.g. formulaic sequences) and pragmatic phenomena (e.g. discourse markers). These Spanish L1 corpora have however been used only scarcely for research with didactic purposes. Some of these studies address how the use of corpora can help with language learning and teaching. The authors of this overview point out that a lot of varieties of Spanish are still poorly represented in existing corpora, even more in corpora for specific purposes. As to language learning and teaching, these Spanish L1 corpora are so far also scarcely used in research on language learning and teaching, highlighting a gap to be overcome.

In Article 2, Yuly Asención-Delaney, Joseph Collentine, Jesus Colmenares, and Alfredo Urzúa describe ways in which pre- and in-service Spanish teachers can be trained on the pedagogical uses of corpora. They highlight ways in which increasing teachers' corpus literacy can help provide them with tools for exploring authentic language use, which can also contribute to increasing their understanding of language variation. The authors provide examples of how they are introducing corpus resources in their own teacher-training programs and helping teachers gain practice creating corpus-informed materials and hands-on data-driven activities for their students. They demonstrate how corpus training was incorporated into graduate courses on Computer-Assisted Language Learning (CALL) and Spanish sociolinguistics, and how this training also led to two students choosing to utilize corpora in their graduate capstone projects. The final example involved the creation of a learner corpus of peer interactive tasks that was part of an action-based research project. In-service teachers utilized the learner corpus to explore *translanguaging* (García and Wei 2014) in peer interactions. The authors conclude by highlighting the importance of longitudinal research examining whether and how teachers continue to utilize corpora in their teaching after their initial training.

In Article 3, Nausica Marcos Miguel and Carrie Bonilla present convincing arguments as to why L1 corpora can be useful in the L2/FL/HL classroom and in the creation of textbooks and pedagogical materials. They also explain the reasons that have prevented the use of corpora for pedagogical purposes from becoming a mainstream practice, pointing to limited teacher training, the prevalence of the communicative language teaching approach, and the difficulty of learning the technological tools required to consult corpora. In response to these challenges, the authors provide avenues to address each of them and emphasize the benefits of using corpora as useful resources for language learning and teaching. Finally, they propose concrete activities that can be done, based on the literature and on their own experience, to introduce and practice various aspects of language learning through specific L1 corpora.

In the first article of the second part, Article 4, Aarnes Gudmestad summarizes the Spanish LCR that has been carried out over the last decade to investigate L2/FL and HL learning in the areas of grammar, vocabulary, and pragmatics. She focuses her review on research utilizing publicly available corpora highlighting a number of advantages to making learner corpora freely accessible to the research community. She demonstrates how Spanish learner corpora have been examined to test different theories and approaches from SLA research (e.g., the *Aspect Hypothesis*, Levelt's Model of Speech Production, Processability Theory, Variationist Approaches), and global and specific aspects of vocabulary (e.g. lexical diversity, collocations) and pragmatics (e.g. discourse markers, general extenders). The second half of her article describes the applications of this research to teaching. In particular, she emphasizes how observations from Spanish LCR could be utilized to design pedagogical interventions. She concludes by describing several areas for future research including the need to examine more varied linguistic phenomena such as prepositions and object pronouns, in addition to creating more oral corpora which would allow for the investigation of unique features of oral language (e.g., prosody).

In article 5, Guillermo Rojo, Ignacio Palacios, María Sampedro Mella and Aurélie Marsily offer an overview of existing Spanish LC and build on this overview to trace future perspectives. The

panorama described in this article is particularly rich since it includes not only information concerning the type of data, such as written/spoken mode, learners' L1 and corpus size, but also information related to the analyses that can be carried out with them. The authors take into account whether the data are available to other researchers, whether they are annotated and whether a specific platform or interface has been created. This review article sheds light on some recurrent problems, such as lack of information relative to corpus size, period and accessibility, but also concerning the variety of Spanish that learners studied and/or whether or not they had been in a Spanish-speaking country for a study exchange. The data obtained for the corpora are often based on specific tasks, which has the advantage of comparability but the disadvantage of lacking authenticity. Many corpora have not been annotated and lack information concerning the level of Spanish of the participants. This article thus offers an interesting overview for those in search of a Spanish LC for their research or pedagogical purposes, as well as routes to further develop and improve Spanish LC.

Finally, in article 6, Cristóbal Lozano and Paloma Fernández Mira describe the many decisions that underlie LC design and how those can impact the type of research that can be done with each corpus. In their article, they start by comparing five LC (i.e. CEDEL2, COWS-L2H, LANGSNAP, CAES and SPLLOC) along five variables: language modality (i.e., oral vs. written), learner profile (e.g., L1, age, educational context), corpus statistics (e.g., number of words, number of participants), subcorpora (e.g., reference L1 corpus, different L1 subcorpora), and metadata (i.e., the information collected about the participants and the tasks). After this initial overview of different corpus characteristics, the authors focus on the design decisions made when designing two specific LC: CEDEL2 and COWS-L2H, providing detailed insights into how those decisions were made, when and why. The article concludes with examples of how both LC can be used to respond to specific research questions concerning grammatical or lexical development, and suggestions about potential uses of these corpora for pedagogical purposes.

5. Limitations and challenges of the use of corpora in Spanish language learning and teaching

The studies included in this special issue shed light on the challenges related to the use of corpora in (research on) Spanish language learning and teaching. As pointed out by Rojo et al. as well as Venegas et al., information about existing corpora is not always easy to find and not all corpora are easily accessible. The general movement towards OpenScience will hopefully encourage the accessibility of corpora. As was also pointed out by Rojo et al., some types of L2 corpora are lacking. There are for instance few longitudinal learner corpora, but LANGSNAP and COWS L2H are leading the way in this direction. As to the type of language production, most corpora include data from set tasks and less from language production in more natural interactions, but COPINE offers an example of how this type of LC can be further developed. Moreover, some L1s are much less represented in the existing Spanish LC than others. While this in part reflects socio-economic realities as well as the fact that L2/FL/HL Spanish is more widespread in certain geographical areas than in others, the creation of Spanish LC representing a wider diversity of L1s, as is the case in CEDEL2, is a promising route for the future, both to broaden overall knowledge on Spanish language teaching and to bring the results of Spanish LC research closer to local teaching practices.

With regards to existing corpora, it seems challenging to transfer the insights we gain from studying (learner) corpora to the actual teaching of L2/FL/HL Spanish. This has multiple causes. A first issue may be the place that the use of corpora holds (or often does not hold) in teacher training. In that respect, Asención-Delaney et al. offer an interesting proposal to enhance the presence of corpora in teacher training and, hence, in the teaching of Spanish. A second one is the fact that curricular constraints may also be an obstacle to the use of corpus techniques in Spanish classes in some countries, especially when there is a very strong focus on communicative skills, a topic discussed in detail in Marcos Miguel and Bonilla. In sum, teachers need to familiarize themselves with the available

(learner) corpora and need to be trained into using them as pedagogical tools in accordance with, or as a complement to, the teaching approaches used in their specific context.

The way ahead for us, corpus enthusiasts, is to create a community of researchers and educators who are motivated to design and compile new (learner) corpora that fill the gaps described in Rojo et al., follow the most updated recommendations in the field in doing so (see Lozano and Fernández Mira), use them to better understand students' difficulties and their linguistic development (see Gudmestad), and share concrete activities that can be done with such corpora in the L2/FL/HL Spanish classroom (see Marcos Miguel and Bonilla for specific examples). Our hope is that this monograph can contribute to these efforts and provide inspiration for future corpus designers and users.

Note

1. The second and third authors appear in alphabetical order. Both have contributed equally to the editing of this special issue.

References

- Academia Mexicana de la Lengua. *Corpus Diacrónico y Diatópico del Español de América (CORDIAM)*. <www.cordiam.org>
- Albelda, M. and M. Estellés. coords. *Corpus Ameresco* <www.corpusameresco.com>
- Anthony, L. 2022. AntConc (Version 4.0.11) [Computer Software]. Tokyo, Japan: Waseda University. Available from <https://www.laurenceanthony.net/software>
- Asención-Delaney, Y., J. G. Collentine, K. Collentine, J. Colmenares, and L. Plonsky. 2015. "El potencial de la enseñanza del vocabulario basada en corpus: optimismo con precaución." *Journal of Spanish Language Teaching* 2 (2): 140-151.
- Bell, P., L. Collins, and E. Marsden. 2021. "Building an Oral and Written Learner Corpus of a School Programme: Methodological Issues." *Learner Corpus Research Meets Second Language Acquisition*: 214-242.
- Benavides, C. 2015. "Using a Corpus in a 300-Level Spanish Grammar Course." *Foreign Language Annals* 48 (2): 218-235.
- Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D., M. Davies, J.K. Jones, and N. Tracy-Ventura. 2006. "Spoken and Written Register Variation in Spanish: A Multi-Dimensional Analysis." *Corpora* 1 (1): 1-37.
- Biber, D. and R. Reppen. 2012. "Introduction." In *Corpus Linguistics* (Vols. 1-4), eds. D. Biber and R. Reppen, SAGE Publications Ltd, <https://doi.org/10.4135/9781446261217>
- Boulton, A. 2010. "Data-Driven Learning: Taking the Computer out of the Equation." *Language Learning* 60 (3): 534-572.
- Bowles, M. A. 2022. "Using Instructor Judgment, Learner Corpora, and DIF to Develop a Placement Test for Spanish L2 and Heritage Learners." *Language Testing* 39 (3): 355-376.
- Callies, M. and M. Paquot. 2015. "Learner Corpus Research: An Interdisciplinary Field on the Move." *International Journal of Learner Corpus Research* 1 (1): 1-6.
- Chen, M., J. Flowerdew, and L. Anthony. 2019. "Introducing In-Service English Language Teachers to Data-Driven Learning for Academic Writing." *System* 87: 102148.
- Colegio de México. *Corpus del Español Mexicano Contemporáneo (CEMC)*. <<http://www.corpus.unam.mx/cemc>>
- Czerwionka, L. and D. J. Olson. 2020. "Pragmatic Development during Study Abroad: L2 Intensifiers in Spoken Spanish." *International Journal of Learner Corpus Research* 6 (2): 125-162.
- Davidson, S., A. Yamada, P.F. Mira, A. Carando, C. H. Sánchez-Gutiérrez, and K. Sagae. 2020, May. Developing NLP tools with a New Corpus of Learner Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 7238-7243).
- Davies, M. 2016. *Corpus del Español: Two Billion Words, 21 countries*. <<http://www.corpusdelespanol.org>>
- Díaz Rodríguez, L. 2002. "Interferencias Discursivas de Hablantes Bilingües Castellano/Catalán: Uso Oral y Escrito. In *Seminari sobre les llengües i educació de l'Estat*, ed. J. Perera. Barcelona: Ice-Horsori.
- Dominguez, L., N. Tracy-Ventura, M.J. Arche, R. Mitchell, and F. Myles. 2013. "The Role of Dynamic Contrasts in the L2 Acquisition of Spanish Past Tense Morphology." *Bilingualism: Language and Cognition* 16 (3): 558-577.
- Fernández-Mira, P., E. Morgan, S. Davidson, A. Yamada, A. Carando, K. Sagae, and C. H. Sánchez-Gutiérrez. 2021. "Lexical Diversity in an L2 Spanish Learner Corpus: The Effect of Topic-Related Variables." *International Journal of Learner Corpus Research* 7 (2): 230-258.
- Gablasova, D., V. Brezina, and T. McEnery. 2019. "The Trinity Lancaster Corpus: Development, Description and Application." *International Journal of Learner Corpus Research* 5 (2): 126-158.
- García, O. and L. Wei. 2014. *Translanguaging: Language, Bilingualism and Education*. London: Palgrave.

- Gilquin, G. 2022. "The Process Corpus of English in Education: Going Beyond the Written Text." *Research in Corpus Linguistics* 10 (1): 31-44.
- Gilquin, G. and S. Granger. 2010. "How Can Data-Driven Learning be Used in Language Teaching?" In *The Routledge Handbook of Corpus Linguistics*, eds. A. O'Keeffe and M. McCarthy, 359-370. London and New York: Routledge.
- Godwin-Jones, R. 2017. "Data-Informed Language Learning." *Language Learning & Technology* 21 (3): 9-27.
- Götz, S. and J. Mukherjee. eds. 2019. *Learner Corpora and Language Teaching*. Amsterdam: John Benjamins.
- Granger S. 2021. "Have Learner Corpus Research and Second Language Acquisition Finally Met?" In *Learner Corpus Research Meets Second Language Acquisition*, eds. B. Le Bruyn and M. Paquot, 243-257. Cambridge: Cambridge University Press.
- Granger, S., E. Dagneaux, and F. Meunier. eds. 2002. *The International Corpus of Learner English: Handbook and CD-ROM*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain. (Available from <http://www.i6doc.com>)
- Gudmestad, A., A. Edmonds, and T. Metzger. 2021. "Moving Beyond the Native-Speaker Bias in the Analysis of Variable Gender Marking". *Frontiers in Communication* 165.
- Hidalgo Navarro, A. 2019. De segmentación y prosodia en la conversación coloquial. In *Pragmática del español hablado: hacia nuevos horizontes*, eds. A. Cabedo Nebot and A. Hidalgo Navarro, 227-238. Valencia: Universitat de València.
- Instituto Cervantes. 2019. *El español: una lengua viva*. Madrid: Instituto Cervantes.
- Jablonkai, R., L. Forti, M. A. Castelló, I. S. Iguenane, E. Schaeffer-Lacroix, and N. Vyatkina. 2020. "Data-Driven Learning for Languages Other than English: The Cases of French, German, Italian, and Spanish." *CALL for Widening Participation: Short Papers from EUROCALL 2020*, 132.
- Johns, T. 1991. "From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning". *English Language Research Journal* 4: 27-45.
- Koike, D. and J. Witte. 2016. "Spanish Corpus Proficiency Level Training Website and Corpus: An Open-Source, Online Resource for Corpus Linguistics Studies." In *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 169-196. Amsterdam: John Benjamins.
- Le Bruyn, B. and M. Paquot. eds. 2021. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press.
- Leńko-Szymańska, A. 2017. "Training Teachers in Data Driven Learning: Tackling the Challenge." *Language Learning & Technology* 21 (3): 217-241.
- Lin, T.-J. 2005. "Corpus de Textos Escritos por Universitarios Taiwaneses Estudiantes de Español." *Lingüística en la Red* 3: 1-58.
- López Meirama, B. 2020. "Variación diatópica y análisis de corpus: algunos casos en la fraseología del español." *Estudios de lingüística* 145-159.
- Lozano, C. 2009. "CEDEL2: Corpus Escrito del Español como L2." In *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente*, eds. C. M. Bretones José, F. Fernández Sánchez, J.R. Ibáñez Ibáñez, M.E. García Sánchez, M.E. Cortés de los Ríos, S. Salaberri Ramiro, M.S. Cruz Martínez, N. Perdu Honeyman, and B. Cantizano Márquez, 197-212. Almería: Universidad de Almería.
- Lozano, C. 2021. "CEDEL2: Design, Compilation and Web Interface of an Online Corpus for L2 Spanish Acquisition Research." *Second Language Research* <https://doi.org/10.1177/02676583211050522>
- Marcos Miguel, N. 2020. "Exploring Tasks-as-Process in Spanish L2 Classrooms: Can Corpus-Based Tasks Facilitate Language Exploration, Language Use, and Engagement?" *International Journal of Applied Linguistics* 31 (2): 211-228.
- Marsily, A. 2018. "¿Es normal que sea un poco difícil de leer la consigna?" La atenuación en las peticiones de hablantes no nativos de español." *ELUA Anexo* 4: 251-268.
- Marsily, M. 2022. *COPINE. Corpus Oral de Peticiones en Interacciones Naturalizadas en Español*. Louvain-la-Neuve: Université catholique de Louvain.
- Martín Butragueño, P. and Y. Lastra. coords. 2011-2015. *Corpus Sociolingüístico de la Ciudad de México (CSCM)*. 1a. ed. Ciudad de México: El Colegio de México.
- McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London and New York: Routledge
- Mendikoetxea, A., S. Murcia Bielsa, and P. Rollinson. 2010. Focus on Errors: Learner Corpora as Pedagogical Tools. In *Corpus-Based Approaches to English Language Teaching*, eds. M. C. Campoy, B. Bellés-Fortuno, and M. L. L. Gea-Valor, 180-194. London: Continuum.
- Miguel, N. M. 2022. "Exploring the Use of Corpus Tools for Teaching Language Variation to L2 Spanish Majors." *Language* 98 (2): e80-e107.
- Minnillo, S., C. H. Sánchez-Gutiérrez, A. Carando, S. Davidson, P. F. Mira, and K. Sagae. 2022. "Preterit-Imperfect Acquisition in L2 Spanish Writing: Moving Beyond Lexical Aspect." *Research in Corpus Linguistics* 10 (1): 156-184.
- Mitchell, R., L. Domínguez, M.J. Arche, F. Myles, and E. Marsden. 2008. "SPLLOC: A New Database for Spanish Second Language Acquisition Research." *EuroSLA Yearbook* 8 (1): 287-304.
- Moreno Fernández, F. 2009. "El estudio sociolingüístico de las hablas hispánicas. Noticias de PRESEEA." In *La investigación dialectológica en la actualidad*, eds. D. Corbella Díaz and J. Dorta Luis, 103-117. Santa Cruz de Tenerife: Agencia Canaria de Investigación y Sociedad de la Información del Gobierno de Canarias.

- Myles, F. 2021. "Commentary: An SLA Perspective on Learner Corpus Research." In *Learner Corpus Research Meets Second Language Acquisition*, eds. B. Le Bruyn and M. Paquot, 258-270. Oxford: Oxford University Press.
- Paquot, M. 2018. "Corpus Research for Language Learning and Teaching." In *Palgrave Handbook of Applied Linguistics Research Methodology*, eds. A. Phakiti, P. De Costa, L. Plonsky, and S. Starfield. London: Palgrave Macmillan.
- Parodi, G., P. Cantos-Gómez, and C. Howe. 2022. *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*. London and New York: Routledge.
- Pons Borderia, S. dir. *Corpus Val.Es.Co 3.0*. <<http://www.valesco.es>>
- Poole, R. 2020. "Corpus can be Tricky": Revisiting Teacher Attitudes towards Corpus-Aided Language Learning and Teaching." *Computer Assisted Language Learning* 1-22. Doi: [10.1080/09588221.2020.1825095](https://doi.org/10.1080/09588221.2020.1825095).
- Real Academia Española. *Banco de datos (CORPES XXI) Corpus del Español del Siglo XXI*. <<http://www.rae.es>>
- Real Academia Española. *Banco de datos (CREA) Corpus de Referencia del Español actual*. <<http://www.rae.es>>
- Real Academia Española. *Banco de datos (CORDE) Corpus Diacrónico del Español actual*. <<http://www.rae.es>>
- Reinhardt, J. 2010. "The Potential of Corpus-Informed L2 Pedagogy." *Studies in Hispanic & Lusophone Linguistics* 3 (1): 239-252.
- Rojo, G., and M. I. M. Palacios. 2016. "Learner Spanish on Computer: The CAES Corpus de Aprendices de Español Project." In *Spanish Learner Corpus Research: Current trends and future perspectives*, ed. M. Alonso Ramos, 55-87. Amsterdam: John Benjamins.
- Sampedro Mella, M. 2021. "Estimado Sr. Vs. Hola hotel: el análisis contrastivo de la interlengua para la enseñanza de la variación pragmático-discursiva." In *La Variación en Español y su Enseñanza: Reflexiones y Propuestas Didácticas*, 133-150. Ediciones Universidad de Salamanca.
- Sánchez-Gutiérrez, C. H. and P. Fernández-Mira. 2022. "Datos Longitudinales en Corpus de Aprendientes de Español." In *Lingüística de corpus en español*, eds. G. Parodi, P. Cantos-Gómez and C. Howe, 374-387. London and New York: Routledge.
- Taguchi, N. 2015. "Instructed Pragmatics at a Glance: Where Instructional Studies Were, Are, and Should Be Going." *Language Teaching* 48 (1): 1-50.
- Tracy-Ventura, N. and A. Huensch. 2018. "The Potential of Publicly Shared Longitudinal Learner Corpora in SLA Research." In *Critical Reflections on Data in Second Language Acquisition*, eds. A. Gudmestad and A. Edmonds, 149-170. Amsterdam: John Benjamins.
- Tracy-Ventura, N., R. Mitchell, and K. McManus. 2016. "The LANGSNAP Longitudinal Learner Corpus: Design and Use." In *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 117-142. Amsterdam: John Benjamins.
- Tracy-Ventura, N. and F. Myles. 2015. "The Importance of Task Variability in the Design of Learner Corpora for SLA Research." *International Journal of Learner Corpus Research* 1 (1): 58-95.
- Tracy-Ventura, N. and M. Paquot. eds. 2021. *The Routledge Handbook of Second Language Acquisition and Corpora*. London and New York: Routledge.
- Vázquez Veiga, N. 2016. "Discourse Markers in CEDEL2 and SPLLOC Corpora of Learner Spanish." *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 267. Amsterdam: John Benjamins.
- Yamada, A., S. Davidson, P. Fernández-Mira, A. Carando, K. Sagae, and C. Sánchez-Gutiérrez. 2020. "COWS-L2H: A Corpus for Measuring Learner Spanish Writing Development." *Research in Corpus Linguistics* 8 (1): 17-32.
- Yanto, E. S. and S.I. Nugraha. 2017. "The Implementation of Corpus-Aided Discovery Learning in English Grammar Pedagogy." *Journal of ELT Research: The Academic Journal of Studies in English Language Teaching and Learning*, 66-83.
- Yao, G. 2019. "Vocabulary Learning through Data-Driven Learning in the Context of Spanish as a Foreign Language." *Research in Corpus Linguistics* 7: 18-46.