EDITORS
- Raquel Freitag (UFS)
- Juliana Bertucci (UFTM)
- Márcia Vieira (UFRJ)

REVIEWERS
- Kendra Dickinson (Rutgers)
- Marta Sousa (IFS)

ABOUT THE AUTHORS
- Katherine Christoffersen
  Conceptualization; Writing – Original Draft; Writing – Review & Editing; Funding Acquisition.
- Isabella Calafate
  Conceptualization; Writing – Original Draft; Writing – Review & Editing; Funding Acquisition.
- Julio Ciller
  Conceptualization; Writing – Original Draft; Writing – Review & Editing; Funding Acquisition.
- Ana Carvalho
  Conceptualization; Writing – Review & Editing; Funding Acquisition.
- Ryan Bessett
  Conceptualization; Writing – Review & Editing; Funding Acquisition.
- Brandon J. Martínez
  Writing – Review & Editing; Data Curation.
- Hannia Rojas Barreda
  Writing – Review & Editing; Data Curation.
- William Flores
  Writing – Original Draft; Writing – Review & Editing.
- Richard Quiroz
  Writing – Original Draft; Writing – Review & Editing.

Received: 05/01/2025
Accepted: 07/17/2025
Published: 08/29/2025

HOW TO CITE
CHRISTOFFERSEN, K.; CALAFATE, I.; CILLER, J.; CARVALHO, A.; BESSETT, R.; MARTÍNEZ, B. J.; ROJAS BARREDA, H.; FLORES, W.; QUIROZ, R. (2025). Sharing and Preserving Sociolinguistic Corpora on the U.S.-Mexico Border. *Cadernos de Linguística*, v. 6, n. 4, e862.

CHECK FOR UPDATES

EXPERIENCE REPORT

SHARING AND PRESERVING **SOCIOLINGUISTIC CORPORA** ON THE U.S.-MEXICO BORDER

Katherine CHRISTOFFERSEN
Department of Writing and Language Studies – University of Texas Rio Grande Valley (UTRGV)
Edinburg, Texas, United States

Isabella CALAFATE
Department of Modern Languages and Cultures – Baylor University (Baylor)
Waco, Texas, United States

Julio CILLER
Department of Writing and Language Studies – University of Texas Rio Grande Valley (UTRGV)
Edinburg, Texas, United States

Ana CARVALHO
Department of Spanish and Portuguese – University of Arizona (UA)
Tucson, Arizona, United States

Ryan BESSETT
Department of Literature – University of California San Diego (UCSD)
San Diego, California, United Statesl

Brandon J. MARTÍNEZ
Department of Spanish and Portuguese – University of Arizona (UA)
Tucson, Arizona, United States

Hannia ROJAS BARREDA
Department of Spanish and Portuguese – University of Arizona (UA)
Tucson, Arizona, United States

William FLORES
Division of Academic Affairs - University of Texas Rio Grande Valley (UTRGV)
Edinburg, Texas, United States

Richard QUIROZ
Division of Academic Affairs - University of Texas Rio Grande Valley (UTRGV)
Edinburg, Texas, United States

ABSTRACT

Since William Labov outlined the methodology for the sociolinguistic interview in 1972, sociolinguistic corpora have been used widely in the field of sociolinguistics to study diverse speech communities and linguistic features. However, most of these invaluable sociolinguistic collections have been available only to the individual researcher or research group, and these data sets usually disappear from use with that individual scholar. More recently, there has been a push towards data sharing in sociolinguistics, reflective of data sharing and the open science movement in other fields. Still, accessible online sociolinguistic corpora are few and far between, in part due to the intense time commitment required to create, sustain, share, and preserve such collections. This paper reviews two accessible online sociolinguistic collections at the U.S.-Mexico border: the Corpus de Español en el Sur de Arizona [Corpus of Spanish in Southern Arizona] or CESA (Carvalho, 2012) and the Corpus Bilingüe del Valle [Bilingual Corpus of the Valley] or CoBiVa (Christoffersen; Bessett, 2019; Christoffersen; Ciller, 2024) in South Texas. We explore these two corpora as case studies for data sharing and preservation through collaboration by detailing the data collection and data management protocols and preservation plans. In doing so, we demonstrate how data sharing in sociolinguistics impacts accessibility, reproducibility, and the democratization of knowledge.

RESUMO

Desde que William Labov desenvolveu a metodologia para a entrevista sociolinguística em 1972, corpora sociolinguísticos têm sido amplamente utilizados no campo da sociolinguística para estudar diversas comunidades de fala e seus aspectos linguísticos. No entanto, a maioria dessas inestimáveis amostras sociolinguísticas tem estado disponível apenas para o pesquisador ou grupo de pesquisa e esses dados geralmente deixam de ser usados após o encerramento da carreira acadêmica do pesquisador. Mais recentemente, tem havido um movimento a favor do compartilhamento de dados em sociolinguística, refletindo o movimento de compartilhamento de dados e ciência aberta em outras áreas. Ainda assim, são raros os corpora sociolinguísticos acessíveis online, em parte devido ao intenso investimento de tempo necessário para criar, manter, compartilhar e preservar tais amostras. Este artigo analisa duas amostras sociolinguísticas disponíveis online com entrevistas coletadas na fronteira entre os Estados Unidos e o México: o Corpus do espanhol no sul do Arizona (Corpus de Español en el Sur de Arizona) ou CESA

(Carvalho, 2012) e o Corpus bilingue do Vale (Corpus Bilingüe del Valle) ou CoBiVa (Christoffersen; Bessett, 2019; Christoffersen; Ciller, 2024) no sul do Texas. Exploramos esses dois corpora como estudos de caso sobre o compartilhamento e preservação de dados através de um trabalho colaborativo, detalhando os protocolos de coleta e gestão de dados e também os planos de preservação. Ao fazer isso, demonstramos como o compartilhamento de dados na sociolinguística impacta a acessibilidade, reprodutibilidade e a democratização do conhecimento.

KEYWORDS
Data Sharing; Preservation; Sociolinguistic Corpora; Open Science; Data Management Plans.

PALAVRAS-CHAVE
Compartilhamento de Dados; Corpora Sociolinguísticos; Ciência Aberta; Planos de Gestão de Dados.

DATA SHARING & SOCIOLINGUISTIC CORPORA

The open science movement has garnered substantial attention and interest in the field of applied linguistics in recent years. Open science is "based on the belief that every aspect of the research process—from hypothesis development to data collection and analysis, to the publication of results—should be transparent and accessible" (García Laborda; Fernández Alvarez, 2024, p. 11). These efforts aim to enhance transparency, rigor, reproducibility, accessibility, and collaboration throughout the research process. The expanding landscape of open science practices includes open access publishing, preregistration, pre-prints, shareable work environments, programming, and code, as well as data sharing. In this paper, we focus on data sharing, specifically in the context of sociolinguistic corpora (see Casillas *et al.*, 2025 for details on a range of other open science practices in linguistics).

Data sharing (or open data) refers to "data collected for research that is freely and easily accessible to anyone interested in accessing it for any purpose" (Open Knowledge, 2023). Researchers are increasingly encouraged to share their linguistic data, often via servers such as the Instruments and Data for Research in Language Studies (IRIS) database (IRIS, 2011). Yet, while data sharing is the norm in some language-related fields, such as human language technologies (Habash *et al.*, 2013) and corpus linguistics (Zinmeister; Breckle, 2013), it remains the exception in linguistics (Bochynska *et al.*, 2023) and sociolinguistics (Cieri, 2014).

The sociolinguistic interview (Labov, 1972) has been one of the most productive and important methods for studying language variation and change for over 60 years now. It has enabled the study of patterns of linguistic behavior across a wide variety of communities and languages. The purpose of the sociolinguistic interview is to elicit naturalistic, vernacular speech patterns, to follow the flow of the conversation and to lessen the observer's paradox. Sociolinguists collect many sociolinguistic interviews from a given speech community, forming a sociolinguistic corpus. Historically, these community-based corpora were private collections accessible only to the individual researcher or research team. In the early years, when technology didn't exist to share digital collections on the internet, some collections might have been shared as tapes, cassettes or records in a special collections archive on campus. Unfortunately, though, we have lost many invaluable sociolinguistic collections from years past, since they were not created with a data sharing or preservation plan.

Data sharing provides opportunities for diachronic and comparative studies which would not be possible otherwise. It also creates opportunities for developing collaborations across contexts or even across disciplinary divides. In addition, data sharing meets requirements from federal funding sources which often require outputs to be publicly accessible. The development of large collections of corpora also responds to trends towards 'big data' (Reyes, 2013). Some recent examples of accessible online sociolinguistic corpora (Bullock; Toribio, 2013; Butragueño; Lastro, 2011; Deuchar *et al.*, 2008) expand the realm of sociolinguistic knowledge and understanding to bilingual and

multilingual language varieties, addressing the monolingual and WEIRD (Western, Educated, Industrialized, Rich, and Democratic) bias in linguistics (Bochynska *et al.*, 2023; Faytak *et al.*, 2024; Nagle *et al.*, 2024).

Yet despite the advantages to data sharing, sociolinguists also face considerable challenges in this pursuit. These often include a lack of time, funding, resources, training, and expertise, along with ethical considerations. These are not inconsequential concerns. After all, sharing linguistic data multiplies the already time-consuming nature of transcription, requiring anonymization of audio, transcripts and various levels of data cleaning and checks before publishing to a website. Scholars may hesitate to undertake such a time-consuming revision, when historically the "time, care, and expertise that go into proper data management in service of reproducible linguistics" or the creation and maintenance of sociolinguistic corpora has not been properly valued (Berez-Kroeker *et al.*, 2022, p. 11). Thankfully, scholars have documented that it is "increasingly common to see the data themselves as an important output, worthy of valuation in hiring, tenure, and promotion" (Berez-Kroeker *et al.*, 2022, p. 11; Alperin, 2022).

Additionally, few sociolinguists have the technical skills required to develop, update, or maintain a website, which is a substantial undertaking. There is also growing attention to the importance of long-term preservation of linguistic data (Berez Kroeker *et al.*, 2022) which would allow corpora to remain available for researchers and allow for both synchronic and diachronic comparisons. Thus, the researcher or research team must develop collaborations with campus Information Technology (IT) and librarians. We also advocate that partnerships with scholars across communities can facilitate data sharing efforts. This paper details how collaborations facilitate the creation, sharing, and preservation of sociolinguistic corpora along the U.S.-Mexico border based on our experience with the Corpus de Español en el Sur de Arizona [Corpus of Spanish in Southern Arizona] or CESA (Carvalho, 2012) in Southern Arizona, and the Corpus Bilingüe del Valle [Bilingual Corpus of the Valley] or CoBiVa (Christoffersen; Bessett, 2019; Christoffersen; Ciller, 2024) in South Texas.

1. DATA COLLECTION AND MANAGEMENT PROTOCOLS

CESA and CoBiVa are two extensive collections of sociolinguistic interviews that document bilingualism on the U.S.-Mexico border—more specifically in Southern Arizona and in South Texas, respectively. In Southern Arizona, approximately 40.9% of the population identifies as Hispanic and 27.7% speaks Spanish at home (United States Census Bureau, 2023). In South Texas, in the region of Rio Grande Valley, the presence of Spanish is even more pronounced. The population is approximately 91.3% Hispanic and 76.1% of the population speaks Spanish as one of their home languages (United States Census Bureau, 2023).

Despite the long-standing presence of Hispanic and Spanish-speaking communities along the U.S.-Mexico border, local bilingual varieties are often delegitimized due to pervasive and harmful language ideologies rooted in purism and standard language norms. These ideologies permeate beliefs of these varieties as impure and inferior, leading to oppressive experiences such as linguistic terrorism (Anzaldúa, 1987; Christoffersen, 2019). In response to this linguistic discrimination, research efforts in the area highlight the richness, complexity, and linguistic integrity of bilingual varieties. Despite the significant work accomplished so far, much remains to be done. Sociolinguistic corpora such as CESA and CoBiVa play a crucial role in advancing and strengthening community-based research, along with its outcomes and implications for both the academic field and the communities involved.

CESA was created in 2012 to document Spanish spoken in Southern Arizona. The project was designed to incorporate a community-engaged fieldwork component, such that students enrolled in Spanish sociolinguistics courses actively participate as collaborators in data collection and transcription. Building on CESA's foundation, CoBiVa was launched in 2017 to document both Spanish and English spoken in the Rio Grande Valley. In CoBiVa, interviews are conducted by students enrolled in community engaged scholarship courses as well as research assistants. Along with documenting bilingual varieties in Southern Arizona and South Texas, the main goals of these corpora include: (1) fostering respect and appreciation for linguistic diversity and local dialects in the U.S.-Mexico borderlands; (2) sharing quality data for linguistic studies with the scholarly community; (3) training students in sociolinguistic research methods and engaging them in community-based research; and (4) establishing and strengthening connections between the local communities and the universities. As research projects involving human subjects, both CESA and CoBiVa underwent Institutional Review Board (IRB) approval at their respective institutions, University of Arizona (UA) and University of Texas Rio Grande Valley (UTRGV). Accordingly, all team members working with the data, including students involved in the data collection process, have completed current training in the Collaborative Institutional Training Initiative (CITI) program to ensure compliance with IRB protocols.

As CESA was initiated, the *CESA Training Handbook* (Bessett; Carvalho, 2015) was developed to guide students through the interview and transcription protocols. With the development of CoBiVa, the training handbook was revised and expanded into the *CESA and CoBiVa Training Handbook* (Bessett; Carvalho; Christoffersen, 2022), which also includes an optimized transcription protocol incorporating a more automated process. The handbook is organized into three main sections: (1) Interview protocol; (2) Forms to complete; and (3) Transcription protocol. The primary goal of the handbook is to provide standardized procedures to ensure consistency in the development of both corpora. With transparency and reproducibility in linguistic research in mind, the handbook can serve as a valuable resource for linguists planning to develop new corpora, offering detailed protocols for the replication of similar data collection and transcription methods. For

additional information, the *CESA and CoBiVa Training Handbook* can be accessed through the official websites of both projects[1]. A resources folder is also available on the CoBiVa website, containing files such as separate Word documents for the different intake forms and other relevant files mentioned below.

As outlined in the handbook, a significant part of building a collection of sociolinguistic interviews involves two interconnected phases: data collection and data transcription. The data collection phase includes recruiting participants, preparing for the interview, conducting the interview, and completing accompanying documentation. Once a participant is recruited and an interview is scheduled, interviewers need to prepare for the interview. As previously mentioned, the sociolinguistic interview is a methodological tool used to elicit natural and spontaneous oral data that reflect as much as possible the linguistic repertoire of a given community. However, the fact that participants are being observed and recorded during the interview presents a significant challenge to collecting spontaneous speech. Following Labov's (1972) model, to reduce the effect of the observer's paradox, these interviews typically last about an hour and focus on the participant's life experiences. For this reason, interviewers begin their preparation by identifying potential conversation topics and creating a plan with questions about the participant's life and interests to encourage them to share narratives as naturally as possible. This planning phase is crucial, as carefully selected questions and a thoughtful interview structure play a key role in minimizing the observer's paradox. We often practice these interviews and emphasize that the word 'interview' is somewhat of a misnomer as we are actually aiming for a casual and comfortable conversation. The "Interview protocol" section of the handbook provides detailed guidelines and examples for interviewers to consider before, during, and after the interview. While CESA interviews are conducted in Spanish, CoBiVa allows members of the research team to choose either Spanish or English as the interview language. In both cases, interviewers are instructed to speak as they naturally would in everyday conversation and follow the flow of the conversation including the interviewees' language choice. These interviews naturally include code-switching and other language-contact phenomena that occur in the borderlands.

In addition to conducting the interview, the data collection process includes gathering information about the interviewee, the interviewer, and the interview (see "Forms to complete" section of the handbook). The interviewee forms include demographic information and a Bilingual Language Profile, which is adapted from the questionnaire developed by Birdsong *et al.* (2012). These documents are fundamental for sociolinguistic analysis. While the first includes information such as place and year of birth, family ancestry, and schooling, among others, the latter consists of

1 CESA website: www.cesa.arizona.edu
CoBiVa website: www.utrgv.edu/cobiva

four sets of questions related to both Spanish and English: language history, language use, language proficiency, and language attitudes. In addition, interviewers are also required to complete a form with demographic information about themselves and a field notes form to document their fieldwork experience including observations that stood out during the interview, as well as details such as the day and time of the interview, the location, the interviewee's attitudes, the formality of the interview, and the levels of silence and comfort. Importantly, before the interview begins, interviewers are required to have the participant read and sign a consent form, which is available in both Spanish and English. This form ensures that the participant understands the study's purpose and procedures, including the anonymization of identifiable information and sharing of the de-identified files on a password-protected website. All these forms are available online in the *CESA and CoBiVa Training Handbook* and on the CoBiVa website (Bessett; Carvalho; Christoffersen, 2022). It is important to note that, as IRB-approved studies, all necessary ethical measures are taken to ensure participant anonymity. Members of the research team remove all identifiable information from the interview by using Audacity, or similar audio editing software, to silence those portions of the audio recording and replace the identifiable information with "XY" in the transcript. Identifiable information includes personal affiliations to specific jobs and schools, home addresses, and personal names.

Once the data collection process is completed, we then move on to the data transcription phase. The transcription process is the most time-consuming and labor-intensive stage, demanding significant time commitment and attention to several details. Initially, transcription was done manually, with the ExpressScribe software often used to facilitate the process. ExpressScribe eases manual transcription through keyboard shortcuts that can allow the insertion of timestamps among other important frequently used features. In order to further facilitate and optimize the transcription process, trials with different technologically aided transcription tools were conducted to determine the most effective transcription software. For more information on this exploration, as well as additional details on the development of the corpora, see Bessett *et al.* (2024). As this project was funded by a grant from the National Endowment of the Humanities (Christoffersen *et al.*, 2020; Christoffersen *et al.*, 2023), it required the development of sustainable solutions which would not require payment or purchasing of transcription services and would be accessible for other scholars and students looking to undertake this work.

This project led to a revised protocol which incorporates a combination of automated and manual transcription techniques, with the goal of streamlining the transcription process. The automated software currently used is Microsoft Stream, a captioning tool. Since Stream is designed for video captions, the first step is the conversion of the audio into a video format. The video is then uploaded to Stream, which automatically generates a transcript. However, the auto-generated transcript includes several non-content lines with system-generated codes and fragmented entries, as expected from a captioning tool. To address these issues, two R scripts have been developed (and are available in the resources folder on the CoBiVa website, along with instructions for use). The first

script is applied after the transcript is generated in Stream to remove unnecessary and distracting lines that interfere with the revision process. After this cleanup, members of the research team revise the transcript. In addition to transcription errors due to possible background noise in the audio, another limitation is Stream's inability to process bilingual audio. If the primary language of the interview is Spanish, for instance, it attempts to interpret the entire audio as Spanish, even English segments, resulting in inaccuracies. Additionally, Stream is not equipped with voice recognition features, requiring the manual assignment of speaker codes. These limitations are addressed during the revision process, when the transcripts are manually revised to add speaker codes which identify speech of the interviewer and interviewee, to correct content issues, and to include additional details following a list of conventions outlined in the handbook. These conventions include marking when a language other than the primary language is used, noting false starts and unintelligible speech, and adding "XY" to personal information that has been silenced. Once revisions are complete, the second R script is run to merge lines corresponding to the same speaker turns, producing the final version of the transcript.

One of the advantages of Stream is that timestamps are automatically included in the auto-generated transcript. These timestamps allow for time-alignment of the audio and the transcript, providing a clickable transcript—a feature already available on the CoBiVa website and soon to be implemented on the CESA website as well. This time alignment significantly enhances data accessibility, enabling researchers to easily locate specific segments in both the audio and the transcript. In addition, it is worth noting that, considering the affordability and sustainability of these corpora, the software used in the transcription process (i.e., Stream and R) is available to members of the research team at no cost. Stream is part of Microsoft Office 365, which is available to students and employees at both institutions (UA and UTRGV), while R is a free, open-source programming language. Thus, the transcription methods present an example of the open science movement, as other research teams may model the development of sociolinguistic corpora projects based on these methods and processes using open-source software.

More recently, the team has been piloting a one step process whereby students input simplified speaker codes on the auto-generated transcript, and then it is processed through one R code. The team has also begun exploring the auto-identification of speaker turns, which is starting to become available through advancements in technology. This would further speed and streamline the transcription and transcription revision processes. All updates will be documented in an updated Handbook as well as on the CoBiVa News website, and these will be further preserved through the preservation process in collaboration with the university libraries (see Section 4).

Currently, the CESA corpus has 78 interviews available on the website, while CoBiVa has 76, with additional interviews for both corpora that are currently in the process of revision. Participant metadata is collected through the required forms completed in the interview and is made available on the CESA and CoBiVa websites. Both corpora feature participants from a range of demographic

profiles, with variation in age, sex, education, and immigrant generation. Unfortunately, the demographic representation of participants is uneven, and a clear pattern emerges across both corpora: most participants are female, young adults, often college-aged, and most were either pursuing or had completed higher education at the time of the interview. A key demographic distinction between the two corpora lies in immigrant generation. Most CoBiVa participants are first generation, meaning they were born in Mexico, whereas CESA primarily includes second-generation participants, that is, individuals born in the U.S. to parents born in other Spanish-speaking countries, predominantly Mexico. As the projects move forward, one of the main goals is to diversify and balance the participant pool. Specifically, efforts are being made to recruit more individuals aged 30 and older, more male participants, and those who are not currently enrolled in college or have not obtained a college degree. Additionally, more first-generation speakers are being recruited in CESA and more second- and third-generation participants in CoBiVa to achieve a better balance in terms of generational groups.

Access to the interviews and accompanying documentation on the CESA and CoBiVa websites is available upon request. In consideration of the ethical implications concerning the availability of linguistic data in both corpora, scholars interested in accessing the data are asked to submit a request by providing their name, email address, institutional affiliation, a copy of their CV, and a brief explanation of their reason for requesting access. This is important because, while we aim to contribute to the sharing of linguistic data with the scholarly community, we also want to ensure that the data is used for legitimate research and educational purposes. Especially in this case, as we are collecting data from linguistically marginalized communities, it is crucial to take these measures to protect the individuals who have generously agreed to participate in the interviews. As mentioned in Bessett *et al.* (2024), there have been attempts to access the corpora for documenting "errors bilinguals commit", for example; however, in these cases, access was not granted. Users granted access to the CESA and CoBiVa corpora are expected to respect the participants, maintain confidentiality, cite the corpora in their work, and share any presentations or publications based on the data with the research teams. For more information and to request access to these corpora, please visit the CESA website (www.cesa.arizona.edu) and the CoBiVa website (www.utrgv.edu/cobiva).

2. ADDITIONAL INFORMATION ON WEBSITES

The websites also host a variety of additional relevant materials and information. In parallel, both the CESA and CoBiVa websites host a list of sociolinguistic corpora, serving as valuable resources for researchers, educators, and community members interested in language variation and multilingualism. The compilation includes several accessible online databases for Spanish, along with other languages such as American English, Brazilian Portuguese, Catalan, and Sign Languages,

among others. These corpora serve as essential tools for sociolinguistic research, providing rich datasets for analyzing language variation and multilingual practices. They enable researchers to explore language use across diverse sociocultural contexts and support a wide range of linguistic inquiries, while also highlighting and supporting other efforts in data sharing.

However, while accessibility is crucial, it must be accompanied by ethical and scholarly use of these resources. Proper citation is a key component of this responsibility. The CESA and CoBiVa research teams explicitly request that researchers follow the citation protocol and credit the corpus in their scholarly work. This is accomplished through a 'Cite the Corpus' page. Acknowledging the intellectual and labor contributions behind these corpora not only ensures credit is given where due but also supports the sustainability of open-access projects. Responsible citation practices reinforce a culture of respect and accountability in the field, underscoring the value of community-engaged linguistic data collection.

To further engage users and enhance their understanding of the project, a welcome video was added to the main CoBiVa website. In this video, we introduce the project, explain its purpose, and describe the types of language phenomena it documents, such as code-switching and the use of loanwords like *troca* or *parquear*. The video also outlines the project's development, significance, and includes testimonials from research assistants who contributed to its creation. Expanding on this multimedia content, the *CoBiVa News* section of the website serves as a platform for disseminating ongoing updates and project achievements. These include updates and developments, announcements of grants, academic awards, team member accomplishments, conference presentations, and related publications, among other highlights.

Building on our commitment to accessibility and outreach, a dedicated module on language ideologies was developed, published on the CoBiVa website, and fully translated into Spanish (De Anda *et al.*, 2023). This module was designed to support educators in incorporating CoBiVa into their teaching and aims to promote greater awareness and appreciation of linguistic variation. The module provides background information on language ideologies, defined as "beliefs, or feelings, about languages as used in their social worlds" (Kroskrity, 2004, p. 498), alongside brief illustrative examples from CoBiVa interviews (audio excerpts with transcripts). To further support classroom use, the module also includes discussion questions and sample activities.

To support research and transparency further, the CESA and CoBiVa teams have compiled and cleaned detailed metadata for their sociolinguistic corpora. These metadata, organized into two Microsoft Excel sheets, contain interviewees' demographic information, linguistic history, language use, language competence, and language attitudes. The CESA sheet includes data from 78 interviewees, while the CoBiVa sheet contains information of 76 CoBiVa participants. More data will be added to these spreadsheets upon review and processing. These datasets are drawn from responses to the forms mentioned in Section 2, more specifically the demographic information of the participant, the Bilingual Language Profile (Birdsong *et al.*, 2012), and field notes. The team is

currently developing data visualizations to represent this data through graphs to highlight patterns and trends. As Unwin (2020) notes, data visualization is essential for identifying structures, outliers, and trends, as well as for presenting findings in a clear and engaging way. We also hope this broadens the reach and accessibility of the findings beyond scholarly audiences to the community.

To support these efforts, the UTRGV IT department developed an Application Program Interface (API) endpoint to allow secure and efficient access to CoBiVa interview data. This API endpoint enables external systems to retrieve and process the data using R (the same free, open-source programming language used to clean the auto-generated transcripts in the interview transcription process), with the ability to display the results as a public-facing table. Importantly, the data remain dynamically linked to the source, ensuring that any updates to interview content are automatically reflected in real time. In parallel, a Shiny web application has been developed and tested locally as a prototype (Posit, s.d.). This tool demonstrates how metadata tables might appear on the public CoBiVa website and showcases the potential for embedding interactive visualizations, offering users an intuitive and engaging experience.

The CoBiVa website has also recently been translated into Spanish; thus, users can toggle between the Spanish or English version of the website. This effort focused on the main menu, key navigation elements, and content of the CoBiVa website. This further serves to elevate Spanish and bilingualism in the context of the university and the local borderlands context, while also enhancing accessibility for Spanish-speaking users.

In the future, the research teams plan to implement language tagging to assign metadata properties indicating the languages used (English, Spanish) in each interview. This enhancement will enable more refined searches and analyses, improving the user's ability to navigate and explore the dataset, as well as our understanding of bilingualism along the U.S.-Mexico border. This project is inspired by the work of Doğruöz *et al.* (2021), which addresses the complexities of code-switching research and language technologies.

3. LONG TERM PRESERVATION

While there is an increased understanding and appreciation for the importance of long-term data preservation of linguistic data (Berez-Kroeker *et al.*, 2022), is too often an afterthought. In our case, a grant from the National Endowment of Humanities (NEH) through the Humanities Collection and Reference Resources (HCRR) program encouraged us to consider the role of preservation early on. While drafting our first foundation level NEH HCRR grant in 2018-2019, we reached out to librarians on both campuses, and a collaborative partnership has continued since then. Our library partners assisted us in drafting short-term and long-term preservation plans along with considering how to preserve other aspects of the project, including webpages, blogs, modules, videos, and data

visualizations. The processes and procedures developed with the UTRGV University Library are described in detail below. The University of Arizona Libraries will be modeling their preservation of CESA based on these plans.

Following our data management protocol (Section 2), the digital files processed through UTRGV University Library are anonymized to conceal personally identifiable and sensitive information. As a short-term backup, the digital files that have been organized are housed in UTRGV's institutional repository known as ScholarWorks. ScholarWorks is powered through Digital Commons and serves as a repository for "working papers, copies of published articles, supplementary files, datasets, and conference papers" (UTRGV ScholarWorks, 2025). Here the interviews can be viewed, while the repository can "distribute electronic copies of the work for the lifetime of the repository, or based upon the agreed timespan, and translate it as necessary to ensure it can be read by computer systems in the future" (UTRGV ScholarWorks, 2024). As of now, the collection is a closed repository, in which users of ScholarWorks cannot access the files themselves and only authorized users can download these digital materials. The ScholarWorks CoBiVa page currently points viewers to the CoBiVa website, and it is used to increase discoverability of the collection (UTRGV ScholarWorks, 2024). These items are also automatically uploaded to Amazon Web Service (AWS) Simple Storage Services (S3) storage bucket, where the library can access the full files of the access copies. These digital materials are stored as information packets/zip files which include audio files, transcription files, agreement forms, and metadata pertinent to the interviewee and the project itself. The memorandum of understanding and the informed consent documents are included but not publicly available. During processing, these files are grouped together to create Digital Information Packages (DIPs). The backups are deposited into ScholarWorks, and the digital files are processed though the University's Library Special Collection and Archives Department digital workflow for long-term preservation.

As part of the Digital Workflow of the UTRGV's Special Collections and Archives department, the DIPs are processed in alignment to recognized archival standards. UTRGV's Digital Preservation Policy ensures to comply with the Reference Model for an Open Archival Information System (OAIS). The audio files are preserved in the .wav format while the accompanying metadata are preserved as Portable Document Format Archiving (PDF/A) files. These formats are sustainable for long-term preservation according to the Library of Congress "Sustainability of Digital Formats" and as described in the following: "The .wav file format is 'widely adopted' and is a 'preferred or recommended format for sound in many long-term archives' and the PDF/A format is 'widely recommended for page-oriented documents as a format that is ready for archiving'" (Sustainability of Digital Formats, 2023).

Regarding the integrity of the digital information packages, they are processed through the Fixity application. This application checks their integrity as they are processed onto M-Discs and the Chronopolis drives. The fixity of the digital files helps with transparency and "assurance that a digital

file has remained unchanged" by using a checksum, which acts as a "digital fingerprint" through a string of characters which would change if the file has been altered or changed in any way (Digital Preservation Coalition, 2025). By assessing the fixity of the digital files through fixity software, the archives will thoroughly evaluate and maintain the sustainability of the digital files for long-term preservation. Furthermore, the preservation plan for the project follows the Linguistic Data Consortium Preserving Data activities, which cover areas such as storage, updates & migration, backups, and data security.

It is outlined in UTRGV's Digital Preservation Policy that "five copies of digital materials will be created" which includes "two access hard copies, two preservation hard copies, and a digital preservation copy" (UTRGV Digital Preservation Policy Framework, 2018, p. 2). The two preservation and two access hard copies are M-Discs. M-Discs are meant for long term preservation of digital media such as photos, videos, and audio materials. They are similar to DVDs except they are created with "materials that are resistant to degradation by UV light and moisture" (Stiemer, 2023). One preservation and one access copy will be stored on the Edinburg campus while the other preservation and access copy will be stored on the Brownsville campus. As stated in the UTRGV Digital Preservation Policy Framework, "the access hard copies will be available in the reading room, while the preservation copies will be kept in the fire vault" (UTRGV Digital Preservation Policy Framework, 2018, p. 2). The fifth digital copy will be stored on the cloud service known as Chronopolis.

The preservation copy through Chronopolis requires setting up metadata forms, file organization, and ingesting the files into ScholarWorks, UTRGV Preservation server systems and Chronopolis. This preservation work was conducted by a graduate assistant and overseen by the Head of Special Collections and Archives as well as a librarian within the Open Scholarship department. In addition, they included quality checks of the digital files and adhered to the workflow for ingesting the files. Chronopolis has several available tools for file transfer such as Data ingest, Data auditing, and the BagIT transfer format. Once processed and uploaded to Chronopolis, the UTRGV Special Collections and Archives workflow is completed with a total of four physical copies created, and one digital copy created on a cloud-drive.

Public-facing digital materials related to the CoBiVa project itself are archived and stored securely. Since these are born digital materials, they are available in ScholarWorks in a public collection, separate from the private CoBiVa collection. These include blog posts, news articles, and press releases which are converted into PDF/A files. Web pages with statements about the terms and conditions of using the CoBiVa site that provide crucial legal information are also converted to PDF/A files. Afterwards, they are saved under a uniformed file naming convention to ensure orderly consistency and deposited into Chronopolis for long term preservation. Other supplementary materials include the *CESA and CoBiVa Training Handbook* and its resources folder currently available on the CoBiVa website. Any future supplementary materials created for the CoBiVa site

such as future press releases, posts, videos, or data visualizations generated for the project will follow the previously mentioned workflow.

5. IMPACT

The development of these sociolinguistic corpora has provided many opportunities for students to participate as collaborators in sociolinguistic research. Since its inception in 2012, over 80 undergraduate and graduate students enrolled in Spanish sociolinguistics courses have contributed sociolinguistic interviews to the CESA corpus. Furthermore, 11 undergraduates and 15 graduate students have worked as research assistants on the project. For CoBiVa, 100 students have participated as interns on the project in community engaged scholarship classes. Additionally, multiple funding opportunities have provided a total of 54 semester-long student research assistantship positions for students at UTRGV. These students have participated in research presentations at local, regional, national, and international conferences, and some have even co-authored papers or completed theses based on the CoBiVa and CESA.

As an accessible sociolinguistic corpus, CoBiVa has had an expansive reach. To date, as of March 2025, 475 people have access to the CESA corpus, including individuals from both the University of Arizona and Arizona State University, 26 other U.S. states (Alabama, California, Colorado, Florida, Georgia, Hawaii, Illinois, Indiana, Iowa, Louisiana, Massachusetts, Michigan, Minnesota, Mississippi, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oregon, Pennsylvania, Tennessee, Texas, Virginia, Wisconsin), 14 other countries (Argentina, Austria, Belgium, Brazil, Canada, Denmark, England, France, Germany, Italy, Mexico, Netherlands, South Korea, Spain), and several community members and educators at the K-12 level. In turn, 491 individuals have requested access to the CoBiVa, and while many of these are UTRGV students, there are individuals from 7 different universities in Texas (University of Texas El Paso, Baylor, University of Texas Austin, Southwestern, Texas A&M University, University of North Texas, Dallas College), from 11 other states (Hawaii, Virginia, Arizona, California, Iowa, North Carolina, Massachusetts, Illinois, Pennsylvania, Wisconsin, Oregon), 4 different countries (Belgium, Mexico, Canada, and Spain) as well as community members.

Both corpora have also been very productive in terms of research. For CESA-based research that we are aware of to-date, publications include 12 articles in peer-reviewed journals, five book chapters, six doctoral dissertations, one master's thesis, and one undergraduate thesis. Furthermore, to our knowledge CESA has been used as the basis for at least three invited talks, one keynote talk, two panel discussions, four posters, 19 conference presentations, and one workshop for heritage language learners. For the CoBiVa, our own research team has produced four publications, a thesis, a handbook, a white paper, four posters, and 21 scholarly presentations based on the CoBiVa. Beyond this work by our research team, we were able to identify three additional publications which cited the CoBiVa.

Recently, scholars from California initiated a sociolinguistic corpus modeled on the CESA and CoBiVa, entitled Multilingual Hispanic Speech in California (MuHSiC), a collaboration recently presented at the Spanish in the U.S. conference (Amengual *et al.*, 2025). Therefore, it is clear that CESA and CoBiVa are providing opportunities for undergraduate and graduate students, providing data for sociolinguistic research, inspiring further development of sociolinguistic corpora, and modelling the practice of data sharing in sociolinguistics, which has been a major goal of this endeavor all along. All of this demonstrates the potential of such initiatives for enhancing access to sociolinguistic data, reproducibility of research, the further development of sociolinguistic corpora, and the democratization of knowledge, which are key goals of the open science movement.

5. DISCUSSION

The collaborations with librarians for long-term preservation will ensure that these collections are accessible beyond any individual researcher's lifetime. These measures value the time, effort and energy that researchers, research assistants, and community members have devoted to this project. Additionally, since "it is within government, within education, within media, within all organizations providing public services that language ideology is most powerfully promoted, and in which there is greatest scope for the reorientation of linguistic values," we argue that the incorporation of linguistically diverse collections in institutional collections is a site of linguistic contestation (Hannaford; Alexander, 2024, p. 94). The preservation of the CESA and CoBiVa within university libraries and on university websites are transformative tools which attribute prestige to bilingual language varieties along the U.S.-Mexico border. The context of these sociolinguistic corpora is particularly poignant as it represents a site of contestation that has seen a history of harsh and ill-informed criticism towards bilingualism in these areas (Christoffersen, 2019; Christoffersen, Under Review). For this reason, students involved in these classes often remark that these sociolinguistic corpora are personally meaningful to them (Bessett *et al.*, 2016; Bessett *et al.*, 2024; Christoffersen; Regalado, 2021; Christoffersen *et al.*, 2023). In one reflection, a student wrote:

> The CoBiVa project is such **a wonderful initiative that makes me feel so validated**. It is so fascinating seeing **the language I hear every day,** the language which is seldom depicted in a positive and accurate way in media, **as the subject of a formal study and considered worth documenting and preserving**. ... [In this project I have experienced] a sense of participation in the documentation of **the language that surrounds me and that is part of my life**. (student reflection in Christoffersen *et al.*, 2023, p. 12).

So, despite all the challenges of sharing sociolinguistic data, the benefits outweigh them all, especially when we can see that it makes a difference to our students. Through collaborations within and across universities, it is possible to develop frameworks for data sharing, preservation and

valorization of local language varieties in ways that impact scholarship, students, the community, and society more broadly.

ACKNOWLEDGEMENTS

We would like to acknowledge Justin White for his work on the library preservation plans. We also wish to thank the many students, research assistants, collaborators, and community members who have participated in this project.

ADDITIONAL INFORMATION

CONFLICT OF INTEREST

The authors declare no competing interests.

STATEMENT OF DATA AVAILABILITY

While no new data were created or analyzed in this study, both sociolinguistic corpora are available upon request on their respective websites. The Corpus Bilingüe del Valle (CoBiVa) is available at utrgv.edu/cobiva, and the Corpus del Español en el Sur de Arizona is available at cesa.arizona.edu.

AI USAGE STATEMENT

The authors used ChatGPT and Gemini for minor grammatical revisions, alternatives for select individual original sentences, and idea generation for synonyms with a minimal contribution to the manuscript. All content was reviewed, revised, and edited by the authors, who assume full responsibility for the final manuscript.

ETHICS AND CONSENT

The CESA and CoBiVa projects were approved by the IRB at each institution (University of Arizona and University of Texas Rio Grande Valley, respectively). Proper informed consent it obtained before conducting interviews.

FUNDING SOURCES

This research was supported by the National Endowment for the Humanities [PW-269430-20, PW-290585-23].

REVIEW AND AUTHORS' REPLY

Review: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID862.R
Author's Reply: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID862.A

REFERENCES

ALPERIN, Juan Pablo; SCHIMANSKI, Lesley A.; LA, Michelle; NILES, Meredith T.; MCKIERNAN, Erin C. The Value of Data and Other Non-traditional Scholarly Outputs in Academic Review. *In:* BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. *The Open Handbook of Linguistic Data Management*. The MIT Press, 2022. p. 171–182.

AMENGUAL, M.; KIM, J. Y.; DAVIDSON, J. (Multilingual Hispanic Speech in California (MuHSiC). *In*: Accessible online sociolinguistic corpora: Challenges, opportunities, and programs, panel organized by Katherine Christoffersen at SPANISH IN THE U.S. CONFERENCE. San Antonio, Texas: University of Texas at San Antonio, 2025.

ANZALDÚA, Gloria. *Borderlands = la frontera: the new mestiza*. San Francisco: Spinsters/Aunt Lute, 1987.

BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; COLLISTER, Lauren B.; KOLLER, Eve. Data, Data Management, and Reproducible Research in Linguistics: On the Need for The Open Handbook of Linguistic Data Management. *In:* BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. *The Open Handbook of Linguistic Data Management*. The MIT Press, 2022. p. 3-8.

BESSETT, Ryan M.; CARVALHO, Ana M.; CHRISTOFFERSEN, Katherine. The CESA and CoBiVa Training Handbook. 2022. https://bit.ly/CESA_CoBiVa_Handbook

BESSETT, R. M.; CARVALHO, A. M.; KERN, J. The full cycle of the sociolinguistic enterprise: Corpus building, student engagement, and critical language pedagogy. *In:* New Ways of ANALYZING VARIATION (NWAV) 45 CONFERENCE. Vancouver, BC, Canada: Simon Fraser University, 2016.

BESSETT, Ryan M.; CHRISTOFFERSEN, Katherine; CARVALHO, Ana M.; CALAFATE, Isabella; VEGA MUDY, Mayte. Developing Community-Based Sociolinguistic Corpora to Promote Social Justice. *In:* LAMAR PRIETO, Covadonga; GONZÁLEZ ALBA, Álvaro. *Digital Flux, Linguistic Justice and Minoritized Languages*. Berlin, Boston: De Gruyter, 2024. p. 195-214. https://doi.org/10.1515/9783110799392-011

BIRDSONG, David; GERTKEN, Libby M.; AMENGUAL, Mark. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. COERLL, University of Texas at Austin, January 2012. https://sites.la.utexas.edu/bilingual/.

BOCHYNSKA, Agata; KEEBLE, Liam; HALFACRE, Caitlin; CASILLAS, Joseph V.; CHAMPAGNE, Irys-Amélie; CHEN, Kaidi; et al. Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics*, 2(1), 2023. http://dx.doi.org/10.5070/G6011239.

BULLOCK, Barbara E.; TORIBIO, Almeida Jacqueline. *The Spanish in Texas Corpus Project.* COERLL, The University of Texas at Austin, 2013. https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/GBYLXB

BUTRAGUEÑO, Pedro Martín; LASTRA, Yolanda. *Corpus sociolingüístico de la Ciudad de México.* México: El Colegio de México, 2011.

CARVALHO, Ana M. *Corpus del Español en el Sur de Arizona (CESA).* University of Arizona, 2012. cesa.arizona.edu.

CASILLAS, Joseph V.; CONSTANTIN-DURECI, Gabriela; ANDREU RASCÓN, Iván; SHAO, Jiawei; RODRÍGUEZ, Stephanie A.; GADAMSETTY, Adrija; MINETTI, Alexandria; LAUNGANI, Krishita; THATCHER, John; GARDERE, Rhode; TAVERAS, Katherine; CHANG, Isabelle; RODRÍGUEZ, Nicole; PARRISH, Kyle; FELIU RIBAS, Meritxell; ESPOSITO, Robert. Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research. *PsyArXiv*, 2025. https://doi.org/10.31234/osf.io/spz4w

CHRISTOFFERSEN, Katherine. (Under Review). "Hablo pocha, ¿no?": Language ideologies of pocho and mocho in the U.S.-Mexico borderlands.

CHRISTOFFERSEN, Katherine. Linguistic terrorism in the borderlands: Language ideologies in the narratives of young adults in the Rio Grande Valley. *International Multilingual Research Journal,* London, v. 13, n. 3, p. 137-151, 2019.

CHRISTOFFERSEN, Katherine; BESSETT, Ryan M. *Corpus Bilingüe del Valle (CoBiVa).* University of Texas Rio Grande Valley, 2019. https://utrgv.edu/cobiva

CHRISTOFFERSEN, Katherine; BESSETT, Ryan M.; CARVALHO, Ana M.; CILLER, Julio; CALAFATE, Isabella. Bilingual voices in the U.S./Mexico borderlands phase 2: Preserving, expanding, and elaborating sociolinguistic collections. National Endowment of the Humanities Award. Humanities Collections & Reference Resources. Implementations Grant. 2023. https://www.neh.gov.

CHRISTOFFERSEN, Katherine; CARVALHO, Ana M.; BESSETT, Ryan M. Bilingual voices in the U.S./Mexico borderlands: Technology-enhanced transcription and community engaged scholarship. National Endowment for the Humanities. Humanities Collections & Reference Resources. Foundations Grant. 2020. https://www.neh.gov.

CHRISTOFFERSEN, Katherine; CILLER, Julio. *Corpus Bilingüe del Valle (CoBiVa).* University of Texas Rio Grande Valley, 2024. https://utrgv.edu/cobiva

CHRISTOFFERSEN, Katherine; REGALADO, Kimberly. "Toda lengua es válida aquí en esta clase": Translanguaging pedagogy and critical language awareness in sociolinguistics courses on the U.S.-Mexico border. *Journal of Bilingual Education Research and Instruction*, v. 23, n. 1, p. 23-71, 2021.

CHRISTOFFERSEN, Katherine; VILLANUEVA, Aubrey; BESSETT, Ryan. Student perceptions of community engaged scholarship courses: Developing a sociolinguistic corpus on the U.S.-Mexico border. *International Journal of Research on Service-Learning and Community Engagement*, v. 11, n. 1, p. 1-19, 2023.

CIERI, Christopher. Challenges and Opportunities in Sociolinguistic Data and Metadata Sharing. *Language and Linguistics Compass*, v. 8, n. 11, p. 472-485, 2014. https://doi.org/10.1111/lnc3.12112

DEANDA, Carolina; CHRISTOFFERSEN, Katherine; CILLER, Julio. Ideologías lingüísticas. 2023. Available at: https://www.utrgv.edu/cobiva/resources/teachers/index.htm. [Online teaching module integrating audio clips and transcript from CoBiVa]

DEUCHAR, Margaret; COUTO, Maria del Carmen Parafita; WEBB-DAVIES, Peredur; DONNELLY, Kevin. *BangorTalk*. 2008. https://bangortalk.org.uk/index.php

DIGITAL PRESERVATION COALITION. Fixity and checksums. Glasgow: Digital Preservation Coalition, 2025. Available at: https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums

DOĞRUÖZ, A. S.; SITARAM, S.; BULLOCK, B. E.; TORIBIO, A. J. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. *In:* PROCEEDINGS OF THE JOINT CONFERENCE OF THE 59TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND THE 11TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (ACL-IJCNLP 2021). Bangkok, Thailand, 2021. p. 1654-1666. https://aclanthology.org/2021.acl-long.131.pdf

FAYTAK, Matthew; KADAVÁ, Šárk;, XU, Chenzi; ÖZSOY, Onur; AKUMBU, Pius W.; CARDOSO, Amanda, ... ROETTGER, Timo B. Big team science for language science: Opportunities and challenges. *Open Science Framework*, 2024. Retrieved from osf.io/3pkj6

GARCÍA LABORDA, Jesús; FERNÁNDEZ ALVAREZ, Miguel. Introducing open science in applied linguistics. *In:* CURADO FUENTES, Alejandro; RICO GARCÍA, Mercedes; FIELDEN BURNS, Laura. *Exploring open science in applied linguistics: Reviews and case studies*. Applied Linguistics Press, 2024. p. 11-34.

HABASH, N.; ROTH, R.; RAMBOW, O.; ESKANDER, R.; TOMEH, N. Morphological Analysis and Disambiguation for Dialectal Arabic. *In:* PROCEEDINGS OF THE 2013 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES (NAACL-HLT). Atlanta, GA, 2013, p. 426-432.

HANNAFORD, Ewan D.; ALEXANDER, Marc. Linguistic diversity in institutional collections: Beyond preservation to valorisation. *International Journal of Language Studies*, v. 18, n. 2, p. 91-112, 2024. https://doi.org/10.5281/ZENODO.104752808

IRIS. IRIS digital repository of instruments and materials for research into second languages. York: University of York, 2011. Available at: https://www.iris-database.org/

KROSKRITY, Paul V. "Language ideologies". *In:* DURANTI, Alessandro. *A Companion to Linguistic Anthropology.* Malden, MA: Blackwell, 2004. p. 496-517.

LABOV, William. *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press, 1972.LIBRARY OF CONGRESS. Format Descriptions. Washington, D.C., 18 June 2024. Available at: https://www.loc.gov/preservation/digital/formats/

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. *Big data: A revolution that will transform how we live, work, and think.* Boston: Houghton Mifflin Harcourt, 2013.

NAGLE, Charlie; BAESE-BERK, Melissa; AMENGUAL, Mark; CASILLAS, Joseph V. Sound communities: A quantitative proposal for studying bilingualism in context. *PsyArXiv*, 2024. https://doi.org/10.31234/osf.io/m67tx

OPEN KNOWLEDGE. The Open Definition. 2023. Retrieved from https://opendefinition.org

POSIT. *Shiny*. (s.d.) Available at: https://shiny.posit.co/. Accessed on: 20 April 2025.

REYES, Angela. Linguistic anthropology in 2013: Super-new-big. *American Anthropologist*, v. 116, n. 2, p. 366-378, 2014.

STIEMER, Friedrich. Ultimate backup: Archival M-Discs store your data for 1000 years. PCWorld, 01 August 2023. Available at: https://www.pcworld.com/article/2015499/storing-data-long-term-m-disc-best-method.html

UNITED STATES CENSUS BUREAU. American Community Survey. Washington, D.C., 2023. Available at: http://data.census.gov

UNIVERSITY OF TEXAS RIO GRANDE VALLEY. Scholarworks: About Institutional Repositories. Rio Grande Valley, Texas, 2024. Available at: https://scholarworks.utrgv.edu/about.html.

UNIVERSITY OF TEXAS RIO GRANDE VALLEY. Scholarworks: CoBiVa. Rio Grande Valley, Texas, 2024. Available at: https://scholarworks.utrgv.edu/cobiva/

UNWIN, Antony. Why is data visualization important? What is important in data visualization. *Harvard Data Science Review*, v. 2, n. 1, p. 1, 2020.

ZINSMEISTER, H.; BRECKLE, M. Starting a sentence in L2 German – discourse annotation of a learner corpus. *In:* PROCEEDINGS OF KONVENS, Saarbrücken, Germany, 2013.