The Multi-ethnic Hong Kong Cantonese Corpus

Alan Yu[1], Nathan Delisle[1], Nicholas Martin[1], Vivienne Zhang[1], Yao Yao[2], and Carol To[3]

[1]University of Chicago, [2]Polytechnic University of Hong Kong, [3]The University of Hong Kong

This presentation introduces the Multi-ethnic Hong Kong Cantonese Corpus (MeHKCC; https://ccds.edu.hku.hk/), a database consisting of audio recordings from three groups of mother-toddler dyads where the mothers came from distinct linguistic backgrounds. The goal of this project aims to investigate how the acquisition of HKC is influenced by the caregivers' linguistic background.

Participants: The participants include (i) 32 local mothers in Hong Kong who speak HKC since birth; (ii) 27 mothers who reported that Putonghua was their strongest language. All except seven had been staying in Hong Kong for 5 to 20 years; the rest has stayed in Hong Kong for two to four years. Their own ratings of the Cantonese proficiency ranged from very good to fair; (iii) 12 South Asian mothers who spoke at least one South Asian language (Urdu [N=8], Punjabi [N=2], Tamil [N=12]) as their first language. Nine of them had been staying in Hong Kong for more than 20 years while the rest had been in Hong Kong for less than 20 years. Their own ratings of the Cantonese proficiency within the group varied substantially from highly proficient to poor. The child participants (31 boys and 37 girls were typically developing children at the age of 6 to 59 months old at the time of study. Their mothers were their main caregivers. All except three were born full-term. None of the children have any diagnosed developmental disorders.

Recordings: The CDS recordings were 45 minutes to 1 hour long with the mothers interacting with their child in a mock living room. Age-appropriate toys and books were provided for interaction. The ADS recordings consists both dialogues and monologues in HKC. Dialogues included face-to-face interviews with the experimenter which consisted of questions and answers regarding child's developmental and social history, daily routine and mother's background, their job or their daily routine. Three mothers in the South Asian group cannot carry out a dialogue solely in Cantonese. The conversation was then accompanied by English. Monologues were elicited via 4 tasks: a film description task, a map description task, a story retelling task and a single-word picture naming task.

Transcription and Forced Alignment: The ADS and CDS recordings were transcribed by a team of phonetically-trained, native Cantonese-speaking research assistants using the software, Phon (Hedlund and Rose, 2019). The research assistants would first determine the temporal boundaries of each utterance, and then transcribe the utterance into Chinese characters or Jyutping symbols (when no standard Chinese characters were available). To ensure transcription reliability, all transcriptions were checked by a separate researcher. The orthographic annotations were submitted to automatic forced alignment using SPeech Phonetization Alignment and Syllabication (SPPAS; Bigi, 2015) where the transcribed utterances were parsed into words, syllables, and segments; SPPAS does not provide a means to encode the tonal information of a form.

Illustrations: As illustrations of the potential of this corpus, this presentation will focus on the ADS/CDS speech samples of Cantonese-speaking Hong Kong born mothers. Specifically, we examine the hyper-speech hypothesis associated with child-directed speech (e.g., Fernald, 2000) by looking at the presence of enhancement effects in CDS relative to ADS in terms of vowel space, tonal space, the obstruent aspiration contrast, and intrinsic consonant F0 effects (i.e. the f0 perturbation effect associated with the aspiration/voicing status of the onset obstruent). Analysis is ongoing, but preliminary results suggest that tone space expansion is present in CDS, but not vowel space expansion.

References

Bigi, B. (2015). Sppas - multi-lingual approaches to the automatic annotation of speech. The Phonetician - International Society of Phonetic Sciences, 111-112:54-69.

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. Phonetica, 57:242:254.

Hedlund, G. and Rose, Y. (2019). Phon 3.0.