# A Study on Implementation of Southern-Min Taiwanese Tone Sandhi System

**Iu$^{n}$ Un-gian**
Dept. of Computer Science and Info. Eng., National Taiwan Univ.
No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
d93001@csie.ntu.edu.tw

**Lau Kiat-gak**
Dept. of Computer Science and Info. Eng., National Taiwan Univ.
No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
kiatgak@gmail.com

**Li Sheng-an**
Dept. of Computer Science and Info. Eng., National Taiwan Univ.
No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
d93005@csie.ntu.edu.tw

**Kao Cheng-yan**
Dept. of Computer Science and Info. Eng., National Taiwan Univ.
No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
cykao@csie.ntu.edu.tw

## Abstract

In the past two hundred years or so, a sizable corpus of Taiwanese text in Latin script has been accumulated. However, due to the political and historical situation 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 by the Taiwanese tone sandhi system by describing a set of computational rules to approximate this system, as well as the results obtained from our implementation. Using the Taiwanese Latinization text as source, we take the sentence as the unit, translate every word into Chinese via a Taiwanese-Chinese dictionary, and obtain the POS information from the CKIP dictionary made by the CKIP group of the Academia Sinica. Using the POS data and tone sandhi rules we formulated based on linguistics, we then tag each syllable with its post-sandhi tone marker. Finally we implemented a Taiwanese tone sandhi processing system which takes a Latinized sentence as input and outputs the tone markers.

We were able to obtain an accuracy rate of 97.56% and 88.90% with training and testing data, respectively. We analyze the sources of error for the purpose of future improvement.

Keywords: written Taiwanese, tone sandhi system, Taiwanese latinization

## 1. Introduction

### 1.1.Background

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. (Tiu$^{n}$ 2001) At present there are several dozen if not more than a hundred proposed phonetic and writing systems for Taiwanese. (Iu$^{n}$&Tiu$^{n}$ 1999) The orthography adopted by this article is Peh-oe-ji (*POJ*, also known as *Latinized Taiwanese* or *Church Latinized 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 Peh-oe-ji Literature Data'. Although a lot of texts have already been lost due to the changing political situation, this project nevertheless revealed nearly 2,000 Peh-oe-ji books and periodicals, with publication sites spread over Taiwan, Xiamen, Shanghai, Guangzhou, Hong Kong, Singapore, Philippine, London, Japan, etc. The amount of publishing peaked in the 1950's and 1960's. (Iu$^{n}$&Tan-Te$^{n}$ 2005) The scope covers not only formally published books and periodicals but also non-published items such as personal letters and medical charts. Later on the government, citing supposed detrimental effects of Peh-oe-ji on Mandarin promotion as a reason, banned its use and thus contributed to the rapid decline of this practice.

We hope that applied information science might enable more people to access the extant materials collected by the above-mentioned project, as well as contribute to basic and applied research in Taiwanese. As most people nowadays are not familiar with Latinized written 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 we 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 "base tones", or the tones of syllables when they are read in isolation. At the level of the word, all syllables but the last one are usually read differently (that is, they manifest tone sandhi). At the level of a whole sentence, under most situations only the last syllables next to the boundary of the phrases or structural markers are read as base tones, the others being read as sandhi tones. In fact, besides the "regular tone sandhi" mentioned above, there exist several other kinds of situations which will be discussed one by one later.

This paper will first formulate the sandhi rules, which are the key to a correct pronunciation. The inputs of our experiment are mainly the data collected by the above-mentioned project; these data are processed by our sandhi system and the final outputs are these data with sandhi markers. Due to the lack of a set of well-marked data, we do not adopt the statistical model in this experiment but the rule-based model. Two of the authors who are long-time experienced Taiwanese users evaluated the outputs for their accuracy.

### 1.2. Tone Sandhi Problem

Tones in Taiwanese are traditionally analyzed as consisting of *piâⁿ, siáng, khì, ji̍p*, each having *im* and *iâng* but for *siáng*, so there are seven tones in all. Following the sequence of *im-piâⁿ, siáng, im-khì, im-ji̍p, iâng-piâⁿ, iâng-khì, iâng-ji̍p*, 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 descriptions of tone pitch are within the parentheses. For tone diacritics please refer to the following examples.

Tone sandhi is a very important characteristic of Taiwanese. At the level of the word, the last syllable is usually read as base tone and the others, as sandhi tones. For example in (1) the underlined syllables are read as base tones, while the others are read as sandhi tones, and dialect difference is represented in parenthesis :

(1) tâi台 (*"palteform"*) / Tâi-gí(gú)台語 (*"Taiwanese language"*)
Tâi-gí(gú)-bûn台語文 ("*written Taiwanese"*)
Tâi-gí(gú) bûn-ha̍k台語文學 ("*Taiwanese literature"*)
Tâi-gí(gú) bûn-ha̍k-sú台語文學史 ("*history of Taiwanese literature"*)

In fact, at the level of the syllable or the word, tone sandhi may manifest 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清清" (7,1) ("*water-white"*)
(ii) tone 7 → tone 3: "chēng-chēng靜靜" (3,7) ("*quiet"*)
(iii) tone 3 → tone 2: "chhiò-chhiò笑笑" (2,3) ("*smily"*)
(iv) tone 2 → tone 1: "léng-léng冷冷" (1,2) ("*cold"*)
(v) tone 5 → tone 7 or 3 (northern Taiwan): "âng-âng紅紅" (7/3,5) ("*red"*)
(vi) tone 4 → tone 8 (-p/t/k) or 2 (-h): like "sip-sip濕濕" (8,4); ("*moisty"*)
"phah-phah打打" (2,4) ("*beat"*)
(vii)tone 8 → tone 4 (-p/t/k) or 3 (-h): like "直直ti̍t-ti̍t" (4,8) (*"straight"*);
"熱熱jo̍ah-jo̍ah" (3,8) (*"hot"*)

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

(3) (i) "A-eng--a阿瑛a" (7,1,1) (the second "a" is a suffix) (*"a person name"*)
(ii) "góa lâi khòaⁿ -- i我來看伊" (1,7/3,3,3) (the base tone of "i伊" is tone 1) (*"I come to see him/her"*)
(iii) "hō--lí [給]你" (7,7) (the base tone of "lí你" is tone 2) (*"give you"*)

(c) Neutral sandhi: the previous syllable is read as base tone, and the tones of the neutral sandhi are read softly as if they were tone 3 or tone 4.

(4) (i) "Tân--sian-siⁿ(sin-seⁿ) 陳先生" (5,3,3) (the original tones of "sian-siⁿ (sin-seⁿ) 先生" are tone 7 and tone 1) (*"Mr. Tân"*)
(ii) "kiâⁿ--chhut-lâi行出來" (5,4,3) (the original tones of "chhut-lâi出來" are tone 8 and tone 5) (*"walk out"*)

(d) Double sandhi: this pattern mostly appears in syllables endng 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 tha̍k-chu[要]讀書" (1,4,1) ("beh 要" is tone 4, but rather than becoming tone 2, it becomes tone 1) (*"want to read books"*)
(ii) "khì gōa-kháu去外口" (1,3,2) ("khì去" is tone 3, but rather than becoming tone 2, it becomes tone 1) (*"go outside"*)

(e) Pre-*á* sandhi: the syllables before *á* are different from the normal sandhi unless they are tone 1 or tone 2.

(6) (i) tone 1 → tone 7: "sun-á孫仔" (7,2) (*"nephew"*)
(ii) tone 2 → tone 1: "chháu-á草仔" (1,2) (*"grass"*)
(iii) tone 3 → tone 1: "tàⁿ-á擔仔" (1,2) (*"stall"*)
(iv) tone 4 → tone 8 (-p/t/k) or tone 1 (-h): "tek-á竹仔" (8,2) (*"bamboo"*)
"thih-á鐵仔" (1,2) (*"iron"*)
(v) tone 5 → tone 7: "lō-á爐仔" (7,2) (*"oven"*)
(vi) tone 7 does not change: "phō-á簿仔" (7,2) (*"tablet"*)
(vii) tone 8 → tone 4 (-p/t/k) or tone 7 (-h): "chha̍t-á賊仔" (4,2) (*"thief"*)
"hio̍h-á葉仔" (7,2) (*"leaf"*)

(f) Triplicated 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清清清" (5,7,1) (*"more water-whiter"*)
(ii) tone 2 → tone 1: like "ún-ún-ún穩穩穩" (1,1,2) (*"more stabler"*)

(iii) tone 3 → tone 2: like "hèng-hèng-hèng興興興" (2,2,3)
(*"more interesting"*)
(iv) tone 4 → tone 8 (-p/t/k) or tone 2 (-h): like "sip-sip-sip濕濕濕" (8,8,4)
(*"more humid"*) "bah-bah-bah肉肉肉" (2,2,4) (*"more fatter"*)
(v) tone 5 → (similar to) tone 5: like "kôaⁿ-kôaⁿ-kôaⁿ寒寒寒" (5,7/3,5)
(*"more colder"*)
(vi) tone 7 → (similar to) tone 5: like "chēng-chēng-chēng靜靜靜" (5,3,7)
(*"more quieter"*)
(vii) tone 8 → (similar to) tone 5: like "ṫit-ṫit-ṫit直直直" (5,4,8)
(*"more straighter"*) "pe̍h-pe̍h-pe̍h白白白" (5,3,8) (*"more whiter"*)

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

(8) "ǒai-siak-chù [白襯衫]" (5,8,3) (*"white shirt"*)
"khǎn-páng[看板]" (5,2) (*"signboard"*)
"hǎn-tó͘-lù [方向盤]" (5,1,3) (*"steering wheel"*)

### 1.3. Historial Review

(Lin 1997) describes an early sandhi system. Its input is Chinese text, and its output is Taiwanese with pronounciation. The corpus is of Chinese news reports. Lin utilized the word segmentation and tagging data from the CKIP and the Taiwanese-Chinese dictionary from Robert Cheng, which was used to map the Chinese news into Taiwanese (in both Han and Latin scripts). The sandhi rules he applied were as follows: a) read the last syllable at the end of a sentence as base tone; b) read the syllable before the particle *ê* as base tone; c) read the last syllable of a noun as base tone; d) read others as normal sandhi tones. An accuracy rate of 82.53% was reported. However, the system did not take Taiwanese as input; word order and semantic ambiguities were not taken into account when converting Chinese text into Taiwanese; and the resulting translation was not quite native-like.

(Liang *et. al.* 2004) is a recent TTS system for Taiwanese. Its input was a large corpus of Chinese news texts in which sentences longer than 20 syllables were removed. It utilized a dictionary to convert the Chinese text into Taiwanese, 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 selected for evaluation by two Taiwanese-speaking experts. The accuracy rate of segmentaion reportedly exceeded 97%; the accuracy rate of pronounciation marking reached 89%; and the accuracy rate of rule-based sandhi processing reached 65%.

Compared with the above two systems, our approach has the following major differences: a Taiwanese corpus balanced for both literary and non-literary sources (about 50% each) was prepared; Chinese-to-Taiwanese translation is not an issue; and sentences of any length can be processed. In addition, because the text is of Latinized script, we do not encounter the problems of word segmentation and phonetic marking. However, compared with text written in Han 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 Peh-oe-ji. Following its orthography, syllables of a word are joined by hyphens, and words are separated with spaces.

The texts are offered by the above-mentioned project. We select parts of four books for the training data. They are:

- ✓ "*Sin-bûn ê chap̍-liok̍*" [News Bulletin] (1913; author unknown; genre: journalism);
- ✓ "*Chap̍-hāng kóan-kiàn"* [Ten Humble Opinions] (1924; author: Chhòa Pôe-hóe; genre: discourse);
- ✓ "*Chháu-tui téng ê bîn-bāng -- jî-tông chong-kàu kò͘-sū"* [Dreams on the Grass Stack -- Religious Stories for Children] (1955; author: N̂g Hôai-un; genre: short stories);
- ✓ "*Tang-pō͘thôan-tō kiàn-bûn ki*" [Record of Preaching in the Eastern Taiwan] (1961; author: Tân Kàng-hâng; genre: journalism)

The published dates of the above sources range from Japan-ruled era to Chinese Nationalist government era. Two sections are selected from each book; there are 614 syllables in total.

In addition to data drawn from the same project, the testing data also include some other sources we collected. Four sources are selected, as well, including:

- ✓ "*Peh̍-ōe-jī ê lī-ek*" [The Benefits of Using Peh-oe-ji] (1885; author: Reverend Ia̍p; genre: discourse);
- ✓ "*Kau-chiàn ê Siau-sit*" [News of the War] (1905; author: the editorial office of *Church News*; genre: report);
- ✓ "*Thiàⁿ lí iâⁿ kè thong sè-kan*" [Caring About You More than the Whole World] (1954; author: Lōa Ji̍n-seng; genre: novel);
- ✓ "*Ài lí kap ài i pîⁿ-á chōe*" [Loving You as Much as Her] (1997, on an Internet forum; author: Lô͘ Tàn-chhun; genre: prose)

The testing data also cover two eras but with a longer time span.

### 2.2. Part of Speech Tagging

Because there is no standard on part of speech (POS) for Taiwanese at this moment, we use that of Chinese instead. We obtain the corresponding Chinese translation for each Taiwanese word by looking up the Taiwanese-Chinese On-line Dictionary. (Iuⁿ 2003) We then look up the POS of the Chinese in the 80,000-word CKIP database. Ambiguity problems encountered include:

(a) homonymy, especially monosyllabic homonyms;
(b) one-to-many mapping when mapping Taiwanese to Chinese;
(c) multiple possible POS for each Chinese word.

On the problem of homonymy, we choose the word with the highest query frequency. After checking with the text we find that this strategy works under most situations. Due to one Taiwanese word mapping to multiple Chinese words, and one Chinese word possibly having 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 CKIP we adopt only the top level. The POS classes that we use are the following 12: A (adjective), C (conjunction), D (adverb), G (postposition), I (interjection), M (special mark), N (noun), P (preposition), R (pronoun), V (verb), S (time), and T (auxiliary). Moreover we 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.

As for indeterminate 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 1. Words with normal sandhi are not marked usually.

**Table 1**: Sandhi Marks

| (t) | # | @ | % | $ | & | ~ | ^ |
|---|---|---|---|---|---|---|---|
| Normal sandhi | Base tone | Following sandhi | Neutral sandhi | Double sandhi | Pre-*á* sandhi | Triplicate sandhi | Rising sandhi |

### 2.4. Tone Sandhi Rules

Tone sandhi rules are the most important part of this research. The algorithm of sandhi marking is shown in Table 2.

**Table 2**: Sandhi Marking Algorithm

1 Apply normal sandhi to all syllables
2 Mark the last syllable as base tone #
3 (Word level) *ê*: Mark the syllable preceding *ê* as base tone #

4 (POS level) A/A Pair (without ambiguity)
  4.1 A/A Pair: Mark the last syllable of the first word as base tone #
5 (POS level) N/V, N/A, N/P, N/R, and N/D Pairs (without ambiguity)
  5.1 N/V Pair: Mark the last syllable of the first word as base tone #
  5.2 N/A Pair: Mark the last syllable of the first word as base tone #
  5.3 N/P Pair: Mark the last syllable of the first word as base tone #
  5.4 N/R Pair: Mark the last syllable of the first word as base tone #
  5.5 N/D Pair: Mark the last syllable of the first word as base tone #

6 (POS level)
  C: Mark the last syllable of the preceding word as base tone #
7 (POS level)
  G: Mark the last syllables of both the preceding word and the word itself as base tones #'s
8 (POS level)
  S: Mark the last syllable of this word as base tone #
9 (Word level) POS R
  9.1 *i* / *in*: Mark them as normal sandhi even if they are the last syllables
  9.2 *góa* / *lí* / *gún* / *góan* / *lán* / *lín* of POS R: Mark them as normal sandhi if they are not the last syllables
10 (Word level) Sentence-final *kóng* [講]: Mark this word as normal sandhi if the delimeter 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)

11 (Syllable level)
  pre-*á*: Mark any syllables just before *á* as pre-*á* sandhi &
12 Double sandhi
  12.1 (Syllable level) *beh*: Mark any *beh* as double sandhi $ unless it appears at the end, including those within a word, such as *kio͘ng-beh, tih-beh*.

12.2 (Word level) *khì* [去]: Mark *khì* as double sandhi $ if the POS of the immediately following word is N or V, unless it appears at the end

12.3 (Syllable level) *koh*: Mark any *koh* as double sandhi $, including those within a word, such as *chiah-koh* or *iáu-koh*, unless it appears at the end

12.4 (Word level) *kah*: Mark any *kah* as double sandhi $ unless it appears at the end

13 (Word level) Neutral sandhi of --: Mark the syllable just before -- as base tone, and mark each syllable after -- as neutral sandhi %

14 (Word level) Triplicate sandhi: Mark the first syllable as triplicate sandhi if that word has 3 syllables of the same spelling

15 (Word level) Special words

15.1 *sím-mih* / *sím-mi̍h*: Change these words into sím-mí (sandhi marks not changed)

15.2 *án-ni* / *àn-ni* / *an-ni* / *an-nī*: Change these words into *án-ni* and to mark its sandhi marks as t#

16 Markers

16.1 (Word level) *iah-sī* / *ah-sī* / *ia̍h-sī* / *a̍h-sī* / *á-sī*: Mark the last syllable before these words as base tone #

16.2 (Pattern level) V *sī* . . . V: Mark the last syllable of the verb that just before *sī* [是] as base tone # if this verb appears again after *sī*

16.3 (Word level) *che* / *he* / *chia* / *hia*: Mark these words as base tone #

16.4 (Word level) *ū-sî* [有時] / *put-sî* [不時] / *kui-khì* [kui 氣] / *óan-jiân* [宛然] / *gôan-lâi* [原來] *chiong-lâi* [將來] / *chiông-lâi* [從來] / *sui-jiân* [雖然] / *sui-bóng* [雖罔] / *sî-siông* [時常] / *hui-siông* [非常] / *si̍t-chāi* [實在] / *sī-chūn* [時陣]: Mark the last syllables of these words as base tone #

16.5 (Word level) *chiū* / *tō* [就] : Mark the syllable of the word just before as base tone # if the POS of the word is A

16.6 (Word level) *sî-kàu* [時 kàu]: Mark both two syllable of this word as base tones

17 (POS level) T: Mark the last syllable of a word as base tone if the word is just before a word of POS T in the end

18 Other sandhi:

18.1 (Word level) *teh* [在]: Mark *teh* or the *teh* in *tī-teh* as other sandhi ^

19 (Word level) Neutral sandhi

19.1 (Word level) *chhut-khì* [出去] *chhut-lâi* [出來] *lo̍h-lâi* [落來] *lo̍h-khì* [落去] *kòe-lâi*/*kè-lâi* [過來] *kòe-khì*/*kè-khì* [過去]: Mark the last syllable of a verb just before these words as base tone #, and mark these words as neutral sandhi %

19.2 *sian-si*$^{n}$/*sin-se*$^{n}$/*sian-se*$^{n}$ [先生]: Mark the word before these words as base tone # and these words as neutral sandhi %, if the first letter of the preceding word is uppercase

19.3 (Word level) *bô* [無] at the end

19.3.1 *á* / *á-sī* / *iah* / *iah-sī* / *ah* /*ah-sī* [或是]: 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 base tone #, and mark *bô* as neutral sandhi %

19.4 (Word level) *bē*/*bōe* [不會] at the end

19.4.1 *ē*/*ôe* [會] *ē-hiáu*/*ôe-hiáu* [會曉]: Mark any final *bē*/*bōe* as neutral sandhi %

19.4.2 *á* / *á-sī* / *iah* / *iah-sī* / *ah* /*ah-sī* [或是]: Mark the *bē*/*bōe* as base 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])

20 (Word level) R at the end

*góa* / *lí* / *i* / *gún* / *góan* / *lán* / *lín* / *in*: Mark the pronoun as following sandhi @ if it appears at the end and there is a verb before it

We set up the algorithm using the following resources:

(a) Tone sandhi rules proposed by linguists;
(b) Rules induced from the training data;
(c) Our understanding as native-speaking observers of sandhi phenomena;
(d) The word segmentation results of the CKIP (noting its POS tagging output);
(e) Taiwanese concordancer system (to check the sandhi phenomena of certain words).

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 over 80% 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.

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

A. The syllable: E.g. *koh* ("and") and *beh* ("want") are read as double sandhi regardless of whether they are part of a word or not, such as *koh-chài* ("again") and *kiông-beh* ("almost").
B. The word: E.g. *che* ("this") and *he* ("that") are read as base tones wherever they appear. Also, the presence of some words (such as *ê* ["of"]) would change the sandhi mark of a preceding word.
C. The POS: E.g. the last syllable of a noun is read as base tone if the POS of its following word is N (noun), A (adjective), D (adverb), P (preposition), R (pronoun) or V (verb). Some types of POS, such as G (postposition), would affect the sandhi mark of the preceding word.
D. The sentence pattern: Certain words exhibit clause-marking function -- for example, *iah-sī* ("or"); the portion before these words are regarded as clauses. Another example is that *bē* in the pattern of *ē...bē* ("whether...will...or not") would be read as neutral tone.

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 你會去國外[不會]" : the last *bē* is marked as neutral sandhi.
"Lí ē khì kok-gōa iah-sī bē 你會去國外[或]是[不會]" : the last *bē* is marked as base 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 (in both Latinized script and mixed Han/Latinized scripts):

(10) Chhin-chhiūⁿ án-ni lâi kóng , chāi lán Tâi-ôan kīn-kīn ch̍it-tiap-á-kú ê kang-hu , ài soaⁿ chiū ū soaⁿ , ài hái chiū ū hái , beh j̍oah chiū ū j̍oah ,kôaⁿ chiū ū kôaⁿ . Só͘-í thang kóng Tâi-ôan sī ch̍it-ê sió Tang-iûⁿ . Lán Tâi-ôan ū chit-khóan thian-jiân ê hó-kéng ,hó khì-hāu , chiong-lâi nā-sī ēng-sim ke lâng ê kang-hu tōa-tōa lâi chéng-tùn , tek-khak ē chiâⁿ-chò Tang-iûⁿ ê tōa kong-hn̂g , hō͘ Tang-iûⁿ ê lâng ch̍ip-óa lâi hióng-hok an-l̍ok .

親像 án-ni 來講，在咱台灣近近一 tiap 仔久 ê 工夫，愛山就有山、愛海就有海；beh 熱就有熱、寒就有寒。所以 thang 講台灣是一個小東洋。咱台灣有這款天然 ê 好景、好氣候，將來若是 koh 用心加人 ê 工夫大大來整頓，的確會成做東洋 ê 大公園，hō͘ 東洋 ê 人集倚來享福安樂。

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) ch̍it-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) j̍oah#(A) chiū(D) ū(D;P;V) j̍oah#(A), kôaⁿ#(A) chiū(D) ū(D;P;V) kôaⁿ#(A). Só͘-í(C) thang(D) kóng(V) Tâi-ôan#(N) sī(D;V) ch̍it-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) ch̍ip-óa(V) lâi(D;V) hióng-hok#(A) an-l̍ok#(A).

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

## 3. Results

### 3.1. Evaluation

Two 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 results are acceptable, and depending on discourse considerations some speakers may opt for one sandhi result over others.

### 3.2. Preliminary Results

Preliminary results are listed in Table 3. There are 614 syllables of training data, 15 errors, giving

an accuracy rate of 97.56%. There are 955 syllables of testing data with 106 mistakes, or an accuracy rate of 88.90%. Some of the errors in traing data output are attributable to the incompleteness of the sandhi rule set. We expect improved results of at least 2.5% once additional rules are appended.

**Table 3**: Accuracy Rates of Sandhi Marks

| | Syllables (A) | Errors (B) | Accuracy Rate (1-B/A) |
|---|---|---|---|
| Training data | 614 | 15 | 97.56% |
| Testing data | 955 | 106 | 88.90% |

## 4. Analysis of Mistakes 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 linguistic question requiring further investigation. Although a few studies of the POS of Taiwanese are available from as early as the 1930's, currently these data have yet to be made available in digital form, 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 comforming 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)打拼做空課" (2,2,2,7,3)
(*"do work hard"*)

b. "kia̍h-ba̍k(V) khòa$^{n}$(V) hn̂g(N)舉目看園" (3,8,2,5)
(*"toss head 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 base 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-chú lêng-kiā$^{n}$電子零件" (*"electronic accessory"*)
b. "thâng-thōa chiáu-chiah蟲豸鳥隻" (*"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 base tone. Currently we see no solution around this.

### 4.5. Error Conditions

Error conditions including those discussed in the previous sections are listed below with possible solution in brackets:

(a) Errors due to dictionary limitation (not having the words); [to increase entries]
(b) Errors due to lack of punctuation marks;
(c) Errors due to wrong POS because of homonymy;
(d) Errors due to indeterminate POS or multiple candidates;
(e) Errors caused by inconsistent orthography in hyphen segmentation; [to revise the sources or deal with the procedures of adding or removing hyphens automatically]
(f) Errors due to incomplete sandhi rule set; [to refine the sandhi rules while avoiding side effects]
(g) Errors associated with quantitative words; [to define rules handling quantitative words]
(h) Errors associated with proper nouns; [to define rules for recognizing proper nouns]
(i) Errors associated with sentence pattern; [to 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 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 of part-of-speech, 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)

## 6. Acknowledgements

Many thanks to the National Museum of Taiwanese Literature for financial support, to two reviewers for their valuable comments, and to Henry H. Tan-Te$^{n}$ for reviewing the English version of this paper.

## 7. References

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

Cheng, Robert. 2002. Tone Sandhi on the Grammar Template--Cognition and Testing. *Proceeding of 2002 International Conference on Teaching and Researching of Taiwanese Romanization.*

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

Iu$^{n}$, Un-Gian. 2003. Taiwanese-Chinese On-line Dictionary -- Discussion of Building Technique and its Utilization. *Proceeding of 3rd International Conference on Internet Chinese Education,*

pp. 132-141.
Iu$^{n}$, U-.G-. and H-.K-. Tiu$^{n}$. 1999. Review and Analysis of Taiwan Ho-lo Language non-Han Character Spelling Symbols. *Proceedings of 1$^{st}$ Conference on the Regeneration and Rebuild of Taiwan Mother Tongue Culture*, pp. 62-76.
Iu$^{n}$, U-.G-. and H-.H-. Tan-Te$^{n}$. 2005. 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. 2004. A Taiwanese Text-to-Speech System with Applications to Language Learning. *Proceedings of the 4$^{th}$ IEEE International Conference on Advanced Learning Technologies,* pp. 91-95.
Lin, Chuan-Jie. 1997. *The Study of a Mandarin-Taiwanese Machine Translation System*. Master Thesis, Natioanl Taiwan University.
Lu, Guang-Cheng. 1999. *The Study of Minnan Vocabulary in Taiwan*. SMC Publishing Inc.
Tiu$^{n}$ Ju-Hong (Chang Yu-hung). 2001. *Principles of POJ or the Taiwanese Orthography: An Introduction to Its Sound-Symbol Correspondences and Related Issues*. Crane Publishing Co.
Tseng, Chin-Chin. 1997. 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.

**Web Site**

*Chinese On-line Word Segmentation System*. http://ckipsvr.iis.sinica.edu.tw
*Taiwanese Concordancer System*. http://iug.csie.dahan.edu.tw/TG/concordance/form.asp
*Taiwanese Package Website.* http://www.phahng.idv.tw , http://taigu.fhl.net/TP/
*Unicode Interface to the Holo-Mandarin Dictionary*. http://lomaji.com/poj/tools/su-tian/index-en.html