Taiwan Child Language Corpus: Data Collection and Annotation

Jane S. Tsay
Institute of Linguistics, Chung Cheng University
Min-Hsiung, Chia-Yi 621, Taiwan
lngtsay@ccu.edu.tw

Abstract

Taiwan Child Language Corpus contains scripts transcribed from about 330 hours of recordings of fourteen young children from Southern Min Chinese speaking families in Taiwan. The format of the corpus adopts the Child Language Data Exchange System (CHILDES). The size of the corpus is about 1.6 million words. In this paper, we describe data collection, transcription, word segmentation, and part-of-speech annotation of this corpus. Applications of the corpus are also discussed.

1 Data Collection

Taiwan Child Language Corpus (TAICORP) is a corpus of text files transcribed from the child speech recorded between October 1997 through May 2000. The target language is Southern Min Chinese spoken in Taiwan.

1.1 Children

All fourteen children participated were from Taiwanese-speaking families in Min-Hsiung Villlage, Chiayi County, Taiwan.

There were nine boys and five girls, aged from one year two months to three years and eleven months at the beginning of the project. More than half of the children were recorded over more than two years.

1.2 Recordings

The recordings were made through regular home visits. Spontaneous speech of these children at play was recorded using Mini Disc recorders. The interval of the sessions was about two weeks. There were totally 431 recording sessions, each 40 to 60 minutes long, totaling about 330 hours.

1.3 Transcription

Each recording session was transcribed into a separate text file, using Chinese orthography. For words that do not have a conventionalized written form, the Taiwan Southern Min romanization system, i.e., Taiwan Southern Min Pinyin was used.

About half of the sessions (from children under two and a half years old) also have phonetic transcription in unicode IPA (International Phonetic Alphabet).

The three primary transcribers, who were also the investigators who did the recordings, were well-trained linguists. All recordings were first transcribed by the investigator of the specific session and then checked by the other two transcribers.

2 Text files in CHILDES format

TAICORP adopts the format of CHILDES (Child Language Data Exchange System), originally set up by Elizabeth Bates, Brian MacWhinney, and Catherine Snow, to transcribe and code the recordings of child speech into machine-readable text (MacWhinney & Snow 1985, MacWhinney 1995).

The main components of CHILDES format are ***headers*** and ***tiers***.

2.1 Headers

Obligatory headers are necessary for every file. They mark the beginning, the end and the participants of the file.
Constant headers mark the name of the file and the background information of the children.

Changeable headers contain information that can change within the file, such as the recording date, duration, coders and so on.

These headers begin with @, for example:

Obligatory headers:
@Begin
@End
@Participants

Constant headers:
@Age of XXX:
@Birth of XXX:
@Coder:
@Educ of XXX:
@Filename:
@ID:
@Language:
@Language of XXX:
@SES of XXX: social and economic status of a specific speaker
@Sex of XXX:
@Warning: the defects of the file

Changeable headers:
@Activities:
@Comment:
@Date:
@Location:
@New Episode:
@Room Layout:
@Situation:
@Tape Location:
@Time Duration:
@Time Start:

2.2 Tiers

The content of a file is presented in tiers, including main tiers and dependent tiers. A main tier, indicated by *, contains the utterance of the speaker.

Main tiers

The main tiers used in TAICORP include the following:

* INV: the utterance of the investigator
* CHI: the utterance of the target child
* MOT: the utterance of mother
* FAT: the utterance of father
* SIS: the utterance of sister
* BRO: the utterance of brother
* GRM: the utterance of grandmother
* GRF: the utterance of grandfather
* OTH: the utterance of other people

The main tier is the most important tier because it is where the utterances are listed. The utterances in the main tier were transcribed in the romanization (pinyin) system of Taiwan Southern Min (to be explained and illustrated in Section 5).

Dependent Tiers

Additional information is given in dependent tiers, indicated by %, following the main tier. Dependent tiers can be changed according to the design and goals of each corpus.

The dependent tiers used in TAICORP include the following:

%ort: transcription in standard orthography
%cod: part-of-speech coding
%pho: phonetic transcription in IPA
%ton: tone value in 5-point scale

For adults' speech, only %ort and %cod tiers are used. For younger children's speech, %pho and %ton tiers are also used. The following text is an example from TAICORP. ([m…] = speech in Mandarin; SHI = 是 "be")

@Begin
@Participants: CHI Lin Target_Child, INV Rose Investigator, MOT Mother, OTH Great Grandmother
@Age of CHI: 2;1.22
@Birth of CHI: 28-AUG-1995
@Sex of CHI: Male
@Coder: Rose, Kay, Joyce
@Language: Taiwanese
@Date: 20-OCT-1997
@Tape Location: Lin D1-1-56
@Comment: Time Duration 37 minutes
@Location: Chiayi, Taiwan
@Transcriber: Rose
@Comment: Track number is D1-1

***INV: bo2lin2@s [:=m], li2 tha5tu2a2 khi3 to2?**
%ort: [m 柏林], 你頭拄仔 去 陀?

%cod: Nb Nh Nd VCL Ncd
***CHI: hia1/hin1.**
%ort: 遐 1.
%cod: Ncd
%pho: h i a
%ton: 55
***INV: hia1 si7 to2ui7?**
%ort: 遐 是 陀位?
%cod: Ncd SHI Ncd
***CHI: hm0.**
%ort: hm0.
%cod: I
%pho: ??
%ton: ??
***MOT: li2 kin1a2 ciah8 bi2ko1 si7 bo0?**
%ort: 你 今仔 1 食 米糕 是 無 3?
%cod: Nh Nd VC Na SHI T
@End

3 Statistics of the corpus

The corpus size is about 1.6 million words (more than 2 million morphemes/Chinese characters). The number of utterances/lines, words, mean length of utterances (MLU) are listed in Table 1.

	Lines	Words	MLU
Children	161,253	434,557	2.695
Adults	336,173	1,211,946	3.605
Total	497,426	1,646,503	3.150

Table 1 Statistics of the corpus

It might be worth mentioning that the MLU of adults in this corpus is relatively short. This could be attributed to the nature of this corpus as being child-directed speech.

4 Part-of speech annotation

Southern Min and Mandarin are both Sinitic languages. They are very similar in their morphology and syntactic structures. Therefore, we adopted the part-of-speech coding system of the Sinica Corpus, Academia Sinica, Taiwan (see various CKIP technical reports). However, among the 115 categories used in the Sinica Corpus (CKIP 1993), only 46 codes were used in TAICORP. In other words, categorization in TAICORP is broader. These codes are listed in Table 2.

Table 2 Part-of-Speech Tagset in TAICORP

Coding	Part-of-speech
A	non-predicative adjective
Caa	coordinate conjunction
Cab	listing conjunction
Cba	conjunction occurring at the end of a sentence
Cbb	following a subject
Da	possibly preceding a noun
Dfa	preceding VH through VL
Dfb	following adverb
Di	post-verbal
Dk	sentence initial
D	adverbial
Na	common noun
Nb	proper noun
Nc	location noun
Ncd	localizer
Nd	time noun
Neu	numeral determiner
Nes	specific determiner
Nep	anaphoric determiner
Neqa	classifier determiner
Neqb	postposed classifier determiner
Nf	classifier
Ng	postposition
Nh	pronoun
I	interjection
P	preposition
T	particle
VA	active intransitive verb
VAC	
VB	active pseudo-transitive verb
VC	active transitive verb
VCL	transitive verb taking a locative argument
VD	ditransitive verb
VE	active transitive verb with sentential object
VF	active transitive verb with VP object
VG	classifactory verb
VH	stative intransitive verb
VHC	stative causitive verb
VI	stative pseudo-transitive verb
VJ	stative transitive verb
VK	stative transitive verb with sentential object

VL	stative transitive verb with VP object
V_2	
DE	*special tag for the word "的"
SHI	special tag for the word "是"
FW	foreign words
*Di/T	*marker following pseudo-transitive active verb
*CIT	*special tag for the word "得 2"

5 Orthography-related issues for a speech-based corpus of Southern Min

5.1 Romanization system

As mentioned in Section 2, utterances in the main tier are transcribed in romanization (Southern Min pinyin). The romanization system used in TAICORP is the Taiwan Southern Min Phonetic Alphabetic (also known as Taiwan Language Phonetic Alphabet, TLPA, originally proposed by the Taiwan Language Society in 1991) announced officially by the Ministry of Education of Taiwan in 1998.

5.2 Standard orthography: Chinese characters

Chinese characters are used in the dependent tier %ort as the standard orthography. This is a reasonable way because most of the Southern Min words are cognates of Mandarin words. However, because Southern Min does not have as conventionalized orthography as Mandarin, quite a few words in Southern Min do not have a consistent way of writing them. Some of them don't even have very obvious corresponding Characters.

In order to ensure consistency in the corpus, Southern Min dictionaries were used. These dictionaries are listed after the References.

This issue is particularly important for a corpus based on spontaneous speech, rather than written text. For example, the following common words in Southern Min have to be checked in the dictionary about their written forms because they do not occur in Mandarin:

蠓罩 /bang2tah4/ "mosquito net"
挽 /ban2/ "to pick"
奇巧 /ki5kha2/ "unusual"

If a written form cannot be found in one of the major Southern Min dictionaries, romanization is used.

Romanization is also used if the written form of a word is found in the dictionary but has so low frequency that it can't be found in the computer coding system.

For homonyms, a number is added after the character to indicate different lemmas. For example:

蓋 1 /kah4/ "to cover with a blanket"
蓋 2 /kham3/ "to cover"
蓋 3 /kua3/ "a cover"

6 The Autosegmentation program and the Spell-checker

In order to speed up the building of the corpus, a word auto-segmentation program is necessary. Yet, when the program is segmenting words from the text, it can also deal with some related problems at the same time, such as the consistency of the transcription, adding romanization, and expanding the lexicon.

The Lexicon Bank

As the basis of the auto-segmentation program and the spell-cheker, a corpus-based lexicon has been constructed which includes the lemma (both in romanization and in Chinese characters), alternative forms, synonyms, and part-of-speech. (See the Appendix for a sample of the lexicon.)

Consistency in the transcription

Taiwanese speech recognition is still developing, so there is no way to transcribe the data with machine. Hence, transcription can only be done manually. The transcribers might be inconsistent in choosing the written form. For example, 按怎 (an3cuann2) "how" can be transcribed as 怎樣, 怎麼樣, 按怎, 怎麼, 什麼 and so on. Therefore, it is very important to design a program can identify the inconsistency.

When the program is segmenting the text, it tries to match a string which matches the

word in the column of "Chinese character" in the lexicon bank. It then segments the word and codes its pinyin. Figure 1 shows the input text in the frame, and Figure 2 shows the output of after segmentation. Word segmentation standard follows mostly that of the Sinica Corpus (Chen et al. 1996).

If the transcription happens to be one of the "other forms," it will be replaced with the standard form listed under the "Chinese character."

Adding new words to the Lexicon

If a word does not exist in the lexicon, it will be added to the lexicon after the file manager confirms its status.

In short, the word auto-segmentation program is able to do four things at the same time:

1 segment words in the text
2 code the pinyin for the characters
3 correct the inconsistent written forms
4 expand the lexicon bank

7 Applications of the corpus

This corpus has been used for studies on various aspects of child language acquisition, including tone acquisition (Tsay and Huang, 1998; Tsay, Myers, and Chen, 2000; Tsay, 2001), consonant acquisition (Liu and Tsay, 2000), classifier acquisition (Myers and Tsay, 2000), final particle acquisition (Hung, Li, and Tsay, 2004), verb acquisition (Lee and Tsay, 2001; Lin and Tsay, 2005), vocabulary acquisition (Tsay and Cheng, in progress). More studies are on the way.

Because this corpus is based on spontaneous speech, it also has its applications in addition to linguistic research. For example, this corpus can be used in extracting important speech features.

This corpus will be released by the Association for Computational Linguistics and Chinese Language Processing, Taiwan, in fall of 2005.

Acknowledgements

This project was supported by grants from the National Science Council, Taiwan. (Grant no. NSC89-2411-H-194-06, NSC90-2411-H-194-031, NSC91-2411-H-194-029). We thank all the children and their families. Research assistants at Chung Cheng University, especially Tingyu Rose Huang, Hui-Chuan Joyce Liu, and Xiao-Chun Kay Chen, have made remarkable contributions to this project.

References

Chinese Knowledge Information Processing Group (CKIP). 1993. Chinese Part-of-Speech Analysis. Technical Report 93-05. Taipei: Academia Sinica.

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu. 1996. SINICA CORPUS: Design Methodology for Balanced Corpora. *Language, Information, and Computation* (PACLIC), 11: 167-176.

Huang, Chu-Ren, Keh-jiann Chen, Feng-yi Chen, and Li-Li Chang. 1997. "Segmentation Standard for Chinese Natural Language Processing" *Computational Linguistics and Chinese Language Processing*, Vol. 2, no. 2, pp. 47-62.

Huang, Chu-Ren, Keh-Jiann Chen and -Shin Lin. 1997. "Corpus on Web: Introducing the First Tagged and balanced Chinese Corpus." *Conference Proceedings of Pacific Neighborhood Consortium 1997*.

Hung, Jia-Fei, Cherry Li, and Jane Tsay. 2004. "The Child's Utterance Final Particles in Taiwanese: A Case Study." *Proceedings of the 9th International Symposium of Chinese Languages and Linguistics*, 477-498. Taipei: National Taiwan University.

Lee, Thomas Hun-tak and Jane Tsay. 2001. "Argument structure in the early speech of Cantonese-speaking and Taiwanese-speaking children." The Joint Meeting of the 10th IACL and the 13th NACCL. June 22-24, 2001. UC Irvine.

Lin, Huei-ling and Jane Tsay. 2005. "Acquiring Causatives in Taiwanese" Paper presented at the 14th IACL. Leiden University.

Liu, Joyce H. C. and Jane Tsay. 2000. "An

Optimality-Theoretic Analysis of Taiwanese Consonant Acquisition." *Proceedings of The 7th International Symposium on Chinese Languages and Linguistics,* 107-126. Chung Cheng University, Taiwan.

MacWhinney, Brian. 1995. *The CHILDES Project: Tools for Analyzing Talk*. 2nd ed. Hillsdale, NJ.: Lawrence Erlbaum Associates Inc., Publishers.

MacWhinney, Brian, and Catherine Snow. 1985. *The Child Language Data Exchange System*. J*ournal of Child Language*, 12: 271-296.

Myers, James and Jane Tsay. 2000 "The Acquisition of the Default Classifier in Taiwanese." *Proceedings of the 7th International Symposium on Chinese Languages and Linguistics,* 87-106. Chung Cheng University, Taiwan.

Myers, James and Jane Tsay. 2002. "Grammar and Cognition in Sinitic Noun Classifier Systems." *Proceedings of the First Cognitive Linguistic Conference*, pp. 199-216. Taipei: Chengchi University

Tsay, Jane. 2001. "Phonetic Parameters of Tone Acquisition in Taiwanese" In Minehru Nakayama (ed.) *Issues in East Asian Language Acquisition*, 205-226. Tokyo: Kuroshio Publishers.

Tsay, Jane and Ting-Yu Huang. 1998. "Phonetic Parameters in the Acquisition of Entering Tones in Taiwanese." *The Proceedings of the Conference on Phonetics of the Languages in China.* 109-112. City University of Hong Kong.

Tsay, Jane, James Myers, and Xiao-Jun Chen. 2000. "Tone Sandhi as Evidence for Segmentation in Taiwanese." *Proceedings of the 30th Child Language Research Forum,* 211-218. Stanford, California: Center for the Study of Language and Information

Dictionaries

Chen, Xiu 1998. Taiwanhua Dacidian [Taiwanese Dictionary]. Taipei: Yuanliu Publishing Co.

Dong, Zhongsi. 2001. Taiwan Minnanyu Cidian [Taiwan Southern Min Dictionary]. Taipei: Wunan Publisher.

Li, Rong. 1998. Xiamen Fangyan Cidian [Xiamen Dialect Dictionary]. Jiangsu: Education Publisher.

Wu, Shouli. 2000. Guotaiyu Duizhao Huoyong Cidian [Mandarin-Taiwanese Comparative Dictionary]. Taipei: Yuanliu Publising Co.

Xu, Jidun. 1992. Changyong Hanzi Taiyu Cidian [Taiwanese Dictionary of Frequently Used Chinese Characters]. Taipei: Culture Department, Zili Evening News.

Yang, Qingchu. 1993. Guotai Shuangyu Cidian [Mandarin-Taiwanese Bilingual Dictionary] Kaohsiung: Duli Publishing Co.

Yang, Xiufang. 2001. Minnanyu Cihui [Southern Min Vocabulary]. Taipei: Ministry of Education.

Southern Min Spell Checker

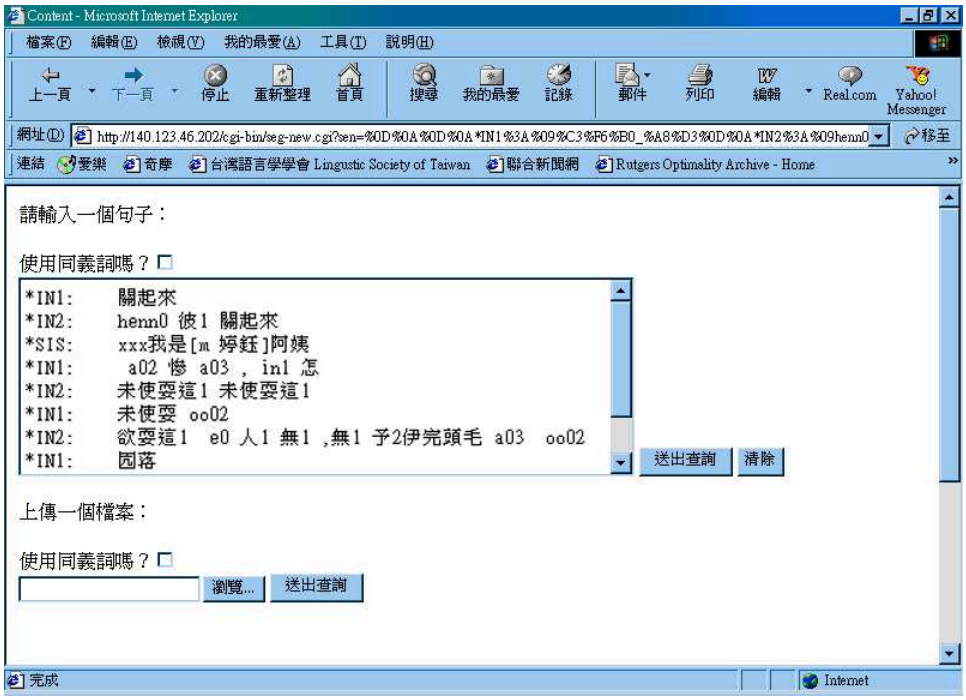

Figure 1 Input text

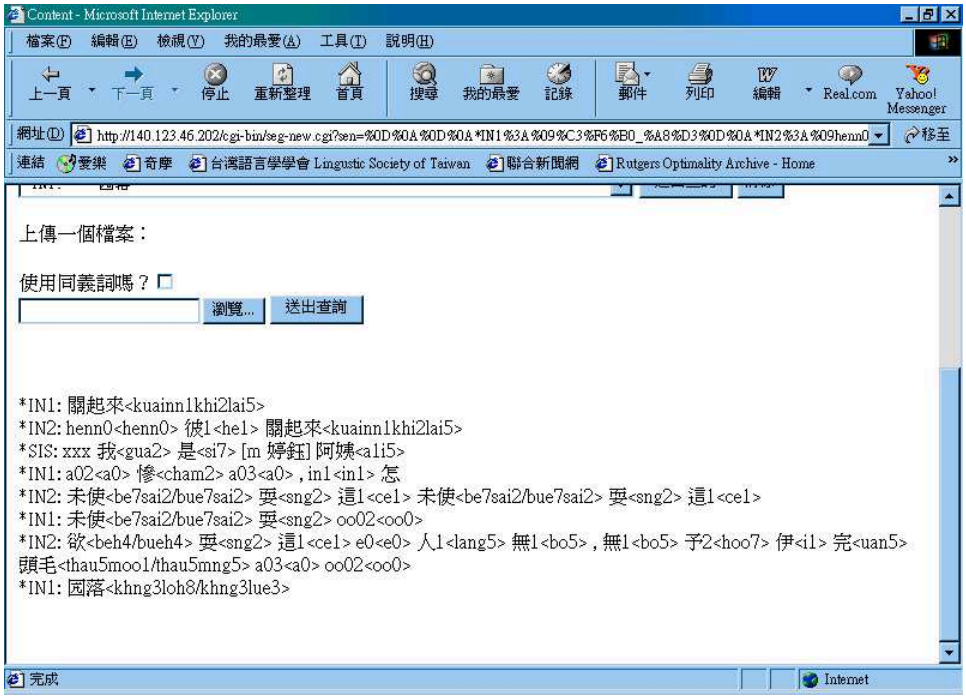

Figure 2 Output text

Appendix Sample of the Lexicon

Chinese character	Southern Min Pinyin	Part-of-speech	Meaning (or Mandarin synonyms)	Example
未記e0	be7ki3e0	VK		
未見笑	be7kian3siau3/ bue7kian3siau3	VH	不要臉	
賣了	be7liau2/ bue7liau2	VB	賣完	
賣了了	be7liau2liau2/ bue7liau2liau2	VB	賣光光	
賣了了去	be7liau2liau2khi3/ bue7liau2liau2khi3	VB	賣光光去	
未使	be7sai2/ bue7sai2	D	不行、不能、不可以、未saih6	你未使去（修飾動詞）
未使	be7sai2/ bue7sai2	VH	不行、不能、不可以、未saih6	伊知道這樣未使
未使得	be7sai2cit4/ bue7sai2cit4	D	不能	
未輸	be7su1/bue7su1	D	未su1、好像	好像
未當	be7tang3/ bue7tang3	D	不能、不可以、不行、未能、未tang3	
未當得	be7tang3cit4/ bue7tang3cit4	D	不能、不可以	
賣掉	be7tiau7/bue7tiau7	VC		
賣掉去	be7tiau7khi3/ bue7tiau7khi3	VB		
未振未動	be7tin2be7tang7	VA	一動也不動	
賣著	be7tioh8/ bue7tioh8	VC		
賣場	be7tiunn5/ bue7tiunn5	Nc		
未拄好	be7tu2ho2/ bue7tu2ho2	VH	配合得不好	
賣完	be7uan5/bue7uan5	VC		
欲	beh4/bueh4	D	會、愛、要、要1	(+動詞) 我要去市仔買菜
欲	beh4/bueh4	D	快要	快來不及了
欲	beh4/bueh4	VC	會、愛、要、要1	(+名詞組）我要這領衫
欲愛	beh4ai3/ bueh4ai3	D	欲愛、想要、欲要、要愛	欲愛 耍 [m 玩具]
欲愛	beh4ai3/ bueh4ai3	VC	欲愛、想要、欲要、要愛	
欲無	beh4bo5	Cbb	bue4無、bue2bo5	（連接詞）本來…但