4-2024

Developing Community-Based Sociolinguistic Corpora to Promote Social Justice

Ryan M. Bessett
University of California - San Diego

Katherine Christoffersen
The University of Texas Rio Grande Valley, katherine.christoffersen@utrgv.edu

Ana M. Carvalho
University of Arizona

Isabella Calafate

Mayte Vega Mudy

Recommended Citation

Bessett, Ryan M., Christoffersen, Katherine, Carvalho, Ana M., Calafate, Isabella and Vega Mudy, Mayte. "Developing Community-Based Sociolinguistic Corpora to Promote Social Justice". Digital Flux, Linguistic Justice and Minoritized Languages, edited by Covadonga Lamar Prieto and Álvaro González Alba, Berlin, Boston: De Gruyter, 2024, pp. 195-214. https://doi.org/10.1515/9783110799392-011

Ryan M. Bessett, Katherine Christoffersen, Ana M. Carvalho, Isabella Calafate, Mayte Vega Mudy

Developing Community-Based Sociolinguistic Corpora to Promote Social Justice

Abstract: This chapter explores the many components that are involved in creating a student-based sociolinguistic corpus. Sociolinguistic corpora can be used as tools for social justice in that they promote local (or often stigmatized) varieties of language and students who speak said varieties often experience heightened language pride or greater esteem for their own language. Using the Corpus del Español en el Sur de Arizona (Carvalho 2012-) and the Corpus Bilingüe del Valle (Christoffersen and Bessett 2019-) as models, this chapter first details how to build the corpus, including the documents needed, the interview protocol, the transcription protocol, and the creation of a website. Next, since the most daunting and time-consuming task is transcription, we report the results of field trials with various technologically-aided transcription methods to help improve the process. Lastly, we explore the ways in which the corpus can be used to promote social justice and how to incorporate the corpus into the classroom. By providing and explaining the tools necessary to create a corpus, we hope this chapter inspires others to create student-based semi-open sociolinguistic corpora throughout the United States and around the world.

Keywords: Corpus, corpora, sociolinguistics, community-based

1 Introduction

The importance of sharing sociolinguistic data has been the subject of workshops (Nagy and Lyskawa 2016), included in publications (Mallinson 2013), and encouraged by funding agencies (NSF 2016). In line with these initiatives is the creation of student-based corpora of U.S. Spanish. For example, in the Spanish in Texas corpus project (Bullock and Toribio 2013), students take part in building sociolinguistic corpora. Student-based corpora not only offer important linguistic data, but also promote social justice in the classroom by providing students with training in sociolinguistic methods and enhanced sociolinguistic awareness, and by fostering an appreciation for local language varieties. Using the Corpus del Español en el Sur de Arizona (CESA) (Carvalho 2012-) and the Corpus Bilingüe del Valle

https://doi.org/10.1515/9783110799392-011

(CoBiVa) (Christoffersen and Bessett 2019-) as models, this chapter discusses the process of creating a sociolinguistic corpus, provides an analysis of different transcription methods to ease the transcription process, and discusses how the corpora can be used as a social justice tool.

Inspired by Labov's (1984) model of neighborhood studies, students in graduate and undergraduate classes in two border communities in the U.S. Southwest participate in developing CESA in Southern Arizona and CoBiVa in South Texas. Students who speak Spanish as a heritage language collect data in their communities and document and analyze their home dialects, while students who speak Spanish as a second language are able to be in contact with the dialect spoken in the communities where they live, instead of the idealized monolingual standards still commonly presented in L2 classes. Students are trained in best practices for conducting sociolinguistic interviews, are provided with sample questions to ask participants, and create a sociolinguistic plan. After conducting sociolinguistic interviews with local Spanish-English bilinguals, they transcribe the interviews for inclusion in the corpora.

The resulting corpora are useful to scholars and are also used to incorporate social justice and sociolinguistic diversity in classes (Schilling 2013; Wolfram 2013). One way in which sociolinguistic corpora accomplish these goals is by countering pervasive standard language ideologies, which present one form of a language as correct or true. In the context of the U.S.-Mexico border, bilinguals often face damaging standard language ideologies which disparage individuals' linguistic varieties (Christoffersen, 2019). By documenting bilingual language practices and teaching students about their highly sophisticated nature, these corpora combat standard language ideologies and the linguistic insecurity they often cause. In this way, community-based corpus building projects contribute to the maintenance of heritage languages (Martínez 2003; Leeman 2018). The corpora are also used to present grammar lessons in Spanish courses, thus positioning the corpora as alternatives to textbooks and delegating prestige to bilingual language varieties on the U.S.-Mexico border. As such, we hope that this chapter will encourage further development of local community-based, community-driven corpora.

To accomplish this aim, the chapter will detail the elements involved in building a corpus, providing examples from CESA and CoBiVa in section 2. Later, section 3 outlines our findings on how to make the most difficult part of corpus building, the transcription process, less time-consuming by employing technologically-aided transcription methods. Finally, section 4 will explore the ways in which the corpus can be used as a tool for social justice.

2 Building the Corpus

Creating a student-based sociolinguistic corpus with semi-open access involves many considerations which can be separated into the following dimensions: the factors to consider before creating the corpus, the creation of protocol for collecting interviews, the development of protocol for transcribing the interviews, and finally, the creation of a website and infrastructure for housing the data. Since students will be collecting interviews, it is important to have all procedures and protocols thoroughly documented before starting the project. In this section we will detail these many considerations using the CESA (Carvalho 2012-) and CoBiVa (Christoffersen and Bessett 2019-) corpora and protocols as models.

2.1 Preliminary Considerations for Corpora Development

Before starting the corpus, it is important to consider factors such as the population to be included, the data collection methods, and the ethical considerations involved. It is essential to identify the population that will be included in the corpus and to determine the geographic limits, the minimum length of residence and any linguistic requirements (for example, do participants need to be bilingual, and how will bilingualism be defined?). After identifying the population that will be included in the corpus, the next step is to complete the Institutional Review Board (IRB) approval process. Before submitting the IRB application, it is recommended to have a sit-down or phone/Zoom meeting with a member of the IRB to discuss the possibility of having the project exempt, or at least to minimize the annual review process, which can be time-consuming and complex. One way to achieve exempt status is by emphasizing the fact that in order to ensure privacy and confidentiality, all names and identifying information will be silenced in the audio file and left unidentified in the transcript, where they will be represented by the placeholder (for example, "XY"). While this may convince the IRB at some institutions to exempt the project, others will still require an annual review or continual review process. Additionally, all personnel (including students) working on the corpus will need to have current Collaborative Institutional Training Initiative (CITI) training to further ensure ethical compliance. When creating the actual website for the corpus, in addition to deidentifying the participant information, since the corpus consists of interviews with members of the community, we highly suggest making the website semi-open. Both CESA and CoBiVa require potential users to create an account and to request access by providing their Curriculum Vitae and briefly describing how they plan to use the corpus. This is to ensure people access the corpus for legitimate research and educational pur-

poses. Especially in the case of CESA and CoBiVa which house samples of stigmatized language varieties (bilingual speech), it is important to protect the speakers and the community from potential bad actors. For example, in both CESA and CoBiVa we have had applicants specifically state intentions like they "wish to document the errors bilinguals commit", and these have not been granted access.

2.2 Interview Protocol

Once the population is established and the IRB approves the project, the next phase is to begin collecting interviews and determining what will be collected from the participants. This subsection is dedicated to the interview protocol and will discuss how to train students as well as the process we use in CESA and CoBiVa. For more detailed descriptions and specific examples of how to train students to participate in the corpus, please visit the "Researcher Resources" section of the CoBiVa website (www.utrgv.edu/cobiva) under the "ReSourses" tab where a link is provided for "The CESA and CoBiVa Training Handbook" (Bessett, Carvalho, and Christoffersen 2021). On the same page, there is also a link to other materials, including all forms discussed in this section.

Before beginning the interview process, it is important to set up the specific protocol that will be used for all interviews that will make up the corpus. In addition to the interviews themselves, this includes what demographic information will be collected and when that information will be collected. In both CESA and CoBiVa, several sets of information are collected from both the person being interviewed and the person conducting the interview. First, there is a Bilingual Language Profile, modified from Birdsong, Gertken, and Amengual (2012), which asks participants about their linguistic background, language use, language proficiency, and language attitudes. For this part, it is important that the interviewer pays attention to the way the interviewee answers the questions to avoid miscalculations of percentages of self-reported language use, such as when the sums for the languages exceed 100%. A second form collects demographic information of the participant including year and place of birth, years in United States, education, languages spoken, ethnicity, religion, and other information that may be useful to future investigators. In addition to the information collected about the participant, demographic information of the interviewer is also included and reports the age, ethnicity, languages and proficiency, dialect, and education. Lastly, a fieldnotes document asks the interviewer to record when and where the interview took place and anything that stood out about the interview (i.e., formality, attitude the participant had, interruptions) or the language used.

After creating the documents that solicit the desired demographic information, the next step is to train the students on how to conduct a sociolinguistic interview. Several reference readings help future interviewers understand the concepts behind the interview process (Tagliamonte 2012; Becker 2013; Myerhoff, Sschleef, and Mackenzie 2015, for example). For the purpose of CESA and CoBiVa, students are taught semi-structured interview methods, loosely based on Labov (1972), in which they ask participants to sit for an interview for approximately 60 minutes to discuss themselves and their community. Students are supplied a list of potential questions to ask the participant, with categories that include describing their childhood, work, dates, family traditions, and life events. These questions are meant to be a guide in case the conversation slows, but students are encouraged to hold a natural conversation as much as possible. This effect is usually achieved when students choose to interview their friends or family, people that they know rather well. However, for some, especially those who are not from the community included in the corpus, the bank of questions is helpful.

While training students on what to expect during the interview process, we separate the procedure into three main parts: before arriving to the interview, when arriving at the interview, and ending the interview. Before arriving to the interview, students need to decide who they will interview and create a tentative plan (a selection of questions they can ask the participant during the interview). It is important here to emphasize that while questions about language use and language attitudes can be interesting, they should be asked only at the end of the interview so that participants do not modify the way they speak. Another issue to consider is the recording device students will use. For the purpose of CESA and CoBiVa, we have found that cellphones are more than adequate. We do have students record an hour of background noise and then attempt to transfer the file from their phone before they conduct the interview in order to ensure that a file of that size and type is easily transferrable. When arriving at the interview, we ask students to have a conversation with the participant until they feel comfortable, especially if they do not know the person they are interviewing well. This helps to keep the formality of the interview low, so as not to affect how the participant speaks. The consent form is given to the participant during this initial conversation. We advise students to avoid giving undue attention to the consenting process, and to explain briefly the content of the consent form. While we do not suggest that the students hide the consent process or have participants sign something they do not understand, it is important to avoid making the process overly formal. In fact, when possible, the students may ask participants to sign the consent form before they meet to conduct the interview to further lessen the formal frame for the sociolinguistic interview. After the interview, we ask students to thank their participant for their time and to fill out the Bilingual Language Profile

and the Demographic Information of the Participant forms mentioned above. Filling out these forms at the end of the interview allows for a more natural conversation and one in which the participant is not hyper-aware of the way in which they speak. However, it is important for the interviewer to ensure that the interviewee does not rush through these documents and that their answers are as accurate as possible. For more details on the specifics of the interview protocol for CESA and CoBiVa, please refer to Bessett, Carvalho, and Christoffersen (2021: 5–22).

2.3 Transcription Protocol

The next stage in the process is to transcribe the interviews once they have been collected. In order to maintain the corpus, a standard transcription protocol must be implemented to ensure consistency. First, establish the level of detail the transcriptions will follow. In CESA and CoBiVa, the transcriptions are word-for-word and include information for false starts and a few levels of pause lengths. For interviews of this length, a phonetic transcription would add a considerable amount of time. For common and salient phonological features that occur in CESA and CoBiVa, students are asked to note them in the "Fieldnotes" form (i.e., the fricative [ʃ] in place of the affricate [tʃ]). Next, a standard form of signaling turns in speech should be determined. For CESA, "E:" is used for the interviewer (entrevistador/a) and "P:" is used for the participant. In CoBiVa, the participant is represented with "<v PAR>" and the interviewer is represented with "<v INV>". A third consideration is the de-identifying of any identifying information, including personal names, schools the participant attended, and street names where the participant lived. In both CESA and CoBiVa, any identifying information is represented by "XY" in the transcript and a silence is added to the audio file using the free program Audacity or any similar program. The last standardization that needs to be determined is orthography and this includes, in addition to spelling, how to mark unintelligible speech, lengthened vowels, false starts, comments added by the transcriber, and how to identify borrowings and codeswitches. In CESA and CoBiVa, other language words/phrases are marked as "OL" for "other language" since it was often hard to be consistent with what counted as a borrowing versus a codeswitch. More details about how to transcribe other linguistic features are available in Bessett, Carvalho, and Christofferesen (2021).

Once there is a standardized protocol in place, the interviews can be transcribed. This is by far the most time-consuming aspect of putting the corpus together. For an inexperienced transcriber, it can take longer than 10 hours to transcribe a one-hour long interview. For best results, encourage transcribers to always wear headphones and to rest frequently (every 10 minutes or so). In order to

facilitate the process, a transcription software program should be used, for example Express Scribe, which offers a free version that works on both PC and Mac operating systems. Express Scribe allows the user to set up "hot keys" which help to pause, play, slow, rewind, and fast forward the audio. Several rounds of transcription help to ensure accuracy. In transcribing, one must be careful to accurately include repeats and false starts. They are not correcting or changing the transcription but instead writing exactly what was said in each interview. Inexperienced transcribers often misunderstand this and need considerable mentoring and review of transcripts. We usually have the course instructor and/or research assistants review student transcripts and provide notes/feedback. For a complete list of transcription conventions used in CESA and CoBiVa as well as the entire detailed transcription protocol, refer to Bessett, Carvalho, and Christoffersen (2021: 36–41). Since transcribing the interviews is a cumbersome task, we looked to technologically-aided transcription methods to improve the experience. Section 3 details our findings and offers what we consider to be the best current option.

3 Technologically-Aided Transcription Methods

Transcribing sociolinguistic data is a very time-consuming process. Conservative estimates suggest that one hour of audio may take at least 10 hours, and if the transcription includes time stamps, bilingual data, and precise formatting, it will take considerably longer. So, it is not surprising that scholars have for some time been examining how technologically-aided tools may speed transcription. This work has studied the usefulness of speech recognition software for conversation analysis (Moore 2015), interview data (Matheson 2007), qualitative analysis (Matheson 2007), and embodied transcription (Brooks 2010) and has examined different software including IBM's Atilla (Moore 2015) and Dragon Naturally Speaking (Evers 2011). However, when it comes to technologically-aided transcription tools for bilingual data (in this case, Spanish/English bilingual data), there are fewer options, and there is also less research on these options. This section outlines the research and testing we undertook to determine the best software for our purposes with CESA and CoBiVa.[1]

1 The work outlined in this section was funded by the National Endowment of the Humanities:

Christoffersen, Katherine; Bessett, Ryan; & Carvalho, Ana. 2021. Bilingual Voices in the U.S./Mexico Borderlands.

National Endowment of the Humanities Award. University of Texas Rio Grande Valley and University of Arizona. https://www.neh.gov.

3.1 Exploration of Technologically-Aided Transcription Methods

We began the investigation into transcription software by gathering a list of available software and comparing them in terms of the following criteria: a) ability to work with both Spanish and English data, b) sustainability, c) type, d) whether it allows for a time-aligned clickable transcript (such as WebVTT format), and e) possible concerns about data confidentiality. Figure 1 shows the results of this preliminary research.

Transcription Method	Spanish &English	Open-Source	Type	WebVTTformat, DataConcerns
Stream	Yes	Yes	Auto	
Youtube	Yes	Yes	Auto	Data concerns
Otranscribe	Yes	Yes	Revoicing	
Speech Notes	Yes	Yes	Revoicing	
Express Scribe	Yes	Yes	Manual	
FTW Transcriber	Yes	Yes	Manual	
Speech Notes	Yes	No	Auto	
Amazon Transcribe	Yes	No	Auto	Data concerns
Dragon	Yes	No	Revoicing	
Live Transcribe (Google)	Yes	No	Revoicing	
InqScribe	Yes	No	Manual	
F4 / F5	Yes	No	Manual	
Voci	Yes	No	Service	
Automatic Sync	Yes	No	Service	Yes
TranscribeWreally	Yes	No	Service	
Ligre	Yes	No	Service	
Trint	Yes	No	Service	
Happy Scribe	Yes	No	Service	
sonix.ai	Yes	No	Service	
Simon Says	Yes	No	Service	
Watson (speech recognition)	Yes	No	Service	
Panopto	English only	Yes	Auto	
Temi	English only	No	Service	
Rev	English only	No	Service	Yes
Otter.ai	English only	No	Service	
Descript	English only	No	Service	
scribblr.ai	English only	No	Service	

Figure 1: Initial exploration of transcription software.

As shown in Figure 1, 21 out of 27 transcription software had the ability to work with both Spanish and English. Only 6 of the software met the criterion of sustainability, or free access by researchers and institutions. Of these, only two had the capacity to auto-generate a transcript upon uploading the file (labelled 'auto'), while another two were only able to generate the transcript as the audio is played from external speakers or spoken into the microphone (labelled as 'revoicing'). Finally, we identified two software options where the transcriber types the audio as it is heard but with the inclusion of several features to aid the process, such as the ability to adjust the volume or speed or set up "hot keys" (or foot pedals) to repeat, rewind, fast forward, or insert timestamps. Manual transcription software has ex-

isted for quite a while, and this served as a relative control. We decided to select one of each type of transcription software to test. We ruled out YouTube due to concerns of privacy and data usage, chose SpeechNotes over Otranscribe due to some additional user-friendly features, and chose ExpressScribe due to our familiarity with this program. Having made these selections, we moved forward to preliminary trials of the following three software programs: 1) Microsoft Stream (autogenerated), 2) SpeechNotes (revoicing), and 3) ExpressScribe (manual).

Microsoft Stream is a platform within the Microsoft Suite, which many universities have access to. To use Stream, one needs to convert the audio file into a video file first, and then proceed to upload it. The program also allows for the choice between English and Spanish, but bilingual phenomena, such as code-switching and lexical borrowings, need to be revised manually after the transcription is produced. Importantly, the transcripts generated by Microsoft Stream had several extra lines, extraneous letters, numbers, and dashes, all of which problematized their readability. (At the end of this section we describe how we developed a two-step process in R Studio to remove these unnecessary lines. For more detailed instructions and the code, refer to the handbook at Bessett, Christoffersen, & Carvalho 2021). Although the Microsoft Stream software is not free to all, it is available free of cost to all faculty, students, and personnel at the universities with agreements with Microsoft.

The second program assessed was SpeechNotes, a speech recognition software application which is available as an app or in a browser. This program also allows for the selection of one language, but the user can change the selection during the transcription relatively easily. This program is meant to pick up voice and dictations and convert them into a transcript. However, during the initial exploration of transcription software options, some tests had success using the SpeechNotes software by playing audio from external speakers, which led us to continue with this option in trials. Although there is a premium version available, the free version was used for the trials and pilots in keeping with the goal of sustainability.

Lastly, we tested ExpressScribe, a transcription software method that has existed for many years. As such, it served as a relative control for the trials and pilot. All the principal investigators had prior experience with this method and used the manual transcription method as a baseline for comparison with the other methods. With ExpressScribe, the user uploads the audio and types the words as they are played. However, it offers several features to ease and speed transcription, including the ability to adjust the volume and speed of the audio. Additionally, the program allows the user to set up “hot keys” and/or foot pedals which are assigned different functions. As with SpeechNotes, the free version was tested, again in keeping with the project’s goal of sustainability.

3.2 Trialing Three Transcription Methods

In order to determine the functionality of the three selected transcription methods, we set up a preliminary trial among research assistants (RAs) working on both CESA and CoBiVa. This allowed us the opportunity to problem-solve issues with instructions and methods before later piloting the methods with a larger group. Six RAs participated in these trials, including two incoming MA students and one recently graduated MA student from the University of Texas Rio Grande Valley and three PhD students from the University of Arizona.

The trials were set up virtually and were divided into six modules held over the course of three weeks. The RAs first went through the training module with various short instructional videos prepared by the investigators and a 30-second transcript practice. During the next four subsequent modules, each RA transcribed a two-minute audio file using each transcription method (Stream, SpeechNotes, ExpressScribe). At the conclusion of each module of audio transcription, the RAs completed a short survey on the use of each method with the specific audio file. Two of these audio files were selected from CESA and two from CoBiVa. After transcribing all four audio files using the three methods, the RAs filled out a final survey where they detailed their experience with each of the methods. The investigators met with the RAs via Zoom three times during this process, once after the training module, once after the second module, and one final meeting and focus group after RAs had completed the final overall survey. In this meeting, the RAs discussed their overall experience with each of the three methods.

These preliminary trials measured speed, accuracy, and ease of use for the three transcription methods. Speed varied across students and trial number (as they got more familiar with the methods, the time was shorter), by order (as they got more familiar with the transcript, the time was shorter), and by audio file (which varied in sound quality, speed, and volume). When taken overall, the average time per audio file was greater for SpeechNotes, at 52.6 minutes, while the average time for Stream and ExpressScribe were comparable at 45.2 and 45.3 respectively. Next, accuracy was examined based on a comparison of the un-edited transcript from the two voice recognition software methods to a corrected version of the transcript for each audio file. In analyzing accuracy, we looked at the three main issues: missing words, incorrect words, and accents. SpeechNotes transcripts were only 5%-56% accurate, while auto-generated (Stream) was 44%-86% accurate. Averaged over all four interviews, SpeechNotes had an average accuracy rate of 26% while Stream had a 65% accuracy rate. Lastly, we assessed the RAs' perceptions of ease of use of the methods to compare their user friendless. This data was derived from the surveys and focus groups. Across the board, all RAs preferred ExpressScribe, probably due to the fact that four out of six had pre-

vious experience with it as well as the fact that the audio files were very short, only two-minutes long. On the other hand, all six RAs unanimously disliked SpeechNotes. During a focus group that gathered RAs to discuss their experiences, they described many problems with the program not capturing the sound, requiring them to repeat the audio into the speaker. Revoicing raised concerns as it does not capture the participant's real speech. Not only can this process result in an inaccurate transcription, it also implies taking the voice from the speakers whose language we attempt to document. The RAs described many problems with the program not capturing the sound, requiring them to repeat the audio into the speakers. Based on low ratings for speed, accuracy, and 'ease of use', SpeechNotes was discarded as a viable option. Stream and ExpressScribe were then further tested on a larger group of students.

3.3 Testing Stream as a Technologically-Aided Transcription Method

In order to determine if Stream was a better option for transcribing than ExpressScribe, larger trials were conducted using students enrolled in a course at the University of Arizona and students enrolled in a course at the University of Texas Rio Grande Valley. Altogether, 41 students were enrolled in the courses: 14 at UA and 28 at UTRGV. However, some students dropped the course and others didn't complete all portions of the pilot study, so they were excluded from the analysis. Thus, the reported results below are from 13 UA undergraduate students along with 14 undergraduate and 6 graduate students at UTRGV.

Each student conducted an hour-long sociolinguistic interview on Zoom due to the Covid-19 pandemic, including the collection of accompanying demographic and linguistic information from the community member. Then, students transcribed 10 minutes of the interview each week over the course of six weeks. During the first three weeks, students used ExpressScribe, and during the last three weeks, students revised the auto-generated Microsoft Stream transcript. All students carefully tracked their time for the transcription and answered a final overall survey on the transcription methods.

Overall, students reported that 10 minutes of transcription took them approximately three hours and 24 minutes for the manual (ExpressScribe) and two hours for the auto-generated transcript revision (Stream). Figure 2 shows that while this speed improved over time, Stream still proved to be a faster method (83.6 minutes) after students became faster with ExpressScribe (130.1 minutes). This and other quantitative results are descriptive and do not attempt to draw generalizations beyond this sample.

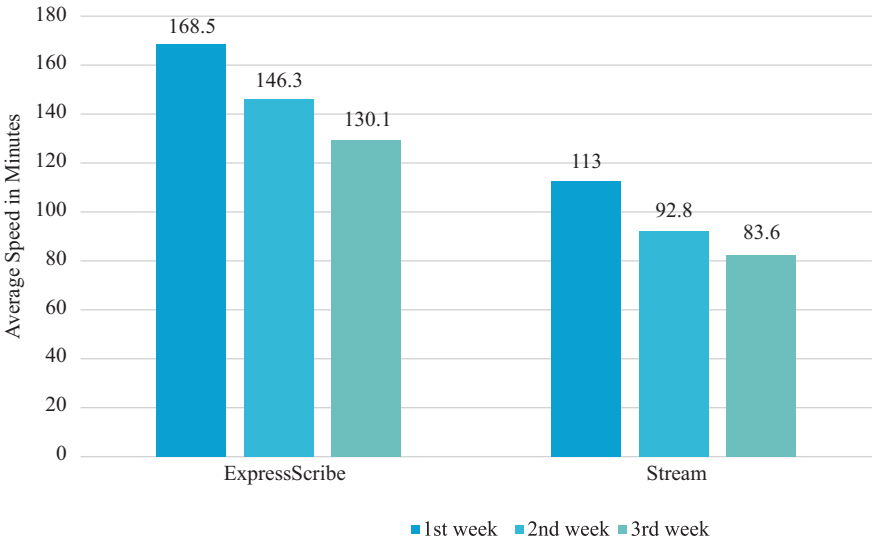

Figure 2: Average speed over time by transcription method (Express Scribe vs Stream) among UTRGV and UA students.

In the survey, students rated each method on a four-point scale for six different characteristics: ease of use, speed, intuitiveness, accuracy, bilingual data, and overall experience. For ease of use, the results were a bit more nuanced. Surprisingly, students rated ExpressScribe more favorably for most of these categories, except for intuitiveness and bilingual data, see Figure 3. However, note that the actual data on speed (above) reveals that Stream is faster; so, the students' perceptions in the survey do not align with the analyzed time tracking data.

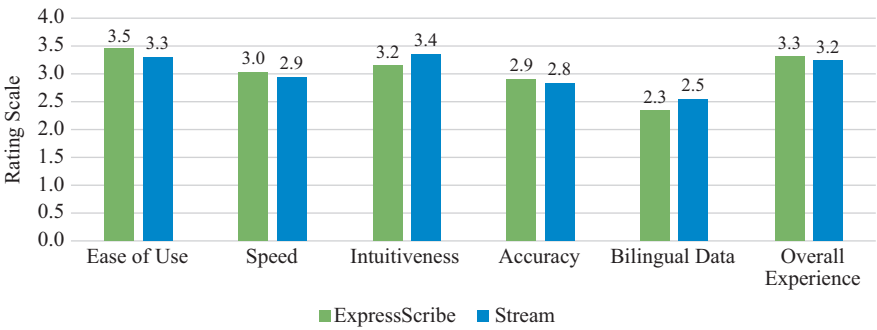

Figure 3: Comparison of method (ExpressScribe vs Stream) by characteristics (ease of use, speed, intuitiveness, accuracy, bilingual data, and overall experience).

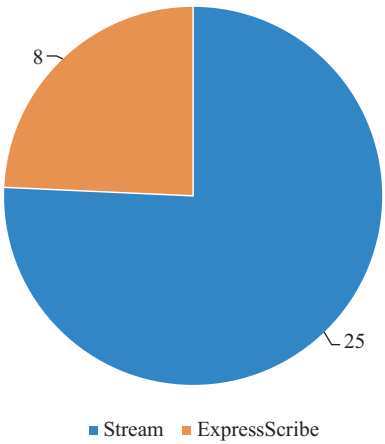

Figure 4: Preferred transcription method by students.

However, when asked which method they preferred, 25 out of 33 students (75.8%) preferred Stream (auto-generated) over ExpressScribe (manual), see Figure 4.

Students who preferred ExpressScribe said that it was fun, that they liked having control over the formatting and transcription. They also stated that it was frustrating to edit instances of code-switching, which represents a problem for the language choice required by Stream. Students who preferred Stream appreciated that it was quicker to edit transcripts. Lastly, five students mentioned that it was helpful to combine both methods and use ExpressScribe as a tool to help revise the autogenerated Stream transcript.

3.4 Final Considerations for Stream

It is clear that Microsoft Stream can be helpful as a first step in creating a transcription, especially in shortening the time it takes to complete a first draft. Nevertheless, the transcriptions that come from Stream still require the time-intensive steps of anonymizing the transcripts and audio files as well as additional rounds of checks for formatting, spelling, grammar, accents, and switches from one language to another. Additionally, as mentioned previously, the transcripts generated by Stream have several extra lines, extraneous letters, numbers, and dashes, all of which problematize their readability and need to be removed/revised. In order to streamline these revisions, the team worked with a consultant to develop a two-step

process using R Studio to clean up the auto-generated Stream transcripts.[2] For detailed instructions and code, refer to Bessett, Carvalho, and Christoffersen (2021: 42–52) and for a copy of the materials needed to run the two-step process, refer to the "CoBiVa Handbook Resources" link on the "Researchers" tab under "Resources" on the CoBiVa website (www.utrgv.edu/cobiva). Having explored all of the necessary elements to creating a semi-open source sociolinguistic corpus, Section 4 explores how this corpus can be used as a tool for social justice.

4 Corpora as a Tool for Social Justice

While creating student-based sociolinguistic corpora has direct applications for research on the speech of the community in which the corpus is based, the existence and use of these corpora also have the potential to be implemented as tools for social justice. Embodying the principles of student and community engagement (Labov 1982; Schilling 2013; Wolfram 2013), the corpus provides the ability for students to work with the Spanish (or language in general) of their community and home for a university project which elevates the language and gives them pride in their variety. Community engagement activities have been proven to better academic outcomes of the students who participate in them (Schulzentenberg 2020; Scales et al. 2006; Maruyama, Furco, and Song 2018) and to increase overall the participation of said students on campus (Patton et al. 2016). Students in both the CESA and CoBiVa corpus have demonstrated positive attitudes towards their experience working on the projects. Students at the University of Texas Rio Grande Valley expressed that the project allowed them to not only gain research experience and skills but to re-evaluate and attribute prestige to local bilingual varieties as can be seen in examples (1) and (2) from Christoffersen, Villanueva, and Bessett (*In Press*):

(1) . . . transcribing the interview allowed me to appreciate and value the interview even more.
(2) I would love to further work with projects like CoBiVa because I believe that field experiences are very important for students to gain more knowledge and confidence by engaging with our community . . .

2 This section of work was funded by the National Endowment of the Humanities and completed by Jessica Draper:

Draper, Jessica; Christoffersen, Katherine; Bessett, Ryan; & Carvalho, Ana. 2021. Bilingual Voices in the U.S./Mexico Borderlands. National Endowment of the Humanities Award. University of Texas Rio Grande Valley and University of Arizona. https://www.neh.gov.

Bringing the corpus into the classroom also provides a natural opportunity to bring bilingual varieties into the classroom through translanguaging pedagogies. In a classroom-based study, Christoffersen & Regalado (2021) document how the use of a translanguaging pedagogy in the corpus building classroom encouraged students to use more translanguaging practices such as code-switching in both spoken and written modalities, breaking down the tradition of English-only in higher education. In one reflection paper, a student wrote the following:

(3) Otra cosa que me gusto y me pareció muy peculiar fue que la clase fue bilingüe, eso es algo que nunca me había tocado ver en ninguna de mis otras clases antes, creo que eso ayudo mucho a la dinámica de la clase . . . fue algo único que creo que de alguna manera termina por ser un formato algo futurístico.

Translation: Another thing that I liked and seemed very unique to me was that the class was bilingual, this is something that I had never seen in any of my other classes before, I think this helped the class dynamic a lot . . . it was something unique and I think that in some way it ended up being an advanced format. (p. 65).

Likewise, students from the University of Arizona had positive experiences while working with CESA that improved their attitudes towards the local variety of Spanish, especially the bilingual phenomena present, as seen in examples (3) and (4) from Bessett, Carvalho, and Kern (2016):

(4) Part of why our discussion was so enlightening was the fact that it completely changed my perspective and attitude toward code-switching. It made so much more sense to me afterward, and I actually embrace it as something cultural that both my family and my region (border region) can take credit for as a positive contribution and cultural identifier in our society.

(5) Something that I definitely came away with from this class was a deeper appreciation of the consistency of grammatical rules in linguistic variation, even when those rules may clash with the form of the language that formal prescriptivists say is correct.

The experience students have in conducting sociolinguistic interviews with community members improves their understanding of the language they speak and creates new appreciation for their community.

To further the connections between the community and the university, using the corpus in the language classroom also raises the prestige of the local dialect. In addition, using the corpus in a heritage language classroom, raises the sociolinguistic awareness (Carreira 2010) of the student and allows an expansion of the students' linguistic repertoires instead of the replacement of them (Valdés 1981; Potowski 2005). Furthermore, bringing the corpus into the classroom teaches students to value different language varieties in different communicative contexts (Zentella 1997; Potowski 2005). These actions promote linguistic pride in the students (Martínez 2003) and encourage students to use their language outside of the

classroom. Possibly one of the best formal examples of bringing corpora into the classroom can be found in Potowski and Shin (2019) in which the authors create grammar explanations by presenting data from Spanish corpora, including corpora of varieties of Spanish spoken in the United States. Another notable example is the Spanish in Texas (SpinTX) Corpus video archive (Bullock & Toribio 2013) which provides a searchable database of grammatical points, language features, topics, and lesson plan/teaching ideas based on excerpts from the corpus. Taking a step in this direction, CoBiVa has a 'Language Ideologies' (DeAnda & Christoffersen 2022) module available in English and Spanish on the website under Teacher Resources. The module provides background information on the topic of language ideologies as well as example excerpts, explanations, discussion questions, and activities. The team plans to create additional modules and work on tagging the corpus for relevant examples, topics, and grammatical features.[3]

5 Conclusion

In this chapter we provided the essential steps to creating a student-based semi-open sociolinguistic corpus. In section 2 we outlined the process of building a corpus which included determining the potential speaker population to be interviewed, creating the materials for collecting demographic information, developing an interview protocol, training students to conduct interviews, and establishing a transcription protocol. In section 3 we presented the results of our work in attempting to streamline the transcription process, and based on these results, suggested using Microsoft Stream as a tool for creating the initial transcription. For more detailed information and to access a copy of all of the documents discussed in this chapter, please visit the CoBiVa website (www.utrgv.edu/cobiva). Lastly, in section 4 we described how creating a corpus with students can be a tool for social justice by benefiting the students and the community in which they live, in addition to providing linguistic data for future research. Thus, building a student-based corpus allows one to complete the full cycle of the sociolinguistic enterprise: data collection, data analysis, student engagement, and classroom practices targeted at raising sociolinguistic awareness.

3 This work will be funded by the National Endowment of the Humanities:

Christoffersen, Katherine; Bessett, Ryan; & Carvalho, Ana. 2023. Bilingual Voices in the U.S./Mexico Borderlands Phase 2: Preserving, expanding, and elaborating sociolinguistic collections.

National Endowment of the Humanities Award. University of Texas Rio Grande Valley and University of Arizona. https://www.neh.gov.

References

Becker, Kara. 2013. "The sociolinguistic interview". In *Data collection in sociolinguistics*: *Methods and applications*, edited by Christine Mallinson, Becky Childs, and Gerard Van Herk, 91–100. New York: Routledge.

Bessett, Ryan M., Ana M. Carvalho, and Katherine Christoffersen. 2021. *The CESA and CoBiVa Training Handbook*. (www.utrgv.edu/cobiva/resources/researchers/index.htm).

Bessett, Ryan M., Ana M. Carvalho, and Joseph Kern. 2016. "The full cycle of the sociolinguistic enterprise: Corpus building, student engagement, and critical language pedagogy." Poster to be presented at New Ways of Analyzing Variation (NWAV) 45, Simon Fraser University, Vancouver, BC, Canada, November 3–6.

Birdsong, David, Libby M. Gertken, and Mark Amengual. 2012. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. COERLL, University of Texas at Austin. https://sites.la.utexas.edu/bilingual.

Brooks, Christine. 2010. "Embodied Transcription: A Creative Method for Using Voice-Recognition Software." *Qualitative Report* 15(5):1227–1241.

Bullock, Barbara and Almeida Jacqueline Toribio. 2013. The Spanish in Texas Corpus Project COERLL, The University of Texas at Austin. http://www.spanishintexas.org.

Carreira, María. 2010. "Validating and promoting Spanish in the U.S.: Lessons from linguistic science." *Bilingual Research Journal* 24(4):423–442.

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

Christoffersen, Katherine. 2019. Linguistic terrorism in the borderlands: Language ideologies in the narratives of young adults in the Rio Grande Valley. *International Multilingual Research Journal*, 13(3):137–151.

Christoffersen, Katherine and Ryan M. Bessett. 2019. Corpus Bilingüe del Valle (CoBiVa). University of Texas Rio Grande Valley. www.utrgv.edu/cobiva

Christoffersen, Katherine, Ryan M. Bessett, and Ana M. Carvalho. 2021. "Bilingual Voices in the U.S./Mexico Borderlands." National Endowment of the Humanities Award. University of Texas Rio Grande Valley and University of Arizona. https://www.neh.gov

Christoffersen, Katherine, Ana Carvalho and Ryan Bessett. 2023. Bilingual voices in the U.S.-Mexico borderlands, Phase 2: Preserving, expanding, and elaborating sociolinguistic collections. National Endowment of the Humanities (NEH) Humanities Award. http://www.neh.gov

Christoffersen, Katherine, Aubrey Villanueva, and Ryan M Bessett. *In Press*. "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*.

Christoffersen, Katherine and Kimberly Regalado 2021. "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* 23(1):23–71.

DeAnda, Carolina and Katherine Christoffersen. 2022. Language Ideologies. https://www.softchalkcloud.com/lesson/serve/hHa6GBrJcNl2E3/html [Online teaching module integrating audio clips and transcript from CoBiVa]

Draper, Jessica, Katherine Christoffersen, Ryan M. Bessett, and Ana M. Carvalho. 2021. "Bilingual Voices in the U.S./Mexico Borderlands." National Endowment of the Humanities Award. University of Texas Rio Grande Valley and University of Arizona. https://www.neh.gov

Evers, Jeanine C. 2011. “From the past into the future. How technological developments change our ways of data collection, transcription and analysis.” *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 12(1). https://doi.org/10.17169/fqs-12.1.1636.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1982. “Building on empirical foundations”. In *Perspectives on Historical Linguistics*, edited by Winifred P. Lehmann and Yakov Malkiel, 17–92. Amsterdam/Phila: John Benjamins.

Labov, William. 1984. “Field methods of the Project on Linguistic Change and Variation.” In *Language in Use*, edited by John Baugh and Jel Sherzer, 84–112. Englewood Cliffs: Prentice Hall.

Leeman, J. (2018). “Critical language awareness in SHL: Challenging the linguistic subordination of US Latin@s.” In *Handbook of Spanish as a Minority/Heritage Language*, edited by Kim Potowski, 345–358. New York: Routledge.

Mallinson, Christine. 2013. “Sharing data and findings.” In *Data Collection in Sociolinguistics*: *Methods and Applications*, edited by Christine Mallinson, Becky Childs and Gerard Van Herk, 253–257. NY: Routledge.

Maruyama, Geoffrey, Andrew Furco, and Wei Song. 2018. “Enhancing underrepresented students’ success through participation in community engagement.” In *Educating for citizenship and social justice*, edited by Tania D. Mitchell and Krista M. Soria, 221–235. Palgrave Macmillan, Cham.

Martínez, Glenn. 2003. “Classroom based dialect awareness in heritage language instruction: A critical applied linguistic approach.” *Heritage Language Journal*, 1(1): 44–57.

Matheson, Jennifer. L. 2007. “The voice transcription technique: Use of voice recognition software to transcribe digital interview data in qualitative research.” *Qualitative Report*, *12*(4):547–560.

Moore, Robert. J. 2015. “Automated transcription and conversation analysis.” *Research on Language and Social Interaction*, *48*(3):253–270.

Myerhoff, Miriam, Erik Schleef and Laurel MacKenzie. 2015. *Doing sociolinguistics: A practical guide to data collection and analysis*. New York: Routledge.

Nagy, Nomi and Paulina Lyskawa. 2016. “Moving forward with multilingual transcription.” Workshop presented at Linguistics Society of America, Washington, DC.

National Science Foundation (NSF). http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816&org=NSF

Patton, Lori D., Kristen A. Renn, Florence M. Guido, Stephen John Quaye, Deanna S. Forney and Nancy J. Evans. 2016. *Student development in college: Theory, research, and practice* (3rd ed.). San Francisco, CA: Jossey-Bass.

Potowski, Kim. 2005. *Fundamentos de la enseñanza del español a hispanohablantes en los EE.UU*. Madrid: Arco/Libros.

Potowski, Kim and Naomi Shin. 2019. *Gramática española: Variación social*. New York: Routledge.

Scales, Peter C., Eugene C. Roehlkepartain, Marybeth Neal, James C. Kielsmeier, and Peter L. Benson. 2006. “Reducing academic achievement gaps: The role of community service and service-learning.” *Journal of Experiential Education*, *29*(1):38–60.

Schilling, Natalie. 2013. *Sociolinguistic fieldwork*. Cambridge, UK: Cambridge University Press.

Schulzetenberg, Anthony J., Yu-Chi Wang, Ashley Hufnagle, Krista M. Soria, Geoffrey Maruyama, and Jason Johnson. 2020. “Improving Outcomes of Underrepresented College Students Through Community-Engaged Employment.” *International Journal of Research on Service-Learning and Community Engagement*, *8*(1):18719.

Tagliamonte, Sali. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Oxford: Wiley-Blackwell.

Valdés, Guadalupe. 1981. “Pedagogical Implications of Teaching Spanish to the Spanish-Speaking in the United States”. In *Teaching Spanish to the Hispanic Bilingual: Issues, Aims, and Methods*, edited by G. Valdés, A. Lozano, and R. García-Moya, 3–20. New York: Teachers College Press.

Wolfram, Walt. 2013. “Community commitment and social responsibility.” In *Handbook of Language Variation and Change 2nd Edition*, edited by J.K. Chambers and Natalie Schilling, 557–576. Malden/Cambridge: Wiley/Blackwell.

Zentella, Ana C. 1997. *Growing up bilingual: Puerto Rican children in New York*. Wiley-Blackwell.