Title: Digital Collections for Latin American and U.S. Latino Spanish Language Research: Phase 2 (Supplemental Funding)

Abstract

Our 12-month project will complete the most significant remaining portion of a current LARRP-supported transcription project that was impeded significantly by the events of 2020 and additional challenges we faced with visa- and immigration-related issues for a key project team member, Julia Orquera Bianco. Our current proposed project will enable Orquera Bianco to complete the orthographic transcriptions for two Spanish sociolinguistic corpora from Santiago, Chile, in the late 1970s and early 1990s. The tape recordings of the two Santiago, Chile, corpora were digitized through a prior LARRP-supported project and include 93 hours of recordings from 49 Spanish speakers in Santiago, Chile, in 1978 and 1992. These digitized recordings are now available via the USC Digital Library, Calisphere, and the Digital Public Library of America (DPLA). Our proposed project will complete transcriptions for 93 hours of Chilean recordings and publish the orthographic transcriptions with the digital audio recordings and research notes in the USC Digital Library, Calisphere, and DPLA. With our current LARRP funding, we expect to complete transcriptions of 63 hours of the recordings from the Santiago corpora by August of 2021. With supplemental LAARP support of $7,569, we can complete transcriptions for the remaining 30 hours of recordings by June 30, 2022.

The corpora included in our project offer rich documentation of Spanish language usage in Santiago and many facets of the daily experiences of Silva-Corvalán's interview subjects. The corpora informed research by Silva-Corvalán, including publications such as *Sociolingüística y pragmática del español* (Georgetown University Press, 2001) and numerous articles on Spanish-speaking communities in Santiago. The audio recordings provide valuable documentation of the social contexts of Spanish language use in Chile for linguists, anthropologists, sociologists, and many others with an interest in language use in Latin American cultures. Upon conclusion of our project, the transcripts will be published with the previously digitized audio recordings for free online public access in Calisphere (https://calisphere.org), DPLA (http://dp.la), and the USC Digital Library (http://digitallibrary.usc.edu/cdm/landingpage/collection/p15799coll22). All recordings, transcripts, and metadata will be preserved in perpetuity in the USC Digital Repository.

Statement on overall value to Latin Americanist research community

These corpora, recorded on audiocassettes during the late 1970s and previously digitized for free online public access via the Digital Public Library of America, have significant value to the Latin Americanist research communities. The transcriptions for the 93 hours of Chilean recordings are needed for comparative linguistic studies of how the Spanish language has changed over time in specific neighborhoods in Santiago and among discrete social groups in this important urban center for Spanish-language cultural expression. The recordings also reveal a wealth of information about the distinctive lifeworlds inhabited by Silva-Corvalán's interview subjects in Santiago during the late 1970s and early 1990s. Transcriptions, created according to professional standards and conventions of linguistic researchers, are necessary for research and teaching, and they will provide searchable text to facilitate discovery via the USC Digital Library, Calisphere, and DPLA. This is a substantial benefit that will enhance the research value of the Santiago corpora from the 1970s and early 1990s.

The data collected by Silva-Corvalán in Santiago constitute an important resource for sociolinguists and anthropologists interested in Chilean language and culture. The

earlier language corpus, which Silva-Corvalán collected in 1978, includes 49 speakers, distributed almost evenly by age group: 15 children (5-6 yrs. of age), 11 adolescents (15-18 yrs. of age), 11 adults (30-55 yrs. of age), and 12 older adults (56-70 yrs. of age). The interview subjects were also evenly divided by sex and three social classes: middle-middle, lower-middle, and working class. The second language corpus was collected by Silva-Corvalán in 1992. It includes 15 of the speakers recorded in 1978 in two age groups: 7 adults (20-50 yrs. of age), and 8 older adults (51+ yrs. of age). The speakers were divided evenly by sex and two social classes: middle-middle and working class. A sample 1992 recording of a Spanish speaker named Paty with transcriptions is available at http://digitallibrary.usc.edu/cdm/compoundobject/collection/p15799coll22/id/281.

The cross section of Spanish speakers in the two Santiago corpora provides a rich digital collection for the study of Chilean Spanish, how it was used by members of different social groups in Santiago, and its changes over time. These two corpora reveal the distinctive phonology and morphosyntax of Chilean Spanish. Its distinctive features include the loss of the final "s" in Chilean usage (e.g. the s marking the plural for words like "libros" or the s marking the second person for "lees" rather than the third person "lee"). In turn, the loss of the final "s" led to lexical and morphosyntactic changes in Chilean Spanish that can be studied through this unique digital collection.

Further, the two corpora allow linguists to study syntactic and morphosyntactic features of Chilean Spanish like:

- (*De*)*queísmo* (*Yo considero* ***de que*** *los padres pueden casarse*. "I think fathers [priests] can get married");
- Loss of case marking in relative clauses (*la casa* ***0*** *que viven es grande* "the house **0** they live in is big");
- Pluralization of *haber* "there to be" and *hacer* "ago" in existential and temporal constructions (*habían muchos afuera* "there were several outside," *hacen años que fui a Chile* "many years ago I went to Chile");
- Redundant verbal clitics in verbal periphrases (***lo*** *quería ver****lo*** "I **him** wanted to see **him"**); and
- The use of a verbal clitic coreferential with a postverbal nominal direct object (*yo* ***lo***$_i$ *encontraba un poco latoso* ***el Quijote***$_i$ *en ese tiempo* "I **it** found **the Quijote** a bit boring at that time").

Like the other Spanish language corpora we digitized with Silva-Corvalán, the Chilean corpora are valuable for documenting actual Spanish usage in specific places and times. As such, they reveal the widespread usage of constructions that may be considered "incorrect" but which reveal the social dimensions of language usage—and how language use marks and is marked by various social groups within Latin American polities. Defining normative Spanish language usage is often difficult in Latin American contexts, and these two Chilean corpora—which are not duplicated by any other online Spanish-language digital resources—provide valuable data points for sociolinguistic research as well as anthropological, sociological, and literary or popular culture studies in which language plays a prominent role.

Further, these two corpora will contribute to a more nuanced understanding of the interplay between societies and language in Chile as well as other Latin American countries. In complex Latin American societies like Chile, where social stratification is

permeable, the transfer of language features across sociolects is common. Phenomena that were once outside the norms for middle-class social groups, or were often avoided, are now widely accepted in both colloquial and formal contexts. Just to give a few examples, it is now far more common to observe phenomena like *voseo* in Argentina, (*de*)*queísmo* in Chile, clitics coreferential with a nominal direct object in a number of countries (including Chile and Argentina), and the personal conjugation of *haber* (at least in the present tense) in almost all Latin American countries. The two Chilean corpora—gathered in 1978 and 1992—provide great insight into these transformations of Chilean Spanish over time. In turn, they will support comparative studies of related changes in other Latin American countries during this time period.

Statement of due diligence

The audio recordings were created by Silva-Corvalán as part of her research and are not duplicated by the holdings of other research archives. The digitized recordings are freely available via the USC Digital Library, Calisphere, and DPLA. However, most of the Chilean recordings lack authoritative orthographic transcriptions. Our proposed project will complete and publish transcriptions for the entire Chilean corpora. These recordings and transcriptions do not exist elsewhere in the online or physical holdings of other archival institutions.

Open access commitment

We are committed to providing open access to the transcriptions that will complement the digitized recordings and metadata created through this project. All materials will be freely available online in the Digital Public Library of America (DPLA), Calisphere, and the USC Digital Library. No access fees of any kind will be charged, and there will be no restrictions on the use of the materials for research and educational purposes.

Description of the project

Scope and content:

Through our proposed project, we will complete orthographic transcriptions of 93 hours of Chilean corpora published in the USC Digital Library via our prior LARRP-supported project. With our current LARRP support, we expect to finish transcriptions for 63 hours of recordings by August of 2021. With our requested supplemental LARRP support, we can finish transcriptions for the remaining 30 hours of recordings by June of 2022. Upon completion of the transcriptions for each recording, we will update the digital objects and Qualified Dublin Core metadata records in the USC Digital Library. In turn, the updated digital objects will be harvested and made available via Calisphere and the DPLA.

With supplemental support from LARRP, we can retain a highly qualified former USC graduate student and Spanish native speaker, Julia Orquera Bianco, to continue her work on the project to create the transcriptions. Experienced USC Digital Library personnel, assisted by Silva-Corvalán, will update the Qualified Dublin Core metadata for the recordings and oversee their publication via the USC Digital Library, Calisphere, and DPLA. In future phases of this project, we will complete similar transcriptions for the remaining Southern California corpora. We will also explore publishing recordings and transcriptions from USC's Spanish Sociolinguistic Research Collection in the Open Language Archives Community (OLAC) and similar shared resources.

Our proposed 12-month project will complete the most critical piece of our current LARRP-supported project, which was delayed by the closure of USC facilities for much

of the year due to the COVID-19 pandemic, the difficulties of remote work, and added delays in resolving visa- and immigration-related issues for Orquera Bianco to enable her continued work on the project. Our proposed project also builds on our previous efforts to digitize Spanish corpora collected by Silva-Corvalán for the USC Digital Library's Spanish Sociolinguistic Research Collection. Silva-Corvalán's field recording work was supported by the National Science Foundation and the Ford Foundation. Our digitization efforts were supported by LARRP, a USC faculty research grant, the Del Amo Fund at USC Dornsife College, and the L.A. Murillo Hispanic Heritage Endowment at the USC Libraries. In addition to making freely available 348 hours of digitized corpora from Santiago and Southern California, these efforts have helped to establish workflows and working relationships among Co-PI Barbara Robinson, Professor Carmen Silva-Corvalán, and Co-PI Wayne Shoaf and other personnel at the USC Digital Library.

In our current LARRP-supported project, Orquera Bianco has played an integral role in creating orthographic transcriptions for the Chilean recordings. We have invested time and effort in training her in USC Digital Library workflows, and she has specialized expertise that will be invaluable in completing our project. For a representative sample transcription that demonstrates the quality and value of Orquera Bianco's work, see the digital recording and transcriptions of a 1992 recording of Spanish speaker named Paty at http://digitallibrary.usc.edu/cdm/compoundobject/collection/p15799coll22/id/281. For the proposed project for supplemental LARRP funding, all transcriptions will be created according to the same standard. After Orquera Bianco creates transcriptions, they are reviewed for accuracy. After any revisions needed to finalize them, Shoaf normalizes them prior to their publication in the USC Digital Library, Calisphere, and DPLA.

In future phases of this project—which will be outside the scope of the proposed project—we hope to create additional orthographic transcriptions of recordings from other corpora and digitize supporting materials (e.g. notes and questionnaire forms) created by Silva-Corvalán. We also hope to publish the recordings in OLAC and other aggregator resources for the study of language. This will require enriched metadata, which is not feasible given the budget for our proposed project.

Access

Upon conclusion of this project, all transcriptions and previously digitized audio recordings will be freely accessible via Calisphere, (www.calisphere.org), the DPLA (http://dp.la) and the Spanish Sociolinguistic Research Collection in the USC Digital Library, http://digitallibrary.usc.edu/cdm/landingpage/collection/p15799coll22.

Copyright and permissions

Professor Carmen Silva-Corvalán, who is working closely with us during this multi-year project, has conveyed the copyright to the audio recordings and all associated research materials. USC will hold the rights to transcriptions created through this project. In her signed gift agreement for all previously digitized materials, Silva-Corvalán has assigned the copyrights to USC and given USC the rights to digitize these materials for free online public access.

To protect the privacy and related rights of Silva-Corvalán's interview subjects and their families, many of whom are still living, we have redacted identifying personal information from the publicly accessible digital audio recordings in the DPLA and USC Digital Library. Similarly, personal information will be redacted from the transcriptions published in the USC Digital Library through this project. We are taking this approach in keeping

with best practices for digital library projects that involve personal information. The USC Digital Library has taken similar approaches to other projects involving personal information, and the research value of the recordings does not depend on providing the names and other identifying information for Silva-Corvalán's interview subjects.

Metadata

USC Metadata and Digital Librarian Wayne Shoaf will update Qualified Dublin Core metadata records for the Chilean corpora after transcriptions are created and finalized. During previous phases of this project, Shoaf created metadata in consultation with Silva-Corvalán and included LARRP as a MetaTag in the project metadata for the Chilean and Southern California corpora published in the DPLA and USC Digital Library at the conclusion of our previous LARRP-supported project.

Long-term sustainability and stewardship

All transcriptions created through this project will be preserved in perpetuity along with the digital audio files—including archival Broadcast Wave Files (BWF) and access mp3 audio files—and metadata created during previous phases of this project. We will use the USC Digital Repository (USCDR). Its systems were created to preserve in perpetuity the 52,000 video testimonies of Holocaust survivors gathered by the USC Shoah Foundation and meet the rigorous standards for data preservation now required by NSF, NIH, and other federal agencies. It complies with all best practices for digital preservation and utilizes checksums to monitor the integrity of digital information. All digital content is regularly monitored, and files are restored from backups if errors or variances of any kind are detected. Long-term plans and resources are in place at the USCDR to migrate to new data storage formats as those become the industry standard. Further information about the USCDR is available at http://repository.usc.edu/.

The USC Libraries devote substantial resources to building and maintaining the USC Digital Library, which includes more than 1.5 million historic photographs, artworks, documents, newspapers, and audiovisual recordings and draws 50,000 unique visitors per month. These resources include a staff of experienced personnel who are familiar with all aspects of digital library projects, USC's community of expertise in digital preservation and the digital humanities, a robust technical infrastructure, and ongoing efforts to improve the user experience and features of the USC Digital Library's online presence and integrations with social media platforms like Twitter and Facebook. The USC Digital Library was selected as the first California content hub for the Digital Public Library of America (DPLA) and is an active participant in the Golden State Digital Network with other California digital libraries. As a result of these partnerships, all items on the USC Digital Library are also published in Calisphere and the DPLA, which bring together the riches of California and America's libraries, archives, and museums, and make them freely available to the world.

Because the transcriptions for the Spanish sociolinguistic corpora from Santiago will be published in public online platforms like the USC Digital Library, Calisphere, and the DPLA with complementary materials and large numbers of monthly unique visitors, researchers will be able to discover and access them via multiple online pathways. In this way, these recordings and orthographic transcriptions—which document Spanish-speaking communities at specific places and times—will enjoy the broadest possible audience. Although they are specialized recordings of greatest interest to linguists, sociologists, anthropologists, and other researchers with an interest in language use among Latin American populations, they have great documentary value and public

interest as a record of how real people spoke and what they spoke about at precise historical moments with a variety of social communities and groups in Santiago.

Accordingly, the USC Libraries will feature the project in our newsblog and active social media outreach via Facebook and Twitter. Co-PI Barbara Robinson will also work with the USC Libraries' communications staff to share information about the project and ensure that Latin Americanists and linguists working in this research area are aware of the digital resource. Our social media outreach and blog posts will share these materials with members of the public who would not otherwise be aware of them. As of April 7, 2021, the USC Libraries have 6,010 people who follow our updates on Facebook and 10,600 followers on Twitter. The USC Digital Library has 2,637 followers on Twitter and 3,775 followers on Pinterest.

Work plan

Project team

Co-Principal Investigator **Barbara Robinson** is the Boeckmann Center for Iberian and Latin American Studies librarian at the USC Libraries' special collections, where she has worked since 1985. Prior to that, she worked for more than a decade at UC Riverside's Tomas Rivera Library. Robinson has extensive knowledge of USC's Latin American collections and has worked closely with team members Carmen Silva-Corvalán and Wayne Shoaf during prior phases of this project. Robinson will oversee the project and the administration of grant funds as well as all communication efforts relating to it.

Carmen Silva-Corvalán is professor emerita in Spanish, Portuguese, and linguistics at the USC Dornsife College. Her research has focused on the social and linguistic forces influencing language variations and changes in monolingual contexts and language permeability in bilingual contexts. Her original field research recordings are the subject of our proposed project, and she will consult closely with Orquera Bianco, who will create orthographic transcriptions via this project.

Co-Principal Investigator **Wayne Shoaf** is Metadata and Digital Librarian at the USC Digital Library and draws on 20 years of experience with digital library projects. Working with Robinson, Silva-Corvalán, and Orquera Bianco, he will normalize the transcriptions for uniformity prior to publication in the USC Digital Library. In addition to updating the digital objects for the Chilean corpora, he will also update the Qualified Dublin Core metadata records for the recordings to reflect the orthographic transcriptions and any digitized paper materials we publish as part of this project (e.g. research notes and questionnaires created by Silva-Corvalán during her projects in Santiago). Shoaf will direct the work of Orquera Bianco and oversee the publication of digital objects in the USC Digital Library. Upon publication in the USC Digital Library, the recordings will be harvested for publication in Calisphere and the Digital Public Library of America.

Julia Orquera Bianco is a former USC graduate student and native Spanish speaker. She holds a MFA from USC and a BFA from la Facultad de Artes en la Universidad del Museo Social Argentino in Buenos Aires. Orquera Bianco has substantial prior experience creating orthographic transcriptions for the materials included in this project, so he has played an integral role in our project to date. Directed by Shoaf with assistance from Robinson and Silva-Corvalán, Orquera Bianco will complete orthographic transcriptions for 93 hours of recordings from the Chilean corpora. In our proposed project for supplemental LARRP funding, Orquera Bianco will complete

transcriptions for 30 hours of recordings from the Chilean corpora. These will be finalized, and we will ensure all personal information is redacted. We estimate Orquera Bianco will spend 315 hours creating the transcriptions at approximately 10 hours of work per hour of recorded audio. We are allowing 15 additional hours to redact personal information from the transcriptions, finalize those transcriptions, and address any unexpected project challenges. If the work goes more quickly than anticipated, we will request permission to create transcriptions from additional corpora in the Spanish Sociolinguistics Collection. However, in the past year, each hour of recorded material has required 10 hours of transcription work. So we are confident in our estimates.

Budget

We are requesting $7,569 in supplemental funding under the LARRP grant program. The USC Libraries will provide $5,694 of in-kind support for the 12-month project, and Silva-Corvalán will contribute her time and expertise to the project. All grant funds will be housed in dedicated, auditable accounts administered by Co-PIs Barbara Robinson and Wayne Shoaf assisted by USC finance and grants administration personnel. USC's office of Sponsored Projects Accounting will provide financial reporting.

We are seeking support under the LARRP program for the wages and benefits of Julia Orqueda Bianco:

- Wages (315 hours @ $18/hr.): $5,670
- Fringe benefits (33.5%): $1,899

Carmen Silva-Corvalán will contribute her time and expertise to the project, and the USC Libraries will provide in-kind support as follows:

- Barbara Robinson (2% of annual salary): $2,018
- Wayne Shoaf (2% of annual salary): $2,247
- Fringe benefits for USC Libraries personnel (33.5% of salaries): $1,429

Timeline

July 2021: Grant Agreement and Project Kick-off

At a project kick-off meeting, we ensure the project team is available and reconfirm our project timetable. We will work with our grants administrators and HR and finance personnel to finalize the grant agreement, establish a grant account, and update Julia Orquera Bianco's status with payroll.

July 2021-June 2022: Transcription

With guidance from Silva-Corvalán and Shoaf and assistance by Robinson, Orquera Bianco will complete orthographic transcriptions for the 93 hours of Chilean recordings. The supplemental LARRP funding will enable her to create transcriptions for 30 hours of recordings along with the 63 hours we expect to transcribe with our current LARRP funding. Orquera Bianco will also ensure all personal information has been redacted.

August 2021-June 2022: Metadata and Publishing

Shoaf will normalize the transcriptions and create updated Qualified Dublin Core metadata records for the 93 hours of recordings from the Chilean corpora, including the 30 hours that Orquera Bianco will transcribe with supplemental LARRP funding through this project. As the records are updated, Shoaf will publish the recordings on the USC Digital Library, Calisphere, and DPLA. We will also update the Spanish Sociolinguistic Research Collection page in the USC Digital Library.

August 2021-June 2022: Dissemination and Outreach
As updated materials are published in the USC Digital Library, Calisphere, and the DPLA, Robinson will update the Latin Americanist research community about the project's progress, and USC Libraries communications staff will conduct outreach via our newsblog and social media presence.

May-June 2022: Assessment and Reporting
We will finalize our evaluation planning after we begin publishing the transcriptions with the digital recordings in the USC Digital Library, Calisphere, and Digital Public Library of America. Our assessment activities will include measuring traffic to the digital collections using Google Analytics and surveys of linguists and Latin Americanists.

Post-project assessment plan
Our primary criteria for judging the success of this project will be whether we meet our proposed timeline and budget for the project and digitize and publish the transcriptions per the high standards of quality we expect for USC Digital Library projects. In addition, we will measure online traffic to the digital collections created by this project using Google Analytics and measure engagement with the collections by the Latin Americanist research community and via social media.

As part of this project, we will administer a survey to linguists and Latin Americanists who work on topics closely related to Spanish language usage in Chile. We will create a brief survey instrument and administer it to 15-20 researchers. The survey instrument will measure their satisfaction with the digital collections and orthographic transcriptions we publish via this project and ask about the features they feel would be most helpful in future phases and online resources to which we should contribute the digital objects.

Additional sources of funding
As part of our proposed 12-month project, the USC Libraries will contribute $5,694 of in-kind costs relating to the participation of the project team, and Carmen Silva-Corvalán will contribute her time and expertise to the project. Silva-Corvalán's field recording work was supported by the Ford Foundation and the National Science Foundation. We were awarded $17,557 by LARRP in 2018 for the creation of transcriptions for the Chilean corpora and a corpus from West Los Angeles in 1978. Due to the challenges associated with the COVID-19 pandemic faced by USC and many other institutions during 2020 as well as added challenges with visa-related issues, we were not able to complete the work as expected. We are therefore requesting $7,569 in supplemental funding to complete the most critical portion of the project.

Previous phases of this larger digitization project were completed with support from a grant for $11,534 from LARRP and an earlier USC Faculty Research Grant. Digitization of the Chilean corpora and other corpora from Silva-Corvalán's research was completed with $10,000 in funding from the Del Amo Fund at USC Dornsife College and the L.A. Murillo Hispanic Heritage Endowment at the USC Libraries.

We will also seek additional funding to support the creation of orthographic transcriptions for 42 hours of recordings of Spanish speakers in West Los Angeles, 21 hours of recordings of bilingual Mexican-American adolescents and 110 hours of audio recordings from Spanish speakers in East Los Angeles. This is our next area of priority for future phases of the Spanish Sociolinguistics Collection digital library project.