Californian Spanish: Building a bilingual linguistic Corpus

Julian Vargo[1] & Justin Davidson, Ph.D.[1]

Department of Spanish & Portuguese[1]

University of California, Berkeley

August 9, 2024

A report on the Multilingual Hispanic Speech in California (MuHSiC) corpus for the UC Berkeley Latinx Research Center's 2023-2024 Faculty Mentored Undergraduate Research Fellowship

1. Spanish-English Bilingualism in California

In recent decades in the United States, bilingualism has faced a significant amount of unjust stigmatization. Just a few generations ago, many people believed that exposing children to more than one language could confuse them, resulting in delays or disorders in their language development. Such bias against bilingualism has had detrimental effects on immigrant families, on education policy, and more broadly on the linguistic and cultural diversity that enriches our society. For instance, in the early 20th century, English-based psychological tests were administered to all U.S. immigrants as a yardstick to "keep out those who are manifestly undesirable" for American society (Sweeney 1922). These specifically English-based psychological tests engendered mass deportations of immigrants who were deemed too cognitively deficient to remain in the country, and non-English-speaking immigrants have since disproportionately been incorrectly relegated to special education classes, on a similar and noncredible assumption that lack of English proficiency equates to mental defect (Romaine 2000). Presently, accented or non-native speakers of English face bias with potential harmful consequences related to employment, housing, racial discrimination, and judicial prejudice (Baugh 2005). Ironically, in California, one of the most linguistically and culturally diverse states in the country, Proposition 227 required that all public schools use only English as the language of instruction, though fortunately this was repealed recently in November, 2016, by the California Non-English Languages Allowed in Public Education Act (Proposition 58).

Modern psycholinguistic research shows a plethora of cognitive benefits derived from bilingualism, nullifying the aforementioned arguments and oppressive policies against non-English speech and underscoring the unjust treatment of bilingual or accented speech (Farhan 2019). Neurolinguistic research using brain imaging such as fMRI has shown that Spanish-English bilinguals have similar linguistic reaction speeds and increased activation in certain regions of the brain when performing linguistic tasks compared to English monolingual speakers (Kovelman et al. 2008).

To further demonstrate the importance of bilingualism in the United States, it is necessary to establish that the U.S. is home to more Spanish speakers than countries such as Spain, Colombia, or Argentina (Instituto Cervantes 2015). The United States now more than ever constitutes an ideal research site for investigations concerning language contact and bilingualism, hosting its own diverse and distinctive dialectal features just as other Hispanophone countries

display. Nonetheless, deep-rooted English monolingualism in American mainstream culture alongside recent anti-immigrant politics, have made it very difficult for children raised in a home with a language other than English to maintain their home language (Fuller 2012). For example, in California, the state with the greatest number of Spanish-speakers (U.S. Census Bureau 2020), foundational work by Silva-Corvalán (1994) regarding the Spanish of Los Angeles revealed abrupt generational practices of abandoning Spanish in favor of English. The youngest generations of Spanish speakers were subsequently characterized by other Spanish speakers as improficient speakers of a simplified and substandard variety of Spanish influenced by English (Silva-Corvalán 1994). In 2010, the *Academia Norteamericana de la Lengua Española* (North American Academy of the Spanish Language) released a book titled *Hablando bien se entiende la gente* (Speaking well, people understand each other), a 188-page reference manual targeted for North American speakers of Spanish to encourage them to choose to speak English and Spanish "well" instead of speaking both "poorly" by "mixing them". On the first page, the manual compares United States Spanish to Tarzan's ape-like speech, painting U.S. Spanish as a primitive, unrefined, and caveman-like aberration of "proper" Spanish. These kinds of characterizations of U.S. Spanish, reflecting Spanglish as an object of social ridicule and overt stereotyping, are discriminatory and often bolster linguistic insecurities in the form of embarrassment, disdain, and even the outright abandonment of the Spanish language on the part of many U.S. Spanish speakers.

The complex and prejudiced history of Spanish-English bilingualism begs the question, what do present-day language trends look like in California? Do Californian bilinguals consider their Spanish to be substandard or impoverished, and how are such ideologies expressed through their use of Spanish and English? Moreover, what kind of specific phonetic, phonological, morphosyntactic, or other linguistic patterns can be uncovered from a wide-reaching study across the entire state regarding Spanish-English bilingualism? Answering such questions will be incredibly consequential for the status of bilingualism in the United States for several years to come. By rigorously studying California Spanish and its diverse community of speakers, the present project, which aims to create a large online library of multilingual speech for academic research, marks a crucial step toward legitimizing United States Spanish and deconstructing previous notions of linguistic prejudice against multilingual speakers across a myriad of academic and non-academic domains.

2. Building a Bilingual Linguistic Corpus

In modern linguistics, the study of accents, grammatical structure, and language attitudes often requires a large database of audio, writing, or other linguistic notes which are compiled in an organized fashion, called a linguistic corpus. The vast majority of linguistic corpora are monolingual, meaning that all of the audio or writing was collected in a single language, as it was common practice in the field to think of monolingual speakers as "ideal" for scientific research, a pattern which was arbitrarily set forth by Chomsky (1965). This means there is a dearth of linguistic corpora featuring more than one language, making the study of bilingual speech particularly challenging, subsequently resulting in the creation of the present corpus.

This project aims to create a corpus of Multilingual Hispanic Speech in California (MuHSiC) by documenting speakers from both Spanish-speaking and English-speaking households across California. By comparing the linguistic features that comprise the speech of Spanish-English bilinguals from diverse backgrounds, this corpus offers a unique source of insight into the cognitive and linguistic abilities of a largely understudied bilingual population that has traditionally been ignored in language research despite the continuous increase of Spanish speakers in California and the US. Moreover, this corpus aims to be the largest and most comprehensive dataset of multilingual sociolinguistic interviews and naturalistic conversations among speakers of diverse social profiles and regional origins in California, obtaining approximately 14,000 minutes of speech from 200 speakers in both Spanish and English, equating to roughly 10 days of high quality audio for scientific analysis.

The corpus consists of 6 principal parts:

- **35 minute sociolinguistic interview of each participant in Spanish**
- **35 minute sociolinguistic interview of each participant in English**
- **Bilingual Language Profile of each participant**: including participants self-described proficiency levels in Spanish & English, educational background and experiences speaking Spanish & English
- **Sociodemographic Questionnaire of each participant**: including information on participants' hometowns, parents' dialects, age, gender, and other relevant demographic information which may affect participant speech patterns
- **Interviewer Field Notes for each participant**: marking any notable anomalies, interruptions or interesting speech patterns which occurred during the interview
- **Interviewer Background:** noting demographic and linguistic information about the interviewer which can potentially alter the speech patterns throughout the recordings.

Constructing a comprehensive, carefully planned linguistic corpus with high quality audio requires a substantial amount of planning, training, data collection, and audio processing. Throughout the 2023-2024 academic year, the first phase of this corpus was completed, which consisted of planning out the design of the corpus, recruiting undergraduate researchers to assist in data collection, training researchers on how to properly use recording equipment, training researchers on how to conduct sociolinguistic interviews, manually checking each interview to ensure high audio quality, assisting undergraduate researchers with technical or logistical problems, and processing the audio and paperwork. During the Fall 2023 and Spring 2024 semesters, 32 undergraduate researchers were successfully trained by Julian Vargo and Dr. Justin Davidson, meeting throughout the academic year with students to assist them with best practices to conduct effective sociolinguistic interviews (eg. keeping the participants at ease, asking participants engaging questions, strategies to elicit naturalistic and casual speech) and to obtain high quality recordings.

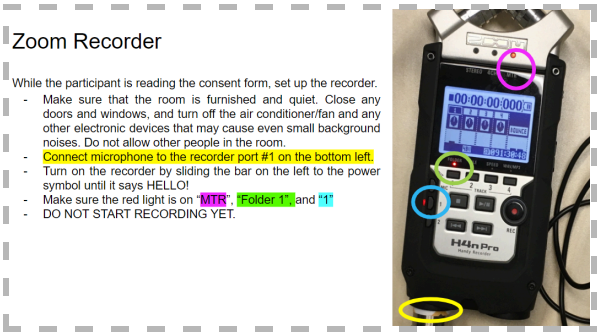

Figure 1. Example slide from student training meetings on using recording equipment

Throughout the academic year, undergraduate researchers would check-out and check-in recording equipment and required paperwork, which was managed and organized jointly by the authors. After each interview was conducted, interviews would be checked to make sure that the sociolinguistic interviews were consistently filled with lots of speech and minimal background noise, as background noise can decrease the accuracy of linguistic tools which rely on automatic detection of speech sounds.

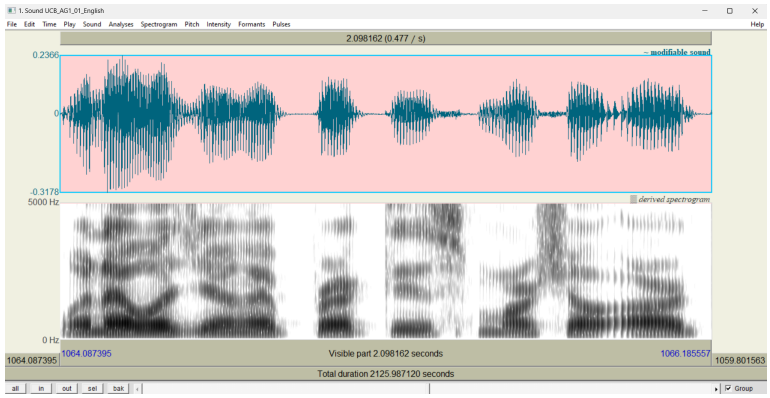

Figure 2. Spectrogram and waveform of two seconds of naturalistic speech from the MuHSiC sociolinguistic interviews, using the phonetic software Praat (Boersma & Weenink 2024)

The final major task for phase 1 of this project, which happened during the Summer 2024 term, consisted of processing the 233 hours of bilingual speech across California, alongside the digitization and sorting of 2,930 pages of metadata written for each interview, primarily done by Julian Vargo. To process the audio, Sonix, a bilingual transcription software, has begun being used to assist in the transcription of the first set of interviews. The 2,930 pages of paperwork consisting of the Bilingual Language Profile, Sociodemographic Questionnaire, Interviewer Field Notes, and Interviewer Background, were also manually scanned and digitized during Summer 2024 and are currently being processed into .csv and .json files in preparation for the public release of the corpus, to ensure that all relevant information about the interview is easily accessible for linguistic, demographic, or sociological research.

Speaker	JV1
English 35 Minute Interview .wav	
Spanish 35 Minute Interview .wav	
Age	20
Hometown	San Marcos, CA
First Language	English
Second Language	Spanish
English Comfort Level	10 / 10
Spanish Comfort Level	9 / 10

Table 1. Abridged example of sorted metadata

3. Future Directions

For such a large project, which has required the authors to oversee data collection and management from a combined 230 researchers and participants, the planned organization and analysis of the materials will continue to be a significant undertaking. Currently, the MuHSiC Corpus project contains an incredibly large quantity of raw audio and unsorted metadata, which are quite difficult to work with from an academic perspective. The principal goal of the second phase of this project is to finish data processing, improve accessibility, and release all of the data onto an original MuHSiC website for public usage.

For the 2024-2025 academic year, the Sonix audio transcriptions will continue to take place, and the 233 hours worth of transcriptions are planned to be manually corrected by trained bilingual undergraduate researchers. Moreover, the audio must carefully be searched for any sensitive personal information which must be redacted from the audio and paperwork files before the data can be ethically published. After the interviews have finished the transcription process, the data must undergo forced-alignment, a crucial step to make the corpus as accessible as possible for other researchers to use, subsequently promoting the furthered study of Californian bilingual speech. Forced alignment is a process used in phonetic research which aligns every single speech sound with a timestamp. In Figure 1, which consists of roughly 2 seconds of audio, there are approximately 20 speech sounds. Each one of these sounds must be converted into a.TextGrid file so researchers can easily identify particular sounds of interest to describe the Californian Spanish-English bilingual accent. Lastly, the design of a user-friendly and accessible website to download and analyze such data is expected to occur during the upcoming academic year, upon which the corpus will be ready to use for the publication of linguistic, sociological, demographic, and other critical social science studies which will continue to improve the study status of Spanish-English bilingualism for decades to come.

References

Academia Norteamericana de la Lengua Española. (2010). *Hablando bien se entiende la gente [People are understood speaking well]*. Santillana USA Publishing Press.

Baugh, J. (2005). Linguistic profiling. *In S. Makoni, G. Smitherman, A. Ball, and A. Spears (eds.) Black linguistics: Language, society, and politics in Africa and the Americas*. pp. 155-168. Routledge Press.

Boersma, P. & Weenink, D. (2024). Praat: Doing Phonetics by Computer.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*. MIT Press.

Farhan, R. (2019). The Benefits of Bilingualism. *Learning to Teach Language Arts, Mathematics, Science, and Social Studies Through Research and Practice. 8*(1). https://openjournals.utoledo.edu/index.php/learningtoteach/article/view/289

Fuller, J. M. (2012). *Bilingual pre-teens: Competing ideologies and multiple identities in the US and Germany*. Routledge.

Instituto Cervantes. (2015). *El español: una lengua viva [Spanish: a living language]*. Departamento de Comunicación Digital del Instituto Cervantes.

Kovelman, I., Baker, S. A., & Petitto, L. (2008). Bilingual and monolingual brains compared: a functional magnetic resonance imaging investigation of syntactic processing and a possible "neural signature" of bilingualism. *Journal of Cognitive Neuroscience*. *20*(1), 153-169. doi:10.1162/jocn.2008.20.1.153

Romaine, S. (2000). *Language in Society: An Introduction to Sociolinguistics*. Oxford University Press.

Silva-Corvalán, C. (1994). *Language Contact and Change: Spanish in Los Angeles*. Clarendon Press.

Sweeney, A. (1922). Mental tests for immigrants. *The North American Review*. *215*(798). pp 600–612.

United States Census Bureau. (2020). ACS 2018 Demographic and Housing Estimates. https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/