Introducing MELI: the Mandarin-English Language Interview Corpus

Suyuan Liu, Molly Babel
Department of Linguistics, University of British Columbia
2613 West Mall, Vancouver, BC V6T 1Z4
suyuan97@student.ubc.ca, Molly.Babel@ubc.ca

Abstract

We introduce the Mandarin–English Language Interview (MELI) Corpus, an open-source resource of 29.8 hours of speech from 51 Mandarin–English bilingual speakers. MELI combines matched sessions in Mandarin and English with two speaking styles: read sentences and spontaneous interviews about language varieties, standardness, and learning experiences. Audio was recorded at 44.1 kHz (16-bit, stereo). Interviews were fully transcribed, force-aligned at word and phone levels, and anonymized. Descriptively, the Mandarin component totals ∼14.7 hours (mean duration 17.3 minutes) and the English component ∼15.1 hours (mean duration 17.8 minutes). We report token/type statistics for each language and document code-switching patterns (frequent in Mandarin sessions; more limited in English sessions). The corpus design supports within-/cross-speaker, within/cross-language acoustic comparison and links acoustics to speakers' stated language attitudes, enabling both quantitative and qualitative analyses. The MELI Corpus will be released with transcriptions, alignments, metadata, scans of labelled maps and documentation under a CC BY-NC 4.0 license.

Keywords: Speech Corpus, Bilingualism, Mandarin, English, Sociophonetics

1. Introduction

Language use naturally varies across individuals and contexts — no two speakers use language in exactly the same way. Such variation is fundamental to capturing linguistic structure and behavior, yet many studies have traditionally minimized its effects, either for methodological simplicity or due to limited data resources. To address this gap, we present the **M**andarin-**E**nglish **L**anguage **I**nterview (MELI) Corpus, an open-source speech corpus comprised of 29.8 hours of high-quality recordings, annotated at the word and phone levels, from 51 Mandarin-English bilingual speakers. The corpus includes both read sentences and spontaneous interviews, in which speakers reflect on their lived experiences with Mandarin and English and share their perceptions of language varieties and ideologies of language standardness. By centring the voices of speakers who are most intimately connected to these linguistic communities, MELI captures authentic linguistic variation and provides a rich resource for examining (1) regional varieties of Mandarin, (2) second-language accents of English, and (3) cross-language dynamics within bilingual speech.

1.1. Variation in Mandarin varieties

"Standard Mandarin" is not a linguistically stable variety. In fact, the term "Mandarin" obscures multiple linguistic concepts (Sanders, 1987). To avoid ambiguity, we use the following terminology in this study:

- **Standard Mandarin:** *Putonghua (*普通话*)*, the idealized form of Mandarin promoted across mainland China.
- **Mandarin varieties,** specifically [Cityname] Mandarin (e.g., Shanghai Mandarin): the local varieties of Mandarin, distinct from the local Chinese languages.
- **Local Chinese languages,** specifically [Cityname] Hua (e.g., Shanghai Hua): the Chinese languages co-existing with local Mandarin varieties, often not mutually intelligible with Standard Mandarin.

Research on "Mandarin variation" has traditionally emphasized local Chinese languages, often termed "dialects", rather than variation *within* Standard Mandarin itself. Consequently, the diversity among Mandarin varieties is often overlooked, reinforcing the misconception that Standard Mandarin is homogeneous.

Recent open-source corpora have expanded access to large-scale Mandarin data across regions, yet notable gaps remain. Many corpora rely on read speech (Bu et al., 2017; Wang and Zhang, 2015; Zhao and Chodroff, 2022), lack detailed documentation of regional varieties (Zhang et al., 2022; Bu et al., 2017), provide limited segment-level annotation (Zhang et al., 2022; Bu et al., 2017), or use sampling rates insufficient for fine-grained acoustic analysis (Yang et al., 2022). These resources are primarily optimized for automatic speech recognition (ASR) and thus contain little social context or the language or its speakers.

The MELI Corpus complements these efforts by offering 44.1 kHz studio-quality recordings of both read and spontaneous speech drawn from interviews on speakers' language attitudes toward

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5896–5904
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Mandarin varieties and Chinese languages. It includes detailed sociolinguistic metadata and hand-corrected annotation at the utterance, word, and segment levels. This design enables both fine-grained phonetic and sociophonetic analyses and qualitative examination of how speakers view linguistic variation, linking the *content* of their speech with its *acoustic* realization.

1.2. Variation in L2 English accents

Research on second-language (L2) English accent variation has a long history in linguistics and speech technology. However, the limited availability of publicly accessible corpora continues to restrict large-scale analyses. Existing L2 English corpora vary in purpose and design, but most were created for developing L2 speech processing systems rather than sociolinguistically-informed phonetic research. Many consist primarily of read speech (Zhao et al., 2018; Chen et al., 2019) or structured monologues (Knill et al., 2024), while conversational L2 English speech remains scarce. Moreover, proficiency information, an essential factor in L2 research, is rarely documented in detail, and few corpora include metadata about speakers' learning histories or linguistic experiences.

The MELI Corpus addresses these gaps with both read sentences and spontaneous English interviews produced by Mandarin-English bilinguals. Each speaker's English learning background and self-reported proficiency are documented, offering valuable context for interpreting variation. The interviews centre on speakers' attitudes toward English varieties through reflections on their learning experiences, yielding rich material for qualitative analysis. The interview content additionally supports computational text analysis of language attitudes, such as sentiment analysis, of the interview transcripts. By linking speakers' L2 English to their Mandarin varieties and language attitudes, MELI supports fine-grained, socially grounded analyses of Mandarin-accented English beyond the constraints of controlled elicitation.

1.3. Variations in Bilingual Speech Across Languages

A key feature of the MELI Corpus is its parallel-mode bilingual design, which includes matched recordings of both Mandarin and English speech from the same set of Mandarin-English bilingual speakers. While research on bilingualism has gained increasing attention, most existing "bilingual" corpora are in fact designed for code-switching research (Lovenia et al., 2022; Li et al., 2022), rather than providing balanced, language-separated sessions. Only a few open-access corpora adopt a parallel-mode approach, such as the Bangor corpora of Spanish-English, Welsh-English, and Welsh-Spanish bilingual speech (Deuchar et al., 2014), and the Speech in Cantonese and English (SpiCE) Corpus (Johnson et al., 2020), which served as an inspiration for MELI.

However, there remains no open-access parallel bilingual corpus for Mandarin and English, limiting comparative analyses of bilingual speech within individuals. The MELI Corpus addresses this gap by providing bilingual recordings with comparable duration, task structure, and recording conditions across both languages. This design enables direct comparison of acoustic and phonetic patterns both within and across languages for the same speakers. Furthermore, MELI integrates linguistic behaviour with sociolinguistic perspectives by linking each speaker's speech data to their explicitly expressed language ideologies, elicited through the interview content. This connection allows for a more comprehensive understanding of bilingualism that bridges acoustic evidence and speakers' self-reported views on language, standardness, and identity.

2. Corpus design and creation

The following section provides a detailed description of the corpus and its creation process. All data was collected between February 2024 and April 2024 in person at the Speech in Context Lab at the University of British Columbia. Transcription and forced-alignment of the interview data started in May 2024 and completed in October 2024.

2.1. Recruitment

Participants were recruited through a variety of methods including posts on the UBC Psychology Paid Studies List, posts on the UBC Graduate Student Community forum, posts on social media, printed flyers, and word of mouth. The recruitment criteria include the following: (1) be living in Metro Vancouver at the time of recording; (2) be born and raised in mainland China[1]; (3) have taken either the TOEFL or IELTS exam[2]; and (4) have attended, or be attending, a university in an English-speaking country at the time of recording.

Before participating in the study, all participants were asked to fill out an eligibility survey. Only those who met all criteria were contacted. The eligibility survey was in both simplified Chinese and English and assumed literacy in reading both. One

[1]This requirement was not further specified and relied on participants' own interpretation.

[2]Both TOEFL and IELTS are language proficiency exams commonly required for admission to English-speaking universities. This requirement was included to approximate a comparable level of English proficiency.

participant who did not read simplified Chinese was excluded from the corpus and subsequent follow-up studies. All participants who completed the interview were compensated with $20 CAD.

2.2. Participants

The corpus consists of speech of 51 Mandarin-English bilinguals (26 women and 25 men). Participants were coded with the following structure: self-identified gender (F or M), last two digits of their year of birth, and a unique letter to differentiate between participants with same identified gender and age. For example, F00A represents a female-identifying participant who were born in the year 2000.

One goal of the MELI corpus is to include speakers from different Mandarin-speaking regions. Figure 1 presents the geographical distribution of participants by province, based on their responses to the interview question "你是哪里人?" (*nǐ shì nǎlǐ rén*). This question captures both participants' hometown ("Where are you from?") and their regional identification ("Which region do you identify with?"). Figure 2 provides an overview of the locations where MELI participants resided for at least 12 consecutive months across different age ranges. If participant moved from mainland China to Canada at the age of 3, both locations would be counted in the 0 to 4 panel. Overall, all participants were born and raised in mainland China and later gained residential experience in an English-speaking country, primarily Canada.

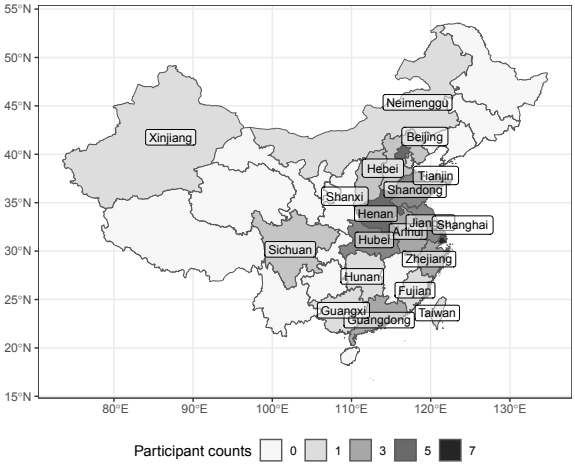

Figure 1: Geographical distribution of MELI particpnats.

Language background is an important component of the corpus and not necessarily reflected through geographical or residential background. To begin with, bilingualism should not be treated as a homogeneous category, as bilingual speakers

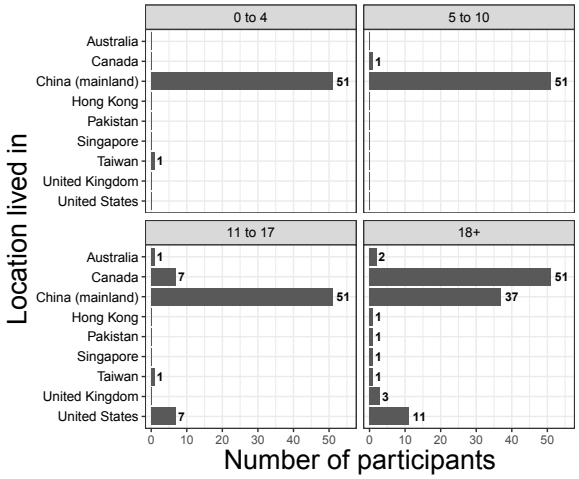

Figure 2: Residential history of MELI participants.

range from simultaneous to late learners. This is reflected in the MELI participants. The data presented here are taken directly from the language background questionnaire completed by participants, where participants rated their speaking, writing, understanding, and reading abilities for each language on a 7-point scale. A score of 1 was described as "very low, almost no ability" (很差), 4 as "average" (平均水平), and 7 as "very high, comparable to a native speaker" (很高, 基本相当于母语者水平). For Mandarin, the average age of acquisition was 1.23 years (range: 1–6). Mean self-ratings were 6.86 for speaking (range: 4–7), 6.70 for writing (range: 4–7), 6.88 for understanding (range: 4–7), and 6.90 for reading (range: 5–7). For English, the average age of acquisition was 6.17 years (range: 1–13). Mean self-ratings were 5.45 for speaking (range: 3–7), 5.39 for writing (range: 3–7), 5.78 for understanding (range: 3–7), and 5.96 for reading (range: 4–7). Together, these results show that MELI participants include both simultaneous and late bilinguals, with native or near-native proficiency in Mandarin and generally high proficiency in English.

2.3. Recording setup

The recording sessions were conducted in a quiet room at the Speech in Context Lab at the University of British Columbia. The participant and interviewer were seated across a table from each other. Both were equipped with SHURE WH20XLR headworn dynamic microphones positioned approximately 3 cm from the corner of the mouth. The microphones were connected to separate channels of a Focusrite Scarlett USBPre2 portable audio interface. Recordings were made in stereo at a 44.1 kHz sampling rate with 16-bit resolution using Audacity (Audacity Team, 2018) on a PC computer.

No	ID	YoB	Interview Order	Gender	Mandarin Lists	English Lists	AoA (Mandarin)	AoA (English)
1	F00A	2000	M → E	F	2, 6	2, 6	1	8
2	F01A	2001	E → M	F	2, 9	2, 9	3	8
3	F01B	2001	E → M	F	3, 10	3, 10	1	4
4	F02A	2002	M → E	F	3, 9	3, 9	1	3
5	F02B	2002	M → E	F	4, 5	4, 5	1	6
6	F02C	2002	E → M	F	4, 6	4, 6	1	4
7	F02D	2002	M → E	F	4, 7	4, 7	1	4
8	F02E	2002	E → M	F	5, 7	5, 7	2	5
9	F03A	2003	E → M	F	2, 5	2, 5	1	5
10	F03B	2003	M → E	F	2, 10	2, 10	1	6
11	F03C	2003	E → M	F	3, 6	3, 6	1	4
12	F05A	2005	E → M	F	6, 9	6, 9	1	8
13	F89A	1989	M → E	F	1, 2	1, 2	1	8
14	F92A	1992	M → E	F	5, 10	5, 10	1	6
15	F92B	1992	M → E	F	6, 8	6, 8	1	10
16	F94A	1994	E → M	F	1, 3	1, 3	1	6
17	F94B	1994	E → M	F	1, 9	1, 9	1	4
18	F94C	1994	M → E	F	5, 8	5, 8	1	6
19	F94D	1994	E → M	F	5, 9	5, 9	1	8
20	F95A	1995	E → M	F	1, 7	1, 7	1	6
21	F95B	1995	M → E	F	5, 6	5, 6	1	7
22	F96A	1996	M → E	F	1, 10	1, 10	1	12
23	F97A	1997	E → M	F	3, 4	3, 4	1	9
24	F98A	1998	E → M	F	3, 8	3, 8	1	3
25	F99A	1999	E → M	F	2, 3	2, 3	1	9
26	F99B	1999	E → M	F	4, 10	4, 10	1	5
27	M00A	2000	M → E	M	3, 5	3, 5	1	6
28	M00B	2000	E → M	M	5, 9	5, 9	1	6
29	M01A	2001	E → M	M	1, 5	1, 5	1	13
30	M01B	2001	M → E	M	4, 9	4, 9	1	5
31	M02A	2002	M → E	M	3, 7	3, 7	1	7
32	M03A	2003	M → E	M	1, 8	1, 8	1	4
33	M03B	2003	M → E	M	5, 8	5, 8	1	4
34	M04A	2004	E → M	M	6, 7	6, 7	1	7
35	M04B	2004	E → M	M	7, 8	7, 8	1	3
36	M04C	2004	M → E	M	9, 10	9, 10	1	1
37	M05A	2005	M → E	M	5, 6	5, 6	1	8
38	M05B	2005	E → M	M	5, 7	5, 7	1	6
39	M89A	1989	M → E	M	2, 8	2, 8	1	11
40	M89B	1989	E → M	M	8, 10	8, 10	1	6
41	M94A	1994	M → E	M	5, 10	5, 10	5	7
42	M95A	1995	E → M	M	2, 7	2, 7	1	6
43	M96A	1996	M → E	M	7, 9	7, 9	6	5
44	M96B	1996	M → E	M	5, 8	5, 8	1	10
45	M97A	1997	M → E	M	1, 4	1, 4	1	5
46	M97B	1997	E → M	M	4, 10	4, 10	1	6
47	M98A	1998	M → E	M	2, 4	2, 4	1	6
48	M98B	1998	M → E	M	6, 10	6, 10	1	7
49	M99A	1999	M → E	M	1, 6	1, 6	1	6
50	M99B	1999	E → M	M	7, 10	7, 10	1	1
51	M99C	1999	M → E	M	8, 9	8, 9	1	3

Table 1: Basic information for each participant's interviews in the MELI Corpus. M → E = Mandarin interview first; E → M = English interview first.

2.4. Recording procedure

After signing the consent form, the audio release form and completing the language background questionnaire, participants were given a quick walk-through of the procedure of the interview. This walk-through verbally (1) went over the general procedure of the recording session, (2) reassured participants that they did not need to answer questions they felt uncomfortable with, and they could withdraw from the study at any point without any consequences, and (3) they could redact any part of the interview before the release of the corpus. Participants then completed two recording sessions in Mandarin and English in one sitting. The order of the languages was counterbalanced (see Table 1 for details).

Each recording session began with a sentence-reading task, followed by a 20–30 minute interview. Participants first familiarized themselves with the sentences before reading them aloud. Both sessions were recorded in a single audio file for each participant. In the finalized corpus, the audio file was segmented by language session.

2.4.1. Sentence reading

The reading materials for the sentence reading task were selected to act as stimuli for a setence-in-noise transcription task in the first author's dissertation work, which will not be discussed in the current paper. The Mandarin sentences were taken from the Mandarin Speech Perception (MSP) sentence test material (Fu et al., 2011), which contains 10 lists. All lists were designed to include a comprehensive set of phonemes in the respective language. The English sentences were 10 lists taken from the English Hearing In Noise Test (HINT) (Nilsson et al., 1994), which was developed for evalua-

tion of speech perception threshold in noise. The original HINT includes 25 lists. Prior to recording, a norming survey was conducted in February 2024. Fourteen Mandarin–English bilingual raters evaluated the naturalness of sentences from all 20 HINT lists (each rater evaluated 13 lists to reduce fatigue) and from the 10 MSP lists on a continuous scale from 1 to 7. The 10 English HINT lists with the highest normalised naturalness ratings were selected for recording (Mean = 0.87, range: 0.83–0.90)[3]. These values are comparable to those of the Mandarin MSP lists (Mean = 0.87, range: 0.82–0.93). Each list contains 10 sentences.

In each language session, participants read two lists of sentences drawn from the selected HINT lists (English session) or MSP lists (Mandarin session). After an initial read-through, participants were instructed to repeat each sentence[4]. This resulted in 40 sentences per language (2 lists $\times$ 10 sentences $\times$ 2 repetitions).

2.4.2. Interview

Interviews were always conducted after sentence reading. The Mandarin interview focused on two main topics: (1) eliciting participant's attitude towards Mandarin varieties, Chinese languages, and Standard Mandarin with relation to their language backgrounds and (2) reflecting on participant's impression of their voice. Inspired by Preston (1982), a draw-a-map task was conducted during the discussion of the Mandarin and Chinese language landscape, where participants were asked to circle the geographical regions that speak the "most standard" and "least standard" Mandarin. The English interview focused on (1) participant's experience learning English and their attitude towards different English varieties and (2) their reflection on how their voices differ across languages, if at all. A full list of sample questions are provided in the corpus documentation.

No instruction or comment on participant's speaking style was given during this session. In particular, no instruction on code-switching or specific language variety was given.

3. Annotation

This section outlines the procedures for processing and annotating the recorded interviews. The pipeline follows these steps: (1) extraction of participant's channel and file segmentation by language, (2) initial transcripts using Whisper text-to-speech (Radford et al., 2023), (3) hand-correction of orthographic transcripts by bilingual research assistants, (4) word and phone level forced alignment using Montreal Forced Aligner (McAuliffe et al., 2017), and (5) anonymization.

3.1. Whisper speech-to-text transcription

The Whisper model (Radford et al., 2023) was used to transcribe both English and Mandarin interviews. Whisper is a pre-trained, Transformer-based sequence-to-sequence model trained for either speech recognition or a combination of speech recognition and speech translation. For this corpus, the multilingual `whisper-large-v3`[5] model was selected, as it was the most current and suitable model for transcribing multilingual data at the time.

Interview recordings were first segmented by language in PRAAT (Boersma and Weenink, 2022). Mandarin and English interviews were processed separately. For each language, the participant's audio channel was extracted as a WAV file and submitted to the Whisper model. A customized Python script, adapted from the `whisper-large-v3` documentation, was used to automate the transcription process. The model was configured for transcription (rather than translation) and implemented a chunked long-form algorithm with a 30-second segment length and a batch size of 16. This procedure divided long audio files into 30-second segments, transcribed each segment individually, and then merged the outputs at segment boundaries. The resulting transcriptions, including utterance-level timestamps, were exported in CSV format.

3.2. Transcription correction and adjustment

The auto-generated transcriptions were then hand-corrected by research assistants. The CSV files were converted to TextGrid using ELAN (Max Planck Institute for Psycholinguistics, 2024) for hand correction. Each TextGrid file contains four tiers: (1) task (sentence reading vs. interview), (2) automatic transcription by Whisper, (3) corrected transcription and (4) notes. Research assistants revised the Whisper-generated transcriptions in the corrected transcription tier and added notes in the notes tier. Notes included flagged identifying information and content participants requested to be redacted before the corpus release. The following conventions were followed during hand-correction:

3.2.1. General conventions

- Unintelligible speech was transcribed as "xxx".

[3]The English sentence lists used were 2, 14, 15, 18, 19, 20, 21, 22, 23, 24 (Nilsson et al., 1994).

[4]The first five participants (F89A, F94A, M97A, M01A, M99A) read each sentence only once, as this repetition procedure was implemented after the initial few sessions.

[5]https://huggingface.co/openai/whisper-large-v3

- Punctuation was generally avoided, except for question marks (“?”) and possessives (marked with “ ’ ”).
- Speech fragments were annotated using “&” followed by English orthography or Pinyin (e.g., “...&con confident...”).
- Non-speech sounds were transcribed using the labels listed in Table 2.
- Language or speech variation such as code-switching or deviation from a participant’s baseline accent was marked with “@” followed by a description (e.g., @e for code-switches to English, @YantaiHua for switches to the local Chinese variety spoken in Yantai). A complete list of language varieties participants switched to can be found in the corpus documentation.
- Numbers were spelled out in full (e.g., “a hundred” or “一百” *yì bǎi* for 100).

Symbol	Meaning
{interviewer}	When interviewer is talking
{laughter}	Laughter
{sil}	Silence
{cough}	Coughing
{chupse}	Sucking air in between teeth
{inhale}	Inhale
{click}	Clicking
{sniff}	Sniff
{exhale}	Exhale
{chchch}	Ch-ch-ch (sound for looking for things)
{micnoise}	Microphone noise
{swoosh}	Swoosh sound
{rhythm}	Demonstrating rhythm, accent, or intonation

Table 2: Labels for non-speech or filler items.

3.2.2. English-specific conventions

- Filler words were transcribed using one of the following standardised forms that best matched the pronunciation: *hm, uhhuh, mmhm, um, uh, mm, huh, ah, em, nn, eh, ih*.
- Capitalisation was used for country names (e.g., China), brand names (e.g., Apple), languages (e.g., English, Mandarin), first-person subject (I), single letters when spelled out (e.g., A-Z), and segmented syllables (e.g., A PPLE).

3.2.3. Mandarin-specific conventions

- The default character for the third-person pronoun was standardised to “她” (*tā*).
- Sentence-final particles and exclamations were transcribed using one of the following standardised forms that best matched participant’s production: 啊 *a*, 吗 *ma (question)*, 呀 *ya*, 吧 *ba*, 呢 *ne*, 咯 *lo*, 嘛 *ma (statement)*, 哎 *ai*, 啦 *la*, 嘿 *hei*, 哇 *wa*, 哦 *o*, 哼 *heng*, 滴 *di*, 嘟 *du*, 耶 *ye*, 哈 *ha*, 呐 *na*, 呗 *bei*, 嘞 *lei*, 咦 *yi*.
- Common filler words were transcribed with the following: 嗯 *en*, 呃 *e*, 嗯哼 *en heng*, 昂 *ang*, 哎哟 *ai yo*.

3.3. Forced alignment

Forced alignment was generated using Montreal Forced Aligner (MFA) v3.0 (McAuliffe et al., 2017), based on the hand-corrected transcriptions. This process required an audio file, an orthographic transcription, an acoustic model, and a pronunciation dictionary. The outputs were word-level and phone-level alignments for the interview audio files.

For forced alignment of English interviews, the `English (US) ARPA acoustic model v3.0.0`[6] was employed. The dictionary, derived from `the English (US) ARPA dictionary v3.0.0`[7], was customized by manually transcribing out-of-vocabulary (OOV) lexical items detected during validation. These transcriptions adhered to the ARPA convention and were incorporated into the existing dictionary.

For forced alignment of Mandarin interviews, word boundaries were first introduced using the `jieba` Python library [8] since they were not marked in the original transcriptions. Forced alignment utilized the `Mandarin (China) MFA acoustic model v3.0.0`[9] and the `Mandarin MFA dictionary v3.0.0`[10], which provides pronunciations in International Phonetic Alphabet (IPA). OOV lexical items were identified during validation, transcribed manually using the same IPA conventions, and added to the dictionary.

The output TextGrid file includes four tiers: (1) task (sentence reading vs. interview), (2) corrected transcription, (3) word tier, and (4) phone tier. Speech fragments, unintelligible speech, and code-switched are transcribed as “spn” in the phone tier by MFA. Figure 3 shows examples of the forced-aligned interviews for both languages.

3.4. Anonymization

Transcribed interviews were returned to participants, who were invited to indicate any portions they wished to have redacted. These segments

[6] MFA English (US) ARPA acoustic model v3.0.0.
[7] MFA English (US) MFA dictionary v3.0.0.
[8] https://github.com/fxsjy/jieba
[9] MFA Mandarin MFA acoustic model v3.0.0
[10] MFA Mandarin (China) MFA dictionary v3.0.0

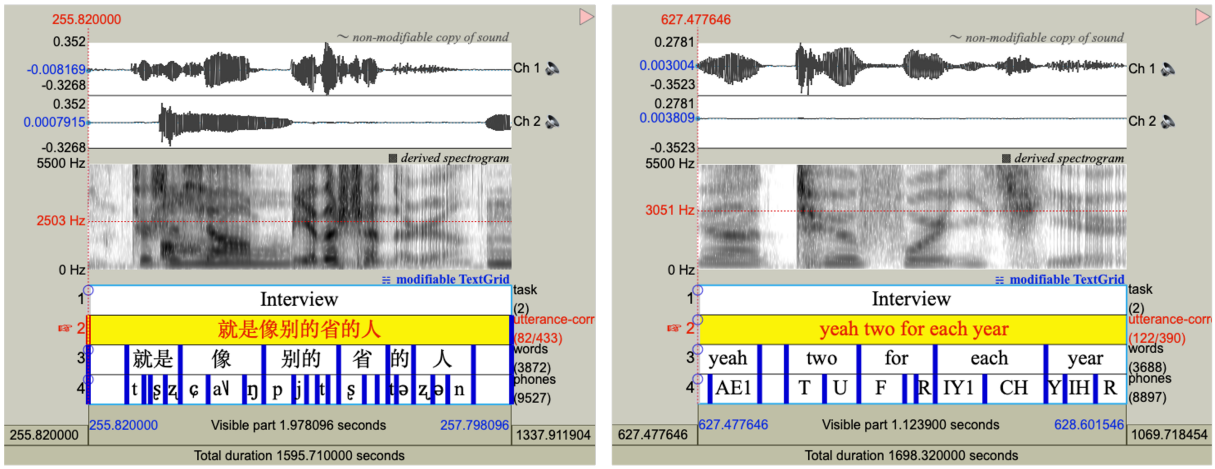

Figure 3: Examples of the forced-aligned interviews (left: Mandarin; right: English).

were marked in the notes tier by two research assistants, incorporating both participant requests made during the review process and any redaction requests expressed during the interview. In addition to participant-specified content, all personal or identifying information (e.g., names of individuals or schools) was redacted, even if not explicitly flagged. All redactions were independently identified and verified by two research assistants. Once the sections requiring redaction were identified, the corresponding audio in both the participant and interviewer channels was muted and replaced with a 1000 Hz tone using ELAN (Max Planck Institute for Psycholinguistics, 2024). The associated intervals in the TextGrid files were replaced with "[REDACTED]" using Praat (Boersma and Weenink, 2022).

4. Descriptive Statistics

This section summarises the basic characteristics of the Mandarin and English interview recordings, including total duration, speaker-level variation, and lexical content in each language component of the MELI Corpus.

4.1. Mandarin Interviews

The Mandarin interviews include 14.7 hours of speech data including both the sentence reading task and the interview. The mean duration of speech data across participants is 17.3 minutes (range: 9.1 minutes – 28.9 minutes). These data exclude silences in the participants' speech.

A total of 5,799 word types (i.e. unique words) and 153,782 word tokens were produced in the Mandarin interview sessions. The total value of word types and word tokens vary across participants with a mean of 586 for word types (range: 334–923) and 3,015 for word tokens (range: 1,304–5,670). This estimation is based on the default dictionary in the `jieba` Python library used for word segmentation and takes all words into consideration regardless of language. This includes all code-switched items, unintelligible speech, and speech fragments. Figure 4A shows a break-down of the types of tokens produced in the Mandarin interview sessions. All participants except for F03C code-switched to English during their Mandarin interview.

4.2. English Interviews

Same as the Mandarin interviews, the duration of English interviews include both sentence reading task and the interview, excluding silences. The English interviews include 15.1 hours of speech data with a mean duration of 17.8 minutes across participants (range: 7.5 minutes – 29.22 minutes).

A total number of 4,543 word types and 129,647 word tokens were produced in the English interview sessions, with a mean word type of 516 (range: 334 to 760) and a mean word token of 2,542 (range: 955 to 5,158) across participants. Figure 4B shows a break-down of the word tokens. 23 out of 51 participants switched to Mandarin during their English interview and 19 out of 51 participants did not code switch at all. In comparison to the Mandarin interviews, MELI participants code switched more from Mandarin to English than from English to Mandarin.

5. MELI corpus release

Following anonymization, the MELI corpus is scheduled to be released in March 2026 through UBC Research Data Collection via Scholars Portal Dataverse under a Creative Commons Attribution 4.0 International License[11], which allows use for non-commercial research and educational purposes with attribution. The initial release includes

[11] https://creativecommons.org/licenses/by/4.0/

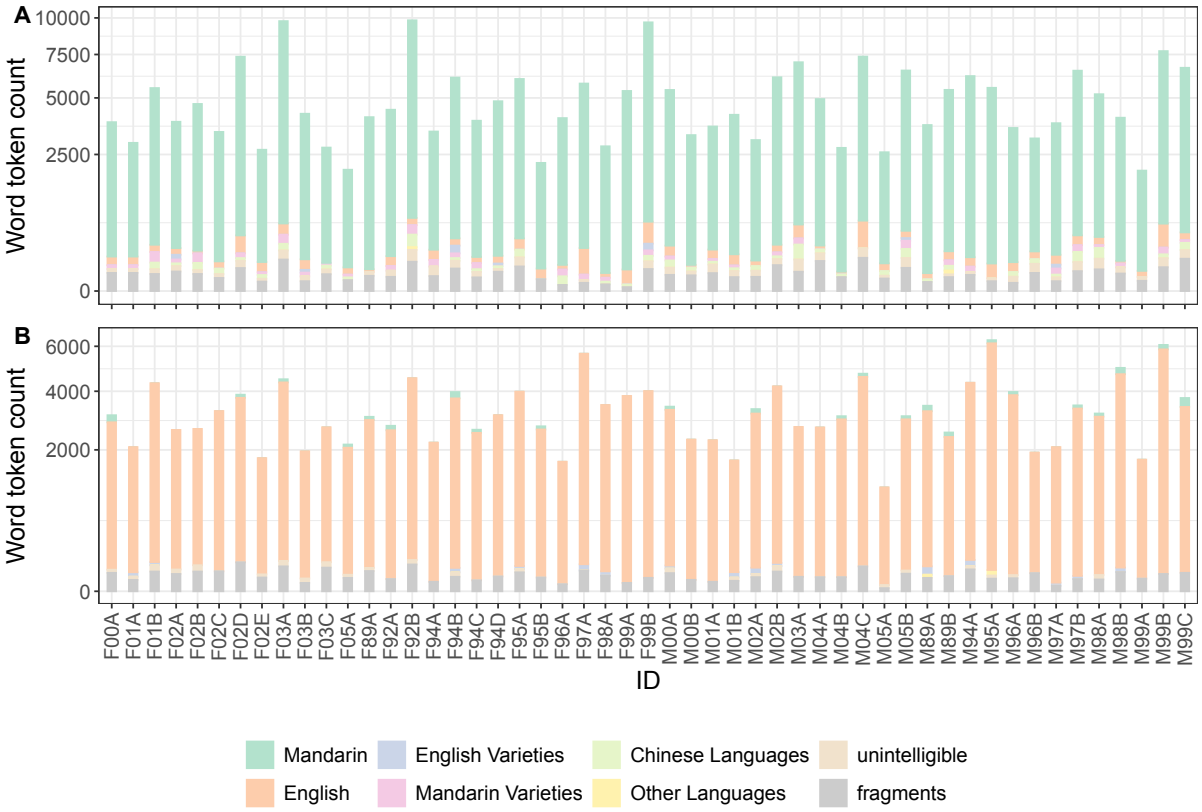

Figure 4: Distribution of code-switching in Mandarin (A) and English (B) interviews.

anonymized audio files, corresponding transcriptions and forced alignments in Praat TextGrid format, scanned copies of maps from the draw-a-map task in the Mandarin interviews, a detailed language background and metadata summary, and a README file documenting the corpus structure. All releases can be found in the online documentation.[12] Future updates will include extended annotations and alignments of the interviewer's speech. Users will be notified of updates through the repository page.

6. Discussion and conclusion

Languages cannot be studied without acknowledging the variations embedded in its speakers, though fully contextualized language analysis can be challenging due to the lack of resources. The **M**andarin-**E**nglish **L**anguage **I**nterview (MELI) Corpus introduced here attempts to provide more resources for this gap, allowing studies of (1) regional varieties of Mandarin, (2) second-language accents of English, and (3) cross-language dynamics within bilingual speech. The unique format of high-quality interview recordings in MELI allow approaches both qualitatively and quantitatively, bridging together digital humanities, phonetics, and computational linguistics, and returns the voice to the people who are most intimately connected to the communities and its languages.

[12] https://meli-corpus.readthedocs.io/

7. Acknowledgements

The creation of the MELI corpus was approved by the University of British Columbia Behavioural Research Ethics Board (H23-03205), and was supported by the UBC Arts Graduate Research Award to the first author and SSHRC grant to the second author. We thank the MELI participants for sharing their time, voice and insights, and many members of the Speech-in-Context lab for their contribution in the creation process, especially Angelina Yuan, Dlorah Lyne Agama, Jeff Li and Sarah Ong.

8. Bibliographical References

Audacity Team. 2018. Audacity (R): Free audio editor and recorder. https://www.audacityteam.org/. Accessed: 2024-07-30.

Paul Boersma and David Weenink. 2022. Praat: doing phonetics by computer [computer program].

http://www.praat.org/. Version 6.2.17, retrieved 23 August 2022.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proc. O-COCOSDA*, pages 1–5.

Yu Chen, Jun Hu, and Xinyu Zhang. 2019. Sellcorpus: an open source multiple accented chinese-english speech corpus for l2 english learning assessment. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7425–7429. IEEE.

Margaret Deuchar, Peredur Davies, Jon Herring, M Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111.

Qian-Jie Fu, Min Zhu, and Xiaoyan Wang. 2011. Development and validation of the mandarin speech perception test. *The Journal of the Acoustical Society of America*, 129(6):EL267–EL273.

Khia A. Johnson, Molly Babel, Ivan Fong, and Nancy Yiu. 2020. SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4089–4095, Marseille, France. European Language Resources Association.

Kate Knill, Diane Nicholls, Mark JF Gales, Mengjie Qian, and Pawel Stroinski. 2024. Speak & improve corpus 2025: an l2 english speech corpus for language assessment and feedback. *arXiv preprint arXiv:2412.11986*.

Chengfei Li, Shuhao Deng, Yaoping Wang, Guangjing Wang, Yaguang Gong, Changbin Chen, and Jinfeng Bai. 2022. Talcs: An open-source mandarin-english code-switching corpus and a speech recognition baseline. *arXiv preprint arXiv:2206.13135*.

Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi, and Pascale Fung. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7259–7268, Marseille, France. European Language Resources Association.

Max Planck Institute for Psycholinguistics. 2024. Elan [computer software]. https://archive.mpi.nl/tla/elan. Version 6.8.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Michael Nilsson, Sigfrid D. Soli, and Jayne A. Sullivan. 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099.

Dennis R Preston. 1982. Perceptual dialectology: Mental maps of united states dialects from a hawaiian perspective. *Working papers in Linguistics*, 14(2):5–49.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christopher McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Robert M Sanders. 1987. *The Four Languages of" Mandarin"*. Department of Oriental Studies, University of Pennsylvania.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A free chinese speech corpus. *CoRR*, abs/1512.01882.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, et al. 2022. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, pages 1547–1551.

Liang Zhao and Eleanor Chodroff. 2022. The mandi corpus: A spoken corpus of mandarin regional dialects. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1985–1990.