Chapter 4 Tone Sandhi Problem and Algorithm

We will propose a rule-based tone sandhi algorithm in this chapter. We address 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 POJ text as the source, we take the sentence as the unit, translate every word into Mandarin via OTMD, and obtain the POS information from CED made by the CKIP group of the Academia Sinica. Using the POS data and tone sandhi rules 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 POJ script sentence as an input and then outputs the tone markers. Our system achieves 97.39% and 88.98% accuracy rate with observation and test data, respectively.

4.1 Tone Sandhi Problem of the Taiwanese Language

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 POJ 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 syllable 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.

4.1.1 Types of the Taiwanese Language Tone Sandhi

Tones in Taiwanese language are traditionally analyzed as consisting of piân ‘平’, siáng ‘上’, khì ‘去’, ji̍p ‘入’, each having im ‘yin 陰’ and iâng ‘yang 陽’ except siáng. So there are seven tones in total. 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. Tones 1,2,3,5,7 are smooth syllables and tones 4,8 checked syllables.

Tone sandhi is a very important characteristic of the Taiwanese language. 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 ‘platform 台’
Tâi-gí ‘Taiwanese language, 台語’
Tâi-gí-bûn ‘written Taiwanese, 台語文’
Tâi-gí bûn-ha̍k ‘Taiwanese literature, 台語文學’
Tâi-gí bûn-ha̍k-sú ‘history of Taiwanese literature, 台語文學史’

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" (7,1) 'clear, 清清'
(ii) tone 7 → tone 3: "chēng-chēng" (3,7) 'quiet, 靜靜'
(iii) tone 3 → tone 2: "chhiò-chhiò" (2,3) 'smiley, 笑笑'
(iv) tone 2 → tone 1: "léng-léng" (1,2) 'cold, 冷冷'
(v) tone 5 → tone 7 or 3 (northern Taiwan): "âng-âng" (7/3,5) 'red, 紅紅'
(vi) tone 4 → tone 8 (-p/t/k) or 2 (-h): "sip-sip" (8,4); 'moist, 濕濕'
"khoeh-khoeh" (2,4) 'crowd, 擁擠'
(vii) tone 8 → tone 4 (-p/t/k) or 3 (-h): like "tit-tit" (4,8) 'straight 直直'; "jo̍ah-jo̍ah" (3,8) 'hot, 熱熱'

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

(3) (i) "A-eng--a" (7,1,1) 'a personal name, 阿英' (the second "a" is a suffix)
(ii) "góa lâi khòaⁿ -- i" (1,7/3,3,3) 'I come to see him/her 我來看他' (the basic tone of "i" '(s)he, 他' is tone 1)
(iii) "hō͘ --lí" (7,7) 'give you 給你' (the basic tone of "lí" 'you, 你' 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-si^{n}" (5,3,3) 'Mr. Tân 陳先生' (the original tones of "sian-si^{n} " 'Mr. 先生' are tone 7 and tone 1)

(ii) "kiâ^{n}--chhut-lâi" (5,4,3) 'walk out 走出來' (the original tones of "chhut-lâi" 'out 出來' 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) 'want to read books 要讀書' ("beh" 'want, 要' is tone 4, but rather than becoming tone 2, it becomes tone 1)

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

(e) Preceding 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) 'nephew, 姪子'

(ii) tone 2 → tone 1: "chháu-á" (1,2) 'grass, 小草'

(iii) tone 3 → tone 1: "tà^{n}-á" (1,2) 'stall, 攤位'

(iv) tone 4 → tone 8 (-p/t/k) or tone 1 (-h): "tek-á" (8,2) 'bamboo, 竹子' "thih-á" (1,2) 'iron, 鐵')

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

(vi) tone 7 does not change: "phō͘-á" (7,2) 'tablet 簿子'

(vii) tone 8 → tone 4 (-p/t/k) or tone 7 (-h): "chhȧt-á" (4,2) 'thief, 賊' "hiȯh-á " (7,2) 'leaf, 葉子'

(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) ‘very clear, 清清清’

(ii) tone 2 → tone 1: like “ún-ún-ún” (1,1,2) ‘very stable, 穩穩穩’

(iii) tone 3 → tone 2: like “hèng-hèng-hèng” (2,2,3) ‘very interesting, 興興興’

(iv) tone 4 → tone 8 (-p/t/k) or tone 2 (-h): like “sip-sip-sip” (8,8,4) ‘very humid, 濕濕濕’ “bah-bah-bah” (2,2,4) ‘very fat, 肉肉肉’

(v) tone 5 → (similar to) tone 5: like “kôa^{n}-kôa^{n}-kôa^{n}” (5,7/3,5) ‘very cold, 冷冷冷’

(vi) tone 7 → (similar to) tone 5: like “chēng-chēng-chēng” (5,3,7) ‘very quiet, 靜靜靜’

(vii) tone 8 → (similar to) tone 5: like “tit-tit-tit” (5,4,8) ‘very straight, 直直直’ “pe̍h-pe̍h-pe̍h” (5,3,8) ‘very white, 白白白’

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

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

We collates above sandhi phenomena in the following table.

Table 4 - 1 The Taiwanese 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			
Preceding 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

4.1.2 Boundary of Tone Sandhi Group

We master the sandhi phenomena of the Taiwanese language in word level, nevertheless, it is difficult for us to realize in sentence level.

Compared with other Chinese languages, the sandhi domain of the Southern Min is more complicated than others. The sandhi domain of the Taiwan Southern Min approximates the phonological phrase (or tone sandhi group) (M. Y. Chen, 2000).

When we want to implement a Taiwanese language pronunciation system, the first thing is to find the boundary of the tone sandhi group. The tone sandhi domain sometimes equals a phrase, but this statement is not definite. A phrase can contain another phrase in it, which is the boundary ? “ê” also is an important marker to tell us the position of the boundary is preceding syllable of “ê”.

If we have a written Taiwanese parser, the problem will become easy. We need to seek another way when the parser has not ready yet. The direction may be syllables, words, phrases, and sentence patterns.

4.2 Implementation of the Taiwanese Pronunciation System

4.2.1 System Diagram

We want to implement the Taiwanese language pronunciation system. We

need a Taiwanese dictionary with syntactic tagged information except for the tone sandhi knowledge. We don't have the tagged Taiwanese but the OTMD which the number of entries is more than 62,000. Therefore, we get the tagged information from CED via the Mandarin translation of OTMD. CED is made by CKIP group, and OTMD is managed by me. The following figure is the system diagram.

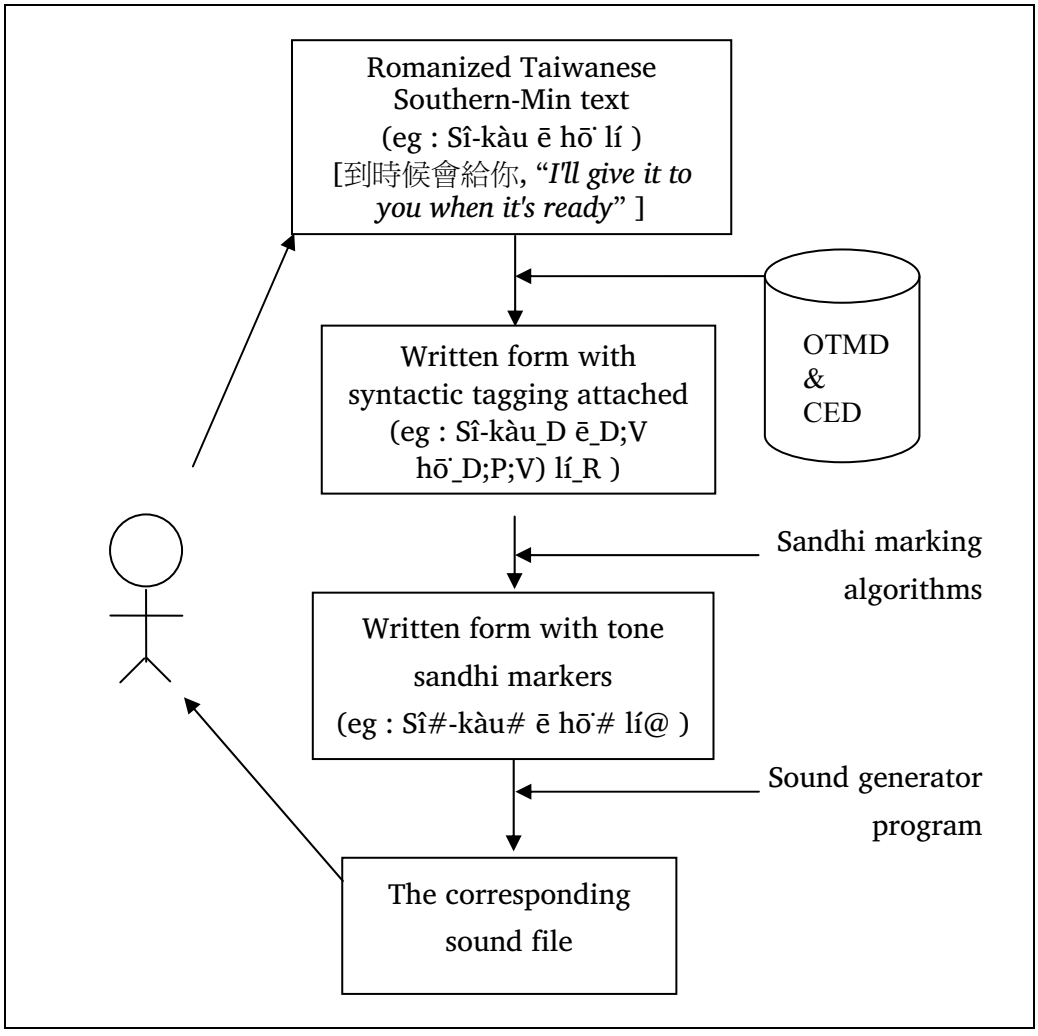

Fig 4 - 1 Taiwanese Tone Sandhi System Diagram

Data source is CCTPLD project carried out by the Department of Taiwanese Literature of Cheng Kung University under the auspices of the NMTL. The

project collected the Taiwanese literature data and typed part of them. The total size is about 2,580,000 syllables. Following POJ orthography, syllables of a word are joined by hyphens, and the words are separated with spaces.

We intend to let the literature data listenable. To achieve this goal, we record each syllable of Taiwanese, write a rule-based tone sandhi algorithm, and write a program to concatenate all the corresponding syllables according to the user input (in numbered POJ) and the tone sandhi algorithm.

4.2.2 Observation Data and Test Data

We select parts of four sources as observation data. The observation data sources are shown in the following table.

Table 4 - 2 Observation Data Sources

Book or Article	Year	Author	Genre
“Sin-bûn ê chap-liok” ‘News Bulletin 新聞的雜錄’	1913	unknown	Journalism
“Chap-hāng kóan-kiàn” ‘Ten Humble Opinions 十項管見’	1925	Chhòa Pôe-hóe ‘蔡培火’	Discourse
“Chháu-tui téng ê bîn-bāng -- Jî-tông chong-kàu kò͘-sū” ‘Dreams on the Grass Stack -- Religious Stories for Children 草堆上的夢—兒童宗教故事’	1955	Ńg Hôai-un ‘黃懷恩’	Short stories
“Tang-pō͘ thôan-tō kiàn-bûn kì” ‘Record of Preaching in Eastern Taiwan 東部傳道見聞記’	1961	Tân Kàng-hâng ‘陳降祥’	Journalism

The published dates of the above sources range from Japanese-ruled era (1895-1945) to postwar era (1945-). Two paragraphs are selected from each book; there are 614 syllables (438 word tokens) in total.

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 4 - 3.

Table 4 - 3 Test Data Sources

Book or Article	Year	Author	Genre
“Pe̍h-ōe-jī ê lī-ek” ‘The Benefits of Using Pe̍h-ōe-jī, 白話字的利益’	1885	Reverend Ia̍p ’葉牧師’	discourse
“Kau-chiàn ê Siau-sit” ‘News of the War 交戰的消息’	1905	the editorial office of Tâi-lâm Prefectural Church News	report
“Thiàⁿ lí iâⁿ kè thong sè-kan” ‘Caring About You More Than the Whole World, 疼愛你勝過全世界’	1955	Lōa Jîn-seng ’賴仁聲’	novel
“Ài lí kap ài i pe̍ⁿ-á chōe” ‘Loving You as Much as Her, 愛妳和她一樣多’	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.

4.2.3 POS Tagging Set

Because there is no standard on POS for Taiwanese at present, we use the standard of Mandarin instead. We obtain the corresponding Mandarin translation for each Taiwanese word by looking up the OTMD. We then look up the POS of the Mandarin in the 80,000-word CED. Ambiguity encountered includes:

(a) homonymy, especially monosyllabic homonyms;

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

(c) multiple possible POSs for each Mandarin 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 facts that one Taiwanese word can map to multiple Mandarin words and one Mandarin word may 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.

The reason why we do not disambiguate the result of POS tagging is that we intend to implement a real-time online Taiwanese tone sandhi system. It is appreciated to save any time since the time of concatenating the corresponging sound files and returning is avoidless. Besides, the POS tagging result is not the only clue in our tone sandhi algorithm.

Of the 46 POSs in the CED, 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 - 4.

Table 4 - 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).

4.2.4 Tone Sandhi Marks

The marks representing tone sandhis are listed in Table 4 - 5. Words with normal sandhi are usually not marked .

Table 4 - 5 Tone Sandhi Marks

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

4.3 Rule-based Tone Sandhi Algorithm

Tone sandhi rules are the most important part of this study. The algorithm of sandhi marking is shown in Table 4 - 6.

Table 4 - 6 Tone Sandhi Marking Algorithm

Rule	Remark
1 Apply normal sandhi to all syllables	
2 Mark the last syllable as basic tone #	
3 **ê** 'of 的': 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 #	POS level , with ambiguity

Table 4 - 6 Tone Sandhi Marking Algorithm

	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 #	
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** '(s)he/they 他(們)': Mark them as normal sandhi even if they are the last syllables 9.2 góa / lí / gún / góan / lán / lín 'I/you/my/our/your 我/你(們)(的)' of POS R: Mark them as normal sandhi if they are not the last syllables	POS/Word level
10	Sentence-final **kóng** 'say 講': 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 observation data
11	Preceding **á** [á is suffix of a word]: Mark any syllables just before á as preceding á sandhi &	Syllable level
12	Double sandhi 12.1 beh 'want 要': Mark any beh as double sandhi $ unless it appears at the end, including those within a word, such as kiông-beh, tih-beh . 12.2 khì 'go 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 koh 'again 再': Mark any koh as double sandhi $, including those within a word, such as chiah-koh 'and then 再' or iáu-koh 'still 還是' , unless it	 Syllable level Word level Syllable level, extended from observation data

Table 4 - 6 Tone Sandhi Marking Algorithm

	appears at the end 12.4 kah ‘and 和’: 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 15.1 sím-mih / sím-mi̍h ‘what 什麼’: Change these words into sím-mí (sandhi marks not changed) 15.2 án-ni / àn-ni / an-ni / an-nī ‘thus 這樣’: Change these words into án-ni and to mark its sandhi marks as t#	Word level, extend from observation data because of not yet standardized
16	Markers	
	16.1 iah-sī / ah-sī / ia̍h-sī / a̍h-sī / á-sī ‘or 或是’: Mark the last syllable before these words as basic tone #	word level, extended from observation data
	16.2 V sī ‘is 是’ 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 observation data
	16.3 che / he / chia / hia ‘this/that/(t)here 這/那(裡)’: Mark these words as basic tone #	word level
	16.4 ū-sî ‘sometimes 有時’ / put-sî ‘from time to time 不時’ / kui-khì ‘just 乾脆’ / óan-jiân ‘like 宛然’ / gôan-lâi ‘originally 原來’ / chiong-lâi ‘future 將來’ / chiông-lâi ‘always 從來’ / sui-jiân/ sui-bóng ‘though 雖然’ / sî-siông ‘, “often 時常’ / hui-siông ‘very 非常’/ si̍t-chāi ‘really 實在’ / sī-chūn ‘(the duration of) time 時候’: Mark the last syllables of these words as basic tone #	word level, extended from observation data
	16.5 chiū / tō ‘as soon as 就’: Mark the syllable of the word just before as basic tone # if the POS of the word is A	word level, induced / extended from observation data

Table 4 - 6 Tone Sandhi Marking Algorithm

16.6 sî-kàu ‘at that time 到時候’: Mark both two syllable of this word as basic tones	word level, induced from observation 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:	word level, our observations
18.1 teh ‘at 在’: Mark teh or the teh in tī-teh as other sandhi ˆ	
19 Neutral sandhi	
19.1 chhut-lâi ‘come out 出來’ / chhut-khì ‘go out 出去’ / lo̍h-lâi ‘come down 下來’ lo̍h-khì ‘go down 下去’ kòe-lâi ‘come up 過來’ kòe-khì ‘pass away 過去’: 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} ‘Mr. 先生’: 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 ‘have nothing 無’ at the end	word level,
19.3.1 á / á-sī / iah / iah-sī / ah /ah-sī ‘or 或是’: if the preceding word is among these words, do nothing	induced / extended from observation data
19.3.2 Otherwise: Mark the last syllable of the word just before bô as basic tone #, and mark bô as neutral sandhi %	
19.4 bē/bōe ‘will not 不會’ at the end	word level,
19.4.1 ē/ōe ‘be good at 會’ ē-hiáu/ōe-hiáu ‘be good at 會’: Mark any final bē/bōe as neutral sandhi %	induced / extended from observation data
19.4.2 á / á-sī / iah / iah-sī / ah /ah-sī ‘or 或是’: 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 (*eg.* bē/bōe ‘sell 賣’)	
20 R at the end góa / lí / i / gún / góan / lán / lín / in ‘ I / me / we / us	word level

Table 4 - 6 Tone Sandhi Marking Algorithm

/ my / our/ you(r)(s)/(s)he / him / her / his / they / them / their 我/你/他(們)(的)': Mark the pronoun as following sandhi @ if it appears at the end and there is a verb before it	

The program is implemented according to the sequence of the above rules.

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 (R. L Cheng, 1997, 2002);

(b) Rules induced from the observation data;

(c) Our intuition as native-speaking observers of sandhi phenomena.

We also consulted:

(d) The CWSTS to examine its word segmentation result and POS tagging output (CKIP, 2004);

(e) The OTCS to check the sandhi phenomena of certain words when we met some questions (Iunn, 2003b).

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 observation data output. In principle sandhi rules are formulated to be applicable to “most situations” -- i.e. an accuracy rate 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" 'Will you go abroad or not 你會不會去國外': the last bē 'will not 不會' is marked as neutral sandhi, whereas
"Lí ē khì kok-gōa iah-sī bē" 'Will you go abroad or not 你會不會去國外': 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.

4.4 Results, Accuracy Rate and Discussion

4.4.1 Experiment Results

The following observation data represents a pre-tagged source (Mandarin 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] ji̍oah[熱] chiū[就] ū[有] ji̍oah[熱] , kôaⁿ[冷]	*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*

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[成爲] 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

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-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) chi̍p-óa(V) lâi(D;V) hióng-hok#(A) an-lo̍k#(A).

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

4.4.2 Accuracy Rate and Related Analysis

There are 614 syllables of observation 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 4 - 7 shows the number of errors and accuracy rate for each paragraph.

Table 4 - 7 Number of Errors and Accuracy Rate for Each Paragraph

Observation 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	96.7	1	130	184	16	91.3
2	42	54	0	100.0	2	56	85	12	85.9
3	44	70	0	100.0	3	53	84	13	84.5
4	33	52	0	100.0	4	96	143	16	88.8
5	38	51	4	92.2	5	66	97	10	89.7
6	85	110	4	96.4	6	63	86	9	89.5
7	97	144	6	95.8	7	32	43	3	93.0
8	72	103	1	99.0	8	38	58	2	96.6
					9	122	182	25	86.3
Total	438	614	16	97.4	Total	656	962	106	89.0

Table 4 - 8 shows the numbers of each rule applied in observation data and test data respectively. We count the number of the affected syllables and accurately affected syllables and accuracy rate of each rule. Note that rule 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 is no triplicated words in our observation and test data.

Table 4 - 8 Affected and Accurately Affected Syllables of Each Rule

	Observation Data			Test Data		
Rule id.	affected syllables	accurately affected	accuracy rate(%)	affected syllables	accurately affected	accuracy rate(%)
1	614	411	66.9	962	662	68.8
2	74	68	91.9	112	105	93.8
3	32	24	75.0	38	26	68.4
4	3	3	100.0	13	7	53.9
5	65	57	87.7	129	90	69.8
6	4	3	75.0	4	3	75.0
7	0	0	--	0	0	--
8	5	5	100.0	3	3	100.0
9	29	29	100.0	25	25	100.0
10	5	5	100.0	0	0	--
11	3	3	100.0	8	8	100.0
12	8	8	100.0	11	11	100.0
13	2	2	100.0	6	5	83.3
14	0	0	--	0	0	--
15	8	8	100.0	6	5	83.3
16	13	13	100.0	4	4	100.0
17	3	3	100.0	2	2	100.0
18	9	9	100.0	3	3	100.0
19	0	0	--	6	6	100.0
20	2	2	100.0	0	0	--

Every syllable affected by at least one rule, and four rules at most. We call the last affected rule as dominant rule because the tone sandhi marker of the syllable is finally determined by this rule. Table 4 - 9 shows the number of dominant rule, accurate dominant rule and accuracy rate.

Table 4 - 9 Number of Dominant Rule, Accurate Dominate Rule and Accuracy Rate

	Observation 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	97.4	616	568	92.2
2	62	62	100.0	104	99	95.2
3	24	23	95.8	34	26	76.5
4	2	2	100.0	6	3	50.0
5	62	57	91.9	126	87	69.1
6	2	2	100.0	4	3	75.0
7	0	0	--	0	0	--
8	4	4	100.0	3	3	100.0
9	27	27	100.0	25	25	100.0
10	5	5	100.0	0	0	--
11	3	3	100.0	8	8	100.0
12	7	7	100.0	11	11	100.0
13	2	2	100.0	6	5	83.3
14	0	0	--	0	0	--
15	6	6	100.0	5	4	80.0
16	13	13	100.0	3	3	100.0
17	3	3	100.0	2	2	100.0
18	9	9	100.0	3	3	100.0
19	0	0	--	6	6	100.0
20	2	2	100.0	0	0	--

After examination, we find that we can add 7 additional rules without too much effort; this way we were able to fix 20 errors and achieve a 91.06% accuracy rate. Table 4 - 10 shows additional rules in order to fix 20 errors in test data.

Table 4 - 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-jip-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

4.4.3 Discussion

In our investigation we use the POS set for Mandarin. 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.

Besides, 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)
’do work hard 努力做工作’
(b) “kiȧh-bȧk(V) khòa^{n}(V) hn̂g(N)” (3,8,2,5)
’lift eyes and see plowland 舉目看園’

(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) “siong-giȧp kóng-kò” ‘commercial advertisement 商業廣告’
(b) “thâng-thōa chiáu-chiah” ‘insects and birds 昆蟲(、)小鳥’

(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 around this.

Regarding error conditions, including those discussed in the previous sections, Table 4 - 11 lists the possible solutions:

Table 4 - 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 it 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

Table 4 - 11 Error Conditions and Possible Solutions

(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	

4.5 Summary and Possible Direction

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) Change the dictionary's POS tags, such as making use of Embree's POS analysis (Embree, 1984).

(e) Improve the sandhi rules.

(f) Find alternative ways of modeling sandhi processing like template theory or optimality theory. Tē^{n}, Liông-úi proposes that the tone sandhi can be observed from the template theory. In template theory, every word can be palced in a cell of the template. The tone sandhi of the

word is dependent on the position of the template (R. L Cheng, 2002). Hsiao Yuchau states that the optimality theory with the dominant tone sandhi constraints Ident-T-R, $*T^2/D$ and the dominant prosodic constraint AlignP-R can help us to determine the tone sandhi (Hsiao, 2000).

(g) Using machine learning method to model tone sandhi processing if we can construct a corpus with tone sandhi markers.