RiCL Research in Corpus Linguistics

COWS-L2H: A corpus of Spanish learner writing

Aaron Yamada[a] - Sam Davidson[b] - Paloma Fernández-Mira[b] -
Agustina Carando[b] - Kenji Sagae[b] - Claudia Sánchez-Gutiérrez[b]
Creighton University[a] / United States
University of California, Davis[b] / United States

Abstract – This paper presents the *Corpus of Written Spanish of L2 and Heritage Speakers* (COWS-L2H), a large corpus of compositions written by North American university students learning Spanish. The goals of this work are to (1) build a large corpus of Spanish learner writing that provides samples of written data from Spanish learners in the context of a North American university, (2) to contribute corpus data collected not only from second language (L2) learners of Spanish but also from learners of Spanish as a heritage language (SHL), and (3) to develop one of the few Spanish learner corpora to provide longitudinal data.

Keywords – L2 Spanish; Spanish as a heritage language; Learner corpus research

1. Introduction

Studies in the field of second language (L2) acquisition benefit from quantities of data that are large enough to aid in the analysis of L2 learning and learner language. Since the 1980s, such quantities of data have been provided by a growing number of learner corpora, or machine-readable databases of naturally produced language spoken or written by L2 learners. These learner corpora have facilitated analyses in various areas of L2 research (see Granger *et al*. 2015 for an overview). However, while corpora of L2 English are widely available, learner corpora in other languages, such as Spanish, are much less common. Granger *et al*. (2015), for example, catalog 137 total learner corpora and note that 60% are of L2 English. This distribution of learner data is incongruent with the fact that there exists a relatively high demand for learning Spanish in North America and across the globe. In 2013, for example, 51% of students enrolled in U.S. university language courses studied Spanish (American Academy of Arts and Sciences 2016) and there are over 21 million learners of L2 Spanish across the globe (Instituto Cervantes 2019). The present paper outlines the development of the *Corpus of*

Research in Corpus Linguistics 8/1: 17–32 (2020). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.08.01.02

Written Spanish of L2 and Heritage Speakers (COWS-L2H), a learner corpus that aims to remedy this shortcoming in available resources by providing a large sample of texts written by students enrolled in Spanish courses at the University of California at Davis, a large public North American university.

In addition to the general shortage of corpus data in L2 Spanish, there are certain gaps that exist in available learner Spanish corpora that make our endeavor necessary. For instance, there are relatively few L2 Spanish longitudinal corpora that collect data from learners at different points in their learning trajectory, in comparison to cross-sectional corpora that provide snapshots of data collected from different L2 learners at different language course levels. Without longitudinal data, researchers know relatively little about how individual learners advance in their L2 from one point in time to the next. Additionally, many current L2 Spanish corpora are relatively heterogeneous at the participant level, having sampled participants from a wide variety of learning contexts (study abroad vs. classroom, high school vs. university, etc.), which puts certain limitations on the explanatory power of the data. Finally, there are still not many available corpora that collect data from students who learn Spanish as a heritage language (that is, Spanish spoken as a minority language in a society with a different dominant language) in university courses specific for that purpose. Although research in various aspects of Spanish as a heritage language (SHL) has seen a surge in recent years (see the various chapters in Pascual y Cabo 2016), scarce works have attempted to measure the development of SHL using corpus data, which is arguably due to the unavailability of the necessary resources. Importantly, this kind of data could be used to measure the development of SHL within a classroom context, advancing what is known about the effects of institutionalized language programs for learners of a heritage language.

We aim to improve the present state of available corpus resources through the development of a new corpus of short compositions written by university students of L2 Spanish and SHL enrolled in Spanish language courses at a large public North American university. The outline of this paper is as follows: we will review presently available L2 Spanish corpora in Section 2, discuss the novel contributions of COWS-L2H in Section 3, describe its make-up and the procedure used to collect data in Section 4, present data describing our initial release in Section 5, and plot some of our future steps for this resource in Section 6.

2. A REVIEW OF L2 SPANISH CORPORA

Several Spanish language corpora have been designed for research purposes related to studies in sociolinguistics and historical linguistics, such as the *Corpus del Español en el Sur de Arizona* (Carvalho 2012), the *Corpus del Español* (Davies 2016), the *Corpus of Mexican Spanish in Salinas, California* (Brown 2017) and the various corpora compiled by the Real Academia Española. These corpora primarily focus on the oral and/or written production of native speakers of Spanish. There are certainly far fewer available corpora built with data produced by Spanish language learners. This is perhaps due to the wider availability of native speaker text and oral data that can be collected online or in other contexts, in comparison with the relative scarcity of learner data and limited access to L2 learners. Undoubtedly, there is a need for Spanish learner corpora in order to better understand the nature of L2 learner language, to elaborate more effective teaching practices, and ultimately contribute meaningful research to an increasingly multilingual North American society.[1] In this section, we provide a brief overview of some of the available L2 Spanish learner corpora and their key features.

Among available written learner Spanish corpora is the *Corpus de Aprendices de Español* (CAES, Rojo and Palacios-Martínez 2016), which contains 570,000 words produced in written texts by learners of all levels of Spanish within the Common European Frame of Reference (CEFR, Council of Europe 2011) except C2. Writing assignments were organized by level, such that students at different proficiency levels had different writing prompts. A variety of native languages are represented, including English, French, Arabic, Portuguese, Russian, and Mandarin. One of the largest corpora of L2 Spanish data is the *Corpus Escrito del Español como L2* (CEDEL2, Lozano 2009), a corpus directed by researchers at the Autonomous University of Madrid and the University of Granada in Spain, in collaboration with several other investigators from many other universities and secondary schools. This ongoing corpus contains written compositions in L2 Spanish from over 1,000 L1 English-speaking participants at universities and high schools around the world, compiling a corpus of currently over 800,000 words, and aiming to collect a total of one million words. Participants choose to write their compositions from a selection of twelve different topics and are also asked to complete a Spanish placement test. Designed to study the development of L2 Spanish

[1] We thank an anonymous reviewer for his/her helpful suggestions regarding this section.

morphology and syntax, CEDEL2 is effectively one of the broadest and most diverse databases of L2 Spanish available in the field.

Four oral learner corpora are currently available in L2 Spanish. The *Spanish Corpus Proficiency Level Training* (Koike and Witte 2016) was developed for language teacher training. Based on the guidelines of the American Council on the Teaching of Foreign Languages (ACTFL), it is designed to help teachers assess students' proficiency levels in Spanish. It consists of 327 videotaped oral interview sessions with 38 learners whose native language was English. It is also one of the only corpora to include learners of SHL, with data from 17 participants. The *Fono.ele Corpus* (Blanco Canales 2011) is a pronunciation-focused collection of 34,316 audio-recordings of 96 learners of a variety of native languages, at all CEFR levels in Spanish except A1 and C2. The *Spanish Learner Language Oral Corpus* (SPLLOC, Mitchell *et al*. 2008) is a corpus of oral L2 Spanish data collected from native speakers of English who completed a battery of elicitation tasks, such as picture description tasks, narratives, and oral interviews. In total, this corpus contains data collected from 60 L2 Spanish learners divided into three different levels based on proficiency and institutional enrollment. The *Corpus Oral de Español como Lengua Extranjera* (CORELE, Campillos Llanos 2014) is a corpus of oral production elicited using narrative and picture description tasks among 40 learners of L2 Spanish at CEFR levels A2 and B1. These learners were of native languages including English, French, Portuguese, and Italian, among several others.

A common limitation of all the corpora described above is that they do not feature longitudinal data. Some of the few Spanish learner corpora which do so are the *Languages and Social Networks Abroad Project* corpus (LANGSNAP, Tracy-Ventura *et al*. 2016) and the *Aprescrilov corpus* (Buyse *et al*. 2016). Designed to collect learner data in and throughout study abroad sojourns, the LANGSNAP corpus contains 300,000 words produced by 27 L1-English speaking university learners who studied abroad in Spain or Mexico. These participants produced oral and written data in a variety of elicitation tasks over a period of 20 months. *Aprescrilov,* in turn, is a large corpus of written data produced by learners of L3 Spanish whose L1 was either Dutch or French, and whose L2 was either Dutch, French, or English. These learners were enrolled in the first, second, or third year of university level Spanish and wrote more than one essay per academic quarter, which is equivalent to roughly three months.

3. MOTIVATION FOR THE PRESENT CORPUS

Despite the considerable utility of the above corpora, they are not without certain limitations. Principally, there is a notable lack of longitudinal L2 Spanish data, here defined as data collected from participants from at least three different points in time, following Ployhart and Vandenberg (2010). While *Aprescrilov* contains longitudinal data, it does so within a very limited timespan (one academic quarter), and collects data from L1 speakers of Dutch or French learning L3 Spanish. It is thus not of great use to those interested in L1 English-speaking learners of L2 Spanish. The LANGSNAP corpus, on the other hand, collects relatively long-term longitudinal data from L2 Spanish learners, but is limited to a small number of participants. There is clearly a need for a large corpus of longitudinally collected L2 Spanish data.

Additionally, we note that many of the above reviewed corpora collected data from relatively small quantities of participants and are thus modest in size. The corpora that are comparatively large, such as CAES and CEDEL2, are also rather heterogeneous in nature. For instance, while CEDEL2 approaches one million words, it does so in collecting data from a variety of different academic institutions (over one thousand different schools and universities), which increases the variability of these data. This is perhaps disadvantageous for researchers wishing to examine the nature of L2 Spanish within specific learning contexts, such as North American universities with large Spanish language programs. Again, we see a need for a large learner corpus that features data from a numerous but relatively homogenous group of Spanish learners, particularly for researchers interested in L2 Spanish development within a canonical university Spanish language course sequence with a uniform set of instructional syllabi and learning objectives.

Lastly, we must reiterate the fact that there are very few corpora that have collected data from SHL learners. Most research in SHL is devoted to analyzing the differences between SHL learners and native speakers of Spanish, or between SHL learners and L2 Spanish learners. Little empirical research, however, has used large quantities of data to measure SHL learners' linguistic development across the course of an academic SHL program. Large amounts of corpus data collected from SHL learners are needed to fill this gap, which is particularly relevant given that more and more institutions in the United States are designing SHL courses.

In short, while there are several learner corpora in Spanish presently available to researchers, there are also certain motivations for the construction of the present corpus. COWS-L2H thus complements the current set of Spanish learner corpora in the following three ways.

(1) COWS-L2H provides longitudinal data collected from individual learners. As described below in Section 4, participants in this corpus are asked to write a total of two compositions at two separate timepoints during the academic quarter and are allowed to participate in more than one academic quarter. Thus, this corpus includes longitudinal data collected from individuals across more than one quarter, and in several cases, more than one year.

(2) Additionally, COWS-L2H limits data collection to a single academic institution. This allows for a fine-grained analysis of the grammatical and lexical development of learners who share the same instructional context, which is that of a Spanish language program in a large public North American university. Although our corpus collects data at only one university, we know exactly which textbook our participants have used, what content is covered in their course syllabus, and what pedagogical methodology is in place in their classrooms. This allows researchers to study learners' L2 as well as the relationship between the L2 and the institutional factors that form the learning context. This is an essential point in the larger-picture notion of using corpus research to advance the effectiveness of language pedagogy.

(3) Finally, COWS-L2H is one of the few Spanish corpora to include data from learners of SHL, who are enrolled in a specific language program designed to address their unique needs.

4. COWS-L2H

In this section we detail the particular institutional assets at hand that help to make our resource unique, and we outline the methodology employed to collect the writing samples that make up COWS-L2H.

4.1. Institutional structure and participants

The present corpus enjoys several institutional advantages that contribute positively to its goals. First, the data are being collected at the University of California at Davis whose Spanish program offers courses in L2 Spanish at three levels: Introductory (corresponding to the first-year courses titled Spanish 1, Spanish 2, and Spanish 3), Intermediate (corresponding to the fourth and fifth-quarter courses Spanish 21 and Spanish 22), and Composition (corresponding to the sixth and seventh-quarter courses Spanish 23 and 24). The learning objectives of the Introductory and Intermediate courses are largely based on communicative competence and interaction with authentic language materials in Spanish, while the Composition courses are designed with a focus on academic writing skills in Spanish. Students can take a placement exam known as Web-based Computer Placement Exam[2] (WebCAPE 2.0) to be placed into these language courses. Table 1 below shows the raw WebCAPE scores necessary to be placed into the corresponding language courses.

WebCAPE Score	**Course Placement**
Below 260	Spanish 1
260-314	Spanish 2
315-373	Spanish 3
374-423	Spanish 21
424-464	Spanish 22
464 and above	Spanish 23

Table 1: WebCAPE 2.0 Spanish raw scores and corresponding course placement

During any given quarter, a total of roughly thirty individual sections across these course levels are offered, with a maximum of twenty-five students enrolled in each section. In general terms, in each quarter there are two to three times as many sections of Introductory Spanish offered than Intermediate or Composition sections. In all, this corpus benefits from a relatively large pool of student enrollment (roughly 750 students per academic quarter) from whom data can be collected.

Additionally, the University of California at Davis is one of few North American universities to offer a multi-level program in SHL, which consists of a three-quarter series of courses denominated Spanish 31, 32, and 33. This is significant because, as Beaudrie (2012) points out, of all U.S. universities with at least 5% Hispanic

[2] https://perpetualworks.com/

enrollment, only 38% offer SHL courses, and typically at only one level. The SHL courses at this university focus on increasing the academic proficiency in Spanish of students who learned Spanish at home and grew up with a mostly colloquial knowledge of the language. In general, these students are able to communicate effectively in an informal or familiar register, but have neither been frequently exposed to more formal registers nor to global varieties of the language. Like other SHL courses, the core emphasis is on Spanish language maintenance, the acquisition of the standard variety, and the move from receptive abilities to productive proficiency (Valdés and Parra 2018). Ultimately, the series as a whole aims at developing advanced literacy, akin to language arts courses in a monolingual context (Colombi and Harrington 2012) by building vocabulary and discursive devices associated with a diversity of dialects, registers, and genres. Generally, five sections of these SHL courses are offered each quarter, serving roughly 400 students per year.

Lastly, we aim to collect data samples continuously on a quarter-by-quarter basis, for a period of at least five years. Thus, as students continue to take courses in Spanish at this university, they can continue to contribute compositions to the corpus, providing longitudinal data that would allow researchers to measure the development of individual students' Spanish as they advance from one course to the next. Students are encouraged to take these courses and advance through the Spanish language program by the requirements of their majors, many of which require them to complete the Introductory series of a foreign language. Additionally, those students who wish to complete an undergraduate degree program in Spanish are required to fulfill the entirety of the Introductory, Intermediate, and Composition sequences, or in the case of SHL learners, the Spanish as a heritage language series. Thus, the unique advantages of compiling a corpus within this university language program are emphasized: (1) this corpus benefits from a very large participant pool; (2) this participant pool is perpetually replenished with new incoming students; (3) the language program includes a series of three consecutive SHL courses; and (4) students are encouraged to participate multiple times in the corpus project quarter after quarter, permitting the study of learner language from both a cross-sectional and a longitudinal approach. We now turn to describe the nature of the learner data collected in COWS-L2H and how these data were gathered.

4.2. Composition themes and data collection

All students enrolled in the aforementioned Spanish courses are offered extra credit as compensation for their participation in this research project. Through the course of the academic quarter, participants are asked to write a total of two compositions in Spanish that adhere to a minimum of 250 and a maximum of 500 words. Students enrolled in the Spanish 1 course are permitted to write compositions with a minimum word count of 150 words, as many of these students are true beginners in L2 Spanish.

To date, the composition data have been collected under four different themes. For the first set of compositions, collected from 2017 to 2018, participants were asked to write about *A famous person* and *A perfect vacation.* For the following set of compositions, collected from 2018 to the present, participants wrote about the themes *A special person in your life* and *A terrible story*. These composition themes are intended to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation on the part of the writer. For the famous person theme, for example, participants have written about famous figures of the present, of the past, and even about what it means to be a famous person. The use of such broad themes thus permits the production of a wide range of verb tenses and vocabulary. Additionally, it is important to note that we wished to choose composition themes that would be accessible to learners of all proficiency levels. In other words, we wanted to implement a broadly themed writing task that learners enrolled in any course level would be able to address. Furthermore, the rationale behind the choice of these themes, and the decision to change the themes, was to allow for certain linguistic contrasts in the data collected. For example, we changed the first theme from *A perfect vacation* to *A terrible story* in order to capture a range of linguistic structures associated with relatively positive experiences, in comparison with relatively negative experiences. Following the same rationale, the second theme was changed from *A famous person* to *A special person in your life* in order to collect a range of linguistic data related to people, one of whom was comparatively more familiar or intimate to the writer than the other.

We must recognize that a potential limitation of this open-ended composition task is that only a single written genre is represented in the corpus, which may indeed affect the findings of future analyses. However, we must also note that the advantage of utilizing a single type of written task allows for more controlled analyses of these data. A plan in place for future data collection would be to adopt a more authentic writing task, wherein instead of writing about a special person in their lives, authors could write a letter or message directly to a special person in their lives. Such an approach would not only allow for the collection of data more reflective of real-life writing tasks, but would also capture a different range of linguistic forms associated with personal address, for example.

This research protocol was approved by the Institutional Review Board at the university where the data are collected. The large-scale collection of our corpus data is made feasible through the use of the Canvas Learning Management System, an online classroom platform that is used at this university. Students who participate in the corpus project enroll in an online Canvas site, where they consent to participate in the research before providing their written samples. They read all necessary instructions regarding the tasks and then electronically submit their typed compositions. This platform organizes their submissions into a spreadsheet database accessible to the research team. Participants are given a window of one week to redact and submit each composition, at a time and place of their choosing. During this time, participants are able to see the given theme and can take as much time as they need, within the week, to write the composition. The instructions stress that participants are to write their compositions without the aid of any other person or materials. However, there is no guarantee that participants do not resort to such aids, which we recognize as a certain drawback to the online collection of such large amounts of data: it is certainly the case that more data is often noisier than less data. We do stress to participants that the quality of their writing samples will not affect the amount of extra credit they receive, nor will their language course instructors have access to these samples. Furthermore, if we find compositions that are exact copies of previously submitted compositions, they are removed from the corpus database and their authors do not receive extra credit for that quarter. We therefore do not believe that students have any clear incentive to cheat.

A period of one month separates the submission window of the first composition from the submission window of the second composition. All participants, regardless of course enrollment, write to the same themes in any given data collection window. For example, for the first data collection point of the academic quarter all participants write about *A special person in your life* and for the second data collection point all participants write about *A terrible story*. We chose a person-based topic for the first composition theme, because this is the theme that participants address during the first data collection point. This is important because the first data collection point takes place relatively early during the academic quarter, and as such, those at lower course levels (such as the Introductory course) generally have only learned vocabulary and grammar related to personal description and family members.

These participants must additionally complete a linguistic background questionnaire, which is hosted as an electronic form within the Canvas platform. This questionnaire is completed by participants once, at the first data collection point, for every academic quarter in which they participate. The linguistic background questionnaire collects information regarding participants' age, gender, institutional course level, instructors, native language, knowledge of other languages, and experience studying abroad in Spanish-speaking countries. It also includes a brief survey asking participants to self-rate on a scale of 1 to 5 their abilities in Spanish speaking, writing, reading comprehension, and listening comprehension. All of this information is coded into the corpus database accompanying the raw composition data.

This data collection procedure is executed each quarter. Students who have already participated in the project in previous quarters, but who wish to participate again, are able and encouraged to do so. These participants write compositions to the same themes but are asked to write entirely new compositions. In other words, a given student can write two compositions in the fall academic quarter, and then write another two compositions in the following winter academic quarter, and so on and so forth. In this manner, we are able to collect longitudinal data from the same student participants, responding to the same prompts, across multiple academic quarters.

5. INITIAL RELEASE: DESCRIPTIVE DATA

The data in COWS-L2H have been collected over the course of eight academic quarters from 2017 to the present date. We will continue to collect data for at least the next five years. In this section we offer basic descriptive information regarding the current status of COWS-L2H. Presently, there are 1,370 unique students who have contributed data to this corpus, including 850 native (L1) English speakers, and notably, 117 L1 Chinese speakers. Several other L1 speakers that cannot be easily clustered at the moment are also represented, such as those of Vietnamese and Tagalog. In terms of the longitudinal data we have collected, 420 participants have submitted compositions in a total of at least two quarters (for a maximum of four writing samples from each of those students), 150 have submitted compositions in at least three quarters (for a maximum of six writing samples from each student), and 38 have submitted compositions in at least four quarters (for a maximum of eight writing samples from each student). The current attrition rate from the first data collection point to the second data collection point in an academic quarter is 11.8%. Table 2 below details the number of compositions collected according to each aggregate institutional course level at this university, the total number of words collected for each aggregate course level, and the total number of participants who submitted compositions in each aggregate course level.

Course Level	**No. of compositions**	**No. of words**	**No. of participants**
Introductory (Spanish 1-3)	2,058	485,435	1,130
Intermediate (Spanish 21-22)	445	120,102	244
Composition (Spanish 23-24)	536	151,197	287
Heritage (Spanish 31-33)	459	130,684	244
Total	**3,498**	**887,418**	**1,905**[3]

Table 2: Descriptive summary of COWS-L2H by course level

In Table 3, we outline the number of total compositions and words written to each of the four themes: *A famous person*, *A perfect vacation*, *A special person in your life*, and *A terrible story*.

Theme	**No. of compositions**	**No. of words**
A famous person	892	224,328
A perfect vacation	806	205,720
A special person in your life	968	239,077
A terrible story	832	218,293
Total	**3,498**	**887,418**

Table 3: Descriptive summary of COWS-L2H by theme

COWS-L2H is freely available in TXT format to all researchers under a Creative Commons license, via a GitHub repository from which researchers can freely download our data.[4] An updated version of the data will be made available at the end of each academic year, once that year's data has been de-identified (that is, the names of participants and student identification numbers are not released, and the data are de-

[3] Note that this figure recounts students who have submitted compositions across different aggregate course levels, and thus differs from the number of unique participants who have submitted compositions to the corpus, which is 1,370.

[4] See https://github.com/ucdaviscl/cowsl2h

identified by hand in such a way that it would not be possible to link them to students' university records).

6. LIMITATIONS AND FUTURE STEPS

As we move forward with the construction of this corpus, one of our primary goals is to attain a greater balance among the different course levels from which we are collecting data. We recognize the challenges that exist with respect to the availability of participants in course levels that are not as numerously offered and/or populated as others and we are, therefore, considering increased recruitment efforts in these areas. It is worth noting, however, that having a larger number of students at the lower proficiency levels is important in that on average they produce fewer words per composition. Additionally, as often is the case in longitudinal data collection, one of the challenges we face is attrition, in that there is no guarantee that students who participate in the corpus once will participate again during the same quarter, or across multiple quarters. We do, however, require that student participants complete both compositions during an academic quarter to receive the extra credit compensation during that quarter.

We are currently undertaking efforts to develop and implement an error-annotation procedure for errors related to gender and number agreement and the use of the Spanish preposition *a* for direct object marking (e.g. *Respeta a los ancianos* '(s)he respects the elderly'). Our hope is that future research studies on these areas of L2 Spanish grammar could benefit from the use of error-tagged data drawn from this corpus. Additionally, we aim to design and launch search tools and an online interface to facilitate the use of the corpus.

In terms of the research which can be conducted with this corpus data, we hope to undertake preliminary analyses regarding vocabulary size and to compare these results with other large corpora. Another area of investigation that will be worth exploring is the relation between students' written production and their classroom materials. Indeed, no study to date has accumulated such a large amount of written production data in a context where these data can be matched with the syllabus and textbooks that were used at the time of writing. This will help us to better understand what the impact of classroom materials actually is on the written expression of the learners. This kind of information is relevant in the development of language teaching programs, and in testing the effect of institutional changes on the writing samples of the learners. Similarly, in terms of SHL data, COWS-L2H will allow us to (1) examine the synchronic characteristics associated with this speech community as a unique and localized variety of Spanish (Otheguy and Stern 2011); (2) track the impact of instruction on the development of academic linguistic devices and advanced literacy among heritage speakers (Colombi 2015) as these students progress through the three-course SHL series; and (3) extract the patterns associated with each level in order to create appropriate and much-needed SHL placement tests, and inform curriculum design targeting different SHL proficiencies (Beaudrie 2012).

7. Conclusion

This paper has presented COWS-L2H, a new learner corpus whose objective is to track the development of written Spanish language skills as observed over the course of a North American university Spanish language program. COWS-L2H aims to collect large amounts of longitudinal data that are currently scarce in the field of learner corpus research. COWS-L2H also collects data from students within a single homogeneous university language program, which is significant in that it provides data collected from students following a uniform set of learning objectives and pedagogical materials. Although we recognize that building a corpus at only one institution imposes certain limitations on our data, it is our hope that the research community would use our corpus in tandem with other available learner corpora. In this sense, one of our goals is to contribute to the larger resource network of learner corpora utilized by researchers seeking to draw generalizable conclusions about L2 learning in North American university settings. Finally, COWS-L2H is among the only corpora to collect large quantities of data from learners of SHL, which will provide valuable information to investigators working to advance research with respect to the development of learners enrolled in university language courses specifically designed for heritage language learners. In total, COWS-L2H is a significant step forward in the current landscape of corpus resources available to researchers working in the fields of Spanish as a second language and Spanish as a heritage language.

References

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

American Academy of Arts and Sciences. 2016. *The State of Languages in the U.S.: A Statistical Portrait.* Cambridge, Massachusetts: American Academy of Arts and Sciences.

Beaudrie, Sara M. 2012. Introduction: Development in Spanish heritage language placement. *Heritage Language Journal*. *Special Issue on Spanish Assessment* 9/1: i–xi.

Blanco Canales, Ana. 2011. *Fono.ele*, una herramienta Web para la investigación de la competencia fónica y la formación de profesores. In Carmen Hernández González, Antonio Carrasco Santana and Eva Álvarez Ramos eds. *La Red y sus Aplicaciones en la Enseñanza-Aprendizaje del Español como Lengua Extranjera*. Servicio de Publicaciones Universidad de Valladolid, 129–140.

Brown, Earl K. 2017. *Corpus of Mexican Spanish in Salinas, California.* http://itcdland.csumb.edu/~eabrown (24 November, 2019.)

Buyse, Kris, Lydia Fernández Pereda and Katrien Verveckken. 2016. The *Aprescrilov* corpus, or broadening the horizon of Spanish language learning in Flanders. In Margarita Alonso-Ramos ed., 143–168.

Campillos Llanos, Leonardo. 2014. A Spanish learner oral corpus for computer aided error analysis. *Corpora* 9/2: 207–238.

Carvalho, Ana M. 2012–. *Corpus del Español en el Sur de Arizona (CESA)*. University of Arizona. https://cesa.arizona.edu/ (18 February, 2020.)

Colombi, María Cecilia. 2015. Academic and cultural literacy for heritage speakers of Spanish. A case study of Latin@ students in California. *Linguistics and Education* 32/A: 5–15.

Colombi, María Cecilia and Joseph Harrington. 2012. Advanced biliteracy development in Spanish. In Sara M. Beaudrie and Marta Fairclough eds. *Spanish as a Heritage Language in the United States: The State of the Field.* Georgetown University Press, 241–258.

Council of Europe. 2011. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. https://www.coe.int/en/web/common-european-framework-reference-languages (24 November, 2019.)

Davies, Mark. 2016–. *Corpus del Español: Two billion words, 21 countries*. http://www.corpusdelespanol.org (24 November, 2019.)

Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier. 2015. Introduction: Learner corpus research– past, present and future. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 1–5.

Instituto Cervantes. 2019. *El Español: Una Lengua Viva.* Madrid: Instituto Cervantes.

Koike, Dale and Jennifer Witte. 2016. Spanish corpus proficiency level training website and corpus: An open-source, online resource for corpus linguistics studies. In Margarita Alonso-Ramos ed., 169–196.

Lozano, Cristóbal. 2009. CEDEL2: Corpus Escrito del Español como L2. In Carmen M. Bretones José Francisco Fernández Sánchez, José Ramón Ibáñez Ibáñez, María Elena García Sánchez, María Enriqueta Cortés de los Ríos, Sagrario Salaberri Ramiro, María Soledad Cruz Martínez, Nobel Perdú Honeyman and Blasina Cantizano Márquez eds. *Applied Linguistics Now: Understanding Language and*

Mind/La Lingüística Aplicada Actual: Comprendiendo el Lenguaje y la Mente. Almería: Universidad de Almería, 197–212.

Mitchell, Rosamond, Laura Domínguez, María J. Arche, Florence Myles and Emma Marsden. 2008. SPLLOC: A new database for Spanish second language acquisition research. *EuroSLA Yearbook* 8/1: 287–304.

Otheguy, Ricardo and Nancy Stern. 2011. On so-called Spanglish. *International Journal of Bilingualism* 15/1: 85–100.

Pascual y Cabo, Diego ed. 2016. *Advances in Spanish as a Heritage Language*. Amsterdam: John Benjamins.

Ployhart, Robert E. and Robert J. Vandenberg. 2010. Longitudinal research: The theory, design, and analysis of change. *Journal of Management* 36/1: 94–120.

Rojo, Guillermo and Ignacio M. Palacios-Martínez. 2016. Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. In Margarita Alonso-Ramos ed., 55–87.

Tracy-Ventura, Nicole, Rosamond Mitchell and Kevin McManus. 2016. The LANGSNAP longitudinal learner corpus. Design and use. In Margarita Alonso-Ramos ed., 117–142.

Valdés, Guadalupe and María Luisa Parra. 2018. Towards the development of an analytical framework for examining goals and pedagogical approaches in teaching language to heritage speakers. In Kim Potowski ed. *The Routledge Handbook of Spanish as a Heritage Language*. London: Routledge, 301–330.

Corresponding author
Aaron Yamada
Creighton University
Hitchcock 110C
2500 California Plaza
Omaha, NE 68178
United States
e-mail: aaronyamada@creighton.edu

received: December 2019
accepted: February 2019