TAIWAN HAKKA LANGUAGES AND TWHK_ToBI ANNOTATION CONVENTIONS

Shao-ren Lyu & Ho-hsien Pan

National Chiao Tung University, Hsinchu, Taiwan
thousandshine.fl96@nctu.edu.tw; hhpan@faculty.nctu.edu.tw

ABSTRACT

This paper proposes a preliminary prosodic annotation system for Taiwan Hakka, "Taiwan Hakka Tones and Break Indices" is called TWHK_ToBI. TWHK_ToBI includes five tiers: ortho, words, tones, breaks, and miscellaneous. The ortho tier contains Romanization of each syllable and dictionary-defined tones; the words tier includes alphabetized SAMPA spellings of each word; the tones tier includes the sandhi tones for each syllable; the breaks tier indicates degree of juncture including words, fused words, intermediate phrase and intonational phrase boundaries; and the miscellaneous tier labels events such as code switching, laugh, and cough.

Keywords: Hakka, Sixian, Hailu, sandhi, ToBI

1. INTRODUCTION

Sixain and Haili are two most common varieties of Taiwan Hakka. Sixian is spoken by 38.4% of Hakka population, Hailu by 25.6% [6]. Although there are some differences between these varieties, including lexical tones, word usage, and pronunciation, people of these two varieties can understand each other easily. Therefore, the prosodic structure of both Sixian and Hailu varieties can be described with a single ToBI system. Though there are approximately 5.8 million Hakka people in Taiwan [7], there is no prosodic annotation system specifically for Hakka.

2. CHARACTERISTICS OF HAKKA

2.1. Syllables

In Hakka, syllable is the tone-bearing unit. A syllable consists of vowel, complex vowel or syllabic nasal nucleus with an optional onset and coda consonant. The onset and coda consonants are not necessary in a syllable as in these four words in the Sixian variety contain no onset or coda: /i31/ 'rain', /oi55/ 'love', /ieu11/ 'distant', and /ŋ11/ ' fish'.

2.1.1. Onset consonants

There are 21 onset consonants in the Sixian variety and 20 onset consonants in the Hailu variety (Table 1). Most of the onset consonants are shared by these two varieties. Yet, In Sixian, there are no palato-alveolar fricatives /ʃ/, /tʃ /, /tʃ^{h}/, and /ʒ/. Words produced with /ʒ/ in other varieties of Hakka are produced with a front high vowel /i/ in Sixian. The palato-alveolar affricate consonants /ʃ/, /tʃ/ and /tʃ^{h}/ are replaced by dental affricate consonants /s/, /ts/, and /ts^{h}/ in Sixian. An interesting phenomenon occurs when the dental affricate consonants /ts/, /ts^{h}/, and /s/ are followed by a front high vowel /i/. In Sixian, the /ts/, /ts^{h}/, and /s/ will be palatalized; that is, /tɕ/, /tɕ^{h}/, and /ɕ/ (Table 1).

Table 1: Onset consonants in Sixian and Hailu varieties. Romanization and IPA in this paper are based on [1], [2], and [3] with modification.

Onset Consonants			
Romanization	**IPA**	**SAMPA**	**Note**
b	p	b	
c	ts^{h}	c	
c(i)	tɕ^{h}	q	Sixian Variety
ch	tʃ^{h}	ch	Hailu Variety
d	t	d	
f	f	f	
g	k	g	
h	h	h	
k	k^{h}	k	
l	l	l	
m	m	m	
n	n	n	
ng	ŋ	ng	
ng(i)	ȵ	NG	
p	p^{h}	p	
r	ʒ	rh	Hailu Variety
s	s	s	
s(i)	ɕ	x	Sixian Variety
sh	ʃ	sh	Hailu Variety
t	t^{h}	t	
v	v	v	
z	ts	z	
z(i)	tɕ	j	Sixian Variety
zh	tʃ	zh	Hailu Variety

2.1.2. Vowels

There are monophthongs, diphthongs, and triphthongs in Taiwan Hakka. There are 22 identical combinations of vowels in both Sixian and Hailu. Though /ɨ is shared by these two varieties, coda consonants can only be attached to the centralized vowel in Sixian. Here is a unique vowel /ɤ/ in Hailu, and an extra triphthong /iui/ in Sixian (Table 2). The /ɤ/is only utilized as a bound phoneme added to the final position of a phrase in Hailu variety, like "thief" (/ts^{h}et3 ɤ55/). The triphthong /iui/ occurs only in the word, "sharp" (/iui55/), in Sixian [4].

Table 2: All possible combinations of vowels in Sixian and Hailu.

Combinations of Vowels			
Romanization	**IPA**	**SAMPA**	**Note**
a	a	a	
ai	ai	ai	
au	au	au	
e	e	e	
er	ɤ	@	Hailu Variety
eu	eu	eu	
i	i	i	
ia	ia	ia	
iai	iai	iai	
iau	iau	iau	
ie	ie	ie	
ieu	ieu	ieu	
ii	ï	!	
io	io	io	
ioi	ioi	ioi	
iu	iu	iu	
iui	iui	iui	Sixian Variety
o	o	o	
oi	oi	oi	
u	u	u	
ua	ua	ua	
uai	uai	uai	
ue	ue	ue	
ui	ui	ui	

2.1.3. Coda consonants

There are only six possible coda consonants in Hakka. They are a group of voiceless stop consonants /p/, /t/, and /k/, and a group of nasal consonants /m/, /n/, and /ŋ/. (Table. 3)

Table 3: All possible coda consonants of Hakka language.

Coda Consonants			
Romanization	**IPA**	**SAMPA**	**Note**
b	p	p	
d	t	t	
g	k	k	
m	m	m	
n	n	n	

2.2. Lexical tones

There are six different tones in Sixian, and seven different tones in Hailu. The Chao-number is employed for the tones of Hakka languages. The highest tone is transcribed as "5", and the lowest tone is transcribed as "1". Each unchecked tone is signaled with two numbers; unchecked tone is subdivided into rising tones, falling tones, and level tones. The checked tones are written in only a single number. While there are the same sets of tones in these two dialects, they are not used in the same set of words. For example, tones 55 and 11 are reversed. The phrase" Taiwan languages" in Sixian is /t^{h}oi11 van11 fa55/, whereas in Hailu is /t^{h}oi55 van55 fa11/. Yet, the reverse of lexical tones seldom causes confusion, because people in these two varieties know the other ones speak in the other variety. A complete list of tones is provided in Table 4.

Table 4: The correspondent tones between Sixian and Hailu varieties, based on [3] with modification.

Lexical Tones							
	Unchecked Tones					**Checked Tones**	
Sixain Variety	11	24	31	55	55	3	5
Hailu Variety	55	53	24	33	11	5	3

2.3. Sandhi regulations

In Sixian variety, the sandhi changes will occur when lexical tone /24/ is followed by the unchecked tone /24/, /55/, or the checked tone /5/. The following tone /24/ will be pronounced as tone /11/.

In Hailu, the unchecked tone /13/, and the checked tone /5/ will change into tone /33/ when followed by any tone in a phrase (Table 5).

Table 5: Dictionary tones and tonal alternatives in common phrases.

Sixian Variety		**Hailu Variety**	
Dictionary Tones[4]	**Sandhi Changes**	**Dictionary Tones**	**Sandhi Changes**
24+24	11+24	24+any tone	33+any tone
24+55	11+55	5+any tone	3+any tone
24+5	11+5		

3. TWHK_ToBI ANNOTATION CONVENTIONS

Currently, there are five tiers in TWHK_ToBI system: ortho, words, tones, breaks, and miscellaneous. The ortho tier includes a Romanization transcription accompanied by a dictionary specification of tones like the ortho tier

in TW_ToBI [5]. For example in Figure 1, the phrase /lo24 ŋin55 pi24 se11 ŋin55 to53/ 'the old men are more than the young children' is transcribed as "lo24 ngin55 bi24 se11 ngin55 do53" in the ortho tier.

The words tier includes SAMPA symbols as shown in Table 1, 2, and 3 without tones. For example in Figure 1, the phrase /lo24 ŋin55 pi24 se11 ŋin55 to53/ is transcribed as "lo NGin bi se NGin do."

The tones tier is for the sandhi tones and tone values of fused syllables. A sandhi tone will be tagged with "s"; whereas an unexpected sandhi tones will be tagged with "u". Tones spanning fused syllables are tagged with "f" (Table 6). For example in Figure 1, the tone of the syllable /lo24/ surfaced as sandhi tone 33, so this syllable is marked with "33s" in the tone tier. In Figure 2, the phrase /ŋi11 ti53 mo55/ is produced with a fused form with tone 11 unexpectedly. Thus this phrase is marked as 11fu in tone tier. Yet, to our surprise, the tones of the three-word fusion are exceptions of sandhi rules. For example (see Figure 2), [nio11] is the fusion of "do you know" [ŋi55 ti53 mo55] (literally: you know interrogative).

Table 6: Symbols of tone tier in TWHK_ToBI.

s	occurrence of expected sandhi tones (Fig. 1)
u	occurrence of unexpected sandhi tones
f	occurrence of fused syllables (Fig 2)

Figure 1: /lo24 ŋin55 pi24 se11 ŋin55 to53/ 'the old men are more than the young children' (literally: 'old people compare young people more') (Hailu variety).

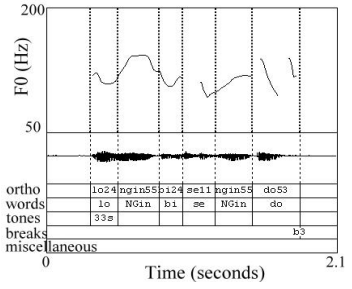
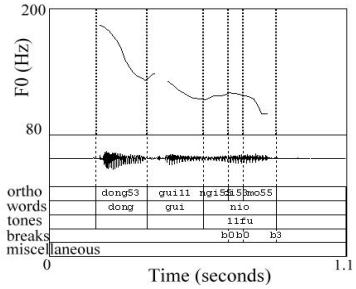


Figure 2: /toŋ53 kui11 ŋi11 ti53 mo55/ 'Do you know that it is very expensive?' (literally: 'very expensive you know interrogative.') (Hailu variety).



The breaks tier is for tagging boundaries including fused syllables, words, intermediate phrases, and intonation phrases. Unlike the other tiers, breaks tier is a point tier. The inventory of break index values is shown in Table 7. For example in Figure 2, the phrase /ŋi11 ti53 mo55/ is produced as a fused form [nio11]. The boundaries between the /ŋi11/ and /ti53/ and between /ti53/ and /mo55/ were marked as "b0" in breaks tier. In Figure 3, the boundaries between /kuet5/ and /ʒin53/, and between /kuet5/ and /lau53/ are marked as "b2" due to vowel lengthening without pauses. However, the final boundary of last /liuk3/ in Figure 3 and Figure 4 is marked as "b3" owing to a clear pause. The boundary between /kai55/ and /tet5/ is also marked as "b3" in Figure 4, because of the obvious vowel lengthening and pause.

Table 7: TWHK_ToBI break index values.

b0	fused syllable (Fig. 2)
(b1)	reserved for the ordinary word boundary, treated as a default value, and not tagged
b2	intermediate phrase boundary (Fig. 3)
b3	full intonation phrase boundary (Fig. 3)

Figure 3: /ŋai55 ʒiu53 hi11 ko11 kai55 tet5 kuet5 ʒin53 kuet5 lau53 tʰai11 liuk3/ 'I visited Germany, England, and China.' (literally: 'I past-tense go aspect the Germany, England, and China.') (Hailu variety).

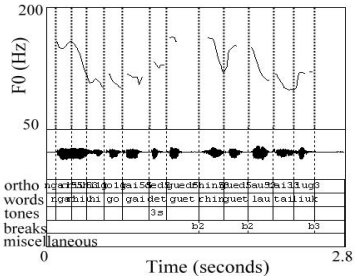
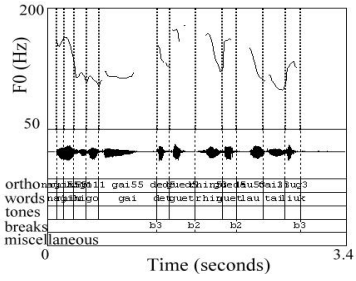


Figure 4: /ŋai55 ʒiu53 hi11 ko11 kai55 tet5 kuet5 ʒin53 kuet5 lau53 tʰai11 liuk3/ 'I visited Germany, England, and China.' (literally: 'I past-tense go aspect the Germany, England, and China.') (Hailu variety).



The miscellaneous tier is for analyzing spontaneous data. This tier is for transcribing laughter, coughs, etc. Code switching is also transcribed in the miscellaneous tier (Table 8).

Table 8: Symbols for code switching, laughter, and cough in TWHK_ToBI.

TMAN	switching to Taiwan Mandarin
THKS	switching to Sixian variety of Taiwan Hakka
THKH	switching to Hailu variety of Taiwan Hakka
TMIN	switching to Taiwan Min
JP	switching to Japanese
ENG	switching to English
C	cough
L	laughter

4. METHOD

4.1. Instrument

A KORG MR-1000 1-bit professional mobile recorder and BEHRINGER ECM8000 microphone were used to pick up all the elicited utterance in a sound treat room. Praat was used to transcribe the sound files.

4.2. Speaker

A 22-year-old male native speaker of Hailu variety Hakka uttered all the sound files.

4.3. Corpus

30 sentences were elicited, including expected and unexpected tone sandhi, fused syllables, intermediate phrases, and intonation phrases.

4.4. Transcribers

Three native Hailu Hakka speakers participated in the transcription. They were students at National Chiao Tung University at time of transcription.

4.4.1. Training process

Transcribers were trained via an hour instruction to be familiar with lexical tones, labels, sandhi rules and break indices of TWHK_ToBI. Three examples of transcriptions were used to train the transcribers.

5. RESULTS

5.1. Agreement of inter-transcribers

The agreement of inter-transcribers in ortho tier, containing 207 boundaries, was 99.03%. The agreement in the words tier, containing 204 boundaries was 99.02%. In the tones tier, the inter-transcriber agreement rate was 92.65%. The disagreement occurred mainly during sandhi tones, especially the unexpected sandhi tones. For the breaks tier was 93.00%. All the disagreements were on b2. There were nine b2s in the corpus, and the agreement of all the three transcribers was only 23.00%; however, the agreement rate for any of the two transcribers was 89.00%.

6. DISCUSSION

This paper inferred the disagreement on the break index b2 (intermediate phrase boundary) might result from insufficient professional training of transcribers or other cues not reported in the current study which may contribute to b2 perception. The agreement of inter-transcribers could be elevated through more training. Besides, the definition of b2, intermediate phrase, might be modified after analyzing more spontaneous data.

7. REFERENCES

[1] Gu, S.-S. 2010. *Keyu Nengli Renzheng Jiben Cihui Zhongji Zhonggaoji Ji Yuliao Xuancui Hailu Qiang [Hakka Basic Vocabulary and Phrases for Intermediate and High-intermediate Level Hakka Proficiency Test in Hailu Variety]*. Taipei, Taiwan. Council of Hakka Affairs, Executive Yuan, XIV-XIX.

[2] Gu, S.-S. 2010. *Keyu Nengli Renzheng Jiben Cihui Zhongji Zhonggaoji Ji Yuliao Xuancui Hailu Qiang [Hakka Basic Vocabulary and Phrases for Intermediate and High-intermediate Level Hakka Proficiency Test in Sixian Variety]*. Taipei, Taiwan. Council of Hakka Affairs, Executive Yuan, XIV-XIX.

[3] Gu. S.-S., Ho, S.-S., Liu, C.-X. 2004. *Keyu Fayin Xiue [Hakka phonics]*. Taipei, Taiwan: Wu-Nan Book Inc., 3-9, 66-72.

[4] Jiaoyubu Taiwan Kejiayu Changyongci Cidian Shiyongban [Taiwan Hakka dictionary of frequent terms]. *http://hakka.dict.edu.tw/hakkadict/index.htm*

[5] Peng, S.-H., Beckmen, M.E. 2003. Annotation conventions and corpus design in the investigation of spontaneous speech prosody in Taiwanese. *Proceedings of SSPR 2003*, 17-22.

[6] Yang, W.-S. 2007. *Quanguo Kejia Minzhun Keyu Shiyung Zhuangkuang [A Survey on National Hakka Languages]*. Taipei, Taiwan: Council of Hakka Affairs, Executive Yuan, 61-62.

[7] Yang, W.-S. 2008. *Quanguo Kejia Renkou Jichu Ziliao Diaocha Yanjiu [A Survey on National Hakka Population]*. Taipei, Taiwan. Council of Hakka Affairs, Executive Yuan, 5-24.