Encyclopedia of Chinese Language and Linguistics

Volume 4

Shā–Z

ENCYCLOPEDIA OF CHINESE LANGUAGE AND LINGUISTICS

Volume 4
Shā–Z

General Editor
Rint Sybesma

Associate Editors
Wolfgang Behr
Yueguo Gu
Zev Handel
C.-T. James Huang
James Myers

BRILL

LEIDEN • BOSTON
2017

Typeface for the Latin, Greek, and Cyrillic scripts: “Brill”. See and download: brill.com/brill-typeface..

ISBN 978-90-04-18643-9 (hardback, set)
ISBN 978-90-04-26227-0 (hardback, vol. 1)
ISBN 978-90-04-26223-2 (hardback, vol. 2)
ISBN 978-90-04-26224-9 (hardback, vol. 3)
ISBN 978-90-04-26225-6 (hardback, vol. 4)
ISBN 978-90-04-26226-3 (hardback, vol. 5)

This book is printed on acid-free paper and produced in a sustainable manner.

Schmidt, Wilhelm, "Die Mon–Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens" [The Mon–Khmer peoples, a link between the peoples of Central Asia and Austronesia], *Archiv für Anthropologie, Braunschweig* 5, 1906, 59–109.

Schmidt, Wilhelm, "Die Beziehungen der austrischen Sprachen zum Japanischen" [The connections of the Austric languages to Japanese], *Wien Beitrag zur Kulturgeschichte und Linguistik* 1, 1930, 239–251.

Shorto, Harry L., "In Defense of Austric", *Computational Analyses of Asian and African Languages* 6, 1976, 95–104.

Wulff, Kurt, *Chinesisch und Tai. Sprachvergleichende Untersuchungen* [Chinese and Tai: comparative linguistic studies], Copenhagen: Levin and Munksgaard, 1934.

Zhāng Jìmín 张济民, *Gēlǎoyǔ yánjiū* 仡佬语研究 [Studies on Gēlǎo languages], Guìyáng 贵阳: Guìzhōu mínzú 贵州民族出版社, 1993.

Zhāng Jūnrú 张均如, *Shuǐyǔ jiǎnzhì* 水语简志 [Sketch of the Sui language], Běijīng 北京: Mínzú 民族出版社, 1980.

Zhōngguó shèhuì kēxuébào 中国社会科学报 [China Social Science News], "Tànxún 'huóde' xiàngxíng wénzì dàshān shēnchù de shuǐjiā yìjīng shuǐshū 探寻 '活的' 象形文字大山深处的水家易经水书" [In search of the living pictorial orthography: the book of change of the Sui people], 2012, http://news.xinhuanet.com/edu/2012-07/06/c_123380689.htm, and http://history.gmw.cn/2012-10/24/content_5469263_4.htm.

Zhōu Fúróng 周芙蓉 and Lǐ Zhōngjiāng 李忠将, "Zhuānjiā fāxiàn 'shuǐshū' kěnéng yǔ zhōuyì yǒu mìqiè liánxì 专家发现 "水书" 可能与《周易》有密切联系" [Experts suggest a link between Sui writing and I-Ching], *Xinhua News* 13-3-2005, http://news.xinhuanet.com/ent/2005-03/13/content_2690151.htm.

Yongxian Luo

Táiwān Spoken Chinese Corpus

1. Introduction

The Taiwan Spoken Chinese Corpus, also known as The NCCU Corpus of Spoken Chinese, is a project of documenting spoken Mandarin, spoken → Hakka (Kèjiā 客家), and spoken Southern Mǐn (Mǐnnán 閩南), whereby open online access to the data is provided for non-profit-making research and teaching. Taken together, Mandarin, Hakka, and Southern Mǐn are spoken by the majority of the Táiwān population, and only a small fraction (about 2%) speaks indigenous languages. Documentation of the spoken varieties of these languages guided by corpus linguistic principles is of great significance, first, because they are undergoing changes in many respects, and, second, because the population of speakers of Southern Mǐn and Hakka is diminishing.

The NCCU Corpus of Spoken Chinese comprises three sub-corpora: the Corpus of Spoken Mandarin, the Corpus of Spoken Hakka, and the Corpus of Spoken Southern Mǐn. As a language documentation project, the corpus focuses on collecting and archiving spoken forms of various types. The recording began in 2006, mostly aiming at spontaneous face-to-face conversations, which is the most common type of language use in daily communication, but other speech varieties have also been collected and documented.

In collecting casual conversational data, a sociolinguistic stratification was adopted as sampling strategy, taking gender and age into consideration. There are three age ranges: from 18 to 30 years old, from 31 to 45 years old, and over 50 years old.

The infrastructure of the NCCU corpus was designed in a simple yet user-friendly way, so that data can be processed efficiently in the database, and users can browse the spoken data from the web. The three corpora share a common scheme of data collection, as detailed below:

Audio-video recording: The participants must have signed a consent form before the recording. The participants are free to develop the topics of the conversation. They are filmed for approximately an hour.

Excerpt selection: One section from each conversation, about twenty minutes long, in which the participants were more comfortable in front of the camera, is then extracted.

Annotation and orthographic transcription: For each excerpt, speaker identity, turns, overlaps, pauses, and code-switching are annotated. Most of the speech sounds are transcribed using Chinese characters. The spoken data is segmented into turns.

Phonetic transcription: The phonetic transcriptions of Mandarin, Hakka, and Southern Mǐn are made using pīnyīn, the Táiwān Hakka Tōngyòng Romanization System proclaimed by the Ministry of Education in 2003, and the Táiwān Southern Mǐn Romanization System proclaimed by the Ministry of Education, respectively. They indicate speakers' actual pronunciation. For instance, some would pronounce the Mandarin distal demonstrative 那 as *nà*, others would say *nèi*. We also try to annotate phonetic details, including the change of tones from 33 to 23 in Mandarin (i.e., when two third-tone words occur in succession, the first word will be pronounced in the second tone, e.g., *nǐ hǎo* 你好 > *ní hǎo*; → Tone Sandhi), the reduction of the Mandarin *zhè-yang* 這樣 'this way' to *jiàng*. Finally, in case of slips of the tongue, the phonetic transcriptions will represent the wrong pronunciations.

Audio clipping: For each turn, the original recording is segmented and saved in MP3 format.

2. The Corpus of Spoken Mandarin

Mandarin is also called Guóyǔ 國語 in Táiwān. The Corpus of Spoken Mandarin currently contains short oral narratives and daily face-to-face conversations. The cartoon narrations were produced by 22 NCCU undergraduate students in 2002. Each subject viewed a 7-minute cartoon episode of the "Mickey Mouse and Friends" series. The soundtrack of the cartoon included music and only a very small amount of dialogue. In the episode in question, Mickey, Minnie, Pluto, and a bull are holding a party at the beach, eating and playing around. They then have a fight with an octopus, which they finally win. The subject immediately recounted the story from memory to a listener after viewing the cartoon. The subject was filmed by a video camera so that speech and manual movements could be recorded. The subjects were not informed about our particular research interests. The elicited cartoon narrations ranged from about 2 to 10 minutes in length; the mean length is 4 minutes 30 seconds. The 6 oral narratives on the web total about 30 minutes of talk. With regard to conversations, casual conversations among family members, friends, and colleagues have been videotaped since 2006. All the participants were paid. The participants were free to find and develop topics of common interest; they were filmed for approximately an hour with a visible camera. One stretch from each talk, of about 20 to 40 minutes, in which the participants were comfortable in front of the camera, was selected for transcription. The 36 conversational excerpts on the web total about 810 minutes of talk, and the first 10 transcripts further include word-for-word English glosses and English translation for every turn.

3. The Corpus of Spoken Hakka

The data in this corpus contain several subdialects of Hakka in Táiwān. The excerpts in the NCCU Corpus of Spoken Hakka comprise TV talk programs, face-to-face conversations and frog stories, base on *Frog, Where are You?* by Mercer Mayer (1969). The copyright of talk programs of Hakka Television was authorized by Council for Hakka Affairs, and those of face-to-face conversations and frog stories were warranted by the participants. Each recording of the TV talk programs and face-to-face conversation was videotaped about 60 minutes, and an excerpt of about 20 minutes was extracted from the most natural and spontaneous part of the conversation. Each recording of the frog stories was videotaped about 10 minutes. Then each excerpt was transcribed and annotated with pause, overlap, and code switch, and was marked with its genre, type, recording information, participant, and topic. In addition, each turn in each excerpt was supplied with a sound file. Then recordings were converted into website pages via a database management system.

So far, we have collected 87 recordings, 44 of which are transcribed and ready for open accessibility, including 15 face-to-face conversations and 29 frog stories. As for the subdialect spoken, 28 belong to Sìxiàn 四縣, 14 to Hǎilù 海陸, 1 to Dàbù 大埔, and 1 being multi-subdialectal (Sìhǎi 四海). A total of 134 participants, 80 females and 54 males, participated in this project. Among the 80 female subjects, 38 speak Sìxiàn, 39 speak Hǎilù, and 3 speak Sìhǎi, while among the 54 male subjects, 31 speak Sìxiàn, 11 Hǎilù, 3 Dàbù, and 9 Ráopíng.

When transcribing, recorders encountered problems such as missing codes and missing characters, as well as questions which character to select. The stopgap for the former was using romanizations based on the Táiwān Hakka Romanization System rather than characters and that for the latter the Hakka Dictionary proclaimed by the Ministry of Education in 2006 was followed. Thus, the codes and characters in our corpus were kept consistent without confusion.

4. The Corpus of Spoken Southern Mǐn

Southern Mǐn (or Mǐnnányǔ 閩南語) is also called "Taiwanese" in Táiwān. Although it is anticipated that the composition of a complete spoken corpus of a language should contain both dialogues and monologues, at the current stage it is the former that constitute the NCCU Corpus of Spoken Southern Mǐn. Also, since people pay least attention in casual and emotionally charged speech, which is considered to be prototypical of a spoken corpus, vernacular Southern Mǐn, which is most frequently used in casual face-to-face spontaneous conversations to talk about topics related to daily affairs and social issues, was chosen to be the primary target of this corpus.

One of the major goals of this corpus is to locate sociolinguistic variation and change of Southern Mǐn in order to examine whether such diversity leads to shifting or even death of Southern Mǐn in Táiwān, in particular the Taipei Area (Taipei 臺北 [Táiběi]; including Taipei City and some towns of Taipei County surrounding Taipei City). Accordingly, an urbanized geographical location with high population density, a mixture of ethnic groups, and high social mobility, was selected for first-step sampling.

In the first stage of data collection, 18 conversations were collected, reflecting part of a matrix of social stratification based on gender, age, education level, familiarity between interlocutors, genre, and formality of situation. All of these 18 conversations were videotaped. Each of them lasts for at least 60 minutes, but only an excerpt of 20 minutes was extracted for orthographic transcription, Romanization, annotation, and English translation. Also, six of these 18 excerpts were transcribed but not yet fully proofread, and thus not ready for online access.

The major problem which this corpus encounters is inconsistent or wrong choices of Chinese characters for orthographic transcription. To solve this problem, this corpus chose to comply with the mapping system used in the Online Dictionary of Southern Mǐn provided by the Ministry of Education. As for orthographic transcription beyond the reach of the Ministry of Education dictionary, other dictionaries and systems are consulted and used systematically.

To facilitate user access and use, the NCCU Corpus of Spoken Chinese provides a Web Concordancer and Frequency Count. The Web Concordancer can be used to retrieve sorted lists of spoken data from the three corpora; Frequency Count can calculate the number of occurrences of a "character" or a "word" in the corpus as a whole. Two frequency count tables for 27 Mandarin conversational excerpts are provided in pdf on the web: '2016 Character Frequency Table' and '2016 Word Frequency Table'.

To conclude, the establishment of a spoken corpus with a sizable amount of data that reflect people's real and habitual use of language is a lifetime work. The use of NCCU Corpus hopefully encourages further more work in documenting the various spoken forms of existing languages.

Bibliography

Hakka Dictionary [Táiwān Kèjiāyǔ chángyòngcí cídiǎn 臺灣客家語常用詞辭典], Ministry of Education of the Republic of China (MOE) [Zhōnghuá Mínguó Jiàoyùbù 中華民國教育部], available on: http://hakka.dict.edu.tw/hakkadict/ (last accessed 9 September, 2016).

Táiwān Hakka Romanization and Pronunciation Learning Network [Táiwān Mǐnnányǔ luómǎ pīnyīn jí qí fāyīn xuéxí wǎng 臺灣閩南語羅馬拼音及其發音學習網], Ministry of Education of the Republic of China (MOE) [Zhōnghuá Mínguó Jiàoyùbù 中華民國教育部], available on: http://www.ntcu.edu.tw/tailo/ (last accessed 9 September, 2016).

The NCC Corpus of Spoken Chinese, National Chengchi University [Guólì zhèngzhì dàxué 國立政治大學], available on: http://spokenchinesecorpus.nccu.edu.tw/ (last accessed 9 September, 2016).

Kawai Chui, Huei-ling Lai & Hui-Chen Chan