Chapter 1

The SpiCE Corpus

1.1 Introduction

Most of our knowledge about spoken language and speech processing comes from monolingual individuals producing scripted speech in laboratory settings. Monolingual lab speech allows for researchers to exercise tight control over the linguistic backgrounds of the speakers and the linguistic material (e.g. reading or repeating sounds, words, or sentences). While highly informative, these controlled monolingual speech samples are not representative of spoken language in the real world. Multilingualism is the norm, not the exception, and individuals regularly make creative linguistic choices.

Crucially, corpus-based research with conversational or spontaneous speech is important in the fields of phonetics and psycholinguistics, as the research conclusions drawn from corpus and lab-based experiments do not always coincide (Gahl et al., 2012). Conversational speech allows for a more accurate empirical description of spoken language, as it represents more realistic and natural productions than scripted laboratory speech, even compared to scripted connected speech. Conversational speech also crucially permits for field testing of speech production theories (Bell et al., 2009; Gahl et al., 2012) in their natural habitats.

The discrepancies between results for conversational and lab speech have been found for monolingual (English) speech, but are likely be found with bilingual speech as well. Resources to query bilingual conversational speech are limited, however, as the necessary resources permitting this type of inquiry are rare. As a step towards filling this gap, this chapter introduces the **SpiCE** corpus of conversational bilingual **Sp**eech **i**n **C**antonese and **E**nglish. This open-access corpus was explicitly developed with phonetic bilingualism research in mind. The corpus design is based on key aspects of widely used existing corpora, such as the Buckeye corpus of conversational speech (Pitt et al., 2005). It crucially includes speech from the same individual in more than one language, as is the case in the Bangor corpora of Spanish-English, Welsh-English, and Welsh-Spanish bilingual speech (Deuchar et al., 2014), but with a more controlled recording setup, allowing for more nuanced acoustic-phonetic measurements.

The primary motivation for collecting this corpus was to have comparable high-quality recordings of conversational speech from early bilinguals in two languages, which in turn enables large scale phonetic analysis on a within-speaker basis.[1] To our knowledge, this type of resource does not yet exist for any pair of languages, much less for a typologically distinct pair like Cantonese (Sino-Tibetan) and English (Indo-European). Furthermore, Cantonese is a relatively understudied language, despite there being approximately 55 million native speakers around the world (Matthews et al., 2013), though this is changing with new corpora (Luke and Wong, 2015; Leung and Law, 2001; Winterstein et al., 2020; Alderete et al., 2019) and tools (Lee, 2018; Yau, 2019).

This chapter provides a detailed overview of the corpus design and collection procedures, a description of the speakers, and the transcription and annotation pipeline. It concludes with descriptive statistics.

1.2 Corpus design and creation

This section provides detail about the speakers (Section 1.2.1), the procedures used to ensure high-quality recordings (Section 1.2.2), and the three tasks that each participant completed in both Cantonese and English (Section 1.2.3).

Data collection took place between November 2018 and March 2020. Ortho-

[1] SpiCE is a large speech corpus for the type of speech involved, and is comparable in size to the widely used Buckeye corpus (Pitt et al., 2005). Larger speech corpora tend to comprise existing recordings of broadcasters, phone conversations, or read speech. While these types of speech corpora are certainly useful, they are designed for different purposes.

No.	ID	Order	Age	Gender	English AoA	Cantonese AoA
1	VF19A	E → C	19	F	0	0
2	VF19B	E → C	19	F	0	0
3	VF19C	E → C	19	F	3	0
4	VF19D	C → E	19	F	2	0
5	VF20A	C → E	20	F	4	0
6	VF20B	C → E	20	F	5	0
7	VF21A	E → C	21	F	0	0
8	VF21B	C → E	21	F	3	0
9	VF21C	C → E	21	F	4	0
10	VF21D	E → C	21	F	0	0
11	VF22A	C → E	22	F	0	0
12	VF23B	E → C	23	F	2	0
13	VF23C	C → E	23	F	0	0
14	VF26A	C → E	26	F	0	0
15	VF27A	E → C	27	F	0	0
16	VF32A	C → E	32	F	3	0
17	VF33B	C → E	33	F	0	0
18	VM19A	E → C	19	M	0	0
19	VM19B	C → E	19	M	2	0
20	VM19C	E → C	19	M	0	0
21	VM19D	C → E	18	M	1	1
22	VM20B	E → C	20	M	0	0
23	VM21A	E → C	21	M	0	0
24	VM21B	E → C	21	M	0	0
25	VM21C	C → E	21	M	0	0
26	VM21D	C → E	21	M	0	0
27	VM21E	C → E	21	M	5	0
28	VM22A	C → E	22	M	4	0
29	VM22B	E → C	22	M	0	0
30	VM23A	E → C	23	M	0	0
31	VM24A	E → C	24	M	3	0
32	VM25A	E → C	25	M	4	0
33	VM25B	E → C	25	M	0	0
34	VM34A	C → E	34	M	0	0

Table 1.1: Basic participant information, including age, gender, age of acquisition (AoA), and the order the interviews occurred.

graphic transcription began shortly after the first interview was recorded, and was completed in April 2021.

1.2.1 Participants

The recordings in this corpus will comprise the speech of 34 early Cantonese-English bilinguals, 17 of which are female. At the time of submission, ten of 17 male participants had been recorded. All participants were between the ages of 19 and 35 (inclusive), reported normal speech and hearing, and resided in Vancouver, Canada at the time of recording. All participants in the study completed an extensive language background questionnaire, which included questions about language background, proficiency, use, and general demographics. A summary of language background information for all recorded speakers in the corpus is provided in Table 1.1. A more comprehensive language background summary will be released along with the corpus audio and transcripts (see Section 1.5 for more details about releases).

Definitions of bilingualism are highly variable in the literature, as there are many different types of bilinguals (Amengual, 2017). For the purposes of this corpus, an early bilingual is someone who acquired both Cantonese and English before starting primary school (approximately age 5), and reports consistent use of both languages since that time. It is important to highlight that the Cantonese-English bilingual community in Vancouver (and Canada more generally) is incredibly diverse, both in terms of dialects spoken, as well as the regions from which families originally emigrated (Yu, 2013). Furthermore, given the prevalence of Cantonese in Vancouver (Statistics Canada, 2017), and longevity of the community (Yu, 2013), immigration from other Cantonese-speaking areas continues today.

This corpus reflects the diverse nature of Cantonese-English bilingualism in Vancouver, as it includes Canadian-born heritage speakers, recent immigrants from Hong Kong, as well as Cantonese speakers from other parts of the Cantonese diaspora. As a result, while all speakers are early bilinguals, various dialects are represented. The most well-represented dialect is Hong Kong Cantonese, as 20 of 27 participants report having at least one caretaker from Hong Kong (14 report only Hong Kong born caretakers).

1.2.2 Recording Setup

Recording took place in a quiet room in the linguistics laboratory building at the University of British Columbia in Vancouver, Canada. Two Cantonese-English bilingual research assistants (the third and fourth authors) and the participant were seated around a table. The interviewer was a female Cantonese-English bilingual from Metro Vancouver. The recording process was monitored by a male Cantonese-English bilingual from Hong Kong, who moved to Vancouver to attend university. The interviewer and participant were outfitted with AKG C520 head-mounted microphones positioned approximately 3 cm from the corner of the mouth. The microphones were connected to separate channels on a Sound Devices USBPre2 Portable Audio Interface. Stereo recordings were made with Audacity 2.2.2 (Audacity Team, 2018) on a PC laptop, and saved according to best archival practices, with a 44.1 kHz sampling rate, and 24-bit resolution.

1.2.3 Recording Procedure

Upon arrival, participants were provided with an overview of the recording session procedures, and informed of the corpus publication process. Subsequently, participants were asked to provide written consent. Upon consent, participants completed a session in English, and a session in Cantonese. The order of languages was counterbalanced across participants (see Table 1.1). Each session consisted of three tasks—sentence reading, storyboard narration, and a conversational interview—described in the following sections. Together, these three tasks took approximately 30 minutes in each language. Along with the consent process, and a break between interviews, participants spent approximately 90 minutes in the lab.

Sentence Reading

Participants first read the sentences listed in Table 1.3 and Table 1.2 aloud, pausing between sentences. Participants were not instructed to speak in a particular style. As participants had varying levels of Cantonese reading ability, they were simultaneously presented with both Cantonese characters and the Jyutping romanization.[2] If necessary, participants could make use of the phrase's English translation. The Cantonese sentences are well-known declarative phrases, typically associated with Chinese New Year. While a more explicitly balanced set of sentences could have been used, participants' familiarity was deemed more important, as many Cantonese-English bilinguals in Canada are not literate in Cantonese. The English sentences included the Harvard Sentences list number 60 (IEEE, 1969), as well as series of holiday-themed declarative sentences to better match the content of the Cantonese sentences. This task was relatively formal, and typically lasted less than one minute.

Sentence reading was included in the session to insure that different participants produced a set of identical items, considering the core of the session was unscripted conversational interview (described in Section 1.2.3). While these sentences do not exhaustively reflect the sound systems of Cantonese and English, they provide samples of identical items for all individuals, which is advantageous for future analyses or projects that require matched utterances.

Storyboard Narration

For the second task, participants narrated a short story from a cartoon storyboard in detail (Littell, 2010). The storyboard followed a simple plot about receiving gifts and writing thank you notes to family members and friends—a topic that Cantonese-English bilinguals in the corpus were expected to be familiar with in both languages. This task was less formal than the sentence reading task, and ensured that different participants produced some of the same words in a more spontaneous context. Similar to the sentences, these same words may be useful for future analyses or projects that require matched utterances. Participants narrated the same cartoon in each language, which ensured that some of the same content was conveyed in each language (e.g., productions of *mother* in both languages). It lasted 4–5 minutes, and allowed participants time to get used to the recording setup and helped them get into the right language mode before the interview. This is important, because language mode is known to affect the degree of crosslinguistic influence in speech production (Simonet and Amengual, 2019).

[2] Jyutping is one of the primary Cantonese romanization systems (Matthews et al., 2013), and is widely used in Cantonese corpus research (Nagy, 2011; Tse, 2019)

	English
1	Stop whistling and watch the boys march
2	Jerk the cord, and out tumbles the gold
3	Slide the tray across the glass top
4	The cloud moved in a stately way and was gone
5	Light maple makes for a swell room
6	Set the piece here and say nothing
7	Dull stories make her laugh
8	A stiff cord will do to fasten your shoe
9	Get the trust fund to the bank early
10	Choose between the high road and the low
11	Wish on every candle for your birthday
12	Deck the halls with boughs of holly
13	Ring in the new year with a kiss
14	Have a spooky Halloween
15	Enjoy the vacation with your loved ones
16	Be filled with joy and peace during this time
17	Relax on your holiday break

Table 1.2: Sentences 1–10 comprise the Harvard Sentences List 60. Sentences 11–17 are holiday-themed original imperatives, designed to thematically match the Cantonese sentences.

No.	Cantonese	Jyutping	English translation

Table 1.3: All Cantonese sentences are widely-known imperatives associated with Chinese New Year.

Conversational Interviews

The conversational interviews formed the bulk of the recording time for each participant, lasting around 25 minutes. Participants were informed of the general interview structure ahead of time. The casual interview format was inspired by the Buckeye corpus of conversational speech (Pitt et al., 2005), and included everyday topics such as family, school, culture, hobbies, and food. These topics were selected to be relevant, interesting, and encourage storytelling, but to not delve into the personal details typically elicited in a sociolinguistic interview (Nagy, 2011). A major goal was for participants—who knew they were being recorded for linguistic inquiry—to feel at ease and freely discuss the questions. Questions were loosely laid out under general topic headings, with optional follow-up questions. While the English and Cantonese interviews had the same structure and general topic areas, the particular questions differed. Furthermore, each interview took its own shape, and was guided by what the participant wanted to talk about, anywhere from three to six topic areas covered—the planned sequence of questions is included in the Appendix. As a result, the speech samples from each language are comparable, but the specific questions differ between interviews and across participants.

Participants were encouraged to code-switch between languages by the interviewer, who included code-switches in some of her questions, and asked about topics that encouraged switches (e.g., Chinese foods in English; university course work in Cantonese). While code-switching was encouraged, it was not a primary focus for the session. As will become apparent later in this chapter, there was substantially more code-switching in the Cantonese part of the session.

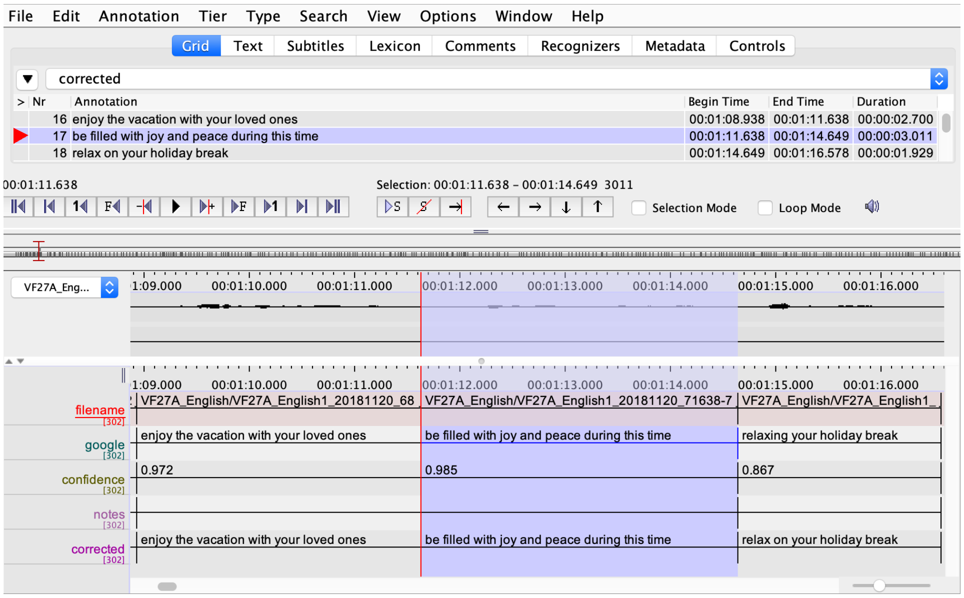
1.3 Annotation

All recordings were processed according to the pipeline outlined in this section. As much as possible, automatic tools were leveraged to expedite hand correction.

1.3.1 Cloud Speech-to-Text

Google Cloud Speech-to-Text was used to produce an initial transcript of the interviews (Google, 2019). This was done using the Short Audio option, with the language variety set to Canadian English (en-CA) or Hong Kong Cantonese (yue-Hant-HK). In order to use this speech recognition product, the participant's speech from the recordings was first segmented into short chunks, typically under 15 seconds in duration.[3] No attention was paid to constituents at this point; rather, breaks were placed at breaths and other pauses. Short chunks were necessary for speech recognition with locally stored files and desirable for transcribers in the subsequent

[3] The interviewer's speech is included in the recordings for the purpose of context, but is not transcribed.



corrected

Nr	Annotation	Begin Time	End Time	Duration
16	enjoy the vacation with your loved ones	00:01:08.938	00:01:11.638	00:00:02.700
17	be filled with joy and peace during this time	00:01:11.638	00:01:14.649	00:00:03.011
18	relax on your holiday break	00:01:14.649	00:01:16.578	00:00:01.929

00:01:11.638

Selection: 00:01:11.638 - 00:01:14.649 3011

	1:09.000 – 00:01:11.000	00:01:12.000 – 00:01:14.000	00:01:15.000 – 00:01:16.000
filename	VF27A_English/VF27A_English1_20181120_68	VF27A_English/VF27A_English1_20181120_71638-7	VF27A_English/VF27A_English1_
google	enjoy the vacation with your loved ones	be filled with joy and peace during this time	relaxing your holiday break
confidence	0.972	0.985	0.867
notes			
corrected	enjoy the vacation with your loved ones	be filled with joy and peace during this time	relax on your holiday break

Figure 1.1: This screenshot from ELAN showcases a sample of hand-corrected English from the sentence reading task for participant VF27A. The audio waveform is displayed in two channels, with one for the participant (top) and the other for the interviewer (bottom). The annotation tiers include (1) the short audio chunk's filename, (2) the raw speech-to-text transcript, (3) the speech-to-text confidence rating, (4) space for transcriber notes, if any, and (5) the corrected transcript. Note that "relaxing" was corrected to "relax on" in the rightmost section displayed.

hand correction phase. With the audio files prepared in this way, speech recognition was completed using the Python client library for Google Cloud Speech-to-Text. The output included both a transcript and a confidence rating for each audio chunk. While the transcripts generated in this fashion were far from perfect, they served the function of expediting the hand-correction process.

1.3.2 Orthographic Transcription Hand-Correction

The automatically generated transcripts were converted into multi-tiered ELAN transcription files (Sloetjes and Wittenburg, 2008), with tiers for the automatically generated transcript, phrase transcription confidence, notes, and corrected transcript. During hand-correction, research assistants adjusted the transcript in the corrected tier, and took note of anything pertinent to the given audio chunk. Figure 1.1 depicts an example of corrected English transcriptions in ELAN (Sloetjes and Wittenburg, 2008). Direct identifiers (e.g., names) were marked during this phase, and silenced from the recordings prior to release. Transcriber guidelines were adapted from the multilingual Heritage Language Variation and Change corpus, which includes Cantonese (Nagy, 2011). Guidelines for Cantonese were developed in collaboration with the bilingual research assistant team.

In both languages, the following conventions were used:

- The placeholder “xxx” denotes unintelligible speech.
- Fragments are transcribed using “&” followed by the fragment produced (e.g., “&s”).
- The “?” symbol marks questions; other punctuation is not used.

Cantonese-specific conventions include:

- Where possible, transcription is in characters.
- Words without a standard character are transcribed with Jyutping (e.g., *jyut6ping3*).
- Words produced in Mandarin Chinese are transcribed in Mandarin characters with “@m” appended to each.[4]

English-specific conventions include:

- Standard spelling is used.
- Proper nouns are capitalized and hyphenated if composed of multiple words (e.g., “British-Columbia”).
- Filled pauses are transcribed with “um”, “er”, “uh”, and other similar forms.
- Numbers are written out in word form (e.g., “one hundred”).

[4] Most participants report knowledge of Mandarin, though age of acquisition, use, and proficiency all vary drastically. In all cases, Mandarin was learned later than Cantonese and English.

1.3.3 Forced Alignment

Force-aligned transcripts were produced with the Montreal Forced Aligner (McAuliffe et al., 2017), using the hand-corrected orthographic transcripts and short audio chunks.

In Cantonese, forced alignment was completed with the Train-and-Align option, as there was no pretrained model available for Cantonese. The Cantonese pronunciation dictionary was generated using the *PyCantonese* Python library (Lee, 2018). As Cantonese orthpgraphy does not separate words with spaces, words segmentation was done according a maximum length matching algorithm... Pronunciations were identified by getting the Jyutping romanization from each character (or using the Jyutping transcribed), separating it into segments, and appending the tone number to the syllable nucleus (i.e., vowel or syllabic nasal). Research assistants supplemented the dictionary with alternative pronunciations for words that participated in syllable fusion. This approach bears some similarity to that of Tse (2019), but differs in that it also includes tonal information—which has been shown to improve forced alignment as long as there are not too many tone-nucleus combinations (Ćavar et al., 2016; Yuan et al., 2014).

Forced alignment in English took advantage of the Montreal Forced Aligner's pretrained English model and pronunciation dictionary, which broadly reflects North American English varieties. The dictionary was supplemented with manual additions.

The force-aligned transcripts were not manually corrected or checked. This means that any short chunk with code-switching or unintelligible speech will likely have poorer alignment. As a result, it is advisable to use stringent exclusionary criteria or perform checks prior to analyzing data from the corpus.

1.4 Descriptive Statistics

As hand-correction is currently underway, the descriptive statistics reported in the section are based on the Google Cloud Speech-to-Text transcripts of all three tasks (sentence reading, storyboard narration, interview) for the 27 participants listed in Table 1.1.

1.4.1 Cantonese Interviews

The Cantonese interviews include approximately 9.43 hours of participant speech.[5] There were a total of 1,836 character types, and 98,401 character tokens. The number of characters varies somewhat drastically by participant, with a mean of 3,514 characters per interview (SD=734, min=2,171, max=5,410). The overall distribution of character frequency in the Cantonese interviews is depicted in Figure 1.2. As expected, there are a relatively small number of characters occurring frequently (e.g. pronouns, function words, etc.), while a majority are mid and low frequency.

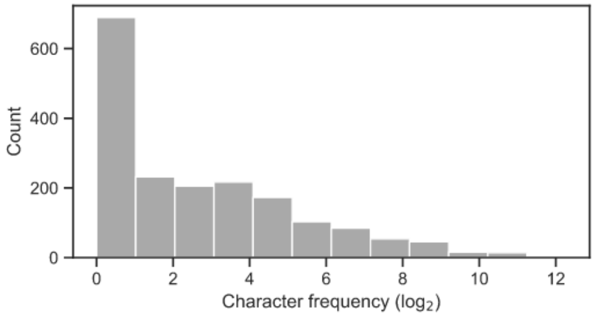


Figure 1.2: The distribution of log character frequency in the Cantonese interviews.

The decision to report descriptive statistics for Cantonese in characters rather than words arises from the difficulty of defining wordhood in Cantonese (Wong, 2006), a lack of tools for parsing written Cantonese,[6] and because written Cantonese does not include spaces between words. An approximation of the word count can be devised by calculating the average word length from the Hong Kong Cantonese Corpus (Luke and Wong, 2015)—1.3 characters. By this estimation, the Cantonese interviews here include approximately 75,115 words.

While Google Cloud Speech-to-Text primarily transcribes Cantonese in char-

[5] Note that this excludes the duration of the interviewer questions, as well as pauses in the participant's speech. Furthermore, with the addition of the remaining seven male participants, this will increase to approximately 12 hours

[6] The *PyCantonese* package is a notable exception to this (Lee, 2018), though it does not include tools for splitting sentences into words.

acters, it also inserts English text.[7] As a result, it is possible to get a rough idea of how much code-switching there is in the Cantonese interviews. There were 1,321 English word types in the Cantonese interviews, and 2,858 word tokens. This does not necessarily indicate the number of switches, as participants may have produced more than one English word in a row for a given switch. Nonetheless, it demonstrates that there is a substantial amount of code-switching from Cantonese to English. The mean number of English words in the Cantonese interviews is 102 (s.d.=46, min=30, max=198).

1.4.2 English Interviews

The English interviews include a total of 5,494 word types and 91,828 word tokens in 9.89 hours of participant speech.[8] As in the Cantonese interviews, the number of words varies substantially by participant, with a mean word count of 3,729 (s.d.=701, min=2,113, max=4,518). The distribution of log word frequency in the English interviews is portrayed in Figure 1.3. Word frequency follows a similar pattern to Cantonese character frequency, with most words occurring infrequently, and a smaller proportion occurring very frequently.

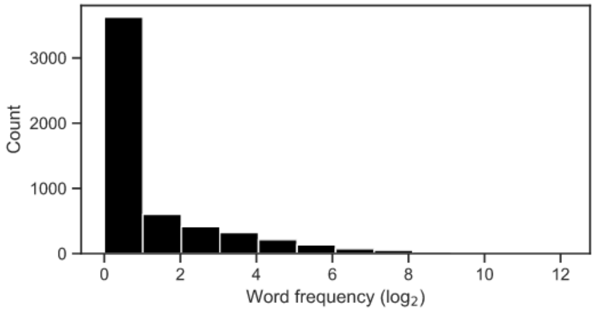


Figure 1.3: The distribution of log word frequency in the English interviews.

Unlike for the Cantonese interviews, it is not possible to get a sense of code-

[7] It is unclear at this point how accurate the English text in the Cantonese interviews is, though anecdotally speaking, it is accurate at least some of the time.

[8] As with the Cantonese interviews, this will increase to approximately 12 hours with the remaining participants.

switching in the English interviews, as Google Cloud Speech-to-Text for Canadian English does not insert Cantonese characters. Anecdotally, the interviewers reported that while most of the English interviews contained some code-switches, there was overall less English-to-Cantonese code-switching. This remains to be quantified when the orthographic hand-correction is completed.

1.5 SpiCE Corpus Releases

The SpiCE corpus will be made publicly available in a series of releases with increasing transcription accuracy and breadth of annotation. The first release is planned for mid-2020, and will include all audio files with accompanying Google Cloud Speech-to-Text transcriptions and a detailed language background questionnaire summary. The second release will include hand-corrected orthographic transcriptions, with force-aligned phonetic transcriptions. These files will be released as they are completed. Further releases will depend on resources, but would likely include phone-level hand correction and/or corrections for a particular subset of phones. All full releases will be described in the online documentation,[9] shared through the LRE Map,[10] and announced on the first author's website.

1.6 Discussion & Conclusion

While various bilingual corpora exist, they lack in different ways. The SpiCE corpus described here enables within-speaker phonetic comparisons across languages. While this would be possible with some of the bilingual speakers in resources like the Bangor corpora (Deuchar et al., 2014), the recording quality limits the scope of phonetic queries. With the release of SpiCE and its high-quality recordings, scholars have the ability to ask and answer empirically and theoretically motivated research questions within the speech and language sciences using more sophisticated phonetic measurement techniques (e.g., spectral measures, in addition to temporal measures). This offers substantial potential for increasing our understanding of bilingual spoken language from both phonetic and psycholinguistic perspectives. While the recording quality of this corpus offers these particular advantages,

[9] https://spice-corpus.readthedocs.io/

[10] http://lremap.elra.info

SpiCE is also suitable for any other standard corpus-based inquiry with conversational speech, whether linguistic or paralinguistic in nature. The opportunities made available with SpiCE are especially important given the typological difference between the languages under consideration, and the fact that Cantonese is an understudied language.

Bibliography

Alderete, J., Chan, Q., and Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, 191:103952. → page 2

Amengual, M. (2017). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *International Journal of Bilingualism*, 23(5):954–970. → page 4

Audacity Team (2018). Audacity (R): Free audio editor and recorder. → page 5

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111. → page 1

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., and Carter, D. (2014). Building bilingual corpora. In Thomas, E. M. and Mennen, I., editors, *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters. → pages 2, 14

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806. → page 1

Google (2019). Cloud speech-to-text. → page 8

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246. → page 6

Lee, J. L. (2018). PyCantonese [Version 2.2.0]. → pages 2, 11, 12

Leung, M.-T. and Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6(2):305–325. → page 2

Littell, P. (2010). Thank-you notes [Version 1.0: Agent focus]. → page 6

Luke, K. K. and Wong, M. L. Y. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics*, 25(2015):309–330. → pages 2, 12

Matthews, S., Yip, V., and Yip, V. (2013). *Cantonese: A Comprehensive Grammar*. Routledge. → pages 2, 6

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner [Version 1.0.1]. → page 11

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata*, 43(1-2):65–84. → pages 6, 8, 10

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95. → pages 2, 7

Simonet, M. and Amengual, M. (2019). Increased language co-activation leads to enhanced cross-linguistic phonetic convergence. *International Journal of Bilingualism*, 24(2):208–221. → page 6

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco. European Language Resources Association (ELRA). → pages 9, 10

Statistics Canada (2017). Proportion of mother tongue responses for various regions in Canada, 2016 Census. → page 4

Tse, H. (2019). *Beyond the Monolingual Core and out into the Wild: A Variationist Study of Early Bilingualism and Sound Change in Toronto Heritage Cantonese*. Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA. → pages 6, 11

Winterstein, G., Tang, C., and Lai, R. (2020). CantoMap: A Hong Kong Cantonese MapTask corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20)*, pages 2906–2913, Marseille, France. European Language Resources Association. → page 2

Wong, W. Y. P. (2006). *Syllable fusion in Hong Kong Cantonese connected speech*. Doctoral dissertation, The Ohio State University, Columbus, OH. → page 12

Yau, M. (2019). PyJyutping. → page 2

Yu, H. (2013). Mountains of gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, pages 108–121. Routledge. → page 4

Yuan, J., Ryant, N., and Liberman, M. (2014). Automatic phonetic segmentation in Mandarin Chinese: Boundary models, glottal features and tone. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2539–2543. → page 11

Ćavar, M., Ćavar, D., and Cruz, H. (2016). Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia. European Language Resources Association (ELRA). → page 11