

Chapter 2 Resources and Survey of Written Taiwanese Processing

This chapter introduces the digital resources for written Taiwanese, including 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.

2.1 Digital Resources for Written Taiwanese

2.1.1 Fonts

The available fonts include both Han character fonts and Taiwanese Romanization fonts. As mentioned before, the orthography has not yet been standardized. If someone chooses to write Taiwanese in Han characters, it is possible to select characters that exceed the range of the Unicode character set. Some Han characters, especially *pún-thó-jī* ‘domestic characters, 本土字’ like “囡 ‘they’ ‘他們,’ ” and “ㄟ ‘be unable to’ ‘不會’ ” *etc.*, are not included in the Unicode character set (The Unicode Consortium, 2006). If you insist on using

them, you must create your own fonts, and will encounter great difficulty when you communicate with others. This dissertation will not discuss this problem because solving it is beyond the scope of this study.

As to the POJ script, not all of the POJ character set was included in the Unicode character set before 2004. Before 2004, the POJ font makers replaced some character graphs that are unused in POJ with the specific POJ character graphs. For example, they replaced “ä” with “ā”. Some fonts, including “HoloWin,” “Taiwanese Serif,” *etc.*, are often attached to Taiwanese language word processing software (Lau & Iunn, 2002).

To date, most Romanization character fonts do not fully support the Unicode standard. The following fonts can support the POJ script: Taigi Unicode, Doulos SIL, Gentium, Charis SIL, DejaVu, *etc.* We adopted the Charis SIL font for this dissertation. These fonts can be downloaded from the Internet (Iunn, 2006b; Laenen, Jacquerye, & Kulev, 2004; Lau, 2005; SIL).

2.1.2 Dictionary

There are three online dictionaries available on the Internet. These will be introduced in this subsection.

(a) Online Taiwanese-Mandarin Dictionary (OTMD):

This Taiwanese-Mandarin translation dictionary was announced and has been online since November 2000. The main data provider is Liông-úi Tēⁿ, but many anonymous contributors also offer entries and correct the typos. This system is maintained by author. There are a

total of more than 62,000 entries. The URL is <http://iug.csie.dahan.edu.tw/q> .

This dictionary offers POJ, HR mixed script, and Mandarin fields, with the POJ field also offering different accents. The pronunciation function was added in 2006, and English translations were added to more than 10,000 entries in 2007 based on (Embree, 1984), which contains English, Mandarin, and POJ fields.

This data can be searched using either the Taiwanese language or Mandarin. Users can also look up entries using a toneless query expansion function. For example, if a user types “hoe-chhia,” they will find “hoe-chhia ‘vehicle decorated with flowers, 花車,’” “hóe-chhia, ‘railway train, 火車,’” and “hòe-chhia, ‘truck, 貨車.’”

This is a popular online system; there have been a total of more than 2.4 million searches from more than 125,000 different IP addresses since December 2002, with more than 2,700 searches per day for the past year. ¹

This dictionary has also been of assistance in many research works (Iunn, 2000, 2002, 2003g).

(b) Online Taiwanese Syllable Dictionary (OTSD):

This dictionary was announced and has been online since 2003. There are more than 22,000 entries, including POJ and Han character fields.

A pronunciation function was added in 2006. Its URL is

¹ Retrieved on December 30, 2008.

<http://iug.csie.dahan.edu.tw/TG/jitian/> .

This online search system offers four kinds of query expansions, including toneless, glottal stop, checked syllable, and vowel. There have been a total of more than 290,000 searches from more than 32,000 different IP addresses since January 2003, with more than 250 searches per day for the past year.² The POJ syllable query expansion will be discussed in Chapter 3 (Iunn, 2003c, 2003f).

- (c) Online Taiwanese-Japanese Dictionary with Taiwanese Translation ‘台日大辭典台語譯本’ (OTJDTT):

The Taiwanese-Japanese Dictionary was published in 1931 & 1932 and edited by Naoyoshi Ogawa ‘小川尙義’. There are more than 90,000 entries. In 2002, Chùn-iòk Lîm began re-typing some of the entries (more than 70,000) and translated the Japanese into Taiwanese. The quality of the entries is higher than the OTMD. The online search system was released in 2005 by Jer-min Tsai. Its URL is <http://taigi.fhl.net/dict/> (C.-C. Cheng, Ho, Hsiao, Chiang, & Chang, 2007; Iunn & Lau, 2007; Iunn et al., 2008).

- (d) MOE Taiwan Southern Min Common Word Dictionary ‘教育部臺灣閩南語常用詞辭典’:

MOE began to edit this dictionary in 2000. There are more than 20,000 entries. This dictionary has been online since October 2008 (MOE, 2008a).

² Retrieved on December 30, 2008.

2.1.3 Text Corpora

This subsection will omit the speech corpora since we are focusing on the written language.

- (a) The Texts Database of Folk Songs in Southern Min Dialect ‘閩南語俗曲唱本「歌仔冊」全文資料庫’:

Sūn-liông Ông released “The Texts Database of Folk Songs in the Southern Min Dialect” in 1999. He hired Chinese typists to enter the data and cooperated with Academia Sinica, who offered the online search system. For some reason they discontinued the online system after a few years. At present, users need to submit an application form to Sūn-liông Ông to use this corpus (Ong, 1999).

- (b) Taiwanese Wikipedia:

Taiwanese Wikipedia (*aka* Holopedia), which was established in 2003 and included in Wikipedia in 2004, could also be considered as a POJ corpus. There are more than 4,000 articles so far (Wikimedia Foundation, 2004).

- (c) Written Taiwanese Corpus:

Ún-giân Iúnn released a POJ corpus and a HR mixed script corpus in 2003. The genres include academic papers, novels, prose, poems, drama, folk literature, interviews, lectures, dialogs, fables, reports, *etc.* Afterwards he got NSC financial support from August 2004 to July 2005 to keep adding contents. He has collected more than 3.4 and 5.8 million syllables for these two corpora so far. He provided the Online

Taiwanese Concordancer System (OTCS), syllable/word frequency/mutual information/correlation reports *etc.* for these two scripts. There are a total of nearly 1,900,000 times of searches from more than 56,400 different IP addresses counted from January 2003, and about 1630 times of searches per day for past one year.³ (C.-C. Cheng et al., 2007; Iunn, 2003b, 2003e, 2005a, 2005b, 2005c; Iunn & Lau, 2007).

(d) Taiwanese Bible Corpus:

Faith Hope Love Information center has offered an online Taiwanese Bible since 2006. They offered the Barclay edition version (both New and Old Testament published in 1916 and 1933, respectively) and red-cover version (only the New Testament, which was translated in the 1970s and was confiscated by the Chinese Nationalist regime) (Faith Hope Love Foundation, 2006).

(e) Digital Archive Database for Written Taiwanese (2nd stage):

“The Collection and Cataloging of Taiwanese POJ Literature Data” (CCTPLD) ‘台灣白話字文學資料蒐集’ project was carried out by Heng-chhiong Li, under the auspices of the NMTL ‘台灣文學館,’ from May 2001 to December 2004. This project collected Taiwanese literature data and entered a portion of it into a database.

The “Digital Archive Database for Written Taiwanese” project was carried out by Cheng-yen Gao from September 2004 to December 2005. This project first recorded each syllable of the Taiwanese language,

³ Retrieved on December 30, 2008.

using a separate mp3 format sound file for each syllable. It then constructed a rule-based tone sandhi algorithm, and wrote a program to concatenate all of the corresponding sounds based on the user input (in numbered POJ) and tone sandhi algorithm. The core problem of this project was the tone sandhi. The tone sandhi algorithm will be explained in Chapter 4 (Iunn, Lau, Tan-Tenn, Lee, & Kao, 2007).

The follow-up project, called the “Digital Archive Database for Written Taiwanese (2nd stage)” (DADWT), was carried out by Ún-giân Iunn, from February 2006 to December 2006. This project integrated the results of the above two projects and continued adding content to create a website, which includes both POJ and HR mixed scripts, using paragraph alignment to display the texts, and provides speech synthesis for each paragraph. It contains a total of about 2,580,000 syllables. There have been a total of more than 1,320,000 page visits since December 2006, with 1,672 page visits per day on average⁴ (Iunn, 2006a, 2007b).

- (f) Taiwanese Vernacular Literature Archive(TVLA) ‘台灣白話字文獻資料館’:

This website is the result of an ongoing project, called the “Taiwan Church News Vernacular Literature Digital Archive Project, 1885~1969 ‘台灣教會公報白話字文獻數位典藏計畫，1885~1969,’” carried out by Khîn-huānn Lí under the auspices of the Content

⁴ Retrieved on December 30, 2008.

Development Division of the Taiwan National Digital Archives Program
'數位典藏內容開發分項' since March 2007.

This project selects articles written in POJ and published in the "Taiwan Church News" from 1885 to 1969, enters the text in POJ, transcribes it into HR mixed script, and scans the images. So far they have entered the text of more than 1,000 articles and scanned more than 1,000 images (B.-c. Li, 2008; K.-h. Li, 2008; Li, Li, & Lau, 2008; Tan, 2008).

(g) Ongoing and others:

The "Taiwanese and Hakka Modern Literature Website" (THMLW) project was carried out by Ūi-bûn Chiú", under the auspices of the Council for Cultural Affairs, from February 2007 to December 2007. This project collected the selected works of 51 writers, for a total of about one million syllables. The content of this website is still being updated (Chiunn, 2008).

The "Southern Min and Hakka Language Archive '閩客語典藏' " (SMHLA) project is being carried out at present by Min-hua Jiang, under the auspices of the Academia Sinica. It began in January 2007 and is scheduled to continue until December 2011. For the Southern Min, they intend to scan and retype some books, including "Doctrina Christiana '基督要理,' " "Chinese-English Dictionary of the Vernacular or Spoken Language of Amoy '廈英大辭典,' " "English and Chinese Dictionary of the Amoy Dialect '英廈辭典,' " "Taiwan Church

News ‘台灣教會公報,’ ” *etc.* We will be involved in this project to help with the POS tagging (Academia Sinica, 2008).

The “Taiwan Child Language Corpus ‘台灣兒童語料庫’ ” (TAICORP) project was carried out by Jane Tsay under the auspices of the National Science Council ‘國家科學委員會’ from August 1997 to July 2000 (data collection). They collected speech samples from 14 children and transcribed these into text with syntax tags that were modified with the tagset of the CKIP group. This text consists of a total of 1,646,503 words (2,097,400 syllables). They also performed some related researches, but there is no online search system for this corpus (Jane S. Tsay, 2005; Jane S. Tsay, 2007).

There are also other small-scale corpora scattered in other places. It is necessary to integrate these resources for the promotion of related researches.

2.1.4 Electronic Books

Taiwanese electronic books are not directly related to written Taiwanese. Nevertheless, it is difficult to find written Taiwanese books in bookstores, and many of these books are now out of print.

Therefore, electronic books are a very important source of written Taiwanese, since the easier it is to find and utilize materials, the more helpful such materials are in written Taiwanese related studies.

Some websites provide Taiwanese electronic books, such as (Taichung

Library).

In June 2007, Ún-giân Iunn initiated a project called “The Memory of Written Taiwanese,” which has collected 567 book or periodical pages, and has so far had a total of more than 1,280,000 page visits⁵ (Iunn, 2007a, 2008).

2.2 Survey of Written Taiwanese Processing

Techniques

2.2.1 Input Method

A word processor is a computer application used for the production of any sort of printable material. Word processors are the bases for written Taiwanese texts, and Taiwanese input methods are the bases for Taiwanese word processors.

Liông-úi Tēⁿ developed TW301 in DOS. Chi-bêng So⁷ developed HOTSYS-HAKSYS in the Windows operating system (updated to Windows ME). These two Taiwanese input systems provided both POJ fonts (ascfont and HoloWin, respectively) and input methods. The above two systems played a very important role in the revitalization of written Taiwanese in the 1990s; however, they are now out of date (Iunn, 2001).

Yuang-chin Chiang developed the “Dai-im input method ‘台音輸入法.’” This tool provided pronunciation, but the Romanization script was not the same

⁵ Retrieved on December 30, 2008.

as POJ, and the Dai-im symbols were in conflict with POJ (Chiang, 2004).

Chùn-iòk Lîm developed the “KKS ‘Kài khin-sang’ ‘介輕鬆’ input method” in 2005. He used the phrase input function built into Microsoft Windows to provide more than 10,000 entries. Other similar input methods include “SKS ‘Siōng khin-sang’ ‘上輕鬆’ ” provided by Pêng-tī Siau in 2007 and “YKS ‘Iah khin-sang’ ‘亦輕鬆’ ” (related to Yahoo!), which was also provided by Chùn-iòk Lîm in 2008 (Lim, 2005, 2008; Siau, 2007).

On October 14, 2006, the Ministry of Education announced the TL “Taiwan Southern Min Lō-má-jī (Romanization) phonetic scheme ‘臺灣閩南語羅馬字拼音方案,’ ” which is compatible with POJ script. Pek-tiong Tân developed the “TL input method ‘台羅輸入法’ ” in 2007 under the auspices of the MOE. This is a cross-platform input method (MOE, 2006, 2007b).

There are also other input methods. Jason Cox and It-kûi Tân developed the “POJ input method” under the Mac platform in 2002, and Pek-tiong Tân and Teng Lâu developed the “Open Vanilla POJ input method” in 2005. The Firefox add-on “Transliterator (ToCyrillic)” included a POJ input method, which allows users to type POJ in the Firefox browser environment (Benenson; Iunn *et al.*, 2008).

2.2.2 Word Segmentatation

Word segmentation is a basic research tool. In 1997, Kim-kim Chan discussed a principle for Taiwanese word segmentation (K.-k. Chan, 1997). However, so far, no one has established a dictionary following the principle she

proposed.

Based on OTMD, Kiát-gák Lâu and Ún-giân Iûⁿ developed an online word segmentation system for Taiwanese HR mixed script. They used the backward maximal matching algorithm to implement the system (Lau & Iunn, 2007).

2.2.3 POS Tagging

Si-yuan Chou used the Brill tagger based on the HMM model to tag words in a T3 treebank. They used a tag set size of 26, and attained tagging accuracy rates of 92.80% and 85.59% for the training and test data, respectively (Chou, 2006).

2.2.4 Scripts Conversion

This subsection will focus on the conversion between POJ script and HR mixed script.

Jer-min Tsai developed systems to convert between POJ script and HR mixed script based on the dictionary. They announced that the systems achieved an accuracy rate of over 90% for religious domain texts (Lim Chun-iok, 2006; J.-m. Tsai).

Un-gian Iunn and Beng-han Lui also developed conversion systems between POJ script and HR mixed script based on the OTMD and the mutual information trained from written Taiwanese corpus (Iunn & Lui, 2009).

2.2.5 Text-to-Speech

Before we can successfully transform written Taiwanese text into its natural

speech-like tonal contour, tone sandhi presents a challenging problem to be solved. This is because the POJ script represents the tones as “basic tones,” or the tones of syllables when they are pronounced in isolation. At the word level, almost all of the syllables except the last one are usually pronounced differently (that is, they manifest tone sandhi). At the level of an entire sentence, only the last syllables by the boundary of the phrases or structural markers are read as basic tones in most situations, with the others being read as sandhi tones. In fact, besides the "regular tone sandhi" mentioned above, there are still other kinds of tone sandhi phenomena, which will be discussed in Chapter 4 (M. Y. Chen, 2000; R. L. Cheng, 1997; Iunn et al., 2007).

Chuan-jie Lin and Hsin-hsi Chen described an early tone sandhi system (Lin & Chen, 1999). Users input Mandarin text, and the system outputs Taiwanese text with the pronunciation. The corpus is news reports in Mandarin. They used the word segmentation and tagged data of the CKIP group and the Taiwanese-Mandarin dictionary, which was provided by Liông-úi Tēⁿ, to map the Mandarin news into Taiwanese (in both Han-Romanization mixed and POJ scripts). The sandhi rules applied were as follows:

- (a) Pronounce the last syllable at the end of a sentence as a basic tone;
- (b) Pronounce the syllable before the particle “ê” as a basic tone;
- (c) Pronounce the last syllable of a noun as a basic tone;
- (d) Pronounce other syllables as normal sandhi tones.

An accuracy rate of 82.53% was reported.

Liang *et al.* describes a text-to-speech system for Taiwanese (Liang, Yang,

Chiang, & Lyu, 2004). Its input was a large corpus of Mandarin news texts, with sentences longer than 20 syllables removed. It utilized a dictionary to convert the Mandarin 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 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.

Pan, Neng-huang *et al.* also implemented a text-to-speech system for Taiwanese (Pan, Yu, & Tsai, 2008). The input is a Mandarin sentence. The system segments the sentence into words using a Mandarin dictionary with about 130 thousand words, and then it translates Taiwanese using a Mandarin Taiwanese bilingual dictionary with about 37 thousand entries, finally outputs the sound. A Mandarin unknown word will be separated into two or more words with less syllables in order to match the entry listed in the dictionary. A Taiwanese unknown word will also be separated into syllables for matching the corresponding syllable of other words to find the pronunciation. For example, “玉山” ‘Mt. Jade’ is an unknown word. The pronunciation then is found from the syllable ‘玉’ of word ‘玉女’ and the syllable ‘山’ of word “入山”. The system will output the sound “gék-soaⁿ” for this unknown word “玉山”.

The size of test data is 221 syllables. There are 22 translation errors and 28 tone sandhi errors. The accuracy rate is 77.38%.

2.2.6 Translation

This subsection will only focus on Mandarin to Taiwanese translation. The above three text-to-speech systems also provide translation functions (Liang et al., 2004; Lin & Chen, 1999; Pan et al., 2008).

Jer-min Tsai provided an online Mandarin to Taiwanese translation system using a Mandarin-Taiwanese dictionary. The main purpose was to semi-automatically generate written Taiwanese text (Lim Chun-iok, 2006; J.-m. Tsai).

2.2.7 Parsing

(Shi, 2006) described the Taiwanese Treebank using a Brill parser. Based on the Mandarin sentences in the book, “Modern Chinese 800 words ‘現代漢語八百詞’ ” by Shu-xiang Lyu ‘呂淑湘,’ they translated into Taiwanese and Hakka to establish the T3 corpus with 23 tags, developed some editing tools to help in the construction of the T3 treebank, and proposed the dotted tag method for phrase tagging to make the parsing tree flatter. They also used a Brill parser, which is an error-driven transformation-based parser, to improve the results (Brill, 1993; Liu, 2005; Shi, 2006).

2.3 Summary

We have introduced the existing electronic resources for written Taiwanese known to us. We have also surveyed written Taiwanese processing techniques.

Compared with Mandarin or English, we think that the results are limited and insufficient. However, these results are the solid foundations that we will use to move forward.