Computational Linguistics and Chinese Language Processing
Vol. 12, No. 4, December 2007, pp. 349-370

Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods

Un-Gian Iunn[*], Kiat-Gak Lau[+], Hong-Giau Tan-Tenn[#],

Sheng-An Lee[*], and Cheng-Yan Kao[*]

Abstract

A sizable corpus of Taiwanese text in Latin script has been accumulated over the past two hundred or so years. However, due to the special status of Taiwan, few people can read these materials at present. It is regrettable that the utilization of these plentiful materials is very low.

This paper addresses problems raised in the Taiwanese Southern-Min tone sandhi system by describing a set of computational rules to approximate this system, as well as the results obtained from its implementation. Using the romanized Taiwanese Southern-Min text as source, we take the sentence as the unit, translate every word into Chinese via an online Taiwanese-Chinese dictionary (OTCD), and obtain the part-of-speech (POS) information from the Chinese Electronic Dictionary (CED) made by the Chinese Knowledge and Information Processing (CKIP) group of Academia Sinica. By using the POS data and tone sandhi rules based on linguistics, we then tag each syllable with its post-sandhi tone marker. Finally, we implement a Taiwanese Southern-Min tone sandhi processing system which takes a romanized sentence as an input and then outputs the tone markers.

Our system achieves 97.39% and 88.98% accuracy rates with training and test data, respectively. Finally, we analyze the factors influencing error for the purpose of future improvement.

[*] Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {d93001, d93005, cykao}@csie.ntu.edu.tw
[+] Phahng Taiwanese Workshop, http://www.phahng.idv.tw
E-mail: kiatgak@gmail.com
[#] Independent scholar
E-mail: chenchen@umdnj.edu

[Received October 31, 2006; Revised July 6, 2007; Accepted July 11, 2007]

Keywords: Taiwanese Southern-Min, Written Taiwanese, Tone Sandhi System, Taiwanese Romanization

1. Introduction

1.1 Background and Motivation

Taiwanese is often used in daily life in Taiwan, but written Taiwanese is less common by far. Even so, the history of written Taiwanese stands at well over a century [Tiunn 2001]. At present, there are several dozen if not more than a hundred proposed phonetic and writing systems for Taiwanese [Iunn and Tiunn 1999]. The orthography adopted by this article is Peh-oe-ji (*POJ*, 白話字, also known as *Latinized Taiwanese* or *Missionary Romanization System for Taiwanese*).

Under the auspices of the National Museum of Taiwanese Literature, the Department of Taiwanese Literature of Cheng Kung University carried out a project titled “The Collection and Cataloging of Taiwanese Peh-oe-ji Literature Data” (CCTPLD). Although many texts have already been lost due to the alternation of political status, this project nevertheless revealed nearly 2,000 POJ books and periodicals, with publication sites spread over Taiwan, Xiamen (Amoy), Shanghai, Guangzhou (Canton), Hong Kong, Singapore, the Philippines, London, Japan, and beyond. The amount of publishing peaked in the 1950’s and 60’s [Iunn and Tan-Tenn, unpublished]. The scope covers both formally published books and periodicals as well as non-published items such as personal letters and medical charts. Later on, the government, citing supposedly detrimental effects of POJ on Mandarin promotion, banned its use and thus caused the rapid decline of this practice.

We hope that the extant materials collected by the above-mentioned CCTPLD project can be accessed by more people, as well as contribute to both basic and applied Taiwanese research. As most people nowadays are not familiar with Latinized Taiwanese, use of state-of-the-art text-to-speech technology would enhance the value of these materials to the general public.

Tone sandhi represents a challenging problem to be solved before one can successfully transform the written Taiwanese text to its natural speech-like tonal contour. This is because the written form of Latinized Taiwanese represents the tones as "basic tones", the tones of syllables when they are pronounced in isolation. At the level of the word, all syllables except the last one are usually pronounced differently (that is, they manifest tone sandhi). At the level of a whole sentence, in most situations only the last syllables next to the boundary of the phrases or structural markers are read as basic tones, the others being read as sandhi tones. In fact, besides the "regular tone sandhi" mentioned above, there are still several other kinds of tone sandhi phenomena which will be discussed in detail later.

We will first formulate the sandhi rules, which are the key to correct pronunciation and the core issue of this paper. The input of our experiment mainly consists of the data collected by the CCTPLD project; these data are processed by our sandhi system to produce sandhi-marked final outputs. Due to the lack of tagged data, we adopt the rule-based model, not the statistical model in this experiment. Figure 1 describes the skeleton of our system, and the webpage http://iug.csie.dahan.edu.tw/nmtl/dadwt/ demonstrates the results. Three of the authors who are native Taiwanese speakers evaluated the outputs for their accuracy.

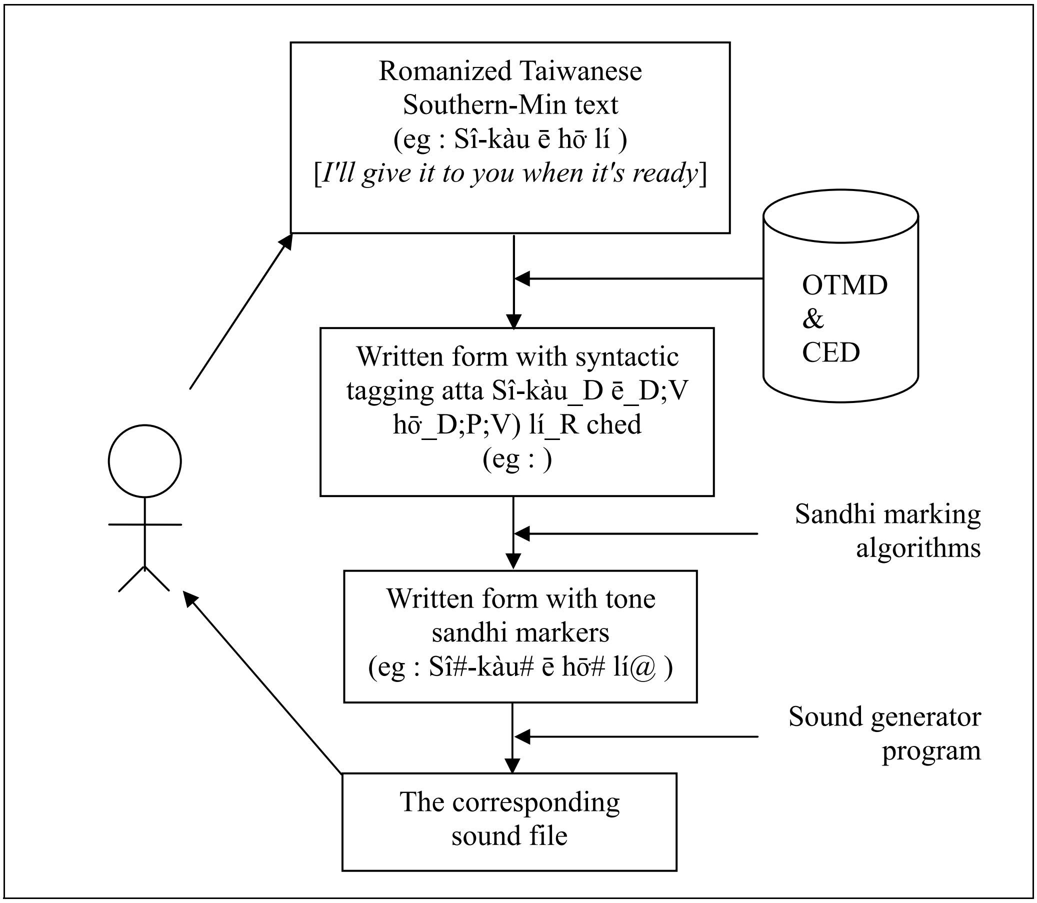
Romanized Taiwanese Southern-Min text
(eg : Sî-kàu ē hō͘ lí)
[I'll give it to you when it's ready]
OTMD
&
CED
Written form with syntactic tagging atta Sî-kàu_D ē_D;V hō͘_D;P;V) lí_R ched
(eg :)
Sandhi marking algorithms
Written form with tone sandhi markers
(eg : Sî#-kàu# ē hō͘# lí@)
Sound generator program
The corresponding sound file

Figure 1. Taiwanese Southern-Min tone sandhi system diagram

1.2 The Tone Sandhi Problem

Tones in Taiwanese are traditionally analyzed as consisting of *piâⁿ* [平], *siáng* [上], *khì* [去], *ji̍p* [入], each having *im* [陰, "*yin*"] and *iâng* [陽, "*yang*"] except *siáng*. So there are a total of seven tones. Following the sequence of *im-piâⁿ*[陰平], *siáng*[上], *im-khì*[陰去], *im-ji̍p*[陰入], *iâng-piâⁿ*[陽平], *iâng-khì*[陽去], *iâng-ji̍p*[陽入], they are numbered 1 (high flat), 2 (high to low), 3 (low), 4 (middle short), 5 (low rising), 7 (middle flat), and 8 (high short). The tone pitch is described within the parentheses. Please refer to the following examples for tone diacritics. In this paper, all examples are written in Taiwanese. For the sake of apprehensibility, we also add the Mandarin and English translations.

Tone sandhi is a very important characteristic of Taiwanese. At the word level, the last syllable is usually pronounced as basic tone and the others as sandhi tones. In example (1), the underlined syllables are pronounced as basic tones, the others as sandhi tones:

(1) tâi [台, “*platform*”]
Tâi-gí [台語, “*Taiwanese language*”]
Tâi-gí-bûn [台語文, “*written Taiwanese*”]
Tâi-gí bûn-ha̍k [台語文學, “*Taiwanese literature*”]
Tâi-gí bûn-ha̍k-sú [台語文學史, “*history of Taiwanese literature*”]

At the level of the syllable or the word, tone sandhi may manifest itself in at least the following several ways:

(a) Normal sandhi: using reduplicated syllables as examples (the numbers within parentheses are reading tones).

(2) (i) tone 1 → tone 7: “chheng-chheng” (7,1) [清清, “*clear*”]
(ii) tone 7 → tone 3: “chēng-chēng” (3,7) [靜靜, “*quiet*”]
(iii) tone 3 → tone 2: “chhiò-chhiò” (2,3) [笑笑, “*smiley*”]
(iv) tone 2 → tone 1: “léng-léng” (1,2) [冷冷, “*cold*”]
(v) tone 5 → tone 7 or 3 (northern Taiwan): “âng-âng” (7/3,5) [紅紅, “*red*”]
(vi) tone 4 → tone 8 (-p/t/k) or 2 (-h): like “sip-sip” (8,4); [溼溼, “*moist*”]
“khoeh-khoeh” (2,4) [擁擠, “*crowd*”]
(vii) tone 8 → tone 4 (-p/t/k) or 3 (-h): like “tit-ti̍t” (4,8) [直直, “*straight*”];
“jo̍ah-jo̍ah” (3,8) [熱熱, “*hot*”]

(b) Following sandhi: this pattern generally occurs with pronouns or the suffix of names. The tone pitch depends on that of the preceding syllable and is either tone 1 (high), 3 (middle), or 7 (low).

(3) (i) "A-eng--a" (7,1,1) [阿英, *a personal name*] (the second "a" is a suffix)

(ii) "góa lâi khòaⁿ -- i" (1,7/3,3,3) [我來看他, "*I come to see him/her*"]
(the basic tone of "i"[他, "*(s)he*"] is tone 1)

(iii) "hō͘ --lí" (7,7) [給你, "*give you*"] (the basic tone of "lí"[你, "*you*"] is tone 2)

(c) Neutral sandhi: the syllable immediately preceding the neutral sandhi (marked orthographically with double hyphens same as (b)) is read as basic tone, and the tones of the neutral sandhi are pronounced softly as if they were tone 3 or tone 4.

(4) (i) "Tân--sian-siⁿ" (5,3,3) [陳先生, "*Mr. Tân*"] (the original tones of
"sian-siⁿ "[先生, "*Mr.*"] are tone 7 and tone 1)

(ii) "kiâⁿ--chhut-lâi" (5,4,3) [走出來, "*walk out*"] (the original tones of
"chhut-lâi" [出來, "*out*"] are tone 8 and tone 5)

(d) Double sandhi: this pattern mostly appears in syllables ending in the glottal stop (-h) and having tone 4. The normal sandhi rules are applied twice in sequence (*i.e*. tone 4 → tone 2 → tone 1):

(5) (i) "beh thȧk-chu" (1,4,1) [要讀書, "*want to read books*"] ("beh" [要, "*want*"]
is tone 4, but rather than becoming tone 2, it becomes tone 1)

(ii) "khì gōa-kháu" (1,3,2) [去外面, "*go outside*"] ("khì"[去, "*go*"] is tone 3,
but rather than becoming tone 2, it becomes tone 1)

(e) Pre-*á* sandhi: the syllables before á do not follow normal sandhi rules unless they are tone 1 or 2.

(6) (i) tone 1 → tone 7: "sun-á" (7,2) [姪子, "*nephew*"]

(ii) tone 2 → tone 1: "chháu-á" (1,2) [小草, "*grass*"]

(iii) tone 3 → tone 1: "tàⁿ-á" (1,2) [攤位, "*stall*"]

(iv) tone 4 → tone 8 (-p/t/k) or tone 1 (-h): "tek-á" (8,2) [竹子, "*bamboo*"]
"thih-á" (1,2) [鐵,"*iron*"]

(v) tone 5 → tone 7: "lô͘-á " (7,2) [爐子, "*oven*"]

(vi) tone 7 does not change: "phō-á" (7,2) [簿子, "*tablet*"]

(vii) tone 8 → tone 4 (-p/t/k) or tone 7 (-h): "chhȧt-á" (4,2) [賊, "*thief*"]
"hiȯh-á " (7,2) [葉, "*leaf*"]

(f) Triplicate sandhi: the first syllable of triplicated words does not follow normal sandhi rules unless it is of tone 2, 3, or 4:

(7) (i) tone 1 → tone 5: like “chheng-chheng-chheng” (5,7,1) [清清清, “*very clear*”]
(ii) tone 2 → tone 1: like “ún-ún-ún” (1,1,2) [穩穩穩, “*very stable*”]
(iii) tone 3 → tone 2: like “hèng-hèng-hèng” (2,2,3) [興興興, “*very interesting*”]
(iv) tone 4 → tone 8 (-p/t/k) or tone 2 (-h): like “sip-sip-sip” (8,8,4)
[溼溼溼, “*very humid*”] “bah-bah-bah” (2,2,4) [肉肉肉, “*very fat*”]
(v) tone 5 → (similar to) tone 5: like “kôaⁿ-kôaⁿ-kôaⁿ” (5,7/3,5)
[冷冷冷, “*very cold*”]
(vi) tone 7 → (similar to) tone 5: like “chēng-chēng-chēng” (5,3,7)
[靜靜靜, “*very quiet*”]
(vii) tone 8 → (similar to) tone 5: like “ti̍t-ti̍t-ti̍t” (5,4,8) [直直直, “*very straight*”]
“pe̍h-pe̍h-pe̍h” (5,3,8) [白白白, “*very white*”]

(g) Rising sandhi: this pattern usually occurs on loanwords from Japanese; the sandhi tone is similar to tone 5.

(8) “ŏai-siak-chù” (5,8,3) [白襯衫, “*white shirt*”]
“khăn-páng” (5,2) [看板, “*signboard*”]
“hăn-tó͘-lù ”(5,1,3) [方向盤, “*steering wheel*”]

We collate the above sandhi phenomena in *Table 1*.

Table 1. Taiwanese Southern-Min tone sandhi phenomena

Normal sandhi	Basic tone of syllable	1	2	3	4	5	7	8
	Sandhi tone	7	1	2	8/2	7/3	3	4/3
Following sandhi	Basic tone of preceding syllable	1	2	3	4	5	7	8
	Sandhi tone	1	3	3	3	7	7	1
Neutral sandhi	Basic tone of preceding syllable	1	2	3	4	5	7	8
	Sandhi tone	3	3	3	4/3	3	3	4/3
Double sandhi	Basic tone of syllable	-	-	3	4	-	-	-
	Sandhi tone			1	1			
Pre-*á* sandhi	Basic tone of preceding syllable of “*á*”	1	2	3	4	5	7	8
	Sandhi tone	7	1	1	8/1	7	7	4/7
Triplicate sandhi	Basic tone of the first syllable of three	1	2	3	4	5	7	8
	Sandhi tone	5	1	2	8/2	5	5	5

1.3 Historial Review

[Lin and Chen 1999] describes an early sandhi system. Users input Chinese texts, and the system outputs Taiwanese texts with pronunciation. The corpus is news reports in Chinese. They used the word segmentation and tagged data from the CKIP group and the Taiwanese-Chinese dictionary from Robert Cheng to map the Chinese news into Taiwanese (in both Han and Latin scripts). The sandhi rules applied were as follows: a) pronounce the last syllable at the end of a sentence as basic tone; b) pronounce the syllable before the particle *ê* as basic tone; c) pronounce the last syllable of a noun as basic tone; d) pronounce other syllables as normal sandhi tones. An accuracy rate of 82.53% was reported. However, as the system did not take Taiwanese as input, word order and semantic ambiguities were not taken into account when converting, the translation was not quite native-like.

[Liang *et al.* 2004] is a recent text-to-speech system for Taiwanese Southern-Min. Its input was a large corpus of Chinese news texts, but sentences longer than 20 syllables were removed. It utilized a dictionary to convert the Chinese text into Taiwanese Southern-Min, followed by word segmentation, phonetic marking, and rule-based sandhi processing to generate speech files. Due to the size of the corpus, only the first 200 sentences generated were evaluated by two Taiwanese-speaking experts. The accuracy rates were 97% for word segmentation, 89% for pronunciation marking, and 65% for rule-based sandhi processing.

Compared with the above systems, our approach has some major differences: a balanced Taiwanese corpus for both literary and non-literary sources (about 50% each) was prepared; no translation from Chinese to Taiwanese; and no limits for length of sentences. In addition, because the text is written in Latinized script, we do not need to manipulate word segmentation and phonetic marking. However, compared to text with Han character script, there is a more rigorous challenge to deal with homonymy, especially with monosyllabic words.

2. Method

2.1 Data

The input data of our system are from the CCTPLD project. Following POJ orthography, syllables of a word are joined by hyphens, and the words are separated with spaces.

We select parts of four sources as training data. The training data sources are shown in *Table 2*.

Table 2. Training data sources

Book or aticle	year	author	genre
“Sin-bûn ê chȧp-liȯk” [新聞的雜錄, “*News Bulletin*”]	1913	unknown	journalism
“Chȧp-hāng kóan-kiàn” [十項管見, “*Ten Humble Opinions*”]	1925	Chhòa Pôe-hóe [蔡培火]	discourse
“Chháu-tui téng ê bîn-bāng -- jî-tông chong-kàu kò͘-sū” [草堆上的夢─兒童宗教故事, “*Dreams on the Grass Stack -- Religious Stories for Children*”]	1955	N̂g Hôai-un [黃懷恩]	short stories
“Tang-pō͘ thôan-tō kiàn-bûn kì” [東部傳道見聞記, “*Record of Preaching in Eastern Taiwan*”]	1961	Tân Kàng-hâng [陳降祥]	journalism

The published dates of the above sources range from Japan-ruled era (1895-1945) to postwar era (1945-). Two paragraphs are selected from each book, with a total of 614 syllables (438 word tokens).

In addition to data drawn from the same project, the test data also include some other sources we collected. Four sources are selected as well. The test data sources are shown in *Table 3*.

Table 3. Test data sources

Book or article	year	author	genre
“Pe̍h-ōe-jī ê lī-ek” [白話字的利益, “*The Benefits of Using Peh-oe-ji*”]	1885	Reverend Ia̍p [葉牧師]	discourse
“Kau-chiàn ê Siau-sit” [交戰的消息, “*News of the War*”]	1905	the editorial office of *Tâi-lâm Prefectural Church News*	report
“Thiàⁿ lí iâⁿ kè thong sè-kan” [疼愛你勝過全世界, “*Caring About You More Than the Whole World*”]	1955	Lōa Jîn-seng [賴仁聲]	novel
“Ài lí kap ài i pîⁿ-á chōe” [愛妳和她一樣多, “*Loving You as Much as Her*”]	1997	Lô͘ Tàn-chhun [盧誕春]	prose

Two or three paragraphs are selected from each book or article. The test data total 962 syllables (656 word tokens) and also cover two eras but with a longer time span.

2.2 Part of Speech Tagging

As there is no standard on part of speech (POS) for Taiwanese at present, we use the standard of Chinese instead (see Results section). We obtain the corresponding Chinese translation for each Taiwanese word by looking up the Taiwanese-Chinese On-line Dictionary. [Iunn 2003] We, then, look up the POS of the Chinese in the 80,000-word CED. Ambiguity encountered includes:

(a) homonymy, especially monosyllabic homonyms;

(b) one-to-many mapping when mapping Taiwanese to Chinese;

(c) multiple possible POSs for each Chinese word.

To resolve homonymy, we choose the word with the highest querying frequency. We found out that this strategy works under most situations. Due to the fact that one Taiwanese word may map to multiple Chinese words, and one Chinese word could possibly have multiple POSs, there may be multiple POSs for one Taiwanese word. We initially retain all candidate POSs in tagging and only attempt to narrow down the list upon applying the sandhi algorithm. Of the 46 POSs in the Chinese Electronic Dictionary, we adopt the top level and adjust certain POSs known to affect tone sandhi. For example, Vh (state intransitive verb, etc.) is marked A, Nh (pronoun) marked R, Ng (postposition) marked G, and Nd (time) marked S. The POS classes we used are shown in *Table 4*.

Table 4. POS classes

POS	statement	POS	statement	POS	statement
A	adjective	I	interjection	R	pronoun
C	conjunction	M	special marker	S	time
D	adverb	N	noun	T	auxiliary
G	postposition	P	preposition	V	verb

As for unknown words, if they are of the form 'XX' or 'XXX' (duplicate or triplicate syllables), we mark them as A (adjective). Other words are marked as N (noun).

2.3 Tone Sandhi Marks

The marks representing tone sandhis are listed in *Table 5*. Words with normal sandhi are usually not marked.

Table 5. Sandhi marks

Symbol	Phenomenon	Symbol	Phenomenon
(none)	Normal sandhi	**$**	Double sandhi
#	Basic tone	**&**	Pre-*á* sandhi
@	Following sandhi	~	Triplicate sandhi
%	Neutral sandhi	^	Rising sandhi

2.4 Tone Sandhi Rules

Tone sandhi rules are the most important part of this study. The algorithm for sandhi marking is shown in *Table 6*.

Table 6. Tone sandhi marking algorithm

Rule		Remark
1	Apply normal sandhi to all syllables	
2	Mark the last syllable as basic tone #	
3	*ê* [的, "*of*"] : Mark the syllable preceding *ê* as basic tone #	*ê* is a special marker
4	A/A Pair 4.1 A/A Pair: Mark the last syllable of the first word as basic tone #	POS level, with ambiguity
5	N/V, N/A, N/P, N/R, and N/D Pairs 5.1 N/V Pair: Mark the last syllable of the first word as basic tone # 5.2 N/A Pair: Mark the last syllable of the first word as basic tone # 5.3 N/P Pair: Mark the last syllable of the first word as basic tone # 5.4 N/R Pair: Mark the last syllable of the first word as basic tone # 5.5 N/D Pair: Mark the last syllable of the first word as basic tone #	POS level , with ambiguity
6	C: Mark the last syllable of the preceding word as basic tone #	POS level
7	G: Mark the last syllables of both the preceding word and the word itself as basic tones #'s	POS level, without ambiguity
8	S: Mark the last syllable of this word as basic tone #	
9	POS R 9.1 *i* / *in* [他(們), "*(s)he/they*"] : Mark them as normal sandhi even if they are the last syllables 9.2 *góa* / *lí* / *gún* / *góan* / *lán* / *lín* [我/你(們)(的), "*I/you/my/our/your*"]of POS R: Mark them as normal sandhi if they are not the last syllable	POS/Word level
10	Sentence-final *kóng* [講, "*say*"] : Mark this word as normal sandhi if the delimiter is among [,：: "] and there is any word of POS R in front of this word (note: this rule needs to be refined in case there is a name in front of this word)	Word level, induced from training data

Table 6. Tone sandhi marking algorithm

11	pre-*á* [*á is suffixof a word*]: Mark any syllables just before *á* as pre-*á* sandhi &	Syllable level
12	Double sandhi	
	12.1 *beh* [要, "*want*"] : Mark any *beh* as double sandhi $ unless it appears at the end, including those within a word, such as *kion̂g-beh, tih-beh* .	Syllable level
	12.2 *khì* [去, "*go*"] : Mark *khì* as double sandhi $ if the POS of the immediately following word is N or V, unless it appears at the end	Word level
	12.3 *koh* [再, "*again*"] : Mark any *koh* as double sandhi $, including those within a word, such as *chiah-koh*[再, "*and then*"] or *iáu-koh*[還是, "*still*"], unless it appears at the end	Syllable level, extended from training data
	12.4 *kah*[和, "*and*"] : Mark any *kah* as double sandhi $ unless it appears at the end	Word level
13	Neutral sandhi of --: Mark the syllable just before -- as basic tone, and mark each syllable after -- as neutral sandhi %	Word level
14	Triplicate sandhi: Mark the first syllable as triplicate sandhi if that word has 3 syllables of the same spelling	Word level
15	Special words	Word level, extend from training data beacuse of not yet standardized
	15.1 *sím-mih* / *sím-mi̍h*[什麼, "*what*"] : Change these words into sím-mí (sandhi marks not changed)	
	15.2 *án-ni* / *àn-ni* / *an-ni* / *an-nī* [這樣, "*thus*"] : Change these words into *án-ni* and to mark its sandhi marks as t#	
16	Markers	
	16.1 *iah-sī* / *ah-sī* / *iáh-sī* / *áh-sī* / *á-sī* [或是, "*or*"] : Mark the last syllable before these words as basic tone #	word level, extended from training data
	16.2 V *sī* [是, "*is*"] V: Mark the last syllable of the verb that just before *sī* as basic tone # if this verb appears again after *sī*	Sentence pattern level, induced from training data
	16.3 *che* / *he* / *chia* / *hia* [這/那(裡), "*this/that/(t)here*"] : Mark these words as basic tone #	word level
	16.4 *ū-sî* [有時, "*sometimes*"] / *put-sî* [不時, "*from time to time*"] / *kui-khì* [乾脆, "*just*"] / *óan-jiân* [宛然, "*like*"] / *gôan-lâi* [原來, "*originally*"] *chiong-lâi* [將來, "*future*"] / *chiông-lâi* [從來, "*always*"] / *sui-jiân*/ *sui-bóng* [雖然, "*though*"] / *sî-siông* [時常, "*often*"] / *hui-siông* [非常, "*very*"] / *si̍t-chāi* [實在, "*really*"] /	word level, extended from training data

Table 6. Tone sandhi marking algorithm

	sî-chūn [時候, "*(the duration of) time*"] : Mark the last syllables of these words as basic tone #	
	16.5 *chiū* / *tō* [就, "*as soon as*"] : Mark the syllable of the word just before as basic tone # if the POS of the word is A	word level, induced / extended from training data
	16.6 *sî-kàu* [到時候, "*at that time*"]: Mark both of the two syllables of this word as basic tones	word level, induced from training data
17	T: Mark the last syllable of a word as basic tone if the word is just before a word of POS T in the end	POS level
18	Other sandhi: 18.1 *teh* [在, "*at*"]: Mark *teh* or the *teh* in *tī-teh* as other sandhi ^	word level, our observations
19	Neutral sandhi	
	19.1 *chhut-lâi* [出來, "*come out*"] / *chhut-khì* [出去, "*go out*"] / *lo̍h-lâi* [下來, "*come down*"] *lo̍h-khì* [下去, "*go down*"] *kòe-lâi* [過來, "*come up*"] *kòe-khì* [過去, "*pass away*"]: Mark the last syllable of a verb just before these words as basic tone #, and mark these words as neutral sandhi %	word level
	19.2 *sian-si^n/sin-se^n/sian-se^n* [先生, "*Mr.*"]: Mark the word before these words as basic tone # and these words as neutral sandhi %, if the first letter of the preceding word is uppercase	word level
	19.3 *bô* [無, "*have nothing*"] at the end 19.3.1 *á* / *á-sī* / *iah* / *iah-sī* / *ah* /*ah-sī* [或是, "*or*"]: if the preceding word is among these words, do nothing 19.3.2 Otherwise: Mark the last syllable of the word just before *bô* as basic tone #, and mark *bô* as neutral sandhi %	word level, indecud / extended from training data
	19.4 *bē*/*bōe* [不會, "*will not*"] at the end 19.4.1 *ē*/*ōe* [會, "*be good at*"] *ē-hiáu*/*ōe-hiáu* [會, "*be good at*"]: Mark any final *bē*/*bōe* as neutral sandhi % 19.4.2 *á* / *á-sī* / *iah* / *iah-sī* / *ah* /*ah-sī* [或是, "*or*"]: Mark the *bē*/*bōe* as basic tone # if any of these words immediately precedes it 19.4.3 Otherwise: Do nothing as it could be ambiguous (*e.g.* *bē*/*bōe* [賣, "*sell*"])	word level, indecud / extended from training data
20	R at the end *góa* / *lí* / *i* / *gún* / *góan* / *lán* / *lín* / *in* [我/你/他(們)(的), "*I/we/my/our/you(r)(s)/(s)he/they/their*"]: Mark the pronoun as following sandhi @ if it appears at the end and there is a verb before it	word level

These sandhi rules work on 4 different levels: the syllable, the word, the part of speech, and the sentence pattern.

The algorithm described above is mainly based on a) tone sandhi rules proposed by linguists; b) rules induced from the training data; and c) our intuition as native-speaking observers of sandhi phenomena. We also consulted d) the word segmentation results of the CKIP (examining its POS tagging output) and e) the Taiwanese concordancer system (to check the sandhi phenomena of certain words) when we met some questions.

It should be noted that some of the sandhi rules proposed by linguists deal with specific contexts and thus cannot be broadly applied; some others carry exceptions. There is, therefore, some difficulty in converting these rules into an algorithm. So, besides (a), we also formulated some rules from (b) and (c) by analyzing errors in the training data output. In principle sandhi rules are formulated to be applicable to "most situations" -- *i.e.* an accuracy rate of over 75% on corpus data. Once applied, the new rules may affect the original rules, so (d) and (e) are our important references in deciding whether or not to apply the new rules.

Some rules have priority. Subsequent rules can supersede previous ones. As an example, rule 9 (pronoun rule) can supersede rule 3 (*of* rule). At the level of sentence pattern, rule 19.4.2 can supersede 19.4.1 as in the following example:

(9) "*Lí ē khì kok-gōa bē*" [你會不會去國外? "*Will you go abroad or not*"]:
the last *bē* [不會, "*will not*"] is marked as neutral sandhi, whereas
"*Lí ē khì kok-gōa iah-sī bē*" [你會不會去國外? "*Will you go abroad or not*"]:
the last *bē* is marked as basic tone.

Moreover, because of the uncertainty in tagging POS, some rules are set to apply only when there is no ambiguity, while some other rules are applied to any matching POSs.

We currently employ 20 rules and expect to refine them or append new ones.

The following training data represents a pre-tagged source (Chinese and English translations added):

(10) Chhin-chhiūⁿ[像] án-ni[這樣] lâi[來] kóng[說], chāi[在] lán[我們] Tâi-ôan[台灣] kīn-kīn[近近] chi̍t-tiap-á-kú[一下子] ê[的] kang-hu[工夫], ài[要] soaⁿ[山] chiū[就] ū[有] soaⁿ[山], ài[要] hái[海] chiū[就] ū[有] hái[海], beh[要] jo̍ah[熱] chiū[就] ū[有] jo̍ah[熱], kôaⁿ[冷] chiū[就] ū[有] kôaⁿ[冷]. Só͘-í[所以] thang[可以] kóng[說] Tâi-ôan[台灣] sī[是] chi̍t-ê[一個] sió[小] Tang-iûⁿ[東洋]. Lán[我們] Tâi-ôan[台灣] ū[有] chit-khóan[這種] thian-jiân[天然] ê[的] hó-kéng[好景], hó[好] khì-hāu[氣候], chiong-lâi[將來] nā-sī[若是] ēng-sim[用心] ke[加] lâng[人] ê[的] kang-hu[工夫] tōa-tōa[大大] lâi[來] chéng-tùn[整頓], tek-khak[的確] ē[會] chiâⁿ-chò[成為] Tang-iûⁿ[東洋] ê[的] tōa[大] kong-hn̂g [公園], hō͘ [讓] Tang-iûⁿ[東洋] ê[的] lâng[人] chi̍p-óa[靠近] lâi[來] hióng-hok[享福] an- lo̍k [安樂].

--- "*Cha̍p-hāng kóan-kiàn*" [十項管見]
by Chhòa Pôe-hóe[蔡培火], 1925

Take this as an example. Here in Taiwan, reachable with a minimum of effort, you have mountains for those who like mountains, seas for those who like seas, hot weather for those who like heat, and cold weather for those who like cold. So you can say Taiwan is a miniature East. Given Taiwan's natural sceneries and fair climate, if you'd take care to rebuild it, it'd surely become the Great Park of the East, where Easterners go for rest or fun.

---"*Ten Humble Opinions*"
by Chhòa Pôe-hóe, 1925

After POS tagging and applying the sandhi rules:

(11) Chhin -chhiūⁿ(D) án-ni#(D;N) lâi(D;V) kóng#(V), chāi(D;A;P;V) lán(R) Tâi-ôan#(N) kīn-kīn(A) chi̍t-tiap&-á-kú#(N) ê(M) kang-hu#(A;N), ài(D;V) soaⁿ# (N) chiū(D) ū(D;P;V) soaⁿ#(N), ài(D;V) hái#(N) chiū(D) ū(D;P;V) hái#(N), beh$(D) jo̍ah#(A) chiū(D) ū(D;P;V) jo̍ah#(A), kôaⁿ#(A) chiū(D) ū(D;P;V) kôaⁿ#(A). Só͘-í(C) thang(D) kóng(V) Tâi-ôan#(N) sī(D;V) chi̍t-$\boxed{\text{ê\#}}$(N) sió(D;A) Tang-iûⁿ#(N). Lán(R) Tâi-ôan#(N) ū(D;P;V) chit-khóan#(D;N) thian-jiân#(A) ê(M) hó-kéng#(N), hó(D;A;C;V) khì-hāu#(N), chiong-lâi#(S) nā-sī(C) ēng-sim#(N) ke(V) lâng#(N) ê(M) kang-hu#(A;N) tōa-tōa(A) lâi(D;V) chéng-tùn#(V), tek-khak(D) ē(D;V) chiâⁿ-chò(V) Tang-iûⁿ#(N) ê(M) tōa(A;N) kong-hn̂g#(N), hō͘(D;P;V) Tang-iûⁿ#(N) ê(M) lâng#(N) chip-óa(V) lâi(D;V) hióng-hok#(A) an-lo̍k#(A).

The letters within the parentheses are the POSs. Incorrectly processed syllables are boxed.

3. Results

3.1 Evaluation

Three authors of this paper, who are skilled native speakers familiar with written Taiwanese, evaluated the correctness of the output. Note that in certain contexts more than one sandhi result is acceptable, and depending on discourse considerations some speakers may opt for one sandhi result over others. For example, “hō lí” [給你,“*give you*”] can be read as (3,2) (normal sandhi) or (7,7) (following sandhi). Telephone number is another example: the number may be divided into various groups, each group containing 2, 3 or 4 digits.

3.2 Preliminary Results

There are 614 syllables of training data, 16 errors, giving an accuracy rate of 97.39%. There are 962 syllables of test data with 106 errors, or an accuracy rate of 88.98%. *Table 7* shows the number of errors and accuracy rate for each paragraph.

Table 7. Number of errors and accuracy rate for each paragraph

training data					test data				
para. id.	no. of words	no. of syllables	no. of errors	accuracy rate	para. id.	no. of words	no. of syllables	no. of errors	accuracy rate
1	27	30	1	96.67%	1	130	184	16	91.30%
2	42	54	0	100.00%	2	56	85	12	85.88%
3	44	70	0	100.00%	3	53	84	13	84.52%
4	33	52	0	100.00%	4	96	143	16	88.81%
5	38	51	4	92.16%	5	66	97	10	89.69%
6	85	110	4	96.36%	6	63	86	9	89.53%
7	97	144	6	95.83%	7	32	43	3	93.02%
8	72	103	1	99.03%	8	38	58	2	96.55%
					9	122	182	25	86.26%
total	438	614	16	97.39%	total	656	962	106	88.98%

Table 8 shows the numbers of each rule applied in training data and test data respectively. We count the number of affected syllables, accurately affected syllables, and accuracy rate of each rule. Note that rules 5 & 6 don't seem work well because of POS ambiguities, rule 7 does not affect any syllables because the word whose POS is G (postposition) also has other POSs, rule 14 does not affect any syllables because there are no triplicated words in our training and

test data.

Table 8. Affected and accurately affected syllables of each rule

	training Data			test data		
rule id.	affected syllables	accurately affected	accuracy rate	affected syllables	accurately affected	accuracy rate
1	614	411	66.94%	962	662	68.81%
2	74	68	91.89%	112	105	93.75%
3	32	24	75.00%	38	26	68.42%
4	3	3	100.00%	13	7	53.85%
5	65	57	87.69%	129	90	69.77%
6	4	3	75.00%	4	3	75.00%
7	0	0	--	0	0	--
8	5	5	100.00%	3	3	100.00%
9	29	29	100.00%	25	25	100.00%
10	5	5	100.00%	0	0	--
11	3	3	100.00%	8	8	100.00%
12	8	8	100.00%	11	11	100.00%
13	2	2	100.00%	6	5	83.33%
14	0	0	--	0	0	--
15	8	8	100.00%	6	5	83.33%
16	13	13	100.00%	4	4	100.00%
17	3	3	100.00%	2	2	100.00%
18	9	9	100.00%	3	3	100.00%
19	0	0	--	6	6	100.00%
20	2	2	100.00%	0	0	--

Every syllable is affected by at least one rule, and is affected by four rules at most. *Table 9* shows the number of dominant rule, accurate dominant rule, and accuracy rate.

Table 9. Number of dominant rule, accurate dominate rule and accuracy rate

	training data			test data		
rule id.	no. of dominant rule	no. of accurate dominant rule	accuracy rate	no. of dominant rule	no. of accurate dominant rule	accuracy rate
1	381	371	97.38%	616	568	92.21%
2	62	62	100.00%	104	99	95.19%

Table 9. Number of dominant rule, accurate dominate rule and accuracy rate

3	24	23	95.83%	34	26	76.47%
4	2	2	100.00%	6	3	50.00%
5	62	57	91.94%	126	87	69.05%
6	2	2	100.00%	4	3	75.00%
7	0	0	--	0	0	--
8	4	4	100.00%	3	3	100.00%
9	27	27	100.00%	25	25	100.00%
10	5	5	100.00%	0	0	--
11	3	3	100.00%	8	8	100.00%
12	7	7	100.00%	11	11	100.00%
13	2	2	100.00%	6	5	83.33%
14	0	0	--	0	0	--
15	6	6	100.00%	5	4	80.00%
16	13	13	100.00%	3	3	100.00%
17	3	3	100.00%	2	2	100.00%
18	9	9	100.00%	3	3	100.00%
19	0	0	--	6	6	100.00%
20	2	2	100.00%	0	0	--

After examination, we find that we can add 7 additional rules without too much effort; in this way, we were able to fix 20 errors and achieve a 91.06% accuracy rate. *Table 10* shows the additional rules in order to fix 20 errors in test data.

Table 10. Additional rules to obtain higher accuracy rate

Rules	Number of corrections in test data
Word suffix “V-tit” (adverbialize the word whose POS is verb)	5
Double sandhi of “khah” [更, “*more*”]	4
Re-process the syllable preceding “ê” [個, *a numerary adjunct*] when the preceding word is a number or “chit/hit/pa̍t” [這/那/別, “*this/that/other*”]	4
“V-ji̍p-lâi” [V 進來, “*V-in*”]: mark as neutral sandhi when sentence-final	3
Word “hut-jiân” [忽然, “*suddenly*”]: mark the last syllable as basic tone in any case	2
Word “kīn-lâi” [近來, “*recently*”]: mark the last syllable as basic tone in any case	1
Word suffix “N-nih”[N 裡, “*inside N*”]: mark as neutral sandhi	1

4. Analysis of Errors and Relevant Issues

Some of the problems we encountered may be taken into account in the future.

4.1 POS

In our investigation, we use the POS set for Chinese. Whether this approach is suitable for Taiwanese is a debatable linguistic question requiring further investigation. Although a few studies of the POS of Taiwanese are available from as early as the 1930s, currently these data have yet to be digitized, and will need to be reviewed by linguists to ensure that they are suitable for dealing with the sandhi problem.

4.2 Word Segmentation Standard and Dictionary

[Tseng 1997] proposes a standard for Taiwanese word segmentation. Unfortunately discussion is lagging. Should a working word segmentation standard emerge, we would also need a dictionary conforming to that standard.

4.3 Standardization of Written Taiwanese

Historically, the use of Han script to represent Taiwanese has suffered from a high degree of idiosyncrasy in character choice. For documents written in Latin script, most of the differences attributed to dialects can be reconciled by referencing existing dictionaries. Orthographic inconsistency in the use of hyphen is more problematic, as it could affect the result of sandhi processing. Manual standardization of hyphen placement is hardly a solution.

4.4 Tone Sandhi Problems Not Solvable by POS Order

We have encountered certain sandhi problems that likely cannot be solved solely by inspecting the POS order. These include verb-verb (VV) and noun-noun (NN) patterns:

(12) a. “phah-pià^{n}(V) chò(V) khang-khòe(khè) (N)” (2,2,2,7,3)
[努力做工作, “*do work hard*”]
b. “kia̍h-ba̍k(V) khòa^{n}(V) hn̂g(N)” (3,8,2,5)
[舉目看園, “*lift eyes and see plowland*”]

(12) is an example of a VV pattern. The final syllable of the first verb in (a) should be marked as sandhi tone, while in (b) it should be marked as basic tone. Differences in the internal structure of these two initial verbs suggest some clues for handling this problem. However, its implementation awaits further research.

(13) a. "tiān-sī kóng-kò" [電視(的)廣告, "*TV advertisement*"]
b. "thâng-thōa chiáu-chiah" [昆蟲(、)小鳥, "*insects and birds*"]

(13) is an example of a NN pattern. Again, the final syllable of the first noun in (a) should be marked as sandhi tone, while in (b) it should be marked as basic tone. Currently, we see no solution to this.

4.5 Error Conditions

Error conditions, including those discussed in the previous sections, are listed below with possible in *Table 11* :

Table 11. Error conditions and possible solutions

Errors	Possible Solutions
(a) Due to dictionary limitation (not having the words)	Increase entries
(b) Due to lack of punctuation marks	Pre-process, but this is very difficult
(c) Due to wrong POS because of homonymy	Apply semantic knowledge
(d) Due to indeterminate POS or multiple candidates	Tagging disambiguity
(e) Caused by inconsistent orthography in hyphen segmentation	Pre-process the sources or deal with the procedures of adding or removing hyphens automatically
(f) Due to incomplete sandhi rule set	Refine the sandhi rules while avoiding side effects
(g) Associated with quantitative words;	Add DM rules
(h) Associated with proper nouns	Detect proper nouns
(i) Associated with sentence pattern	Add sandhi rules for sentence patterns
(j) Possibly other sources of error yet to be identified	

5. Future Work

A three-year-old child native speaker can process tone sandhi correctly and apparently without effort, yet it is rather more difficult for a computer system to do so. Clearly, a practical system for sandhi processing of Taiwanese remains out-of-reach and a cause for future research. Some suggestions for future work:

(a) Solicit assistance from linguists. It is hoped that linguistics will define a standard for part-of-speech analysis and word segmentation, and that a dictionary conforming to such a standard will be built.

(b) Improve word segmentation, especially the processing of morphology, quantitative words, and proper nouns.

(c) Improve the processing of POS tags to account for ambiguity.

(d) Improve the dictionary's part-of-speech data, such as making use of Embree's POS analysis [Embree 1984].

(e) Improve the sandhi rules.

(f) Find alternative ways of modeling sandhi processing, such as Cheng's grammar template model. [Cheng 2002]

Acknowledgements

This work is supported by National Museum of Taiwanese Literature in Taiwan. We also thank the anonymous reviewers for their constructive opinions.

References

Cheng, R., *Taiwanese and Mandarin Structures and Their Developmental Trends in Taiwan Book I : Taiwanese Phonology and Morphology*, Yuan-liou Publishing Co., 1997.

Cheng, R., "Tone Sandhi on the Grammar Template--Cognition and Testing," *Proceeding of 2002 International Conference on Teaching and Researching of Taiwanese Romanization*, 2002, pp. I1-I19.

Embree, B. L.M.A., *A Dictionary of Southern Min*. Taipei Language Institute, 1984.

Iunn, U.-G., "Taiwanese-Chinese On-line Dictionary -- Discussion of Building Technique and its Utilization," *Proceeding of 3rd International Conference on Internet Chinese Education*, 2003, pp. 132-141.

Iunn, U.-G. and H.-K. Tiunn, "Review and Analysis of Taiwan Ho-lo Language non-Han Character Spelling Symbols," *Proceedings of 1st Conference on the Regeneration and Rebuild of Taiwan Mother Tongue Culture*, 1999, pp. 62-76.

Iunn, U.-G. and H. H.Tan-Tenn, "A Survey of Media and Data Processing Development for Written Taiwanese," Accepted by *International Journal of the Sociology of Language, Special Issues on Taiwanese*.

Liang, M.-S., J.-C. Yang, Y.-C. Chiang, and R.-Y. Lyu, "A Taiwanese Text-to-Speech System with Applications to Language Learning," *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 91-95.

Lin, C.-J. and H.-H. Chen, "A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan," *International Journal of Computational Linguistics and Chinese Language Processing*, 4(1), 1999, pp. 59-84.

Lu, G.-C., *The Study of Minnan Vocabulary in Taiwan*. SMC Publishing Inc., 1999.

Tiunn, J.-H. (Chang, Y.-H.). *Principles of POJ or the Taiwanese Orthography: An Introduction to Its Sound-Symbol Correspondences and Related Issues*, Crane Publishing Co., 2001.

Tseng, C.-C., "The Discussion of Taiwanese Word Segmentation Principles," *The Project Report for the Collecting, Cataloging and Select Editing of Taiwanese Liturature Publications*, pp. 47-73, Council for Culture Affairs, 1997.

Online Resources

Chinese On-line Word Segmentation System, http://ckipsvr.iis.sinica.edu.tw

Digital Archive Database for Written Taiwanese, http://iug.csie.dahan.edu.tw/nmtl/dadwt

Taiwanese Concordancer System, http://iug.csie.dahan.edu.tw/TG/concordance/form.asp

Taiwanese Package, http://www.phahng.idv.tw or http://taigu.fhl.net/TP/