TRANSCRIBING CHINESE LANGUAGE CLASSROOM TALK

To build a computer corpus

Liu Yongbing, Goh Hock Huan and Zhao Shouhui

Abstract

With the completion of classroom observation and data collection of the CRPP Core Project in mid-2005, about 120 hours of classroom audio data was collected from 38 units of Chinese language lessons in Singapore Primary 5 and Secondary 3 classrooms. As designed in the CRPP Core Project (Panel 4; see Luke et al., 2004), the collected audio recordings were to be transcribed and the Mandarin part of the Singapore Corpus of Research in Education was established (hereafter referred to as SCoRE Chinese). Since the Chinese classroom audio files are language specific and different from other subjects areas (in English), the Chinese language research team initiated this specific focus project to provide the transcribed audio data to the Singapore Corpus of Research in Education.

In view of the language uniqueness in these Chinese audio recordings, the transcription standards and conventions designed for the CRPP Core Project were adapted. With the completion of adapting the transcription standards and conventions, two Chinese transcription training workshops were given to 15 potential transcribers and 10 of them were employed on part-time basis for the transcribing task. Though with demonstration and transcription standards and convention, there were many transcription errors and unforeseen data errors identified by our quality controller in the transcribed files from transcribers. Therefore, systematic checking and data-cleaning were conducted and a Data Checking and Cleaning Manual was developed in the process.

While transcribing of audio data and cleaning of transcribed data, an annotation scheme was designed for SCoRE Chinese, namely Annotation of IRF. This scheme was designed with reference to various studies of IRF in the research literature (Swan, 1995; Tsui, 1985; White & Lightbrown, 1984). Annotation can be done by using this scheme on transcribed data for further analysis of classroom interactions.

In this project, 70 hours of audio recordings from 25.5 units of Chinese language lessons (i.e., 79 audio files) have been transcribed, cleaned and handed over to the CRPP Corpus team for the establishment of the SCoRE Chinese, and an annotation scheme of IRF has been developed for the SCoRE trial purpose. Due to both the labour-intensive nature of transcription and the prolonged data checking and cleaning process, 25 hours of classroom recordings from 8.5 units of Chinese language lessons (i.e., 29 audio files) of the total are left (not yet transcribed).

The full version of this report is not available due to the confidential nature of the data.
Acknowledgement
This paper makes use of data from the research project “Transcribing Chinese Language Classroom Talk: To Build a Computer Corpus” (CRP 23/05 LYB), funded by the Centre for Research in Pedagogy and Practice, National Institute of Education, Singapore. The views expressed in this paper are the author’s and do not necessarily represent the views of the Centre or the Institute.

About the authors
Liu Yongbing was Associate Professor at the National Institute of Education, Nanyang Technological University, Singapore. He is currently a consultant with the Centre for Research in Pedagogy and Practice.

Goh Hock Huan is a Research Associate with the Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore.

Zhao Shouhui is Associate Professor with the Centre for Research in Pedagogy, National Institute of Education, Nanyang Technological University, Singapore.

Contact us
For further information, please email: reports@nie.edu.sg