Taiwan Journal of Linguistics
Vol. 6.2, 119-144, 2008

THE NCCU CORPUS OF SPOKEN CHINESE: MANDARIN, HAKKA, AND SOUTHERN MIN*

Kawai Chui and Huei-ling Lai

ABSTRACT

In Taiwan, most people speak Mandarin, Southern Min, or Hakka. Not only are the three Chinese dialects undergoing linguistic changes, but the population of Southern Min and Hakka is also diminishing. The NCCU Corpus of Spoken Chinese is thus a project of language documentation whereby open online access to Mandarin, Hakka, and Southern Min data is provided for non-profit-making research.

As a language documentation project, the NCCU spoken corpus focuses on collecting and archiving spoken forms of various types. It consists of three sub-corpora, namely the Corpus of Spoken Mandarin, the Corpus of Spoken Hakka, and the Corpus of Spoken Southern Min. The three corpora share a common scheme for the collection of spoken data, mostly in the form of spontaneous face-to-face conversations. The infrastructure of the corpus is 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 directly from the web. We hope that our work can encourage more people to engage in building up spoken corpora from different perspectives and for different purposes.

1. INTRODUCTION

Establishing spoken corpora based on audio-video recordings of people's daily use of language in face-to-face communication is the most

* The earlier version of this paper was presented at the 7th Annual Wenshan International Symposium on May 19, 2008. We would also like to thank the two anonymous referees for offering valuable comments and suggestions. All errors of interpretation are our own responsibility.

effective way to document the contemporary forms of languages. According to Sinclair (1991:15), "most corpora keep well away from the problems of spoken language—with some honorable exceptions—and, for a corpus which in any way purports to reflect a 'state of the language,' this is most unfortunate." A large number of teachers as well as language scholars believe that the spoken form of the language is a better guide to the fundamental organization of the language than the written form. Sinclair (1991) further claims that "there is no substitute for impromptu speech." Similarly, Crowdy (1993) states that "the importance of conversational dialogue to linguistic study is unquestionable: it is the dominant component of general language both in terms of language reception and language production."

Many spoken English corpora have been established for different research and educational purposes. The London-Lund Corpus of Spoken English contains spoken British dialogues and monologues. The Lancaster Speech, Writing and Thought Presentation Spoken Corpus, contains conversations and oral narratives with annotations for categories of speech, thought, and writing. Ten percent of the British National Corpus consists of orthographic transcriptions of informal conversations and verbal interactions in formal business, government meetings, radio shows, and phone-ins. The corpus of British Academic Spoken English contains lectures and seminars recorded in many different university departments. The spoken corpora of the Cambridge International Corpus consist of the Cambridge and Nottingham Corpus of Discourse in English, the Cambridge and Nottingham Spoken Business English corpus, and the Cambridge Cornell Corpus of Spoken North American English. As to spoken American English, the Michigan Corpus of Academic Spoken English provides transcripts of academic speech events at the University of Michigan. The Corpus of Spoken Professional American-English has two sub-corpora: one consists of academic discussions; the other contains transcripts of White House press conferences. The Santa Barbara Corpus of Spoken American English records natural conversations from throughout the United States. The International Corpus of English contains spoken and written corpora collected in twenty countries and regions. In addition to English, corpora have also been built up for other languages, such as Bosnian (Oslo Corpus of Bosnian Texts), Bulgarian (Corpus of Spoken Bulgarian), Dutch (Spoken Dutch Corpus), German (NEGRA Corpus), Israeli (Corpus of Spoken Israeli Hebrew), Italian (CORIS Corpus), Spanish

(CRATER Spanish Corpus), and Swedish (Spoken Language Corpus of Swedish).

Chinese is a group of languages of the Sino-Tibetan family. The seven major dialects are Mandarin, Min, Hakka, Wu, Cantonese, Xiang, and Gan. Despite its large population, there are very few spoken archives of the wide varieties of Chinese (see Section 3). In Taiwan, most people speak Mandarin, Southern Min, or Hakka, while about 1.9% of the total population speak indigenous languages. The three Chinese dialects are not only undergoing linguistic changes, but the population of Southern Min and Hakka is also diminishing. The NCCU Corpus of Spoken Chinese is thus a project of language documentation whereby open online access to spoken Mandarin, spoken Hakka, and spoken Southern Min data is provided for non-profit-making research. This paper introduces the project for the establishment of spoken corpora for the three major Han dialects spoken in Taiwan at National Chengchi University (NCCU). We hope to encourage more involvement in documenting the various spoken forms of existing languages. Section 2 discusses some of the Chinese corpora published on the internet. Section 3 introduces the content of the NCCU Corpus of Spoken Chinese and its overall infrastructure. The last section provides concluding remarks.

2. CHINESE CORPORA

The Chinese corpora to be introduced here are published on the internet. They include different types of linguistic data in different dialects. First, the corpus of Chinese Pear Stories provides oral narratives of the Pear Story (Chafe 1980) across the seven Chinese dialects with audio files. Storytellings in Mandarin, Cantonese, Hakka are represented by Chinese characters and phonetic transcription; those in Wu, Min, Xiang and Gan include an orthographic transcription only. The data in the Hong Kong Cantonese Adult Language Corpus (HKCAC) comprises both the orthographic and phonetic transcriptions of call-in programs and forums on the radio from 1998 to 2000. The corpus also provides 'single character' and 'sentence' online search functions. Data can be downloaded and displayed in Excel. The Southern Wu dialect spoken in Wenzhou is documented in the Wenzhou Spoken Corpus. It consists of data from face-to-face conversations, phone calls, Wenzhou news

commentaries, internet chats, stories, and Wenzhou songs. Concordance and collocates searches are provided with technical support from the Text Analysis Portal for Research, a collaboration by six Canadian universities to build a centralized gateway to texts and sophisticated text analysis tools.

The Lancaster Corpus of Mandarin Chinese (LCMC) is designed as a Chinese match for the FLOB and FROWN corpora for modern British and American English in order to carry out cross-linguistic studies. The written texts include news, literary texts, academic prose, and official documents published in mainland China in the early 1990s. The full corpus can be accessed online using the LCMC Corpus Web Concordancer. The UCLA Corpus of Written Chinese is a recent update of the Lancaster Corpus of Mandarin Chinese for cross-linguistic and diachronic studies of written Chinese. The corpus consists of written texts available on the internet from 2000 to 2005. The data can also be accessed via the LCMC Corpus Web Concordancer.

In Taiwan, Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) inclues texts which have been collected since 1990 from different areas. Spoken data are just a small part of the corpus, consisting of speech drafts, play scripts, actors' lines, conversations, and written records of speeches and meetings. The corpus is open to the research community through internet. Just like the Sinica Corpus, the Mandarin spoken corpora project is also part of the Language Archives Project at Academia Sinica. Three corpora are being developed: the Mandarin Conversational Dialogue Corpus (MCDC), the Mandarin Topic-oriented Conversation Corpus (MTCC), and the Mandarin Map Task Corpus (MMTC). The three corpora each have their own respective annotation systems. All sound recordings are segmented and saved as audio files. The MCDC is now distributed by the Association for Computational Linguistics and Chinese Language Processing. In addition to these Mandarin Chinese corpora, many others have been built up for Mandarin and Southern Min mainly by linguists for individual research, and are not open for public access.

However, despite the fact that Mandarin is the dominant language in Taiwan, its spoken varieties are still under-documented. Even less has been done on spoken Hakka and Southern Min. Therefore, a spoken corpus, consisting of authentic data ranging over different sub-dialects, should be constructed for each of the three languages.

3. THE NCCU CORPUS OF SPOKEN CHINESE

The NCCU spoken corpus consists of three sub-corpora, namely the Corpus of Spoken Mandarin, the Corpus of Spoken Hakka, and the Corpus of Spoken Southern Min. As a language documentation project, the corpus focuses on collecting and archiving spoken forms of various types. Most of the data has been documented since 2006,[1] and is mostly in the form of spontaneous face-to-face conversations—the most common type of language use in daily communication. Other spoken varieties have also been collected and documented in the different sub-corpora. This section first introduces the content in the Mandarin and Hakka spoken data, as well as the overall infrastructure.

3.1 Mandarin data

The Corpus of Spoken Mandarin includes two types of data, namely daily face-to-face conversations and oral narratives. First, Hui-chen Chan[2] devised a sociolinguistic stratification to consider gender and age in collecting casual conversational data. There are three age ranges: from 18 to 30 years old, from 31 to 45 years old, and over 50 years old. The Mandarin spoken corpus currently contains thirty casual recordings. Ten conversations among college students who knew each other, totaling about 3.5 hours, are available on the web. Table 1 is a list of the ten Mandarin excerpts with information about the participants and the length of the texts.

[1] This language documentation project is mainly funded by a grant to National Chengchi University from the Aim for the Top University and Elite Research Center Development Plan from the Ministry of Education. Other grants have also been received from the Humanities Research Center of the National Science Council and the Office of Research and Development at National Chengchi University.

[2] Hui-chen Chan, a member of the team, will discuss the Corpus of Spoken Southern Min in her forthcoming article.

Table 1. Details of the Mandarin conversations

	Participants		Length of excerpt		
	Female	Male	Minutes	Number of turns	Number of characters
M001	2	0	14	422	8688
M002	2	1	24	568	13117
M003	3	0	21	466	8724
M004	3	0	21	405	9869
M005	2	0	40	730	22258
M006	1	1	20	440	9707
M007	2	1	20	386	8463
M008	2	2	20	307	6584
M009	2	1	19	341	8908
M010	2	0	20	394	9163
Total:	21	6	219	4459	105,481

Storytelling is as common as face-to-face talk in daily communication. The Mandarin corpus also comprises short oral narratives. The recordings took place in 2002. They were audio-video-taped in two circumstances (Aboudan and Beattie 1996): 'a social/monologue condition' "where the speaker narrated the story in the presence of a silent interviewer who neither interrupted nor made any verbal exchanges with the speaker" (275), and 'a social/dialogue condition' in which the interviewer "asked questions during the course of the narrations to clarify the story" (276). The participants were undergraduate students of National Chengchi University; the gender of the students was also considered. Each student viewed a 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, Mickey, Minnie, Pluto and a bull are holding a party at the beach, and eating and playing around. Then, they have a fight with an

octopus, which they finally win. After viewing the cartoon, the subject immediately recounted the story from memory to a listener. The elicited cartoon narrations ranged from about two to ten minutes in length, totaling about 1.5 hours. See the details in Table 2.

Table 2. Details of the Mandarin narratives

Social/monologue condition

	Minutes	Seconds	Number of characters		Minutes	Seconds	Number of characters
F1	4	20	2176	M1	6	17	4086
F2	3	25	1738	M2	3	2	1556
F3	2	20	1222	M3	4	20	2587
F4	1	21	738	M4	10	2	6436
F5	1	31	866	M5	3	12	1656
F6	2	45	1493		26	53	16321
	13	162	8233				

Social/dialogue condition

	Minutes	Seconds	Number of characters		Minutes	Seconds	Number of characters
F1	4	44	3702	M1	8	18	6110
F2	6	1	3851	M2	5	2	3084
F3	3	43	2489	M3	3	57	2925
F4	9	3	6451	M4	4	30	3312
	22	91	16493	M5	7	21	6082
				M6	3	38	2879
				M7	6	5	5051
					36	171	29443

The content of the first three turns in the excerpt M001 about schoolwork counseling can be seen in Figure 1.

Figure 1. Screenshot of Mandarin data on the web

3.2 Hakka data[3]

Hakka accents in Taiwan vary from area to area. The Hakka population in Taiwan can be categorized into four accents/sub-dialects according to the areas from which the predecessors of the speakers came: Sixian (mainly used in Taoyuan, Miaoli, Pingdong and Gaoxiong), Hailu (mainly used in Xinzhu and part of Taoyuan), Dapu (exclusively used in Taizhong), and Shaoan (mainly used in Yunlin and Nantou). Other accents/sub-dialects, such as Raoping and Yongding, are also used in a small number of areas. A survey conducted by the Council for Hakka Affairs (henceforth CHA) (2004) shows that most Hakka speakers use more than one accent in their daily communication. Among the various accents used in their speech, Sixian accounts for 49%; Hailu 47.1%; Dapu 8.6%; and the other three below 5%.[4] Another study also conducted by the CHA in 2002 indicates that the use of Hakka has decreased by 5% since 1994.[5] Table 3 represents the language codes among child-parent interaction in Hakka families. In terms of using Hakka as a channel for communication, the percentage of speakers from 13 to 18 years old is almost three times less than that of speakers whose age is above 60.

3 Originally the term *Hakka* was not used to refer to a certain ethnic group living in a particular area. Emerging during the Song Dynasty (960-1279), it was used to indicate "guests" who had left their homelands to settle down in other parts of the country, in contrast to residents originally from the area. As to the formation of Hakka people, two views are held in the literature. One view holds that Hakka people originated from the Central Plains of China and moved southwards mainly to the Southern areas of China due to foreign invasions, civil wars and other historical reasons. Then, some Hakka people migrated to Taiwan around the middle of the century (Hashimoto 1973, Wu 1995, Luo 1998). Another view holds that the Hakka originated in the area of the southern Gan in the Song dynasty, with the Hakka dialects bearing features similar to non-Chinese languages such as She and Yao. Thereafter, some Hakka migrated southwards to Taiwan in the early Qing dynasty (Chappell 2001). A related view also holds that Yue, Hakka and Southern Gan are subdialects of a Guangzhou dialect type (Lau 1999).

4 For more detailed information, please refer to the web page of the Council for Hakka Affairs (CHA) http://www.hakka.gov.tw/public/Attachment/512722563971.pdf.

5 Please refer to http://www.hakka.gov.tw/ct.asp?xItem=29832&CtNode=1657&mp=256&ps=.

Table 3. Language used in daily conversation with parents (from CHA, 2002)

No. of Person / %

	Sampling	Hakka	Mandarin Chinese	Southern Min	Contextually
Total	1986	73.36	16.11	7.91	2.62
Gender					
Male	1002	76.05	13.47	7.98	2.50
Female	984	70.63	18.80	7.83	2.74
Age					
13-18	137	34.31	59.85	4.38	1.46
19-29	274	41.97	43.80	8.39	5.84
30-39	441	71.20	15.65	9.98	3.17
40-49	464	81.68	6.68	9.70	1.94
50-59	411	89.78	3.41	5.84	0.97
above 60	259	89.96	1.54	5.79	2.70

Further studies (conducted by the CHA 2005, 2006) indicate that although 43.6% of native Hakka speakers have no problem with listening comprehension or speaking ability, the ratio of fluent speakers under thirteen years old only accounts for 11.6%.

The above statistics disclose that the Hakka language is encountering a severe situation such that it may vanish in due course without any language policies. In order to halt the declining trend, various government and also folk organizations and academics have set up policies in the past seven years.[6] As a growing number of scholars and cultural workers have begun to pay attention to the research on Hakka, in which linguistic or language-related studies play an important part, the need for the collection and systematic storage of a considerable amount

6 The CHA, under the Executive Yuan, was established with the purpose of preserving Hakka culture and language. Moreover, Hakka Television (Channel 17) offers programs about Hakka cultures and languages. In addition, folk organizations such as Hakka Magazine, Hakka News Magazine and Hakka Taiwanese Special Magazine as well as the groups of academics such as the Taiwan Languages and Literature Society also contribute to Hakka affairs and enlarge the development of Hakka studies. For more information about organizations and academics concerned, please refer to the web sites listed in our references.

of authentic language data has been brought to the foreground. Cross-linguistically, researchers have started to rely on corpora consisting of naturally occurring languages to conduct quantitative linguistic studies. However, compared with the large scale research on Mandarin Chinese in Taiwan, the research resources for Hakka studies are quite insufficient, and consequently it is rare to find a corpus-based study on Hakka. More overall and in-depth exploration of Hakka remains a challenging task because of this scarcity of language materials. Therefore, a Hakka corpus, consisting of a great deal of authentic data ranging over different sub-dialects should be constructed in order to pave the ground for Hakka studies. Equipped with systematic and easily-accessible data, the corpus can help researchers not only to present a complete picture of the Hakka language, but also to examine the similarities and differences among each sub-dialect of Hakka. In addition to strengthening the foundation of linguistic studies in Hakka, the construction of the corpus can also enhance research in other fields, including dialectology, dialect geography, language and culture, ethnology, sociology and literature. The effect brought about by the use of a Hakka corpus is highly-anticipated. Noticing both the usefulness and the urgency of constructing corpora of endangered languages, the Academia Sinica has initiated long-term projects for the compilation of languages archives, including Hakka and Southern Min. In the six-year longitudinal project (2007-2012), data for Hakka and Southern Min will be collected from various resources, including the lyrics for songs, and other literary works. However, up to now little attention has been paid to work on spoken Hakka or Southern Min corpora. While setting up similar goals, the NCCU Corpus of Spoken Hakka was started, to include the challenging task of constructing a spoken corpus for Hakka.

Regarding the Corpus of Spoken Hakka, the excerpts comprise TV talk programs and face-to-face conversations. The copyright of the talk programs from Hakka Television is authorized by the CHA. Ten recordings have been collected: five of ten are in Sixian, three are in Hailu, one is in Dapu, and one is multi-sub-dialectal. A total of 26 participants, 14 females and 12 males, participate in this project. Among the female subjects, ten speak Sixian, and four speak Hailu and among the male subjects, seven speak Sixian, two speak Hailu, and three speak Dapu. The distribution of our data is shown in Figure 2 and Figure 3 below.

Figure 2. Percentage of recordings of different sub-dialects

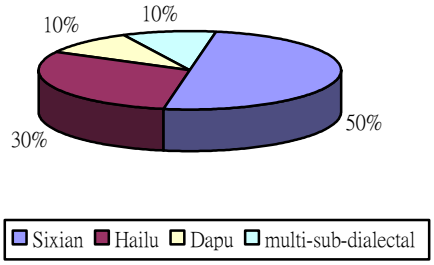
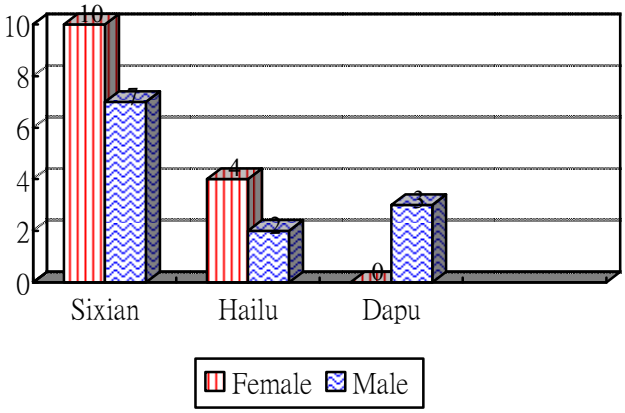


Figure 3. Gender of participants and type of sub-dialect



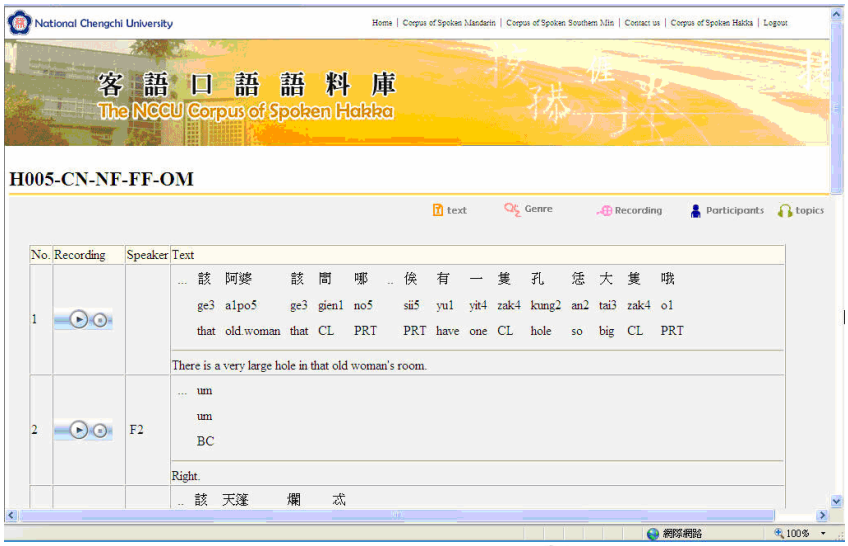
Presently, the corpus includes five Hakka excerpts, one from a TV talk program and four from face-to-face conversations. Table 4 lists the information about the five excerpts.

Table 4. Details of Hakka excerpts

	Participants		Length of excerpt		
	Female	Male	Minutes	Number of turns	Number of characters
H001	2	1	20	769	11848
H002	2	0	20	384	10849
H003	3	0	22	417	12460
H004	0	2	23	389	11129
H005	2	0	20	670	14931
Total:	9	3	105	2629	61,217

Figure 4 illustrates the first two turns in the excerpt 'Origins of Substitutes for Dad and Mom'.

Figure 4. Screenshot of spoken Hakka data on the web



During the construction of the NCCU Corpus of Spoken Hakka, the fact that Hakka is an endangered language increases the difficulty of data collection. As demonstrated in Table 3 above, the lower frequency of the middle and young age groups in speaking Hakka may result in unbalanced sampling and data distribution. Finding subjects in the lower age groups is a task to be pursued in the future. In addition, another future task is to obtain data from different genres. Possible arrangements of various genres are shown in the following table:

Table 5. Future planning of various genres in the database

Dialogues	▪ News Discussions ▪ Call-in TV programs
Monologues	▪ Lectures ▪ Storytelling ▪ Oral History ▪ News Manuscripts

3.3 Corpus infrastructure

The infrastructure of the NCCU corpus is 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 below:

Audio-video recording

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

Excerpt selection

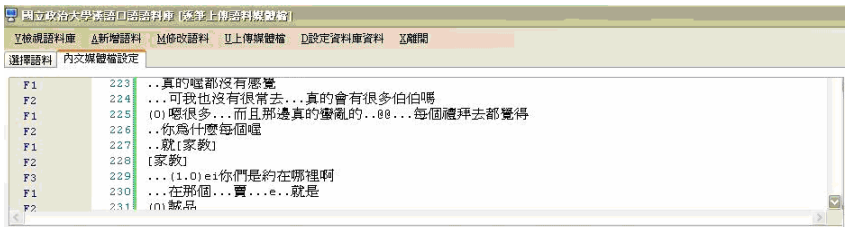
One section from each conversation, about twenty minutes of talk, in which the participants were more comfortable in front of the camera, is then extracted.

Annotation and orthographic transcription

For each excerpt, speaker identity, turns, overlaps, pauses, and code-switching are annotated. (Please see Appendix B.) Speech sounds

are transcribed into Chinese characters. The spoken data is segmented into turns. Figure 5 indicates nine turns of interaction among three Mandarin female speakers.

Figure 5. Screenshot of annotation and orthographic transcription of Mandarin data in the database

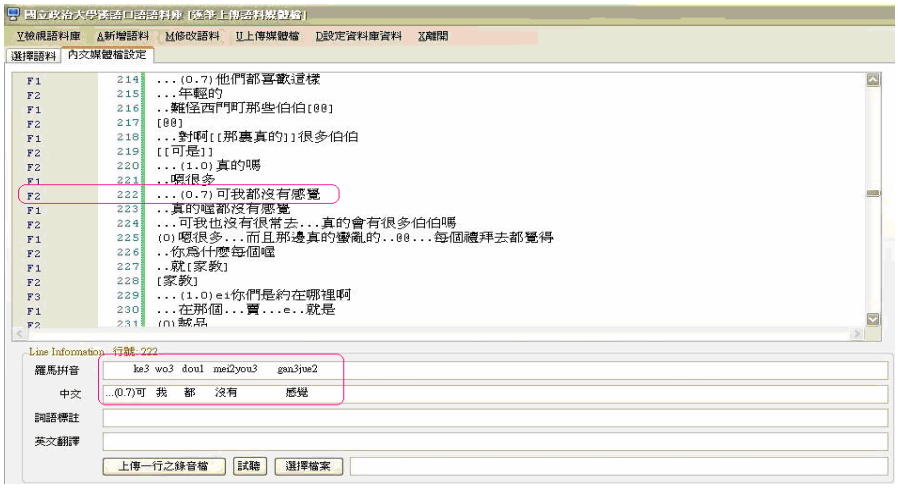


Phonetic transcription

The phonetic transcriptions of Mandarin, Hakka, and Southern Min follow the Pinyin system, the Taiwan Hakka Tongyong Romanization System proclaimed by the Ministry of Education in 2003, and the Taiwan Southern Min Romanization System proclaimed by the Ministry of Education, respectively. They by and large indicate speakers' actual pronunciation. For instance, some would pronounce the Mandarin distal demonstrative as *na*; some would say *nei*. We also try to annotate as many phonetic details as possible, including the change of tones from '33' to '23' in Mandarin, the reduction of the Mandarin proximal demonstrative from *zhe-yang* to *jiang*. Finally, in case of slips of the tongue, the phonetic transcriptions still represent the wrong pronunciations which would be glossed as 'REPAIR'.

Figure 6 exemplifies the phonetic transcription for Turn 222 可我都沒有感覺 'but I had no feeling at all' in the database.

Figure 6. Screenshot of phonetic transcription in the database



English glossing

Word-for-word glosses in English are provided. As shown in Figure 7, the words in Turn 222 are provided with English glosses or grammatical abbreviations (see Appendix A).

Figure 7. Screenshot of English glossing of Mandarin data in the database

English translation

For every turn, English translation is provided. The English translation for the whole of Turn 222 can be seen in Figure 8.

Figure 8. Screenshot of English translation of Mandarin data in the database

Audio clipping

For each turn, the original recording is segmented and saved in MP3 format. See Figure 9.

Figure 9. Screenshot of audio clippings of Mandarin data in the database

Figure 10 and Figure 11 indicate the documentation of Hakka and Southern Min data in the database, respectively.

Figure 10. Screenshot of phonetic transcription, English glossing, and English translation of Hakka data in the database

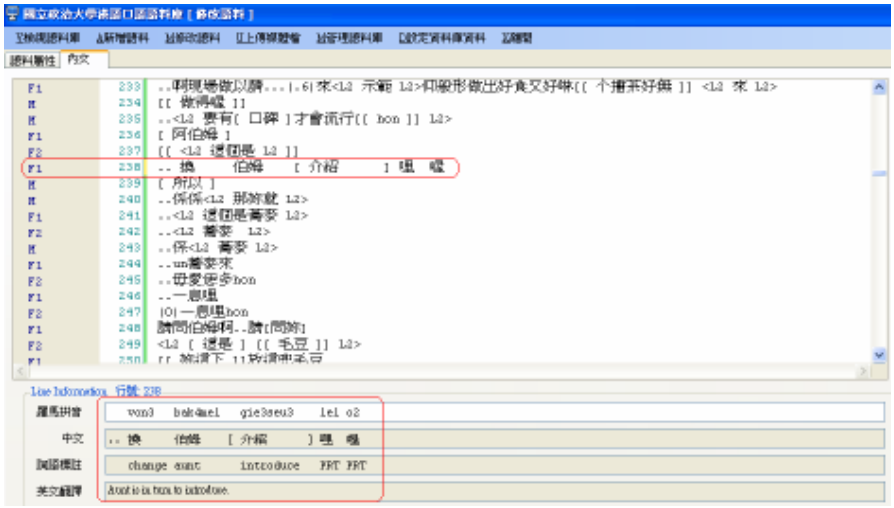
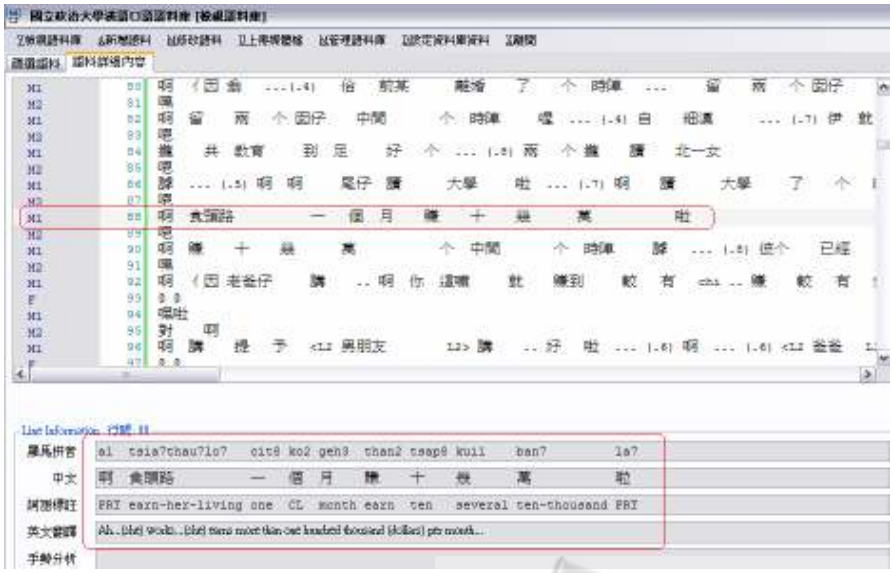


Figure 11. Screenshot of phonetic transcription, English glossing, and English translation of Southern Min data in the database



4. CONCLUDING REMARKS

International efforts have been undertaken to establish spoken corpora for under-documented and endangered languages. For instance, the Survey of California and Other Indian Languages, a research project in the UC Berkeley Department of Linguistics, has conducted language documentation and archiving since 1952. The Language Documentation Training Center at the University of Hawai'i at Manoa provides linguistic workshops, language documentation techniques, and relevant computer software to help students document their own languages. The Hans Rausing Endangered Languages Project at SOAS, University of London, grants funds for documentation projects, teaches the theories and skills necessary for language communities and academic researchers to carry out linguistic fieldwork and documentation, to preserve documentation materials, and to use such materials. The Volkswagen Foundation provided funding for the DOBES program (Documentation of Endangered Languages) at the Max-Planck-Institute for Psycholinguistics in 2000 as an international research project for the documentation of endangered languages. In Taiwan, the Language Archive Project at Academia Sinica, which started in 2001, also grants funds to support archivial work of the languages on the island.

It takes a very long time to accomplish the establishment of a spoken corpus with a sizable amount of conversational data. Nonetheless, such data serves to provide examples of spontaneous verbal interactions that reflect people's real and habitual use of language. However, although spoken Mandarin is collected for the NCCU language project and the Sinica Mandarin Spoken Corpora project, the released data so far is far from sufficient to provide the complete usage of this major language spoken in Taiwan. Although Min data is collected for both the NCCU language project and the Southern Min Archives under the Language Archive Project at Academia Sinica, the Southern Min Archives is a database containing Min dramas, plays, and operas, and even Hakka songbooks, for the purpose of studying historical change and language distribution. The NCCU language project, on the other hand, aims to document people's daily use of Southern Min and Hakka in conversational interactions.

We have been working on language documentation for the NCCU spoken corpus since 2006. The corpus does not contain a large amount of data. Nor does it include a lot of textural markups, linguistic tagging and

annotations, or sophisticated computing infrastructure. However, despite the small quantity, our spoken data should suffice for carrying out preliminary observations or pilot studies in research. As a lifetime work, we will continue our archive work in order to gain the entire contemporary profile of using Mandarin, Hakka, and Southern Min in daily communication. We also hope that the use of our work can encourage more people to engage in building up spoken corpora from different perspectives and for different purposes.

REFERENCES

Aboudan, Rima, and Geoffrey Beattie. 1996. Cross-cultural similarities in gestures: the deep relationship between gestures and speech which transcends language barriers. Semiotica 111.3-4, 269-94.

Chafe, Wallace. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

Chappell, Hilary. 2001. Synchrony and diachrony of Sinitic languages: A brief history of Chinese dialects. *Sinitic Grammar*, ed. by Hilary Chappell, 3-28. Oxford: Oxford University Press.

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8.4: 259-265.

Hashimoto, Mantaro J. 1973. *The Hakka Dialect: A Linguistic Study of its Phonology, Syntax, and Lexicon*. Cambridge: Cambridge University Press.

Lau, Chunfat. 1999. Huanyu fangyan de fenlei biaozhun yu Kejiahua zai Hanyu fangyan de fenleishang de wenti [Criteria for the classification of Chinese dialects and the question of the status of Hakka]. Paper presented at the Eighth International Conference on Chinese Languages and Linguistics. Melbourne: University of Melbourne.

Leung, M.-T., and S.-P. Law. 2001. HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics* 6, 305-325.

Luo, Mei-zhen. 1998. The continuity and variation of Hakka language and culture in Taiwan. Proceedings of the Fourth International conference on Hakkaology: Hakka and Modern World, 275-284. Taipei: Academia Sinica.

Lyu, Ren-yuan, Min-siong Liang, and Yuang-chin Chiang. 2004. Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka, and Mandarin. *Computational Linguistics and Chinese Language Processing* 9.2: 1-12.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford; Oxford University Press.

Wang, H. C., F. Seide, C. Y. Tseng, and L. S. Lee. 2000. Mat-2000—design, collection, and validation of a Mandarin 2,000-speaker telephone speech database. Paper presented at the International Conference on Spoken Language Processing 2000. China: Beijing.

Wu, Zhong-jie. 1995. Keyu cifangyan yu keyu jiaoxue [Hakka subdialects and Hakka teaching]. *Papers from the 1994 Conference on Language Teaching and Linguistics in Taiwan Vol. II: Hakka*, ed. by Feng-fu Tsao, and Mei-hui Tsai, 289-306. Taipei: Crane.

Xu, Zhao-quan. 2003. *Hakka Dictionary of Taiwan*. Taipei: SMC Publishing Inc.

WEBSITE RESOURCES

Academia Sinica Balanced Corpus of Modern Chinese
http://www.sinica.edu.tw/SinicaCorpus/index.html

British Academic Spoken English (BASE) corpus
http://www2.warwick.ac.uk/fac/soc/celte/research/base

British National Corpus
http://www.natcorp.ox.ac.uk/

Brown University Corpus
http://icame.uib.no/

Cambridge International Corpus
http://www.cambridge.org/elt/corpus/international_corpus.htm

Chinese Pear Stories
http://pearstories.org/

Collins Cobuild
http://www.collins.co.uk/books.aspx?group=153

CORIS/CODIS Corpus
http://corpora.dslo.unibo.it/coris_eng.html

Corpus of Spoken Bulgarian
http://www.hf.uio.no/ilos/studier/studenttjenester/Nettressurser/bulg/mat/Nikolova/

Corpus of Spoken Israeli Hebrew
http://www.tau.ac.il/humanities/semitic/cosih.html

Corpus of Spoken Professional American-English
http://www.athel.com/cpsa.html

Council for Hakka Affairs
http://www.hakka.gov.tw/

CRATER Spanish Corpus
http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

Formosan Language Archive
http://formosan.sinica.edu.tw

Hakka Television
http://www.hakkatv.org.tw/

Hakka Magazine
http://www.hakka.url.tw/yellowpage/

Hakka News Magazine
http://www.pts.org.tw/php/html/hoga/main.php

Hakka Taiwanese Special Magazine
http://home.i1.net/~alchu/hakka/hakkafa1.htm

Helsinki Corpus of English Texts
http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html

Hong Kong Cantonese Adult Language Corpus
http://shs.hku.hk/corpus/index.htm

International Corpus of English
http://www.ucl.ac.uk/english-usage/ice/index.htm

Lancaster Corpus of Mandarin Chinese
http://score.crpp.nie.edu.sg/laohong/LCMC.htm

Lancaster Speech, Writing and Thought Presentation Spoken Corpus
http://www.ahds.ac.uk/catalogue/collection.htm?uri=lll-2464-1

Lancaster/Oslo-Bergen Corpus
http://icame.uib.no/

Lancaster-Los Angeles Spoken Chinese Corpus
http://www.corpus4u.org/showthread.php?t=834

Language Archives Project
http://languagearchives.sinica.edu.tw/

London-Lund Corpus of Spoken English
http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

Mandarin spoken corpora project
http://mmc.sinica.edu.tw/home_e.htm

Michigan Corpus of Academic Spoken English
http://quod.lib.umich.edu/m/micase/

NEGRA Corpus
http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/

Oslo Corpus of Bosnian Texts
http://www.tekstlab.uio.no/Bosnian/Corpus.html

Santa Barbara Corpus of Spoken American English
http://www.linguistics.ucsb.edu/research/sbcorpus.html

Southern Min Archives
http://SouthernMin.sinica.edu.tw/

Spoken Dutch Corpus
http://lands.let.kun.nl/cgn/ehome.htm

Spoken Language Corpus of Swedish
http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=1

Survey of California and Other Indian languages
http://www.linguistics.berkeley.edu/Survey/index.html

Taiwan Languages and Literature Society
http://210.240.194.138:8080/index.asp

UCLA Corpus of Written Chinese
http://score.crpp.nie.edu.sg/laohong/UCLA.htm

Wenzhou Spoken Corpus
http://corpora.tapor.ualberta.ca/wenzhou/

York-Toronto-Helsinki Parsed Corpus of Old English Prose
http://www-users.york.ac.uk/~sp20/

Kawai Chui
Department of English
National Chengchi University
Taipei, Taiwan 116, ROC

Huei-ling Lai
Department of English
National Chengchi University
Taipei, Taiwan 116, ROC

APPENDIX A: ABBREVIATIONS OF LINGUISTIC TERMS

1PL	first person plural
1SG	first person singular
2PL	second person plural
2SG	second person singular
3PL	third person plural
3SG	third person singular
ACMPL	accomplishment aspect
ASSC	associative morpheme
BA	the morpheme BA
BC	backchannel
CL	classifier
COMPARE	compare morpheme
COMPL	complementizer
COP	copula verb
DLM	delimitative aspect
EMP	emphatic adverbial
EXP	experiential aspect
NEG	negative morpheme
PF	pause filler
POSS	possessive
PRF	perfective aspect
PROG	progressive aspect
PRT	discourse particle
QST	question particle
REPAIR	repair phoneme(s)
SELF	reflexive morpheme

APPENDIX B: TRANSCRIPTION CONVENTIONS

:		speaker identity/turn start
[]		speech overlap
...(N)		long
...		medium
..		short
(0)		latching
@		laughter
<L2	L2>	code-switch to English
<L3	L3>	code-switch to Southern Min
<L4	L4>	code-switch to Japanese
<L5	L5>	code-switch to Hakka
<L6	L6>	code-switch to Mandarin