Computational Linguistics and Chinese Language Processing
Vol. 12, No. 4, December 2007, pp. 411-442

Construction and Automatization of a Minnan Child Speech Corpus with some Research Findings

Jane S. Tsay*

Abstract

Taiwanese Child Language Corpus (TAICORP) is a corpus based on spontaneous conversations between young children and their adult caretakers in Minnan (Taiwan Southern Min) speaking families in Chiayi County, Taiwan. This corpus is special in several ways: (1) It is a Minnan corpus; (2) It is a speech-based corpus; (3) It is a corpus of a language that does not yet have a conventionalized orthography; (4) It is a collection of longitudinal child language data; (5) It is one of the largest child corpora in the world with about two million syllables in 497,426 lines (utterances) based on about 330 hours of recordings. Regarding the format, TAICORP adopted the Child Language Data Exchange System (CHILDES) [MacWhinney and Snow 1985; MacWhinney 1995] for transcribing and coding the recordings into machine-readable text. The goals of this paper are to introduce the construction of this speech-based corpus and at the same time to discuss some problems and challenges encountered. The development of an automatic word segmentation program with a spell-checker is also discussed. Finally, some findings in syllable distribution are reported.

Keywords: Minnan, Taiwan Southern Min, Taiwanese, Speech Corpus, Child Language, CHILDES, Automatic Word Segmentation

1. Introduction

Taiwanese Child Language Corpus is a corpus based on spontaneous conversations between young children and their adult caretakers in Minnan speaking families in Chiayi County, Taiwan. This corpus is special in several ways. First, it is a Minnan corpus. Minnan is Southern Min Chinese spoken in Taiwan (also known as Taiwanese in linguistic literature). It is less studied, especially when compared with Mandarin Chinese. Second, it is a speech-based corpus. The scripts in the corpus were transcribed from recordings of

* Institute of Linguistics, National Chung Cheng University, 168 University Road, Min-hsiung, Chiayi County, Taiwan 62102, ROC Phone: 886-5-2720411 ext. 31502 Fax: 886-5-2721654
E-mail: Lngtsay@ccu.edu.tw

[Received November 1, 2006; Revised August 14, 2007; Accepted August 16, 2007]

spontaneous speech. Third, it is a corpus of a language that does not yet have a conventionalized orthography. Fourth, it is a child corpus. It's a collection of longitudinal child language data. Fifth, it is currently one of the largest child corpora in the world. It contains about 2 million syllables/characters in 497,426 lines (utterances) based on about 330 hours of recordings. Finally, it is a corpus that uses an international platform. This platform is the Child Language Data Exchange System (CHILDES) [MacWhinney and Snow 1985; MacWhinney 1995] for transcribing and coding the recordings into machine-readable text.

The goals of this paper are:

(1) to introduce the construction of this speech-based child language corpus, TAICORP (Section 2);

(2) to introduce the automatization process of this corpus and discuss some issues encountered during the implementation of the system (Section 3);

(3) to present some research findings based on this corpus (Section 4).

2. Taiwanese Child Language Corpus

Taiwanese Child Language Corpus (TAICORP) contains scripts transcribed from about 330 hours of spontaneous speech from fourteen young children acquiring Taiwan Minnan as their first language. A brief introduction to this corpus was reported at the 5th Workshop on Asia Language Resources [Tsay 2005a]. In this extended paper, in addition to a more detailed description and more discussion about the corpus and related issues, findings in syllable type distribution and tone type distribution are also presented.

There are about 1.6 million words (over 2 million syllables/characters) in this corpus, as shown in Table 1.

Table 1. The size of TAICORP

	Lines (utterances)	Words	Syllables	
			Syllables (in words) 1,558,408	Syllables (in particles) 538,992
Total	497,426	1,646,503	2,097,400	

Since some words do not have corresponding Chinese characters and are presented in romanization notation (Minnan Pinyin) in this corpus, the syllable might be a more precise unit than the more traditional unit *zi* 字 (Chinese character).

Note that we divide the syllables into two categories: syllables in words (*e.g.*, chia 車) and syllables in particles (*e.g.*, la 啦). Among all the 2,097,400 syllables, 538,992 syllables (about 26%) are in particles. This is a very interesting fact and will be discussed in more detail in Section 4.

In this section, TAICORP is introduced in the following aspects:

2.1. Motivation

2.2. Data collection

2.3. Text files in CHILDES format

2.4. Transcribing sound files into text files

2.5. Annotations

2.1 Motivation

From the linguistics point of view, there is an urgent need to construct a Minnan child language corpus, partly because there has not been any such corpus available and partly because it may be getting more and more difficult to find young children learning Minnan as their first language, especially in the cities. On top of that, the significance of a large collection of longitudinal child language data for linguistic studies goes beyond saying.

Mandarin and Minnan are the two major Chinese languages in Taiwan. For over forty years, Mandarin was the only official language for instruction at school in spite of the fact that about 73% of the population belonged to the Minnan ethnic group [Huang 1993]. Young children in kindergartens and elementary schools were not allowed to speak Minnan even if Minnan was the language spoken at home. This policy caused a decrease in the number of young children learning Minnan as their first language.

Although the situation has changed in recent years and other local languages besides Mandarin, including Minnan, Hakka, and the aboriginal (Formosan) languages have been included in the curriculum of elementary schools, there is still a serious concern about the decrease of native Minnan speakers. This concern can be supported by a more recent survey. Tsay [2005] reports that in a survey of all 8th graders in Chiayi City in Southern Taiwan, an area where the population should be overwhelmingly Minnan, only about 26% of 14 year-olds used Minnan in their daily life, although over 80% of their grandparents and over 70% of their parents were native Minnan speakers.

Under this consideration, Minnan was chosen as the target language. The project was conducted in a rural area in Chiayi County in Southern Taiwan with the hope to find young children who were raised in a Minnan-speaking environment.

2.2 Data collection

Data collection took place over a period of around three years between August 1997 and July 2000 under the support of the National Science Council in Taiwan (NSC 87-2411-H-194-019, NSC 88-2411-H-194-019, NSC 88-2418-H-194-002).

Child participants

Young Children from Minnan-speaking families were recruited in Min-hsiung Village, Chiayi County, in Southern Taiwan. Nine boys and five girls from the following villages in Min-hsiung Xiang participated in this project: Fengshou（豐收村）, Sanxing（三興村）, Dongxing（東興村）, Xidibu（溪底部）, and Zhenbei（鎮北村）. They aged from one year and two months (1;2) to three years and eleven months (3;11) old at the beginning of the recording. More than half of the children were recorded over more than two years. The age range at the offset of the recordings is between 2;7 and 5;3.

Recording

Regular home visits were conducted every two weeks for younger children and every three weeks for children older than three years old. The recording setup was children at play at home interacting naturally with the adult(s), usually one of their caretakers (parents, grandparents, or, in very few cases, the nanny) and/or the investigator. The activities were children's daily life at home: playing with toys or games, reading picture books, or just talking without any specific topics. Since we hoped to have the most natural environment, Mini-disc recorders and microphones were used so that it was easier for the recorder (the investigator) to follow the child wherever she/he went. Usually, each recording session lasted from 40 to 60 minutes.

Information about the child participants and the recordings is given below.

Table 2. Recording Information of TAICORP

Name	Sex	Age range	Sessions	length (min.)
YDA	M	3;11.02 – 4;04.26	9	540
YCX	M	3;10.16 – 4;00.16	6	285
LJX	M	3;09.20 – 4;02.24	8	530
CQM	M	2;09.07 – 4;06.22	30	1584
LMC	F	2;08.07 – 5;03.21	50	2045
YJK	M	2;06.11 – 2;0626	2	105
CEY	F	2;01.27 – 3;10.00	37	1728
HBL	M	2;01.22 – 4;00.03	45	1889
LWJ	F	2;01.08 – 3;07.03	36	1777
WZX	M	2;01.17 – 4;03.15	44	1757
YSW	M	1;07.17 – 2;07.14	21	1210
TWX	F	1;05.12 – 3;06.15	44	1829
HYS	M	1;02.28 – 3;04.12	51	2280
LYC	F	1;02.13 – 3;03.29	48	2255
Total	M=9 F=5		431	about 330 hours

Sound file editing

There were a total of 431 recording sessions. Each session was saved as a separate sound file. The sound files were first edited so that the empty or noisy parts could be cleared. In order to have easier searching and locating the content of the recordings, each sound file was segmented into several tracks and the tracked marks were tagged.

2.3 Text Files in CHILDES Format

The sound files were transcribed into text files in CHILDES format. CHILDES (Child Language Data Exchange System) was originally set up by Elizabeth Bates, Brian MacWhinney, and Catherine Snow to transcribe and code recordings into machine-readable speech text [MacWhinney and Snow 1985; MacWhinney 1995].

CHILDES has been widely accepted as the standard system for child language data. TAICORP adopted the format of CHILDES so that it will be easy to exchange and share data with researchers around the world. CHILDES includes a transcription system, CHAT, and a set of programs, CLAN, for various analyses. In this section, we introduce a simplified version of the format of text files in CHAT. For details, please refer to MacWhinney [1995] or the official website of CHILDES at http://childes.psy.cmu.edu/.

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

Headers

There are three kinds of headers: obligatory headers, constant headers, and changeable headers.

> ***Obligatory headers***: Obligatory headers are necessary for every file. They mark the beginning, the end, and the participants of the file.
>
> ***Constant headers***: They mark the name of the file and the background information of the children.
>
> ***Changeable headers***: They contain information that may change across files, such as the recording date, duration, coders, and so on.

These headers all begin with @. Some examples are given below:

Obligatory headers:

@Begin	to mark the beginning of a file
@End	to mark the end of a file
@Participants	to list all the participants in a file

Constant headers:

@Age of XXX:	the age of speaker
@Birth of XXX:	the birthday of the speaker
@Coder:	the file coder's name
@Educ of XXX:	the highest education of the speaker
@Filename:	filename
@Language:	the main language used in the file
@Language of XXX:	the language used by the speaker
@Sex of XXX:	the sex of the speaker
@Warning:	the defects of the file

Changeable headers (optional):

@Activities:	Activities involved in the situation
@Bck:	background information of the utterance
@Comment:	the comment of the investigator
@Date:	the date of the interaction
@G:	gems
@Location:	the location of the interaction
@New Episode:	the new episode of the recording starts
@Room Layout:	room configuration and positioning of furniture
@Situation:	the situation of the interaction
@Tape Location:	the specific ID, side and footage
@Time Duration:	the length of recording time
@Time Start:	the starting time of recording

Tiers

The content of a file is presented in tiers in CHILDES. There is a main tier and several dependent tiers for each line (utterance).

The main tier, marked with *, is the speech of the speaker. Three capital letters indicate the status of the speaker, *e.g.*, *CHI is the child, *MOT the mother, and *INV the investigator.

Minnan Pinyin is used in the Main tier. Words are separated by a space. Therefore, an utterance "I want to water the vegetables" from a child would be:

*CHI:	gua2	beh4	ak4	chai3.
	I	want	water	vegetable

The additional information is given in dependent tiers that are marked with % at the beginning of a new line. The seven dependent tiers used in TAICORP are given below.

%ort:	the utterance in logographic orthography (*i.e.*, Chinese characters)
%pro:	the actual target pronunciation of the utterance (dialectal variation)
%syl:	syllable type coded with C and V (*e.g.* CVV for /gua/)
%cod:	part-of-speech coding
%pho:	phonetic transcription in Unicode IPA (for child speech only)
%syc:	syllable type of the child's pronunciation
%ton:	tone value in 5-digit scale

For the adult speech, there are only four dependent tiers: %ort, %pro, %syl, and %cod because no phonetic transcription was done on the adult speech. For the child speech, there are up to seven dependent tiers as shown in the following example.

(main tier)	***CHI:**	**gua2**	**beh4^bueh4**	**ak4**	**chai3.**
(depnt tier)	%ort:	我	欲	沃	菜.
(depnt tier)	%pro:	gua2	beh4	ak4	chai3.
(depnt tier)	%syl :	CVV	CVK	VK	CVV
(depnt tier)	%cod:	Nh	D	VA	
(depnt tier)	%pho:	guaŋ	be	ak	t'ai
(depnt tier)	%syc:	CVVN	CV	VK	CVV
(depnt tier)	%ton:	55	55	5	21

2.4 From Sound Files to Text Files

All sound files were transcribed into text files. Transcriptions included (1) orthographic transcription; and (2) phonetic transcription (in IPA, International Phonetic Alphabet).

There were two kinds of systems used in orthographic transcription. One was the logographic orthography (*i.e.*, traditional Chinese writing system Hanzi 漢字), and the other was a spelling-based romanization system for Minnan (called Minnan Pinyin). Thus, each sound file was transcribed into a separate text file in both Chinese characters and Minnan Pinyin.

2.4.1 Orthographic Transcription in Chinese Characters

The reason that the sound files were first transcribed into Chinese characters was because this written form is closest to most native speakers' intuition. Therefore, by transcribing [tsetɕʰia] into "坐車", it makes it much easier for the user to read.

Although romanization notation (Minnan Pinyin) in the Main tier (*e.g.*, *CHI tier in the above example) makes it easier to run the analyzing programs in CHILDES and might also be easier for non-Chinese users of the corpus, having a tier with Chinese characters would be more convenient for those who know Chinese. Therefore, a dependent tier %ort was added to present the utterances in Chinese characters. This is a reasonable method because most Minnan words are cognates of Mandarin words. Still, there are quite a few words that either do not have their corresponding Chinese characters or their corresponding Chinese characters are so obsolete that they cannot be found in the software for typing Chinese characters.

Since Minnan does not have as conventionalized orthography as Mandarin, quite a few words in Minnan do not have a consistent way of writing them. In order to help build consensus in Minnan cognates（閩南語本字）, Minnan dictionaries were consulted. At least seven dictionaries were used as listed after the references.

There are several possibilities regarding Chinese characters used in Minnan:

First, they are exactly the same as those used in Mandarin, for example, 色筆/sik4pit4/ "color pens".

Second, they are synonyms of Mandarin words, but use different characters, for example, 挽 /ban2/ "pluck; pick up" is a synonym of Mandarin 摘 /zhai1/ or 採 /cai3/; 鼻芳 /phinn7phang1/ "smelling the fragrance" is a synonym of 聞香 /wen2xiang2/.

Third, although the Chinese characters in Minnan can be found in the dictionary, they might be so obsolete that one has to use special software to make the character forms, as in the first character of the following word meaning "good morning".


早 /gau5ca2/

This is very inconvenient for users and is very hard to process, too. In such cases, Minnan Pinyin is used and the above word would be presented as *gau5 早*.

Fourth, when Chinese characters cannot be found at all for Minnan words, Minnan Pinyin is used, as in the first morpheme is the word *chua7 路* /chua7loo7/ "leading the way" or *chit4tho5* /chit4tho5/ "playing around".

For homonyms that share the same Chinese character, a number is added to 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/lid"

2.4.2 Orthographic Transcription in Minnan Pinyin

The reason for transcribing the sound files into Minnan Pinyin was twofold: (1) to encode the sounds in a spelling system, and (2) to make it easier for the machine (computer program) to read and to do analyses such as syllable frequency counts.

The Minnan Pinyin system used in TAICORP is the Taiwan Southern Min Phonetic Alphabetic officially announced by the Ministry of Education in Taiwan in 1998.[1] Like most romanization systems, the Minnan Pinyin system labels sounds at the phonemic level.

The Minnan Pinyin notation system with examples is given in Table 3 (consonants) and Table 4 (vowels) below. Note that '-' before a symbol indicates the coda position, as in a checked (Rusheng) syllable. It is necessary to make such a distinction because of the asymmetry in the distribution of consonants. For example, [b] cannot occur in the coda position, although it can occur in the onset position. Following the IPA convention, a dot under a symbol is used to denote a syllabic consonant. Nasal vowels are denoted with "nn". Therefore, the word [tī] "sweet" is transcribed as /tinn/ in this system.

[1] The system released by the Ministry of education adopted the Taiwan Language Phonetic Alphabet (TLPA) originally proposed by Taiwan Languages Society in 1994. Revisions can easily be made if it becomes necessary.

Table 3. Minnan Pinyin System (Consonants)

Minnan Pinyin	IPA	Example	Glossary
p	p -p	pit4 筆 ciap4 汁	pen juice
ph	p^h	phue5 皮	skin
b	b	be2 馬	horse
m	m -m m̩	moo1 毛 sim1 心 a1m2 阿姆	fur heart aunt
t	t -t	to1 刀 that4 踢	knife kick
th	t^h	thau5 頭	head
l	l	lai5 來	come
n	n -n	ni5 年 sin1 新	year new
k	k -k	kau2 狗 kak4 角	dog horn
kh	k^h	kha1 跤	foot
g	g	gu5 牛	cow
ng	ŋ -ŋ ŋ̍	nge7 硬 sing1 升 ng5 黃	hard ascend yellow
h	h -ʔ	hue1 花 bah4 肉	flower meat
c	ts tɕ	cu2 煮 cit8 一	cook one
ch	ts $tɕ^h$	chai3 菜 chit4 七	vegetable seven
s	s ɕ	sai1 獅 si3 四	lion four
j	z	jit8 日	sun

Table 4. Minnan Pinyin System (Vowels)

Minnan Pinyin	IPA	Example	Glossary
i	i	ti1 豬	pig
e	e	be2 馬	horse
a	a	ka7 咬	bite
oo	ɔ	koo1 姑	aunt
o	o/ə	to1 刀	knife
u	u	gu5 牛	cow
inn	ĩ	tinn1 甜	sweet
enn	ẽ	chenn1 星	star
ann	ã	sann1 三	three
onn	ɔ̃	honn3ki5 好奇	curious
ia	ia	khia7 倚	stand
io	io/iə	kio5 橋	bridge
iu	iu	kiu5 球	ball
iann	ĩã	kiann5 行	walk
iunn	ĩũ	kiunn1 薑	ginger
ai	ai	lai5 來	come
au	au	chau2 草	grass
ainn	ãĩ	phainn2 歹	bad
ui	ui	cui2 水	water
ue	ue	hue2 火	fire
ua	ua	kua1 歌	song
uann	ũã	suann3 線	string
iau	iau	iau1 枵	hungry
uai	uai	kuai1 乖	submissive
uainn	ũãĩ	kuainn1 關	close
uinn	ũĩ	khuinn3uah8 快活	joyful

There are seven lexical tones in Minnan spoken in Chiayi, Taiwan. These tone categories are denoted by digits 1 to 8, except for Tone 6 Yangshang (陽上) which has been merged into other tone categories due to historical sound change. Morphemes (or syllables) without underlying tones are marked with '0'. Interjections and particles, which do not have an underlying tone and their surface tones might vary due to different contexts, are all marked with '0', for example, a0 啊, le0 咧. Loan words, for example, too0sang0 多桑, borrowed from the Japanese word for "father", are also marked with the '0' tone category. Tones deviating from the seven lexical tones are categorized into the '9' tone category, for example, tones derived by syllable concatenation, bo5iau3kin2 → bua9kin2 不要緊 "not matter".

Table 5. Minnan Tones

Tone Category		Example	Glossary
0	toneless	oo0 哦	(interjection)
1	Yinping 陰平	si1 詩	poem
2	Yinshang 陰上	si2 死	death
3	Yinqu 陰去	si3 四	four
4	Yinru 陰入	sik4 色	color
5	Yangping 陽平	si5 時	time
7	Yangqu 陽去	si7 寺	temple
8	Yangru 陽入	sik8 熟	ripe
9	others	bua9kin2 不要緊	not-matter

2.4.3 Phonetic Transcription in IPA

As mentioned above, Minnan Pinyin is a notation system at the phonemic level. The adult speech is considered the target as well as the input of the child language. We assume that the adult speech is "standard", and no phonetic transcription was done for the adult speech due to the limitation of manpower. In general, it is appropriate to represent the adult speech phonemically, unless one wants to know the allophonic variation or idiosyncratic characteristics of the adult speakers. In those cases, detailed phonetic transcription would be required.

However, we are most concerned with the child speech. The most important aspect in child language is its deviation from the ambient adult speech. Therefore, narrow phonetic transcription has to be available to understand the pattern and development of child language.

Narrow phonetic transcription was conducted for sound files of children under two and a half years old using Unicode IPA. The following are two sample utterances from the child

WZX at 2;1.17. The child's segmental pronunciation is shown in the %pho (phonetic) tier, and tonal pronunciation is shown in %ton (tone) tier using a 5-point scale. Note that the child's pronunciation was different from the standard speech of the adult. For example, /gua/ "I" was pronounced as [ua], /cing/ [tsiŋ] "plant" was pronounced as [t'iŋ], etc. This is to record truthfully what the child actually said. Such data are very important for studying children's phonological development.

Example 1

*CHI:	gua2	gua2	koh4	peh4
	我…	我	擱	擘
	I	I	again	split
	"I want to split it again."			
%pho:	ua	ua	kəʔ	pe
%ton:	55	55	4	32

Example 2

*CHI:	he1	a1po5	cing3	e0
	彼	阿婆	種	e
	that	grandma	plant	(Relative clause marker)
	"That was planted by grandma."			
%pho:	he	aʔ pə	t'iŋ	ẽ
%ton:	44	3 55	21	21

2.5 Annotations

Two kinds of annotations are described in this section: part of speech (POS) annotations and discourse annotations.

2.5.1 Part of Speech Annotation

Minnan and Mandarin are both Sinitic languages and are very similar in their morphology and syntactic structures. Therefore, the POS coding system of a Minnan corpus should be very similar to that of the Sinica Corpus of Mandarin (see various technical reports by the Chinese Knowledge Information Processing Group (CKIP) [CKIP 1993, 1998; Chen *et al*. 1996]. There are a total of 46 codes listed as simplified codes and 115 corresponding codes for Mandarin in Sinica Corpus [CKIP 1998].

Table 6. Tagset in Sinica Corpus [CKIP 1998]

Simplified codes (total 46 codes)	Corresponding CKIP codes (total 115 codes)
A	A
Caa	Caa
Cab	Cab
Cba	Cbab
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb
Da	Daa
Dfa	Dfa
Dfb	Dfb
Di	Di
Dk	Dk
D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj
Na	Naa, Nab, Nac, Nad, Naea, Naeb
Nb	Nba, Nbc
Nc	Nca, Ncb, Ncc, Nce
Ncd	Ncda, Ncdb
Nd	Ndaa, Ndab, Ndc, Ndd
Neu	Neu
Nes	Nes
Nep	Nep
Neqa	Neqa
Neqb	Neqb
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi
Ng	Ng
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc
I	I
P	P*
T	Ta, Tb, Tc, Td
VA	VA11, 12, 13, VA3, VA4
VAC	VA2
VB	VB11, 12, VB2
VC	VC2, VC31, 32,33
VCL	VC1
VD	VD1, VD2
VE	VE11, VE12, VE2
VF	VF1, VF2
VG	VG1, VG2
VH	VH11, 12, 13, 14, 15, 17, VH21
VHC	VH16, VH22
VI	VI1, 2, 3
VJ	VJ1, 2, 3
VK	VK1, 2
VL	VL1, 2, 3, 4
V_2	V_2
DE	/的, 之, 得, 地/
SHI	/是/
FW	/外文標記/

To avoid arbitrary classification of words into the morpho-syntactic categories, we adopted the simplified version with 46 morph-syntactic codes, instead of the finer 115 categories used in the Sinica Corpus. In other words, categorization in TAICORP is broader. These codes (tagset) are listed in the table below.

Table 7. Tagset of TAICORP

Tagging	POS	POS (Chinese terms)
A	non-predicative adjective	非謂形容詞
Caa	coordinate conjunction	對等連接詞
Cab	listing conjunction	連接詞
Cba	conjunction occurring at the end of a sentence	連接詞
Cbb	following a subject	關聯連接詞
Da	possibly preceding a noun	數量副詞
Dfa	preceding VH through VL	動詞前程度副詞
Dfb	following adverb	動詞後程度副詞
Di	post-verbal	時態標記
Dk	sentence initial	句副詞
D	adverbial	副詞
Na	common noun	普通名詞
Nb	proper noun	專有名稱
Nc	location noun	地方詞
Ncd	localizer	位置詞
Nd	time noun	時間詞
Neu	numeral determiner	數詞定詞
Nes	specific determiner	特指定詞
Nep	anaphoric determiner	指代定詞
Neqa	classifier determiner	數量定詞
Neqb	postposed classifier determiner	後置數量定詞
Nf	classifier	量詞
Ng	postposition	後置詞
Nh	pronoun	代名詞
I	interjection	感嘆詞
P	preposition	介詞

Tagging	POS	POS (Chinese terms)
T	particle	語助詞
VA	active intransitive verb	動作不及物動詞
VAC		動作使動動詞
VB	active pseudo-transitive verb	動作類及物動詞
VC	active transitive verb	動作及物動詞
VCL	transitive verb taking a locative argument	動作接地方賓語動詞
VD	ditransitive verb	雙賓動詞
VE	active transitive verb with sentential object	動作句賓動詞
VF	active transitive verb with VP object	動作謂賓動詞
VG	classificatory verb	分類動詞
VH	stative intransitive verb	狀態不及物動詞
VHC	stative causative verb	狀態使動動詞
VI	stative pseudo-transitive verb	狀態類及物動詞
VJ	stative transitive verb	狀態及物動詞
VK	stative transitive verb with sentential object	狀態句賓動詞
VL	stative transitive verb with VP object	狀態謂賓動詞
V_2		有
DE	*special tag for the word "的"	的
SHI	special tag for the word "是"	是
FW	foreign words	外文標記
*Di/T	*marker following pseudo-transitive active verb	*le01
*CIT	*special tag for the word "得 2"	*得 2
*Comp	*complementizer	*補語連詞

2.5.2 Discourse Annotations

The texts in TAICORP are based on spontaneous conversations. Therefore, it is necessary to have discourse annotations. As a speech-based corpus, it is full of incomplete, repeated, repaired, and interrupted utterances. We tried to code these in the scripts. Since discourse analysis is not the primary focus of this paper, we only list some the discourse codes that were used in TAICORP.

(1) Codes for unidentifiable material

(a) xxx/xx: unintelligible speech (utterance/word).

(b) yyy/yy: unintelligible speech at the phonetic level.

(c) www/ww: untranscribed speech to be used in conjunction with a note to explain the situation

(2) Repetition

[/]: repetition of either one or more words

(3) Basic utterance terminators

The basic utterance terminators are the period, the question mark, and the exclamation mark. Each utterance must end with one of these three utterance terminators.

(4) Special utterance terminators: these terminators all begin with the + symbol and end with one of the three basic utterance terminators. For example,

(a) +… Incomplete but not interrupted utterance

(b) +/. Incomplete utterance due to interruption

(c) +//. Self-interruption: breaking off an utterance and starting up another by the same speaker

(d) +?. Interruption of a question: the utterance being interrupted is a question

(e) +, Self-completion: to mark the completion of an utterance after an interruption

(5) Scoped symbols

(a) [=! text] Paralinguistic material: marking paralinguistic events or actions, such as coughing, laughing, telling, crying, singing, and whispering.

(b) [>] Overlap follows

(c) [<] Overlap precedes

(d) [/] Retracing without correction

(e) [//] Retracting with correction

The following is a sample of discourse coding in TAICORP

@Begin

@Participants: CHI Lin Target_Child, INV Rose Investigator

@Age of CHI: 2;9.22

@Birth of CHI: 28-AUG-1995

@Coder: Rose, Kay, Joyce
@Filename: HBL17ipa
@Language: Taiwanese
@Sex of CHI: Male
@Date: 19-JUN-1998
@Tape Location: Lin D4-30-41
@Comment: Time Duration: 35 minutes
@Location: Chiayi, Taiwan
@Transcriber: Rose
@Comment: Track number is D4-30

*INV: a1lin5 [/] a1lin5, li2 koh4 kong2 cit8 kai2.
%ort: 阿林 [/] 阿林, 你 擱 講 一 1 改.
%cod: Nb Nb Nh D VE Neu Nf

*INV: <li2 kong2> [//] li2 thau5tu2a2 kong2 a1ma2 khi3 toh4?
%ort: <你 講> [//] 你 頭拄仔 講 阿媽 去 陀?
%cod: Nh VE Nh Nd VE Na VCL Ncd

*CHI: khi3 sio1kim1 la0.
%ort: 去 燒金 la0.
%cod: VCL VA T
%pho: i t,j i o t,j i,n ng ng a,n
%ton: 55 33 55 21

*INV: hann0/hannh0?
%ort: hann0?
%cod: I

*CHI: khi3 sio1kim1.
%ort: 去 燒金.
%cod: VCL VA
%pho: kh i t,c\ i o t,c\ i ng

%ton: 55 33 55
*INV: khi3 sio1hiunn1 oo0?
%ort: 去 燒香 oo02?
%cod: VCL VA

*CHI: li2 bo5 +/.
%ort: 你 無 1 +/.
%cod: Nh D
%pho: i b o
%ton: 55 33

*INV: a0 i1 u7 cah4 <sann2mih8hue3 khi3> [>]?
%ort: a01 伊 有 紮 <啥物貨 去> [>]?
%cod: Dk Nh V_2 VC Nep VCL

3. Automatization

Constructing a speech-based corpus requires a lot more steps than constructing a corpus based on written texts. The most labor-intensive and time-consuming work is devoted to transcribing the sound files into text files. In the first stage of the construction of TAICORP, every step was done manually. These steps are shown in Figure 1 below.

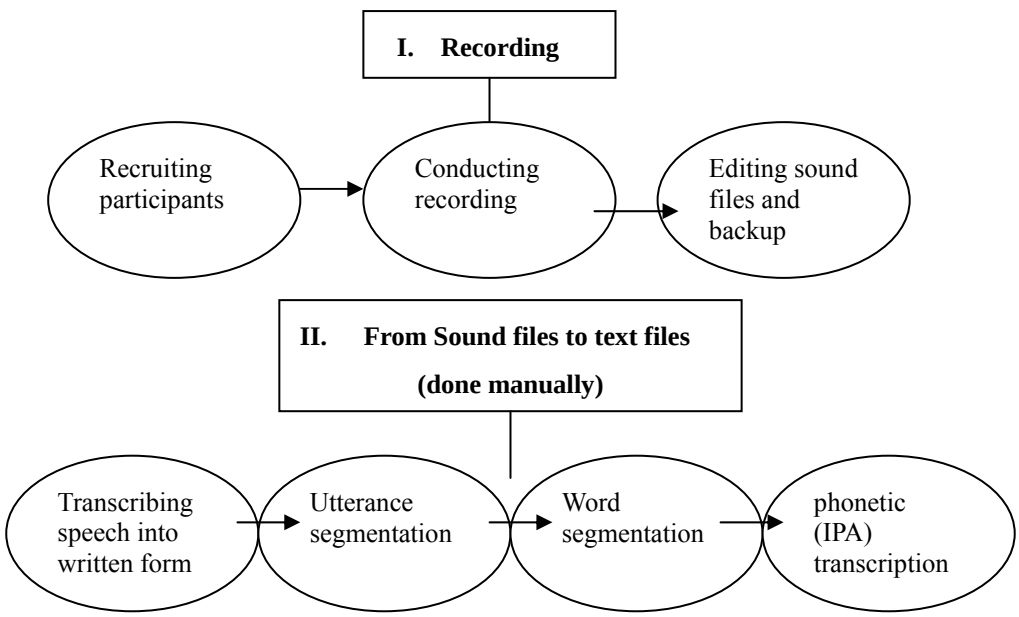

Figure 1. Steps in manual construction

3.1 Automatic Word Segmentation

After all the hard work, it was hoped that the corpus could contribute to the automatization of the procedure. Under this consideration, an automatic word segmentation program has been developed.[2] As the basis of the automatic word segmentation program, a corpus-based lexicon has been constructed manually, which includes the lexical item (both in Minnan Pinyin and in Chinese characters), alternative forms, synonyms, and part-of-speech labels.

Thus, the lexical bank contains the following information for each lexical item:

Logographic orthography: the word in Chinese characters

Spelling-based orthography: the word in Minnan Pinyin

Part-of-speech: the POS coding of the word

Alternative forms/synonyms: alternative written forms of the word

Since the orthography convention has not reached consensus in the Minnan-speaking community, the transcribers might not be consistent in their uses of the written form. Their non-standard uses of the written form are also listed as "alternative forms" so that they can be used in searching for such mistakes by the transcribers and thus can be corrected.

A sample of the lexical bank is given in Table 8 below.

Table 8. A sample of the lexical bank

Chinese characters	Minnan Pinyin	POS	Meaning/synonym (or alternative forms)
未記	be7ki3	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

[2] Programmers who helped out with the development of this program at different stages were Ming-Chung Chang and Charles Jie.

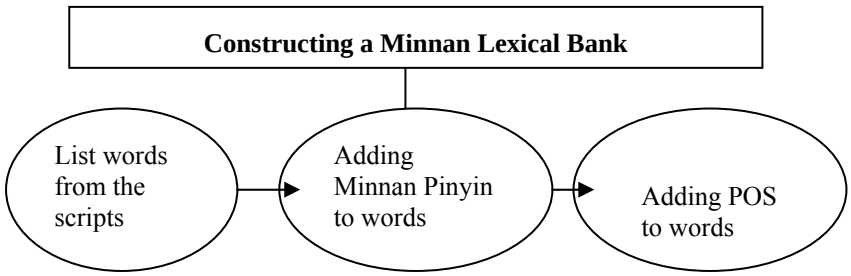

Figure 2. Constructing a Minnan lexical bank

After the lexical bank was established, an automatic word segmentation program was developed. This program also converts the word into Minnan Pinyin after segmentation. The way the program works is to identify a string of sounds that match the word in the column "Chinese characters" in the lexical bank. It then segments the word from the text and codes it in Minnan Pinyin. The word segmentation standard mostly follows that of the Sinica Corpus [Huang *et al*. 1997].

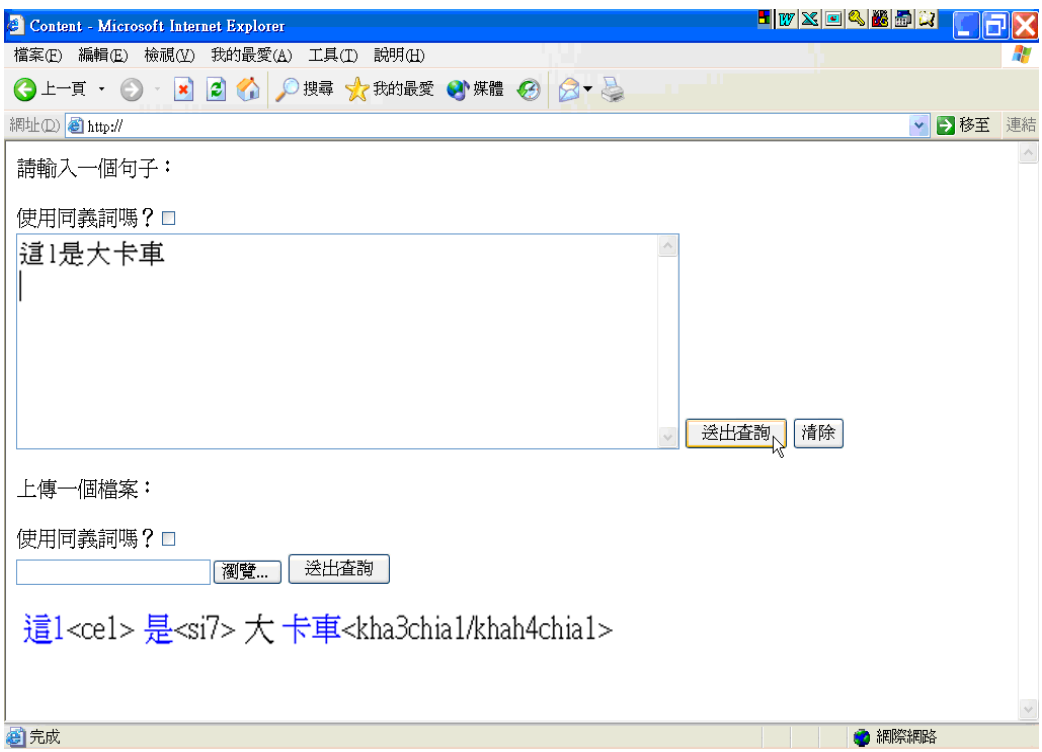

Figure 3. Automatic Word Segmentation

After word segmentation and Minnan Pinyin conversion at the %ort tier, POS codes are tagged to the word at the %cod tier.

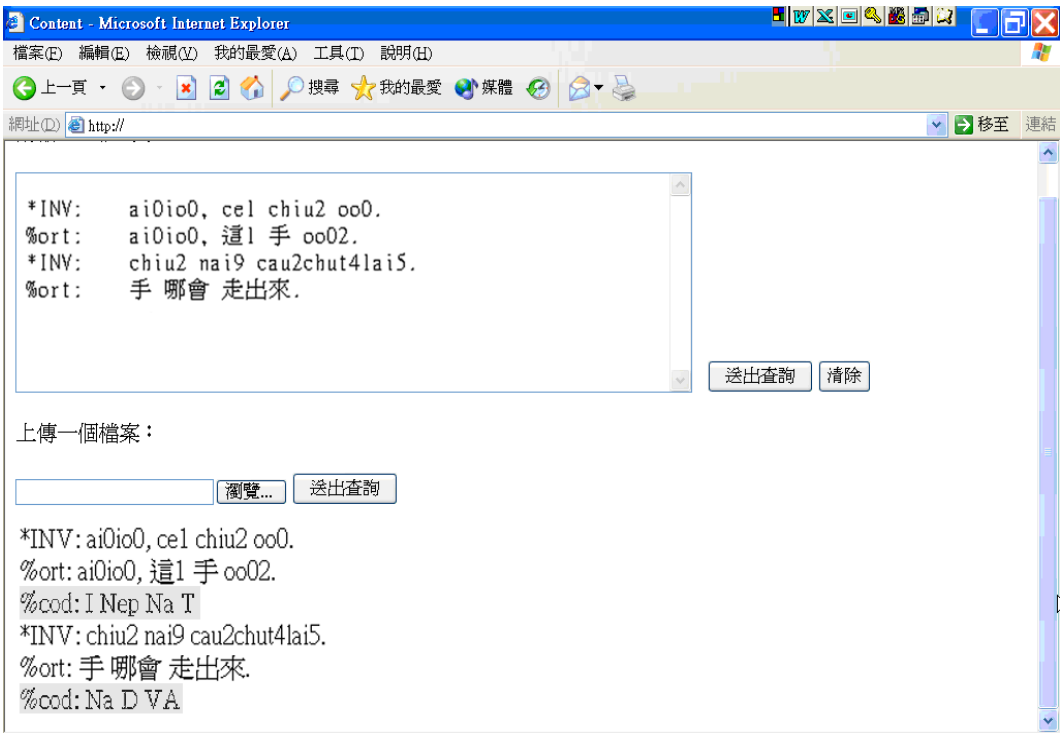

Figure 4. POS tagging after autosegmentation

During the process of constructing this corpus, we found some issues that only occur in speech-based corpora and not in corpora that are based on written texts. The first was the issue of multiple pronunciations of the same word due to dialectal variation. For example, the word for "want to" 欲 is pronounced /beh/ for some people, but is pronounced /bueh/ for others. Since they belong to the same word (same lexical entry), they have to be listed under the same lexical entry in the lexicon as /beh^bueh/. This does not happen in corpora based on written texts.

This phenomenon is especially common in Minnan because Taiwan Minnan speakers originally came from different areas in Fujian Province, China, including Zhangzhou 漳州, Quanzhou 泉州, and Xiamen 廈門. Therefore, dialect variations are very common. Words with multiple pronunciations (mostly dialectal variations) are all listed but connected by ^.

This is not a problem when the speech is transcribed into Chinese characters because there is only one orthographic form for each word. This problem is also not too serious when the speech is transcribed in a romanization notation, like Minnan Pinyin, manually by researchers. Transcribing speech into Minnan Pinyin is slow and an automatic converter is preferred. However, when an automatic word segmentation program uses the lexicon for word

segmentation, it will automatically retrieve a multiple pronunciation form like /beh^bueh/.

Take the following utterance as an example.

*CHI: gua2 beh4^bueh4 ak4chai3.
%ort: 我 欲 沃菜.

The word meaning "want" 欲 has two pronunciations *beh4* and *bueh4* and they show up as *beh4^bueh4* as in the main tier.

When counting words, they are counted as one word. That is, they are the same word in the lexicon and do not cause trouble in word frequency counts. However, when counting syllable token frequencies, they will be double counted. Besides, these two pronunciations have different syllable types, CVC and CVVC, respectively. Moreover, it is necessary to know the real target pronunciation of the specific speaker. Therefore, we need to have another tier %pro to show the actual pronunciation of the specific speaker based on the recording. Unfortunately, this can only be done manually.

*CHI: gua2 beh4^bueh4 ak4chai3.
%ort: 我 欲 沃菜.
%pro: gua2 beh4 ak4chai3.

3.2 The Inconsistency Issue and the Spell-Checker

Minnan speech recognition systems are still being developed. Hence, transcription can only be done manually. As mentioned above, Minnan does not have a conventionalized orthography, so transcribers might be inconsistent in choosing the written form. For example, /an3cuann2/ "how" can be transcribed as 怎樣, 怎麼樣, 按怎, 怎麼, 什麼, and so on. As shown by Minnan dictionaries, 按怎 is listed in the lexicon as the standard form in Minnan. Therefore, it is very important to design a program that can check for inconsistency in the written form.

A spell-checker for Minnan was thus developed.[3] This spell-checker works together with the automatic word segmentation program. When the program is segmenting the text, it searches for words in the columns of "Chinese character" and "alternative forms" in the lexical bank. It then segments the word and adds Minnan Pinyin to the word.

[3] This program was designed by the author and James Myers, and was implemented by Ming-Chung Chang.

The most challenging situation for the spell-checker is probably a case where the transcriber uses a form translated from Mandarin. That is, the form is not a standard Minnan written form. For example, the form 早上 "morning" is not a standard Minnan written form. However, the spell-checker finds that the form 早上 matches an alternative form (in the fourth column) in the lexical bank. In other words, it is very likely a Mandarin form being borrowed by the transcriber. The spell-checker then finds all the Minnan words that have listed 早上 as an alternative form. These are 早起 /ca2khi2^cai2khi2/, 早時 /ca2si5/, 兮早仔 /e1cai2a2^e7ca2a2/,兮早起/e1cai2khi2^e1ca2khi2/, and 透早/thau3ca2/, as shown in Table 9 below.

Table 9. Inconsistency in orthographic transcriptions

Chinese characters	Minnan Pinyin	POS	Synonym/ Alternative forms
早起	ca2khi2^cai2khi2	Nd	***早上***
早時	ca2si5	Nd	***早上***
兮早仔	e1cai2a2^e7ca2a2	Nd	e5早仔、今早、***早上***、下早仔
兮早起	e1cai2khi2^e1ca2khi2	Nd	e5早仔、今早、***早上***、下早仔
透早	thau3ca2	Nd	***早上***

The user can then decide which of the forms matches the pronunciation presented in Minnan Pinyin in the second column.

In summary, the automatic word segmentation program is able to do four things at the same time:

(1) segment words in the text

(2) code Minnan Pinyin for the words already transcribed in Chinese characters

(3) correct inconsistent written forms

(4) expand the lexical bank by adding new words

4. Some Findings from Research based on TAICORP

In this section, preliminary findings based on this corpus are reported. Since the syllable is a fundamental phonological unit, we will focus on findings on syllable distributions, including token and type frequencies.

As mentioned in the introduction, there are about two million syllables in TAICORP. The frequencies of syllables in words and syllables in particles are given in Table 10.

Table 10. Syllable Frequency Counts in TAICORP

	Syllables	
	syllables (in words) 1,558,408	syllables (particles) 538,992
Total	2,097,400	

One interesting finding about syllable distribution is that about 26% of the syllables are particles. There are 26 different syllables found in the corpus, as shown in Table 11 below. Note that, although it is possible to write these particles in Chinese characters, most of them still do not have conventionalized written forms. Also note that some syllables might represent more than one particle. In this case, a digit is added to distinguish among them in the text, *e.g.*, "a1 (啊 1)", "a2 (啊 2)", "a3 (啊 3)." The ones with very low frequencies could be considered idiosyncratic of the speakers.

Table 11. Particles and their token frequencies in TAICORP

Particle	Token frequencies
a 啊	240,103
oo 哦	118,450
la 啦	48,398
le 咧	41,137
hoonn	22,811
ne 呢	19,932
hoo	17,773
u	8,436
hann	7,588
m	6,760
ma 嘛	2,232
hannh	2,144
ue 喂	712
o	667
pa	508
io	466
lioo	440
noo	244
ng	204

Particle	Token frequencies
onn	150
loo	126
oonn	122
me	80
oi	24
na 哪	20
ni	10

These particles mainly serve pragmatic functions [Li 1999; Hung 2003; Hung *et al.* 2004]. Since particles do not have underlying tones, they play a more crucial role in prosody and pose more challenges for speech recognition. This is an area that deserves more attention.

As to syllables in words, the syllable token frequencies of adults and children are given below.

Table 12. Syllable Token Frequencies in TAICORP

	Adults	%	Rank	Children	%	Rank
CV	382760	33.2	1	140028	34.5	1
CVC	260358	22.6	2	79976	19.7	2
CVV	209672	18.2	3	79763	19.7	3
V	122111	10.6	4	47426	11.7	4
CVVC	71852	6.2	5	20092	5.0	5
Subtotal		90.8			90.6	
VC	28341	2.4	6	8438	2.1	6
VV	26126	2.3	7	8389	2.1	7
CN	21392	1.9	8	7661	1.9	8
N	12293	1.1	9	5812	1.4	9
CVVV	8723	0.8	10	3563	0.9	11
VVC	8655	0.8	11	4278	1.1	10
VVV	490	0.0	12	209	0.1	12
Total	1152773			405635		

Note that the top-five most frequent syllable types are the same for both adults and children. They are CV, CVC, CVV, V, and CVVC.

Regarding type frequencies, there are totally 624 different syllables found in both adults' and children's speech. However, different syllables might belong to the same syllable type. For example, the ten words listed in the following table are different syllables with the same

syllable type CV.

Table 13. Examples of CV syllables

example	IPA	Minnan Pinyin	Coding
抱 "hold"	p^ho	pho	CV
馬 "horse"	be	be	CV
霧 "fog"	bu	bu	CV
坐 "sit"	tse	ce	CV
飼 "feed"	$tɕ^h$i	chi	CV
好 "good"	ho	ho	CV
虎 "tiger"	hɔ	hoo	CV
雞 "chicken"	ke	ke	CV
去 "go"	k^hi	khi	CV
三 "three"	sã	sann	CV

In order to obtain syllable type frequencies, it is necessary to code the syllable types first. After coding the syllables, we found that there were a total of 12 different syllable types in Minnan. The syllable type frequencies are as follows.

Table 14. Syllable Type Frequencies in Minnan

Syllable Type	Total
CVC	218
CVV	131
CV	109
CVVC	83
VC	19
CVVV	17
VV	12
CN	12
V	11
VVC	8
VVV	2
N	2
Total	624

To summarize the findings:

(1) The most frequent syllable type is CV. This is consistent with theories in the phonology literature where CV has been considered the core syllable. This is also consistent with the findings in infant vocalization. In another words, this is a very unmarked pattern and might also be a cross-linguistic universal pattern.

(2) The second most frequent syllable type is CVC. This result is not surprising because in speech perception, a CVC syllable might be the easiest to perceive with the acoustic cues from formant transitions of the preceding as well as the following consonant of the nucleus vowel.

(3) Both adults and children have the same top five syllable types, *i.e*., CV > CVC > CVV > V > CVVC. Also note that, CV and CVC syllables count more than half of the total syllables. Even more strikingly, the five most frequent syllable types account for more than 90% of the syllables.

(4) Since the adults show the same patterns as the children, there is a possibility that the children were influenced by the adults (*i.e*., the input lexicon), although this needs to be confirmed by further research.

(5) Compared with data from Dutch children, there is a great similarity between in syllable types. Boersma and Levelt [2000] and Levelt, Schiller, and Levelt [1999] found that the order of acquisition in Dutch children was CV > CVC > V > VC. (These two languages differ in that Dutch does not allow VV syllables and that Minnan does not allow CC consonant clusters.）

(6) We have collapsed the sonorant coda with the obstruent coda (so-called Rusheng or checked syllables), *i.e*., collapsing CVN and CVK into CVC. Since the obstruent codas seem to behave differently [Tsay and Huang 1998], it might be interesting to have an alternative analysis. As Zamuner *et al*. [2005] point out, there seems to be a difference between syllables with sonorant coda and syllables with obstruent coda in English. Some cross-linguistic studies might be worth pursuing.

5. Concluding Remarks

We have introduced the construction of TAICORP, a speech-based corpus of Taiwan Minnan. We have also addressed some issues related to transcribing sound files into text files in Minnan, including multiple pronunciations and the orthographic problems. The automatization process using the corpus has also been illustrated.

This corpus has been used for studies on various aspects of child language acquisition, including tone acquisition [Tsay and Huang 1998; Tsay *et al*. 2000; Tsay 2001], consonant acquisition [Liu and Tsay 2000], vowel development [Lee 2007], classifier acquisition [Myers and Tsay 2000, 2002], final particle acquisition [Hung *et al*. 2004], verb acquisition [Lee and Tsay 2001; Huang 2005; Lin and Tsay, to apper], noun acquisition [Kuo *et al*. 2005],

vocabulary acquisition [Lin 2004; Tsay and Cheng, in preparation]. The corpus is being coded with more phonological annotations such as syllable boundary and tone groups for studying prosodic acquisition. A potential proposal on the WordNet of child language is also being explored. As this corpus is based on spontaneous speech, it also has applications for speech research, for example, analyzing phonetic characteristics of disfluency in child speech.

Acknowledgements

This research was supported by research grants from the National Science Council (NSC 92-2411-H-194-015, NSC 93-2411-H-194-025, NSC 94-2411-H194-002). The author would like to thank the research assistants who participated at different stages of this research, especially Ting-yu Huang, Hui-chuan Liu, Xiao-jun Chen, Peiyu Hsieh, and Yunwei Li. Comments and suggestions from the three anonymous reviewers are also highly appreciated.

References

Boersma, P., and C. Levelt, "Gradual Constraint-Ranking Learning Algorithm Predicts Acquisition Order," *The Proceedings of the Thirtieth Annual Child Language Research Forum*, 2000, pp, 229-237.

Chen, K.-J., C.-R. Huang, L.-P. Chang, and H.-L. Hsu, "SINICA CORPUS: Design Methodology for Balanced Corpora," *Language, Information, and Computation*, 11, 1996, pp. 167-176.

CKIP, "Chinese Part-of-Speech Analysis," Technical Report No. 93-05, Institute of Information Science Academia Sinica, Taipei, 1993.

CKIP, "Content and Instruction of the Sinica Balanced corpus (revised version)," Technical Report No. 95-02, Institute of Information Science Academia Sinica, Taipei, 1998.

Huang, C.-R., K.-J. Chen, F.-Y. Chen, and L.-L. Chang, "Segmentation Standard for Chinese Natural Language Processing," *Computational Linguistics and Chinese Language Processing*, 2 (2), 1997, pp. 47-62.

Huang, S., *Yuyan, Shehui yu Zuqun Yishi [Language, Society, and Ethnic Awareness]* Crane Publishing, Taipei, 1993. 黃宣範，語言、社會與族群意識 -- 台灣社會語言學的研究，文鶴出版公司，台北，1993。

Huang, Y.-C., The Child's Acquisition of Verbs in Taiwanese, MA thesis, National Chung Cheng University, Taiwan, 2005.

Hung, C. C.-F., The Child's Utterance Final Particles in Taiwanese: A Case Study, MA thesis, National Chung Cheng University, Taiwan, 2003.

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

Kuo, J., J. Tsay, and J. Peng, "Basic Level Effects in Taiwanese Noun Acquisition." In *Proceedings of 2005 Conference on Taiwan Culture: Linguistics, Literature, Culture and Education,* 2005, Chia-yi: National Chiayi University. pp. 43-54.

Lee, T. H.-T., and J. Tsay, "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.

Lee, Y.-W., Vowel Development of a Child Acquiring Taiwan Southern Min, MA thesis, National Chung Cheng University, Taiwan, 2007.

Levelt, C. C., N. O. Schiller, and W. J. M. Levelt, "A developmental grammar for syllable structure in the production of child language," *Brain and Language*, 68, 1999, pp. 291-299.

Li, Y. C., *Utterance-final Particles in Taiwanese: a Discourse-pragmatic Analysis*, Crane Publishing, Taipei, 1999.

Lin, P.-C., The Acquisition of Nouns and Verbs in Taiwanese, MA thesis, National Chung Cheng University, Taiwan, 2004.

Lin, H.-L., and J. Tsay, "Acquiring Causatives in Taiwan Southern Min," *Journal of Child Language*, to appear.

Liu, J. H. C., and J. Tsay, "An Optimality-Theoretic Analysis of Taiwanese Consonant Acquisition," In *Proceedings of the 7th International Symposium of Chinese Languages and Linguistics*, 2000, National Chung Cheng University, Chiayi, Taiwan, 2000, pp. 107-126.

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

MacWhinney, B. and C. Snow, "The Child Language Data Exchange System," *Journal of Child Language*, 12, 1985, pp. 271-296.

Myers, J., and J. Tsay, "The Acquisition of the Default Classifier in Taiwanese," In *Proceedings of the 7th International Symposium of Chinese Languages and Linguistics*, 2000, National Chung Cheng University, Chiayi, Taiwan, pp. 87-106.

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

Tsay, J., "Phonetic Parameters of Tone Acquisition in Taiwanese," *Issues in East Asian Language Acquisition*, ed. by Minehru Nakayama, Kuroshio Publishers, Tokyo, 2001, pp. 205-226.

Tsay, J., "Taiwan Child Language Corpus: Data Collection and Annotation," In *Proceedings of 5th Workshop on Asia Language Resources*, Jeju Island, Republic of Korea, 2005a, pp. 56-61.

Tsay, J., "The Language Issue ," Documentation of Chaiyi City, the Language and Literature Volume, Vol. 8, Chiayi City Hall, Chiayi, Taiwan, 2005b, pp. 1-66. (蔡素娟撰，嘉義市政府編印，嘉義市志語言文學志語言篇，嘉義，2005b，pp. 1-66)

Tsay, J. and C. C. Cheng, "Productivity in Young Children's Vocabulary," in preparation. Manuscript, National Chung Cheng University, Taiwan.

Tsay, J. and T.-Y. Huang, "Phonetic Parameters in the Acquisition of Entering Tones in Taiwanese," In *The Proceedings of the Conference on Phonetics of the Languages in China,* City University of Hong Kong, China, 1998, pp. 109-112.

Tsay, J., J. Myers, and X.-J. Chen, "Tone Sandhi as Evidence for Segmentation in Taiwanese," In *Proceedings of the 30th Child Language Research Forum,* Center for the Study of Language and Information, Stanford, California, 2000, pp. 211-218.

Zamuner, T. S., L. Gerken, and M. Hammond, "The Acquisition of Phonology Based on Input: A closer look at the relation of cross-linguistic and child language data," *Lingua*, 115(10), 2005, pp. 1403-1426.

Minnan Dictionaries

Chen, X., Taiwanhua Dacidian [Taiwanese Dictionary], Yuanliu Publishing, Taipei, 1998. 陳修，台灣話大辭典：閩南話漳泉二腔系部分，遠流出版事業股份有限公司，台北，1998。

Dong, Z., Taiwan Minnanyu Cidian [Taiwan Southern Min Dictionary], Wunan Publisher, Taipei, 2001. 董忠司編纂，國立編譯館主編，台灣閩南語辭典，五南圖書出版有限公司，台北市，2001。

Li, R., Xiamen Fangyan Cidian [Xiamen Dialect Dictionary], Education Publisher, Jiangsu, 1998. 李榮，廈門方言詞典，江蘇教育出版社，江蘇省，1998。

Wu, S., Guotaiyu Duizhao Huoyong Cidian [Mandarin-Taiwanese Comparative Dictionary], Yuanliu Publising, Taipei, 2000. 吳守禮主編，國台語對照活用辭典—詞性分析、詳註廈漳泉音（上冊, 下冊），遠流出版有限公司，台北，2000。

Xu, J., Changyong Hanzi Taiyu Cidian [Taiwanese Dictionary of Frequently Used Chinese Characters]. Culture Department, Zili Evening News, Taipei, 1992.許極燉編著，常用漢字台語辭典，自立晚報社文化出版部，台北市，1992。

Yang, Q., Guotai Shuangyu Cidian [Mandarin-Taiwanese Bilingual Dictionary], Duli Publishing, Kaohsiung, 1993. 楊青矗主編，國台雙語辭典，敦理出版社，高雄，1993。

Yang, X., Minnanyu Cihui [Southern Min Vocabulary], Ministry of Education, Taipei, 2001. 楊秀芳撰稿，教育部國語推行委員會編輯，閩南語字彙（一, 二）修訂版，教育部，台北市，2001。