SME Corpus

Samantha Sie
Theoretical and Applied Linguistics Section
Faculty of Modern and Medieval Languages and Linguistics
University of Cambridge
slws2@cam.ac.uk

Participants: 175
Type of study: narrative
Location: Malaysia; United Kingdom
Media type: audio

Citation Information

Sie, S. (2023). *Investigating the role of an indigenised variety of English in the acquisitional and sociolinguistic contexts of the Malaysian ecology*. PhD thesis: University of Cambridge. DOI: https://doi.org/10.17863/CAM.96583

Sie, S. and Tsimpli, I. M. The effects of L1 on the use of grammatical finiteness in Standard English: An investigation into the multilingual ecology of Malaysia. *In preparation*.

In accordance with TalkBank rules, any use of data from this corpus must be accompanied by at least one of the above references.

Corpus Description

The Standard Malaysian English (SME) corpus comes from the author's main PhD project (supervised by Prof. Ianthi Maria Tsimpli), which examined crosslinguistic influence in the ultimate acquisition of Standard English in the Postcolonial Englishes context of Malaysia. Malaysia was chosen as the primary research site for two main reasons. Firstly, it has a linguistically diverse makeup, with Malay, Chinese, and Tamil being some of the most widely spoken first languages (L1s) amongst the local speech communities. Secondly, it has a long-standing history with the English language, which not only enjoys an elevated albeit restrictive status of an official second language (L2) but has also undergone structural indigenisation due to linguistic and sociolinguistic factors such as protracted language contact, L2 acquisitional mechanisms, identity rewritings, and bilingual creativity. Accordingly, the PhD project set out to investigate the extent to which different L1s – including nativized English, which sees a growing number of L1 speakers – played a facilitative or adverse role in the ultimate acquisition of Standard English.

The corpus comprises 175 elicited narratives produced by adult Malaysians (n = 145; mean age = 20 years, SD = 1.21) and British controls (n = 30, mean age = 21, SD =

2.54). These participants were mostly university students studying in Malaysia (i.e., University of Malaya) and the UK (i.e., University of Cambridge), respectively. There are altogether 103,607 words from about 14 hours of audio recording in this corpus.

The narrative task was carried out on a one-to-one basis. Participants were asked to watch an animated silent film called "Snack Attack" (2012) and narrate the story to the researcher. As the morphosyntactic features under inquiry were the English finiteness markers (i.e., tense inflections, copula and auxiliary BE, auxiliary DO), four questions were presented in the following order to elicit them from participants:

Q1 Tell me what happened in the video.
Q2 If you were the young man, what would you do when the old lady took your packet of cookies?
Q3 If you were the old lady, what would you do when you found out that you ate the young man's cookies?
Q4 What is the moral of the story?

As the data collection had to be conducted in two phases due to Covid disruptions, the narrative sessions took place in different mediums. This means that the quality of the audio recordings was affected to a certain degree. In the first phase (November 2019 – March 2020; pre-Covid), narrative sessions were recorded in person using a Sony ICD-UX560F sound recorder and were held in a quiet, public room (e.g., research office, seminar room) on campus. In the second phase (November 2020; during Covid), narrative sessions were conducted via Zoom and were recorded using the audio-recording function provided by the video conferencing programme.

The PhD project was carried out in line with the ethical guidelines set out by the Ethics Committee of the Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge. Participants were informed well in advance about the aims and requirements of the study and took part on a voluntary basis. Most of them gave signed consent to have their anonymised audio files made available in an open-access language corpus, such as TalkBank. There were, however, three individuals who did not give permission for their audio files to be uploaded publicly. Therefore, the transcripts of the three individuals are not audio-linked whereas the rest are. Finally, for anonymisation purposes, unique IDs were assigned to all participants.

Notes in the transcripts' @Comment line include:

- L1 of the participant
- Onset age of acquisition (AoA) in English (in years; applicable to Malaysians only)
- Latest formal English Language assessment (applicable to Malaysians only; e.g., Malaysian University English Test [MUET]; International English Language Testing System [IELTS])
- Proficiency score (in %) from the British Council Online English Level Test

Reference:

Snack Attack (2012). Cadelago, A. Metanoia Films and Arc Productions. Available at: https://snackattackmovie.com/ [Accessed 25 August 2023]

Usage Restrictions
Embargo