Corpora and L2 acquisition: the L1 Portuguese – L2 Spanish subcorpus of CEDEL2

Cristóbal Lozano, Joana Teixeira**, Ana Madeira****

*Universidad de Granada
**FLUP & CLUNL
***NOVA FCSH & CLUNL

Abstract

This paper presents the L1 Portuguese – L2 Spanish subcorpus of *Corpus Escrito del Español L2* (CEDEL2), a new methodological resource for second language acquisition (SLA) research, which is freely searchable and downloadable (http://cedel2.learnercorpora.com). CEDEL2 is a large-scale, multi-L1 learner corpus of L2 Spanish which contains written productions from learners at all proficiency levels as well as 6 native control subcorpora (total size: over 1,100,000 words from over 4,000 participants). CEDEL2 follows strict corpus design criteria (Sinclair, 2005) and learner corpus design recommendations (Tracy-Ventura & Paquot, 2021a). In its current version (CEDEL2 v. 2), its Portuguese component includes an L1 Portuguese – L2 Spanish subcorpus, with 21,662 words written by 164 participants, and an L1 Portuguese native subcorpus, with 3,500 words from 16 L1 speakers of European Portuguese. Thanks to their design features (e.g., same design across subcorpora, inclusion of metadata about SLA-relevant variables, dual native control subcorpora) and freely available web interface, CEDEL2 and its Portuguese subcorpora allow researchers to investigate a wide range of topics in SLA.

***Keywords*: L2 acquisition, learner corpora, Spanish, Portuguese

Palavras-chave: aquisição de L2, corpora de aprendizagem, espanhol, português

1. Introduction

In recent decades, second language acquisition (SLA) research has been based mainly on experimental data. However, there is a growing consensus that more natural and spontaneous data are needed to obtain a more complete picture of the process of SLA. For this reason, interest in learner corpus research has increased in recent years (e.g., Gilquin, 2020; Granger, 2009; Granger, Gilquin & Meunier, 2015; Tracy-Ventura & Paquot, 2021) and several large-scale learner corpora have emerged, including the *Corpus Escrito del Español L2* (CEDEL2) (Lozano, 2009a; Lozano & Mendikoetxea, 2013; Lozano, in press). This corpus contains data of second language (L2) Spanish learners at all proficiency levels (beginner to advanced) from 11 different L1 backgrounds, including Portuguese, as well as 6 native control subcorpora.

This paper presents the design, compilation and free online web interface of the L1 Portuguese – L2 Spanish subcorpus of CEDEL2 and discusses its potential uses. The paper is organised as follows: section 2 introduces the concept of learner corpus and discusses its potential advantages and limitations in SLA research. In section 3, we describe the design features of CEDEL2 and how they adhere to Sinclair's (2005) corpus design principles and Tracy-Ventura and Paquot's (2021a) recent recommendations for learner corpus design. Section 4 presents the L1 Portuguese – L2 Spanish subcorpus of CEDEL2 and the free, web-based search engine interface where it is available (http://cedel2.learnercorpora.com). Finally, in section 4, we discuss some of the potential uses of CEDEL2, in general, and its L1 Portuguese – L2 Spanish subcorpus, in particular.

2. Learner corpora

SLA research aims to describe and explain the nature of non-native language representations and developmental processes and understand the factors which constrain interlanguage development. SLA research conducted within a generative framework has tended to privilege the use of data collected through a variety of experimental methods (Mackey & Gass, 2005; Mendikoetxea & Lozano, 2018). This type of methodology allows researchers to assemble different types of data through various experimental procedures (e.g., elicited production, acceptability judgement or comprehension tasks), designed to obtain information regarding particular linguistic features. Experimental data is thus highly controlled targeted data, allowing researchers to test specific hypotheses, and is particularly important in the case of linguistic features which are rarely found in naturally occurring speech. The use of experimental methods also enables researchers to collect information not only on what structures or forms learners allow, but also on what they consider to be impossible.

Despite the importance of experimental data, it has become evident that the use of more natural and less controlled data is necessary for a more complete characterisation of learners' interlanguages and their development. For this reason, interest in corpora has been increasing in SLA research in recent years (e.g., Myles, 2005; Granger, 2012; Lozano & Mendikoetxea, 2013; Lozano, 2021b; for an overview of corpus-based SLA studies, see, e.g., Myles, 2015), following a trend long observed in L1 acquisition studies (e.g., Demuth, 1998). Learner corpora are electronic collections of written or oral language data from L2 learners, designed according to well-defined criteria (e.g., Granger, 2008, 2009; Callies & Paquot, 2015). Learner corpus data is not necessarily authentic data, i.e., uncontrolled language produced for communicative purposes in natural settings, as it often consists of language samples produced in classroom contexts and elicited through particular linguistic tasks (Granger, 2002, 2015; Días Negrillo & Thompson, 2013). Unlike language elicited through experimental methods, however, linguistic structures and forms found in a corpus are always contextualised, occurring as part of a complete written or oral text.

The use of learner corpora has important advantages and can contribute significantly to SLA research (Myles, 2005, 2015; Granger, 2008; Lozano & Mendikoetxea, 2013; Mendikoetxea, 2014; Lozano, 2021b, in press). Given that they are typically made up of free and (relatively) spontaneous language samples, produced with a focus on content rather than on form, learner corpora can reflect the use of language under natural conditions more accurately and are less subject to the interference of explicit knowledge than most experimental tasks used in SLA research. Moreover, because corpora typically consist of very large data sets, they allow the identification of recurrent patterns of language use that may not be captured in small-scale experimental studies (Myles, 2005; McEnery *et al*., 2019). The fact that *corpus* data is always contextualized also makes it ideal for investigating linguistic phenomena at the interfaces, for example, taking into account discourse and pragmatic factors (Lozano, 2021b). Additionally, this is a resource which can be shared widely and used for many different purposes –building a large learner corpus is a costly and laborious process but, once assembled, the corpus can be made widely available, enabling a large number of researchers to investigate many different aspects of learner language (Myles, 2005; McEnery et al., 2019).

For a learner corpus to be useful for SLA research, it must follow specific design criteria (Gilquin, 2015). The design features of learner corpora should be determined by general corpus design principles (Sinclair, 2005) and simultaneously take into account variables which are relevant to SLA research – learner variables, such as L1, proficiency level, gender, chronological age, age of exposure to the L2, learning context, knowledge of other languages and patterns of language use, as well as task variables, such as type of text and conditions under which the text is produced (Granger, 2015). Information on these variables is presented in the corpus in the form of metadata. A well-designed corpus can contribute significantly to a better understanding of the role of different factors in L2 acquisition. For a detailed discussion of the general corpus design principles proposed by Sinclair (2005) and of how they interact with specific learner corpus design recommendations (Tracy-Ventura & Paquot, 2021a) and SLA-relevant variables, see section 3.

Learner corpus data has some limitations. For example, it is not always possible to draw conclusions about learners' competence from lack of production, as learners may be avoiding the use of a particular structure or form, or there may be an insufficient number of occurrences in the corpus (e.g., because there are not enough contexts in which it would be natural to use a given structure or form) to allow for an accurate characterisation of learners' knowledge. One way to overcome this potential limitation of corpus data is to guarantee that the data is collected using a variety of tasks in order to ensure that the language produced is as representative of learners' language use as possible (Tracy-Ventura & Myles, 2015). Moreover, learner corpus data should in many cases be combined with experimental data, which can provide a more accurate picture of learners' competence than learner corpora (Lozano & Mendikoetxea, 2013; Tracy-Ventura & Myles, 2015; Mendikoetxea & Lozano, 2018; Lozano, 2021b).

The first large-scale learner corpus, the *International Corpus of Learner English* (ICLE), was developed at the University of Louvain in Belgium by Sylviane Granger and colleagues (Granger, Dagneux & Meunier, 2002). It initially contained around 2 million words and consisted of short argumentative texts written by advanced learners of English from different L1 backgrounds. This corpus was later expanded and supplemented by several other corpora, including a control L1 English corpus – the *Louvain Corpus of Native English Essays* (LOCNESS) – and an oral learner corpus – the *Louvain International Database of Spoken English Interlanguage* (LINDSEI)[1]. Since the appearance of the Louvain corpora, there has been an increasing interest in corpus-based research in the field of SLA (e.g., Granger, 2009; Granger, Gilquin & Meunier, 2015) and a significant number of new learner corpora covering a wide range of L2s have emerged.[2] One of the languages for which a number of learner corpora have appeared in recent years is Spanish (Alonso-Ramos, 2016; Mendikoetxea, 2014) (a fairly representative list can be found in *Indexador de Corpus de Aprendices de Español*[3]). Among these Spanish learner corpora are included the following (Lozano, 2021a,b): the *Corpus de aprendices de español como lengua extranjera* (CAES), a corpus of written texts produced by A1-C1 learners of Spanish at the Cervantes Institute from different L1 backgrounds (Rojo & Palacios Martínez, 2016); the *Spanish Learner Language Oral Corpora* (SPLLOC), which is composed of oral language samples elicited through a variety of tasks from L1 English speakers at three different proficiency levels who were learning Spanish in UK schools and universities (Mitchell *et al*., 2008); the *Languages and Social Networks Abroad Project* (LANGSNAP), a longitudinal corpus of L2 Spanish which consists of spoken and written data from English-speaking learners in a study-abroad context (Tracy-Ventura, Mitchell & McManus, 2016); and the *Corpus Oral de Español como Lengua Extranjera* (CORELE), which consists of oral data collected through interviews from A2 and B1 learners of Spanish with different L1 backgrounds (Campillos Llanos, 2014). These corpora are relatively small (for example, CAES, the largest of the four, contains data from 1,423 speakers and a total of 573,718 words), do not always include a native control corpus (this is the case of CAES) and include a limited number of learner and task variables in their design (Lozano, 2021b).

The *Corpus Escrito del Español L2* (CEDEL2) (Lozano, 2009a; Lozano & Mendikoetxea, 2013; Lozano, in press), a large-scale written learner corpus which contains data from L2 Spanish learners with different L1 backgrounds and at all proficiency levels (beginner to advanced), complements the existing corpora and adds important features. Version 1 of CEDEL2 included more than 800,000 words from about 2,600 speakers and consisted of three subcorpora: L1 English – L2 Spanish, L1 Greek – L2 Spanish and L1 Spanish. In its present

[1] These corpora can be accessed at https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html.

[2] For a fairly comprehensive list of learner corpora around the world, see https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[3] Available at http://repositorios.fdi.ucm.es/corpus_aprendices_espa%C3%B1ol/view/paginas/view_paginas.php?id=1.

version (version 2), CEDEL2 includes data from learners from other L1 backgrounds, namely German, Arabic, Chinese, French, Dutch, Italian, Japanese, Russian and Portuguese (over one million words and 4,300 participants). For additional statistical details, see Lozano (in press) and the Statistics section in CEDEL2 website (http://cedel2.learnercorpora.com/statistics).

3. CEDEL2: Design and structure

The version under discussion in this paper is CEDEL2 version 2 (for a full overview of its design and features, see Lozano, in press). It was publicly released in 2021 at the website http://cedel2.learnercorpora.com, where it can be freely searched and downloaded.

The basic design of CEDEL2 could be defined, in Gilquin's (2020) words, as "a mono-L2 and multi-L1 learner corpus, made up of Spanish learner data produced by speakers of various L1s" (p. 295). Importantly, since its inception in 2006, the first version of CEDEL2 (Lozano, 2009a; Lozano & Mendikoetxea, 2013) was designed following the design principles standardly applied to large monitor corpora (Sinclair, 2005), though these principles were adapted to the peculiarities of learner corpora (Lozano, 2009a; Lozano & Mendikoetxea, 2013). Additionally, the current version of CEDEL2 (version 2) (Lozano, in press), which incorporates the L1 Portuguese – L2 Spanish (as well as 10 other learner subcorpora and 6 native subcorpora), adheres to the learner corpus recent recommendations by Tracy-Ventura et al. (2021). CEDEL2 was designed with an SLA agenda, i.e., the final product (corpus and web interface) should be maximally usable by a large number of users, particularly (but not exclusively) SLA researchers who want to investigate learners' interlanguage development and acquisition, as recommended by Díaz-Negrillo & Thompson (2013) – see section 5 for a discussion of potential users and uses of CEDEL2. Given such an SLA agenda, CEDEL2 recorded a wide range of SLA-motivated and SLA-informed variables pertaining to the learners' linguistic profile and their task so as to better understand the factors that influence L2 acquisition, as recommended by Granger (2008). In what follows, we will discuss in detail the design principles of CEDEL2 and how they interact with the learner corpus recommendations and its SLA-motivated variables.

Regarding the basic **corpus design principles**, CEDEL2 adheres to 'good practice' corpus design principles typically used in the design of large monitor corpora (McEnery et al., 2006; McEnery & Hardie, 2012; Wynne, 2005). In particular, in his seminal paper *How to build a corpus*, Sinclair (2005) proposes ten 'good practice' corpus design principles to build a well-designed corpus that can ultimately be maximally usable by language researchers and users in general. These ten principles were adapted to the peculiarities of SLA and L2 learners, which laid the foundations for the design of CEDEL2 (Lozano, 2009a; Lozano & Mendikoetxea, 2013). In what follows, we will discuss the ten principles briefly (the reader is referred to the two papers by Lozano (2009a) and Lozano and Mendikoetxea (2013) for details).

1. ***Content selection:*** The first principle states that the content or texts of the corpus should be selected according to external but not internal criteria. External criteria relate to the communicative function of the texts, whereas internal criteria relate to the language of the texts. In other words, corpus builders should collect texts irrespective of the language they contain. This means that in a learner corpus, for example, the tasks should not prompt for particular linguistic phenomena (e.g., structures like passive voice, certain types of vocabulary content, etc.). CEDEL2 was originally designed to elicit a wide variety of linguistic phenomena and vocabulary, which is achieved via the inclusion of 14 different task titles (see *recommendation 5* below for further details).
2. ***Representativeness:*** If internal criteria are used, the second principle of representativeness could be compromised since the corpus texts may be biased towards a particular linguistic structure/phenomenon, which will be overrepresented in the text, whereas other phenomena may be underrepresented. So, the L2 texts should be as representative as possible of the learner language they

intend to represent. A way of ensuring maximum representativeness is to design a multi-task corpus, i.e., a corpus containing several tasks that vary in certain aspects: narrative/argumentative/descriptive genres, written/oral modality, monologues/dialogues, different topics to talk about, etc. Certainly, a corpus will always be a sample of the language, so it cannot contain all potential linguistic phenomena, but corpus builders should at least aim at collecting a representative sample. The notion of representativeness could be also extended to relate to the representativeness of the corpus speakers/writers. For example, a learner corpus of secondary-school students of L2 Spanish will only represent teenage learners who are acquiring the L2 in the classroom, so results and findings can never be extrapolated to other L2 populations like adults who are acquiring the L2 in a naturalistic setting. As will be explained below, the multi-task design of CEDEL2, together with its variety of L2 learners and natives, renders CEDEL2 a relatively well-balanced and representative corpus of the variety/varieties it intends to sample.

3. ***Contrast:*** Sinclair's third principle stipulates that only those components/subcorpora of the corpus that have been designed to be contrasted should be contrasted. All the learner and native subcorpora in CEDEL2 follow the same design principles. This means that CEDEL2 allows for what is called *Contrastive Interlanguage Analysis*, CIA (Granger, 2015), which basically refers to the fact that SLA researchers can reliably compare and contrast several subcorpora. For example, researchers may be interested in comparing several Romance L1s (L1 Portuguese – L2 Spanish vs. L1 Italian – L2 Spanish vs. L1 French – L2 Spanish) or even comparing two typologically unrelated languages (L1 Portuguese – L2 Spanish vs. L1 Japanese – L2 Spanish). This is possible due to the same design in all subcorpora and it allows researchers to investigate L1-specific vs. universal patterns of influence in L2 learners' interlanguage.
4. ***Structural criteria:*** The fourth principle stipulates that the criteria for determining the structure of a corpus should be just a few. In the case of CEDEL2, the structural criteria relate to the main criteria that ultimately determine the major structure of the corpus. These are the L1 of the learners (11 L1s: English, German, Dutch, Portuguese, French, Italian, Greek, Russian, Arabic, Japanese, Chinese) and the natives (6 L1s: Spanish, English, Portuguese, Greek, Arabic, Japanese), which renders 11 learner subcorpora and 6 native subcorpora. Future versions of CEDEL2 will incorporate additional learners' L1s (for example, we are currently collecting data from L1 Turkish – L2 Spanish and L1 Estonian – L2 Spanish) and additional native languages so that all the learners' L1s are ultimately represented. There are certainly other learner and task variables that can act as secondary structural criteria, particularly learners' proficiency level (lower/upper beginner, lower/upper intermediate, lower/upper advanced) and the task (12 different types of tasks).
5. ***Annotation:*** Sinclair's fifth principle states that the actual texts of the corpus (i.e., the raw texts in TXT format) and their linguistic annotation (i.e., their tagging) should be stored separately. This recommendation should be understood in the context of early corpora in the 80's and 90's, where the texts contained inline annotations in such a way that the actual raw text and the annotations relating to the grammatical categories of each word, were typically annotated between pointed brackets. This rendered raw texts that were not easily legible as the raw text was intermingled with cryptic annotation symbols and codes. CEDEL2 files can be downloaded in two formats: the raw texts (in TXT format), which correspond to the written or spoken task productions, and the raw texts with metadata, which correspond to the raw texts accompanied by all the headers representing the learner variables (e.g., learners' L1, age, proficiency level in L2 Spanish, etc.) and the task variables (e.g., task title, written/oral format, tools used to produce the task, etc.). Regarding the annotation of CEDEL2 (version 2), only the Spanish and English texts have been automatically annotated with the Freeling (http://nlp.lsi.upc.edu/freeling) software. The tags are separate from the raw text since the tags are a

built-in feature of the CEDEL2 web-based search interface (http://cedel2.learnercorpora.com). Via menu-driven, drop-down lists users can search for specific linguistic tags, also known as parts-of-speech (POS) tags, such as *adjective* + *noun* order vs. *noun* + *adjective* order to investigate word order patterns within the noun phrase; or *past tense* verbs; etc. There are many more sophisticated combinations, so the reader is referred to section 4 below as well as the CEDEL2 website for details. Note that the automatic Freeling tagger annotates words based on native dictionaries. Therefore, L2 learners' productions have been automatically tagged and their errors have not been tagged since Freeling misses them. For additional information on tagging and learners' errors, see the CEDEL2 website > User Guide > Tags (http://cedel2.learnercorpora.com/user_guide/tags).

6. ***Sample size:*** The sixth principle states that the important fact for every text is that it should be an *entire* text (i.e., a complete text not edited/shortened/cut off by the corpus builder) rather than a long or short text. This principle was relevant in the early days of corpus linguistics since corpus builders often had to shorten (i.e., cut off) the actual texts due to size limitations of computer files. In CEDEL2, texts are of varying sizing, some texts ranging from just one paragraph with a few lines, which are often typical in lower beginner learners, to long texts containing several paragraphs, which are typical of very advanced learners. The CEDEL2 texts were never shortened or cut off, i.e., the integrity of the texts is always guaranteed, in line with Sinclair's principle.
7. ***Documentation:*** The seventh principle relates to the fact that the design, structure and composition of a corpus should be fully documented. In this respect, in the CEDEL web interface, the user can find an online user guide with detailed information about crucial aspects of the corpus, including the use of the web interface to search and download the corpus, information about the design of CEDEL2, details about the metadata of the texts, the transcription conventions for the oral data, details about the tags (annotation), precise statistics about different aspects of the corpus, information about the CEDEL2 team, etc..
8. ***Balance:*** The eighth principle is closely related to principle 2 (*representativeness*). Sinclair's definition of balance is rather vague, but it refers to the fact that the different subcorpora or components of the corpus should be balanced to achieve representativeness. In this respect, the current version of CEDEL2 is not fully balanced in the sense that, for example, the L1 English – L2 Spanish subcorpus is the largest corpus (558,731 words from 1,931 texts), whereas the rest of the learner subcorpora are smaller (e.g., L1 Portuguese – L2 Spanish 21,662 words from 164 texts). This is due to the fact that CEDEL2 originated as an L1 English – L2 Spanish corpus, whose data were collected for a period of 10 years (2006-2016), hence its large size. For its second version, the CEDEL2 learner subcorpora were expanded for a shorter period (2017-2021), hence their smaller size. In any case, balance in CEDEL2 is also achieved via the application of the same design principles and the collection of the same variables across subcorpora, which is closely related to the principle of contrast.
9. ***Topic:*** According to Sinclair's ninth principle, corpus builders should not exert any control over the topic or the vocabulary of the texts, in line with the first principle of content selection. Sinclair complains that "it seems strange to many people that it is essential that the vocabulary should not be directly controlled" (Sinclair, 2005, p. 9). As explained above, CEDEL2 was designed to potentially elicit different linguistic phenomena and varied vocabulary. This was achieved via 14 different tasks to choose from (see recommendation number 5 below for details).
10. ***Homogeneity:*** The last design principle recommends that the texts should be as homogeneous as possible, avoiding 'rogue' texts which stand out as completely different from the majority of texts. This is a rather intuitive but simple notion. CEDEL2 data were collected via online forms (see the web portal http://learnercorpora.com, where learners can participate). After the data collection was over, all data and texts were manually checked so as to discard any rogue texts (e.g., some low-level

learner texts had been clearly copied and pasted by the learners from an internet source, so the text was not theirs; other very low-level texts were written in the learners' L1, so these texts were also discarded, although we did not discard those texts written in L2 Spanish that contained occasional words from the learners' L1 as this phenomenon can shed light on processes related to code switching/missing).

Let us focus now on the seven **learner corpus design recommendations** by Tracy-Ventura *et al.* (2021) and how these interact with the general design principles outlined above and with the SLA-informed learner and task variables collected in CEDEL2. According to Tracy-Ventura *et al*. (2021), learner corpora should: focus on L2s other than English; include learners at all proficiency levels, with varied L1s, from different ages, and from different learning backgrounds and settings; promote cross-linguistic comparisons; include more learner and task variables (metadata); include varied tasks, some of which promote hypothesis-testing research; consider different perspectives on what a 'control' corpus is; be freely available to the research community (Open Science). Let us discuss now each of the recommendations in some detail.

1. ***Recommendation 1** (Learner corpora should focus on L2s other than English):* Traditionally, most learner corpora have indeed targeted L2 English learners (see overviews in, e.g., Gilquin, 2020; Granger et al., 2015; Tracy-Ventura et al., 2021). The past decade has witnessed an interest in L2 Spanish acquisition research (see overviews in Geeslin, 2014; Montrul, 2004, 2013), which has triggered the creation of large L2 Spanish corpora (Alonso-Ramos, 2016; Lozano, 2021a; Mendikoetxea, 2014). CEDEL2, together with other L2 Spanish corpora, appears in this context to provide answers to L2 Spanish acquisition researchers and other types of users (see section 5 for a discussion of users and uses).
2. ***Recommendation 2** (Learner corpora should include learners at all proficiency levels, with varied L1s, from different ages, and from different learning backgrounds and settings):* This recommendation relates to a series of factors that are crucial (proficiency level: variety of: learners' L1s, chronological ages, learning backgrounds and settings), each of which will be analysed in detail below. They are related to the principle of representativeness discussed above. The learners' **proficiency** level is a key aspect in learner corpus design. In his early paper on learner corpus design, Tono (2003) already recommended using standardised and objective measures to ascertain the learners' proficiency level. CEDEL2 takes three measurements of the learners' proficiency level in L2 Spanish: an objective measure via a standardised placement test (University of Wisconsin, 1998), a subjective measure (learners are asked to self-rate their proficiency level in L2 Spanish in the four skills: speaking, writing, listening, reading) and an additional though optional measure (learners are asked to state any L2 Spanish language certificates they may hold). Following recommendation 2, CEDEL2 holds data from all proficiency-level learners. Another feature stemming from recommendation 2 is the learner's **variety of L1s**. CEDEL2 incorporates, to our knowledge, the widest range of learner subcorpora (by L1) and native control subcorpora of any L2 Spanish corpus currently available (see principle 4 *Structural criteria* above). Regarding the chronological **age** of the learners (and natives as well), CEDEL2 samples speakers from different ages, ranging from secondary-school aged learners to learners over their 70s. However, the most frequent age range is 18-30 years old. Another basic biodata of the learners' (and natives') is their **sex** (male/female). As for **learning background and settings**, most learners in CEDEL2 are instructed learners (particularly at universities in the US and UK), though there is a wide variety of learners in non-instructed and naturalistic settings. According to Gilquin (2015, 2020), this variety of learners and learning environments/settings is a feature that makes CEDEL2 stand out in contrast with other learner corpora, which typically sample university-level L2 students only. This variety in CEDEL2's design

is in line with principle 2 (representativeness) due to the variety of interlanguage types sampled, as well as with principle 7 (documentation) since, for a given learner, a wide range of SLA-relevant variables are registered.

3. ***Recommendation 3** (Learner corpora should promote cross-linguistic comparisons):* Cross-linguistic contrasts are an essential linguistic feature in any learner corpus. This feature is clearly related to Sinclair's third principle of contrast seen above. As we explained earlier, all CEDEL2 subcorpora are equally designed, so specific cross-linguistic comparisons are allowed across learner subcorpora with L1s that can be typologically (un)related (e.g., L1 Portuguese – L2 Spanish intermediates vs. L1 Chinese – L2 Spanish intermediates for a given task like the narration of a silent Charles Chaplin video clip); or the contrast between different native subcorpora (e.g., Spanish natives vs. Portuguese natives); or even the comparison between different varieties within a native subcorpus (e.g., peninsular Spanish vs. Mexican Spanish vs. Argentinian Spanish vs. other varieties of native Spanish in Latin America). Future versions of CEDEL2 will allow us to compare different varieties of native Portuguese, e.g., European Portuguese vs. Brazilian Portuguese, since currently (version 2) only European Portuguese native data have been collected.
4. ***Recommendation 4** (Learner corpora should include more learner and task variables):* CEDEL2 goes beyond recommendation 2 as it incorporates an additional series of learner- and task-related variables that are SLA-motivated and can help researchers address some classic questions in SLA (e.g., age-related and critical period effects, language use/dominance effects, proficiency-level effects, educational effects, medium (spoken/written) effects and task effects). These are discussed in detail with reference to the L1 Portuguese – L2 Spanish subcorpus in section 4.
5. ***Recommendation 5** (Learner corpora should include varied tasks, some of which promote hypothesis-testing research):* While learner corpora have been traditionally used in descriptive, hypothesis-finding L2 acquisition research (Myles, 2005, 2015, 2021 for an overview), they have recently started to be used in explanatory, hypothesis-testing L2 studies (Lozano, 2021b for an overview). Hypothesis-testing research is more conducive in corpora that either (i) are multi-task corpora (as opposed to mono-task corpora) because different tasks can be contrasted so as to yield a fully-rounded picture of learners' interlanguage, as shown in a task-variability study on L2 Spanish (Tracy-Ventura & Myles, 2015), or (ii) contain tasks specifically designed to test particular hypotheses or phenomena, as done with aspectual contrasts in L2 Spanish corpus data (Domínguez *et al.*, 2013). Given that CEDEL2 was designed with a clear SLA agenda in mind, it abides by these two rules since it contains 14 different tasks that range in several respects (degree of difficulty, tense-aspectual contrasts, genres, spoken/oral medium, etc.) and tasks that were purposely used to investigate anaphora resolution in L2 Spanish acquisition due to the very nature of the tasks (e.g., task 2 in CEDEL2 version 1, and the two new tasks of CEDEL2 version 2, tasks 13 and 14), as shown in a series of hypothesis-testing studies (Lozano, 2009b, 2016; Lozano & Quesada, in prep; Martín-Villena & Lozano, 2020) – see also Lozano (in press) for a discussion of how CEDEL2 has been used in other hypothesis-testing studies. The 14 tasks in CEDEL2 version 2 are: *1. Describe the region where you live; 2. Talk about a famous person; 3. Describe a film you have recently seen; 4. What did you do last year during your holidays?; 5. Which are your plans for the future?; 6. Describe a trip you have recently made; 7. Narrate an experience you have lived; 8. Talk about the problem of terrorism in the world; 9. What do you think about the anti-tobacco law?; 10. Do you think gay couples have the right to get married and adopt children?; 11. Do you think marihuana should be legalised?; 12. Analyse the main aspects of immigration; 13. Describe the frog story (picture-bask task); 14. Describe the Charles Chaplin silent video clip*. Researchers can thus have a more fully-rounded picture of the task produced by learners via variables such as the task title / task type: 14 tasks titles to choose from, ranging in degree of

difficulty (from tasks typically suitable for beginners to complex tasks suitable for advanced learners, though learners could choose the task title) and in textual type (descriptive, narrative, argumentative). The reader is referred to Lozano in press and Lozano and Mendikoetxea (2013) for a justification of these three text types. The medium (written/spoken) is another variable: though CEDEL2 is mainly written, it contains spoken audios (with their corresponding transcriptions) coming mainly from the same learners that have done the same task in the written modality, which is a useful feature to compare written vs. spoken language while maintaining the task and learner constant. For additional details on task variables, see section 5 below. Note, incidentally, that this wide variety of task types is also in line with principle 1 (content selection), since the wider the range of task types, the wider the linguistic phenomena they will elicit when analysed together, and principle 9 (topic), since the higher the variety of task topics, the higher the variety of vocabulary.

6. ***Recommendation 6*** *(Learner corpora should consider different perspectives on what a 'control' corpus is):* This is a crucial recommendation that has been rather overlooked in most learner corpora, as these typically lack a native control corpus, which can serve as a benchmark against which L2 learners can be compared. In those few corpora that include a native control subcorpus, it typically samples the language of native speakers of the target language (i.e., the language that the learners are acquiring). This is the case of CEDEL2, which includes a large native Spanish subcorpus of 304,211 words coming from 1,112 Spanish-speaking natives (Peninsular Spanish and Latin American Spanish varieties). This renders CEDEL2 a corpus of native Spanish in its own right. But CEDEL2 goes beyond this standard notion of a control corpus and, spurred by recommendation 6, it also includes several native subcorpora of the learners' L1, which follow the same design principles as the learner subcorpora, in line with principle 3 (contrast). This 'dual' native-corpus design allows researchers to check the effects of the learners' typologically (un)related L1s as well as the effects of the target language (Spanish) they are acquiring. This particular design of CEDEL2 is also motivated by principle 3 (contrast), since the native subcorpora follow the same design principles as the learner subcorpora,
7. ***Recommendation 7*** *(Learner corpora should be freely available to the research community):* The final recommendation pertains to the recent scientific trend (Open Science) that encourages researchers to make their research data freely available to the international research community. Following this philosophy, version 2 of CEDEL2 is released under a Creative-Commons license (CC BY-NC-ND 3.0 ES). It is freely searchable and fully downloadable via a dedicated web interface from the CEDEL2 website at http://cedel2.learnercorpora.com (see the end of section 4 for further details).

4. The L1 Portuguese – L2 Spanish subcorpus of CEDEL2

CEDEL2 includes an L1 Portuguese – L2 Spanish subcorpus for two main reasons. The first is that this language combination is still understudied, even though the number of learners of Spanish in Portugal has significantly increased over the last decade, thanks to the inclusion of Spanish in basic and secondary education curricula in the late 2000s (Ministerio de Educación y Formación Profesional, 2020). The second reason is that Portuguese is typologically related to Spanish – they are Romance languages. Consequently, this language combination can help researchers discriminate between L1 transfer and universal effects and explore the role of typological proximity in L2 acquisition.

Following Sinclair's (2005) principle of contrast and Tracy-Ventura et al.'s (2021) third recommendation (see section 3), the L1 Portuguese – L2 Spanish subcorpus has the same design as the other subcorpora of CEDEL2 to ensure maximum comparability. Data have been collected online via Google Forms from a dedicated website (http://learnercorpora.com). The form is written in Portuguese to ensure full understanding. It includes four sections: instructions and informed consent, learner profile, task, and placement test. Let us see each of these in detail:

1. **Learner profile:** the form collects metadata about a wide range of learner variables that are relevant to SLA research, namely: chronological **age**; **sex** (male/female); **educational background** (educational institution, major degree if any, year at university if any); **learner's L1**(s) and variety of Portuguese; **L1(s) and variety of Portuguese of the learner's father and mother**; **language(s) spoken at home**; **age of exposure** to L2 Spanish (i.e., age at which the learner was exposed to Spanish either naturalistically or in instructed settings); the **length of instruction in Spanish** (i.e., years studying Spanish in formal settings); **stays** in Spanish-speaking countries (in months), which includes the **location** (country), the **periods of residence** and the **length of residence**; Spanish **language certificates** (if any); and **proficiency level self-evaluation** in Spanish and in additional foreign languages (speaking, listening, writing, reading). Registering all these metadata is crucial to enable researchers to investigate a wide range of SLA phenomena on the basis of the L1 Portuguese – L2 Spanish subcorpus (e.g., L1 influence, interference from additional languages, age-related effects, among others – see section 5) and have control over potential confounding factors such as age of exposure (e.g., child L2 acquisition differs from adult L2 acquisition in some respects and the boundaries between them are set on the basis of learners' age of exposure – cf. Schwartz, 2004; Meisel, 2011) or the variety of Portuguese spoken by the learner and his/her parents (e.g., European Portuguese is significantly different from Brazilian Portuguese at a lexical and morphosyntactic level).
2. **Task profile**: Five task variables are recorded in the L1 Portuguese – L2 Spanish subcorpus: (i) **task title**; (ii) **task tex**t (written text produced by the learner); (iii) approximate **time** to produce the task (in minutes); (iv) **place** where the task was done (in class, outside class, both); and (v) **resources** used to produce the task (help from Spanish native, bilingual dictionary, monolingual dictionary, spellchecker, grammar book, background readings, none).To elicit participants' written production, we have used a film retelling task which involves watching a short video clip from Charlie Chaplin's silent movie *The Kid* and writing a summary of the story with at least 400 characters. This task has been extensively used in SLA research (e.g., Perdue, 1993) and in all subcorpora of CEDEL2 (Lozano, in press), which allows for comparisons across subcorpora and with other studies. In the future, we plan to use other tasks of CEDEL2 to collect data from Portuguese-speaking learners of Spanish. This is an important step to ensure that this subcorpus is as representative as possible of the learner language it intends to represent (cf. the principle of representativeness in section 3). Finally, the **medium** (written/spoken) variable is also part of the task profile.
3. **Placement test**: In addition to the measures of proficiency in Spanish collected in the learner profile section (proficiency level self-evaluation in Spanish and Spanish language certificates), the L1 Portuguese – L2 Spanish subcorpus, like all learner subcorpora of CEDEL2, includes an objective independent measure: a standardised placement test from the University of Wisconsin (1998), which consists of 43 multiple-choice questions. This is the last section to be completed. Based on the test, learners are placed at one of the following proficiency levels: lower beginner, upper beginner, lower intermediate, upper intermediate, lower advanced, and upper advanced. The placement score eventually becomes part of the learner profile.

At present, the L1 Portuguese – L2 Spanish subcorpus contains 21,662 words written by 164 participants, with ages ranging from 13 to 53. 98% of these participants have European Portuguese as their only L1. The remaining 2% are L1 speakers of either Brazilian or Angolan Portuguese. As shown in Figure 1, most participants are upper intermediate (29%), lower advanced (29%) or upper advanced (27%) speakers of L2 Spanish. These are also the proficiency levels with the highest number of words in the subcorpus (cf. Figure 2). As beginner and lower intermediate learners are currently underrepresented in the L1 Portuguese – L2 Spanish subcorpus, one of our top priorities is to recruit learners at these lower proficiency levels.

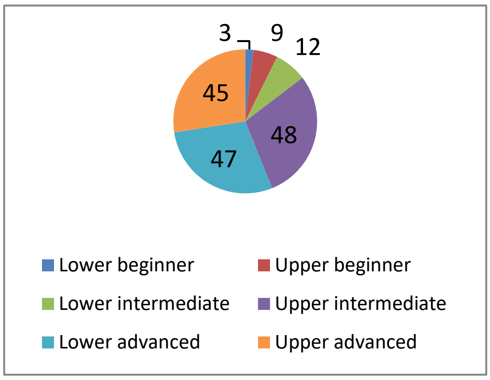


Figure 1: Number of participants in the L1 Portuguese- L2 Spanish subcorpus per proficiency level

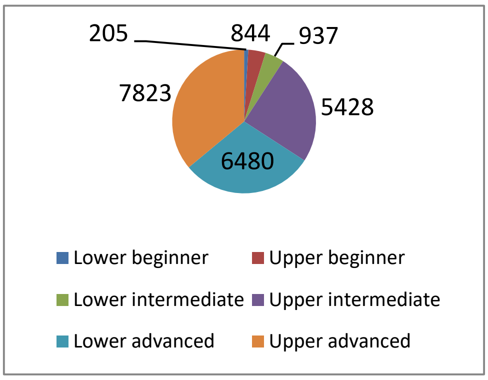


Figure 2: Number of words in the L1 Portuguese- L2 Spanish subcorpus per proficiency level

In addition to the L1 Portuguese – L2 Spanish learner subcorpus, the Portuguese component of CEDEL2 includes a control subcorpus with texts written in Portuguese. This subcorpus contains around 3,500 words from 16 native speakers of European Portuguese, with ages between 14 and 46, who completed the same film retelling task used to collect data from Portuguese learners of Spanish. The L1 Portuguese subcorpus was included in CEDEL2 because researchers need to contrast learner subcorpora with two types of control subcorpora: a native subcorpus of the learners' L1, to analyse possible L1 transfer effects, and a native subcorpus of the learners' L2, to analyse potential effects of input on L2 acquisition, determine whether learners' performance is native-like, and examine whether variability in learner language is either genuine learner variability triggered by representational/processing deficits or a reflection of the variability present in native grammars. Corpus data may reveal that the linguistic phenomenon under research is more variable in Spanish than assumed in theoretical models. Using two native control subcorpora thus allows researchers to better understand how the linguistic features of the learners' L1 and L2 shape their interlanguages.

To ensure maximum comparability, the L1 Portuguese subcorpus has the same design as the corresponding learner subcorpus, in line with Contrastive Interlanguage Analysis (Granger, 2015), as justified above. There are just two differences between learner and native subcorpora: (i) no placement test was administered; and (ii) the learner profile questionnaire only includes questions about age, gender, education, the participant's L1(s) and variety of Portuguese, the L1(s) and variety of Portuguese spoken by the participant's father and mother, the language(s) spoken at home, and additional languages (including proficiency level self-evaluation).

As shown in Figure 3, the two Portuguese subcorpora are the fourth largest learner and native subcorpora of CEDEL2 (version 2). Of all the learner subcorpora which combine two typologically similar languages, the L1 Portuguese – L2 Spanish subcorpus is the largest one. With a view to expanding the Portuguese component of CEDEL2, we are currently collecting more data from learners and native speakers. So far our focus has been on European Portuguese, but, in the future, data collection will be expanded to other varieties of Portuguese.

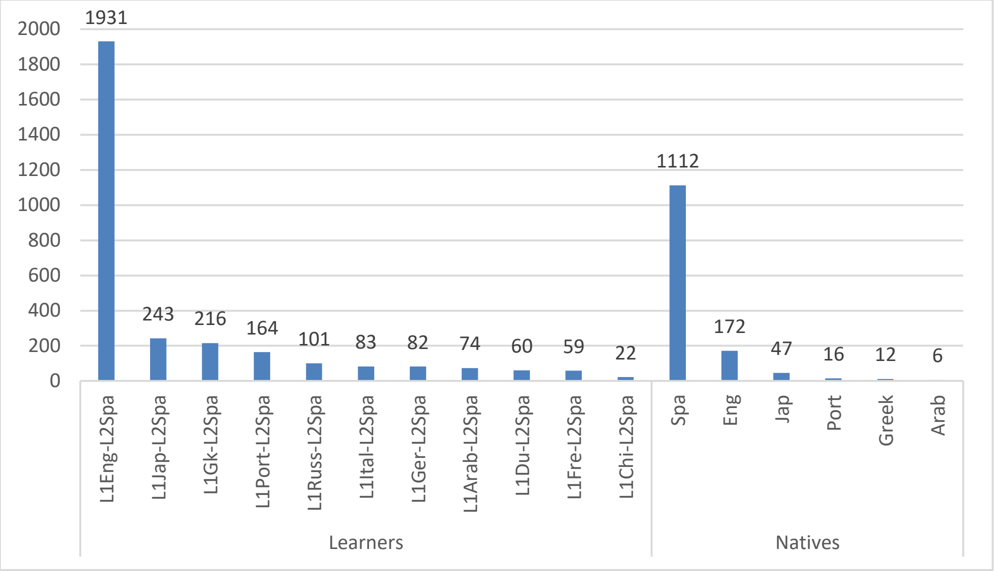
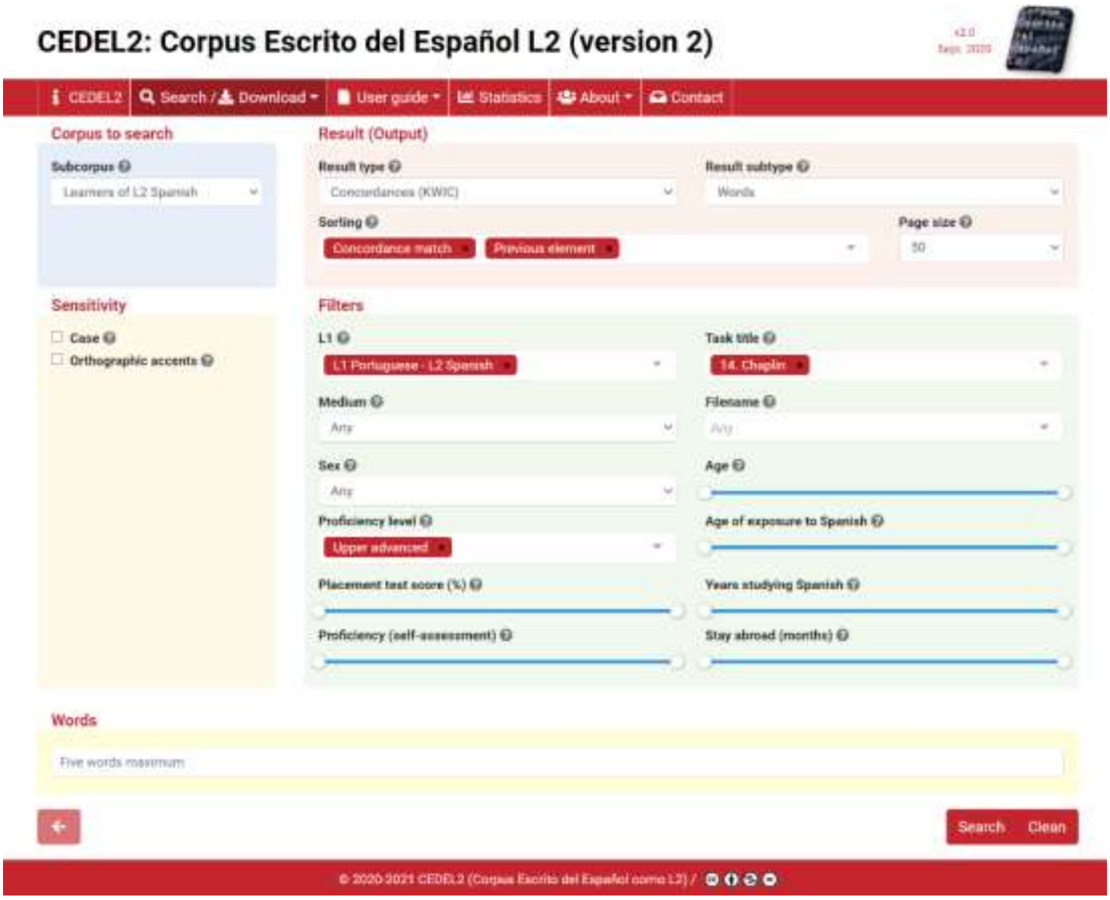


Figure 3: Number of files per subcorpus (CEDEL2 v.2)

Both the L1 Portuguese – L2 Spanish subcorpus and the L1 Portuguese subcorpus are available in open access on the newly developed CEDEL2 v.2 web-based interface: http://cedel2.learnercorpora.com (cf. Figure 4). This interface generates various types of outputs: (i) texts, which can be downloaded in several formats (TXT, TXT with metadata, CSV); (ii) concordances; (iii) simple frequency (frequency of the searched elements, e.g., word, lemma, out of the total number of words in the corpus and out of the total number of documents in the corpus); or (iv) full frequency (frequency of the searched element(s) according to eleven variables, including, e.g., L1, proficiency level, age of exposure to Spanish, among others). For concordances and simple/full frequency, the searchable element(s) can be words and word combinations, grammatical elements (part-of-speech tags and lemmas), words proxim (searches for a first word separated N words from a second word), and grammatical elements proxim (searches for a first grammatical word separated N words from a second grammatical word). The output can be filtered according to various learner/task filters which are relevant to SLA research, including learners' L1, proficiency level, placement test score in Spanish, self-evaluated proficiency level in Spanish on a 6-point scale, age, age of exposure to Spanish, length of instruction in Spanish, and length of residence in a Spanish-speaking country (for further details, cf. Lozano, in press).



CEDEL2: Corpus Escrito del Español L2 (version 2)

Corpus to search

Subcorpus: Learners of L2 Spanish

Result (Output)

Result type: Concordances (KWIC)

Result subtype: Words

Sorting: Concordance match, Previous element

Page size: 50

Sensitivity

□ Case

□ Orthographic accents

Filters

L1: L1 Portuguese - L2 Spanish

Task title: 14. Chaplin

Medium: Any

Filename: Any

Sex: Any

Age

Proficiency level: Upper advanced

Age of exposure to Spanish

Placement test score (%)

Years studying Spanish

Proficiency (self-assessment)

Stay abroad (months)

Words

Five words maximum

© 2020-2021 CEDEL2 (Corpus Escrito del Español como L2) /

Figure 4: CEDEL2 v.2 web-based search engine

5. Potential uses of CEDEL2 and its Portuguese subcorpus in L2 acquisition research

Learner corpora have a wide variety of users and uses (Díaz-Negrillo & Thompson, 2013), as illustrated in Figure 5. Users include theoretical and applied linguists (e.g., grammarians, SLA and foreign language teaching (FLT) researchers), professionals involved or interested in language teaching (e.g., materials writers, L2 teachers, language testers) and L2 learners. Learner corpora data can be used for applied purposes in the language-teaching industry (e.g., to develop dictionaries, teaching materials and profiles of learner performances at different levels) as well as in theoretical research, in particular in the domain of SLA (e.g., to trace the acquisition of an L2 by learners and to test particular hypotheses or models) (cf. Lozano, 2021b for a discussion).

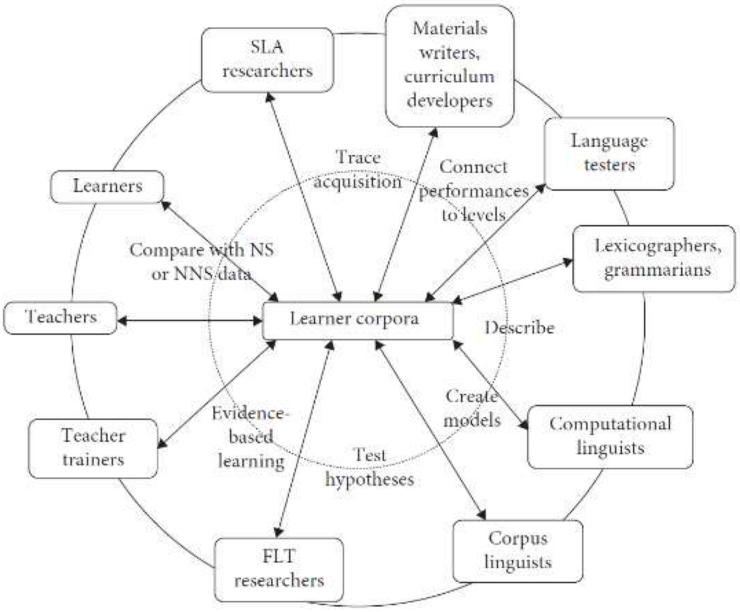


Figure 5: Users and uses of learner corpora (Díaz-Negrillo & Thompson, 2013, their Fig 2)

Thanks to their design features (e.g., same design across subcorpora, inclusion of metadata about SLA-relevant variables, dual native control subcorpora), CEDEL2 and its Portuguese subcorpora, in particular, allow us to investigate a wide range of topics in SLA, including but not limited to the following:

(i) L1 transfer effects (e.g., by comparing the L1 Portuguese – L2 Spanish subcorpus against the L1 Portuguese subcorpus and/or comparing the performance of Portuguese-speaking learners with that of learners with a different L1 background but the same level of proficiency);

(ii) L1 transfer effects vs. universal effects (by comparing learners whose L1 is similar to Spanish to learners whose L1 is different with respect to the linguistic phenomena under study);

(iii) language dominance effects (by considering variables such as the L1 of the learner's father, L1 of the learner's mother and language(s) spoken at home);

(iv) influence of other non-native languages (e.g., by comparing learners with different L2s for whom Spanish is an L3/Ln, which is often the case of Portuguese learners of Spanish);

(v) the role of typological proximity in L2 acquisition (by comparing same-proficiency learners with L1s that are typologically related to and distant from Spanish) and in L3 acquisition (by comparing L3/Ln Spanish learners with different L1-L2 combinations, where the two languages differ in their degree of typological proximity to Spanish, e.g., L1 Portuguese – L2 English and L1 English – L2 Portuguese);

(vi) sensitivity to microvariation in Romance languages (e.g., by comparing L2 Spanish learners who are L1 speakers of different Romance languages);

(vii) age-related and critical-period effects (by comparing learner groups with different ages and/or ages of exposure to L2 Spanish);

(viii) effects of exposure to naturalistic input on L2 development and ultimate attainment (by considering variables such as length of residence and the recency of the periods of residence in Spanish-speaking country/countries);
(ix) effects of instruction (by considering the length of instruction in Spanish);
(x) effects of construction-frequency and input variability (e.g., indirectly through the analysis of the L1 Spanish control subcorpus);
(xi) L2 developmental path (by comparing learners at different levels of proficiency);
(xii) endstate of L2 acquisition and fossilization phenomena (e.g., by analysing speakers with an upper advanced level and a long period of residence in a Spanish-speaking country vs. the L1 Spanish subcorpus).

CEDEL2 is thus a rich resource for SLA research, which makes available learner and native corpus data that can be used to complement experimental data, widening the empirical base of SLA research and strengthening the reliability and validity of its findings (Callies & Paquot, 2015). The triangulation of corpus and experimental data to study a given phenomenon is particularly effective when done in a cyclic fashion (Mendikoetxea & Lozano, 2018): the insights gained from corpus data can be tested experimentally and, in turn, experimental results may inform subsequent corpus analyses.

References

Alonso-Ramos, Margarita (Ed.) (2016) *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. John Benjamins. https://doi.org/10.1075/scl.78

Callies, Marcus & Magali Pacquot (2015) Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research 1*(1), pp. 1-6.

Campillos Llanos, Leonardo (2014) A Spanish oral learner corpus for computer-aided error analysis. *Corpora 9* (2), pp. 207-238.

Demuth, Katherine (1996) Collecting spontaneous production data. In. D. McDaniel, C. McKee & H. S. Cairns (Eds.) *Methods for Assessing Children's Syntax*. Cambridge, Mass.: The MIT Press, pp. 3-22.

Díaz-Negrillo, Ana & Paul Thompson (2013) Learner corpora: Looking towards the future. In. A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, pp. 9-29.

Domínguez, Laura, Nicole Tracy-Ventura, María J. Arche, Rosamond Mitchell & Florence Myles (2013) The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition 16* (03), pp. 558-577. https://doi.org/10.1017/S1366728912000363

Geeslin, Kimberly (Ed.) (2014) *The Handbook of Spanish Second Language Acquisition*. Wiley-Blackwell. http://doi.org/10.1002/9781118584347

Gilquin, Gaëtanelle (2015) From design to collection of learner corpora. In. S. Granger, G. Gilquin, & F. Meunier (Eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, pp. 9-34. https://doi.org/10.1017/CBO9781139649414.002

Gilquin, Gaëtanelle (2020) Learner corpora. In. *A Practical Handbook of Corpus Linguistics*. Springer, pp. 283-303. https://doi.org/10.1007/978-3-030-46216-1_13

Granger, Sylviane (2008). Learner corpora. In. A. Lüdeling & M. Kytoe (Eds.) *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, pp. 259-275.

Granger, Sylviane (2009) The contribution of learner corpora to second language acquisition and foreign language teaching. In. K. Aijmer (Ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins, pp. 13-32.

Granger, Sylviane (2012) How to use foreign and second language learner corpora. In. A. Mackey, & S. M. Gass (Eds.) *Research Methods in Second Language Acquisition: A Practical Guide*. Oxford: Wiley-Blackwell, pp. 5-29.

Granger, Sylviane (2015) Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research 1* (1), pp. 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, Sylviane, Estelle Dagneux & Fanny Meunier (2002) *The International Corpus of Learner English.* Louvain, Belgium: Université Catholique de Louvain.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (Eds.) (2015) *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Lozano, Cristóbal (2009a) CEDEL2: Corpus Escrito del Español como L2. In. C. M. Bretones & et al (Eds.) *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente*. Universidad de Almería, pp. 197-212.

Lozano, Cristóbal (2009b) Pronominal deficits at the interface: New data from the CEDEL2 corpus. In. C. M. Bretones & *et al* (Eds.) *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente*. Universidad de Almería, pp. 213-227. https://dialnet.unirioja.es/servlet/libro?codigo=520015

Lozano, Cristóbal (2016) Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. In. M. Alonso Ramos (Ed.) *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. John Benjamins, pp. 235-265. https://doi.org/10.1075/scl.78.09loz

Lozano, Cristóbal (2021a) Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. In. M. Cruz Piñol (Ed.) *E-Research y español LE/L2: Investigar en la era digital*. Routledge, pp. 138-163. http://doi.org/10.4324/9780429433528-9

Lozano, Cristóbal (2021b) Generative approaches. In. N. Tracy-Ventura & M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. Routledge, pp. 213-227. https://doi.org/10.4324/9781351137904

Lozano, Cristóbal (in press) CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research.*

Lozano, Cristóbal & Amaya Mendikoetxea (2013) Learner corpora and second language acquisition: The design and collection of CEDEL2. In. A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, pp. 65-100. https://doi.org/10.1075/scl.59.06loz

Lozano, Cristóbal & T. Quesada (in prep) *Anaphora resolution in L2 Spanish: What corpus data reveal about the Position of Antecedent Strategy (PAS)*.

Mackey, Alison & Susan Gass (2005) *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Martín-Villena, Fernando & Crsitóbal Lozano (2020) Anaphora resolution in topic continuity: Evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. In. J. Ryan & P. Crosthwaite (Eds.) *Referring in a Second Language: Studies on Reference to Person in a Multilingual World*. Routledge, pp. 119-141. http://doi.org/10.4324/9780429263972-7

McEnery, Tony & Andrew Hardie (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McEnery, Tony, Richard Xiao & Yukio Tono (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.

McEnery, Tony, Vaclav Brezina, Dana Gablasova & Jayanti Banerjee (2019) Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, *39*, 74-92.

Meisel, Jürgen M. (2011) *First and second language acquisition*. Cambridge: Cambridge University Press.

Mendikoetxea, Amaya (2014) Corpus-based research in second language Spanish. In. K. L. Geeslin (Ed.) *The Handbook of Spanish Second Language Acquisition*. Wiley-Blackwell, pp. 11-29.

Mendikoetxea, Amaya & Crsitóbal Lozano (2018) From corpora to experiments: Methodological triangulation in the study of word order at the interfaces in adult late bilinguals (L2 learners)". *Journal of Psycholinguistic Research 47* (4), pp.871-898.

Ministerio de Educación y Formación Profesional (2020) *El mundo estudia español*. Madrid: Secretaría General Técnica.

Mitchell, Rosamond, Laura Domínguez, María Arche, Florence Myles & Emma Marsden (2008) SPLLOC: A new database for Spanish second language acquisition research. In. L. Roberts, F. Myles & A. David (Eds.) *EUROSLA Yearbook 8*. Amsterdam/Philadelphia: John Benjamins, pp. 287-304.

Montrul, Silvina (2004) *The Acquisition of Spanish: Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and Adult L2 Acquisition*. John Benjamins. https://doi.org/10.1075/lald.37

Montrul, Silvina (2013) *El bilingüismo en el mundo hispanohablante*. Wiley-Blackwell.

Myles, Florence (2005) Interlanguage corpora and second language acquisition research. *Second Language Research 21* (4), pp. 373-391.

Myles, Florence (2015) Second language acquisition theory and learner corpus research. In. S. Granger, G. Gilquin, & F. Meunier (Eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, pp. 309-332.

Myles, Florence (2021) An SLA perspective on Learner Corpus Research. In. B. Le Bruyn & M. Paquot (Eds.) *Learner Corpus Research Meets Second Language Acquisition*. Cambridge University Press, pp. 258-273. https://doi.org/10.1017/9781108674577

Perdue, Clive (ed.) (1993) *Adult language acquisition: Cross-linguistic perspectives*. Vol. 1. Cambridge: Cambridge University Press.

Rojo, Guillermo & Ignacio Palacios Martínez (2016) Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. In. M. Alonso Ramos (Ed.) *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam/Philadelphia: John Benjamins, pp. 55-87.

Schwartz, Bonnie (2004) Why child L2 acquisition? In. J. V. Kampen & S. Baauw (eds.) *Proceedings of GALA 2003*. Utrecht: Netherlands Graduate School of Linguistics (LOT), pp. 47-66.

Sinclair, John (2005) How to build a corpus. In. M. Wynne (Ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow books, pp. 79-83.

Tono, Yukio (2003) Learner corpora: Design, development and applications. In. D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.) *Proceedings of the 2003 Corpus Linguistics Conference*. UCREL Technical Paper number 16, pp. 800-809.

Tracy-Ventura, Nicole & Florence Myles (2015) The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research 1* (1), pp. 58-95. https://doi.org/10.1075/ijlcr.1.1.03tra

Tracy-Ventura, Nicole, Florence Myles & Magali Paquot (2021) The future of corpora in SLA. In. N. Tracy-Ventura & M. Paquot (Eds.) *The Routledge Handbook of Second Language Acquisition and Corpora*. Routledge, pp. 409-424. https://doi.org/10.4324/9781351137904

Tracy-Ventura, Nicole & Magali Paquot (Eds.) (2021) *The Routledge Handbook of SLA and Corpora*. Routledge. https://doi.org/10.4324/9781351137904

Tracy-Ventura, Nicole, Rosamond Mitchell & Kevin McManus (2016) The LANGSNAP longitudinal learner corpus: Design and use. In. M. Alonso Ramos (Ed.) *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam/Philadelphia: John Benjamins, pp. 117-142.

University of Wisconsin. (1998) *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. University of Wisconsin Press. http://testing.wisc.edu/centerpages/spanishtest.html

Wynne, M. (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books.