THE OXFORD HANDBOOK OF

CORPUS PHONOLOGY

Edited by

JACQUES DURAND, ULRIKE GUT,

and

GJERT KRISTOFFERSEN

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© editorial matter and organization Jacques Durand,
Ulrike Gut, and Gjert Kristoffersen 2014
© the chapters their several authors 2014

The moral rights of the authors have been asserted

First Edition published in 2014
Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2014933501

ISBN 978-0-19-957193-2

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

CHAPTER 32

A PHONOLOGICAL CORPUS OF L1 ACQUISITION OF TAIWAN SOUTHERN MIN*

JANE S. TSAY

32.1 INTRODUCTION

THIS chapter describes data collection, transcription, and annotations for the Taiwanese Child Language Corpus. Computer programs developed for specific phonological analyses will also be described briefly.

The Taiwanese Child Language Corpus (TAICORP hereafter) is a corpus of spontaneous speech between young children growing up in Taiwanese-speaking families and their carers. The target language Taiwanese is a variety of Southern Min Chinese spoken in Taiwan.[1] Taiwanese and Southern Min are used interchangeably in this chapter.

In the literature on child language acquisition, studies have focused primarily on universal innate patterns, described in current phonological theories with markedness constraints. For example, Optimality-theoretic (OT) models of child language acquisition make specific predictions about markedness (e.g. Prince and Smolensky 1997; Tesar and Smolensky 1998). A number of studies have followed this line (e.g. Barlow and Gierut

* This project has been supported by research grants from the National Science Council, Taiwan, for more than ten years (NSC87/88/89-2411-H-194-019, NSC89/90/91-2411-H-194-067, NSC92/93/94-2411-H-194-015, NSC95-2411-H-194-022-MY3, NSC98-2410-H-194-086-MY3). We thank the children and their families for their participation. Without the research assistants over the years, especially Rose Huang, Joyce Liu, and Kay Chen, this project would not have gone so far.

[1] Southern Min (or Minnan) originally referred to the southern area of Min (Fujian Province, China), including Xiamen (or Amoy), Zhangzhou, and Quanzhou. Most of the early immigrants to Taiwan more than 300 years ago came from the Zhangzhou and Quanzhou areas.

1999; Barlow 2001; Boersma and Hayes 2001; Dinnsen 2001; Dinnsen and Gierut 2008; among others).

However, learning phonology also requires learning the language-specific sound patterns found in the adult language's specific lexicon. In particular, the lexicon contains crucial information about frequency. Therefore, we expect both universal markedness and language-specific lexical properties to be available for the child. This point of view has also been recognized in recent years (e.g. Gierut 2001; Zamuner et al. 2005; Lleó 2006; Fikkert and de Hoop 2009; Rose 2009; Tessier 2009). Beckman and colleagues have also emphasized different levels of phonological abstractness and the role of vocabulary size in phonological acquisition (Edwards et al. 2004; Beckman et al. 2007; Edwards and Beckman 2008a, b; Beckman and Edwards 2010; Edwards et al. 2011).

However, vocabulary size and frequency information regarding sound patterns in child language are difficult to obtain due to methodological limitations. To obtain frequency information, we need child language corpora that are very large, but that also present a great amount of phonological detail. While a lexicon provides type frequency information, a corpus can provide token frequencies. It is therefore very desirable to adopt a corpus-based approach towards child phonology acquisition (e.g. Zuraw 2007, and works cited therein).

Although the significance of a large-scale collection of longitudinal child language data for linguistic studies goes without saying, there was an additional motivation for studying the acquisition of Taiwan Southern Min.

Until almost the turn of the century, for over forty years, Mandarin was the only official language for instruction in schools in Taiwan in spite of the fact that about 73 per cent of the population belonged to the Southern Min ethnic group (Huang 1993). Young children in kindergartens and elementary schools were not allowed to speak Southern Min at school even though it was the language spoken at home. Although the situation has changed in recent years and languages other than Mandarin, including Southern Min, Hakka, and the aboriginal (Formosan) languages have been included in the curriculum of elementary schools, there still is a serious concern about the dwindling numbers of native Southern Min speakers. This concern can be supported by a more recent report by Tsay (2005) which found that, in a survey of all 8th graders in Chiayi City in Southern Taiwan, only about 26 per cent of these 14-year-olds used Southern Min in their daily life, compared with results showing that over 70 per cent of their parents were native Southern Min speakers.

Given these considerations, we saw some urgency in the study of the first language acquisition of Southern Min, which motivated the construction of our corpus, beginning more than ten years ago.

Data collection was done through regular home visits during a period of three years. A total of fourteen children (four 1-year-olds, seven 2-year-olds, three 3-year-olds) participated in this longitudinal study during the three-year data collection phase. There were about 330 hours of recordings from the 431 recording sessions (see below for details on the recording sessions). These recordings were transcribed into 431 text files which

together contain 1 646 503 word tokens (about 2 million syllables/morphemes/characters)[2] in 497 426 lines (utterances).

The recordings were transcribed into machine-readable text in the format of the Child Language Data Exchange System (CHILDES) (MacWhinney and Snow 1985, MacWhinney 1995, Rose and MacWhinney, this volume).

Annotations in the text files include part of speech, narrow phonetic transcription of the child speech, and syllable types. For some young children, the sound files were also synchronized with the annotated text using a function in CLAN of CHILDES. Discourse annotations were only minimally coded due to manpower limitations.

The following sections describe data collection, transcription, text files in CHILDES format, annotated phonological information, and data analysis programs.

32.2 Data Collection

Data collection took place over a period of three years. Fourteen children (9 boys and 5 girls) were recruited from Southern Min-speaking families in Min-hsiung Township, Chiayi County in Southern Taiwan.

Home visits were conducted at two-week intervals at the child's home by one investigator of the project, accompanied by a carer (parent or grandparent for most children, the nanny for one child). The children's spontaneous speech while at play or interacting with the carer and/or the investigator was recorded using a digital MiniDisc recorder and a microphone. Each recording session lasted from 40 to 60 minutes. The activities were children's daily life at home—playing games or playing with toys, reading picture books, or just talking without any specific topics.

Three research assistants participated in this project, each being responsible for the longitudinal recording of three to four children during the three-year data collection phase. These research assistants were also the first-round transcribers of the recordings.

Each child had his/her own recording schedule and had different participation durations. One child, YJK, was recorded only twice because he was speaking more Mandarin than the target language Southern Min. Three children, LJX, YCX, and YDA, were recorded for half a year until they went to preschool and started picking up Mandarin. The other ten were recorded for at least one year and seven months. Among them, six children were recorded for more than two years. The gender, age range, participation duration, number of recording sessions, and recording time of each child are given in Table 32.1.

There were a total of 431 recording sessions. Each session was saved as a separate sound file. Because the recordings were done in a natural setting, long periods of silence

[2] Like other Sinitic languages, most morphemes in Taiwanese are monosyllabic, each corresponding to one Chinese character in the orthography.

Table 32.1 Information about the children and their recording sessions

Name	Sex	Age range	Duration of participation	Recording Sessions	Length (min.)
LYC	F	1;2.13–3;3.29	2yr 2mo	48	2255
HYS	M	1;2.28–3;4.12	2yr 3mo	51	2280
TWX	F	1;5.12–3;6.15	2yr 2mo	44	1829
YSW	M	1;7.17–2;7.14	1yr 1mo	21	1210
LWJ	F	2;1.08–3;7.03	1yr 7mo	36	1777
WZX	M	2;1.17–4;3.15	2yr 3mo	44	1757
HBL	M	2;1.22–4;0.03	2yr 0mo	45	1889
CEY	F	2;1.27–3;10.00	1yr 10mo	37	1728
YJK	M	2;6.11–2;6.26	0yr 1mo	2	105
LMC	F	2;8.07–5;3.21	2yr 8mo	50	2045
CQM	M	2;9.07–4;6.22	1yr 10mo	30	1584
LJX	M	3;9.20–4;2.24	0yr 6mo	8	530
YCX	M	3;10.16–4;0.16	0yr 6mo	6	285
YDA	M	3;11.02–4;4.26	0yr 6mo	9	540
Total	M = 9 F = 5			431	330 hours

and background noises were inevitable. So the sound files were first edited to delete the long silences and noisy parts. In order to permit easier searching and locating of the content on the recordings, each sound file was segmented into several tracks, which were then tagged.

32.3 Text Files

32.3.1 Transcription

All sound files were transcribed into text files in both orthographic transcription and phonetic transcription.

32.3.1.1 Orthographic Transcription

There are two kinds of systems used in orthographic transcription: the logographic orthography (i.e. Chinese characters—traditional Chinese writing system) used in the main tier and a spelling-based romanization system for Taiwan Southern Min (Minnan

Pinyin) used in the dependent tier %ort. (The names and descriptions of the tiers will be explained shortly.)

It should be noted that some problems were encountered in transcribing the speech into Chinese characters. Although all Sinitic languages use Chinese characters as the writing system, only Mandarin has an almost perfect mapping between the spoken words and the written words. This is probably the consequence of Mandarin being the 'official' spoken language assigned by the government since the turn of the twentieth century. The mass media (especially the newspapers) have also helped in conventionalizing the written form. By contrast, Taiwan Southern Min does not yet have as conventionalized an orthography and quite a few words in Taiwan Southern Min do not have a consistent written form.

In order to increase the consistency in Taiwan Southern Min writing conventions, several Southern Min dictionaries were consulted. A program for checking the inter-transcriber consistency of the orthography was developed by Galvin Chang, James Myers, and Jane Tsay (see Tsay 2007 for more detailed discussion).

32.3.1.2 *Phonetic Transcription*

Phonetic transcription was first done by the investigator who made the recording for a particular session. Another investigator would do a second-round transcription. A third investigator then checked discrepancies between the first two transcriptions. Segments were transcribed in IPA in the %pho tier and tones were transcribed using a 5-point scale (where 1 = lowest pitch, 5 = highest pitch) in the %tone tier. The child speech in about 180 out of the 330 hours of total recordings was transcribed phonetically.

32.3.2 The Text Format

The format of the text files in CHILDES is introduced very briefly in this chapter. For details, including the tools of transcription, please refer to the official CHILDES website at http://childes.psy.cmu.edu/.

The main components of text files in the CHILDES format are *headers* and *tiers*. There are three kinds of headers: obligatory headers, constant headers, and changeable headers. Obligatory headers are necessary for every file. They mark the beginning and the end of the file. Constant headers mark the name of the file and background information of the children, while changeable headers contain information that may change across files, such as language, participant ID, recording date, transcribers, and so on. These headers all begin with @.

```
Headers

  @Begin
  @Languages:zho-nan, zho
  @Participants:CHI    LYC    Target_Child,    IN1    Kay
     Investigator, IN2 Rose Investigator, SIS Ci Sister
  @ID:zho-nan|Tsay|CHI|3;3.29|female|||Target_Child||
  @ID:zho-nan|Tsay|IN1|||||Investigator||
```

```
@ID:zho-nan|Tsay|IN2|||||Investigator||
@ID:zho-nan|Tsay|SIS|||||Sister||
@Transcriber:Rose, Kay, Joyce
@Date:31-MAY-2000
@Media: LYC30329, audio
@Tape Location:Yi D17-1-10
@Comment:Time duration is 40
@Location:Chiayi, Taiwan
@Transcriber:Kay
@Comment:Track number is D17-1
```

The content of the speech is presented in tiers: the main tier and the dependent tiers. The main tier, marked with *, contains the utterance of the speaker, for example, *CHI the target child, *MOT the mother, and *INV the investigator. The speech in the main tier in TAICORP is transcribed in Chinese characters.

The dependent tiers, marked with %, are for additional information about the utterance in the main tier. The names and description of the dependent tiers in TAICORP are given below.

```
%ort: utterance in Southern Min romanization (Minnan
  Pinyin)
%cod: part-of-speech coding
%eng: English gloss of the words
%syl: syllable types (in CV notation) of the target (pho-
  nemic) pronunciation
%pho: phonetic transcription of the child speech
%syc: syllable types of the actual production of the
  child speech
%ton: phonetic transcription of the tones on a
  5-point scale
```

The following example is from the child HYS at 2;3.9.[3]

```
Tiers
*CHI:你 食.
%ort:li2 ciah8.
%cod:Nh VC
%eng:you eat
%syl:CV CVVK
%pho:i kia
%syc:V CVV
%ton:55 32
```

[3] Digits in the %ort tier denote the lexical tone categories (to be explained in the next section) of the syllable, while digits in the %ton tier are pitch values of the tone categories in the child's actual pronunciation.

Table 32.2 Lexical tones in Taiwanese

Tone Category	Traditional terms	Tone values in junctureor isolation	Example	Gloss
Tone 1	Yinping	55	si^{55}	Poem
Tone 2	Yinshang	53	si^{53}	Death
Tone 3	Yinqu	21	si^{21}	Four
Tone 4	Yinru (short tone)	33	sik33	Colour
Tone 5	Yangping	13	si^{13}	Time
Tone 7	Yangqu	33	si^{33}	Yes
Tone 8	Yangru (short tone)	5	sik5	Ripe

32.4 Phonological Information

32.4.1 The Sound System of Taiwanese

Like all languages in the Sinitic (Chinese) family, Taiwanese has the following characteristics regarding syllable and tone: (1) virtually all morphemes in Taiwanese are monosyllabic; (2) lexically contrastive tones occur with almost all syllables, except for some function words like particles which do not have an underlying tone and usually surface with a neutral tone.

Syllable structure is relatively simple in Taiwanese. No consonant clusters are allowed. The consonant in the coda position of a syllable is very restricted—only nasals and stops are allowed.

There are seven lexical tones in Taiwan Southern Min. Tone categories are referred to by digits, including five long tones (T1, T2, T3, T5, and T7) and two short (abrupt) tones (T4 and T8).[4] Short tones only occur with the so-called checked syllables (which end with an unreleased obstruent coda -p, -t, -k, or glottal stop) and are called Rusheng or Entering Tone in the Chinese philology tradition. Tone values in 5-point scale notation for syllables/morphemes in juncture position (including in isolation) are given in Table 32.2 with short tones underlined.[5]

In addition to the seven lexical tones described in the previous paragraph, there is also a neutral tone, labelled Tone 0, which occurs in underlyingly toneless words such

[4] Tone 6 does not exist in Taiwanese any more due to historical sound change. For the purpose of diachronic comparison as well as synchronic dialectal studies, this gap is still respected and preserved in the numbering of the tone categories.

[5] Surface tone values are different in juncture (including in isolation) vs. non-juncture (context) positions, a phenomenon called tone sandhi in the literature.

as particles. These particles usually serve pragmatic functions. The actual realization of the particles might vary according to pragmatic situations. Another non-lexical tone, labelled Tone 9, is a non-contrastive tone, usually derived from tone contraction due to the coalescence of two adjacent syllables into one syllable.

The consonants in Taiwanese are: /p, pʰ, b, m, t, tʰ, l, n, k, kʰ, g, ŋ, h, ʔ, ts, tsʰ, s, dz/. As mentioned above, the coda stops /-p/, /-t/,. /-k/, and /-ʔ/ are unreleased and are the only obstruents that can appear in the coda position. The other coda consonants are nasals /-m/, /-n/, and /-ŋ/. The labial nasal and velar nasal can also be syllabic as in /am̩/ 'aunt' and /ŋ̍/ 'yellow', respectively.

There are six single vowels: /i, e, a, ɔ, o (or ə in some dialects), u/. Single vowels can be combined into diphthongs and triphthongs. Except for /o/, all single vowels also have a nasalized counterpart.

32.4.2 Phonological Coding and Phonological Analysis

Regarding phonological analysis, syllable types and tone types have been the main concern in our research. Two programs *SYLLABLE* and *ToneFreq* were developed to study these issues.

32.4.2.1 Coding Syllable Types and Counting Syllable Token Frequencies

The program *SYLLABLE* was designed by Chienyu Hsu to code syllable types in the %ort tier (the target/phonemic sound in Minnan Pinyin). Since all syllables in the %ort tier end with a digit which denotes the tone categories, these digits mark syllable boundaries at the same time. Therefore, by replacing the digits at the end of the syllable with a space, all syllables in the %ort tier were segmented.

The isolated syllables were then coded with C (consonant) or V (vowel). Twelve syllable types in Taiwan Southern Min were found: V, CV, VC, VV, CVC, CVV, VVC, VVV, CVVC, CVVV, CN, N.[6] Token frequencies for each syllable type can therefore be obtained by running the FREQ program in CLAN in CHILDES, treating the separated syllables as separate words.

Syllable type coding in the %ort tier is based on the target speech (i.e. adult speech), while syllable type coding in the %pho tier is based on the actual production of the child. Thus by comparing (mapping) the syllable type coding in these two tiers, the child's syllable errors (i.e. the mismatches) can be identified. For example, consider the passage given in Section 3.2 above, where the child was supposed to say *li ciah* (CV CVVK) 'you eat', but said *i kia* (V CVV) instead. We can see that there is an onset deletion in the first syllable and a coda deletion in the second syllable. The program *SYLLABLE* compares these two tiers and gives the results of the deviation of the child speech from the adult speech.

[6] In addition to standard C and V codes for consonants and vowels, coda Cs could be further coded as N (nasal coda) or K (obstruent coda) when this distinction in coda becomes relevant.

32.4.2.2 *Tone Distribution*

Another program *ToneFreq*, also developed by Chienyu Hsu, was designed to count tone frequencies at both the syllable level and the word level. Both type frequencies and token frequencies of the tone categories can provide interesting information about the characteristics of lexical tone. For example, it was found that syllables with certain onsets do not occur with certain tones, a potential phonotactic constraint very likely caused by historical sound change. Whether this kind of sound pattern plays a role in tone acquisition is an issue worth pursuing.

32.4.2.3 *Digitized Audio Linkage*

Using a tool in CLAN of CHILDES, the sound file and the text file of a recording can be synchronized. A command in the phonic mode called 'insert bullet' can link the sounds to the computerized transcript line by line. This makes it much easier for the user to hear the sounds while reading the transcript. It also makes acoustic analysis much easier, although the manual insertion of the bullets remains labour-intensive. In TAICORP, this work has been completed for the three youngest children (about 106 hours).

32.5 POS and Discourse Annotations

32.5.1 Automatic Word Segmentation and POS Tagging

Constructing a speech-based corpus requires a lot more steps than constructing a corpus based on written texts. The most labour-intensive and time-consuming work is transcribing the sound files into text files. In the first stage of the construction of TAICORP, every step was done manually. After a lexical bank was built based on the manual transcription, automatic word segmentation and part-of-speech (POS) tagging became possible.

The POS coding system used in TAICORP is a revised version of the system used in the Sinica Corpus of Mandarin (see various technical reports by the Chinese Knowledge Information Processing Group (CKIP) (CKIP 1993, 1998; Chen et al. 1996). For details about the POS coding system of TAICORP, see Tsay (2007).

A sample of the lexical bank is given in Table 32.3, which contains the following information for each lexical entry: the orthography (both in Chinese characters and in Minnan Pinyin), English gloss, POS tag(s), synonyms in Mandarin, and an example from the corpus.

The programs listed in Table 32.4 were developed by Galvin Chang for the automatization of the transcription and coding processes.

It should be noted that, morphological (including both derivational and inflectional) marking is very limited (with very few exceptions, e.g. plural marker *men* for a limited

Table 32.3 A sample of the lexical bank of TAICORP

Chinese	MinnanPinyin	EnglishGloss	POSTag	Synonyms in Mandarin	Example from the corpus
公	kang1	male	A	公	你這尾魚仔是公e0抑母e0?
抑無	ah8bo5^ah4bo5	otherwise	Cbb	要不然,否則, 不然, 要 不	抑無你看這張
大概	tai7khai3	approximately	D	大概	大概是按呢la0.
畫粧	ue7cong1	make up	VA	化妝、化粧	早起起來畫(粧) [/] 畫粧e0時陣伊創啥?

Table 32.4 Programs developed to automatize the transcription and coding processes

Program	Functions	Programmer
Spell-Checker	Check the accuracy of the orthography, including the spelling in Minnan Pinyin and Chinese characters.	Galvin Chang
Automatic Word Segmentation	Segment an utterance into words and automatically insert Minnan Pinyin for each word in Chinese characters	Galvin Chang
Automatic POS Tagging	Code the POS tags automatically after word segmentation	Galvin Chang

set of nouns). Therefore, for word forms that have more than one syntactic category (e.g. *water* could be a noun or a verb in English), it is ambiguous when they occur without context. This might cause problems in word frequency counts as will be discussed in the next section. For the automatization procedure for transcription and coding, and details about some issues associated with transcribing into Chinese characters and Minnan Pinyin, see Tsay (2007).

32.5.2 Data Analysis Programs

A program called *WordClassAgent* was developed by Ziyang Wang mainly for word frequency counts (Table 32.5). There are two sub-programs in *WordClassAgent*: one selects words in a specific main tier (i.e. speech for that specific speaker only); the other matches the words with their syntactic categories in the %cod tier where POS is coded. The former gives the word frequency counts for a specific speaker and the latter gives the word frequency counts for a specific syntactic category.

Table 32.5 Programming for word frequency counts

Program	Functions	Programmer
WordClassAgent	Mainly for word frequency counts	Ziyang Wang
Sub-program 1: SpeakerSelection	Sort data by different participants/speakers (main tier)	
Sub-program 2: ChildWordClass	Count word frequencies with specific constraints, e.g. verbs vs. nouns (i.e. words with specific syntactic categories at the %cod tier)	

32.5.3 Discourse Annotation

Due to manpower limitations, discourse annotation was minimal. Discourse codes used in TAICORP are from the conventions provided in CHILDES.

(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, crying, singing, and whispering.
(b) [>] Overlap follows
(c) [<] Overlap precedes
(d) [/] Retracing without correction
(e) [//] Retracing with correction

32.6 Final Remarks

TAICORP was originally intended for the study of phonological acquisition and has generated endless tasks at that level. What is encouraging is that even though there will never be an end to improving a speech corpus, many studies have been able to use the corpus for various topics, for example, phonology acquisition (Tsay et al. 2000; Liu and Tsay 2000; Tsay 2001; Myers and Tsay 2011), classifier acquisition (Myers and Tsay 2000, 2002), causative acquisition (Lin and Tsay 2008), acquisition of syntactic categories (Tsay and Cheng 2011), pragmatics acquisition (Hung et al. 2004), and several in-progress studies. Part of the corpus has been donated to the PhonBank. We hope that with the availability of the corpus, more people will be able to use the data collected.