AESOP-ILAS Corpora

1. **Name:** AESOP-ILAS (Asian English Speech cOrpus Project - Institute of Linguistics, Academia Sinica) Corpora

2. **Language:** L2 English by Taiwan Mandarin speakers, corresponding speech by L1 American speakers, and some corresponding L1 Mandarin by Taiwan speakers

3. **Aims:** Prosodic features of Taiwan L2 English

4. **Content:** AESOP-ILAS 1 and AESOP-ILAS 2

5. **Time of Recording:**
   (1) AESOP-ILAS 1: 2009.07-2010.09
   (2) AESOP-ILAS 2: 2011.11-2012.07

6. **Data Type:** Microphone speech (Sennheiser PC155 headset microphone)

7. **Number of Speakers:** 540 in total; the number of male and female speakers is approximately equal in both groups
   (1) AESOP-ILAS 1: 12 native speakers of North American English (L1) and 488 native speakers of Taiwan Mandarin (L2)
   (2) AESOP-ILAS 2: 10 native speakers of North American English (L1) and 30 native speakers of Taiwan Mandarin (L2)

8. **Data Size:** 14 GB (approximately 812 hours) in total; including sound files, corresponding text, and automatic HTK aligned files for phone boundaries, some of which have been manually adjusted
   (1) AESOP-ILAS 1: 8.64 GB; 500 hours (approximately 1 hour per speaker)
   (2) AESOP-ILAS 2: 5.42 GB; 312 hours (approximately 7.8 hours per speaker)

9. **Data Content:**
   (1) AESOP-ILAS 1: 6 elicited read speech tasks, 1 fully aided computer-prompted dialogue task, and 1 partially aided picture description task
   (2) AESOP-ILAS 2: 4 elicited read speech tasks (including one Taiwan Mandarin task) and 1 fully aided computer-prompted dialogue task

10. **Introduction:**
    The AESOP-ILAS speech corpus is especially designed for the Taiwan division of the multinational research project AESOP (Asian English Speech Corpus

Project), featuring L2 English speech by native speakers of Taiwan Mandarin. It is funded by the Chiang Ching-kuo Foundation for International Scholarly Exchange (DB002-D-08. 2009.7.1-2012.12.31.). The principal investigator of this project is Dr. Chiu-yu TSENG, Distinguished Research Fellow and Director of the Institute of Linguistics, Academia Sinica.

The project aims to build up a corpus of the English spoken in Taiwan as an open resource and to investigate a wide range of communicative phonetic and prosodic features in Taiwan English at the segmental, lexical, phrasal, and discourse levels, rather than focusing on specific and individual phenomena.

AESOP-ILAS is 14 GB in total, featuring approximately 812 hours of sound files, corresponding text, and automatic HTK aligned files for phone boundaries, some of which have been manually adjusted. It is separated into 2 parts: AESOP-ILAS 1 and AESOP-ILAS 2. AESOP-ILAS 1 is 8.64 GB (500 hours) and includes L1 English speech data by 12 American English native speakers and L2 English speech by 488 Taiwan Mandarin speakers. The recording time of each speaker is approximately 1 hour. AESOP-ILAS 2 is 5.42 GB (312 hours) and includes L1 English speech data by 10 American English speakers and L2 English speech data by 30 Taiwan Mandarin speakers. The recording time of each L1 speaker is approximately 5.25 hours and 8.7 hours for each L2 speaker. The speech corpus should be useful for research and development in language teaching, language modeling, phonetic research and applications to speech synthesis and recognition.

AESOP-ILAS is released in April, 2015 for use of non-commercial academic research only. ACLCLP is authorized to release it. Applicants are supposed to apply by signing the license agreement and complying with the terms on the license agreement.