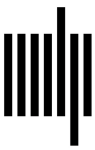
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

17 Managing Legacy Data in a Sociophonetic Study of Vowel Variation and Change

James Grama

1 Introduction: The focus of this data management use case

This data management use case provides a description of the workflow of a sociophonetic study of language variation and change using legacy data (Grama 2015). This study investigated change in the vowel system of Pidgin (known to linguists as Hawai'i Creole), an English-lexified creole spoken in Hawai'i. In that study, two existing corpora were used to conduct a trend study that compared Pidgin speakers at two time points: the 1970s and the 2000s. Analysis was conducted on acoustic vowel measurements taken from speech elicited using sociolinguistic-style interviews (cf. Labov 1972). This use case serves as a meta-analysis of the methods applied in the Grama study and is intended for researchers interested in using naturalistic legacy data to identify longitudinal change. Throughout the chapter, methodological considerations are made that concern sociophonetic studies, the management of legacy data (especially when that data was not originally designed to address sociophonetic research questions), and further issues that are relevant to variationist research.

The chapter is organized as follows: first, the language setting for the current study is described, along with a discussion of my positionality to Pidgin (section 2). Then, I discuss the way archived interviews were selected for analysis, as well as characteristics of the archived corpora used in Grama (2015) (section 3). This is followed by a description of the way vowel data was transcribed (section 4.1), force aligned (section 4.2), manually checked (section 4.3), extracted (section 4.4), cleaned (section 4.5), and normalized (section 4.6). Issues concerning data transparency (cf. Gawne & Styles, chapter 2, this volume), figure interpretability (section 5.1), and storing and sharing of data (section 5.2) are then briefly discussed. Throughout, the Grama study is evaluated in the context of best practices in sociophonetics, and deviations from optimal procedure are noted, where relevant.

I now turn to a brief discussion of the social setting of this study, as the social landscape of any variety is paramount to interpreting the results and understanding why the study took the shape it did.

2 A brief history of Hawai'i and the development of Pidgin

Hawai'i's complex history of contact began from its discovery. The islands were originally settled by Polynesian seafarers between 1190 and 1293 CE (Wilmshurst et al. 2011:1816; Walworth 2014:258), and the following two centuries saw intercultural movement among neighboring Polynesian cultures (Collerson & Weisler 2007). It is generally accepted that sustained contact between Hawai'i and other eastern Polynesians declined sharply in the fifteenth century (Drager 2012b:62). In 1778, Hawai'i was irrevocably altered by James Cook's arrival. European contact opened the floodgates for foreign influx at an unprecedented scale; the islands were quickly exploited for their sandalwood and used both as a strategic launching point during the whaling trade and a stopover point in the fur trade of the early 1800s (Reinecke 1969:24). Foreign presence was debilitating for Hawaiians and their language ('ōlelo Hawai'i).

The research discussed in this paper was made possible by funding from the Russell J. and Dorothy S. Bilinski Dissertation Fellowship Award and the University of Hawai'i at Mānoa. Many thanks are due to the editors of this handbook, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, to two anonymous reviewers, and to Melody Ann Ross for providing helpful feedback on earlier drafts. All remaining errors are my own. Greatest thanks go to the participants who graciously lent their voices to these collections.

Throughout the 1800s, the number of native Hawaiians dropped precipitously due largely to foreign disease (Stannard 1990:330). An increase in Christian missionaries in 1820 further contributed to the decline of ʻōlelo Hawaiʻi, as English was elevated as the language of the church, economic advancement, and social capital (Drager 2012b:63). The overt prestige of English was further entrenched by the steady growth of US influence in schools. A wealthy, English-speaking minority affected policies that gradually forced ʻōlelo Hawaiʻi schools to switch to English as the primary language of instruction. All of this was designed to prepare the youth of Hawaiʻi for "participation in an American-type community" (Stueber 1964:144). In July 1887, a group of wealthy US businessmen coerced King Kalākaua under threat of force to sign a new constitution that stripped the Hawaiian monarchy of its authority and disenfranchised native Hawaiians. A mere five years later, the Kingdom of Hawaiʻi was overthrown by a wealthy, white minority, and Hawaiʻi was illegally annexed by the US in 1898.

The rapidly changing social climate of Hawaiʻi in the nineteenth century was spurred on by the establishment of sugarcane plantations in 1835. The plantations lured laborers worldwide, initially, Cantonese and Portuguese, then Japanese, laborers from the Philippines, and later, from Germany, Korea, Puerto Rico, Spain, and islands throughout the Pacific. While the plantation foremen originally spoke ʻōlelo Hawaiʻi, English soon took its place, reflecting a changing social climate that was increasingly dominated by an English-speaking minority. Because many languages were used on the plantations with no shared first language among the workers, an English-based pidgin arose to facilitate communication. Children raised in this context commonly spoke both the language(s) of their parents and a creolized version of English—Hawaiʻi Creole, known locally as Pidgin (Kawamoto 1993). As English speakers solidified a position of overt economic power, Pidgin increasingly took on the role of lingua franca among Hawaiʻi Locals, and within three generations (by the 1930s), Pidgin was a distinct language from English (Roberts 2004).

Pidgin has become closely linked with a Local identity, due in part to its development alongside Hawaiʻi's changing demography. This identity encompasses many cultural backgrounds and ethnicities, the extent of which is beyond the scope of this use case (for detailed accounts, see Fujikane 1997; Rohrer 1997; Ohnuma 2002), but a crucial opposition places *Local* at conceptual odds with *Haole*. For many, the category "Local" comprises a non-white person, born in Hawaiʻi, with a familial connection to the plantations and Pidgin, while the category "Haole" comprises a white person whose presence on the islands is the result of historical exploitation via colonialism. "Haole" is a construct characterized by external forces (e.g., whiteness, the military, tourism, English dominance), while "Local" evinces solidarity against that force (Ohnuma 2002). Attitudes toward Pidgin and English reflect this opposition. Pidgin is marginalized by many as "broken English," while English is upheld as "proper" (Drager & Grama 2014). Moreover, English is often cited as a language of economic mobility, and despite Pidgin's use in publicly visible domains (e.g., politics, the news), it is nevertheless viewed by many as a language whose use should be restricted to the home and close personal relationships (see Marlow & Giles 2008, 2010).

2.1 A note on researcher positionality

The history of colonialism in Hawaiʻi cannot be ignored when discussing the present study. In keeping with the observation that data cannot be divorced from its source (Holton, Leonard, & Pulsifer, chapter 4, this volume), the researcher's positionality to the variety under study is key. In the context of Hawaiʻi, it matters greatly that I am a white, non-native Pidgin speaker from California. My very presence as a researcher in Hawaiʻi was the result of settler privilege. Because of this, the likelihood that I could elicit representative Pidgin speech data is low, given the long history of language hegemony (see Marlow & Giles 2008, 2010). Therefore, it was both prudent and methodologically necessary to make as much use of existing data as possible in lieu of collecting new data.[1]

Having established context, I now move on to describe the goals of the original study, as well as how legacy data were used to address those goals.

3 Using legacy data to achieve research goals

Building on phonological work on Pidgin (e.g., Bickerton 1976; Sakoda & Siegel 2008), the focus of Grama (2015) was to acoustically characterize changes in the vowel system of Pidgin over time, with considerations to both internal and external factors.[2] While phonological

descriptions of Pidgin highlight that inter- and intra-speaker variation is clearly present in the language, relatively few studies have applied variationist methods to acoustic data to describe this variation (for exceptions in other creoles, see Kraus 2017; Lesho 2014; Rosenfelder 2009; Sabino 1996, 2012; Veatch 1991; Wassink 1999, 2001, 2006).

To characterize changes over time in Pidgin, Grama (2015) employed both real- and apparent-time data taken from archival recordings from two existing corpora on Kaipuleohone, the University of Hawai'i at Mānoa's digital ethnographic archive (Albarillo & Thieberger 2009; Berez 2013); one collection was recorded in the 1970s (with speakers born 1896–1946), and one was recorded in the 2000s (with speakers born 1947–1988). These data were appropriate for answering questions of language change over time because they represented two independent samplings of the Hawai'i speech community approximately thirty years apart, with speakers born over a ninety-year time frame. This allowed for real-time comparisons to be made across corpora and apparent-time comparisons to be made within each corpus. That the data were already extant was particularly beneficial, given that many of the recordings were conducted between native Pidgin speakers, and that my position as a migrant, white, non-native Pidgin speaker in Hawai'i limited my ability to reliably collect new data. These corpora—the Bickerton Collection and the Influences and Variation in Hawai'i Creole English—are discussed in the next section.

3.1 The corpora

3.1.1 The 1970s corpus (the Bickerton Collection) The Bickerton Collection includes materials elicited from Pidgin speakers and from speakers of a range of other languages. Recordings were made between 1970 and 1980 and include people born between the mid-1890s to the late 1940s. Speakers were represented across these birthdates, but they tended to be either in their mid-to-late sixties, or in their early thirties to mid-forties; this created a natural break in the corpus, where about half the speakers were older than fifty years of age and half were younger. The collection was amassed to describe the linguistic structure of Pidgin, with special attention paid to morpho-syntactic and phonological alternations (for studies based on this corpus, see Bickerton 1976; Odo 1975, 1977). Interviews were conducted by both Pidgin and non-Pidgin speakers; the apparent dominant language of the interviewer became an important criterion for interview exclusion from Grama (2015) (see section 3.1.3). In all, the collection includes 168 recordings, which vary widely in duration. While many of the recordings are in the style of traditional sociolinguistic interviews, a number of them are recordings of radio and television programs or advertisements. Throughout the remainder of the chapter, this collection will be referred to as the 1970s corpus.

3.1.2 The 2000s corpus (Influences and Variation in Hawai'i Creole English) The Influences and Variation in Hawai'i Creole English collection comprises sociolinguistic-style interviews with Pidgin speakers. Recordings were made in the early-to-mid-2000s and include people born between the mid-1940s to the mid-1980s. Speakers were represented across these birthdates, but they tended to be either in their late forties, or in their early twenties, again creating a natural break in the corpus; here, approximately half the speakers were older than thirty years of age and half were younger. The corpus was amassed by Jeff Siegel and a number of research assistants to examine variation in Pidgin and the role played by external influences on language change (for studies based on this corpus, see Sakoda & Siegel 2003, 2008; Siegel 2007). Interviews were conducted by Pidgin speakers, with Pidgin speakers, and interviewees were typically friends, family members, colleagues, or otherwise previously known to the interviewer. In all, the collection includes 117 recordings, ranging from around six minutes to one hour and forty minutes. Throughout the remainder of the chapter, this collection will be referred to as the 2000s corpus.

3.1.3 Interview selection Selecting recordings from these corpora that were appropriate for a sociophonetic study of language change over time proved labor-intensive, largely because the original intent of each corpus was not to facilitate sociophonetic research. Recordings were prioritized that met five constraints. First, sufficient information had to be included in the metadata or in the recording itself to indicate that interviewees were born and raised in Hawai'i and not people who emigrated to Hawai'i later in life.[3] This constraint was implemented to be as certain as possible that speakers included in the study were native Pidgin speakers, and

not second-language learners, as later acquisition can impact phonological realizations (see, e.g., Flege, Shirru, & MacKay 2003). Second, interviews were prioritized where the interviewer was also a Pidgin speaker. Given the history of linguistic hegemony in Hawai'i, people are less likely to use Pidgin if their interlocutor does not also speak Pidgin. This is especially true in more formal domains (e.g., in a recorded conversation). Third, recordings with one interviewee were desirable because recordings with multiple interlocutors yield overlapping speech and uneven turn-taking, which makes preparing data for forced alignment more labor-intensive. Fourth, recordings needed to be of high-enough quality to undergo acoustic analysis. Quality issues rendered many recordings unusable; wind, excessive background noise, static, feedback, clipping, and quiet interviewees yielded recordings that were unlikely to produce reliable formant measurements. Fifth, recordings had to feature enough speech to map a speaker's vowel space.[4] Through conducting the original study, it became clear that interviews lasting around twenty minutes, or two thousand words, reliably produced enough vowel tokens across vowel category to accurately map the vowel space. Interviews that met each of these five constraints were candidates for inclusion in Grama (2015).

There was also a desire for a balanced number of speakers to ensure equal representation across demographic category. These categories consisted of *corpus* (1970s vs. 2000s), wherein real-time, longitudinal change could be tested; *relative age* within the corpus (relatively older vs. relatively younger speakers), so that change in apparent time could be tested, where relatively older speakers represent the language as it was in the past, and relatively younger speakers represent the language as it is spoken at the time of recording (Bailey 2004); and binary *sex* (female, male), to assess how females and males differed (if at all) in their participation in the identified changes. Pruning the two corpora based on these constraints resulted in a total of thirty-two total recordings. The distribution of speakers across the tested demographic categories included in Grama (2015) is reported in table 17.1.

Despite these restrictions, inherent differences exist between the two collections, particularly with respect to interview styles. In the 2000s corpus, interviewers and interviewees tended to be close friends or family members. The familiarity between speakers resulted in very conversational interactions. By contrast, 1970s interviewers were less likely to be previously known to interviewees and tended to feature more monologic speech styles. Additionally, differences arose that reflected Hawai'i's changing social climate over a thirty-year period. Young 2000s speakers as a group, for example, completed more formal schooling than 1970s speakers, who typically did not achieve more than a high school education. By contrast, 1970s speakers were more likely to have worked in positions that required manual labor than 2000s speakers did. Appendix A provides a more detailed breakdown of demographic information for each speaker.

Table 17.1
Speaker numbers with age information across corpus, relative age, and sex

Corpus	Relative age	Sex	Mean age at time of recording	Mean birthdate	*N*
1970s	Old	Female	61	1913	4
		Male	65	1911	4
	Young	Female	40	1935	4
		Male	33	1940	4
2000s	Old	Female	49	1958	4
		Male	48	1959	4
	Young	Female	22	1985	4
		Male	22	1985	4
				Total	32

4 Data processing

The following section describes how data were processed in Grama (2015), including how interviews were transcribed and force aligned; how the resulting alignments were manually checked; and how vowel formants were extracted, cleaned, and normalized.

4.1 Transcription

After selection, interviews were orthographically transcribed at the utterance level. Research concerned with the acoustic properties of vowels ultimately relies on transcription of some kind. While there have been significant strides in automated speech recognition, fully automated transcription is far from accurate enough to replace human transcribers. Thus, manual transcription remains a necessary bottleneck to acquiring large amounts of high-quality vowel data (cf. Seifert et al.

2018:335). Currently, the industry standard for transcription is ELAN (Lausberg & Sloetjes 2009), which is under continual development. For this study, however, interviews were transcribed and time aligned in Transcriber (Barras et al. 2001). While Transcriber's interface has a gentler learning curve than ELAN's, this choice was made largely because Transcriber, not ELAN, interfaced with LaBB-CAT (Language, Brain and Behaviour – Corpus Analysis Tool) (Fromont & Hay 2012), the forced aligner used in Grama (2015) (see discussion in section 4.2). However as of writing, Transcriber suffers from a lack of upkeep. Prior to an update in March 2017, Transcriber had not been updated since 2005. Given that ELAN now interfaces with LaBB-CAT, it would be my choice of transcription software were I to undertake the study today.

A mean of twenty-two minutes, or two thousand words, was transcribed for each speaker for the original study. Transcription was of temporally contiguous sections, increasing the likelihood that an interviewee used roughly the same speech style. Overt discussions about Pidgin were avoided when possible, as this motivated some speakers to shift toward English; issues of recording quality also made it necessary to skip (usually short) sections of the interview until conditions improved. Table 17.2 summarizes the transcription statistics from each interview, noting word count as is typical in corpus descriptions.

Table 17.2
Transcription metadata

Age	Sex	Speaker pseudonym	Time transcribed (hr:min:sec)	Word count
Old 1970s	Male	Joseph	00:21:16	2,744
		Kawika	00:19:42	2,358
		Kimo	00:30:30	1,508
		Manny	00:32:06	2,831
	Female	Kaimana	00:18:45	2,324
		Keiko	00:15:13	1,689
		Kaimana	00:18:45	2,324
		Malia	00:27:56	2,766
Young 1970s	Male	Danny	00:19:36	1,735
		Eddie	00:19:23	2,245
		Glen	00:19:15	1,777
		Victor	00:14:42	2,045
	Female	Delia Jane	00:18:17	2,099
		Eddie	00:19:23	2,245
		Mona Lisa	00:18:40	1,858
		Teresa	00:26:58	1,930
Old 2000s	Male	Grant	00:17:06	1,976
		Keoni	00:14:11	1,952
		Kevin	00:18:36	1,910
		Palani	00:30:05	2,063
	Female	Carla	00:13:32	1,927
		Kahea	00:15:09	2,201
		Lani	00:14:05	1,727
		Pua	00:28:50	1,707
Young 2000s	Male	Alika	00:11:55	2,142
		Eric	00:27:37	2,018
		Kaleo	00:17:23	2,230
		Myko	00:21:48	2,038
	Female	Lena	00:27:30	1,941
		Mina	00:33:42	1,840
		Sarah	00:20:24	1,984
		Starla	00:24:09	1,966
		Total	11:16:29	66,100

4.2 Forced alignment

After interviews were transcribed, they were force aligned. Forced alignment refers to the process of automatically creating segmentations at the level of the phoneme using the acoustic signal (usually a .wav file) and an orthographic transcription (in this case, a .trs file). The rise of computational methods in the past decade has revolutionized research methods in phonetics, drastically increasing the speed of getting analyzable data (cf. Schiel, Draxler, & Harrington 2011). Choosing a forced aligner is therefore crucial, as they are now standard practice of workflows in many production-based sociophonetic studies.[5]

In Grama (2015), corpus storage and forced alignment were performed using a server build of LaBB-CAT housed at the Sociolinguistics Server (SOLIS) at the University of Hawai'i at Mānoa (Drager 2012a).[6] This choice was made both because of LaBB-CAT's ability to automatically generate corpus annotations and because I had prior experience using the system. Transcriber files were uploaded along with their accompanying .wav files to SOLIS, and files were tagged with available metadata. Each uploaded transcript was automatically checked for words that were not in the English CELEX dictionary (Baayen, Piepenbroock, & Gulikers 1995), with which LaBB-CAT is designed to interface. Unfamiliar items in the dictionary were added using the grapheme-to-phoneme (G2P) mapping system that CELEX employs for British English.[7] Errors in the transcript (e.g., misspellings, unrecognized characters) were corrected in SOLIS with LaBB-CAT's transcription editing protocols. Forced alignment

was produced using the Hidden Markov Model Toolkit, HTK (Young et al. 2009), via the default train/align procedure (see Fromont & Watson 2016), which produced automated boundaries around phonemes according to lexical data from CELEX.

Force-aligned data in LaBB-CAT can be accessed in two main ways. First, the user can access the *transcripts* tab to interface directly with the transcript and toggle available annotation layers on and off. These layers are programmed by the user and generated automatically. Information can then be extracted as a time-aligned file in several formats, for example, as a Praat TextGrid (Boersma & Weenink 2019). While the entire interview file can be exported as one Praat TextGrid (which is preferable if one-to-one correspondence between file and speaker is desired), Praat often has issues with larger files, making them difficult to work with if processing power is at a premium. The second way to extract the force-aligned data is via LaBB-CAT's *search* function. By executing a search across the segments layer, LaBB-CAT returns all instances that fit the specified criteria. Individual segments can then be extracted, a process which yields TextGrids and .wav files for all aligned intervals that correspond to interval breaks in Transcriber or annotations in ELAN. For example, the user could specify one or more speakers in LaBB-CAT and query a specific vowel. Figure 17.1 shows an example of such a search executed for Myko, a young 2000s male.

This process was performed independently for each vowel category and individual speaker. The pairs of .wav files and TextGrids (between 25 and 350 files per vowel per speaker) were then extracted from SOLIS and stored on my personal computer in a folder organized by speaker by vowel. In accordance with best practices for data management, these files will be archived in the future in one of three locations: SOLIS, Kaipuleohone, or the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

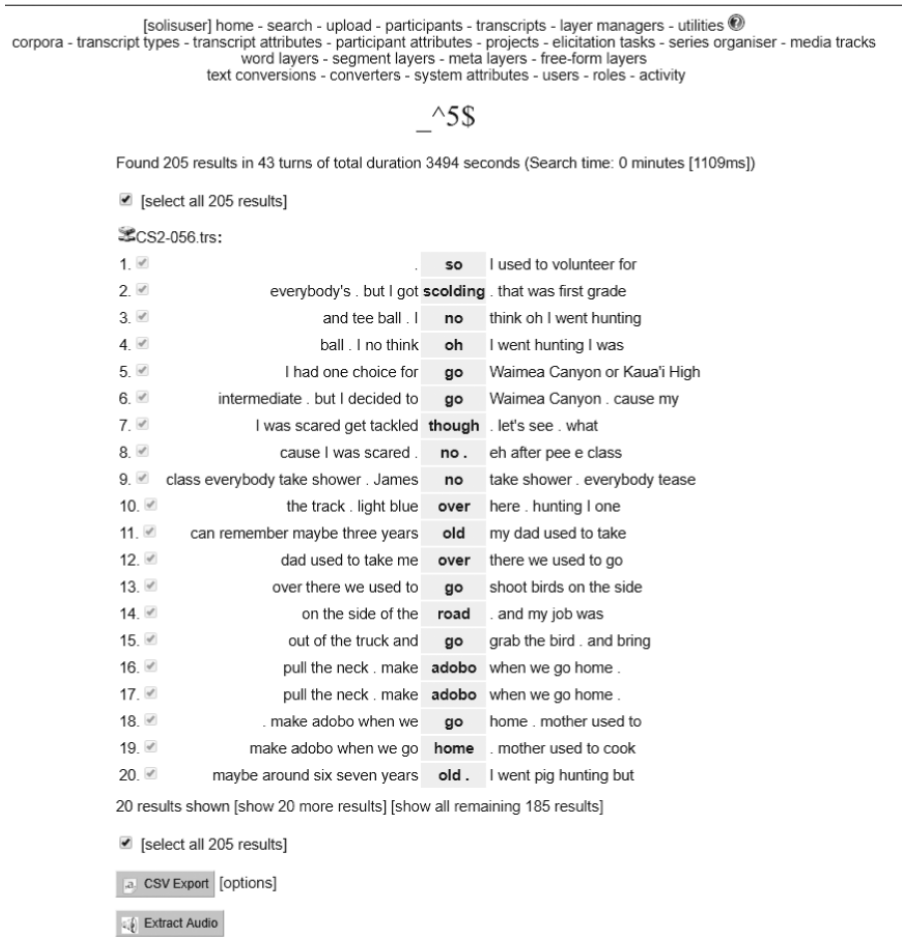


Figure 17.1
LaBB-CAT's output for queried /o/ (represented by 5); first twenty tokens displayed.

4.3 Checking data

After forced alignment and extraction, vowel alignments were manually checked for accuracy and coded following the protocol described here. There is some debate in sociophonetics as to whether, and under what circumstances, manual checking of force-aligned vowel data is strictly necessary. For many, this comes down to what the research question is, trust in the force-aligned output, or more practical considerations (e.g., time). There are reasonable arguments for and against manually checking vowel alignments for accuracy. While some work demonstrates that force-aligned output produces inconsistent boundaries between vowels and sonorants (Strelluf 2016; Gonzalez, Grama, & Travis 2020), other work suggests that, given enough data, correcting force-aligned output only marginally improves formant measurements (Labov, Rosenfelder, & Fruehwald 2013:37–38; Gonzalez & Docherty 2018). However, much of the work that evaluates taking formant measurements from uncorrected data assesses the observed improvement in static vowel measures (e.g., single F1/F2 measurements for each vowel). Complications can arise if more measurements are taken from single vowels, as forced aligners are not as good at segmentation as humans are, despite doing the work much faster (Fromont & Watson 2016). In Grama (2015), there was the added complication of legacy data, where recordings were not performed in ideal scenarios, or where audio degradation impacted the quality of recordings prior to digitization. With such data, there is a higher risk that completely automated methods could yield spurious measurements. A middle ground is to check a random sample of vowel tokens, particularly those that occur in phonological contexts that disproportionately impact alignment (see Gonzalez, Grama, and Travis 2020). In my view, claims about the behavior of vowel trajectories necessitate more accurate alignments, given these concerns.

Given that the original study utilized legacy data and investigated vowels both statically and dynamically, it was prudent to take a more conservative approach and manually check all the force-aligned output. Each vowel was checked by hand and boundaries were corrected if necessary; then, the vowel was tagged with its appropriate lexical set (cf. Wells 1982). As is typical of sociophonetic studies, only vowels in stressed content words were prepared for analysis. Boundary correction was done using a strict set of criteria, detailed in sections 4.3.1–4.3.3, and boundaries were placed at, or as close to a zero crossing in the waveform as possible. If the cues discussed in the following sections were not available, the token was excluded.

4.3.1 Obstruents For vowels preceded by stops, the burst and aspiration were included in the consonant segment, not the vowel segment. Vowel onsets were marked where the waveform indicated periodicity and the spectrogram showed clear formant structure (e.g., rising F1). A perceptible decrease in amplitude served as an additional cue for vowels adjacent to voiced stops. Boundaries for vowels adjacent to voiceless fricatives were placed where formant structure was clear, and where the cessation of aperiodic energy coincided with a change in amplitude.

4.3.2 Sonorants For vowels bordering nasals and laterals, boundaries were placed where decreased amplitude coincided with formant dampening or a lowered F1. Boundaries between vowels and pre-vocalic /r/ were marked at maximum F3, or F2, if F3 was unavailable. Where post-vocalic /r/ was present, boundaries were placed where a dip in F3 indicated oral closure consistent with /r/ articulation (see, e.g., Johnson 2012:140). Along with amplitude, boundaries for /w/ were placed where F1 and F2 began to diverge, and boundaries for /j/ were evaluated based on where F2 and F3 began to diverge.

4.3.3 Word-initial and word-final vowels Vowel-initial words at the beginning of an utterance were often bordered by silence or glottal closure. The starting point of the vowel was therefore marked as the first relatively high-amplitude vocal pulse evident in the waveform. Vowel-final words at the end of an utterance were marked at the last high-amplitude vocal pulse evident in the waveform. In both cases, Praat's ability to track the formant structure factored into boundary placement.

4.4 Extracting formant values

Vowel checking and formant extraction was done for each speaker and vowel independently. Before extraction, a 15% subsample of each speaker's vowel category was checked to ensure that settings in Praat accurately captured formant behavior. While it is commonplace to use a standard set of values for vowel extraction

(e.g., five formants under 5,500 hertz for females), this was not possible in every case given the aforementioned issues regarding recording quality. Therefore, manual assignment of formant settings during formant extraction was necessary. The specific extraction values supplied to the Praat script can be found in Grama (2015:296–307).

The Praat script used for this study was based on a script written by Mietta Lennes, which had been modified by Abby Walker and Katie Drager.[8] This script extracted vowel identity, the word in which the vowel appeared, the preceding and following phonological segments, the duration of the vowel, the fundamental frequency, and readings of the first three formants—F1, F2 and F3—from seven equidistant points, starting from 20% of the duration through the vowel and terminating at 80%. Speech rate was assessed using de Jong and Wempe's (2009) Praat script, which calculates speech rate as a function of the number of amplitude peaks over the duration of an interval. This process yielded two different types of data for each vowel: midpoint data, which were useful for assessing the overall picture of the vowel space and the target position of monophthongs, and transition data, which were useful for quantifying formant contours over the vowel's duration.

4.5 Cleaning the data

Speaker demographic information, along with formant and speech rate measurements were compiled into a single .csv file using R (R Core Team 2018) in wide format (i.e., where each row in the data frame corresponds to a single token). This yielded 11,551 vowel tokens over fourteen vowel categories across thirty-two speakers. Formant measurements were checked in R over the duration of the vowel to ensure that accurate readings were taken by the Praat script. Radical deviations from expected patterns (e.g., an F2 in /u/ of 1,200 hertz at 30% of the vowel, followed by an F2 of 500 hertz at 40%) were treated as spurious measurements, and resulted in the exclusion of the vowel token. Each speaker's vowel space was plotted and checked for outliers. Tokens with formant measurements that fell outside the range of plausible adult physiological limits (e.g., an F1 of 2,000 hertz or an F2 of 300 hertz) were removed. Tokens that were phonologically unlikely (e.g., /i/ in the low back area of the vowel space) were evaluated in Praat and removed if they were judged to be measurement errors. Finally, a script was written to remove tokens whose formant measurements fell more than three standard deviations outside of a speaker's vowel distribution, calculated within speaker, within vowel (cf. Hughes 2014, discussed in Foulkes et al. 2018). Figure 17.2 presents an example of a speaker's /i/ and /æ/ before and after filtering.

A total of 353 vowel tokens were removed following these processes, yielding 11,198 tokens for analysis. A breakdown of the number of vowels per vowel class can be seen in table 17.3.

4.6 Normalizing vowel formants

Even when thoroughly cleaned, raw data are not usually the focus of direct analysis at the group level (cf. Han, chapter 6, this volume). Normalization is often a necessary step to interpreting patterns in vowel data. Vowel normalization seeks to neutralize differences between

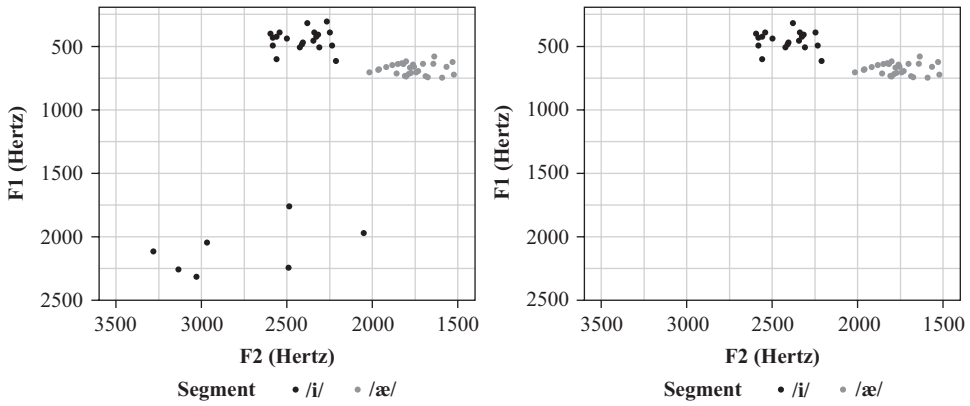


Figure 17.2
Raw formant measurements of /i/ (dark) from Eddie, a young 1970s man, before (left panel) and after (right panel) outlier filtering; /æ/ (light) included for reference.

Table 17.3
Distribution of vowel tokens in Grama (2015)

Vowel	*n*	Vowel	*N*
/i/	1,053	/u/	731
/ɪ/	1,093	/ʊ/	380
/e/	1,037	/o/	978
/ɛ/	1,158	/ʌ/	798
/æ/	1,154	/ɔ/	552
/aɪ/	899	/ɑ/	854
/aʊ/	412	/oɪ/	99

speakers that result from differences in physiology. For sociophoneticians, this typically means accounting for variation that stems from vocal tract length, a vital determiner of formant values; performing vowel normalization in this fashion means that any observed variation can be confidently ascribed to other factors.[9]

When and how to implement normalization is something of an ongoing discussion. If comparisons are made across groups (e.g., females to males, children to adults), normalizing vowels is uncontroversial (but consider methodological challenges pertaining to the automated vowel analysis of non-binary speakers discussed in Miles-Hercules & Zimman 2019), but normalization is typically unnecessary when investigating individual vowel spaces. Some argue against the need for normalization if comparisons are kept within group (e.g., males are compared to other males), pointing out that normalization warps the vowel space in such a way as to remove real patterns that emerge from the data (see Watson & Harrington 1999). Proponents of normalization argue that analyzing raw values is anticonservative, and that assuming the comparability of raw formant frequencies even across speakers who share similar physiologies leads to uninterpretable data (cf. Watt, Fabricius, & Kendall 2010). Given that comparisons across speakers is precisely the goal of many sociophonetic studies, this is an especially vital point to consider. Practically speaking, my view is that normalizing is a relatively low-cost step to ensure the veracity of observed patterns, and it is often the case that well-normalized data bear considerable resemblance to unnormalized data.

The question of what normalization method is most appropriate is also a somewhat thorny issue in sociophonetics, and there are dozens of algorithms to choose from. Work that directly compares the efficacy of normalization techniques (e.g., Adank, Smits, & Van Hout 2004; Flynn & Foulkes 2011) tends to find that methods that are vowel-extrinsic, formant-intrinsic, and speaker-intrinsic are best at reducing variation that arises as a result of physiology, while preserving sociolinguistic variation. Vowel-intrinsic methods perform comparably worse by these same metrics. However, any vowel normalization technique shows improvement over raw hertz comparisons (Flynn & Foulkes 2011:686). A popular normalization choice is the Lobanov method (Lobanov 1971; but see Barreda & Nearey 2018), as it is vowel-extrinsic and produces interpretable plots.

The vowel data in this study were normalized using the Lobanov method, which requires that relatively equal samples of all (monophthongal) vowel categories be included to avoid artificial skewing. Lobanov converts raw hertz values to normalized *z*-scores by subtracting a speaker's mean formant frequency (μ_i) from a raw measurement (F_i), and then diving this by the standard deviation for that speaker's formant (σ_i), as in equation (17.1).

$$F_i^N = \frac{F_i - \mu_i}{\sigma_i} \tag{17.1}$$

Because Lobanov produces values that are centered on (0, 0), it is not uncommon to scale these values back to hertz (see, e.g., Labov, Rosenfelder, & Fruehwald 2013:36), though this should only be performed after all values have been normalized. While easily performed manually or via script, normalization can also be performed using the *vowels* package (Kendall & Thomas 2018) in R, as can the Bark difference (Traunmüller 1997), ANAE (Labov, Ash, & Boberg 2006), Nearey (1977), and Watt and Fabricius (2002) methods.

5 Reporting, storing, and sharing the data

5.1 Reporting data

Transparent, clear reporting of data is an important aspect of any sociophonetic study. To this end, I include the raw formant values measured along the vowel's duration for all vowels across speakers (Grama 2015:308–337), as well as the normalized formant values across vowel, age, and sex (338–345). Moreover, clear figures and statistical analyses are key to achieving interpretable and reproducible findings. R is a powerful option for graphics creation

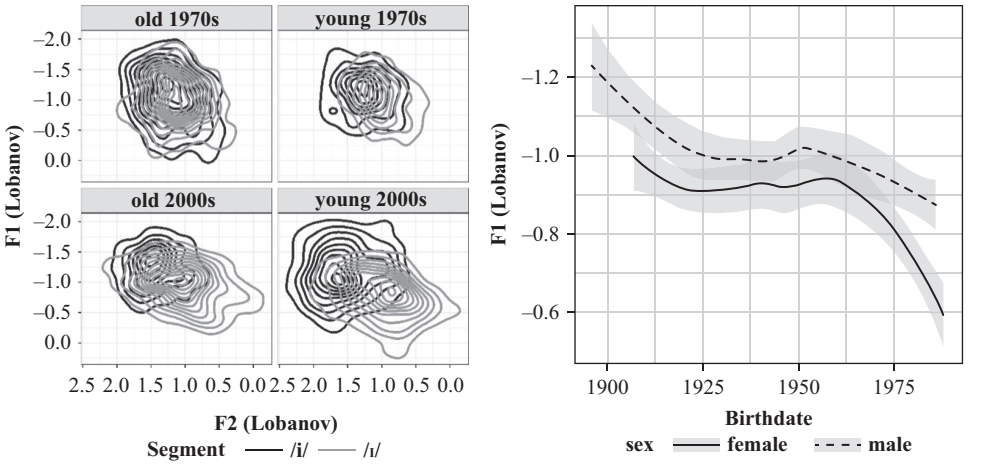


Figure 17.3
(Left) Two-dimensional kernel density plot of /i/ (dark) and /ɪ/ (light) over time across corpora and relative age. (Right) Local polynomial regression of /ɪ/ over birthdate for males (dashed line) and females (solid line).

and fitting statistical models. For graphing, ggplot2 (Wickham et al. 2019) is a popular and highly customizable option; for fitting statistical models, lme4 (Bates et al. 2019) allows the user to fit mixed models, and both lmerTest (Kuznetsova, Brockhoff, & Bojesen 2019) and pbkrtest (Halekoh & Højsgaard 2017) allow the user to derive *p*-values and interpretable model outputs from lme4.[10]

The original study strove to achieve clarity and statistical accountability. For example, the two-dimensional kernel density plots in the left panel of figure 17.3 show the distribution and overall shape of the data, giving the reader an understanding of the distribution of the overall vowel categories; this aids the interpretability of the plot. However, not every plot was equally well conceived. The right panel in figure 17.3 shows a localized polynomial regression with standard errors for the F1 of /ɪ/ over time. The regression indicates a sex difference; however, the figure lacks the data points on which the regression is based. Therefore, it is unclear how closely the regression fits the data and whether the data points vary consistently across the observed effect. A solution to this would be to include the individual data points in the plot in the right panel.

5.2 Storing and sharing the data

SOLIS served as the repository for transcribed interviews and the boundaries produced by forced alignment. Using this server allowed me to store raw data used in the original study in a password-protected online repository as I progressed through the collections. While this minimized the risk of data loss, the rest of my data (e.g., individual formant measurements, R scripts, and statistical models) is on personal repositories. And while formant values are reported in Grama (2015), no officially archived data sheet exists. This is a key area where the original study was not in line with best practices (see Andreassen, chapter 7, this volume). One way to solve this issue would be to make further use of SOLIS. Because SOLIS uses LaBB-CAT architecture, it allows for detailed annotation stores. Formant values and other information could be uploaded to the server to allow the precise study to be replicated by anyone with access. In addition, manually checked Praat TextGrids could be uploaded to SOLIS to enrich the files currently housed on SOLIS. This would also constitute a step toward sharing the data, as access is, as of writing, restricted to SOLIS users or achieved through personal requests to me.

6 Conclusion

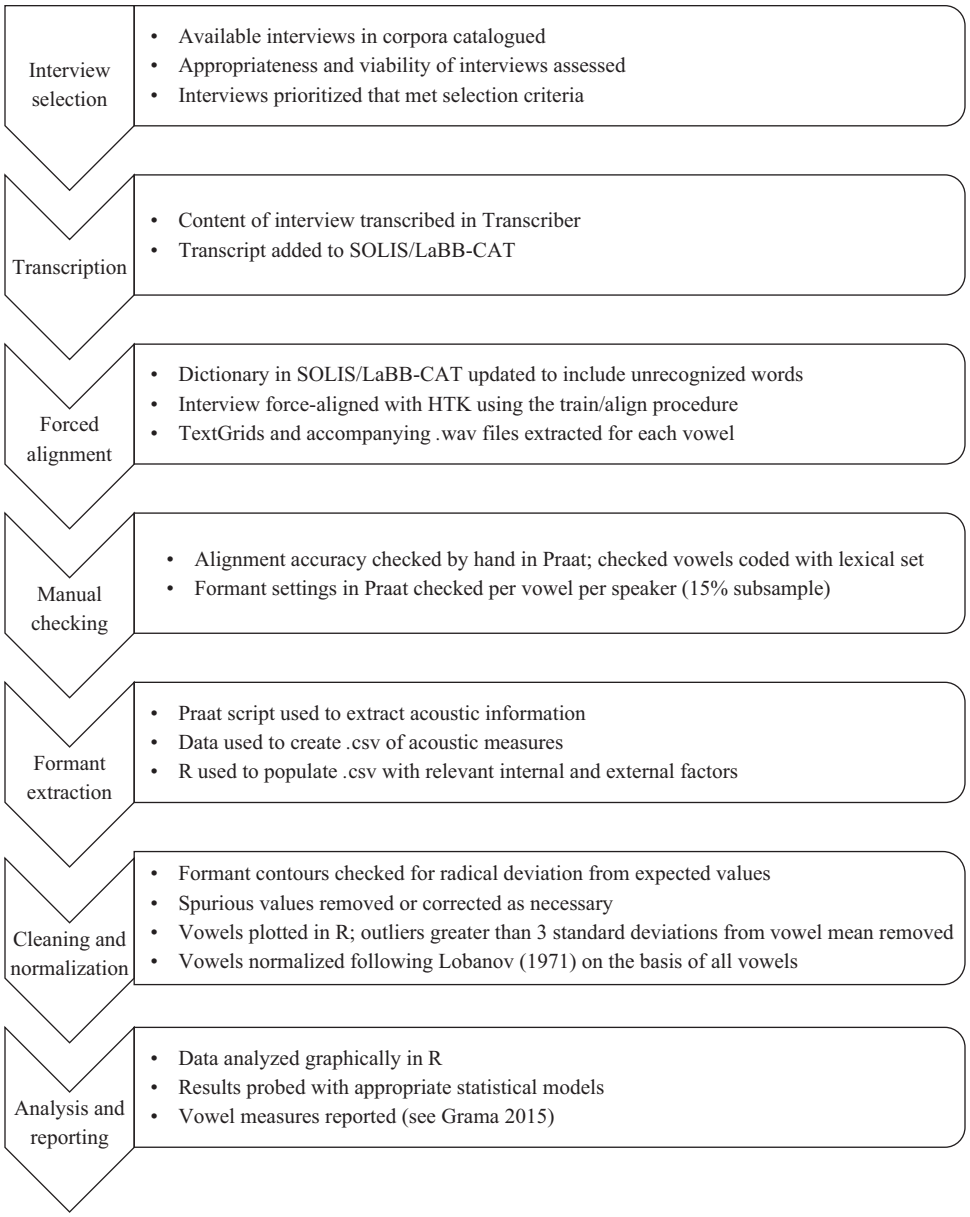
This use case serves as a guide for those wishing to undertake similar longitudinal studies of language variation and change, especially where legacy data are required. As a short-hand reference, a schematized version of the methodological steps detailed in this chapter is available in appendix B. As a final caveat, this methodology was employed between 2014 and 2015, meaning that there are aspects of it that may not stand the test of time. Nevertheless, the steps for processing the data discussed here, many of which are maximally conservative, should remain relevant in the future.

Appendix A: Detailed speaker demographic information in Grama (2015)

Table 17.4
Speaker demographic and social information in Grama (2015), including the collection file name, the corpus, and the speaker's relative age, assigned pseudonym, binary sex, age, date of birth, island of residence, listed ethnicity, highest education achieved, and occupation information

Collection file name	Corpus and relative age	Speaker pseudonym	Sex	Age	Date of birth	Island	Ethnicity	Highest education	Occupation
DB1–072	Old 1970s	Kawika	M	79	1896	Kauaʻi	Hawaiian	na	Retired motel owner
DB1–122	Old 1970s	Joseph	M	69	1906	Hawaiʻi	Portuguese	No high school	Retired plantation worker
CS1-JA-03	Old 1970s	Manny	M	58	1922	Hawaiʻi	Filipino	High school	Farmer, real estate
DB1–164	Old 1970s	Kimo	M	54	1921	Oʻahu	Part Hawaiian	High school	Retired roofer, plantation
DB1–073	Old 1970s	Miki	F	68	1907	Kauaʻi	Japanese	High school	Retired barber
DB1–066	Old 1970s	Malia	F	64	1911	Kauaʻi	Hawaiian	High school	Housewife
DB1–162	Old 1970s	Kaimana	F	57	1918	Oʻahu	Hawaiian, Haole	High school	Retired
DB1–056	Old 1970s	Keiko	F	55	1918	Kauaʻi	Japanese	High school	Home management
DB1–165	Young 1970s	Eddie	M	39	1936	Oʻahu	Part Hawaiian	High school	Construction worker
DB1–059	Young 1970s	Victor	M	37	1938	Kauaʻi	Portuguese	High school	NA
CS1-EA-03	Young 1970s	Danny	M	30	1942	Oʻahu	Filipino	High school	Floorer
CS1-GN-02	Young 1970s	Glen	M	25	1944	Hawaiʻi	Japanese	High school	Contract laborer
DB1–074	Young 1970s	Mona Lisa	F	48	1927	Kauaʻi	Filipino	High school	NA
DB1–075	Young 1970s	Leilani	F	42	1933	Kauaʻi	Hawaiian	High school	Housewife, retired
DB1–120	Young 1970s	Delia Jane	F	35	1940	Hawaiʻi	Filipino	High school	Adult education instructor
DB1–065	Young 1970s	Teresa	F	35	1940	Kauaʻi	Filipino	College	Air national guard
CS2–053	Old 2010s	Grant	M	56	1951	Oʻahu	Japanese	College	Government worker
CS2–029	Old 2010s	Kevin	M	52	1955	Hawaiʻi	Hawaiian	NA	Unemployed, ex-military/farmer
CS2–017	Old 2010s	Palani	M	44	1963	Hawaiʻi	Part Hawaiian	NA	Shop owner
CS2–030	Old 2010s	Keoni	M	40	1967	Hawaiʻi	Part Hawaiian	High school	NA
CS2–037	Old 2010s	Pua	F	58	1949	Oʻahu	Part Hawaiian	High school	NA
CS2–040	Old 2010s	Lani	F	49	1958	Oʻahu	Part Hawaiian	High school	Housewife
CS2–027	Old 2010s	Carla	F	46	1961	Hawaiʻi	Portuguese	High school	Unemployed
CS2–011	Old 2010s	Kahea	F	42	1965	Kauaʻi	Part Hawaiian	High school	Ranch worker
CS2–052	Young 2010s	Kaleo	M	22	1985	Maui	Hawaiian, Korean, Haole	College	Student
CS2–056	Young 2010s	Myko	M	22	1985	Kauaʻi	Portuguese	College	Student
CS2–046	Young 2010s	Eric	M	21	1986	Hawaiʻi	Chinese, Filipino	College	Student
CS2–051	Young 2010s	Alika	M	21	1986	Hawaiʻi	Japanese	College	Student
CS2–048	Young 2010s	Sarah	F	24	1983	Oʻahu	Chinese	College	MA Student
CS2–019	Young 2010s	Starla	F	23	1984	Hawaiʻi	Hawaiian, Chinese, Japanese	High school	NA
CS2–055	Young 2010s	Mina	F	21	1986	Kauaʻi	Japanese, Haole, Chinese, Hawaiian	College	Student
RK01	Young 2010s	Lena	F	19	1988	Kauaʻi	Filipino, Japanese	College	Student

F = female; M = male; NA = not available.



Appendix B: Data workflow for Grama (2015)

The chart schematizes the workflow for Grama (2015) from interview selection to analysis. It is meant as a short-hand reference for those wishing to replicate the methodology of this study. The transcription software and forced aligner listed could easily be altered to accommodate researcher preference.

Notes

1. This was especially true because time and financial considerations made hiring research assistants impractical during the completion of the original study.

2. One aim of Grama (2015) was to quantify vowel variation along the creole continuum (cf. DeCamp 1971; Sato 1993), a focus that falls outside the scope of the present chapter.

3. Metadata were missing or inconsistent across recordings; it is a researcher's duty to report these gaps, while being as vigilant as possible in our own studies to record metadata as thoroughly and representatively as possible (cf. Mattern, chapter 5, this volume). In some cases, accompanying corpus notes were mined for missing information to account for gaps in the metadata. When information crucial to the test variables was absent, the interview was excluded.

4. Recordings were excluded if they were shorter than ten minutes on these grounds.

5. Space does not allow for a fuller description of the range of factors that might govern the selection of a forced aligner; however, aligner choice should be based on the researcher's needs, as well as aligner performance and suitability. For a more thorough discussion of these factors, I direct the reader to Fromont and Watson (2016), McAuliffe et al. (2017), MacKenzie and Turton (2020), and Gonzalez, Grama, and Travis (2020).

6. The steps described here would be identical to those were the reader to use a local build of LaBB-CAT.

7. At the time, CELEX seemed preferable to the CMU Pronouncing Dictionary (the latter of which is optimized for US varieties), given the variable absence of post-vocalic /r/ in Pidgin. However, recent work demonstrates that automated alignment is not strongly affected by variety if the variety under study is not markedly different from the training variety (MacKenzie & Turton 2020), or if vowel coercion is performed (Gonzalez, Grama, & Travis 2020).

8. The original script is collect_formant_data_from_files.praat, available at https://github.com/lennes.

9. Some phoneticians (e.g., Rosner & Pickering 1994) argue that normalization is done to model cognitive processes that underpin vowel perception, though there is some debate in the phonetics literature as to whether listeners normalize vowels at all (see, e.g., Pisoni 1997). This chapter does not weigh in on this debate.

10. Because deriving *p*-values from linear and logistic mixed-effect models is not straightforward, I recommend consulting Luke (2017) for best practices.

References

Adank, Patti, Roel Smits, and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116:3099–3107. https://doi.org/10.1121/1.1795335.

Albarillo, Emily E., and Nick Thieberger. 2009. Kaipuleohone, University of Hawai'i's digital ethnographic archive. *Language Documentation and Conservation* 3 (1): 1–14. http://hdl.handle.net/10125/4422.

Baayen, R. H., R. Piepenbroock, and L. Gulikers. 1995. *The CELEX Lexical Database* (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bailey, Guy. 2004. Real and apparent time. In *The Handbook of Language Variation and Change*, ed. J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 312–332. Malden, MA: Blackwell. https://doi.org/10.1002/9780470756591.ch12.

Barras, Claude, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication* 33 (1–2): 5–22. https://doi.org/10.1016/S0167-6393(00)00067-4.

Barreda, Santiago, and Terrance M. Nearey. 2018. A regression approach to vowel normalization for missing and unbalanced data. *Journal of the Acoustical Society of America* 144 (1): 500–520. https://doi.org/10.1121/1.5047742.

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, and Gabor Grothendieck. 2019. lme4: Linear mixed-effects models using "Eigen" and S4. R package version 1.1–21. https://github.com/lme4/lme4/.

Berez, Andrea L. 2013. The digital archiving of endangered language oral traditions: Kaipuleohone at the University of Hawai'i at C'ek'aedi Hwnaz in Alaska. *Oral Tradition* 28 (2): 261–270. doi:10.1353/ort.2013.0010.

Bickerton, Derek. 1976. *Change and Variation in Hawaiian English: Final Report on National Science Foundation Grant no. GS-39748*. Honolulu: Social Sciences and Linguistics Institute, University of Hawaii.

Boersma, Paul, and David Weenink. 2019. Praat: Doing phonetics by computer (computer program). Version 6.1. http://www.praat.org/.

Collerson, Kenneth D., and Marshall I. Weisler. 2007. Stone adze compositions and the extent of ancient Polynesian voyaging and trade. *Science* 317 (5846): 1907–1911. doi:10.1126/science.1147013.

DeCamp, David. 1971. Towards a generative analysis of post-creole speech continuum. In *Pidginization and Creolization of Languages*, ed. Dell Hymes, 349–370. Cambridge: Cambridge University Press.

de Jong, Nivja H., and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41 (2): 385–390. doi:10.3758/BRM.41.2.385.

Drager, Katie. 2012a. New directions in sociolinguistic methods. Paper presented at the *Linguistic Data Consortium 20th Anniversary Workshop*, Philadelphia, September 6–7.

Drager, Katie. 2012b. Pidgin and Hawai'i English: An overview. *Journal of Language, Translation and Intercultural Communication* 1 (1): 61–73. http://dx.doi.org/10.12681/ijltic.10.

Drager, Katie, and James Grama. 2014. "De tawk dakain ova dea": Mapping language ideologies on O'ahu. *Dialectologia* 12 (12): 23–51.

Flege, James E., Carlo Shirru, and Ian R. A. MacKay. 2003. Interaction between the native and second language phonetic subsystems. *Speech Communication* 40:467–491.

Flynn, Nicholas, and Paul Foulkes. 2011. Comparing vowel formant normalization methods. In *Proceedings of the 17th ICPhS*, ed. W. S. Lee and E. Zee, 683–686. Hong Kong: City University of Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Flynn/Flynn.pdf.

Foulkes, Paul, Gerry Docherty, Stefanie Shattuck Hufnagel, and Vincent Hughes. 2018. Three steps forward for predictability: Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. *Linguistics Vanguard* 42 (2): 1–11. https://doi.org/10.1515/lingvan-2017-0032.

Fromont, Robert, and Jennifer Hay. 2012. LaBB-CAT: An annotation store. In *Proceedings of the Australasian Language Technology Workshop*, 113–117. Otago University, December 4–6. https://www.aclweb.org/anthology/U12-1015.pdf.

Fromont, Robert, and Kevin Watson. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora* 11 (3): 401–431. https://doi.org/10.3366/cor.2016.0101.

Fujikane, Candace. 1997. Reimagining development and the Local in Lois-Ann Yamanaka's "Saturday Night at the Pahala Theater." *Social Process in Hawai'i* 38:40–61.

Gonzalez, Simon, and Gerry Docherty. 2018. Measuring a stabilization point of uncorrected forced-aligned data. Paper presented at *SocioPhonAus2*, Brisbane, Australia, July 16–17.

Gonzalez, Simon, James Grama, and Catherine Travis. 2020. Comparing the performance of major forced aligners used in sociophonetic research. *Linguistics Vanguard* 6 (1). https://doi.org/10.1515/lingvan-2019-0058.

Grama, James. 2015. Variation and change in Hawai'i Creole vowels. PhD dissertation, University of Hawai'i at Mānoa.

Halekoh, Ulrich, and Søren Højsgaard. 2017. pbkrtest: Parametric bootstrap and Kenward-Roger based methods for mixed model comparison. R package version 0.4–7. https://myaseen208.github.io/pbkrtest/.

Hughes, Vincent. 2014. The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison. PhD dissertation, University of York.

Johnson, Keith. 2012. *Acoustic and Auditory Phonetics,* 3rd ed. Malden, MA: Wiley-Blackwell.

Kawamoto, Kevin Y. 1993. Hegemony and language politics in Hawai'i. *World Englishes* 12 (2): 193–207. https://doi.org/10.1111/j.1467-971X.1993.tb00021.x.

Kendall, Tyler, and Erik R. Thomas. 2018. Vowels: Vowel manipulation, normalization, and plotting. R package, version 1.2–2. https://cran.r-project.org/web/packages/vowels/.

Kraus, Janina. 2017. A sociophonetic study of the urban Bahamian Creole vowel system. PhD, dissertation, Ludwig-Maximilians-Universität München.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen. 2019. lmerTest: Tests in linear mixed effects models. R package version 3.1–0. https://cran.r-project.org/web/packages/lmerTest/index.html.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press. https://doi.org/10.1017/S0047404500004528.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English*. New York: Mouton de Gruyter.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89 (1): 30–65. https://doi.org/10.1353/lan.2013.0015.

Lausberg, Hedda, and Han Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, and Computers* 41 (3): 841–849. doi:10.3758/BRM.41.3.841.

Lesho, Marivic. 2014. The sociophonetics and phonology of the Cavite Chabacano vowel system. PhD dissertation, Ohio State University.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49 (2): 606–608. https://doi.org/10.1121/1.1912396.

Luke, Steven G. 2017. Evaluating significance in linear mixed-effects models in R. *Behaviour Research Methods* 46 (4): 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.

MacKenzie, Laurel, and Daniel Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6 (1). https://doi.org/10.1515/lingvan-2018-0061.

Marlow, Mikaela L., and Howard Giles. 2008. Who you tink you, talkin propah? Hawaiian Pidgin demarginalized. *Journal of Multicultural Discourses* 3 (1): 53–68. https://doi.org/10.1080/17447140802153535.

Marlow, Mikaela L., and Howard Giles. 2010. "We won't get ahead speaking like that!" Expressing and managing language criticism in Hawai'i. *Journal of Multilingual and Multicultural Development* 31 (3): 237–251. https://doi.org/10.1080/01434630903582714.

McAuliffe, Michaela, Michael Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sondregger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502. doi:10.21437/Interspeech.2017-1386.

Miles-Hercules, DeAndre, and Lal Zimman. 2019. Normativity in normalization: Methodological challenges in the (automated) analysis of vowels among non-binary speakers. Paper presented at New Ways of Analyzing Variation 48 (NWAV48), Eugene, OR, October 10–12.

Nearey, Terrance M. 1977. Phonetic features system for vowels. PhD dissertation, University of Alberta. (Reprinted 1978 by the Indiana University Linguistics Club.)

Odo, Carol. 1975. Phonological processes in the English dialect of Hawai'i. PhD, dissertation, University of Hawai'i at Mānoa.

Odo, Carol. 1977. Phonological representations in Hawaiian English. *University of Hawai'i Working Papers in Linguistics* 9 (3): 77–85.

Ohnuma, Keiko. 2002. Local Haole—a contradiction of terms? The dilemma of being white, born and raised in Hawai'i. *Cultural Values* 6 (3): 273–285. https://doi.org/10.1080/1362517022000007211.

Pisoni, David B. 1997. Some thoughts on "normalization" in speech perception. In *Talker Variability in Speech Processing*, ed. Keith Johnson and John W. Mullennix, 9–32. San Diego: Academic Press.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Reinecke, John R. 1969. *Language and Dialect in Hawaii*. Honolulu: University of Hawai'i Press.

Roberts, Sarah J. 2004. The emergence of Hawai'i Creole English in the early 20th century: The sociohistorical context of creole genesis. PhD dissertation, Stanford University.

Rohrer, Judy. 1997. Haole girl: Identity and white privilege in Hawai'i. *Social Process in Hawai'i* 38:138–161.

Rosenfelder, Ingrid. 2009. Sociophonetic variation in educated Jamaican English: An analysis of the spoken corpora of ICE-Jamaica. PhD dissertation, University of Freiburg, Germany.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite. Version 1.2.2. https://zenodo.org/record/9846#.X9hzldhKiUk.

Rosner, B. S., and J. B. Pickering. 1994. *Vowel Perception and Production*. Oxford: Oxford University Press.

Sabino, Robin. 1996. A peak at death: Assessing continuity and change in an underdocumented language. *Language Variation and Change* 8 (1): 41–61. https://doi.org/10.1017/S095439450000106X.

Sabino, Robin. 2012. *Language Contact in the Danish West Indies: Giving Jack His Jacket*. Leiden: Brill Publishers. https://doi.org/10.1163/9789004230705.

Sakoda, Kent, and Jeff Siegel. 2003. *Pidgin Grammar: An Introduction to the Creole English of Hawai'i*. Honolulu: Bess Press, Inc.

Sakoda, Kent, and Jeff Siegel. 2008. Hawai'i Creole: Phonology. In *A Handbook of Varieties of English*. Volume I: *Phonology*, ed. Berndt Kortmann and Edgar W. Schneider, 729–749. Berlin: Mouton de Gruyter.

Sato, Charlene J. 1993. Language change in a creole continuum: Decreolization? In *Progression and Regression in Language: Sociocultural, Neuropsychological and Linguistic Perspectives*, ed. Kenneth Hyltenstam and Åke Viberg, 122–143. Cambridge: Cambridge University Press.

Schiel, Florian, Christian Draxler, and Jonathan Harrington. 2011. Phonemic segmentation and labelling using the MAUS technique. Paper presented at New Tools and Methods for Very-Large-Scale Phonetics Research Workshop, University of Pennsylvania, January 28–31.

Seifert, Frank, Nicholas Evans, Harald Hammarström, and Steven C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94 (4): e324–e345. doi:10.1353/lan.2018.0070.

Siegel, Jeff. 2007. Recent evidence against the Language Bioprogram Hypothesis: The pivotal case of Hawai'i Creole. *Studies in Language* 31 (1): 51–88. https://doi.org/10.1075/sl.31.1.03sie.

Stannard, David E. 1990. Disease and infertility: A new look at the demographic collapse of Native populations in the wake of Western contact. *Journal of American Studies* 24 (3): 325–350.

Strelluf, Christopher. 2016. Overlap among back vowels before /l/ in Kansas City. *Language Variation and Change* 28 (3): 379–407. https://doi.org/10.1017/S0954394516000144.

Stueber, Ralph K. 1964. Hawaii: A case study in development education 1778–1960. PhD dissertation, University of Wisconsin.

Traunmüller, Hartmut. 1997. Auditory scales of frequency representation. http://www.ling.su.se/staff/hartmut/bark.htm.

Veatch, Thomas C. 1991. English vowels: Their surface phonology and phonetic implementation in vernacular dialects. PhD dissertation, University of Pennsylvania.

Walworth, Mary. 2014. Eastern Polynesian: The linguistic evidence revisited. *Oceanic Linguistics* 53 (2): 256–272. https://doi.org/10.1353/ol.2014.0021.

Wassink, Alicia B. 1999. A sociophonetic analysis of Jamaican vowels. PhD dissertation, University of Michigan.

Wassink, Alicia B. 2001. Theme and variation in Jamaican vowels. *Language Variation and Change* 13 (2): 135–159.

Wassink, Alicia B. 2006. A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *Journal of the Acoustical Society of America* 119 (4): 2334–2350. https://doi.org/10.1121/1.2168414.

Watson, Catherine I., and Jonathan Harrington. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America* 106 (1): 458–468.

Watt, Dominic J., and Anne H. Fabricius. 2002. Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1~F2 plane. *Leeds Working Papers in Linguistics and Phonetics* 9 (9): 159–173.

Watt, Dominic J., Anne H. Fabricius, and Tyler Kendall. 2010. More on vowels: plotting and normalization. In *Sociophonetics: A Student's Guide,* ed. Marianna Di Paolo and Malcah Yaeger-Dror, 107–118. London: Routledge.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. ggplot2: Create elegant data visualisations using the grammar of graphics. R package version 3.2.0. https://cran.r-project.org/web/packages/ggplot2/index.html.

Wilmshurst, Janet M., Terry L Hunt, Carl P. Lipo, and Atholl J Anderson. 2011. High precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proceedings of the National Academy of Sciences* 108 (5): 1815–1820. https://doi.org/10.1073/pnas.1015876108.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, et al. 2009. *The HTK book (for version 3.4)*. Cambridge: Cambridge University Engineering Department.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363