The Corpus of Mexican American Language (COMAL)

Ziyun Chew

Abstract. Integrating the Mexican American (MA) historical context and the individuals language acquisition into the study of MA sociolinguistic performance enables us to further our understanding of variation within and across languages, identity construction, and multilingual linguistic behavior. The racialization and discrimination experienced by MAs has enduring effects on their ability to access, acquire, and ultimately use languages and styles. These processes give rise to historic, localized, and individual language ideologies among MAs, which in turn influence individuals' linguistic behavior in English and/or Spanish. Previous research in multilingual phonetics and sociolinguistics enables us to understand how multilingualism and sociolinguistic variation can both give rise to the variation seen in MA speech, but may not appropriately capture the range of MA linguistic behavior and its relationship with culture, experience, and identity. The Mexican American Socio/Historic/Linguistic (MASHL) Framework was developed In order to accordingly consider the multiplex factors during the data collection process and subsequent interpretation of MA speech, thus serving as a structure to develop a corpus. The Corpus Of Mexican American Language (COMAL) aims to provide researchers with viable linguistic data in both English and Spanish that documents multiplex MA identities, experiences, and linguistic behaviors, while reflexively considering how these may interact with conventional linguistic methodologies. COMAL currently consists of de-identified recordings, transcripts, and metadata of one-on-one interactions in English and Spanish between the author and 4 adult MAs living in El Paso, TX. Preliminary analyses of the data, including rudimentary English vowel formant and stylistic analysis, indicate its viability for linguistic research.

1. Introduction. While movement across North America has existed for centuries at least, the creation of Mexico and the United States as discrete countries has designated the movement of people from Mexico to the US as immigration by Mexicans. This immigration also constitutes a change in language environment despite neither country having an official language, as the de facto language in Mexico is Spanish, while the US's is English. The presence of Spanish in the US dates back from Spanish colonization, including several US varieties of Spanish such as New Mexican Spanish. As of 2022, the American Community Survey estimates that 78% of Americans of Mexican origin speak Spanish at home (U.S. Census Bureau 2022a). This preservation of Spanish through the centuries among Mexican Americans (MAs) has accordingly been attributed to ongoing immigration and transmission across generations (Jenkins 2018). Such long-lasting maintenance led to intense language contact between English and Spanish, which may have generated language contact features in both Mexican American populations and non-Hispanic populations (Fought 2002). Additionally, the use of both languages and resulting language contact has fostered the creation of and/or contribution of elements to now-enregistered varieties of English and Spanish associated with MAs (Ornstein-Galicia 1987; González 1988; Fought 2002). Such varieties have connections to urban, working-class youth, and have been memorialized by pachucx, cholx, and Chicanx art and music (Baker 1969; Mendoza-Denton 2008)

In addition to these processes, MAs have contended with historic discrimination and racialization that affects their language use to this day. According to Glenn (2015), anti-Mexican racism

in the US is a project to construct them as the "undesirable exogenous others," regardless of their nationality or origin. MAs have been and continue to be racialized through containment, erasure, terrorism, and removal (Glenn 2015), which have in part relied on creating an image of non-Anglophones with "disorderly" speech and conduct (Hill 1998). Such policies and ideologies have and continue to be the seed of intergenerational trauma. One of the most salient repercussions is the interruption of transmission of Spanish and Mexican heritage to younger generations, due to schoolteachers in the 1900s violently punishing MAs for speaking Spanish (Dowling 2014). Consequently, MAs in the modern age carry historic trauma regarding their racialization, which affect their ability to access, acquire, and use languages, and styles within languages (Hurtado & Gurin 1987; Velázquez 2009).

While it is certain that not all MAs experience all of these processes or their effects, all MAs have linguistic ideologies surrounding their racial identity that are historic, local, and/or individualized. These ideologies, varying access to language acquisition, and other elements of their identities and related ideologies combine to influence each MA's individual linguistic performance, decisions in identity construction, and consequently, stylistic choices (Fought 1999, 2010). Social, historic, and linguistic experiences are therefore crucial characteristics that must be taken into account when studying MA linguistic practices and performance.

This paper will explore linguistic research on MA populations and their limitations; a research framework to address said limitations; the development of a corpus of MA speech based on this framework; and the viability of the corpus.

2. Previous Research and Similar Corpora. Much linguistic research on MAs and other racialized multilingual populations in the US has come from the subfields of phonetics and phonology, language acquisition, sociolinguistics, linguistic anthropology. Though other aspects of language are equally as important to the study of MAs, this paper and corpus creation methodology focus on phonetic/phonological production. This section describes previous research contributions and limitations.

2.1. MULTILINGUAL PHONETICS/PHONOLOGY AND LANGUAGE ACQUISITION. Research into multilingual behavior at its core has often been considered to be an exploration of human language acquisition, as the questions and approaches to multilinguals rely on the assumption that the timing and manner in which a language was acquired affects its use (De Houwer 2009). This is largely upheld by the proliferation of the Critical Period Hypothesis (CPH) and related theories of acquisition in various forms and strengths, which generally assume a window of time in infancy during which the brain is best able to acquire language(s) and/or certain aspects of language(s) (Lenneberg 1967). The study of acquisition of multiple languages is accordingly concerned with relating types of language acquisition with linguistic behavior. This facilitated the creation of research paradigms seeking to distinguish between monolingual and multilingual acquisition and behavior, and to distinguish between different types of multilinguals according to their age of acquisition and behavior. Generally speaking, academics have described several generalized types of language acquisition based on the CPH that correlate with linguistic behavior to some degree: monolingual, simultaneous/from-birth multilingual (or bilingual first language), early (before adulthood, after infancy) L2, and adult L2 (De Houwer 2009). Note that early L2 and adult L2 acquisition implies previous monolingual language acquisition experience. It is also important to note that children of immigrants to the US are often called "heritage speakers" in the literature. This term broadly describes speakers who only acquire the language "corresponding

to their heritage" until they enter the US school system, afterwards becoming early L2 English acquirers (Benmamoun et al. 2013).

The literature on acquisition of the phonological systems of multiple languages seeks to explore phenomena including language change and contact, and rates of "accentedness" in comparison to monolinguals. Three models have been proposed to explain how multilinguals acquire multiple sound systems, while also accounting for variation due to age of acquisition and distinction from monolinguals. These are the Perceptual Assimilation Model (PAM) (Best et al. 1994), the Speech Learning Model (Revised) (SLM-r) (Flege & Bohn 2021), and the Native Language Magnet theory (NLM) (Kuhl et al. 2008). All three models postulate that the acquirer's perception of similarity of new L2 sounds to L1 sounds affect their ability or manner of acquisition of the new sound. The SLM-r in particular claims that bilinguals have phonetic systems that differ from monolinguals, regardless of their language acquisition history (Flege 2007). These results suggest that bilinguals may be the impetus for language contact and change, though it is important to further explore how the individual bilingual contributes to language change (Yao & Chang 2016).

However, prior research on multilingual acquisition of sound systems may not completely address the range of MA speech, and may in fact contribute to discriminatory ideologies surrounding said speech. The degree to which they are used to account for non-Standard production is limited due to the conflation of "native" acquisition with monolingualism and normative language use. Additionally, conventional experimental methodology may only include the most privileged population members, and may be uncritical of social/historical factors affecting production described above. Cheng et al. (2021) note several issues in this framing of multilingualism, as this perpetuates the idea that "there exists some threshold of language ability available to 'native speakers' that non-native speakers' cannot reach, as a result of age of language onset" (pg.3). This is particularly problematic for the inclusion of marginalized people such as MAs due to multiple factors.

Firstly, not all individuals have equal opportunities to access and acquire languages and styles. There is ample evidence indicating that monolingual children of color are less likely to acquire Standard formal registers, which impacts their ability to perform on standardized measures (Dudley-Marling & Lucas 2009). This is further complicated by the diversity of input and acquisition experience possible in multilingual communities/families of color. Multilinguals of color may then not be able to perform language in the same manner as a monolingual white person would. Furthermore, these multilinguals of color may not be considered viable candidates for studies seeking monolinguals or bilinguals due to language acquisition histories that do not correspond to nebulous ideas of first- or second-language acquisition. MA "Heritage speakers" of Spanish and/or indigenous languages exist, as do MAs who experience unbalanced exposure and acquisition of multiple languages, and many other types of acquisition experiences.

The type of speech data collected for conventional phonetic research and related corpora are usually produced in lab settings using materials "standardized" across several languages (Bradlow n.d.). However, lab speech may not be representative of an individual MA's repertoire, the variation in the group, or variation across different populations of multilinguals. Due to the whiteness of academia and trauma related to academia, the lab setting may become a White Public Space where people of color may feel compelled to edit their speech or refuse to speak a language other than English (Hill 1998). Some may even decline to enter due to their stigmatized language use, as they feel they are not a suitable subject for linguistic study (Cheng et al.

2021). In the case that a multilingual of color does participate in a study based on lab speech, their non-Standard variation and choices may not be seen as such, and instead viewed as a deficit in acquisition. Given that marginalized individuals are prone to dealing with such issues, their data might not be comparable to that of privileged populations. Consequently, theories produced by conventional phonetic research and related research/corpora may not be able to faithfully depict marginalized multilingual populations as they are likely underrepresented in the data. Such considerations should be made in future research, and can be enhanced with the addition of sociolinguistic and linguistic anthropological methods, though these methods must also be submitted to scrutiny.

2.2. SOCIOLINGUISTICS AND LINGUISTIC ANTHROPOLOGY. Historically, much sociolinguistic research on Latinx populations in the US has focused on identifying and linking phonetic variation in English with country-of-origin identity. Emphasis has been placed on variants that researchers associate with individual/community contact with Spanish, as well as those that "depart" from norms in white populations residing in or near the same areas. Influential works such as Homegirls (Mendoza-Denton 2008) and Chicano English in Context (Fought 2002) validated non-Standard language use by MAs by reporting the existence of MAs with "native acquisition and proficiency" in an ethnolect named Chicano English. The validation was particularly important as many of these "native" Chicano English speakers were multiply marginalized: they were working class, lived in heavily policed neighborhoods, and victimized by the school-to-prison pipeline due to their racialization. Not only did these speakers contradict dominant ideologies that viewed them as languageless (Rosa 2016), they were also prolific innovators of language in both English and Spanish. These research works also demonstrated the possibility of stylistic variation and use of persona by MAs in English. Although some speakers reported English monolingualism or used variants associated with Standard English, they were documented using variants previously only associated with bilingualism or contact with Spanish. Mendoza-Denton's ethnography particularly emphasized that this phenomenon was intimately tied to mutable social factors such as identity and social networks.

The validation of this population's speech was critical in furthering sociolinguistic understanding of multiply marginalized populations. Nonetheless, work still remains in these frameworks to address the privileging of "native" language users and their English. Furthermore, the use of ethnolectal models to describe linguistic variation has been critiqued in relation to other marginalized racial/ethnic populations, as it has been demonstrated to obscure important dimensions of variation and identity (King 2020, 2021).

Unlike much research on phonetic systems of multilinguals, Mendoza-Denton and Fought reframed the definition of "native speaker" to include those who acquired non-Standard varieties of languages during the Critical Period. While this tactic was intended to be inclusive of multiply marginalized populations, it can inadvertently reinforce "native speakers" as a standard of acquisition. Participants who were "late-acquirers" of English were excluded from phonetic analysis or not included as an object of study, as they were considered to be categorically unlike the "native speakers." This was upheld even if "late-acquirers" exhibited similar sociophonetic production and personas as the "native speakers" during the researchers' data collection. This is difficult to reconcile given that MAs have a wide range of experiences on multiple dimensions that defy categorization, and the strength of the Critical Period hypothesis is hotly debated (Schouten 2009). When "late-acquirers" continue to be excluded from sociolinguistic analyses, their sociolinguistic

variation may be erased and attributed to language acquisition errors despite participating in or even innovating linguistic change.

It is also important to note that descriptions of Chicano English in Fought and Mendoza-Denton's work is based on that of Californian MAs in the '90s and '00s, and reflects a specific subculture or persona/style associated with the era and region: the cholx. This style, which includes elements such as loose chino pants, flannels, low-rider cars, and Chicano-style tattoos, has come to signify a "response to racial stigmatization, class subordination, and criminalization" (Aldama et al. 2012). While the cholx has been emblematic for many MAs, particularly those who identify as Chicanx, this is only one of many personas available for use in the MA stylistic repertoire (Rosas 2010). Much like the conflation of African American Vernacular English with particular Black American personas (King 2020), the conflation of the cholx persona and Chicano English with the MA experience is problematic in that it can prevent researchers from recognizing an MA's use of any other stylistic variation or persona to index their multiplex and localized identities.

By considering the ways in which conventional paradigms of multi/monolingualism has also influenced sociolinguistic and linguistic anthropological research, we can continue to develop more nuanced research frameworks to address multilingual populations of interest.

2.3. CORPUS LINGUISTICS. The field of linguistic corpora continues to grow, and with it comes a wide selection of collection techniques, data, and participants available for research. Data contained in corpora may be textual as in the case of text corpora for computational linguistics and Natural Language Processing, or recordings of speech along with transcriptions and annotations. Corpora data are created through multiple methods, including the compilation of found data (such as in CHILDES (MacWhinney 2014)), and the collection of data through consented solicitation of data. Solicited data may be created through experimental methods as seen in the ALLSSTAR corpus (Bradlow n.d.), or through the recording of unscripted speech such as interviews or conversations.

There is a broad range of types of data contained in speech corpora about multilinguals and their languages, as there are many facets to these objects of study. These repositories feature one or more registers of speech in one or multiple of the multilingual's languages, and document social/demographic factors that has previously been associated with linguistic behavior, e.g. age, age of acquisition, and places lived. The most common registers of speech documented are unscripted conversational, and read speech. Interlocutor types differ between corpora as these can include interviewers or other participants, which ultimately influence the type of conversations held during recordings. Linguistic corpora are important for the landscape of linguistics as a field, as they enable higher access to large collections of data that are prohibitively expensive and labor-intensive for individual researchers to create, the possibility of observing emergent behavior through corpus linguistics, and the possibility of replicability through public access of the dataset.

While the variety of corpora allows for a multitude of uses and analyses, a single corpus may not address an end user's multiple needs due to extensive curation of the data or uniqueness of found data. For example, although SpinTX (Bullock & Toribio 2013) is comprised of sociolinguistic interviews with Spanish speakers in Texas, many of whom are bilingual English/Spanish speakers, it does not capture bilinguals' speech in English or other registers of speech in Spanish. Quirks of multilingual corpora include the issue of "multilingual" modes throughout the data due

to the endeavor to capture speech in a multilingual's multiple languages during a small window of time. This may concern researchers who are interested in how "monolingual" versus "multilingual" modes affect a multilingual's speech production (Sancier & Fowler 1997). However, researchers are beginning to make the effort to address and document multiple issues and factors affecting multilingual behavior within corpora, such as Nagy (2011)'s corpus focused on language contact and variation.

Although the motivations and goals of linguistic corpora differ from experimental and ethnographic research, the methods and underlying philosophy of science are similar and therefore face similar issues regarding faithful or constructive representations of participants or the population of study. Corpora using experimental methods in particular may deal with the problems seen in Section 2.1. Most importantly, data in experiments and ethnographies are usually processed and categorized by the same researcher(s) who conduct the data analyses and synthesize the findings. The use of corpora in research departs from the conventional process as the intended end users of the data and annotations are not those who conducted the data collection and/or annotation. This creates an unconventional situation in which the corpus creators can be held accountable to a higher degree for how the data and its originators (the participants) are portrayed and annotated. However, the advent of corpus linguistics also allows for end users to accept, challenge, or disregard author annotations, which enables expansive inquiries into the data.

2.4. LINGUISTIC IDEOLOGIES AFFECTING RESEARCH. The similarity of issues outlined in previous linguistic research and corpora is no accident, as they are undergirded by three discriminatory linguistic ideologies: Standard Language Ideology (SLI) (Milroy 2001), the Deficiency Model of Bilingualism (DMB) (Cheng et al. 2021; Kutlu et al. 2022), and Languagelessness (Rosa 2016).

The notion of disorderly speech relies on the presupposition that speech can or should be orderly, which is the foundation of Standard Language Ideology (SLI) as described by Milroy (2001). As language is infinitely variable due to its productivity as a system, motions to standardize language necessarily require the prescription of only a subset of possible language. This results in the forms of language associated with the privileged group being designated as Standard, namely that of white upper-class men in the US. Consequently, all forms not adhering to the Standard are marked and considered non-Standard, non-normative, and/or deficient in comparison to the Standard. A final core assertion of the SLI assumes that all speakers are oriented towards and seek to align their language use with the Standard.

In the case of the Deficiency Model of Bilingualism (DMB), the foundational claim is that monolingual language acquisition is the purest form of language acquisition, as bilingualism-from-birth necessarily requires the acquirer's time to be split between two languages. As such, monolingual language is considered a Standard and exemplary of native language acquisition. Bilinguals, particularly those who experience dynamic multi-language environments during childhood language acquisition, are then scrutinized against the monolingual Standard and often may not be seen as a native speaker of a language acquired in childhood. In the most extreme form of the DMB, bilinguals may be seen as deficient as a result of a non-monolingual acquisition, particularly if they are not able to perform language as if they were two monolinguals in a single body (Cheng et al. 2021).

These ideologies in themselves are problematic, but when combined with the racialization of Latinxs, they create a new ideology that seeks to brand Latinxs as languageless (Rosa 2016). This

is achieved by framing non-normative language use by Latinxs as disorderly and non-white (Hill 1998), and non-native, thus casting their linguistic competence in any language into doubt. Rosa (2016) describes how Languagelessness has far-reaching consequences for Latinx livelihoods, as it is one of the bases for which Latinxs are discriminated against in arenas such as education and the criminal justice system.

The SLI, DMB, and Languagelessness are intertwining and self-reproducing linguistic ideologies that negatively affect all humans to different degrees. Linguists and anthropologists are advocating for more nuanced and anti-discriminatory research paradigms by explicitly addressing how such ideologies affect not only the well-being of our society, but also the continued advancement of knowledge in linguistics and other social sciences. Given these concerns, researchers should consider how discriminatory ideologies may affect participant behavior, as well as the possibility that they themselves may have inadvertently internalized these ideologies and reproduced them in their research. Articles such as King (2020) and Cheng et al. (2021) make explicit suggestions for the process of developing reflexive, nuanced, and anti-discriminatory research. The following section describes the process of applying the knowledge gained from these anti-racist scholars to a new framework developed to investigate a MA population's speech.

3. The Mexican American Socio/Historic/Linguistic (MASHL) Framework. In order to fully address the MA population's nuances through linguistic research, I developed the Mexican American Socio/Historic/Linguistic (MASHL) Framework. Although research on MAs has contributed crucial insights into language use and identity construction, they may have also unknowingly propagated the SLI, DMB, and Languagelessness ideologies, as well as ethnolectal models that can ultimately encroach on the goals that were originally proposed. The MASHL framework adapts elements from sociolinguistics, anthropology, and multilingualism to: actively combat discriminatory research ideologies such as the DMB and SLI; approach MA speech with a more nuanced understanding; and contribute to the "discovery of more precise factors which influence language understanding and use" (Cheng et al. 2021). These tenets are directly influenced by the framework and questions in King (2020) and Cheng et al. (2021). The following paragraphs describe the major recommendations put forth by the framework.

MASHL advocates for the combining of multiple subfield approaches while contextualizing MA language use in historical and modern processes. From third wave sociolinguistics, the acknowledgement of the full extent and capability of individuals to situationally shift their language production encourages the researcher to consider their agency over their linguistic performance (Eckert 2008). Additionally, viewing variation as indexical as opposed to the result of incomplete or non-conventional language acquisition invites new perspectives on multilingual language use. Researchers will be better equipped to identify and understand this possible indexical variation by contextualizing MAs and their data in socio/historical processes. Although shifting is possible, research from multilingualism indicates that there are variable limits to this. Whether this is due to individual differences (Paradis 2023) or general human cognitive limits (Flege & Bohn 2021), researchers should be open to possibilities for both large and small ranges of shifting in the individual and the population.

MASHL also advises researchers to consider whether their research frame may uphold SLI, DMB, and Languagelessness ideologies, and how researchers influence data collection and analysis. It is important to scrutinize research questions and methodology to determine which groups or individuals are explicitly or implicitly positioned as a control, a Standard, or worth studying.

Elevating those who use Standard forms or acquired a language during the Critical Period, and framing Standard forms as the goal of language acquisition not only propagate harmful ideologies, but also prevent academics from documenting valuable behavior of marginalized peoples. MASHL encourages researchers to consider different ways of thinking about language profiles that are not based on discrete categories such as "native"/"non-native" and "high"/"low proficiency." Finally, reflexive thought on how the researcher's positionality influences participant behavior is critical in the commitment to understanding the population of study (Holmes 2020).

I then developed the Corpus of Mexican American Language (COMAL) in an effort to carry out a research project using the MASHL Framework as much as possible. The Framework requires a more holistic approach than is usually possible in traditional experimental methods, as social variables are multiplex and difficult to discretize. As a result, the goal of COMAL is the following:

- To document a wide range of Mexican American socio/cultural experiences, language profiles, and linguistic ideologies, while mitigating the harmful effects of SLI, DMB, and Languagelessness ideologies on the researcher's behalf, and acknowledging the author's influence on data collection.
- To provide viable data for quantitative and qualitative analysis using methods from various subfields.

The following section describes how I achieved these goals during the development of COMAL.

4. COMAL Data Collection and Annotation. The procedures and participant recruitment protocol detailed here were conducted under the approval of the Northwestern University IRB.

4.1. PARTICIPANTS. I chose to recruit participants from the city of El Paso, Texas due to multiple reasons. El Paso is a city of historical significance on the border of Texas and the state of Chihuahua, Mexico (Romo 2014). The city is part of the larger metroplex of Paso Del Norte, which also emcompasses Ciudad Juárez, Chihuahua, and Las Cruces, New Mexico. Not only has it been one of the major crossing points between Mexico and the US for centuries (Timmons 1990), but it is also said to be home to the world's largest bilingual workforce on the planet (Chamberlain 2007). El Paso has been documented to be a majority Hispanic city since the '70s (Hedderson 1987; U.S. Census Bureau 2022b), though oral histories indicate that despite intense segregation, people of Mexican and indigenous descent have always existed in the area (personal correspondence). The most recent American Community Survey estimates that in 2022 El Paso was 81% Hispanic, and 67% of the total population over 5 years of age spoke Spanish (U.S. Census Bureau 2022b). El Paso's unique situation makes for a complex MA population, where it may be difficult to distinguish between a "true" monolingual and bilingual due to the historic intermingling of English and Spanish (Baker 1969; Holguín Mendoza 2011). Finally, I chose El Paso as my field site as it is my hometown: I consider myself to be inculcated in regional norms and have personal connections, which I leveraged to develop connections with my participants.

Participants were recruited through posting flyers in English and Spanish on public internet forums such as Facebook and Craigslist, and through snowball sampling. Interested individuals were guided to complete a pre-screening survey to determine whether they met the following criteria:

- Lives in El Paso, Texas
- Identifies as having Mexican ancestry
- Speaks English and/or Spanish
- Is 18+ years old
- Has a computer with access to the internet

Those who qualified were asked to enter their contact information and language preference, which I then used to arrange a two-hour interview. Participants received a $15 virtual gift card for their participation.

At this moment, the recordings available in COMAL represent eleven diverse individuals (ten participants and myself). Full transcriptions and annotations for the first four participants have been completed as of writing. All ten participants lived in El Paso or its exurbs in Texas at the time of their interview; reported racial and/or ethnic identifiers of Hispanic, Latinx, and/or Mexican; and reported being able to speak both English and Spanish. Ages ranged between 25 and 77. Other demographic characteristics reported varied more widely. Table 1 details selected reported and emergent demographic characteristics that differed among participants and myself. Due to the possibility that a participant's pre-interview relationship with me may influence their linguistic production in the recording, the corpus will also contain information about the participant's relationship to me (ex. stranger, close friend, etc.). Participants chose or were assigned a pseudonym, and are additionally labeled using the paradigm "EP#" according to the order in which they were interviewed.

Participant	Age	Immigration Depth	Gender + Sexual Orientation	Occupation
Author	20s	2nd/circular migrant family	trans masculine / queer	PhD student
EP1	20s	3rd+	female / straight	Prof. Doct. student
EP2	30s	adoptive parents: 3rd+	female / lesbian	PhD student
EP3	20s	1st/circular migrant family	female / pansexual	PhD student
EP4	70s	1st/circular migrant family	female / men	government employee
EP5	60s	3rd+	female / straight	retired
EP6	40s	2nd	female / straight	professor
EP7	20s	1.5	female / anything	administration
EP8	20s	1st	male / bi	masters student
EP9	20s	2nd	female / normal	undergraduate student
EP10	20s	1st/circular migrant family	male / straight	PhD student

Table 1. Selected reported and emergent demographic characteristics for each participant and the author.

4.2. RECORDING METHOD AND PROCEDURE. All interviews were conducted by myself, and were held virtually in order to expand my access to participants in El Paso. Participants were asked to be in a quiet room by themselves, and to use headphones or earbuds during the interview if possible in order to capture less ambient noise or speaker echo. Visual contact was maintained

with participants through Zoom, while my and the participant's audio were recorded on separate tracks through Cleanfeed (Sanker et al. 2021) with a backup recording of the participant's audio in Audacity. In the event that the participant was not able to use Cleanfeed due to technical issues, only my audio was recorded through Cleanfeed, and the participant's Zoom audio was recorded through Audacity. Five of the seven interview sets are Cleanfeed-recorded; EP3 and EP5 are Zoom/Audacity-recorded only.

After consent was obtained, audio recording was begun, and preliminary data was recorded. The participant was asked which language they would prefer that I speak while conducting the interview (English, Spanish, or "Spanglish" [i.e. code-switching]). The participant was then asked for a preferred pseudonym and their microphone/computer setup details. Afterwards, I used a script to introduce myself and establish my image as a fellow El Pasoan with an interest in representing the city's culture in my research. Finally, the participant was asked brief questions about how they came to participate in the study, and what (if any) knowledge they had of me. The latter question intends to capture potential motivations for phonetic accommodation during the interview.

All visual stimuli were presented to the participant through the screen sharing function on Zoom.

Tracks of my audio during all interviews, as well as my own recording of attempting the last two tasks will be provided in COMAL.

4.3. TASKS. All tasks were designed to accommodate a range of language profiles and comfort levels by allowing participants to choose if and when to speak either language. Tasks are presented to the participant in order of "least monitored" to "most monitored" in the Labovian sense (Labov 1973). The reason for ordering of tasks according to this paradigm is to encourage trust in me as the interviewer. The type of interview created for COMAL may solicit personal and traumatic memories caused by discrimination and SLI/DMB/Languagelessness. Opening the interaction with what could be perceived as a high judgment task may frame me as an arbiter of language. This in turn affects the participant's willingness to discuss their true personal relationship to language and culture, or reveal non-Standard practices that are valuable and previously undocumented in linguistic research. Participants were reminded throughout the interview that they were welcome to answer as much as they felt comfortable, decline to answer or participate in tasks, or guide the conversation as desired.

These tasks were chosen for the purpose of collecting different types of production data. The final task involved the same read passage in both English and Spanish in order to have at least one recording that is comparable across all participants. A general breakdown of task characteristics, data collected, and language settings are described in Table 2. The English version of all task and interview prompts, questions, and documentation are included in the appendix of this paper, with the exception of the MINT Sprint prompts.

4.3.1. SEMI-STRUCTURED INTERVIEW. This interview is designed to depart from the Labovian sociolinguistic interview (Labov 1973), as it places emphasis on the participant's agency in self portrayal and self-report of speech, language and cultural experience, and identity. A priori researcher-imposed categories are minimized by using open-ended questions and instructions. The protocol required me to ask questions regarding the participant's demographic information, the places they have lived in throughout their life, and the amount of time they spoke or heard the languages important to them throughout their life. The second and third topics are intended to

Task	Language	Based On	Potential Data Collected
Semi-Structured Interview	participant choice	sociolinguistic interview demographic surveys D'Onofrio & King (n.d.) Cheng et al. (2021) King (2020, 2021)	demographic info self-portrayal experience/opinions metalinguistic commentary "unmonitored" speech
ADPs	variable	Bell (1984) Labov (1973)	language/style shifting experience/opinions commentary on shifting "semi-monitored" speech
MINT Sprint	encourage both	Garcia & Gollan (2022) word lists, controlled	language dominance word lists lexical variation "monitored" speech
Reading Passage	encourage both	Zimman (2012) Bradlow (n.d.)	self-portrayal read speech, controlled "monitored" read speech

Table 2. Task breakdown according to language spoken during the task, previous experimental and theoretical paradigms they are based on, and types of data that may be collected using these tasks.

elicit a spontaneous storytelling register of speech as they are asked to recount their life experience.

During the latter portion of the interview, I initiated conversation by either asking questions from a list, or by asking about something the participant had mentioned that I was interested in. I would then either allow the participant to guide the conversation, or picked a new question from the list to discuss. This list of questions contains questions pertaining to: culture, community, place, race/ethnicity, and their connection to language; and metalinguistic commentary on my, the community's, or the participant's own speech. By creating a conversational environment with the participant, this portion of the interview captured a conversational register of speech.

The total amount of recording time for each participant ranges between 1 and 1.5 hours. However, due to the sensitive nature of the content, only de-identified clips of both the participant's and my audio tracks may be distributed in the publicly published version of COMAL. Clips will be annotated with trigger and content warnings such as death, description of violence or discrimination, and so forth. For more information on plans to release de-identified clips from the semi-structured interview, please see the section on Future Directions.

4.3.2. AUDIENCE DESIGN PROMPTS (ADPS). ADPs are a novel task designed to elicit style or language shifts when only one interviewer is available during data collection. Participants may not desire to or be able to consciously speak a different language or style with an interviewer, making it difficult to capture their linguistic range/repertoire. According to Bell's Audience Design theory, interlocutors may style shift as a function of their intended audience. I hypothesize this may still be the case when different hypothetical audiences are presented, though a real audience (the interviewer) is still present throughout.

The basic methodology is as follows.

The participant was presented with a hypothetical interlocutor with particular demographics, and then was asked to explain a cultural touchstone in their community to said hypothetical person. After completion, the participant was asked to describe their impression of who the hypothetical interlocutor is based on the description, and how they changed their speech and explanation based on this impression. The process is repeated using a hypothetical interlocutor with different demographics. For COMAL, the ADP was presented three times in the language the participant asked me to speak in at the beginning of the interview. It was presented as follows: "Imagine you've been introduced to a new person who lives in PLACE, and has never left this place. This person wants to know more about Chico's Tacos in El Paso. What would you tell them?"

The PLACE characteristic was replaced each time using each of the following prompts:

- A small town in the US
- A small town in Mexico
- El Paso

This method is being piloted during data collection for COMAL to determine whether it reliably elicits style/language shifting and/or hypothetical-audience-motivated content modification. The additional question regarding the participant's idea of the hypothetical interlocutor allows the documentation of conscious commentary behind an audience-motivated shift, as well as the perception of a hypothetical person based on demographic information. The register of speech documented by ADPs may be variable depending on the participant's reaction to the prompts.

The total amount of recording time during the ADPs for each participant ranges between 5 and 20 minutes. De-identified tracks will be provided for both me and the participant.

4.3.3. MINT SPRINT. MINT Sprint (Garcia & Gollan 2022) is a language dominance test involving naming 80 pictures in each language, where each row of pictures progresses in terms of vocabulary specialization. The administrative instructions specify that the test must be administered in the participant's dominant language. Prior to administering the test, the participant was asked which language they believe they are dominant in. Although doing so seems circular, this was done to comply with administrative instructions and maintain consistency across participants. Participants are asked to skip pictures if they cannot remember the "correct" word.

This task also functions as a word list, as all participants were presented with the same images, and participants are not required to be literate in either language in order to be able to produce the words. Finally, MINT Sprint may capture lexical and phonetic variation, including home words.

Although using MINT Sprint as a task in COMAL may flout portions of the MASHL Framework by imposing a Standard through progression by "correctly" identifying pictures, it is necessary to include it in order to make this corpus comparable to other research. Harm is mitigated through the addition of lexical variants as possible "correct" answers, and prefacing instructions to the participant using a reminder of the project's intention to document instead of pass judgment. Participants are also reminded that they may decline to participate in all tasks or not complete them.

The total amount of recording time for each participant ranges between 20 and 30 minutes. Two versions of the participant's track will be provided: the de-identified audio of the testing procedure including unsolicited commentary, speech errors, and response time; and a spliced track only containing target words. My de-identified audio of the testing procedure is also included.

4.3.4. READING PASSAGE. The final task was a reading passage conducted in a style pioneered by Zimman (2012) using the North Wind and the Sun passage. Participants were first asked to silently read the passage in the language they had identified as being dominant in, while considering and planning how they would like to present themselves through their speech while reading the passage out loud. After reading out loud, they were asked questions about how they wanted to portray themselves, and what they did to achieve this. The process was then repeated using the version of the passage in the particpant's non-dominant language. In addition to documenting a register associated with reading aloud and scripted storytelling, this may also capture the participant's portrayal of a style/persona they identify with. Most importantly, this task serves to produce recordings of the same utterances under nearly identical conditions for all participants.

This reading task may also flout portions of the MASHL Framework as it requires literacy skills in both languages that disadvantaged MAs may not possess. This is further exacerbated by a possible mismatch in vocabulary between each version of the passage, as some participants commented that they had never heard or read many of the words in the Spanish version. Participants may feel distressed as this task can be reminiscent of their experience with disenfranchisement in academic settings. The task instructions and debrief with the participants is designed to discuss, validate, and address the participant's feelings on the task and their experiences, while highlighting that this project does not intend to measure how "well" they speak.

The total amount of recording time for each participant ranges between 5 and 20 minutes. Two versions of the participant's track will be provided: the de-identified tracks for both the participant and me, and the participant's track only involving the read passage.

4.4. TRANSCRIPTION. Segments containing individual utterances were created using PRAAT (Boersma 2001). Utterance boundaries were created every time there was an audible breath or long pause between utterances.

All files were transcribed by hand by me using ELAN (Lausberg & Sloetjes 2009). This was to have consistency between the perception of the conversation and the transcription, particularly as I have personal experience with common El Pasoan speech norms (such as code-switching, non-Standard pronunciations, local lexemes) and familiarity with the area. Guidelines were adapted from the Chicagoland Project (D'Onofrio & King n.d.) transcription guidelines, with additional conventions created to address non-Standard speech. For each hour of a single talker recording, transcription took approximately two hours. All transcriptions will be de-identified for public distribution.

COMAL transcriptions also include innovative approaches to transcribing and annotating unintelligible speech, and utterances that are difficult to determine whether they are English or Spanish due to non-Standardness or partial intelligibility. In my personal experience, words such as "weenie" are commonly used in otherwise monolingual Spanish contexts in El Paso despite the existence of equivalent Standard Mexican Spanish words (in this case, "salchicha"). There is a case to be made that such words have become Spanish words as a result of borrowing from English, though it could also be argued to be the result of code-switching. Ambiguous and unintelligible utterances are exemplified by the pair of phrases "but I"/"pero I" as they may sound

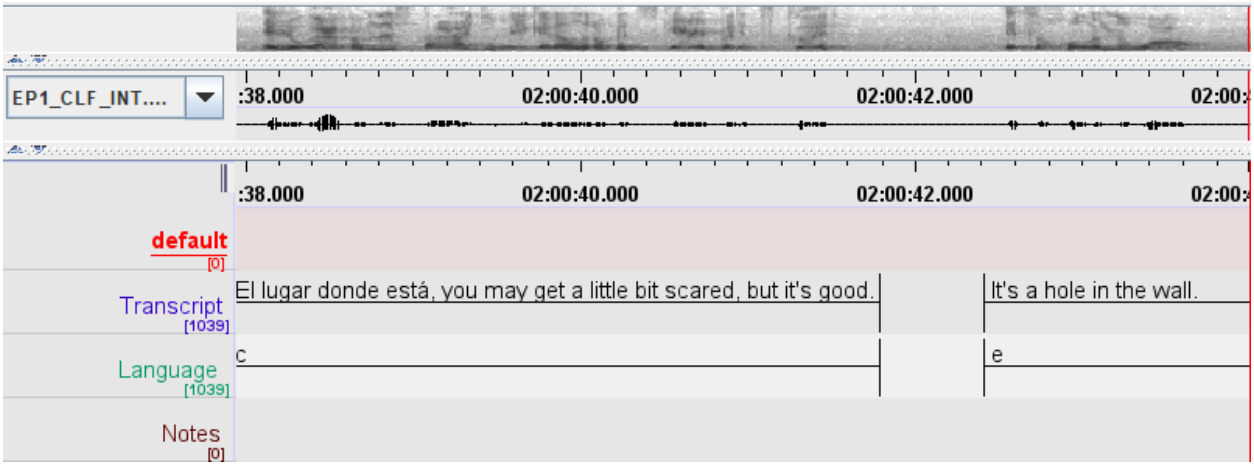


Figure 1. A screenshot of ELAN transcription and annotation tiers for EP1's interview recording file.

similar in fast speech. I have developed innovative conventions in part to acknowledge how my transcriptions are influenced by my perception of the participant. I believe that transcriptions are extended perception tasks; although I advocate for holistic views of each individual participant, my perception of a participant as Spanish dominant may influence me to believe they are saying "pero I" when they may actually be saying "but I." I would prefer to leave interpretations of these ambiguous and unintelligible utterances up to the end users, as their particular needs may require a different approach from mine.

A list of detailed transcription conventions is included in the appendices of this paper.

4.5. ANNOTATION. COMAL is provided with two tiers of annotations: utterance language, and notes. The language tier denotes the language of the utterance as English, Spanish, code-switching, indeterminate, guesses for the utterance's language when the utterance is indeterminate, or Not Applicable (such as for filled pauses).

Due to the ambiguity of some words as described above, and potential of propagating SLI, DMB, or Languagelessness by creating an ontology of which ambiguous or indeterminate words "are" English or Spanish, I have decided to code such words and utterances containing them as indeterminate. COMAL users are encouraged to make their own decisions in accepting or rejecting any transcriptions or annotations I provide in COMAL.

The Notes Tier contains phonetic spellings of mispronounced words, proper nouns, and indicates whether a discourse marker was used in the utterance.

Note: due to the utterance segmentation protocol, multiple segments may represent an entire sentence. As a result, all annotations are utterance-level as opposed to sentence-level.

In addition to the annotations, COMAL will provide force aligned transcripts for all utterances annotated as English and Spanish. These were created by using the Montreal Forced Aligner (McAuliffe et al. 2017). As I remain agnostic as to the language coding of many utterances, only the English and Spanish coded utterances are suitable for force alignment.

5. Limitations. While COMAL largely attains its goals, the encapsulated data may not be useful for all applications or analyses. This is largely due to quirks regarding the modality, how the interview process developed with each participant, and the transcriptions and annotations I intend

to provide. COMAL also currently cannot faithfully capture Mexican/American indigeneity, and essentializes the connection of Spanish with Mexican ethnicity/heritage.

5.1. MODALITY. The use of Cleanfeed as a recording studio requires all users to have technical knowledge of microphone and internet browser settings. Participants who were not computer literate were often unable to troubleshoot microphone issues, even with my guidance. As a result, two participant recordings consist of Zoom audio captured using Audacity. Those who are interested in comparing phonetic or acoustic qualities across all participants may choose to discard these participants' recordings due to issues when comparing across recordings made with different equipment (Sanker et al. 2021).

While recordings in uncontrolled environments are typically vulnerable to environmental noise, internet-based recordings must additionally contend with noise or signal interruptions due to internet lag and equipment malfunction. Some sections of recordings in COMAL suffer from issues such as clicks and voice distortion, which may impact phonetic analyses.

Although the virtual modality made it possible to interview participants from afar, it also makes a large portion of the population of interest inaccessible to me. Access to a computer with stable internet and a quiet environment requires a level of privilege that many MAs in El Paso do not have, particularly those who are most marginalized. Additionally, the most marginalized MA El Pasoans are often Spanish speakers who prefer to or only speak Spanish, while the most privileged are English speakers. The current sample represented in COMAL are all self-identified English/Spanish bilinguals who have completed at least a bachelor's degree, and spoke English for most of the interview. COMAL is currently not representative of all MA experience, but may still serve to demonstrate variation within a convenient sample of demographically similar individuals.

5.2. BALANCE OF LANGUAGES AND STYLES/REGISTERS SPOKEN. The COMAL interview process is designed to maintain the participant's comfort levels due to possible stress or discomfort surrounding the topics addressed in conversation, as well as to mitigate the awkwardness of forcing participants to switch to converse in different languages. Participants are given the freedom to choose which language they and I spoke, which results in some recordings of unscripted conversational speech only containing one language, or not having any "monolingual" utterances.

In order to conduct the lengthy interview process with multilingual tasks, and without burdening participants, the participants were asked or encouraged to switch between English and Spanish multiple times for 3 out of 4 tasks. They also received input from me in both English and Spanish throughout the entire interview process. Recordings thus only have "multilingual" contexts as participants are exposed to and encouraged to language shift and code-switch. This may influence analyses or comparisons with other studies centered on "monolingual" contexts in an experimental setting (Sancier & Fowler 1997).

The balance of languages and styles/registers spoken by the participants is also influenced by my positionality and behavior while conducting interviews. Most notably, due to my position as the child of a university professor and as an academic in the US, I often resorted to speaking an English associated with upper-middle class academics. In my experience/opinion, this behavior causes participants to want to use more formal registers and not speak Spanish with me. My full positionality statement and behavioral notes for each interview will be included in the corpus.

An unintended consequence of the reading passage task design was that many participants reported they felt extreme pressure to perform Standard language in order to "sound like a formal

storyteller." This was not my intention, and the resulting recordings may not correspond with the registers I had claimed in Section 4.3.4 would appear. Nevertheless, these recordings will still be included as they are viable for linguistic study.

5.3. TRANSCRIPTIONS AND ANNOTATIONS. In keeping with the MASHL Framework, I have decided to annotate a large portion of the transcriptions as indeterminate language, and only provide force alignment for utterances coded as English or Spanish. As a result, COMAL may not be a suitable corpus for those requiring ready-to-use corpora for research purposes. While this may be a deterrent, I hope to encourage more researchers to make more considerate choices based off MASHL by reviewing the data provided. Ultimately, the decision to accept or reject my transcriptions and annotations is the end user's to make, which I anticipate and welcome.

5.4. EXCLUSION OF INDIGENEITY. The effects of the subjugation of indigenous peoples in Mexico must also be taken into consideration for COMAL to fulfill its anti-oppressive aspirations. I designed COMAL to only capture English and Spanish language use, which normalizes and essentializes ties between Mexican ethnicity/heritage and Spanish. Doing so contributes to the ongoing erasure of indigenous peoples in Mexico, particularly those who immigrate to the US. MAs from indigenous communities often experience more instability due to their indigeneity, and may not speak English or Spanish, which severely limits their ability to access resources and participate in the wider society.

6. Proof Of Concept. In order to demonstrate COMAL's viability as a specialized, curated corpus, this section will detail a possible phonetic analysis uniquely served by COMAL data.

6.1. BROAD QUESTIONS: TEASING BIDIRECTIONAL INFLUENCE AND STYLISTIC VARIATION APART IN EL PASO. As described in Sections 1 and 2, MA speech production may be influenced by multiplex factors including social, historical, personal, and cognitive. Given El Paso's uniqueness as city historically comprised of a majority Hispanic and Spanish-speaking population, the speech documented in COMAL may be a testing ground to tease apart these factors, or discover causes and correlations for variation observed within. Previous research indicates that bilingual MAs may have bidirectional phonetic influence in their speech, and MAs regardless of their language profile are able to perform stylistic variation. Therefore it would behoove us to compare general trends among the speakers in COMAL with those observed in prior literature, while also assessing whether individual differences improve our understanding of MA performance of speech. This raises the following broad questions:

- Do El Pasoans speak similarly to other MA or English/Spanish bilingual populations in the literature?
- Are there general patterns that indicate El Pasoan English phonetic production is influenced by Spanish?
- Is degree of Spanish use, dominance, or ability correlated with Spanish-influenced English production?
- Can an approach based on theories of style/persona also adequately account for the variation observed?

6.2. NARROW QUESTIONS: RELATIONSHIPS BETWEEN ENGLISH AND SPANISH VOWELS. Much work has been undertaken to understand the difference and relationships between English and Spanish vowels in multiple varieties and among bilinguals (see Bradlow 1995; Fought 2002; Flege & Wayland 2019). Almost all Spanish vowels are included in the English vowel inventory (/i, e, o, u/) with the exception of /a/, but they may or may not be produced in the same acoustic space by a bilingual, depending on whether these categories are merged in accordance with the SLM. As a result, the following narrow questions are proposed:

- Do COMAL speakers have similar vowel spaces in English and Spanish as other MA or English/Spanish bilingual populations in the literature?
- Does each COMAL speaker produce stressed English /i, e, o, u/ vowels similarly to their stressed Spanish /i, e, o, u/ vowels?
- Is the production of stressed English or Spanish vowels similar across all COMAL Speakers?
- Does vowel space relate to the degree of English/Spanish use or dominance?
- Are there styles/personas in the speech recordings that are observable through the vowel space of each speaker?

All of these questions may be addressed using the data in COMAL. However, for the purpose of this paper, only the third, fourth and fifth questions will be further explored.

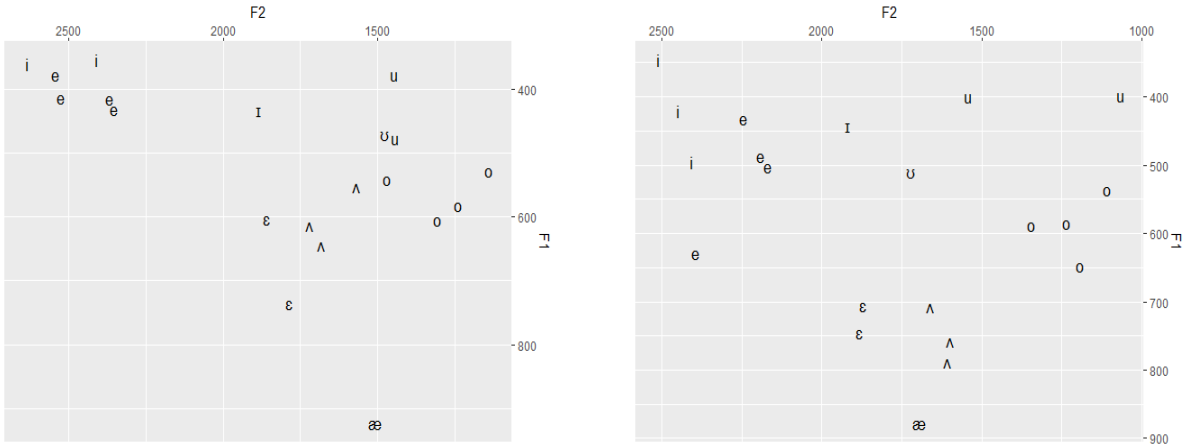
6.3. HYPOTHESES. Although the third and fourth narrow questions have conventionally been assessed without consideration for extra-linguistic factors, I believe that they cannot be assessed without also considering the fifth question. Therefore, my hypotheses are as follows:

- Speakers will not have similar vowel spaces for stressed English vowels.
- Speakers' English vowel spaces will not necessarily correlate with their English/Spanish use or dominance.
- Speakers' English vowel spaces will demonstrate patterns associated with well-documented English styles/personas (such as Valley Girl, cholx, etc), and yet-to-be-documented styles/personas.

6.4. METHODOLOGY FOR PROPOSED QUESTIONS. COMAL provides comparable recordings of several registers and genres for all speakers, including the North Wind and Sun passage and partial word lists through MINT Sprint in both English and Spanish. These more formal recordings contain viable tokens for most of the vowels of interest in stressed positions. F1 and F2 values of tokens from the recordings of the English read passage from myself, EP1, EP2, and EP4 were measured by hand at 50% and graphed in R. These participants' recordings were used for this preliminary analysis as they were captured through Cleanfeed and have undergone the annotation process. Tokens were vowels in stressed syllables and between obstruents, with the exception of /o/ and /æ/ as they did not appear in these environments. All /e/ tokens had a velar obstruent in their environment, and only one token of /æ/ was available ("last").

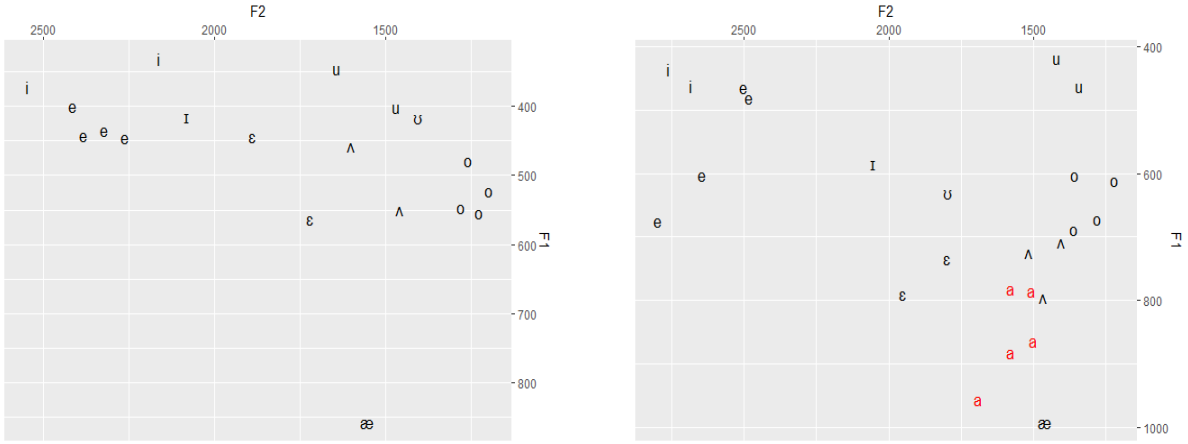
I chose to only examine the relative backness of /æ/ as it does not exist in Spanish, and its relative positioning can index several enregistered sociolinguistic varieties associated with the

California Vowel Shift and the Northern Cities Vowel Shift. Relative backness was calculated by setting the F2 value of the front-most /i/ token as 0% back, and the F2 value of the back-most /o/ token as 100% back, then calculating the relative percentage of the sole /æ/ token for each speaker.



(a) EP1 (b) EP2

Figure 2. English Vowel plots for participants EP1 and EP2. Graph scales are optimized for each speaker's vowel space.



(a) EP4 Vowel Plot (b) Author Vowel Plot

Figure 3. English Vowel plots for participant EP2 and the author, with added Spanish /a/ vowels for the author. Graph scales are optimized for each speaker's vowel space.

6.5. POSSIBLE EXPLANATIONS FOR RELATIVE BACKNESS OF /Æ/. Despite not having comparable tokens for all English vowels, the pilot graphs shown in Figures 2 and 3 are coherent visual representations of four individuals' vowel systems consistent with the general vowel space seen in previous research on US English. In order to compare across speakers, further analyses should focus on measuring vowel dispersion or on norming vowel plots across speakers. Although cross-speaker analyses of vowel spaces are not possible using these plots as each plot's

scale is optimized for each speaker's vowel space, we can calculate the relative positioning of the single /æ/ token across speakers. Table 3 displays relative backness for each speaker using percentage. These results must be considered in light of /l/ preceding the /æ/ token, as coarticulation effects of /l/ may cause the following vowel to be more back. Regardless of this caveat, the calculations indicate that not only is there a sizeable range of backness for /æ/ in the sample, all four speakers produced /æ/ closer to /o/ than to /i/ as they all had percentages at or above 58%.

Speaker	/i/ F2	/æ/ F2	/o/ F2	/æ/ % Back
EP1	2632	1507	1141	75%
EP2	2511	1693	1106	58%
EP4	2545	1552	1199	74%
ZLC	2762	1459	1219	84%

Table 3. F2 values for the front-most /i/ token, back-most /o/ token, single /æ/ token, and the relative backness of /æ/ compared to these measures in terms of percent. Speakers EP1, EP2, EP4, and the author (ZLC) are represented.

It is interesting to note that EP1 and EP4 produced /æ/ at nearly the same relative position (75% and 74%, respectively). As seen in Table 4, these two speakers use English and Spanish at similar rates in similar contexts, unlike EP2 and myself. Perhaps use of English and Spanish may correlate to some degree with this variable. However, although at the time of recording I spoke much less Spanish than these speakers, my /æ/ is relatively much backer. This result is further complicated by the Bilingualism Index, detailed in Table 5, as the two speakers with indexes suggesting relatively more "uneven" language dominance (EP2 and myself) produced the front-most and back-most /æ/. While these results can be further supported by statistical analyses, this demonstrates that COMAL is able to support analyses similar to those conducted in conventional research on multilingual phonetics.

Speaker	School/Work	Family	Friends	Community	Overall Percentage Eng/Spn
EP1	Eng = Spn	Spn	Eng = Spn	Eng = Spn	50%/50%
EP2	Eng = Spn	Eng	Eng	Eng = Spn	90%/10%
EP4	Eng = Spn	Spn	Eng = Spn	Eng = Spn	50%/50%
ZLC	Eng	Eng $<$ Spn	Eng $>$ Spn	Eng $>$ Spn	95%/5%

Table 4. Language(s) spoken in different contexts; relative degree to which each language is spoken in each context; and reported overall percentage of how much each language is spoken on a daily basis by each speaker.

Given that language dominance and use appear to not fully explain /æ/ backness, there may be acquisition-related or sociolinguistic motivations for these differences. COMAL's design allows us to further explore these motivations by examining the responses to the open-ended questions of the semi-structured interview recordings in tandem with the reading passage vowel plots. For example, I discussed with EP1 that I feel conflicted over whether I can say El Paso is my hometown. Even though I lived in El Paso for virtually my entire childhood and adolescence, I

Speaker	English MINT Sprint Score	Spanish MINT Sprint Score	Bilingualism Index
EP1	71	59	0.83
EP2	74	39	0.53
EP4	72	73	0.99
ZLC	79	63	0.80

Table 5. MINT Sprint scores in English and Spanish (of of 80 possible points). Bilingualism Index denotes how "balanced" the person's language proficiency is across their two languages, and is calculated by dividing the lower score by the higher score. A Bilingualism Index of 1 indicates a "balanced bilingual" as they score the same points for both sections.

was born in California, and lived there for 3+ years as an adult. Not only is California important to my identity, but it is also important to my English acquisition and maintenance history, as I was exposed to the California Vowel Shift during the Critical Period and during a formative period as a young adult. Perhaps these connections to California could explain the backness of my /æ/ in the reading passage recording, either through acquisition of the California Vowel Shift in childhood and/or the use of a Californian persona at this point in my life (Hinton et al. 1987).

With regard to the similarity between EP1 and EP4, this could be traced back to the use of similar yet-to-be-discovered personas and styles. During the interview, both speakers discussed their hybrid identities as people from the borderlands who are capable of communicating, existing, and blending in to both El Paso/the US and Juárez/Mexico - "ni de aquí ni de allá." Finally, I believe that EP2's relatively front /æ/ could be incidental to the use of a style/persona related to Chicano English. EP2 was distinct from the other speakers in several dimensions. These included an almost categorical use of stops [t, d] in place of interdental fricatives [θ,] (Fought 2002), and her discussion of her experience as a lifelong resident of the El Pasoan neighborhood that she believes is best known for its association with cholx, pachucx, and Chicanx culture/speech. In addition to my perception of her style as one similar to older working class cholxs/Chicanxs/MAs like actors Danny Trejo and Cheech Marin, EP2 said that she believes that some El Pasoans, including herself, "sound like cholos" despite not being cholx. This is another indicator that language dominance may not necessarily be the only factor influencing phonetic production, as her style/persona during the interview is likely dissimilar from the other speakers.

Although the four speakers represented in these analyses would be listed on Census records with almost identical demographic characteristics, I have demonstrated in this section that they did not behave entirely similarly in several linguistic and social dimensions during their recorded interviews. Moreover, the discovery of these social dimensions and probable connections with speech was possible by also analyzing the semi-structured interviews, which include overt and covert discussions of the participants experience of identity, language, and culture. These preliminary analyses and results indicate that at present COMAL contains a wealth of data that is viable for sociolinguistic and phonetic analyses, including intra- and inter-speaker analyses. Investigations into style/persona for this population will be more robust once all participants' recordings are transcribed, annotated, and force aligned in order to gather more viable tokens, and identify vocalic variation and themes repeated throughout the corpus.

7. Future Directions. As COMAL is still in its infancy, I intend to collect more than 10 sets of Cleanfeed-recorded interviews from El Pasoans. This goal will allow future users the possibility of conducting a robust phonetic analysis on the recordings provided. Participant recruitment will continue to proceed as described in Section 4.1. However, I will request snowballers to attempt to recruit a larger variety of individuals, such as non-binary people and people who prefer to speak Spanish, in order to diversify the demographics reflected in COMAL.

Once at least 10 sets of Cleanfeed-recorded interviews have been completed, transcribed, and annotated, I intend to publish COMAL in a venue suitable for its needs. Because participants may reveal private or traumatic histories during semi-structured interviews, this information and related data may not be suitable for distribution to the public even if de-identified. Participants may still be identifiable though the particulars of the events described, or may face backlash for personal beliefs. More consideration must be made to decide which portions of the semi-structured interview (if any) will be made available to others, and to whom they will be released.

As of the writing of this paper, I am considering publishing all ADP, MINT Sprint, and reading passage recordings and associated data in a publicly available format like the ALLSSTAR corpus on Northwestern University's Speechbox (Bradlow n.d.). The decision to release de-identified clips from the semi-structured interview remains under deliberation. Current options include granting access on a case-by-case basis, or publishing de-identified clips of basic demographic and language profile information. While this strategy chafes against the spirit of open access and prevents contextualization of the participants against their data, I feel that my duty is primarily to their participants' safety. I intend to continue discussing potential repercussions with participants and other academics while preparing COMAL for release. If I ultimately decide to never distribute any portion of the semi-structured interviews, the availability of the other tasks' recordings and associated data to the public would nevertheless attain many goals set by the MASHL Framework and COMAL.

In addition to the materials discussed above, COMAL will also provide more detailed information on the data collection and annotation process, including my personal notes for each interview and a positionality statement.

Future iterations of the interview protocol may be modified in order to address the limitations discussed above, or as MASHL evolves as a framework.

8. Discussion and Conclusion. This paper describes my process of developing the Corpus of Mexican American Language (COMAL), a corpus intended to document the range of linguistic and experiential possibilities within the Mexican American (MA) experience for use in research. Previous research and corpora addressing the MA and similar populations revealed important findings regarding multilinguals' phonetic production and sociolinguistic behavior. However, these research endeavors may have been affected by discriminatory linguistic ideologies in a manner that prevented a more holistic understanding of the populations of study. I created the Mexican American Socio/Historical/Linguistic (MASHL) Framework in response to guide the development of COMAL and advocate for combining methods from various fields for a more integrative approach. As a result, COMAL currently consists of recordings of interviews between me and 10 MA participants living in or around El Paso Texas; 4 of the recordings are accompanied by transcription, annotation, and documentation data. Preliminary phonetic analyses of the data indicate COMAL's viability for future use by other researchers. Though the current corpus is limited, my future plans for COMAL include expanding the sample size represented, publish-

ing the corpus in a public venue and/or distributing certain portions of interviews, and modifying portions of the interview and tasks.

By using methods of data collection from multilingual phonetics/phonology, sociolinguistics, and linguistic anthropology, the data collected in COMAL is appropriate for research in multiple fields. The specificity and amount of data for each participant allows for case studies; furthermore, the eventual breadth of demographics and number of participants will enable corpus and generalizable studies. Such investigations can include questions on identifying El Pasoan varieties of languages, influence of Spanish on English production, and identification of previously undocumented styles/personas in the MA repertoire. Additionally, COMAL's availability will empower researchers to compare studies and interpretations of data, which encourages transparency and comparability in both quantitative and qualitative fields.

References

Aldama, A.J., C. Sandoval, P.J. Garca, M. Daz-Snchez, M. Lugones, A. Chabram, K.M. Davalos, C. Gutirrez-Jones, D.V. Taylor-Garca, N.E. Cant & others. 2012. *Performing the US Latina and Latino Borderlands* Cine y espectaculos. Indiana University Press. https://books.google.com/books?id=a_rfBtZBHwYC.

Baker, George Carpenter. 1969. *Pachuco: An american-spanish argot and its social functions in tucson, arizona*. University of Arizona Press (Tucson, AZ).

Bell, Allan. 1984. Language style as audience design. *Language in society* 13(2). 145–204.

Benmamoun, Elabbas, Silvina Montrul & Maria Polinsky. 2013. Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical linguistics* 39(3-4). 129–181.

Best, Catherine T et al. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words* 167(224). 233–277.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot. Int.* 5(9). 341–345.

Bradlow, Ann R. 1995. A comparative acoustic study of english and spanish vowels. *The Journal of the Acoustical Society of America* 97(3). 1916–1924.

Bradlow, Ann R. n.d. Speechbox https://speechbox.linguistics.northwestern.edu.

Bullock, Barbara E. & Almeida Jacqueline Toribio. 2013. The spanish in texas corpus project http://www.spanishintexas.org.

Chamberlain, Lisa. 2007. 2 cities and 4 bridges where commerce flows. https://www.nytimes.com/2007/03/28/realestate/commercial/28juarez.html?pagewanted=all.

Cheng, Lauretta S. P., Danielle Burgess, Natasha Vernooij, Cecilia Sols-Barroso, Ashley McDermott & Savithry Namboodiripad. 2021. The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology With Inclusive and Accurate Measures. *Frontiers in Psychology* 12. 715843. https://doi.org/10.3389/fpsyg.2021.715843. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.715843/full.

De Houwer, Annick. 2009. *Bilingual first language acquisition*, vol. 2. Multilingual Matters.

D'Onofrio, Annette & Sharese King. n.d. The chicagoland language project. https://sites.northwestern.edu/chilang/.

Dowling, Julie A. 2014. *Mexican americans and the question of race*. University of Texas Press.
Dudley-Marling, Curt & Krista Lucas. 2009. Pathologizing the language and culture of poor children. *Language Arts* 86(5). 362–370.
Eckert, Penelope. 2008. Variation and the indexical field 1. *Journal of sociolinguistics* 12(4). 453–476.
Flege, James E. 2007. Language contact in bilingualism: Phonetic system interactions. *Laboratory phonology* 9(353-381).
Flege, James E & Ratree Wayland. 2019. The role of input in native spanish late learners production and perception¡? br?¿ of english phonetic segments. *Journal of second language studies* 2(1). 1–44.
Flege, James Emil & Ocke-Schwen Bohn. 2021. The revised speech learning model (slm-r). *Second language speech learning: Theoretical and empirical progress* 3–83.
Fought, Carmen. 1999. A majority sound change in a minority community:/u/-fronting in chicano english. *Journal of sociolinguistics* 3(1). 5–23.
Fought, Carmen. 2002. *Chicano english in context*. Springer.
Fought, Carmen. 2010. Language as a representation of Mexican American identity. *English Today* 26(3). 44–48. https://doi.org/10.1017/S0266078410000131. https://www.cambridge.org/core/journals/english-today/article/language-as-a-representation-of-mexican-american-identity/82987A7DCF93F06884E60D59FF7C8462. Publisher: Cambridge University Press.
Garcia, Dalia L & Tamar H Gollan. 2022. The mint sprint: Exploring a fast administration procedure with an expanded multilingual naming test. *Journal of the International Neuropsychological Society* 28(8). 845–861.
Glenn, Evelyn Nakano. 2015. Settler colonialism as structure: A framework for comparative studies of u.s. race and gender formation. *Sociology of Race and Ethnicity* 1(1). 52–72. https://doi.org/https://doi.org/10.1177/2332649214560440.
González, Rafael Jesús. 1988. Pachuco: The birth of a creole language .
Hedderson, John. 1987. The population of el paso county by age, sex and ethnicity, 1980-1990. *Southwest Journal of Business and Economics* 5(2). 1.
Hill, Jane H. 1998. Language, race, and white public space. *American anthropologist* 100(3). 680–689.
Hinton, Leanne, Birch Moonwomon, Sue Bremner, Herb Luthin, Mary Van Clay, Jean Lerner & Hazel Corcoran. 1987. It's not just the valley girls: A study of california english. In *Annual meeting of the berkeley linguistics society*, vol. 13, 117–128.
Holguín Mendoza, Claudia. 2011. Language, gender, and identity construction: Sociolinguistic dynamics in the borderlands .
Holmes, Andrew Gary Darwin. 2020. Researcher positionality–a consideration of its influence and place in qualitative research–a new researcher guide. *Shanlax International Journal of Education* 8(4). 1–10.
Hurtado, Aida & Patricia Gurin. 1987. Ethnic Identity and Bilingualism Attitudes. *Hispanic Journal of Behavioral Sciences* 9(1). 1–18. https://doi.org/10.1177/0739986387030901O1. https://doi.org/10.1177/07399863870309010. Publisher: SAGE Publications Inc.
Jenkins, Devin. 2018. Spanish language use, maintenance, and shift in the united states. *The Routledge handbook of Spanish as a heritage language* 53. 65.

King, Sharese. 2020. From african american vernacular english to african american language: Rethinking the study of race and language in african americans speech. *Annual Review of Linguistics* 6. 285–300.
King, Sharese. 2021. Rethinking race and place: The role of persona in sound change reversal. *Journal of Sociolinguistics* 25(2). 159–178.
Kuhl, Patricia K, Barbara T Conboy, Sharon Coffey-Corina, Denise Padden, Maritza Rivera-Gaxiola & Tobey Nelson. 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1493). 979–1000.
Kutlu, Ethan, Samantha Chiu & Bob McMurray. 2022. Moving away from deficiency models: Gradiency in bilingual speech categorization. *Frontiers in psychology* 7428.
Labov, William. 1973. *Sociolinguistic patterns* 4. University of Pennsylvania press.
Lausberg, Hedda & Han Sloetjes. 2009. Coding gestural behavior with the neuroges-elan system. *Behavior research methods* 41(3). 841–849.
Lenneberg, Eric H. 1967. The biological foundations of language. *Hospital Practice* 2(12). 59–67.
MacWhinney, Brian. 2014. *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.
McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386.
Mendoza-Denton, Norma. 2008. *Homegirls: Language and cultural practice among latina youth gangs*. Blackwell.
Milroy, James. 2001. Language ideologies and the consequences of standardization. *Journal of sociolinguistics* 5(4). 530–555.
Nagy, Naomi. 2011. A multilingual corpus to explore variation in language contact situations. *A Multilingual Corpus to Explore Variation in Language Contact Situations* 65–84.
Ornstein-Galicia, Jacob L. 1987. Chicano caló: description and review of a border variety. *Hispanic Journal of Behavioral Sciences* 9(4). 359–373.
Paradis, Johanne. 2023. Sources of individual differences in the dual language development of heritage bilinguals. *Journal of Child Language* 50(4). 793–817.
Romo, David Dorado. 2014. *Ringside seat to a revolution: An underground cultural history of el paso and juárez: 1893-1923*. Cinco Puntos Press.
Rosa, Jonathan Daniel. 2016. Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology* 26(2). 162–183.
Rosas, Gilberto. 2010. Cholos, chúntaros, and the criminalabandonments of the new frontier. *Identities: Global Studies in Culture and Power* 17(6). 695–713.
Sancier, Michele L. & Carol A. Fowler. 1997. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics* 25(4). 421–436. https://doi.org/https://doi.org/10.1006/jpho.1997.0051. https://www.sciencedirect.com/science/article/pii/S0095447097900516.
Sanker, Chelsea, Sarah Babinski, Roslyn Burns, Marisha Evans, Jeremy Johns, Juhyae Kim, Slater Smith, Natalie Weber & Claire Bowern. 2021. (don't) try this at home! the effects

of recording devices and software on phonetic analysis: Supplementary material. *Language* 97(4).
Schouten, Andy. 2009. The critical period hypothesis: Support, challenge, and reconceptualization. *Studies in Applied Linguistics and TESOL* 9(1).
Timmons, Wilbert H. 1990. *El paso: A borderlands history*. University of Texas Press.
U.S. Census Bureau. 2022a. ACS 1-year estimates public use microdata sample Retrieved from https://data.census.gov/mdat//search?ds=ACSPUMS1Y2022cv=HHL(0,1,2)rv=HISPwt=PWGTP.
U.S. Census Bureau. 2022b. Selected characteristics of the native and foreign-born populations. U.S. Census Bureau. Accessed on 25 October 2023. https://data.census.gov/table/ACSST1Y2022.S0501?q=el+paso+texas.
Velázquez, Isabel. 2009. Intergenerational spanish transmission in el paso, texas: Parental perceptions of cost/benefit. *Spanish in Context* 6(1). 69–84.
Yao, Yao & Charles B Chang. 2016. On the cognitive basis of contact-induced sound change: Vowel merger reversal in shanghainese. *Language* 92(2). 433–467.
Zimman, Lal. 2012. *Voices in transition: Testosterone, transmasculinity, and the gendered voice among female-to-male transgender people*: University of Colorado at Boulder dissertation.

9 Appendix

9.1. Opening Interview Questions and Script.

9.1.1. Pre Consent.

- Make small talk, could talk about how their day is
- Verify name and pronouns
- Short intro
- Establish language for consent form: "The first thing we need to do is go over the consent form. I would like to make sure that you understand the form so that you can give your full consent. Would you rather I speak English or Spanish to discuss the consent form with you? You can change your mind later if you want."

9.1.2. Consent.

- Reminder: "You can choose to not participate in all portions of the study. If you choose to not participate in all parts of the study, you will still receive the $15 gift card."
- Go over optional elements: "This study has two optional elements that you can consent to. Do you consent to...?"
- Obtain consent verbally, write down necessary fields

9.1.3. Set up Cleanfeed and Audacity.

- "I'm going to send you a link through Zoom chat. The website is going to record your audio, so please make sure that you do not close that window. Your web browser will ask you to give permission to use your microphone, please click on yes."
- Once done: "I'm going to mute us on Zoom. Just to remind you, I am also recording this session using Audacity on my own computer."

9.1.4. Briefing and Housekeeping Questions.

- Establish the language for the interview: "As you saw in the consent form, the first thing we're going to do is the interview. For this part, we can talk in English, Spanish, and Spanglish. Do you have a preference if I speak one or the other?" "If you end up changing your mind about the language, please let me know."
- Ask for a pseudonym or assign one: "To guard your privacy, we need to give you a pseudonym for our files. If you haven't picked one out for yourself, we can assign one to you. Is there a pseudonym you'd like to use?"
- Recording instrument questions: "We'd like to know about your computer and microphone setup. What type (brand and model if you know) of computer are you using to participate in this interview?" "I need to ask about your microphone. On Cleanfeed you'll see that there's a gear to the right of the bar that says "Participant". When you click the gear you'll see that there is a new menu that probably says "Use browser setting," click on this and let me know what it says. " (figure out which one should be used)
- Introduce myself: "To tell you a little bit about myself..." Reiterate my name (and pronouns), "I'm doing my PhD, this is my study. I'm from El Paso, I grew up in the Northeast. I lived there until I went to college, but I still go back often to see my parents and friends. I'm mixed race, I'm Chinese and Mexican. I have a lot of ties to EP and Mexican culture, which is why I wanted to do this research."
- Establish schedule for the interview: Task 1 (1hr 20min), Task 2 (10 min), Task 3 (20 min), Task 4 (10 min)
- Recruitment Questions: "How did you find out about this study? What made you decide to participate? Have you ever heard of me? What have you heard about me? Did you prepare or think about any of the questions sent to you in the email? Which ones?"

9.2. Task 1: Semi-Structured Interview.

9.2.1. Demographic Questions.

"Now we can start the interview. I'm going to ask you a few demographic questions. There are lots of ways in which these questions can be answered, but for this interview I'm not asking you to answer the way you'd answer a Census. I want you to answer however YOU want to answer, and tell me what is important to YOU. You don't have to tell me anything you don't want to. You can also take these questions as a jumping off point to talk about things you think are relevant."

- "Can you tell me about your race/ethnicity?"
- "Can you tell me about your Mexican ancestry?"
- "What part of EP do you live in?"
- "Can you tell me about where your family is from?"
- "What languages do you speak?"

- "Can you tell me about your occupation?"
- "Can you tell me about your educational history?"
- "Can you tell me about your gender?"
- "What are your pronouns?
- "Can you tell me about your sexual orientation?"
- "Is there any other aspect about yourself or your identity that I didn't ask about that's important to you?"

9.2.2. AUTOBIOGRAPHY QUESTIONS. "For this part of the interview I'm going to ask you about your life experience. First, I want to hear about where you've lived. I want you to narrate your life in terms of where you've lived, and other places you've spent time in that are important to you. Do you have any questions?"

"The next thing I'd like to have you talk about is what languages are important to you, when you have heard them throughout your life, how much you heard or spoke these languages, who spoke these languages to you, and who you would speak these languages to. You only need to talk about the languages that are important to you."

9.2.3. OPEN ENDED QUESTIONS. These questions are conversation starters. The interviewer may choose to use any of these questions, participants may also guide the conversation.

Culture/Community/Place Questions

- "Is there a culture, or cultural elements, that you grew up with or identify with the most?"
- "Does your cultural identification have anything to do with the languages you speak or understand?"
- "Would you say that you are "from" the place where you currently live?"
- "What does it mean to be a person with Mexican heritage in El Paso?"
- "Do you have to speak Spanish to be Hispanic?"
- "What do you consider to be your community? Who is included?"
- "Is your community tied to El Paso as a whole, or specific parts of El Paso?"
- "What language(s) does your community speak?"
- "How do you engage with your community/the people around you?"
- "What is your role in your community/EP?"
- "What kinds of people exist in EP?"
- "What is EP part of (Texas, etc)?"

- "Does anyone ever say you are like (insert person)?"
- "Based on this interaction with me, who do you think I am? Am I like anyone you've met?"

Speech Evaluation

- "What do you think about the way you speak?"
- "What about in (other language)?"
- "Which language do you think you're dominant in?"
- "Does anyone ever say anything about the way you speak?"
- "Does anyone change the way they act towards you based on how you speak?"
- "Can you tell me about how people in El Paso speak?"
- "Are there any language things that are only heard in El Paso, like specific words, accents?"
- "What do you think about the way I speak? Do I speak like anyone you know?"
- "Did you change anything about the way you speak during this interview based on who you think I am?"

9.3. TASK 2: ADPS. This task was presented as described in Section 4.3.2 with the preface instructions as follows: "I'm going to ask you to do something three times, and there will be something slightly different each time. Here's the first time:"

9.4. TASK 3: MINT SPRINT. This task was administered using the guidelines provided in the materials for MINT Sprint (Garcia & Gollan 2022), and with additional instructions. The task was prefaced with the following script: "Next, we're gonna do a picture task, I'm going to ask you to do this in both English and Spanish, starting with (dominant language). I want to remind you that this study looks at how your relationship with your culture and society influences the way you speak, so I'm not trying to see how "well" you speak either language. I'd really like you to try your best to do this task in both languages, and it is okay if you feel like you can't do it. You are always welcome to tell me if you want to decline to do something. I'm going to read the instructions to you first, and then we will start."

The administrative instructions for MINT Sprint were used, but included this instruction: "However, I'd like you to not say "uh/um" as much as possible, try to space each word out, and ONLY say the word, like this: "banana... mouse.""

9.5. TASK 4: READING PASSAGE. "The last thing we're gonna do today is have you read a passage out loud, first in (dominant language) then in (non dominant language). Before we start, I'd like to remind you that this is not a test, I'm not trying to see how "well" you speak one language or the other. This research project is focused on how different people speak, and whether they might have different cultural reasons for speaking in a certain way.

"I'm going to show you the passage in a second, and I'm going to give you some time to read it silently. I want you to think about how you want to present yourself through your speech. Does this make sense?

"You can start reading out loud whenever you want." " Can you tell me about your thoughts while doing this? How did you want to portray yourself, and how did you speak in order to do so?"

"Now we're gonna do the same thing in (non dominant language). Same as before, take some time to read this passage to yourself and think about how you want to portray yourself through your speech. You can start reading out loud whenever you want."

9.6. POST INTERVIEW. "We're done! Before we wrap up and I send you the gift card, do you have any last questions about the consent form, or the procedures or the questions I asked?"