Chapter 1 Introduction

This dissertation will focus on Taiwan Southern Min text processing. This chapter will introduce the background of this language and some related issues.

1.1 Background

1.1.1 Language Population in Taiwan

There are a total of 22 living languages in Taiwan (Gordon, 2005). The following table ranks these in descending order.

Table 1 - 1 Living Languages in Taiwan

Language	Population	Language	Population
Min Nan	15,000,000	Taroko	4,750
Mandarin	4,323,000	Yami	3,384
Hakka	2,366,000	Tsou	2,127
Amis	137,651	Kavalan	24
Atayal	84,330	Kanakanabu	6 to 8
Taiwan Sign Language	82,558	Amis, Nataoran	5
Paiwan	66,084	Saaroa	5 to 6
Bunun	37,989	Thao	5 to 6
Rukai	10,543	Babuza	3 to 4
Puyuma	8,487	Kulon-Pazeh	1
Saisiyat	4,750	Japanese	
Source: (Gordon, 2005)			

As can be seen in Table 1 - 1, the Min Nan (Southern Min) population is in the majority. Six of these languages are nearly extinct because their speaking

population is less than 100. In fact, at least 10 languages, mainly aboriginal languages, have disappeared in Taiwan over the past several hundred years. In addition, some languages, like Vietnamese, Thai, *etc.*, have migrated to Taiwan because of mixed marriages or migrant work.

1.1.2 Southern Min Language Population

Let us take another viewpoint and examine the worldwide population of Southern Min speakers (Gordon, 2005). The following table shows the populations of Southern Min speakers in different nations in descending order of percentages.

Table 1 - 2 Southern Min Language Populations around the World.

Nation	Population	Total Population	Percentage (%)
Taiwan	15,000,000	22,057,144	68.01
Singapore	1,170,000	2,619,376	44.67
Malaysia	1,946,698	15,676,135	12.42
Brunei	12,147	341,573	3.56
China	25,725,000	1,226,606,396	2.10
Thailand	1,081,920	53,478,502	2.02
Philippines	592,200	70,556,507	0.84
Indonesia	700,000	218,607,876	0.32
Total	**46,227,965**		
Source: (Gordon, 2005), the “Percentage” field is added by author.			

As Table 1 - 2 shows, Taiwan is the most representative country for the Southern Min language.

Shuan-fan Huang estimated that the percentage of Southern Min speakers in Taiwan was over 70%, while the population was about 17 million (Huang,

1995). Therefore, Taiwan has the highest percentage of Southern Min speakers in the world.

However, according to (Grimes, 2000), the previous edition of (Gordon, 2005), there were a total of about 49 million speakers. We do not know why the number of Southern Min speakers decreased by about 3 million from 2000 to 2005. Maybe it was due to a switch in identification. In addition, these number do not consider emigration from Taiwan to the US, Japan, *etc.* The reasons for emigration may be political or commercial factors.

According to (Grimes, 2000), if we list languages by the sizes of their speaking populations, Southern Min is ranked 21. It is thus an important language that has received very little attention in the world!

1.1.3 Another Investigation: the Taiwan Southern Min Viewers

Ún-giân Iûnn established a Taiwanese language website in 1998. The main contents of this site are related to the written form of Taiwan Southern Min (Iunn, 1998). We used Google Analytics to view the web traffic statistics, and found that this statistical data could help us to determine where the visitors come from.

The following data came from Google Analytics. It was calculated in May 2008. The “pages” and “total time” fields were added by us. The data is sorted in descending order by total time:

Table 1 - 3 Visits to the Taiwan Southern Min Contents Website

	Country/Territory	Visits	Pages/ per Visit	pages	Avg Time (sec)	total time	New Visits(%)
1	**Taiwan**	15,927	44.73	712,415	1,704	27,139,608	46.00
2	United States	683	40.94	27,962	2,312	1,579,096	43.63
3	Japan	337	68.97	23,243	2,336	787,232	34.42
4	**China**	1,105	9.37	10,354	482	532,610	75.20
5	Canada	119	113.45	13,501	2,733	325,227	52.94
6	Hong Kong	328	20.16	6,612	804	263,712	82.01
7	Netherlands	32	87.44	2,798	6,001	192,032	28.12
8	South Korea	51	107.53	5,484	3,064	156,264	47.06
9	**Singapore**	131	24.41	3,198	1,135	148,685	59.54
10	United Kingdom	64	23.94	1,532	1,536	98,304	67.19
11	Germany	49	45.35	2,222	1,741	85,309	48.98
12	**Thailand**	15	101.73	1,526	2,939	44,085	20.00
13	**Malaysia**	51	28.1	1,433	729	37,179	90.20
14	Australia	33	21.73	717	1,104	36,432	87.88
15	Panama	15	22.93	344	1,546	23,190	0.00
16	Bermuda	2	117	234	10,604	21,208	100.00
17	**Philippines**	31	6.97	216	478	14,818	41.94
18	Vietnam	19	19.05	362	729	13,851	100.00
19	Ireland	13	22.62	294	943	12,259	92.31
20	Macao	18	13.39	241	570	10,260	83.33
21	Spain	1	75	75	7,494	7,494	100.00
22	Austria	3	27	81	1,334	4,002	100.00
23	Brazil	2	39.5	79	1,950	3,900	100.00
24	Peru	7	30.57	214	543	3,801	100.00
25	Poland	5	11.6	58	580	2,900	40.00
26	**Indonesia**	6	6	36	344	2,064	100.00
27	New Zealand	8	5.12	41	252	2,016	100.00
28	France	16	3.31	53	94	1,504	100.00
29	**Brunei**	1	16	16	1,283	1,283	100.00
30	Belgium	3	7.33	22	377	1,131	100.00

Note: The data was counted in May 2008 by Google Analytics. The nation/ territory field is bold if it appears in Table 1 - 2.

Who are interested in websites where the contents are mainly in Taiwan Southern Min? We could not analyze the users in detail. Maybe they are:

(a) People who speak Southern Min;

(b) Taiwanese language researchers;

(c) Web robots, or someone else.

1.1.4 The Confusing Name of This Language

Khîn-huānn Lí pointed out that this language has 17 distinct names (Li & Ang, 2007). The divergent names reflect the status and situation of this language.

At first, foreigners called this language Amoy '廈門話' because they encountered it in the commercial harbor of Amoy. In the period of Japanese rule, the Japanese called the language Hokkienese '福建話', then Taiwanese '台灣話'. In the postwar era after 1945, the Chinese called it the Southern Min language '閩南語'. On the other hand, the Hakka people called it the Hok-lo language '福佬話'.

The official name is currently "Taiwan Southern Min" '台灣閩南語.' In civil society, people often call it Taiwanese. As mentioned in Table 1 - 2, since the language population is the majority in Taiwan, and Taiwan has the highest percentage population compared with other countries, it is adequate to refer to it as "Taiwanese."

However, someone may claim that calling it "Taiwanese" is prejudiced. That is why the name is so confusing.

We sought suggestions via Google search in relation to this problem. We searched for the following keywords to determine the number of web pages in Taiwan.

Table 1 - 4 The Names of “Taiwan Southern Min”

Name	Number of Web Pages
台語 ‘Taiwanese’	4,770,000
閩南語 ‘Southern Min’	680,000
台灣閩南語 ‘Taiwan Southern Min’	350,000
福佬話 ‘Hok-lo language’	229,000
Remark: Retrieved on June 1, 2008.	

As Table 1 - 4 shows, “Taiwanese” is also the term most often used on the internet. Therefore, we will call this language “Taiwanese” in this dissertation unless we wish to distinguish between it and some other native language of Taiwan, like Hakka or the Austronesian languages.

1.2 Different Types of Written Taiwanese Scripts

Taiwanese and Mandarin are different but related languages, differing in phonological, morphological, and syntactic features (H.-k. Tiunn, 1998). Liông-úi Tē^{n} (*aka* Robert L. Cheng) lists the following phonological and morphological characteristics of Taiwanese:

(a) Preservation of Ancient Chinese morphemes.

(b) Characters with distinct colloquial vs. literary readings.

(c) Taiwanese morphemes without standardized characters.

(d) Japanese and English loans, with most of the English loans being borrowed via Japanese.

(e) Loans that are written with Japanese characters but have Taiwanese pronunciations.

(f) Contractions (Robert L. Cheng, 1990).

The characteristics of Taiwanese, as listed above, should be taken into consideration in developing a written system for the language.

How many types of written Taiwanese systems are there? (Iunn & Tiunn, 1999) estimated that there were at least 64 systems in existence. These systems can be classified into 4 types: Han characters, phonetic symbols, Kana, and Romanized characters. In this section, we will introduce the main systems, including the Han character script, one of the Romanization scripts (Pe̍h-ōe-jī, vernacular writing, *abbrev*. as POJ) and the Han-Romanization mixed script. In addition, other systems will be briefly described.

1.2.1 The Han Characters Script

The earliest preserved work in the Han character script was published in 1566 and found in the Southern Min area (Gou, 1995). At that time, Han characters were primarily employed in the classical language, not in the service of the written vernacular. Also, the the songbooks 'koa-á-chheh 歌仔冊' were spread throughout Taiwan civil society in the 19th century (Klöter, 2005).

The above materials were not colloquial Taiwanese writing but only a special genre. The first complete colloquial writing in Han characters, "Doctrina Christiana" '基督要理,' was found in the Philippines (Iunn, 2007a, 2008).

The Han characters used for writing Taiwanese fall into four categories:

(a) Hùn-thȯk-jī ‘semantic borrowing characters 訓讀字’;

(b) Pún-jī ‘etymological characters 本字’;

(c) Chioh-im-jī ‘phonetic borrowing characters 借音字’;

(d) Pún-thó͘-jī ‘domestic character 本土字’ (H.-k. Tiunn, 1998).

The first Taiwan Southern Min recommended orthographic word list, which contained 300 words ‘臺灣閩南語推薦用字（第 1 批）300 字詞,’ was announced by the Ministry of Education on May 29, 2007 (MOE, 2007a). Siok-lîng Luā used 12 Taiwanese dictionaries to investigate the usage of Han characters, and found a total of 698 different usages, with an average of 2.33 Han characters used for a common Taiwanese word (Lua, 2008).

Another problem is that a Han character usually has two or more pronunciations. For example, when readers see the word “大人,” it is difficult for them to know whether to say “tōa-lâng” ‘adult,’ or “tāi-jîn” ‘policeman’ without the context. Since “ē/ōe ‘can’ ‘會’ ” was pronounced as “kōe” in the 19th century, if you write this word using the Han character “會,” others cannot determine the actual pronunciation. Worse still, the Han character “會” has seven different sounds in total (ōe/òe/kōe/kòe/hōe/hē/ē), according to the Online Taiwanese Syllable Dictionary (*abbrev*. OTSD) (Iunn, 2003c).

There is also kau-phòa, ‘a character with two or more pronunciations, 破音字’ in Mandarin, but it is more ambiguous in Taiwanese. Ún-giân Iûnn examined the OTSD, and found that there are a total of 22,080 entries and 11,635 distinct Han characters, with an average of two pronunciations per Han character.

However, when he only counted the common Han character set '常用字集' (5,401 characters in total), there were a total of 13,176 entries and 5,337 distinct Han characters, with an average of 2.5 pronunciations per Han character (Iunn, 2003d).

The advantage of the Han character script is that most people are educated in Han characters in Taiwan, without being formally taught the Romanized script, so they are often afraid of alphabetic writing. They are more willing to guess at the meaning when reading written Taiwanese in the Han character script than the Romanized script.

1.2.2 The Romanized Scripts

Among the dozens of Romanized scripts, three have been used as major systems in recent years. They are:

(a) POJ '*abbrev*. of Pe̍h-ōe-jī, vernacular writing, 白話字';

(b) TLPA '*abbrev*. of Taiwanese Language Phonetic Alphabetic, 台灣語言音標方案';

(c) TY '*abbrev*. of Tong-yong 通用' (Iunn, 2003a).

A new system, TL '台灣閩南語羅馬字拼音方案,' which is essentially a mixture of POJ and TLPA, was announced by the Ministry of Education on October 26, 2006 (MOE, 2006).

POJ is traceable to 1832. Historically, "Taiwan Church News" (originally "Tâi-ôan Hú-siâⁿ Kàu-hōe-pò" 'Taiwan Prefectural City Church News' '台灣府城教會報') was the first Taiwanese newspaper written in POJ. This paper was

founded in 1885. As the longest lasting newspaper in Taiwan, it is in the unique position of having documented Taiwanese society during a century of Manchurian, Japanese, and Chinese rule, and to have done so in a major language of the masses (J.-h. Tiunn, 2001).

The Han character script and Romanized script play complementary roles. Some disadvantages of the Han character script can be solved by the Romanized script, and vice-versa (H.-k. Tiunn, 1998).

1.2.3 The Han-Romanization Mixed Script

The idea of Han-Romanization mixed script writing was first introduced to Taiwan by Liông-úi Tē^{n} through publications using this system and the exposition of its theory in the late 1980s. The Han-Romanization mixed script is used in the writing of poems, novels, and prose, as well as in academic writing, Taiwanese textbooks, and religious works. It appears in newspapers, bulletins, and books. It is the writing system preferred by most of the advocates of written Taiwanese (Chhong-bi Memorial Foundation; Iunn, Tiunn, & Li, 2008).

For convenience, the “Han-Romanization mixed script” will be abbreviated as “HR mixed script” in the following sections.

1.2.4 Other Scripts

There are also other scripts that have appeared over the past several decades, most of which are oriented toward phonetic transcriptions. These scripts can be classified into four categories, including Romanized scripts, Kana

scripts '假名', Hangul scripts '諺文', and phonetic symbol scripts, like ㄅㄆㄇ (Iunn & Tiunn, 1999).

1.2.5 Target Scripts in This Dissertation

We intended to collect a sufficient amount of written Taiwanese material and then process it. We selected POJ and the HR mixed script as the target scripts in this dissertation. The Romanized script portion of the Han-Romanization mixed script was also POJ. In total, we have so far collected about 10 million syllables in these two scripts.

1.3 Issues Related to Written Taiwanese Processing

In order to process our two target scripts, it was necessary to understand some related issues.

(a) Written Taiwanese has not yet been standardized;

(b) Although MOE announced Taiwan Southern Min recommended orthographic word lists in May 2007 and May 2008 (MOE, 2007a, 2008b), the orthography of Han characters has not yet been standardized either;

(c) Some Han characters especially the *pún-thó͘-jī* 'domestic characters,' are not in the Unicode character set (The Unicode Consortium, 2006);

(d) All of the POJ characters have been in the Unicode set since 2004 (ISO/IEC JTC1/SC2 & WG2, 2004), with some of the characters being composed of two or three Unicode characters, but the characters are

separated into different zones, including basic Latin, Latin-1, Latin Extended-A, Latin Extended-B, and Latin Extended Additional (Lau, 2002);

(e) Using an internal plain text representation to record the Romanized script could be more convenient for searching;

(f) The word segmentation of the HR mixed script is more complicated than Mandarin because the usage of Han characters has not yet been standardized and this script is mixed with Romanization;

(g) Taiwanese tone sandhi is a difficult problem (M. Y. Chen, 2000; R. L Cheng, 1997), the Taiwanese corpus annotation recommends annotating the phonetic and tone sandhi markers;

(h) The technician who develops Taiwanese language related tools needs to interact with the people who are interested in the Taiwanese language in order to satisfy their needs.

1.4 Organization of This Dissertation

This dissertation is divided into six chapters.

We introduce the overall background in Chapter 1. A researcher with a background in computer science may not be familiar with the Taiwanese language, given the monolingual education in Taiwan. Therefore, we devote space to describing the background of the language, including its history, language population, different types of scripts, and abbreviations.

Chapter 2 describes the resources and our survey of written Taiwanese

processing. We omit the plentiful research results in the Mandarin and English fields for the sake of space. Written Taiwanese processing is an almost uncultivated field, and has received very little attention. Generally speaking, most journal editors are not interested in this field; therefore, we cite numerous websites rather than academic papers. In regards to the digital resources of written Taiwanese, we introduce fonts, dictionaries, corpora, electronic books, *etc*. We also introduce recent written Taiwanese processing techniques, including input method, word segmentation, tagging, script conversion, text-to-speech, translation, and parsing techniques.

In Chapter 3, we introduce the coding, I/O of POJ, and text processing for written Taiwanese. English and Mandarin have their own processing problems. For example, it is necessary to manipulate the word stemming problem and the modifier of a prepositional phrase in English processing, and the Han character encoding and word segmentation problem for Mandarin. As to POJ, it is necessary to solve some fundamental problems, including encoding, display, and search, which are not the same as English and Mandarin. We first introduce the POJ character code, and mention numbered POJ as the interchange code for various POJ encodings. Then, we propose a two-stage search strategy: perform string matching and then filter the results. In addition, we propose query expansions, including toneless, glottal stop, checked syllable, and vowel search, because it is difficult for someone with a Mandarin education to distinguish the differences. We also describe the display method for POJ, and some POJ word processing utilities, including phoneme segmentation, spelling checker, and

syllable/word/sentence count utilities. At the end of this chapter, we describe a word segmentation method for HR mixed script.

In Chapter 4, we propose a rule-based tone sandhi algorithm. We address some 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 took a sentence as the unit, translated every word into Mandarin via OTMD, and obtained POS information from the CED made by the CKIP group of the Academia Sinica. Using the POS data and tone sandhi rules formulated based on linguistics, we then tagged each syllable with its post-sandhi tone marker. Finally, we implemented a Taiwanese tone sandhi processing system that takes a POJ script sentence as the input and outputs the tone markers. Our system achieved accuracy rates of 97.4% and 89.0% with the observation and test data, respectively.

For example, if a user inputs the POJ sentence:

> “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 jo̍ah chiū ū jo̍ah, kôaⁿ chiū ū kôaⁿ”

Our tone sandhi algorithm adds the tone sandhi markers:

> “Chhin-chhiūⁿ án-ni# lâi kóng#, chāi lán Tâi-ôan# kīn-kīn h̍it-tiap& -á-kú# ê kang-hu#, ài soaⁿ# chiū ū soaⁿ#, ài hái# chiū ū hái#, beh$ jo̍ah# chiū ū jo̍ah#, kôaⁿ# chiū ū kôaⁿ#.”

We then concatenate all of the sound files for the corresponding syllables to

an MP3 format sound file and return it to the user. The purpose of the Taiwanese tone sandhi algorithm is to implement a real-time Taiwanese tone sandhi system.

In Chapter 5, we propose a POS tagging method using the OTMD and 10 million Mandarin words as training data to tag Taiwanese. The literary written Taiwanese corpora have both POJ script and HR mixed script, with genres that include prose, novels, and drama. We followed the tagset drawn up by CKIP. We developed a word alignment checker to assist with the word alignment work for the two scripts, and then used the OTMD to find the corresponding Mandarin candidate words, selected the most adequate Mandarin word from the Mandarin training data using an HMM probabilistic model, and finally tagged the word using an MEMM (Maximal Entropy Markov Model) classifier. We achieved an accuracy rate of 91.5% in the Taiwanese POS tagging work and analyzed the errors.

For example, the original data was a paragraph by paragraph parallel corpus with POJ and HR mixed scripts, like:

góa chiong chháu-bō-á kòa tī piah-téng, hêng-lí khêng khêng leh, chē tòa sió-tiàm ê tha-tha-mì téng-kôan, …	我將草帽仔掛tī壁頂，行李 khêng khêng leh，坐tòa小 店ê tha-thá-mì頂kôan，…

First, our word alignment program rearranged the data as:

“我 [góa] 將 [chiong] 草帽仔 [chháu-bō-á] 掛 [kòa] tī[tī] 壁頂 [piah-téng]，[,] 行李[hêng-lí] khêng[khêng] khêng[khêng] leh[leh]，[,] 坐[chē] tòa[tòa] 小店[sió-tiàm] ê[ê] tha-thá-mì[tha-tha-mì] 頂 kôan[téng-kôan]，[,] …”

Second, we referenced the OTMD and added the Mandarin translation(s) for every word. We called these Mandarin translation(s) candidate words. We performed this task because we intended to use the Mandarin language model:

“我[góa]{我} 將[chiong]{將} 草帽仔[chháu-bō-á]{@草帽仔} 掛[kòa]{帶;掛;戴} tī[tī]{在} 壁頂[piah-téng]{牆壁上}，[,]{，} 行李[hêng-lí]{行李} khêng[khêng]{收拾;盤點} khêng[khêng]{收拾;盤點} leh[leh]{咧} ，[,]{，} 坐[chē]{坐} tòa[tòa]{住} 小店[sió-tiàm]{@小店} ê[ê]{的} tha-thá-mì[tha-tha-mì]{塌塌米} 頂 kôan[téng-kôan]{上面} ，[,]{，} …”

Note that the words “草帽仔” and “小店” are not found in OTMD, we treat the HR mixed script as the Mandarin candidate word. Third, we use Hidden Markov Model to select the most suitable Mandarin word from the candidate words:

“{我}＜我＞ {將}＜將＞ {@草帽仔}＜草帽仔＞ {帶;掛;戴}＜帶＞ {在}＜在＞ {牆壁上}＜牆壁上＞ {，}＜，＞ {行李}＜行李＞ {收拾;盤點}＜收拾＞ {收拾;盤點}＜收拾＞ {咧}＜咧＞ {，}＜，＞ {坐}＜坐＞ {住}＜住＞ {@小店}＜小店＞ {的}＜的＞ {塌塌米}＜塌塌米＞ {上面}＜上面＞ {，}＜，＞ …”

Note that, since the words “草帽仔” and “小店” are not found in the OTMD, we treated the HR mixed script as the Mandarin candidate words. Third, we used the Hidden Markov Model to select the most suitable Mandarin word from the candidate words:

“＜我＞(Nh) ＜將＞(D) ＜草帽仔＞(Na) ＜帶＞(VC) ＜在＞(P) ＜牆壁上＞(Nc) ＜，＞(COMMACATEGORY) ＜行李＞(Na) ＜收拾＞(VC) ＜收拾＞(VC) ＜咧＞(T) ＜，＞(COMMACATEGORY) ＜坐＞(VA) ＜住

＞(VCL) ＜小店＞(Na) ＜的＞(DE) ＜塌塌米＞(Na) ＜上面＞(Ncd) ＜，(COMMACATEGORY)＞ …”

Finally, we got the Taiwanese POS tagging result:

“我[góa](Nh) 將[chiong](D) 草帽仔[chháu-bō-á](Na) 掛[kòa](VC) tī[tī](P) 壁頂[piah-téng](Nc) ，[,](COMMACATEGORY) 行李[hêng-lí](Na) khêng[khêng](VC) khêng[khêng](VC) leh[leh](T) ，[,](COMMACATEGORY) 坐[chē](VA) tòa[tòa](VCL) 小店[sió-tiàm](Na) ê[ê](DE) tha-thá-mì [tha-tha-mì](Na) 頂kôan[téng-kôan](Ncd) ，[,](COMMACATEGORY) …”

We hope that this POS tagging system can assist us to develop a Taiwanese parser.

A summary of our work will be given in Chapter 6. This dissertation is not the end of our work on written Taiwanese processing tasks. Chapter 6 will also propose future directions for written Taiwanese processing research.