second
language
research

Review Article

CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research

Second Language Research
1–19
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02676583211050522
journals.sagepub.com/home/slr
SAGE

Cristóbal Lozano
Universidad de Granada, Spain

Abstract

This article presents and reviews a new methodological resource for research in second language acquisition (SLA), CEDEL2 (*Corpus Escrito del Español L2* 'L2 Spanish Written Corpus'), and its free online search-engine interface (cedel2.learnercorpora.com). CEDEL2 is a multi-first-language corpus (Spanish, English, German, Dutch, Portuguese, Italian, French, Greek, Russian, Japanese, Chinese, and Arabic) of L2 Spanish learners at all proficiency levels. It additionally contains several native control subcorpora (English, Portuguese, Greek, Japanese, and Arabic). Its latest release (version 2) holds material from around 4,400 speakers, which amounts to over 1,100,000 words. CEDEL2 follows strict corpus-design criteria (Sinclair, 2005) and L2 corpus-design recommendations (Tracy-Ventura and Paquot, 2021), and all subcorpora are equally designed to be fully contrastable, as recommended by Contrastive Interlanguage Analysis (Granger, 2015). Thanks to its design and web interface, CEDEL2 allows for complex searches which can be further narrowed down according to its SLA-motivated variables, e.g. first language (L1), proficiency level, self-reported proficiency level, age of onset to the L2, length of exposure to the L2, length of residence in a Spanish-speaking country, knowledge of other foreign languages, type of task, etc. These CEDEL2 features allow L2 researchers to address SLA questions and hypotheses.

Keywords

L2 acquisition research, L2 corpora, L2 Spanish corpus, learner corpora, second language acquisition (SLA)

I Introduction: Learner corpora and SLA

In learner corpus research (LCR), learner corpora are defined as systematic collections of authentic and contextualized written/spoken language produced by second language

Corresponding author:
Cristóbal Lozano, Universidad de Granada, Facultad de Filosofía y Letras, Campus de Cartuja, Granada, 18071, Spain.
Email: cristoballozano@ugr.es

(L2) learners (Callies and Paquot, 2015b: 1) and assembled according to explicit design criteria (Granger, 2009: 14). They ‘can contribute to SLA [second language acquisition] theory by providing a better description of interlanguage . . . and a better understanding of the factors that influence it’ (Granger, 2008: 259). CEDEL2 contributes to SLA by (1) incorporating these ‘factors’ into its corpus design via metadata, (2) adhering to specific design principles (Sinclair, 2005), and (2) following LCR recommendations (Tracy-Ventura and Paquot, 2021), which state that L2 corpora should:

1. focus on L2s other than English;
2. include learners at all proficiency levels, with varied L1s, from different ages, and from different learning backgrounds and settings;
3. promote cross-linguistic comparisons;
4. include more learner and task variables (metadata);
5. include varied tasks, some of which promote hypothesis-testing research;
6. consider different perspectives on what a ‘control’ corpus is;
7. be freely available to the research community (Open Science).

SLA researchers have traditionally favoured experimental and controlled data to test SLA hypotheses (Mackey and Gass, 2016). However, learner corpora are on the increase (Granger, 2012; Mendikoetxea, 2014; Myles, 2015) and recent corpus studies have started to test SLA hypotheses too (Lozano, 2021b). Large L2 corpora: (1) can offer a wide empirical base to test SLA theories (Mendikoetxea, 2014); (2) are ecologically valid since they sample unconstrained language, i.e. learners can choose their own wording (Granger, 2008); (3) contain highly contextualized language that allow researchers to go beyond the lexical/sentential level and explore discursive/pragmatic aspects (Myles, 2015); and (4) can be interrogated to find patterns leading to the formulation of hypotheses (hypothesis-finding) and also to test hypotheses against the corpus data (hypothesis-testing) (Mendikoetxea, 2014; Myles, 2005). In this context, we will discuss CEDEL2, which ‘represents a laudable attempt to bring learner corpus research and SLA closer together’ (Gilquin, 2015: 24).

This article aims to present the CEDEL2 design, compilation and free web interface in the context of SLA and L2 Spanish acquisition. CEDEL2 takes an SLA-motivated approach to learner corpus design with a view to offering answers to SLA questions. This article does neither intend to discuss the role of learner corpora in SLA theory (for discussions, see Lozano, 2021b; Lozano and Mendikoetxea, 2013; Mendikoetxea, 2014; Myles, 2015, 2021) nor to present an overview of L2 Spanish corpora (see overviews in Alonso-Ramos, 2016; Mendikoetxea, 2014; Rojo, 2021, and the L2 Spanish corpus index at http://repositorios.fdi.ucm.es/corpus_aprendices_español).

The article is structured as follows: Section II contextualizes CEDEL2. Section III presents its design principles, current holdings, data collection, web-based search interface, and some of its limitations. Section IV concludes with thoughts on the way forward in L2 (Spanish) corpus research.

II Some representative L2 Spanish acquisition corpora

While L2 English corpora predominate in the LCR field, the increasing interest in L2 Spanish has triggered the creation of corpora for L2 Spanish acquisition research. To

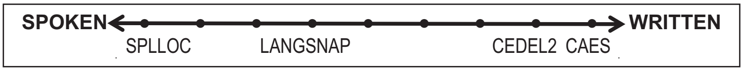


Figure 1. Four representative L2 Spanish corpora plotted along the spoken–written continuum.

contextualize CEDEL2, we will focus on a few representative corpora used in L2 Spanish acquisition research: *Spanish Learner Language Oral Corpora* (SPLLOC: Mitchell et al., 2008); *Language and Social Networks Abroad Project* (LANGSNAP: Tracy-Ventura et al., 2016); and *Corpus de Aprendices de Español* 'Spanish Learner Corpus' (CAES: Rojo and Palacios Martínez, 2016). We will see how CEDEL2 v.2 ultimately complements these corpora in terms of written-spoken data (next paragraphs) and adds functionalities in terms of SLA-informed design features (Section III).

Regarding medium (Figure 1), the four corpora range from the entirely spoken SPLLOC (100% of spoken files) to the entirely written CAES (100% written), with the mixed LANGSNAP (65% spoken, 35% written) and the mostly written CEDEL2 (97.5% written, 2.5% spoken) in between. Crucially, spoken data collection (and their corresponding transcriptions) are more labour intensive and costly than written data collection (Callies, 2015; Tracy-Ventura and Paquot, 2021). This results in spoken corpora like LANGSNAP (331,554 words) and SPLLOC (unreported number of words) being smaller in word size than written corpora like CAES (573,718) and CEDEL2 (1,106,013). For SLA, however, representativeness is more relevant than corpus size since the corpus is designed to represent the language type it intends to sample (see 3.1 for details on representativeness).

Concerning the spoken vs. written dichotomy, LCR researchers have often assumed that spoken data reflect learners' competence better than written data (e.g. Myles, 2015). However, many written learner corpora have been used to investigate competence and to test SLA hypotheses (Granger et al., 2015) and most of the learner corpus studies in the seminal book by Le Bruyn and Paquot (2021) use written corpora to 'provide powerful evidence that written data can make a significant contribution to some major SLA issues' (Granger, 2021: 248). A written corpus is thus a valid and valuable source of evidence to tap into learners' competence and to test SLA hypotheses (Lozano, 2021b; Mendikoetxea, 2014), despite 'SLA's insistence on the supremacy of spoken data' (Granger, 2021: 248). LCR researchers could move beyond this dichotomy by exploring new avenues like the triangulation of spoken and written data from different corpora. For example, Vázquez Veiga (2016) triangulated spoken SPLLOC data and written CEDEL2 data to investigate discourse markers at the lexicon-pragmatics interface. This approach ultimately yields a more fully-rounded picture of the linguistic phenomenon under investigation. Triangulation is even more promising when the spoken and written data come from the same learners, same task and same corpus (Granger, 2021), as is the case in CEDEL2 (see Section III.1).

Regarding their speakers, SPLLOC (n = 150 speakers) samples British secondary-school and university learners of L2 Spanish. LANGSNAP (n = 37) samples a cohort of British university learners of L2 Spanish longitudinally over two academic courses at three points in time (*before, during* and *after* their year abroad in Spain/Mexico). CAES (n = 1,423) samples instructed learners of Spanish at the *Instituto Cervantes* from six different L1 backgrounds. CEDEL2 (n = 4,334) samples more heterogeneous learners, with eleven L1s and diverse backgrounds in terms of countries, proficiency levels, chronological ages, ages of exposure to L2 Spanish, lengths of instruction in Spanish and learning environments. L2 Spanish researchers have thus at their disposal cross-sectional data with a variety of instructed and naturalistic backgrounds, plus longitudinal data (LANGSNAP).

As for the number of variables sampled, SPLLOC and LANGSNAP register 5 variables each (4 learner variables and 1 task variable), CAES 10 (9 learner, 1 task) and CEDEL2 25 (20 learner, 5 task). In terms of the number of subcorpora, SPLLOC and LANGSNAP contain 2 subcorpora each (1 learner, 1 native), CAES 6 (learner only) and CEDEL2 16 (11 learner, 6 native). Further details can be found in Appendix 1 in supplemental material.

III The CEDEL2 (version 2) corpus

We describe CEDEL2 next: its general corpus design principles and SLA-motivated features (3.1), its holdings (3.2), the data collection (3.3), its online web search interface (3.4), the process and product (3.5) and its limitations (3.6).

1 CEDEL2 corpus design

While many learner corpora are not built according to specific design principles (Gilquin, 2015; Tono, 2016), Gilquin (2015) proposes CEDEL2 as a good-practice case in learner corpus design since it rests on 10 corpus-design principles (Sinclair, 2005) that were adapted for SLA purposes (Lozano and Mendikoetxea, 2013). Sinclair's (2005) most relevant principles are:

1. content selection: select corpus contents based on external (communicative function of the texts) and not internal (the language of the texts) criteria;
2. representativeness: select contents to be as representative as possible of the language it samples;
3. topic: select the subject matter of the corpus based on external criteria;
4. contrast: compare only those subcorpora that have been equally designed;
5. documentation: fully document the contents of the corpus (i.e. its linguistically-motivated variables or metadata).

Based on Sinclair's design principles and the seven recommendations by Tracy-Ventura and Paquot's (2021) seen above, CEDEL2 showcases nine features that make it a suitable and valuable tool for L2 Spanish acquisition research.

Feature 1: Same design across subcorpora to ensure maximum comparability. The principle of contrast requires all subcorpora to be equally designed and balanced. In line with

Tracy-Ventura and Paquot's (2021) second recommendation and Granger's (2015) Contrastive Interlanguage Analysis (CIA), all CEDEL2 subcorpora follow the same design principles to ensure maximum comparability. Both between-subcorpora and within-subcorpus contrasts allow to test different effects (e.g. L2 development, L1 transfer, universal mechanisms, ultimate attainment, exposure, bimodality, bidirectionality, etc.), as will be discussed in the following subsections.

Feature 2: SLA-motivated variables (learner profile). Though learner variables 'have not been controlled consistently across corpora and are seldom incorporated in metadata' (Díaz-Negrillo and Thompson, 2013: 13), CEDEL2 registers, in line with Tracy-Ventura and Paquot's (2021) fourth recommendation, 20 learner variables (metadata) that can be exploited in the investigation of many SLA phenomena (Table 1).

Proficiency level is a key learner variable that is often lacking in learner corpora (Gilquin, 2015: 24). Some corpora use *ad hoc* criteria like course year as a proxy for proficiency level (Callies, 2021), but '[b]eing in the same year group at school is not always a sufficiently rigorous indication, and it is advisable to carry out independent measures of proficiency' (Myles, 2015: 316). CEDEL2 uses four measures: (1) objective independent measure: learners do a 43-point standardized Spanish placement test (University of Wisconsin, 1998); (2) subjective measure: learners self-rate each of their skills (speaking, writing, listening, and reading) according to a 6-point scale (lower/upper beginner, lower/upper intermediate, lower/upper advanced). Gilquin (2015) argues that the 'double proficiency measure in CEDEL2 is therefore a major asset (also because . . . it makes it possible to compare self-rated and real proficiency)' (p. 24). Additionally, CEDEL2 records (3) language certificate (learners state any certificates they hold, if any); and (4) length of instruction (LoI) in Spanish (learners report how long they have been learning Spanish for). All these metadata provide a good estimation of the learner's proficiency level.

Feature 3: SLA-relevant variables (task profile). Five task metadata were recorded:

1. Task title (14 tasks to choose from; see Appendix 1 in supplemental material);
2. Task text (written text/spoken text transcription with audio file);
3. Approximate time to produce the task (in minutes);
4. Where the task was done (in class/outside class/both);
5. Resources used to produce the task (help from Spanish native/bilingual dictionary/monolingual dictionary/spellchecker/grammar book/background readings/none).

Representativeness relates to both the sampled language (see this subsection) and the sampled speakers (see Feature 9 below). Lozano and Mendikoetxea (2013) argue that L2 corpus design should adhere to external criteria to ensure that their language is representative and authentic. This is achieved via 'tasks that allow learners to choose their own wording rather than being requested to produce a particular word or structure' (Granger, 2008: 261), which ultimately leads to 'a high degree of inclusiveness and a low degree of language bias' (Mendikoetxea, 2014: 14). Following these and Tracy-Ventura

Table 1. Learner variables (numbered list) and likely second language acquisition (SLA) phenomena to investigate (bulleted list) in CEDEL2.

① Learner's L1
② L1 of the learner's father
③ L1 of the learners' mother
④ Languages spoken at home

- L1 transfer effects by contrasting the 11 different L1s of the learner subcorpora
- L1 transfer effects vs. the effects of general/universal cognitive mechanisms by comparing L1s that are typologically similar/different
- Indirect L2 input effects by analysing the native Spanish control subcorpus
- Likely language-dominance patterns via the parents' L1 and language spoken at home

⑤ Standardized placement test score (1–43 points)
⑥ Proficiency level based on the placement score (lower/upper beginner, lower/upper intermediate, lower/upper advanced)
⑦ Proficiency-level self-evaluation on each skill in Spanish (speaking, listening, writing, reading)
⑧ Proficiency-level self-evaluation on each skill in an additional foreign language (speaking, listening, writing, reading)
⑨ Spanish language certificates held, if any

- L2 developmental effects by comparing beginner vs. intermediate vs. advanced learners
- Interlanguage knowledge at a given proficiency level within a learner subcorpus or across learner subcorpora
- Likely influence from the learners' additional foreign language
- Correlation between the learners' objective placement test score and their self-evaluation score

⑩ Sex

- Sex-related linguistic differences (males/females)

⑪ Age (chronological)

- Maturational effects due to age (e.g. linguistic abilities in young/adult/senior learners)
- Linguistic phenomena across the lifespan

⑫ Age of Exposure (AoE) to L2 Spanish

- Likely AoE effects (critical periods in L2)

⑬ Years studying Spanish (length of instruction, LoI)

- LoI effects (or lack thereof)

⑭ Stay(s) in Spanish-speaking countries for longer than 1 month? (yes/no)
⑮ Stay(s): Where?
⑯ Stay(s): When? (periods of residence)
⑰ Stay(s): How long? (length of residence)

- Effects of Length of Residence (LoR) and the recency of the stays in Spanish-speaking country/countries
- Effects of exposure to naturalistic input in a naturalistic setting
- Ultimate attainment effects in near natives (e.g. learners with high proficiency levels and long stays in a Spanish speaking country) vs. Spanish natives

⑱ School/University/Educational institution
⑲ Major degree (if any)
⑳ Year at university (if any)

- Likely effects of general educational background (school/university)
- Likely effects of educational background in Spanish (for those majoring in Spanish/Hispanic Studies vs. those who are not)

and Paquot's (2021) fourth and fifth recommendations, CEDEL2 tasks meet five criteria so as to potentially elicit authentic and varied linguistic phenomena within the text types it samples (see additional details in Appendix 1 in supplemental material):

1. Range of text types: descriptive (task 1 and 2 on the description of your region and a famous person), narrative (3–7 on the narration of a film, your last holidays, your future plans, a recent trip and a life experience; 13–14 picture- and video-based narratives), and argumentative (8–12 on terrorism, anti-tobacco law, gay marriages, marijuana legalization, and immigration). Some tasks may trigger a blending of descriptive-narrative styles (e.g. tasks 3–7).
2. Range of control: Most tasks are relatively open-ended (tasks 1–12), but tasks 13 and 14 impose a certain degree of control since participants narrate the same visual prompts.
3. Range of difficulty: Some descriptive tasks are linguistically undemanding and suitable for beginners (e.g. task 1), while narrative tasks require tense–aspect contrasts and argumentative tasks (e.g. 8–12) are linguistically demanding.
4. Pedagogical and replication criteria: Tasks 1–12 were selected from essay topics found in mainstream Spanish language textbooks. Tasks 13 (picture-based narrative *Frog, where are you?*) and 14 (short video from Charles Chaplin's *The kid*) have been extensively used in SLA research, which allows for replication.
5. Tense–aspect contrasts are triggered by different tasks, e.g. 3 *Describe a film you have recently seen* and 6 *Describe a trip you have recently made* (present perfect) vs. 4 *What did you do last year during your holidays?* (preterite) vs. 5 *Which are your plants for the future?* (future).

Purpose-designed, theoretically-motivated tasks are also necessary in learner corpora (Callies and Paquot, 2015a; Myles, 2015; Tracy-Ventura and Myles, 2015; Tracy-Ventura and Paquot, 2021). Domínguez et al. (2013) and Tracy-Ventura and Myles (2015) showed that such tasks in SPLLOC revealed tense/aspect contrasts that would have gone undetected if only one generic past-tense narrative task had been used. These contrasts are also allowed in CEDEL2; see point (5) above.

Theoretically-motivated CEDEL2 tasks (7 *Retell a recent film you have seen recently*; 13 *Retell the frog story* and 14 *Retell the Chaplin video*) trigger different information-status contexts (topic continuity vs. topic shift with varying number of potential antecedents) that constrain anaphora resolution in native and L2 Spanish. Such tasks have shed light on SLA theoretical issues like: (1) the Pronominal Feature Geometry hypothesis (Lozano, 2009b) since learners' well-known deficits with anaphora resolution selectively affect only 3rd person human anaphoric pronouns; (2) the Pragmatic Principles Violation Hypothesis (Lozano, 2016; Martín-Villena and Lozano, 2020) since learners are more redundant (in topic-continuity contexts) than ambiguous (in topic-shift contexts); and (3) the Position of Antecedent Hypothesis and the Accessibility Hierarchy (Georgopoulos, 2017).

Feature 4: Multiple L1 backgrounds. A classic debate in SLA is L1 influence vs. language universals. Following Tracy-Ventura and Paquot's (2021) first and second recommendations, CEDEL2 samples L2 Spanish learners from eleven L1s (English, German, Dutch, Portuguese, Italian, French, Russian, Greek, Japanese, Chinese, and Arabic). Additionally, based on the principle of contrast, CIA and Tracy-Ventura and Paquot's (2021) third recommendation, all CEDEL2 subcorpora were equally designed to allow contrasts between typologically-(un)related L1s (e.g. Germanic: English vs. German vs. Dutch; Romance: Portuguese vs. Italian vs. French; Germanic vs. Romance vs. Slavic vs. Semitic vs. Sino-Tibetan vs. Japonic).

Feature 5: Cross-sectional, developmental corpus. Following Tracy-Ventura and Paquot's (2021) second recommendation, CEDEL2 samples learners at all proficiency levels (based on a standardized placement test, see Feature 2 above), so L2 development can be traced. Unlike CAES, CEDEL2 does not restrict composition topics to different proficiency levels. CEDEL2 is a cross-sectional corpus, like SPLLOC and CAES, but unlike the longitudinal LANGSNAP.

Feature 6: Bidirectionality. CEDEL2 has an equally-designed, mirror-image L2 English corpus, COREFL (*Corpus of English as a Foreign Language*: http://corefl.learnercorpora.com) (Lozano et al., 2021), so it adheres to the principle of contrast and to Tracy-Ventura and Paquot's (2021) third recommendation. The same phenomenon can be explored bidirectionally (L1↔L2), e.g. L1 English–L2 Spanish (CEDEL2) vs. L1 Spanish–L2 English (COREFL), which allows to uncover L1-specific vs. universal effects that are independent of the L1–L2 combinations. Bidirectionality is an under-researched area in SLA/LCR.

Feature 7: Bimodal contrasts. As discussed above, spoken data are costly to collect and process, hence written corpora are the norm in LCR (Tracy-Ventura and Paquot, 2021). While CEDEL2 is predominantly a written corpus, it incorporates some data from participants who did the same task twice (written then spoken) with a two-week gap to avoid task-habituation effects (n = 104 participants: 26 L1 English–L2 Spanish learners, 59 L1 Spanish native controls, and 19 L1 English native controls; see Appendix 1 in supplemental material). The classic argument that spoken data better reflect learners' competence than written data (see discussion in Section II) can be reliably tested in CEDEL2 since medium varies (spoken/written) but task and speaker are constant, as recommended by Granger (2021: 248): 'bimodal corpora allow interesting comparisons, especially when the data are collected from the same learners'.

Feature 8: 'Dual' native control subcorpora. Some learner corpora include a native control subcorpus as a benchmark of the (variety of) language learners are exposed to. Control corpora are justified in LCR –see the 'comparative fallacy' vs. 'comparative hypocrisy' debate (Granger, 2009; Tracy-Ventura and Paquot, 2021). Following the recommendation of using several native norms in learner corpus research (Gilquin, 2021b) and Tracy-Ventura and Paquot's (2021) sixth recommendation, CEDEL2 samples data from 1,112 natives across Spanish-speaking countries (Peninsular and Latin American varieties), which turns it into a Spanish native corpus in its own right. Importantly, to determine whether learners' L2 knowledge is due to their L1, L2 input, or universal cognitive mechanisms, two native control subcorpora are required: (1) the learners' target (L2)

Table 2. Current native control subcorpora in CEDEL2 v.2.

Native control subcorpus 1 (learners' mother tongue)	Learner subcorpus	Native control subcorpus 2 (learners' target language)
L1 English	L1 English–L2 Spanish	L1 Spanish
L1 Portuguese	L1 Portuguese–L2 Spanish	
L1 Greek	L1 Greek–L2 Spanish	
L1 Arabic	L1 Arabic–L2 Spanish	
L1 Japanese	L1 Japanese–L2 Spanish	
L1 German*	L1 German–L2 Spanish	
L1 Dutch*	L1 Dutch–L2 Spanish	
L1 Italian*	L1 Italian–L2 Spanish	
L1 French*	L1 French–L2 Spanish	
L1 Russian*	L1 Russian–L2 Spanish	
L1 Chinese*	L1 Chinese–L2 Spanish	

Note. * under development.

language to check for potential effects of input on L2 acquisition; (2) the learners' L1 to check for possible L1 transfer effects. Using two native control subcorpora provides more information about the likely sources of knowledge than using only one control corpus, as demonstrated in L2 Spanish fluency (Huensch and Tracy-Ventura, 2017) and L2 English reference (Kang, 2004). CEDEL2 incorporates this 'dual' native-corpus perspective (Table 2).

Feature 9: Heterogeneous sample. LCR researchers complain that many current learner corpora oversample argumentative essays produced in university settings by young adults who are advanced learners (Paquot and Plonsky, 2017), so, following Tracy-Ventura and Paquot's (2021) second recommendation, CEDEL2 samples heterogeneous learners: eleven typologically (un)related L1s, as well as different proficiency levels, chronological ages, AoE to Spanish, LoI and LoR, learning environments (instructed/uninstructed) and educational backgrounds (universities/secondary schools). Such CEDEL2 variety is argued to be an asset (Gilquin, 2015). The CEDEL2 web-based search interface (Section III.4) allows researchers to filter results according to 12 metadata (e.g. L1, age, proficiency level, AoE, LoR, LoI), so researchers can select representative samples of the population they intend to investigate.

2 CEDEL2 holdings: Subcorpora and statistics

CEDEL2 was designed and compiled by Cristóbal Lozano, who directs the project since 2004 (Lozano, 2009a; Lozano and Mendikoetxea, 2013). Online written data collection started in 2006. CEDEL2 v.2 currently holds 4,399 written and spoken files coming from 4,334 speakers, amounting to over one million word-tokens (n = 1,105,936) (see statistics in Appendix 1 in supplemental material and on the CEDEL2 website) thanks to the voluntary data-collection collaboration of both local (Universidad de Granada, UGR)

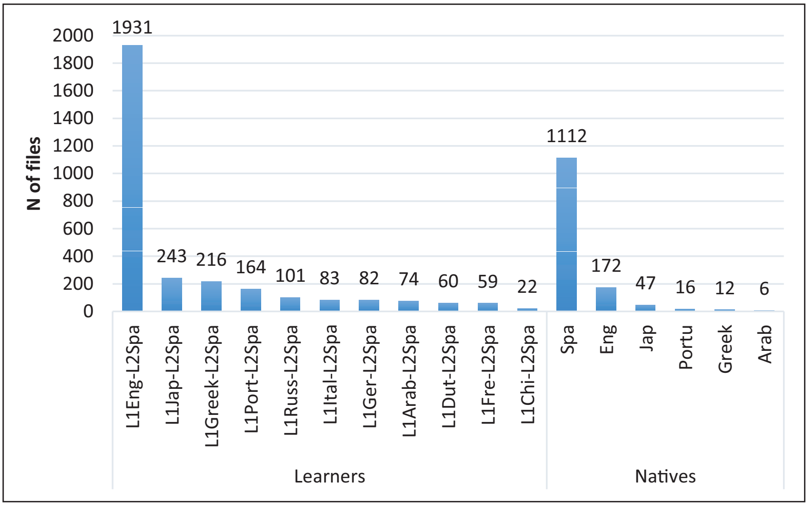


Figure 2. Number of files per subcorpus (CEDEL2 v.2).

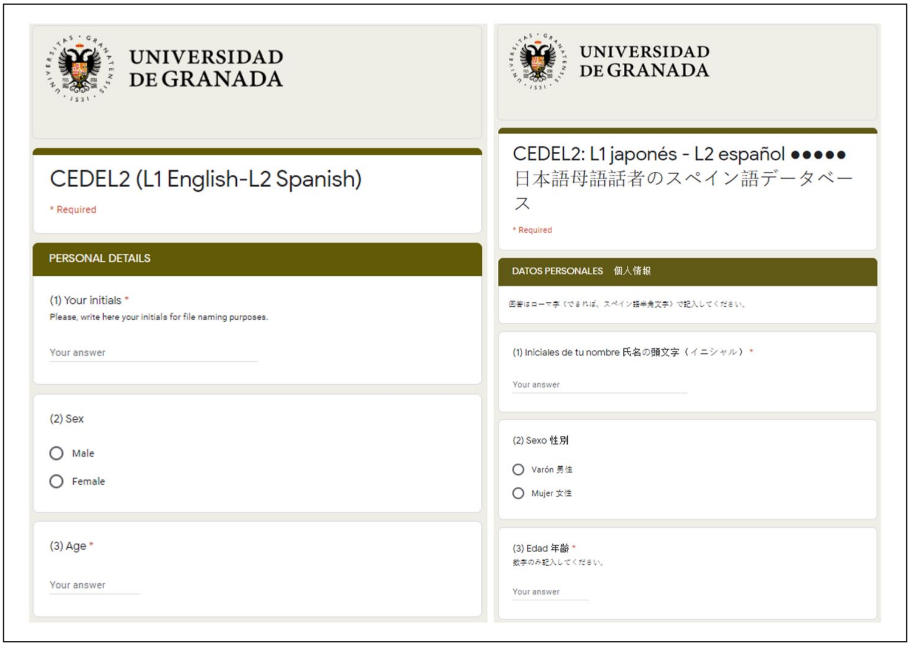
and international collaborators (for details, see Appendix 2 in supplemental material). Its 17 subcorpora (Figure 2) are still growing for the future CEDEL2 version 3.

3 CEDEL2 data collection

CEDEL2 written data were collected via dedicated online forms (v.1) from 2006 to 2016 and via Google Forms (v.2) from late 2016 (http://learnercorpora.com). L2 corpus online data collection has been argued to be ‘exciting for conceptualizing new avenues for data collection that could reach L2 users, rather than learners, that have largely been ignored’ (Bell and Payant, 2021: 64).

Forms are written in the participant’s native language to ensure full understanding (Figure 3). Participants completed three sections (instructions and informed consent, learner profile, and task profile), and learners additionally completed a Spanish placement test. CEDEL2 participation is voluntary. Ethics approval was granted by the Human Research Ethics Committee at the Universidad de Granada. Calls for participation were advertised in distribution lists (Linguist List, Infoling, Corpora List, social media, etc.). In return for their written participation, learners received their placement-test score and, upon request, a written statement of participation.

An international team (Appendix 2 in supplemental material) collaborated in the data collection and in the translation of the forms into the participants’ mother tongue. Approximately 27% of the data (1,185 files out of 4,440 files) were collected by collaborators and 73% (3,215 out of 4,399 files) were collected by C. Lozano. Data collection was coordinated from the University of Granada (89% or 3,936 out of 4,399 files) and the remaining 11% from other international universities (464 out of 4,400 files). These figures represent the coordinating institutions and not the actual data collection institution.



UNIVERSIDAD DE GRANADA

CEDEL2 (L1 English-L2 Spanish)

* Required

PERSONAL DETAILS

(1) Your initials *
Please, write here your initials for file naming purposes.

Your answer

(2) Sex

○ Male
○ Female

(3) Age *

Your answer

UNIVERSIDAD DE GRANADA

CEDEL2: L1 japonés - L2 español ●●●●●
日本語母語話者のスペイン語データベース

* Required

DATOS PERSONALES　個人情報

回答はローマ字（できれば、スペイン語半角文字）で記入してください。

(1) Iniciales de tu nombre 氏名の頭文字（イニシャル）*

Your answer

(2) Sexo 性別

○ Varón 男性
○ Mujer 女性

(3) Edad 年齢 *
数字のみ記入してください。

Your answer

Figure 3. CEDEL2 v.2 forms.

While written data were collected online from participants all over the world, spoken data were collected *in situ* at the Universidad de Granada in a quiet room with recording equipment (Audio Technica AT2020: Cardioid condenser microphone, 74 dB, 1 kHz at 1 Pa) to ensure optimal audio quality. Audio files were orthographically transcribed into text files. Transcripts followed basic transcription conventions (see the *User guide* in the CEDEL2 website for details) e.g. *xxx* for unintelligible words, / for silent pauses, *eh* for filled pauses, and = for false starts).

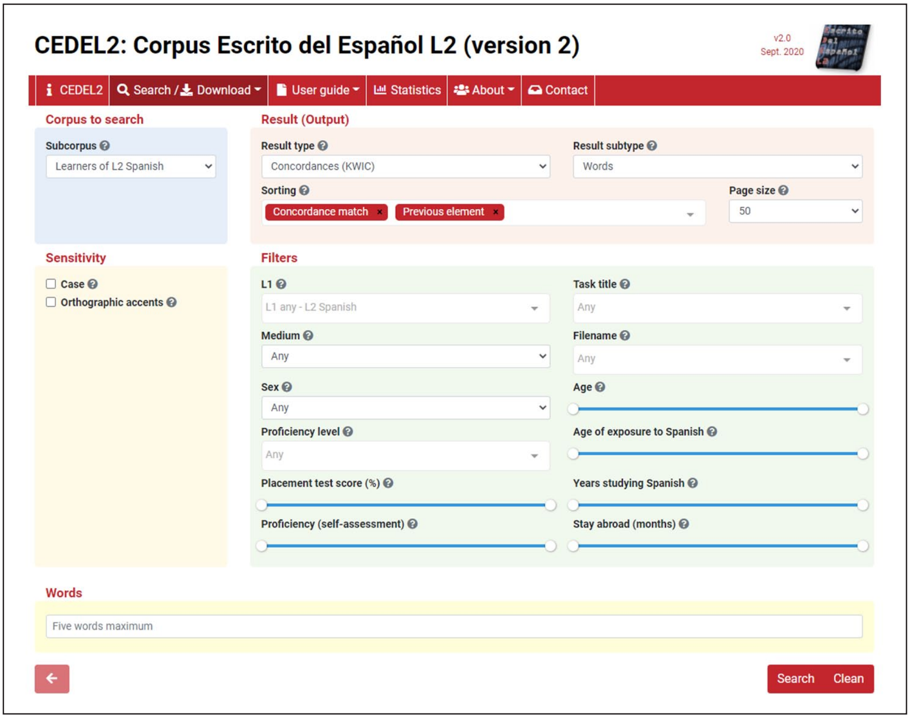
4 CEDEL2 search and download engine

Following the Open Science philosophy and Tracy-Ventura and Paquot's (2021) seventh recommendation, the newly developed CEDEL2 v.2 web-based interface (Figure 4) was freely released in September 2020 under a Creative Commons license (CC BY-NC-ND 3.0 ES) at http://cedel2.learnercorpora.com.

The CEDEL2 interface offers multiple and sophisticated search and download options (for details, see Appendix 1 in supplemental material, the *user Guide* and the circled question mark (?) on the CEDEL2 website). There are several (sub)types of results (i.e. outputs), which can be refined by the 12 filters Figure 4.

a Output type

1. Texts: The output shows a tabulated list of corpus files with 10 columns representing variables (Figure 5). Each text can be visualized by clicking on the



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 any - L2 Spanish

Task title: Any

Medium: Any

Filename: Any

Sex: Any

Age

Proficiency level: Any

Age of exposure to Spanish

Placement test score (%)

Years studying Spanish

Proficiency (self-assessment)

Stay abroad (months)

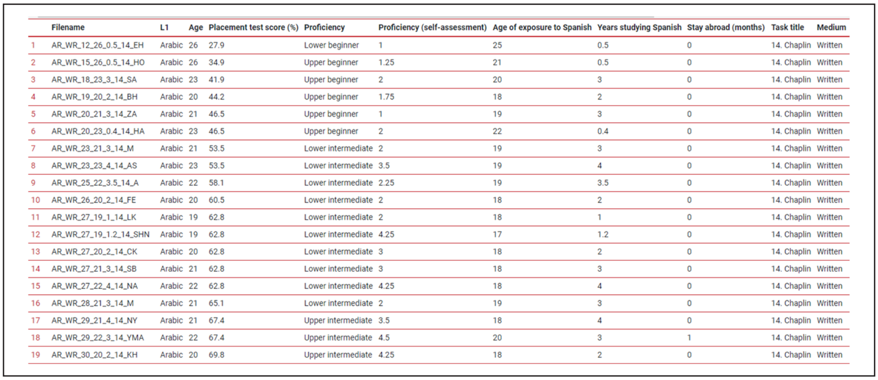
Words

Five words maximum

Figure 4. CEDEL2 v.2 web-based interface (http://cedel2.learnercorpora.com).

tabulated list. The list of texts can be additionally filtered (see Section III.4.c below), sorted according to certain criteria (Figure 6, left image), and downloaded in several formats: TXT (actual written text or spoken text transcription), TXT with metadata (text together with the 20 linguistic-profile variables and the 4 task variables), CSV (Comma Separated Values) for Excel, CSV for other software. The MP3 audio files can also be downloaded.

2. Concordances: The searched element is displayed in the centre accompanied by its surrounding context, i.e. keywords in context (KWIC, Figure 7). Concordances can be filtered (see Section III.4.c below), sorted (Figure 6, right image) and downloaded.
3. Simple frequency: The output shows the frequency of the searched element(s) (e.g. word, lemma, grammatical word) out of the total number of words and documents in the corpus, e.g. lemma *SER* ‘to be’ (21,016/851,675 results [24.68 million] in 2,713/3,034 documents) vs. lemma *ESTAR* ‘to be’ (5,038/851,675 results [5.92/million] in 1,752/3,034 documents).
4. Full frequency: The output shows the frequency of the searched element(s) according to eleven variables (L1, medium, L1 and medium, proficiency level, text title, years studying Spanish (LoI), stay abroad (LoR), age of exposure (AoE) to Spanish, age, sex, and placement test score).



	Filename	L1	Age	Placement test score (%)	Proficiency	Proficiency (self-assessment)	Age of exposure to Spanish	Years studying Spanish	Stay abroad (months)	Task title	Medium
1	AR_WR_12_26_0.5_14_EH	Arabic	26	27.9	Lower beginner	1	25	0.5	0	14. Chaplin	Written
2	AR_WR_15_26_0.5_14_HO	Arabic	26	34.9	Upper beginner	1.25	21	0.5	0	14. Chaplin	Written
3	AR_WR_18_23_3_14_SA	Arabic	23	41.9	Upper beginner	2	20	3	0	14. Chaplin	Written
4	AR_WR_19_20_2_14_BH	Arabic	20	44.2	Upper beginner	1.75	18	2	0	14. Chaplin	Written
5	AR_WR_20_21_3_14_ZA	Arabic	21	46.5	Upper beginner	1	19	3	0	14. Chaplin	Written
6	AR_WR_20_23_0.4_14_HA	Arabic	23	46.5	Upper beginner	2	22	0.4	0	14. Chaplin	Written
7	AR_WR_23_21_3_14_M	Arabic	21	53.5	Lower intermediate	2	19	3	0	14. Chaplin	Written
8	AR_WR_23_23_4_14_AS	Arabic	23	53.5	Lower intermediate	3.5	19	4	0	14. Chaplin	Written
9	AR_WR_25_22_3.5_14_A	Arabic	22	58.1	Lower intermediate	2.25	19	3.5	0	14. Chaplin	Written
10	AR_WR_26_20_2_14_FE	Arabic	20	60.5	Lower intermediate	2	18	2	0	14. Chaplin	Written
11	AR_WR_27_19_1_14_LK	Arabic	19	62.8	Lower intermediate	2	18	1	0	14. Chaplin	Written
12	AR_WR_27_19_1.2_14_SHN	Arabic	19	62.8	Lower intermediate	4.25	17	1.2	0	14. Chaplin	Written
13	AR_WR_27_20_2_14_CK	Arabic	20	62.8	Lower intermediate	3	18	2	0	14. Chaplin	Written
14	AR_WR_27_21_3_14_SB	Arabic	21	62.8	Lower intermediate	3	18	3	0	14. Chaplin	Written
15	AR_WR_27_22_4_14_NA	Arabic	22	62.8	Lower intermediate	4.25	18	4	0	14. Chaplin	Written
16	AR_WR_28_21_3_14_M	Arabic	21	65.1	Lower intermediate	2	19	3	0	14. Chaplin	Written
17	AR_WR_29_21_4_14_NY	Arabic	21	67.4	Upper intermediate	3.5	18	4	0	14. Chaplin	Written
18	AR_WR_29_22_3_14_YMA	Arabic	22	67.4	Upper intermediate	4.5	20	3	1	14. Chaplin	Written
19	AR_WR_30_20_2_14_KH	Arabic	20	69.8	Upper intermediate	4.25	18	2	0	14. Chaplin	Written

Figure 5. Text output.

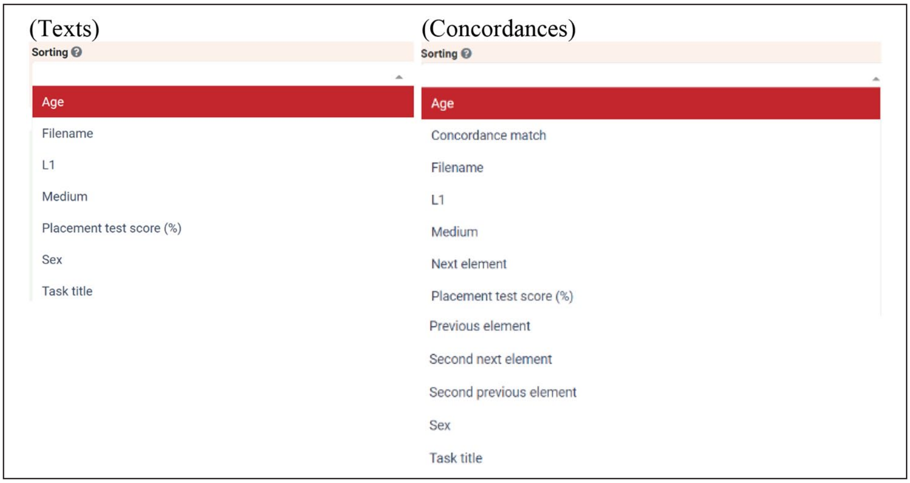
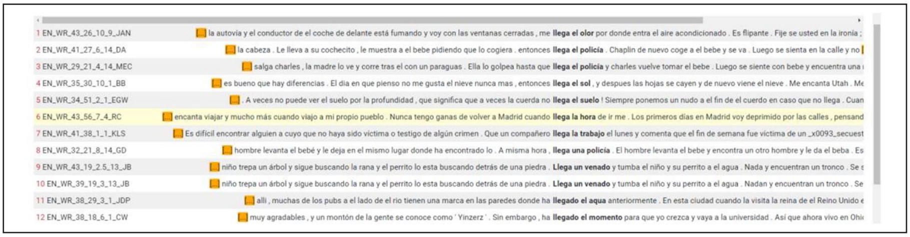


Figure 6. Sorting options.

b *Output subtype.* For concordances and frequencies, the searchable element(s) can be:

1. Words: Strings (characters, word, word combinations) that can incorporate wildcards: * (any number of characters), *?* (for any 1 character), | (for either one string or the other).
2. Grammatical elements: In CEDEL2 v.2, all words in the Spanish and English texts have been automatically tagged (i.e. part-of-speech annotated) with Freeling (http://nlp.lsi.upc.edu/freeling), which is an automatic annotator (see details on the CEDEL2 website's *User guide*). The searchable grammatical elements can be:

a. Part-of-speech (POS) tags, available from a drop-down menu containing word categories (e.g. Noun, Adjective, Verb, Adverb, etc.) and subcategories (e.g. noun.masculine.plural; verb.indicative.imperfect; pronoun.personal.3rd.singular.



```
1 EN_WR_43_26_10_9_JAN          [...] la autovia y el conductor de el coche de delante está fumando y voy con las ventanas cerradas , me llega el olor por donde entra el aire acondicionado . Es flipante . Fije se usted en la ironia ;
2 EN_WR_41_27_6_14_DA                  [...] la cabeza . Le lleva a su cochecito , le muestra a el bebe pidiendo que lo cogiera . entonces llega el policía . Chaplin de nuevo coge a el bebe y se va . Luego se sienta en la calle y no
3 EN_WR_29_21_4_14_MEC                     [...] salga charles , la madre lo ve y corre tras el con un paraguas . Ella lo golpea hasta que llega el policía y charles vuelve tomar el bebe . Luego se siente con bebe y encuentra una
4 EN_WR_35_30_10_1_BB             [...] es bueno que hay diferencias . El dia en que pienso no me gusta el nieve nunca mas , entonces llega el sol , y despues las hojas se cayen y de nuevo viene el nieve . Me encanta Utah . Me
5 EN_WR_34_51_2_1_EGW                  [...] . A veces no puede ver el suelo por la profundidad , que significa que a veces la cuerda no llega el suelo ! Siempre ponemos un nudo a el fin de el cuerdo en caso que no llega . Cuan
6 EN_WR_43_56_7_4_RC      [...] encanta viajar y mucho más cuando viajo a mi propio pueblo . Nunca tengo ganas de volver a Madrid cuando llega la hora de ir me . Los primeros días en Madrid voy deprimido por las calles , pensand
7 EN_WR_41_38_1_1_KLS       [...] Es difícil encontrar alguien a cuyo que no haya sido víctima o testigo de algún crimen . Que un compañero llega la trabajo el lunes y comenta que el fin de semana fue víctima de un _x0093_secuest
8 EN_WR_32_21_8_14_GD                 [...] hombre levanta el bebé y le deja en el mismo lugar donde ha encontrado lo . A misma hora , llega una policía . El hombre levanta el bebe y encontra un otro hombre y le da el beba . Es
9 EN_WR_43_19_2.5_13_JB           [...] niño trepa un árbol y sigue buscando la rana y el perrito lo esta buscando detrás de una piedra . Llega un venado y tumba el niño y su perrito a el agua . Nada y encuentran un tronco . Se s
10 EN_WR_39_19_3_13_JB            [...] niño trepa un árbol y sigue buscando la rana y el perrito lo esta buscando detrás de una piedra . Llega un venado y tumba el niño y su perrito a el agua . Nadan y encuentran un tronco . Se
11 EN_WR_38_29_3_1_JDP                      [...] alli , muchas de los pubs a el lado de el rio tienen una marca en las paredes donde ha llegado el aqua anteriormente . En esta ciudad cuando la visita la reina de el Reino Unido e
12 EN_WR_38_18_6_1_CW                     [...] muy agradables , y un montón de la gente se conoce como ' Yinzerz ' . Sin embargo , ha llegado el momento para que yo crezca y vaya a la universidad . Así que ahora vivo en Ohi
```

Figure 7. Concordance output.

masculine; etc.). Researchers can search for, e.g. imperfect/perfect past tense contrasts to test the well-known Aspect Hypothesis; 3rd person singular personal pronouns to test hypotheses on anaphora resolution (see above); etc.

b. Lemmas (e.g. *SER* and *ESTAR* would search for all verbal forms of *ser* and *estar* 'to be') to test hypotheses that can account for this well-known verbal contrast in L2 Spanish.

3. Words proxim.: It searches for a first word separated N words from a second word. The user can define the separation (N), e.g. *yo + 2 + estoy* would retrieve cases like *yo no|siempre|nunca|ahora estoy*.
4. Grammatical elements proxim.: It searches for a first grammatical word separated N words from a second grammatical word, e.g. *determiner-article-masculine-singular + 1 word + noun-feminine-singular*, which would find cases of masculine articles followed by feminine nouns (*el acción, el alcantarilla* . . .). This can help researchers test hypotheses on gender (mis)agreement. Other sophisticated lemma/tags combinations are possible, e.g. *lemma LLEGAR + ≤2 + noun*, would retrieve instances of postverbal subjects: *llegar* 'arrive' followed by either a noun or by another word (e.g. article) before the noun, as in *Llega un policía, Llega otro personaje*, etc.) to test the well-known unaccusative hypothesis.

c Filtering options. The output (results) can be filtered according to 12 filters (Figure 4): learners' L1, task medium (written/spoken/written and spoken by the same person), sex, proficiency level on a 6-point scale (lower/upper beginner, intermediate, and advanced), placement test score in Spanish (range: 0%–100%), self-evaluated proficiency level in Spanish on a 6-point scale, task title (14 tasks to choose from), filename, age, AoE to Spanish, LoI in Spanish (in years), and LoR in a Spanish-speaking country (in months). Filters allow to target those elements (learners, concordances, texts, grammatical words, etc.) that meet the researcher's criteria.

5 CEDEL2 process and product

Learners/natives participate in the written component via online forms (http://learnercorpora.com). Data and metadata are received in spreadsheet format. Based on a protocol, the corpus managers manually clean, standardize and join the spreadsheet data of the different subcorpora, which are then exported to CSV files for automatic validation. The

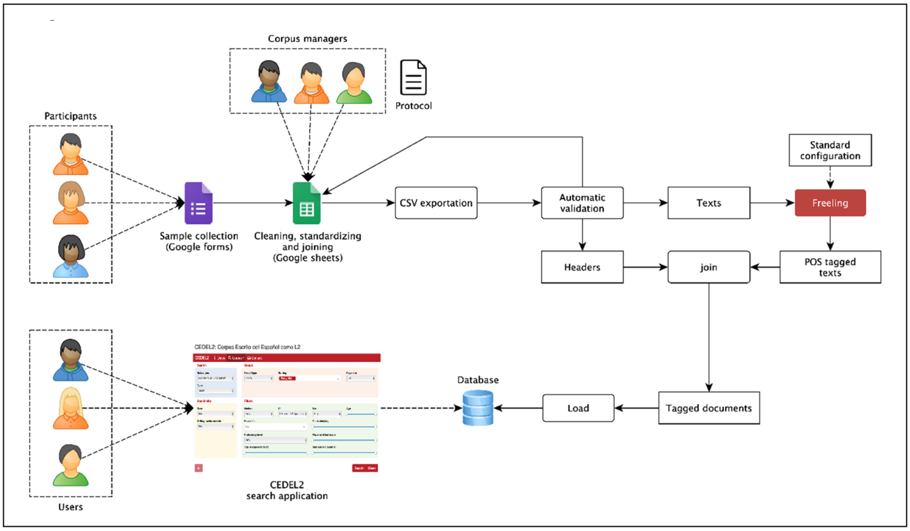


Figure 8. CEDEL2 flowchart.
Source. Copyright of this flowchart by NLPGo (www.nlpgo.com).

task texts are POS-annotated with Freeling and, together with their metadata (headers), are joined. The resulting tagged documents are loaded onto a database, which can finally be searched/filtered/downloaded by users via the online interface (see Figure 8).

At the time of printing of this paper, CEDEL2 data are the source of over 50 publications and dissertations covering SLA phenomena like anaphora, collocations, lexicon, morphology, orthography, reflexives, unaccusativity, (in)transitivity, determiners, lexicon-pragmatics interface, lexicon-discourse interface, error analysis, interference and transfer, automatic proficiency-level classification, natural language processing, and computer-assisted language learning (see http://cedel2.learnercorpora.com/about/studies).

6 CEDEL2: Current limitations and future improvements

Corpus balance requires having equally-proportioned samples of each language variety (Mendikoetxea, 2014: 14). Data size in each cell (i.e. cells resulting from crossing the corpus-design factors) must be balanced to ensure representativeness across subcorpora (Lozano, 2021a: 143). Future versions of CEDEL2 will therefore need to strike a balance between: (1) learner subcorpora, since the L1-English–L2-Spanish subcorpus is the largest in word size, speakers and number of tasks when compared to the other L2 subcorpora, which are smaller in size and contain either one task (*Chaplin*) or two tasks (*Chaplin* and *Frog*); (2) control subcorpora, since not all learner subcorpora have their equivalent native control subcorpus (Table 2); (3) spoken/written subcorpora, since only a small spoken sample has been included in CEDEL2 v.2. As for tagging, it has been done automatically with the Freeling tagger, which sometimes misclassifies words (particularly learners' novel words) into an incorrect word category. This feature is intended to be a useful search tool for SLA researchers, though it needs to be improved in future versions.

Finally, a promising feature for future versions is the use of automated text metrics (e.g. lexical sophistication/diversity, text cohesion, grammatical complexity, text readability, etc.) as an additional measure of proficiency level.

IV Learner corpora: The way forward

LCR has a lot to offer to SLA (Granger, 2021; Tracy-Ventura and Paquot, 2021), particularly to SLA theoretical models (Lozano, 2021b for an overview). This can be achieved via SLA-motivated corpus design and fine-grained, theoretically-informed tagsets to test particular hypotheses, as done with CEDEL2 (e.g. Georgopoulos, 2017; Lozano, 2009b, 2016; Martín-Villena and Lozano, 2020) and SPLLOC (Domínguez et al., 2013; Tracy-Ventura and Myles, 2015).

Most L2 Spanish corpora are cross-sectional, so longitudinal corpora like LANGSNAP are needed (Tracy-Ventura and Paquot, 2021). More spoken and written data coming from the same learner and task would be ideal to test the effect of medium (Granger, 2021). Multi-task (as opposed to single-task) corpora are also welcome since task variability is a key factor to understand learners' interlanguage (Domínguez et al., 2013; Tracy-Ventura and Myles, 2015).

SLA/LCR researchers argue for 'complementing the corpus data with experimental data' (Tracy-Ventura and Myles, 2015: 89). Such triangulation is gaining momentum (Gilquin, 2021a) since complementing large quantities of corpus data with controlled experimental data can offer SLA theories a more solid empirical base (Callies and Paquot, 2015b). Researchers have already triangulated experimental and spoken SPLLOC corpus data to advance our understanding of tense–aspect in L2 Spanish (Domínguez et al., 2013). Triangulation is most fruitful when done in a cyclic fashion to investigate the same phenomenon: the corpus results can provide insights that can be later implemented in an experiment, the results of which can additionally shed light on new aspects that can be later interrogated in the corpus, and so on, as done in the investigation of word order in L2 English under the unaccusative hypothesis with corpus data (Lozano and Mendikoetxea, 2010) and with experimental data (Mendikoetxea and Lozano, 2018). Cyclic triangulation is therefore a welcome step in the contribution of LCR to SLA.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: CEDEL2 has been publicly funded over the past 15 years by several research project grants from the Spanish Government, which we gratefully acknowledge: FFI2016-75106-P (Ministerio de Economía y Competitividad); FFI2012-30755 (Ministerio de Economía y Competitividad); FFI2008-01584/FILO (Ministerio de Ciencia e Innovación); HUM2005-01728/FILO (Ministerio de Educación y Ciencia).

ORCID iD

Cristóbal Lozano https://orcid.org/0000-0003-0068-154X

Supplemental material

Supplemental material for this article is available online.

References

Alonso-Ramos M (ed.) (2016) *Spanish learner corpus research: Current trends and future perspectives*. Amsterdam: John Benjamins.

Bell P and Payant C (2021) Designing learner corpora: Collection, transcription, and annotation. In Tracy-Ventura N and Paquot M (eds) *The Routledge handbook of second language acquisition and corpora*. Abingdon: Routledge, pp. 53–67.

Callies M (2015) Learner corpus methodology. In Granger S, Gilquin G, and Meunier F (eds) *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, pp. 35–55.

Callies M (2021) Proficiency. In Tracy-Ventura N and Paquot M (eds) *The Routledge handbook of second language acquisition and corpora*. Abingdon: Routledge.

Callies M and Paquot M (2015a) An interview with Yukio Tono. *International Journal of Learner Corpus Research* 1: 160–71.

Callies M and Paquot M (2015b) Learner corpus research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research* 1: 1–6.

Díaz-Negrillo A and Thompson P (2013) Learner corpora: Looking towards the future. In Díaz-Negrillo A, Ballier N, and Thompson P (eds) *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins, pp. 9–29.

Domínguez L, Tracy-Ventura N, Arche MJ, Mitchell R, and Myles F (2013) The role of dynamic contrasts in the L2 acquisition of Spanish past tense morphology. *Bilingualism: Language and Cognition* 16: 558–77.

Georgopoulos A (2017) Anaphora resolution in the interlanguage of Greek and English learners of Spanish: A corpus study. *Studies in Greek Linguistics* 37: 239–52.

Gilquin G (2015) From design to collection of learner corpora. In Granger S, Gilquin G, and Meunier F (eds) *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, pp. 9–34.

Gilquin G (2021a) Combining learner corpora and experimental methods. In Tracy-Ventura N and Paquot M (eds) *The Routledge handbook of second language acquisition and corpora*. Abingdon: Routledge.

Gilquin G (2021b) One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*. Epub ahead of print 31 March 2021. DOI: 10.1017/S0261444821000094.

Granger S (2008) Learner corpora. In Lüdeling A and Kytoe M (eds) *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter, pp. 259–75.

Granger S (2009) The contribution of learner corpora to second language acquisition and foreign language teaching. In Aijmer K (ed.) *Corpora and language teaching*. Amsterdam: John Benjamins, pp. 13–32.

Granger S (2012) How to use foreign and second language learner corpora. In Mackey A and Gass SM (eds) *Research methods in second language acquisition: A practical guide*. Oxford: Wiley-Blackwell, pp. 5–29.

Granger S (2015) Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1: 7–24.

Granger S (2021) Have learner corpus research and second language acquisition finally met? In Le Bruyn B and Paquot M (eds) *Learner corpus research meets second language acquisition*. Cambridge: Cambridge University Press, pp. 243–57.

Granger S, Gilquin G, and Meunier F (eds) (2015) *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

Huensch A and Tracy-Ventura N (2017) Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics* 38: 755–85.

Kang JY (2004) Telling a coherent story in a foreign language: Analysis of Korean EFL learners' referential strategies in oral narrative discourse. *Journal of Pragmatics* 36: 1975–90.

Le Bruyn B and Paquot M (eds) (2021) *Learner corpus research meets second language acquisition*. Cambridge: Cambridge University Press.

Lozano C (2009a) CEDEL2: Corpus Escrito del Español como L2. In Bretones CM et al. (eds) *Applied linguistics now: Understanding language and mind / La lingüística aplicada actual: Comprendiendo el lenguaje y la Mente*. Almería: Universidad de Almería, pp. 197–212.

Lozano C (2009b) Selective deficits at the syntax–discourse interface: Evidence from the CEDEL2 corpus. In Leung Y-I, Snape N, and Sharwood-Smith M (eds) *Representational deficits in second language acquisition*. Amsterdam: John Benjamins, pp. 127–66.

Lozano C (2016) Pragmatic principles in anaphora resolution at the syntax–discourse interface: Advanced English learners of Spanish in the CEDEL2 corpus. In Alonso Ramos M (ed.) *Spanish learner corpus research: Current trends and future perspectives*. Amsterdam: John Benjamins, pp. 235–65.

Lozano C (2021a) Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2 [Textual learner corpora for investigating the acquisition of Spanish as a second language]. In Cruz Piñol M (ed.) *E-Research y español LE/L2: Investigar en la era digital*. Abingdon: Routledge, pp. 138–63.

Lozano C (2021b) Generative approaches. In Tracy-Ventura N and Paquot M (eds) *The Routledge handbook of second language acquisition and corpora*. Abingdon: Routledge, pp. 213–27.

Lozano C and Mendikoetxea A (2010) Interface conditions on postverbal subjects: A corpus study of L2 English. *Bilingualism: Language and Cognition* 13: 475–97.

Lozano C and Mendikoetxea A (2013) Learner corpora and second language acquisition: The design and collection of CEDEL2. In Díaz-Negrillo A, Ballier N, and Thompson P (eds) *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins, pp. 65–100.

Lozano C, Díaz-Negrillo A, and Callies M (2021) Designing and compiling a learner corpus of written and spoken narratives: COREFL. In Bongartz C and Torregrossa J (eds) *What's in a narrative? Variation in story-telling at the interface between language and literacy*. Bern: Peter Lang, pp. 21–46.

Mackey A and Gass SM (2016) *Second language research: Methodology and design*. 2nd edition. Mahwah, NJ: Lawrence Erlbaum.

Martín-Villena F and Lozano C (2020) Anaphora resolution in topic continuity: Evidence from L1 English–L2 Spanish data in the CEDEL2 corpus. In Ryan J and Crosthwaite P (eds) *Referring in a second language: Studies on reference to person in a multilingual world*. Abingdon: Routledge, pp. 119–41.

Mendikoetxea A (2014) Corpus-based research in second language Spanish. In Geeslin KL (ed.) *The handbook of Spanish second language acquisition*. Oxford: Wiley-Blackwell, pp. 11–29.

Mendikoetxea A and Lozano C (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: 871–98.

Mitchell R, Domínguez L, Arche M, Myles F, and Marsden E (2008) SPLLOC: A new database for Spanish second language acquisition research. In Roberts L, Myles F, and David A (eds) *EUROSLA Yearbook 8*. Amsterdam: John Benjamins, pp. 287–304.

Myles F (2005) Interlanguage corpora and second language acquisition research. *Second Language Research* 21: 373–91.

Myles F (2015) Second language acquisition theory and learner corpus research. In Granger S, Gilquin G, and Meunier F (eds) *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, pp. 309–32.

Myles F (2021) An SLA perspective on learner corpus research. In Le Bruyn B and Paquot M (eds) *Learner corpus research meets second language acquisition*. Cambridge: Cambridge University Press, pp. 258–73.

Paquot M and Plonsky L (2017) Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3: 61–94.

Quesada T (2021) *Studies on Anaphora Resolution in L1 Spanish – L2 English and L1 English – L2 Spanish Adult Learners: Combining Corpus and Experimental Methods*. PhD dissertation, Universidad de Granada, Granada.

Rojo G (2021) *Introducción a la lingüística de corpus en español [Introduction to corpus linguistics in Spanish]*. Abingdon: Routledge.

Rojo G and Palacios Martínez I (2016) Learner Spanish on computer: The CAES ‘Corpus de Aprendices de Español’ project. In Alonso Ramos M (ed.) *Spanish learner corpus research: Current trends and future perspectives*. Amsterdam: John Benjamins, pp. 55–87.

Sinclair J (2005) How to build a corpus. In Wynne M (ed.) *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books, pp. 79–83.

Tono Y (2016) What is missing in learner corpus design? In Alonso Ramos M (ed.) *Spanish learner corpus research: Current trends and future perspectives*. Amsterdam: John Benjamins, pp. 33–52.

Tracy-Ventura N and Myles F (2015) The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1: 58–95.

Tracy-Ventura N and Paquot M (2021) The future of corpora in SLA. In Tracy-Ventura N and Paquot M (eds) *The Routledge handbook of second language acquisition and corpora*. Abingdon: Routledge.

Tracy-Ventura N, Mitchell R, and McManus K (2016) The LANGSNAP longitudinal learner corpus: Design and use. In Alonso Ramos M (ed.) *Spanish learner corpus research: State of the art and perspectives*. Amsterdam: John Benjamins, pp. 117–42.

University of Wisconsin (1998) *The University of Wisconsin College-Level Placement Test: Spanish (Grammar) Form 96M*. Madison, WI: University of Wisconsin Press. Available at: http://testing.wisc.edu/centerpages/spanishtest.html (accessed September 2021).

Vázquez Veiga N (2016) Discourse markers in CEDEL2 and SPLLOC corpora of learner Spanish: Analysis of some lexical–pragmatic failures. In Alonso Ramos M (ed.) *Spanish learner corpus research: Current trends and future perspectives*. Amsterdam: John Benjamins, pp. 267–97.