This is an Accepted Manuscript of the following paper published in JSLT (Taylor & Francis / Routledge):
Lozano, C., & Fernández-Mira, P. (2022). Designing, compiling and interrogating corpora in L2 Spanish acquisition research. Journal of Spanish Language Teaching, 9(2), 190-206. https://doi.org/10.1080/23247797.2022.2157086

Designing, compiling and interrogating corpora in L2 Spanish acquisition

ABSTRACT

Despite the burgeoning field of Spanish second language acquisition (SLA) research, large Spanish learner corpora (LC) are not common practice yet. We present a general yet practical introduction to the multiple decisions Spanish as a second language (L2) researchers should consider before creating their own LC. We focus on (i) two freely available Spanish LC (CEDEL2 and COWS-L2H), (ii) their general design principles, (iii) crucial variables to collect (learner and task variables), (iv) ways of collecting and compiling LC data, and (v) the final product (the corpus interface). We explore different ways of interrogating the two corpora, illustrating them with specific (morpho)syntactic and lexical examples from L2 Spanish, as well as potential curricular and teaching applications of LC. We conclude with a recommendation for the triangulation of LC data with experimental research and a summary of future directions that the field of LC research may take. Our ultimate aim is to equip researchers with the basic theoretical and methodological tools to build and collect their own, well-designed LC.

KEYWORDS

Second language acquisition, Spanish learner corpora, learner corpus design, CEDEL2 (*Corpus Escrito del Español L2*), COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*)

RESUMEN

A pesar del reciente auge del campo de la investigación de la adquisición de español como segunda lengua (L2), el uso de corpus de aprendices (CA) sigue sin ser una práctica habitual. En este capítulo presentamos, de manera general a la vez que práctica, las múltiples decisiones a las que se enfrentan los investigadores de español L2 a la hora de crear su propio corpus. Nos centramos en (i) dos CA de español de acceso gratuito (CEDEL2 and COWS-L2H), (ii) sus principios de diseño, (iii) las variables relativas a los aprendices y a las tareas, (iv) maneras de recoger y compilar los datos y (v) el producto final (interfaces de búsqueda). Exploramos diferentes maneras de interrogar los corpus, ilustrándolas con ejemplos lingüísticos, y describimos posibles usos de esos datos tanto en la investigación como en la enseñanza. Concluimos con una recomendación de triangular datos de CA y experimentos y un resumen de los próximos pasos en el campo de la investigación de CA. Nuestra finalidad es equipar a los investigadores con herramientas básicas para compilar exitosamente su propio CA.

PALABRAS CLAVE

Adquisición de segundas lenguas, corpus de español L2, diseño de corpus de aprendices, CEDEL2 (*Corpus Escrito del Español L2*), COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*)

1. Introduction

Learner Corpus Research (LCR) is a growing field framed half-way between corpus linguistics and Second Language Acquisition (SLA) (Callies and Paquot 2015; Granger, Gilquin, and Meunier 2015; Le Bruyn and Paquot 2021; Tracy-Ventura and Paquot 2021). Learner corpora (LC) are large, systematic and purposely-designed linguistic databases that contain L2 learners' spoken/written language produced under conditions that range from classroom environments to researcher-led projects where, crucially, learners are free to choose their own means of expression. Importantly, LC contain linguistic and extra-linguistic variables (metadata) that provide information about the learner and the task.

L2 Spanish acquisition research has grown exponentially over the last decades (Montrul 2013, 2016; Geeslin 2014). However, even though some large Spanish LC have recently been created to contribute to research in this field, their use is not yet commonplace (Mendikoetxea 2014; Alonso-Ramos 2016; Rojo 2021). Indeed, SLA researchers have typically favoured experimental methods over corpus methods (Mackey and Gass 2016) due to the control over linguistic features under investigation in experimental settings and to the favouring of *deductive*, hypothesis-testing approaches (i.e., a hypothesis is formulated, and linguistic stimuli are then designed to test it). In LCR, instead, participants will choose their own wordings, with little control left to the researcher, and researchers will often adopt a more *inductive*, hypothesis-finding, data-driven approach. Despite their apparent polarization, LCR and experimental approaches are not dichotomic but rather gradient (Figure 2) and LC are increasingly being used to answer classic SLA questions and to test SLA hypotheses (see Lozano 2021b for a discussion and overview). For example, if well designed, a LC can address questions such as whether spoken data better reflect learners' competence than written data (cf. section 2.1.1); whether learners' production can be used as an additional measure of learners' proficiency level (apart from other objective measures like a placement test) (cf. discussion in section 4). To summarise, LC and experimental methods can be complementary, particularly when they are triangulated (see our final discussion).

[INSERT LOZANO FIGURE 2]

Figure 2. The inductive/deductive gradience in SLA.

In order to incentivize the use of LC and invite more researchers to create additional Spanish LC that offer useful tools to respond to SLA questions, this article will focus on two such corpora, CEDEL2 (version 2) (Lozano 2009, 2021c; Lozano and Mendikoetxea 2013) and COWS-L2H (Davidson et al. 2020; Yamada et al. 2020; Fernández-Mira et al. 2021), and contextualise them in relation to three other corpora: CAES (Rojo and Palacios Martínez 2016), LANGSNAP (Tracy-Ventura, Mitchell, and McManus 2016) and SPLLOC (Mitchell et al. 2008). The five LC have in common that they have been through a corpus design process, contain researcher-initiated tasks, and are freely available. Additional Spanish LC

can be found in the preceding chapter (this volume), in the references in the studies cited above, and at the *Indexador de Corpus de Aprendices de Español*[1].

Whereas these five LC differ (Figure 1), such differences make them complementary:

i. *Language modality*: They range from oral (SPLLOC) to written (CAES, COWS-L2H), with the others falling in between: LANGSNAP (1/3 written, 2/3 spoken), CEDEL2 (97% written, 3% spoken).

ii. *Learner profiles*: SPLLOC targets secondary-school and university British learners of Spanish. LANGSNAP includes British university learners before, during and after their immersion in a Spanish-speaking country. CAES samples learners at the Instituto Cervantes in different countries (6 L1s). CEDEL2 is varied in terms of L1s (11), setting (instructed/naturalistic), ages, and educational background. COWS-L2H's samples university US learners and heritage speakers (HS), and thus are relatively homogenous in age, L1 (mostly English and Spanish), and setting (instructed).

iii. *Corpus statistics*: The five corpora differ in size (Figure 1). Spoken corpora are logically smaller in size than written corpora due to the extra manual work required for the transcriptions (cf. section 2.1.1). According to Sinclair's (2005), the representativeness of the corpus is more important than its size (since corpora will vary in size). Texts in CEDEL2 and COWS-L2H vary in length (e.g., beginner texts tend to be shorter than very advanced learners) but, importantly, the texts have not been edited or shortened in any way, as also recommended by Sinclair.

iv. *Subcorpora*: The design of a corpus (multi- or mono-L1) determines its subcorpora. Figure 1 displays the number of learner and control subcorpora in each of the five focus LC.

v. *Metadata:* The higher the number of variables (metadata), the more filters researchers can use to select/discard participants or tasks that do not meet certain criteria. Figure 1 shows the variability in written and oral LC regarding this point. We discuss below the importance of including SLA-motivated variables (cf. Lozano and Mendikoetxea 2013; Lozano 2021c for full discussions).

[1] http://repositorios.fdi.ucm.es/corpus_aprendices_español/

[INSERT LOZANO FIGURE 1]

Figure 1. Comparison of five representative Spanish LC.

2. Theoretical framework

In this section, the basic process from the establishment of design principles (2.1) to data collection and compilation (2.2) and the creation of a final web interface (2.3) is discussed.

2.1. LC design principles

Table 1 shows a summary of design principles that will guide our discussion on LC design in the following three subsections.

Table 1. Summary of design principles discussed in subsections 2.1.1.-2.1.4.

Basic types	Medium (written/spoken) Genre Target language L1 (mono-/multi-L1) Diachrony (longitudinal/cross-sectional) Availability (commercial/free)
Basic design principles	Content selection Representativeness Same-design and contrast principles
Design recommendations	Homogeneity/heterogeneity of learners Single/double control corpora SLA-motivated learner and task variables
Learner's variables	Age, L1, proficiency level, age of exposure to L2, length of exposure to L2, length of residence in target country
Task variables	Title, resources, conditions, text

2.1.1. Basic types of LC

LC can be classified according to several criteria: medium, genre, target language, L1, diachrony, and availability (Gilquin 2015). In what follows, we discuss each in turn.

Medium: One of the key differentiating factors in any corpus is language medium: spoken vs. written. Logistically, the compilation of spoken LC requires substantial human resources, with the data collection typically happening *in* situ, in an individual manner, and audios having to be transcribed. By contrast, written data can be collected massively (i.e., a whole group of learners can participate simultaneously). Written data, if collected via online forms as in CEDEL2, or a Learning Management System (LMS) as in COWS-L2H, do not require any transcription or manipulation. For these reasons, spoken corpora (SPLLOC, LANGSNAP) are smaller in size than written corpora (CEDEL2, CAES, COWS-L2H), and written LC are more predominant in the field.

Written LC have greatly contributed to SLA in general and LCR in particular despite "SLA's insistence on the supremacy of spoken data" (Granger 2021, 248). Importantly, an innovative feature of some corpora like COREFL (Lozano, Díaz-Negrillo, and Callies 2021), which is being incorporated in CEDEL2, is the possibility of comparing written vs. spoken data by keeping the task and participant constant. This key characteristic of CEDEL2 allows researchers to investigate a classic question in SLA, namely, whether spoken data reflects L2 learners' competence better than written data, as traditionally assumed. It is also relevant to mention that, to the best of our knowledge, no Spanish LC has focused yet on micro-messaging and social media discourse which may present an interesting middle-ground between oral and written discourse.

Genre: Another criterion in corpus design is genre. As Paquot and Plonsky (2017) note, many LC contain university-level argumentative essays, which reflects the influence of one of the key LC in the past decades: the *International Corpus of Learner English (ICLE)* (Granger et al. 2020). However, increasing genre variety is beneficial because it allows SLA researchers to explore (or test) linguistic phenomena across genres. The focus corpora in this paper showcase genre variety: CEDEL2 contains argumentative, narrative and descriptive essays, while COWS-L2H includes the two latter genres. CAES adds in the epistolary genre in the form of letters, emails, and notes. LANGSNAP has oral interviews, oral picture-based narratives and written argumentative essays. Lastly, SPLLOC collects narrative, argumentative, and descriptive oral texts. Another possibility, which has not yet been explored much in LCR, is collecting data with genres more characteristic of naturalistic settings, such as tandem encounters or online L2 communications.

Target language: Most LC over the past two decades have included learners of L2 English, as can be seen in the statistics of the *Learner Corpora Around the World* website (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html). However, Spanish LC are on the rise due to the growing interest in Spanish SLA research.

L1: Depending on the participants' L1, corpora are classified into mono-L1 (SPLLOC) or multi-L1 (LANGSNAP, COWS-L2H, CAES, and CEDEL2). Multi-L1 LC allow researchers to investigate the influence of learners' L1, provided that all subcorpora are equally designed. An interesting aspect is whether the corpus contains typologically (un)related L1s, so that SLA researchers can test language-specific vs. language-universal L2 phenomena.

Diachrony: LC may sample the same group of learners over a period of time, as in the longitudinal LANGSNAP (which samples learners before, during and after their stay abroad) or COWS-L2H (which welcomes repeated participation as learners advance through their university language program), but this is logistically complex and time-consuming. Cross-sectional corpora are the norm as it is simpler to compile samples from different learners at different proficiency levels. Needless to say, an ideal developmental corpus would be

longitudinal, since the acquisition of single learners can be traced across time, which might uncover new perspectives on the variability of individual learner trajectories.

Availability: LC can be commercial (accessible for a fee or completely private) or freely available. The platforms of the five Spanish LC under review in this paper adhere to the Open-Science philosophy, and are thus freely available to the scientific community. Free access to data allows researchers, amongst other things, to conduct replication studies with the same data, which ultimately increases the validity and generalizability of the findings (Porte, 2012).

2.1.2 Design principles and recommendations

General design principles: Corpus builders have discussed general design principles for general corpora (Egbert, Larsson, and Biber 2020; Rojo 2021) and for LC (Granger 2008; Tono 2003, 2016). More specifically, Sinclair's (2005) 10 design principles for general corpora were adapted by Lozano and Mendikoetxea (2013) and Lozano (2021c) to the creation of LC. Following these authors, the essential design principles for LC design are:

i. **Content selection principle**. Design the corpus according to the *communicative function* and not the *language* of the texts. Therefore, the tasks should not be intended to elicit a particular linguistic construction or phenomenon. COWS-L2H intentionally keeps prompts brief and open-ended in order to elicit naturalistic texts and make them accessible for learners of any proficiency level (Yamada et al. 2020). CEDEL2 (version 1) originally contained 12 varied tasks to potentially elicit a wide range of vocabulary, linguistic structures, and verbal tenses (Lozano and Mendikoetxea 2013). Its current version 2 additionally contains two controlled tasks to elicit contexts with anaphora resolution (amongst other structures) (Lozano 2021c), in line with Tracy-Ventura and Myle's (2005) recommendations to incorporate both language-generic and language-specific tasks in LC.
ii. **Representativeness principle**. The corpus should represent the language it samples. In LCR, corpus builders have traditionally sampled written argumentative essays, as was mentioned earlier. Therefore, many LC will be representative of argumentative written language, but may provide little information about other genres and text types (spoken, dialogic, etc). CEDEL2 includes 14 tasks that sample different text types (argumentative, narrative, and descriptive). Similarly, COWS-L2H collects descriptive and narrative texts in all proficiency levels, although it is expanding into letters and emails.
iii. **Same-design and contrast principles.** When having different subcorpora within a large LC, ensure all subcorpora are equally designed. This allows for Contrastive Interlanguage Analysis (CIA) (Granger 2015), so between-subcorpus contrasts can be carried out. For example, the existence of difference L1 subcorpora in CEDEL2 would allow to compare beginner L1 English-L2 Spanish and beginner L1 Japanese-L2 Spanish learners on a given task. In COWS-L2H, the presence of a HS subcorpus would make it possible to compare HS and L2 learners who are enrolled in the same course level.

Basic recommendations*:* LC builders also propose the following design recommendations (Lozano 2021c; Tracy-Ventura, Paquot, and Myles 2021):

i. **Homogeneity vs heterogeneity of learners**. Most LC sample university-level learners of an L2, so these learners have been traditionally overrepresented (Paquot and Plonsky 2017). Gilquin (2015) recommends sampling varied and heterogeneous learners, as done in CEDEL2, which samples learners from 11 typologically (un)related L1s, in different learning settings (school/university/naturalistic), with different chronological ages (from teenage to senior learners), different types and

periods of exposure (years in classroom, months of stays/residence in Spanish-speaking countries), and proficiency levels (lower beginner to upper advanced and near-natives). COWS-L2H collects texts by learners enrolled at all university levels in language courses, including HS of Spanish, a subset of the learner population that is not usually represented in LC.

ii. **Single vs. double control corpora.** Well-designed LC should incorporate control corpora as it is practical to have a reference corpus of the language variety the learners are aspiring to acquire (Granger 2021). This way, researchers can study whether learners' output approximates to/distances from/deviates from natives' output. CEDEL2 contains data from native speakers who use different varieties of Spanish, such as peninsular Spanish and Spanish from all Spanish-speaking Latin American countries. This type of reference corpus is particularly interesting because: (i) it reflects different varieties of the target language learners are aspiring to acquire; (ii) it can account for some interlanguage phenomena which may be result from learners' contact with a given variety of Spanish. For example, it is well known that varieties of Caribbean Spanish are prone to overusing personal pronouns. Therefore, learners who have been in contact with these varieties may also show an increase in personal-pronoun production in relation to learners who have not. At the time of writing, COWS-L2H is collecting data to build a reference corpus that will sample speakers from different varieties of Spanish, including US Spanish as spoken in California, where the data is collected. Additionally, LC designers may benefit from incorporating an additional control corpus of the learners' L1, as is the case in CEDEL2. This double-control model includes: (i) L1 control subcorpora for each of the L1s, (ii) native control subcorpora in the L2 of the learners, and (iii) learner subcorpora organized by L1. Such multi-control structure allows researchers to detect (i) L1 influence, (ii) target-language influence, and (iii) interlanguage features that cannot be attributed to either the learners' L1 or their target language.

iii. **SLA-motivated variables**. For an LC to be maximally usable by SLA researchers, the learner and task variables to be collected must be motivated by key questions in SLA (Lozano 2021c). The metadata of each file must thus contain variables related to the learners' profile and their task, as will be described in the following two subsections.

2.1.3. Learner variables

Many LC collect ad-hoc learner variables (e.g., course/class year, L1, age, etc.), but, for the corpus to be maximally usable by SLA researchers, it should register SLA-motivated variables that allow researchers to address fundamental SLA questions (see Lozano 2021c, Table 10, for a detailed discussion). These variables or metadata can be later used to filter learners that do not meet certain criteria, for example. Potentially relevant learner variables can be classified into:

i. **Basic biodata:**
   a. *Chronological age*: to test for cognitive-development effects on linguistic development (e.g., discourse coherence in teenage vs adult learners; development of a linguistic phenomenon across the lifespan).
   b. *Sex*: to test for male/female linguistic differences.

ii. **Linguistic background variables:**
   a. *L1(s)*: to examine the influence of the learners' L1(s).
   b. *Country*: to test for variation in the learners' L1 (e.g., L1 British English vs. American English).
   c. *Language dominance*: to test for bilinguals' dominance/use of one language over the other, as well as likely influence of the father's/mother's L1,

languages spoken at home/work/school, etc. The dominance variable can be further enriched via scores from purpose-built tests like the Bilingual Language Profile (BLP) (Birdsong, Gertken, and Amengual 2012).

d. *Proficiency level*: to test developmental effects across proficiency levels. Traditionally, many LC have used ad-hoc measures to categorize learners into different proficiency levels, such as the course/year the learners are in, but this approach has obvious limitation and needs to be complemented with objective and subjective proficiency measures, such as standardized and validated placement tests, as done in CEDEL2, and/or learners' self-ratings, as in CEDEL2 and COWS-L2H.

e. *Other foreign languages, proficiency level, and sequence of acquisition (L2, L3, Ln):* to test for possible effects of the L2 on the L3 or vice versa. In CEDEL2, learners self-rate (in each of the four skills: speaking, writing, listening, reading) their proficiency level in any other additional foreign language other than L2 Spanish via a 6-point scale (lower/upper beginner, lower/upper intermediate, lower/upper advanced).

iii. **Linguistic experience variables:**

a. *Age of Onset* (AoO) in a naturalistic setting or *Age of Exposure* (AoE) to the L2 in the classroom: to test how it affects ultimate attainment in critical-period studies that compare, e.g., early vs. late arrivals in the target country.

b. *Length of Exposure* (LoE) or Length of Instruction (LoI) in an academic setting: to test how length affects the learners' linguistic outcomes.

c. *Length of Residence* (LoR) in the target country/countries: to test whether residence (in months/years) positively affects linguistic outcomes.

iv. **Socio-educational variables**:

a. *Educational level* and *socioeconomic status* (SES): to test whether the learners' (/parents') educational level and/or their status affects L2 acquisition success.

b. *Educational background*: to test whether specific educational knowledge (e.g., learners in language degrees vs. science degrees) impacts on L2 success.

2.1.3. Task variables

In addition to learner metadata, a LC should incorporate metadata pertaining to the task, e.g.:

i. **Task title**: Details about the title or the visual prompts used to elicit the data.

ii. **Task resources/tools**: Whether the learners had access to resources prior to, or while, completing the task, and, if so, the types of resources used.

iii. **Task conditions**: Whether the task was done in class or outside class, whether it was timed or untimed, whether it was part of an exam or a free production, the conditions under which the task was recorded or details about the interviewer (for spoken tasks), etc.

iv. **Text type**: The actual text produced by learners can be of different types (narrative, descriptive, argumentative, etc).

The five focus corpora offer the possibility of controlling for the abovementioned task variables, which is important because different aspects of the task have been shown to impact learner output. For instance, different SLA studies have reported on an effect of task complexity and narrative time (present, past, future) on writing outcomes (Ellis and Yuan 2004; Kuiken and Vedder 2008; Cho 2019). Castañeda-Jiménez and Jarvis (2014) found significant differences in lexical diversity (LD) depending on whether the samples were argumentative or descriptive. Fernández-Mira et al. (2021) also found differences in L2 Spanish learners' LD when analyzing writing samples with the same genre but different topic. Tracy-Ventura and Myles (2015) found that the degree of control of the task has an impact on

learners' performance when using past tenses. In short, LC should include a wide range of tasks to reflect the full range of learners' competence.

Spoken or partially spoken corpora such as SPLLOC and LANGSNAP, because of the investment of resources they require, tend to be carefully crafted and provide participants with a series of shorter tasks that are inserted in an interview or picture-based narrative format, or both. As for written corpora, they tend to vary a lot in genre, topic, and writing conditions. As was mentioned before, CEDEL2 has 14 topics for learners to choose from. CAES has two or three prompts depending on the level, all of them guided tasks. COWS-L2, with eight prompts so far, keeps topics consistent across proficiency levels. Providing the same prompts to the entire pool of participants has the advantage of making cross-sectional comparisons more reliable, as the effect of proficiency level and topic can be assessed separately.

In short, the group of available Spanish LC compared in this article present a high level of heterogeneity at the task level. Such variety, paired with large enough corpora, is enriching since it allows researchers to test for multiple task factors that may have an impact on learners' production by using and comparing data from the different LC.

2.2. LC data collection and compilation

Once the corpus builder determines the corpus design principles/recommendations/variables, data collection can be launched.

2.2.1. Data collection

Since 2006, CEDEL2 written data are being collected via web-based forms (http://learnercorpora.com), which is a valid research method in SLA (Wilson and Dewaele 2010) and allows for larger participant samples from varied contexts that are not limited to L2 classrooms (Gilquin, 2015). Importantly, the online participation forms are in the participant's L1 to ensure that even beginners can understand what they are requested to do. The forms are divided into six sections:

i. Section 1 informs participants about the CEDEL2 project, how they can participate and what they can get from it.
ii. Section 2 is a detailed consent form.
iii. Section 3 asks for the basic learners biodata (all anonymized).
iv. Section 4 asks for the learners' linguistic background metadata.
v. Section 5 asks for the task metadata.
vi. Section 6 is a standardized Spanish placement test (University of Wisconsin 1998).

The online-form's metadata is automatically collected in a spreadsheet. The spreadsheet needs a considerable amount of manpower and manual work (standardization of certain metadata, cleaning irrelevant data, etc.) it can be used as a data source for a search engine (see section 2.3 for details on the web interface of CEDEL2 and COWS-L2H). The last step is the design of a platform/software that should contain a user-friendly and attractive search engine.

When collecting data online, volunteers need an incentive to participate. CEDEL2 participants get their placement-test score and a certificate of participation (written participation), as well as monetary compensation in the case of spoken participation (see section 2.2.3 below on conducting oral recordings). The corpus data collector can advertise the 'calls for participation' via distribution lists (e.g., the Linguist List), social media, or participant-recruitment systems (e.g., Amazon Mechanical Turk).

COWS-L2H collects data within a single university in the US through their LMS. Every academic term, all students enrolled in Spanish language courses receive a call for participation that offers extra credit points in exchange for written responses to the corpus' prompts. Participants sign up for the COWS-L2H Canvas site, where a consent form, a linguistic background questionnaire, and the prompts for two writing assignments are posted.

The first assignment is completed during the fourth week of the academic term, and the second during the eighth week. Participants have a one-week window to submit each of their writing samples. At the end of each term, the texts and corresponding metadata are downloaded by the corpus manager team.

COWS-L2H's process has been streamlined to collect large amounts of data with minimal time investment on the part of the corpus managers. Their tasks every given term are: 1) launching the call for participants, either in written or face-to-face form, 2) designing the LMS site reusing the content from previous terms, 3) selecting two new prompts if the current prompts have been in place for four academic terms, 4) scheduling regular reminders for the participants to complete the assignments, 5) giving out extra credit information to the instructors, and 6) downloading the data. Crucially, this repetitive and systematic data collection set up encourages students' repeated participation, resulting in a large amount of longitudinal data. At the moment, 34% of participants have completed writing assignments in two or more terms.

2.2.2. Data processing and anonymization

After collecting the samples, the job of the corpus team is far from being finished and two distinct steps of lengthy manual intervention may be required. First, data need to be cleaned, anonymized and standardized following a consistent protocol for maximum homogeneity. For example, COWS-L2H's data are anonymised each summer, the time of year when enough resources can be allocated to this task, and published one year after they were originally collected. CEDEL2 data are anonymised prior to the launching of a new version on its online platform (http://cedel2.learnercorpora.com). The anonymization step is particularly important according to current data protection legislation and to abide by the consent form signed by participants and approved by an ethical committee (see ethical and consent considerations in Costa et al. 2019; Thomas and Pettitt 2017). Second, since in both CEDEL2 and COWS-L2H, participants are not supervised when writing the research teams need to manually check every single text to discard ungenuine writing that resulted from the use of online translators, for example.

Concerning this last point, CEDEL2 specifically asks participants whether they used spelling checkers and/or (monolingual/bilingual) dictionaries, which can provide interesting data in and of itself. It is noteworthy that these unsupervised data collection conditions contrast with the convenience sampling strategy used in other corpora, which compile already existing texts, such as in-class compositions or exams. Additionally, from the point of view of participants, writing for a corpus/research project (a low-stakes situation with no incentive to "cheat") is very different from writing for class (for the instructor to read and grade). These divergent conditions need to be adequately considered and described in the compilation and publication phase, and certainly need to be accounted for when drawing comparisons between LC.

2.2.3. Conducting oral recordings

As was mentioned earlier, LC are typically either written or spoken, and bimodal corpora (written + spoken) are far from being the norm, despite their benefits. CEDEL2 is mainly written, though the spoken component is increasing, through in-situ and online oral recordings that are allowing to obtain data from learners in different countries with very diverse linguistic backgrounds. Importantly, participants and tasks are kept constant across modalities so that the same learner completes the same task twice: once orally and once in writing (cf. the paragraph *Medium* in section 2.1.1. for the written vs. spoken discussion in LCR).

In these bimodal corpora, and in more traditional oral LC, the transcription phase requires considerable human resources, as all transcribers should follow a strict protocol to

ensure homogeneity across transcriptions. Transcription conventions (e.g., pauses, repetitions, false starts, incomprehensible speech, etc.) should be reflected in the protocol (cf. CEDEL2 transcription conventions at http://cedel2.learnercorpora.com/user_guide/conventions).

2.3. LC final product: web interface

After designing and compiling the corpus, the final step is to reach the highest number of potential users, so the authors should make it publicly available in line with the Open Science philosophy and under a Creative Commons licence (https://creativecommons.org), either via repositories (like COWS-L2H) or via dedicated websites with a purpose-built interface that includes a search engine (CEDEL2).

The CEDEL2 interface (http://cedel2.learnercorpora.com) contains several sections: basic info about CEDEL2; an online user guide (user manual plus technical details: corpus design, metadata, transcription conventions, and tags); general corpus statistics; information about the research team and publications that have used the corpus as a source of data. Importantly, it contains the actual search engine where *basic searches* (strings of characters or words, including the use of wildcards) can be conducted in any/all of the learner/native subcorpora. The *advanced search* layout is also user friendly (Figure 3) and allows for several search (as well as download) options (cf. the online user guide and Lozano 2021c for details). The search output (concordances/texts) can be downloaded in several formats (txt, txt with metadata, CSV, audio files, etc.), which allows users to search/download the following:

i. The query output can have different formats: concordances (i.e., traditional keyword-in-context, KWIC), texts (written texts or the transcription of the spoken texts), simple/complex frequencies of the words/expressions to be searched.
ii. Apart from words, CEDEL2 allows for sophisticated, linguistically-informed queries, e.g., lemmas (e.g., ESTAR would yield all verbal forms of *estar* 'to be'), grammatical elements (e.g., word-order combinations like *Adjective + Noun* vs. *Noun + Adjective*), and other sophisticated search combinations.
iii. The output can be sorted according to several variables (e.g., L1, medium, age, placement score, task title, filename). Concordance outputs can be additionally sorted by several criteria (concordance match, previous/next element, second previous/next element).
iv. Importantly, searches can be filtered according to a series of variables. Recall that these filters are part of SLA-motivated learner and task metadata (cf. 2.1.2). Users can thus target concordances/texts that meet certain specific criteria based on metadata (learners' L1, sex, age, proficiency level, placement test score, proficiency self-assessment, AoE, LoE, LoI, LoR, medium (written/spoken/written&spoken by the same person), task title, and filename).

[INSERT LOZANO FIGURE 3]

Figure 3. CEDEL2 search and download interface.

COWS-L2H shares its data in a public repository hosted on GitHub (https://github.com/ucdaviscl/cowsl2h), which is organized in folders, one for each corpus prompt, and further divided into subfolders, one for each data collection period. Learner essays are available in raw TXT format, and metadata has been extracted from learner questionnaires and is also provided as TXT files–both essay and metadata documents are titled with the participant's ID. In addition to the raw data, some subfolders also contain annotated and corrected versions of the essays. Annotations were manually performed following an annotation scheme (also in the repository) for two types of common L2 Spanish errors: gender/number agreement and accusative 'a'. The corrected essays are alternate versions of the learner essays that use more native-like language, as if an instructor were to "translate" a student's production. More annotation and correction efforts are currently ongoing, through a collaboration with researchers from the Universidad de Salamanca.

There are two additional directories in the repository: one containing CSV files that aggregate the metadata and essays for a specific prompt and data collection period, and another including Python scripts that have been used to process the essays, calculate inter-annotator agreements, extract lexical diversity measures, or perform automatic part-of-speech tagging with Freeling. This setup requires that researchers looking to work with COWS-L2H use coding skills to load, read, parse TXT or CSV files, run the available scripts and edit them as necessary for their purposes. Alternatively, researchers with no programming skills could manually open the CSV files as spreadsheets and filter data to search for specific information.

[INSERT LOZANO FIGURE 4]

Figure 4. Flowchart: from corpus design to final product.

3. Uses and implications for L2 Spanish research and teaching

We conclude with two practical issues: (i) how to interrogate the corpora, and (ii) how to use them for curriculum design purposes.

3.1. L2 Spanish research: Interrogating and exploiting CEDEL2 and COWS-L2H corpora

In this section, we illustrate how to interrogate the corpora by doing (morpho)syntactic searches in CEDEL2 and calculating lexical diversity (LD) indexes in COWS-L2H, even though both corpora allow for more sophisticated searches (cf. their respective websites for further instructions).

In CEDEL2, we will focus on three phenomena: (1) Adjective-Noun (Adj-N) order; (2) Article-Noun gender agreement; (3) *ser*/*estar* contrasts. Additional examples of phenomena that can be studied in CEDEL2 (e.g., Verb-Subject order like *Existe un problema aún más grave; Llega un policía*) can be found in Lozano (2021a, 2021c).

Adj-N order

Justification: L1 English-L2 Spanish learners typically produce Adj-N word order (**acogedores bares* 'cozy bars'*; *alcohólicas bebidas* 'alcoholic drinks') (see Lozano 2014 for an overview of this and other syntactic phenomena in L2 Spanish). In the CEDEL2 interface (Figure 3), follow the steps in (1) to get the concordance output (Figure 5). As can be observed, lower beginner L1 English-L2 Spanish learners produce ungrammatical Adj-N order (*diferente personas, fabuloso hora, famosa persona*). This search can be contrasted against a different subcorpora (e.g., L1 English vs. L1 Italian) to study the effect of the L1 syntax on the L2 syntax.

Search in CEDEL2: CEDEL2 website > Search > Advanced Search > Result type: *Concordances* > Result subtype: *Grammatical elements* > Sorting: *Concordance match* > L1: *L1 English-L2 Spanish* > Proficiency level: *Lower beginner* > Grammatical elements: click on the *Tag* symbol and choose *Adjective* from the Category drop-down list > Click on the plus symbol (✚) to add a new grammatical-element search term > Click on the *Tag* symbol and choose *Noun* from the Category drop-down list > Click on *Search*.

[INSERT LOZANO FIGURE 5]

Figure 5. CEDEL2 concordance (Adj-N order).

Article-Noun gender agreement

Justification: Spanish learners often agree masculine nouns ending in *-a* (a morpheme typically marking feminine gender) like *clima* with the feminine article *la* (**la clima* 'the climate') instead of the grammatically correct masculine article *(el clima*). In CEDEL2 we can search for this, (3). The concordance output (Figure 6) will contain **la clima,***la día* 'the day', **la idioma* 'the language'. It is thus possible to study whether this error decreases as proficiency level increases by selecting the corresponding proficiency levels in the Filters section.

Search in CEDEL2: CEDEL2 website > Search > Advanced Search > Result type: *Concordances* > Result subtype: *Grammatical elements* > Grammatical elements: click on the

Tag symbol and choose the following: Category *Determiner*, Type *Article*, Gender *Feminine* > Click on the plus symbol (✚) to add a new grammatical-element search term > Click on the *Tag* symbol and choose the following: Category *Noun*, Gender *Masculine* Number *Singular*, and finally type in "**a*" in the Word box (which will look for any noun ending in *a*) > Click on *Search*.

[INSERT LOZANO FIGURE 6]

Figure 6. CEDEL2 concordance (*la* + masculine noun ending in *-a*).

***Ser*/*estar* contrasts**

Justification: L2 Spanish learners often confuse the two copular verbs *ser* and *estar* that can both be translated by 'to be', and produce ungrammatical sentences like **Yo era en Londres* (instead *Yo estaba en Londres* 'I was in London'). In this context, we can search for the lemma *ESTAR* (i.e., all verb forms of the verb estar: *soy, eres, es … era, eras, … serás, … seríamos, … fuésemos …*) followed by the preposition *en*, as in (3), which will retrieve sentences such as **Cuando era en el colegio* 'When I was at school', **Todos los días era en Nueva York* 'Every day I was in New York'. The full concordance for this search can be found in Figure 7. You can then filter the concordances by L1 to investigate whether this error is L1-dependent or rather L1 independent.

Search in CEDEL2: CEDEL2 website > Search > Advanced Search > Result type: *Concordances* > Result subtype: *Grammatical elements* > Click on any filters you wish > Type in "ser" in the *Lemma* box > Click on the plus symbol (✚) to add a new grammatical-element to the search > Type in "en" in the *Grammatical element* box > Click on Search.

[INSERT LOZANO FIGURE 7]

Figure 7. CEDEL2 concordance (lemma *SER* + preposition *en*).

LC can also provide valuable insights into learners' productive vocabulary knowledge. Specifically, lexical diversity (LD), or how many different words (total number of types) are used in a text in relation to how long it is (total number of tokens), is a lexical measure that has been widely used in LCR and vocabulary research as a strong predictor of L2 lexical abilities (Castañeda-Jiménez and Jarvis 2013). Additionally, it has the advantage of not requiring part-of-speech tags and not being language-specific.

What follows details how to calculate LD indexes using COWS-L2H's publicly available data and scripts. The reader is referred to Fernández-Mira et al. (2021) and Sánchez-Gutiérrez and Fernández-Mira (2022) for additional analyses and interpretations of the influence of learner and task variables on LD scores in COWS-L2H texts.

Search in COWS-L2H: COWS-L2H GitHub repository[2] > "Code" (on the top right corner) > "Download ZIP" (see Figure 8).

[2] https://github.com/ucdaviscl/cowsl2h

[INSERT LOZANO FIGURE 8]

Figure 8. How to download COWS-L2H.

After decompressing the ZIP file, open the computer's command prompt/terminal and navigate to the location of the corpus, named "cowsl2h-master", and then to the subdirectory "scripts." Now, to compare LD scores of texts with the prompts "A special person in your life" ("special") and "A terrible story" ("terrible"), run the Python file "lexdiv_special_terrible.py" by typing the name of the file followed by two elements: 1) the directory for the "special" prompt and 2) the directory for the "terrible" prompt. Figure 9 shows an example of that command in the first line of the terminal. The output should be the average LD by prompt and course, in list and graph form, as shown in Figure 9.

[INSERT LOZANO FIGURE 9]

Figure 9. Result of running lexdiv_special_terrible.py.

This script can be used as a template and edited in a Python interpreter in order to analyse and compare the LD indexes from prompts in the corpus and using any other metadata.

3.2. L2 Spanish teaching: The curriculum and Spanish LC

In addition to its contributions to SLA research, the field of LCR has great potential to influence the curricular design and teaching of L2 courses. This section provides an overview of the pedagogical applications of COWS-L2H, a LC that is tied to a single higher-education institution. Said university follows a uniform curriculum and textbook for each Spanish language series: Introductory, Intermediate, Composition and Heritage, which allows to match each student sample with the contents studied during the specific week (and the previous weeks) they are writing for the corpus. Given these characteristics, COWS-L2H can be used to learn more about, for instance, how students enrolled in a given course write before and after a linguistic topic is introduced and practiced, how two series (such as Composition and Heritage, which give access to the same upper-division courses) compare in terms of writing skills, or how individual students' writing develops longitudinally as they progress through the series. This kind of research can inform pedagogical decisions such as when and how to teach certain topics or how to supplement observed deficits and gaps. This is especially relevant in a context where many sections of the same course are offered at once (as is the

case for this university), given that any curricular changes need to be implemented at a large scale and therefore carefully considered in advance.

Conversely, it is also possible to observe and evaluate how specific curricular or instructional innovations affect students' written production outside of their normal coursework. COWS-L2H has been collecting data every term since 2017. In 2019 the Introductory Spanish series went through a radical change in structure, content and textbook, and in 2020 the COVID-19 pandemic hit, and all courses had to be delivered in a virtual format for a year and a half. Data from the corpus could be used to study the impact of these events on learners' writing skills, and learn lessons for future planned or unplanned transitions. Similarly, as online media and discourse begin playing a bigger role in L2 instruction, we could see COWS-L2H written texts moving along the written-spoken continuum. In this context, the collection of similar corpora at different institutions would allow for useful comparisons across contexts and to differentiate developmental pathways that depend on specific curricula from those that are shared by students at various universities, for example.

4. Spanish LC: The way ahead

Spanish LC in SLA research: While LC provide highly contextualised and ecologically valid naturalistic data (Figure 2), they are but one of the many research methods used in SLA and have traditionally been underused in this type of research. However, researchers are increasingly advocating for the triangulation of different research methods, including LC, to investigate language acquisition phenomena through different perspectives (Gilquin 2021; Mendikoetxea and Lozano 2018). This new perspective on SLA research may encourage the creation of more Spanish LC that come to complement the ones already in existence and promote the development of new LCR projects in languages other than English.

Spanish LC in research on individual differences: It would be desirable that new LC include more metadata that tap into fine-grained variables such as aptitude, attitude, memory, or language dominance. However, there is a trade-off between how long it takes participants to complete all parts of a study and how many participants end up taking part in it. By increasing the number of questions that participants respond to, the amount of information about each of them is greater, but additional questions may also deter many potential participants from participating altogether. The creation of more LC, if all compile similar metadata, could be an avenue to multiply sources of information even if each corpus is not as large on its own.

Spanish LC and technology: Spanish LCR has room to grow at its intersection with the field of computational linguistics (particularly natural language processing) and educational technology. These disciplines, historically less advanced in languages other than English, require large amounts of data, obtained from the Web, in order to train deep learning models that classify or predict texts. Since learner language is in many cases indistinguishable from native language on the Web, Spanish LC data represent an invaluable resource for computational applications such as automatic proficiency level assessment, error correction, or tailored feedback. However, if LC are to make an impact in SLA research, more similarly-designed LC need to be created to increase the number of samples linked to the same pieces of metadata, since it is about time that the explanation of L2 acquisition be backed up at a macro level. While big data is an essential component of many scientific disciplines, big LC data should ultimately be the result of careful corpus design and data collection processes, as illustrated in this paper.

Spanish LC and curricular development: At an institutional level, LC help language programs better understand learners' skills in order to improve their curriculum. The creation of more LC at single institutions could allow said institutions to, for instance, create their own

placement tests based on LC data, or fill in gaps in learner knowledge when enrolled in different courses.

References

Alonso-Ramos, M. 2016. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. John Benjamins.

Birdsong, D., L. M. Gertken, and M. Amengual 2012. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. COERLL.

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.

Castañeda-Jiménez, G. and S. Jarvis 2013. "Exploring Lexical Diversity in Second Language Spanish." In *The Handbook of Spanish Second Language Acquisition*, ed. K. Geeslin, 498–513. Malden, MA: Wiley & Sons.

Cho, M. 2019. "The Effects of Prompts on L2 Writing Performance and Engagement." *Foreign Language Annals* 52 (3): 576–594.

Costa, P. I. D., J. Lee, H. Rawal, and W. Li 2019. "Ethics in Applied Linguistics Research." In *The Routledge Handbook of Research Methods in Applied Linguistics*, eds. J. McKinley and H. Rose, 122–130. Routledge.

Davidson, S., A. Yamada, P. Fernández-Mira, A. Carando, C. H. Sánchez-Gutiérrez, and K. Sagae 2020. "Developing NLP Tools with a New Corpus of Learner Spanish." *Proceedings of the 12th Language Resources and Evaluation Conference 2020*: 7238–7243.

Egbert, J., T. Larsson, and D. Biber 2020. *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Cambridge University Press.

Ellis, R. and F. Yuan 2004. "The Effects of Planning on Fluency, Complexity, and Accuracy in Second Language Narrative Writing." *Studies in Second Language Acquisition* 26 (1): 59–84.

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.

Geeslin, K. L. 2014. *The Handbook of Spanish Second Language Acquisition*. Wiley-Blackwell.

Gilquin, G. 2015. "From Design to Collection of Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, eds. S. Granger, G. Gilquin, and F. Meunier, 9–34. Cambridge University Press.

Gilquin, G. 2021. "Combining Learner Corpora and Experimental Methods." In *The Routledge Handbook of Second Language Acquisition and Corpora*, eds. N. Tracy-Ventura and M. Paquot, 133–144. Routledge.

Granger, S. 2008. "Learner Corpora." In *Corpus Linguistics: An International Handbook*, eds. A. Lüdeling and M. Kytoe, 259–275. Mouton de Gruyter.

Granger, S. 2015. "Contrastive Interlanguage Analysis: A Reappraisal." *International Journal of Learner Corpus Research* 1 (1): 7–24.

Granger, S. 2021. "Have Learner Corpus Research and Second Language Acquisition Finally Met?" In *Learner Corpus Research Meets Second Language*, eds. B. Le Bruyn and M. Paquot, 243–257. Cambridge University Press.

Granger, S., M. Dupont, F. Meunier, H. Naets, and M. Paquot 2020. *The International Corpus of Learner English. Version 3.* Presses universitaires de Louvain.

Granger, S., G. Gilquin and F. Meunier 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Kuiken, F. and I. Vedder 2008. "Cognitive Task Complexity and Written Output in Italian and French as a Foreign Language." *Journal of Second Language Writing* 17 (1): 48–60.

Le Bruyn, B. and M. Paquot 2021. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge University Press.

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 et al., 197–212. Universidad de Almería.
Lozano, C. 2014. “Word Order in Second Language Spanish”. In *The Handbook of Spanish Second Language Acquisition*, ed. K. L. Geeslin, 287–310. Wiley-Blackwell.
Lozano, C. 2021a. “Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2”. In *E-Research y español LE/L2: Investigar en la era digital*, ed. M. Cruz Piñol, 138–163. Routledge.
Lozano, C. 2021b. “Generative Approaches”. In *The Routledge Handbook of Second Language Acquisition and Corpora*, eds. N. Tracy-Ventura and M. Paquot, 213–227. Routledge.
Lozano, C. 2021c. “CEDEL2: Design, Compilation and Web Interface of an Online Corpus for L2 Spanish Acquisition Research.” *Second Language Research*.
Lozano, C., A. Díaz-Negrillo, and M. Callies 2021. “Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL”. In *What’s in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*, eds. C. Bongartz and J. Torregrossa, 21–46. Peter Lang.
Lozano, C. and A. Mendikoetxea 2013. “Learner Corpora and Second Language Acquisition: The Design and Collection of CEDEL2”. In *Automatic Treatment and Analysis of Learner Corpus Data*, eds. A. Díaz-Negrillo, N. Ballier, and P. Thompson, 65–100. John Benjamins.
Mackey, A. and S. M. Gass 2016. *Second Language Research: Methodology and Design* (2nd ed.). Routledge.
Mendikoetxea, A. 2014. “Corpus-based Research in Second Language Spanish”. In *The Handbook of Spanish Second Language Acquisition*, ed. K. L. Geeslin, 11–29. Wiley-Blackwell.
Mendikoetxea, A. and C. 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): 871–898.
Mitchell, R., L. Domínguez, M. Arche, F. Myles, and E. Marsden 2008. “SPLLOC: A New Database for Spanish Second Language Acquisition Research”. In *EUROSLA Yearbook 8*, eds. L. Roberts, F. Myles, and A. David, 287–304. John Benjamins.
Montrul, S. 2013. *El bilingüismo en el mundo hispanohablante*. Wiley-Blackwell.
Montrul, S. 2016. *The Acquisition of Heritage Languages*. Cambridge University Press.
Myles, F 2015. “Second Language Acquisition Theory and Learner Corpus Research”. In *The Cambridge Handbook of Learner Corpus Research*, eds. S. Granger, G. Gilquin, and F. Meunier, 309–332. Cambridge University Press.
Paquot, M. and L. Plonsky 2017. “Quantitative Research Methods and Study Quality in Learner Corpus Research.” *International Journal of Learner Corpus Research* 3 (1): 61–94.
Porte, G. 2012. “Introduction”. In *Replication Research in Applied Linguistics*, ed. G. Porte, 1-17. Cambridge University Press.
Rojo, G. 2021. *Introducción a la lingüística de corpus en español*. Routledge.
Rojo, G. and I. Palacios Martínez 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. John Benjamins.
Sánchez-Gutiérrez, C. and P. Fernández-Mira 2022. “Datos longitudinales en los corpus de aprendientes de español”. In *The Routledge Handbook of Spanish Corpus Linguistics*, eds. G. Parodi, P. Cantos, and L. Howe. Routledge.
Sinclair, J. 2005. “How to Build a Corpus”. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 79–83. Oxbow books.

Thomas, M. and N. Pettitt 2017. “Informed Consent in Research on Second Language Acquisition.” *Second Language Research* 33 (2): 271–288.
Tono, Y. 2003. “Learner Corpora: Design, Development and Applications”. In *Proceedings of the 2003 Corpus Linguistics Conference*, eds. D. Archer, P. Rayson, A. Wilson, and T. McEnery, 800–809. UCREL Technical Paper number 16.
Tono, Y. 2016. “What is Missing in Learner Corpus Design?” In *Spanish Learner Corpus Research: Current Trends and Future Perspectives*, ed. M. Alonso Ramos, 33–52. John Benjamins.
Tracy-Ventura, N., R. Mitchell, and K. McManus 2016. “The LANGSNAP Longitudinal Learner Corpus: Design and Use”. In *Spanish Learner Corpus Research: State of the Art and Perspectives*, ed. M. Alonso-Ramos, 117–142. 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 2021. *The Routledge Handbook of SLA and Corpora*. Routledge.
Tracy-Ventura, N., M. Paquot, and F. Myles 2021. “The Future of Corpora in SLA”. In *The Routledge Handbook of Second Language Acquisition and Corpora*, eds. N. Tracy-Ventura and M. Paquot, 409–424. Routledge.
University of Wisconsin 1998. *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. University of Wisconsin Press.
Wilson, R. and J.-M. Dewaele 2010. “The Use of Web Questionnaires in Second Language Acquisition and Bilingualism.” *Second Language Research*, 26 (1): 103–123.
Yamada, A., S. Davidson, P. Fernández-Mira, A. Carando, K. Sagae, and C. Sánchez-Gutiérrez 2020. “COWS-L2H: A corpus of Spanish Learner Writing.” *Research in Corpus Linguistics* 8 (1): 17–32.