台語變調系統實作研究

A Study on Implementation of Taiwanese Tone Sandhi System

楊允言 Iûⁿ Ún-giân[1],李盛安 Li Sheng-an[2],劉杰岳 Lâu Kiát-gák[3],高成炎 Kao Cheng-yan[4]
國立台灣大學資訊工程系 台北 台灣
d93001[1] d93005[2] cykao[4] @csie.ntu.edu.tw kiatgak[3]@gmail.com

摘要

台語羅馬字在過去近兩百年來，累積了數量相當可觀的文本，然而目前能流利閱讀台語羅馬字者並不多，使這些資料的利用價值大大降低。

本文主要處理台語的變調問題，實作出台語變調系統。我們採用台語羅馬字書寫的台文語料，以句子爲單位，透過台華對譯辭典找出中文翻譯，再從中研院資訊所詞庫小組的八萬目辭典中取得詞類訊息，接著利用我們訂出的變調規則，標記出每個音節的變調註記。台語變調情形有很多種，文中也有較詳盡的敘述。

研究結果顯示，訓練語料得到 97.56%的變調正確率，測試語料則有 88.90%的變調正確率。我們討論了錯誤的原因，希望持續做改進，以達到更高的正確率。

關鍵詞 台語文 Written Taiwanese、變調規則 Tone Sandhi Rule、台語羅馬字 Taiwanese Romanization

1 背景說明

在台灣，台語是日常生活中常被使用的語言，而台語書面語則較不常見；雖然如此，台語書面語已有百年以上的歷史。[Tiuⁿ 2001]目前台灣社會則存在數十、甚至超過百套的台語書寫系統。[Iûⁿ & Tiuⁿ 1999]本文採用的書寫系統，爲台語羅馬字（又稱爲白話字、教會羅馬字）。

根據國家台灣文學館籌備處委託成功大學台灣文學系所執行的「台灣白話字文學資料蒐集整理計畫」，[1]雖然歷經政治變局，許多資料已遺失，但在該計畫的努力下，仍蒐集到近兩千種台語羅馬字相關書刊，出版地也遍及台灣、廈門、上海、廣州、香港、新加坡、菲律賓、倫敦、日本、...等地；根據蒐集到的出版品其出版年代的統計，1950、1960 年代是相關書刊在台灣出版的高峰。除了正式出版的書刊，也有一般民眾的書信、醫師所寫的患者病歷資料等使用台語羅馬字。然而後來政府以阻礙國語推行爲由強加禁止，導致台語羅馬字的急速式微。

上列計畫成果，我們希望透過資訊科技，讓更多人能夠運用這些資料，促成台語文的基礎或

[1] 計畫主持人爲呂興昌教授，執行期間是 2001.5~2004.12。

應用研究。鑑於一般人對於台語羅馬字並不熟悉，如果利用語音合成技術，將這些文字資料轉成語音，可以讓這些寶貴的資料獲得更高的使用價值。

不過台語的文字要轉成聲音，最大的挑戰在於台語的變調問題。台語羅馬字表記本調，實際發音時，在語詞的層次，大多數情形是最後一音節讀本調，其餘讀變調。然而在句子的層次，大部分情形是在詞組或是結構標誌的分界處的最後一音節讀本調，其餘讀變調（包含語詞的最後一音節也讀變調）。實際上，除了規則變調外，變調又有好幾種情形，本文將一一討論。

本文將重點放在變調規則的處理，這將是其後輸出的語音、聲調是否正確的主要關鍵。本文主要採用上列計畫所蒐集到的語料做爲輸入，實作台語變調系統，輸出爲輸入的語料加上變調註記。由於沒有正確變調結果的訓練語料集，我們暫時不採用統計學習的方法，而是用規則式學習，並由本文其中兩位長久參與台語文工作的作者，來評估變調結果正確率。[2]

2 台語變調說明

台語的聲調，依據傳統的說法，平、上、去、入分陰陽，但上聲不分，所以共有七個聲調，根據陰平、上、陰去、陰入、陽平、陽去、陽入的順序，[3]用數字表示分別是 1(高平)、2(高降)、3(低)、4(中短)、5(低升)、7(中平)、8(高短)。刮號中描述調值。聲調符號請參酌下面的例子。

變調是台語非常重要的特色。在語詞層次，通常最後一音節讀本調，其餘讀變調。下例中的五個語詞，畫底線者讀本調，其餘則讀變調：[4]

例 1 tâi 台 / Tâi-gí(gú)台語 / Tâi-gí(gú)-bûn 台語文
Tâi-gí(gú) bûn-hák 台語文學 / Tâi-gí(gú) bûn-hák-sú 台語文學史

實際上，在音節或語詞的層次，台語變調至少包括下列幾種：

(1) 規則變調：以疊詞做說明，刮號內的數字爲實際讀出的聲調。

例 2 (i) 1 聲→7 聲：如「chheng-chheng 清清」(7,1)
(ii) 7 聲→3 聲：如「chēng-chēng 靜靜」(3,7)
(iii) 3 聲→2 聲：如「chhiò-chhiò 笑笑」(2,3)
(iv) 2 聲→1 聲：如「léng-léng 冷冷」(1,2)
(v) 5 聲→7 聲或 3 聲(台北)：如「âng-âng 紅紅」(7/3,5)
(vi) 4 聲→8 聲(-p/t/k)或 2 聲(-h)：如「sip-sip 濕濕」(8,4)「phah-phah 打打」(2,4)
(vii) 8 聲→4 聲(-p/t/k)或 3 聲(-h)：如「直直 tit-tit」(4,8)「熱熱 jòah-jòah」(3,8)

(2) 隨前變調：一般爲代名詞或人名的後綴，前面一音節讀本調，此音節的聲調視前面聲調而定，爲 1 或 3 或 7 聲。

例 3 (i) 「A-eng--a 阿瑛 a」(7,1,1) (第二個"a" 是後綴)
(ii) 「góa lâi khòaⁿ -- i 我來看伊」(1,7/3,3,3) (「i 伊」原來是第 1 聲)
(iii) 「hō--lí [給]你」(7,7)(「lí 你」原來是第 2 聲)

[2] 本文第一作者從事台語文工作將近 20 年，第三作者將近 10 年，除了發表技術性論文及相關軟體開發，也有台語刊物編輯、台語文章創作等各方面相關的經歷。
[3] 這樣的順序，方便與漢語系其它語言（如客語等）做對應。
[4] 「台語文學」和「台語文學史」是一個詞還是兩個詞（「台語 / 文學」、「台語 / 文學史」）也許仍值得商榷，在這裡我們暫時視爲一個詞。

(3) 輕聲：輕聲前讀本調，輕聲的部分讀 3 聲或 4 聲(入聲)。

例 4 (i) 「Tân--sian-siⁿ(sin-seⁿ)陳先生」(5,3,3)(「sian-siⁿ(sin-seⁿ)先生」原來聲調是 7,1)
(ii) 「kiâⁿ--chhut-lâi 行出來」(5,4,3) (「chhut-lâi 出來」原來聲調是 8,5)

(4) 再變調：多半出現在喉塞音(-h) 4 聲，規則變調兩次(4→2→1)。

例 5 (i) 「beh thȧk-chu[要]讀書」(1,4,1) (「beh[要]」4 聲應變 2 聲，實際變 1 聲)
(ii) 「khì gōa-kháu 去外口」(1,3,2) (「khì 去」3 聲應變 2 聲，實際變 1 聲)

(5) á[仔]前變調：á 前的音節，只有 1、2 聲同規則變調，其餘不同。

例 6 (i) 1 聲→7 聲：如「sun-á 孫仔」(7,2)
(ii) 2 聲→1 聲：如「chháu-á 草仔」(1,2)
(iii) 3 聲→1 聲：如「tàⁿ-á 擔仔」(1,2)
(iv) 4 聲→8 聲(-p/t/k) 或 1 聲(-h)：如「tek-á 竹仔」(8,2)「thih-á 鐵仔」(1,2)
(v) 5 聲→7 聲：如「lô-á 爐仔」(7,2)
(vi) 7 聲→7 聲：如「phō-á 簿仔」(7,2)
(vii) 8 聲→4 聲(-p/t/k)或 7 聲(-h)：如「chhȧt-á 賊仔」(4,2)「hio̍h-á 葉仔」(7,2)

(6) 三連音變調：三連音疊詞的第 1 音節，2、3、4 聲同規則變調，其餘不同。

例 7 (i) 1 聲→5 聲：如「chheng-chheng-chheng 清清清」(5,7,1)
(ii) 2 聲→1 聲：如「ún-ún-ún 穩穩穩」(1,1,2)
(iii) 3 聲→2 聲：如「hèng-hèng-hèng 興興興」(2,2,3)
(iv) 4 聲→8 聲(-p/t/k)或 2 聲(-h)：如「sip-sip-sip 濕濕濕」(8,8,4)
「bah-bah-bah 肉肉肉」(2,2,4)
(v) 5 聲→5 聲：如「kôaⁿ-kôaⁿ-kôaⁿ寒寒寒」(5,7/3,5)
(vi) 7 聲→5 聲：如「chēng-chēng-chēng 靜靜靜」(5,3,7)
(vii) 8 聲→5 聲：如「tit-tit-tit 直直直」(5,4,8)「pe̍h-pe̍h-pe̍h 白白白」(5,3,8)

(7) 升調：通常發生在日語借詞，變調後是 5 聲。

例 8 「ŏai-siak-chù [白襯衫]」(5,8,3)「khăn-páng[看板]」(5,2)「hăn-tó-lù [方向盤]」(5,1,3)[5]

3 文獻探討

在台灣，從事台語文計算語言學的研究團隊，包括長庚大學資訊系呂仁園教授主持的多媒體訊號處理實驗室、[6]台大資訊系陳信希教授主持的自然語言處理實驗室、交大電信系陳信宏教授主持的語音處理實驗室、成大電機系王駿發教授主持的多媒體通訊 IC 系統設計實驗室、成大資訊系吳宗憲教授主持的多媒體人機通訊實驗室、台大資訊系高成炎教授主持的台語文研究室、... 等。以下探討有提出變調正確率的兩篇論文。

[Lîm 1997]是較早實作的台語變調系統，輸入是中文，輸出是台語文及發音，語料爲中文新聞資料，利用中研院資訊所詞庫小組的斷詞、標記結果，以及鄭良偉提供的台華對譯辭典，用資料庫查詢華文所對應的台文，台文有漢字及台語羅馬字。變調規則用到： a)句尾讀本調 b) "ê[的]" 前讀本調 c) 名詞詞尾讀本調 d)其餘規則變調。其變調的正確率有 82.53%。不過系統並非使用台語文做爲輸入，將中文轉成台文時，語詞排列順序及詞義排歧並沒有處理，翻譯出來的台語文，

[5]正式的台語羅馬字聲調符號並不包括 ŏ ă，我們參酌[Tiuⁿ 2001]採用此符號。
[6] 清大統計所江永進教授也加入此研究團隊。

與實際 Native Speaker 所講的台語有些差距。儘管如此，仍然是先驅性的台語變調實作系統。

[Liang et. 2004] 是最近發表的台語語音合成實作系統，輸入爲大量的中文新聞語料，去除多於 20 個音節的句子，利用辭典轉成台文後，經過斷詞、標記發音、做變調規則處理後轉成聲音檔並播放。因爲資料量大，所以挑選前兩百句，由兩位台語專家做正確率的評估，結果是：斷詞正確率超過 97%，標記發音有 89%的正確率，變調規則處理則有 65%的正確率。

本文所實作的系統與上述論文所提的系統，主要的不同點在於，我們採用台語文語料，文學類與非文學類大致各半，沒有觸及中文翻譯成台文的問題，且任何長度的句子都處理；另外，因爲使用台語羅馬字，也少了斷詞及標記發音的問題。不過，與漢字相比，發音相同的歧異語詞數大大增加，特別是單音節語詞，這是較大的挑戰。

4 研究步驟

4.1 語料

本文所採用的台語文語料以台語羅馬字書寫，台語羅馬字的書寫形式是以詞爲單位，同一個語詞的音節以連字符(hyphen)連接，語詞和語詞間以空白間隔。

語料由上列計畫提供。訓練語料部分，我們挑選四本書，分別是：

- 1913 年出版的《Sin-bûn ê chảp-lio̍k 新聞的雜錄》(不知作者，類別：報導)；[7]
- 1924 年出版的《Chảp-hāng kóan-kiàn 十項管見》(作者: 蔡培火, 類別: 論述)；[8]
- 1955 年出版的《Chháu-tui téng ê bîn-bāng -- jî-tông chong-kàu kò-sū 草堆頂的眠夢—兒童宗教故事》(作者：黃懷恩，類別：小說) ；
- 1961 年出版的《Tang-pō thôan-tō kiàn-bûn kì 東部傳道見聞記》(作者：陳降祥，類別：報導)[9]

上述語料涵蓋日本時代和國民政府時代，每本書挑出兩段，總共 614 音節。

測試語料除了上列計畫所提供的之外，部分來源爲我們蒐集的台語羅馬字語料。同樣挑選四份資料，分別是：

- 1885 年出版的《Pe̍h-ōe-jī ê lī-ek 白話字的利益》(作者：葉牧師，類別：論述)[10]
- 1905 年出版的《Kau-chiàn ê Siau-sit 交戰的消息》(作者：教會公報編輯室，類別：報導)[11]

---

[7] 雖然的報導類，不過其寫作風格與現在一般的報導文章不盡相同，我們挑選的兩段，都是以第一人稱在敘述，看起來很像一般的散文。

[8] 蔡培火是日本時代台灣政治社會運動的重要人物之一，文化協會成立後，他曾寫文章、演講、辦夏季學校來鼓吹台語羅馬字的使用，可惜其理念不見容於台灣總督府。戰後，蔡培火加入國民黨，成爲籠絡台灣人的象徵，位高卻無權。有興趣者可參考台灣史料基金會出版的《蔡培火全集》http://www.twcenter.org.tw/a02/a02_08/a02_08_01.htm。

[9] 對整個台灣來說，東部是相對特殊的地方，漢人比例少，清國時代只有領台最後 20 年才真正管理東部，日本時代有計畫地移入大量日本人；因爲邊陲，對於東部的（漢字、日文）文字論述多只有官方立場，反而台語羅馬字書寫的相關東部消息，提供比較貼近尋常百姓的觀點。

[10] 這篇文章在談論台語羅馬字和「孔子字」（漢文）的優缺點，是篇擲地有聲的論述，出處爲台灣府城教會報（目前的名稱是台灣教會公報）。

[11] 這篇文章在談日俄戰爭，出處也是台灣教會公報。討論日本時代的台灣政治社會運動，會提及台灣人所辦的第一份刊物是 1920 年的《台灣青年》，當時還無法在台灣發行。如果也把台語羅馬字文獻也考慮進來，這些政治社會運動的論述

- 1954 年出版的《Thiàⁿ lí iâⁿ kè thong sè-kan 疼你贏過通世間》(作者：賴仁聲，類別：小說)[12]
- 1997 年發表在 BBS 的《Ài lí kap ài i pîⁿ-á chōe 愛妳及愛伊平仔多》(作者：盧誕春，類別：散文)

同樣是涵蓋兩個時代，年代的範圍更廣。

4.2 詞類標記

詞類標記的部分，由於目前尚未有台語詞類的標準，我們只好暫時借用華語的成果。我們將語料，以語詞爲單位，查詢台文華文線上辭典（有台語羅馬字、台語漢羅寫法、華語對譯、查詢頻率、...等欄位），對應到華語的詞彙後，再從中研院詞庫小組的八萬詞目資料中找到此語詞的詞類標記。這裡會遇到歧義(ambiguity)問題，包括：

(a) 同音詞，特別是單音節同音詞；
(b) 台語對華語的翻譯是一對多；
(c) 華語語詞本身有多重詞類。

同音詞的問題，我們只取查詢頻率最高的，[13]實際檢視資料發現，在大部分情形下是對的；因爲一個台語詞可能對應到多個華語詞，而且一個華語詞本身可能有多重詞類，因此一個台語詞可能會有多重詞類，目前在詞類標記階段我們皆加以保留，留待變調標記階段取捨。八萬詞目的詞類標記，是簡化後的詞類標記，有 46 個詞類，實際上我們只取一層，其中某些詞類做了些調整（第二層的訊息、且會影響變調結果的詞類），如將 Vh（狀態不及物動詞、狀態使動動詞）改成 A（形容詞），Nh（代名詞）改成 R，Ng（後置詞）改成 G，Nd(時間詞)改成 S。

至於未知詞，如果是「XX」或「XXX」（音節重覆）的形式，我們暫時標記成 A(形容詞)，其餘標記爲 N(名詞)。

所以，我們所用的詞類標記包括：A 形容詞、C 連接詞、D 副詞、G 後置詞、I 感嘆詞、M 特別標記、N 名詞、P 介詞、R 代名詞、V 動詞、S 時間詞、T 語助詞等 12 個標記。

4.3 變調規則

變調規則是本研究最重要的部分。目前的變調規則演算法請參考表 1：

表 1 變調規則演算法

```
1   變調註記 lóng 先填 t (規則變調)
2   Siōng 尾一個改做本調
3   (語詞層次)"ê" ê 處理：kā 頭前 ê 詞尾改做 #(本調)
```

也許有需要改寫。

[12] 賴仁聲在 1920 年代曾出版兩本台語羅馬字的小說《A-niâ ê bák-sái 阿娘的目屎》及《Khó-ài ê siû-jîn 可愛的仇人》。1950 年代出版的這本書是 2002 年才「出土」的。台語文學史上應有其重大意義。

[13] 一般的詞頻是指某語詞出現在語料中的頻率，這裡的查詢頻率指的是此語詞在線上辭典被使用者查詢到的次數，使用者查詢方式，包括利用台語羅馬字、台語漢字以及華語等三種。

4 (詞類層次) chhē A/A Pair (無 ambiguity ê 情形)

4.1 chhē A / A Pair (頭前 A、後壁 A) ：kā 頭前 A ê 詞尾改做 #(本調)

5 (詞類層次) chhē N/V N/A N/P N/R N/D Pair (無 ambiguity ê 情形)

5.1 Chhē N / V Pair (頭前 N、後壁 V)：kā N ê 詞尾改做 #(本調)

5.2 Chhē N / A Pair (頭前 N、後壁 A) ：kā N ê 詞尾改做 #(本調)

5.3 Chhē N / P Pair (頭前 N、後壁 P) ：kā 頭前 N ê 詞尾改做 #(本調)

5.4 Chhē N / R Pair (頭前 N、後壁 R) ：kā 頭前 N ê 詞尾改做 #(本調)

Chhē N / D Pair (頭前 N、後壁 D) ：kā 頭前 N ê 詞尾改做 #(本調)

6 (詞類層次)C(連接詞) ê 處理

詞類是 C，kā 頭前 hit 個詞 ê 詞尾改做 #(本調)

7 (詞類層次) G(後置詞) ê 處理

詞類是 G，kā 頭前 hit 個詞 ê 詞尾改做 #(本調)，G 這個語詞 ê 詞尾 mā 改做#(本調)

8 (詞類層次) S(時間詞) ê 處理

詞類是 S，kā 這個語詞 ê 詞尾改做#(本調)

9 (語詞層次) R(代名詞) "góa / lí / i / gún|góan / lán / lín / in" ê 處理(愛 tī 第三條後壁)

9.1 語詞是"i / in"包括 tī句尾：變調註記改做 "t"(規則變調) (這條以後需要進一步討論)

9.2 語詞是"góa / lí / gún|góan / lán / lín "而且無 tī句尾：變調註記改做 "t"(規則變調)

10 (語詞層次)句尾 "kóng[講]" ê 處理：假使 Delimeter 是[，, ：:"] ，而且"kóng" ê 頭前 ê 語詞（ùi 頭前一直檢查 kàu 句首）有發現詞類是 R(代名詞) ê，變調註記改做 "t"(這個規則以後需要進一步修改，處理人名(Unknown Word)出現 tī頭前 ê 情形)

11 (音節層次)á 前變調處理

所有 ê 內容，若有包含"á"而且m̄是 tī詞頭 ê，將 "á"頭前 hit 個音節的變調註記改做 & (á 前變調)

12 再變調處理

12.1 (音節層次)"beh" ê 再變調處理：假使"beh" m̄是出現 tī句尾，變調註記改做 $(再變調) (包括 tī詞內底 ê "beh",親像 "強 beh/ tih-beh/愛 beh" ...)

12.2 (語詞層次) "khì[去]" ê 再變調處理："khì[去]"若m̄是出現 tī詞尾，伊 ê 後壁 ê POS 若是 N iah 是 V，tō 標記$(再變調) (這個規則以後可能需要進一步修改)

12.3 (音節層次)"koh" ê 再變調處理：假使"koh" m̄是出現 tī句尾，變調註記改做 $(再變調) (包括 tī詞內底 ê "koh",親像 "koh 再/ chiah-koh/iáu-koh" ...)tō 標記「再變調」(這個規則以後需要進一步修改)

12.4 (語詞層次)"kah" ê 再變調處理：假使"kah" m̄是出現 tī句尾，變調註記改做 $(再變調)

13 (語詞層次)輕聲處理：若是語詞內底出現"--" ê 所在，kā "--"頭前 ê 第一個音節標記本調，"--" 後所有 ê 音節 lóng 標記做輕聲

14 (語詞層次)三連音處理：假使一個語詞 lóng 總三個音節，而且三個音節 lóng kāng 款，第一個標記做"~"

15 (語詞層次) 特殊詞處理

15.1 句內若出現"sím-mih / sím-mih " chia ê 語詞，kā 語詞改做 "sím-mí"，變調註記 mài 修改。

15.2 句內若出現"án-ni /àn-ni"，kā 語詞改做"án-ni"，變調註記是 t#；若出現"an-ni / an-nī"，kā 語詞改做"án-ni"，變調註記是 t#。

16 句內 ê Marker 處理

16.1 (語詞層次)假使句內有 "iah-sī / ah-sī / iáh-sī / áh-sī /á-sī"，頭前 hit 個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.2 (句型層次)"sī[是]" ê 處理：假使句內出現"sī[是]"、而且頭前 hit 個語詞 ê 詞類是 V(動詞)、而且頭前 hit 個詞 tī "sī"後壁（一直 kàu 句尾）koh 有出現，將頭前 hit 個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.3 (語詞層次)假使句內有 "che / he / chia / hia"，變調註記改做#(本調)

16.4 (語詞層次)假使句內有"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] / sit-chāi[實在] / sī-chūn[時陣]"，這個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.5 (語詞層次)假使句內有"chiū tō[就] "，而且頭前語詞 ê 詞類是 A(形容詞)，頭前語詞 ê 詞尾 ê 變調註記改做#(本調)

16.6 (語詞層次)假使句內有"sî-kàu[時 kàu]"，這個語詞 ê 兩個音節 ê 變調註記 lóng 改做#(本調)

17 (詞類層次) T(語助詞)處理：若是 siōng 尾 ê 詞類是 T(語助詞)，將頭前 hit 個語詞 ê 詞尾變調註記改做#(本調)

18 其它變調 ê 處理：

18.1 (語詞層次)"teh[在]" ê 處理：語詞若是"teh / tī-teh"，kā "teh" ê 變調註記改做$(以後改做^)(其它變調，暫時 kah 再變調 kāng 款)

19 (語詞層次)輕聲 ê 處理：

19.1 (語詞層次)句尾若有"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]" 而且頭前 hit 個語詞 ê 詞類是 V，請將頭前 hit 個詞 ê 詞尾 ê 變調註記改做#(本調)，句尾這個詞 ê 所有音節 ê 變調註記 lóng 改做%(輕聲)

19.2 句內若"sian-si^{n}/sin-se^{n}/sian-se^{n}[先生]"m̄是出現 tī第一個語詞，而且頭前一音節是單音節、第一字母大寫，請將頭前音節變調註記改做#(本調)，這個語詞 ê 所有音節 ê 變調註記 lóng 改做%(輕聲)

19.3 (語詞層次)句尾 ê "bô[無]" (是m̄是 beh 輕聲 ê 處理)

19.3.1 頭前一個詞若是"á / á-sī / iah / iah-sī / ah /ah-sī [或是]"，無需要修改(維持原來，讀本調)

19.3.2 其它 ê 情形，頭前 ê 語詞詞尾修改做#(本調)，"bô[無]"改做%(輕聲) (部分會錯誤)

19.4 (語詞層次)句尾 ê "bē/bōe[不會]" (是m̄是 beh 輕聲 ê 處理)

19.4.1 句內若有出現"ē/ōe[會] ē-hiáu/ōe-hiáu[會曉]"，修改做%(輕聲)(tī 19.5.2 進前做)

19.4.2 頭前一個詞若是"á / á-sī / iah / iah-sī / ah /ah-sī [或是]"，改做#(本調)

19.4.3 (其它 ê mài 修改，因爲有可能是 ambiguity(賣))

20 (語詞層次) R(代名詞) "góa / lí / i / gún|góan / lán / lín / in" tī句尾 ê 隨前變調處理(愛 tī第 9 條以後做) 假使這個代名詞 tī句尾，而且頭前 ê 語詞是動詞（只要有出現 tō 會 sái），變調註記改做'@'(隨前變調)

我們利用下列資源建立變調規則演算法，包括：

(a) 語言學家整理的台語變調規則；

(b) 從訓練語料歸納出的規則；

(c) 我們本身對台語變調規則的理解；

(d) 中研院資訊所詞庫小組的中文斷詞系統（參考其詞類標記結果）；

(e) 台語文語詞檢索系統（看台語某些語詞的變調情形）。

值得一提的是，語言學家整理的台語變調規則，有的只針對某些狀況處理而非全體適用，有的規則存在許多例外情形，這些因素導致實作上的困難。因此，語言學家整理的台語變調規則之外，本文其中兩位作者，兼具資訊背景及台語文背景，我們對台語變調規則有所了解，針對訓練語料及現有變調規則演算法得出錯誤變調註記的部分，進行錯誤分析，進一步來補充變調規則。補充變調規則時，也根據經驗，考慮到其它沒有在訓練語料中出現，但是相關的變調規則，也一併補充進來。訂定變調規則時，以適用大部分情形爲原則（例如可以讓語料庫中80%以上正確），不要求完全正確。而新規則加入後，可能影響部分原來的規則，於是，詞庫小組的中文斷詞系統及台語文語詞檢索系統成爲我們決定是否加入新規則的重要參考工具。

變調規則處理音節、語詞、詞類、句型等四種層次的變調問題，舉例來說：

(1) 音節層次，例如「koh[又]」、「beh[要]」，不管是否爲語詞的一部份（如「kiông-beh 強[要]」、「koh-chài[又]再」等），都標記爲再變調。

(2) 語詞層次，例如「che[這]」、「he[那]」，不管出現在何處，一律標記爲本調。有的情形則是，某個語詞出現時（如「ê[的]」，）會去改變前面語詞的變調標記。

(3) 詞類層次，例如詞類爲 N(名詞)，之後的詞類若爲 A(形容詞)、D(副詞)、P(介詞)、R(代名詞)或 V(動詞)，則此名詞詞尾音節標記爲本調。有時是某個詞類出現時（如 G 後置詞），也會改變前面語詞的變調標記。

(4) 句型層次，有些語詞出現時（如「iah-sī[或]是」）此句子此語詞前的部分，可以視爲一個子句；又例如「ē...bē 會...[不會]」的句型出現時（「bē」出現在句尾，句中出現「ē 會」），則將「bē」標記爲輕聲。

有些規則有先後順序，後面的規則可覆蓋原來的規則，如代名詞（「lí 你」、「góa 我」、「i 伊」）的變調處理規則可以覆蓋「ê[的]」的變調處理規則；或是上述句型層次的兩個例子中，第一個規則可以覆蓋第二個規則：

例 9 「Lí ē khì kok-gōa bē 你會去國外[不會]」，句尾的「bē」標記爲輕聲
「Lí ē khì kok-gōa iah-sī bē 你會去國外[或]是[不會]」，句尾的「bē」標記爲本調

另外，因爲詞類歧異的情形沒有處理，所以這些規則在詞類層次的部分，有的規則註明要在沒有歧異時才適用，有的規則是只要存在此詞類就適用。

目前我們訂定 20 條變調規則，並打算持續修改及增加。

舉例來說，下列的訓練語料：

例 10 Chhin-chhiūⁿ án-ni lâi kóng , chāi lán Tâi-ôan kīn-kīn chi̍t-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 . Só-í thang kóng Tâi-ôan sī chi̍t-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-hȏng , hō Tang-iûⁿ ê lâng chi̍p-óa lâi hióng-hok an-lo̍k .

(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，hō 東洋 ê 人集倚來享福安樂。)[14]

經過詞類標記和變調規則處理後，輸出爲：

例 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). Só-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-hȏng#(N), hō(D;P;V) Tang-iûⁿ#(N) ê(M) lâng#(N) chi̍p-óa(V) lâi(D;V) hióng-hok#(A) an-lo̍k#(A).

其中，刮號內爲詞類標記，台語羅馬字之後若沒有符號表示規則變調，標記「#」表示讀本調，「&」表"á"前變調，「$」表再變調，台語羅馬字以粗體加框，表變調規則錯誤的部分（正確應該讀規則變調）。目前所使用的變調註記請參考表 2：

表 2 變調註記

(t)	#	@	%	$	&	~	^
規則變調	本調	隨前變	輕聲	再變調	á 前變調	三連音第一音節	其它變調

4.4 評估正確率

如前所述，本文其中兩位作者是台語文專家，我們有能力確認變調結果的正確率。有些句子，不同的兩種變調結果都是可接受的（有的人變其中一種，有的人變另外一種），則系統的輸出，只要符合其中一種都視爲正確。例如「góa ài lí 我愛你」，「góa ài# lí@」（「góa 我」規則變調，「ài

[14] 出處爲 1924 年出版的《Chha̍p-hāng kóan-kiàn 十項管見》，作者是蔡培火。
http://iug.csie.dahan.edu.tw/TG/chu/10HKK/10HKK.asp

愛」本調，「lí 你」隨前變調）和「góa ài lí#」（「góa 我」、「ài 愛」規則變調，「lí 你」本調）都是正確的，這當中牽涉的語意問題或句子的焦點問題（如上例，句子的重點是動作「ài 愛」還是對象「lí 你」，變調結果應該不同）我們暫不考慮（而有些例子，不牽涉語意和焦點，仍有兩個不同的合法變調結果）。

5 初步研究結果

初步結果請參考表 3。訓練語料共有 614 個音節，經過人工檢查，共有 15 個音節變調標記錯誤，正確率爲 97.23%；測試語料共有 955 個音節，共有 106 個音節變調標記錯誤，正確率爲 88.90%。

表 3 變調標記正確率

	音節數(A)	標記錯誤數(B)	正確率(1-B/A)
訓練語料	614	15	97.56%
測試語料	955	106	88.90%

其中，測試語料中的錯誤，有些是因爲變調規則尚未完整，沒有處理到的，如果把這些變調規則再補上，應該至少可以提高 2.5%的正確率。

6 錯誤分析及相關問題討論

我們希望再做改進，讓變調系統的正確率能夠提高。在此提出我們所遇到的問題。

6.1 台語的詞類

目前我們使用中文的詞類，中文的詞類是不是適合台語，可能需要語言學家給我們答案。[15] 日本時代，1934 年陳輝龍出版《台灣語法》，[16]戰後，1950 年李獻璋在日本出版《福建語法序說》，[17]這些書都有討論台語的詞類；此外，1984 年中華語文研習所出版、Embree 的《台英辭典》，詞條有詞類的資料，這些應該都是可供參考的資料，當然，這些資料需要建立電子檔，還需要語言學家的協助，檢視這些資料對於處理台語變調問題是否合用。

6.2 台語的分詞標準及辭典

[Chan 1997]根據中研院詞庫小組的中文分詞標準，提出了台語分詞標準，可惜沒有引起進一步的討論。如果台語分詞標準可行，我們還需要一部符合分詞標準的辭典，目前還沒有，希望將來可以建置完成。

6.3 書寫的標準化問題

如果是漢字表記的台語文，這個問題十分嚴重。至於台語羅馬字的文獻、語料，腔調基本上沒有問題，已可透過辭典查詢來解決，連字符號(hyphen)的使用則不完全一致，這多少造成一些系統在處理變調問題的麻煩，把一個語詞分開寫（如「chit-ê 這個」寫成「chit ê」）或把兩個語詞

[15] 鄭良偉曾訂定台語的詞類，這是我們之後要斟酌的方向之一。
[16] 陳輝龍所列出的詞類有名詞、代名詞、數詞及助數詞、形容詞、動詞、助動詞、副詞、前置詞、語尾詞、接續詞、感嘆詞等 11 類。
[17] 李獻璋所列出的詞類有名詞、代名詞、(以上屬實體詞)動詞、助動詞、(以上兩類屬敘說詞)形容詞、副詞、(以上兩類屬限定詞)介詞、連接詞、(以上兩類屬關係詞)助詞(屬情態詞)等 9 類。

連起來寫（如「só siá 所寫」寫成「só-siá」）對系統而言，會產生不同的變調標記，其中一個是錯的。然而，動手修改語料並不是好方法。

不過，從資訊處理的角度而言，也許可以將這些問題做較完整的整理，並提出書寫的建議規範。

6.4 詞類無法解決的變調問題

以目前的作法，在檢視變調標記錯誤的地方時，有些可能無法透過詞類的排列順序來決定變調與否，目前看到的，包括並列的動詞及並列的名詞，例如：

例 12　「phah-piàⁿ(V) chò(V) khang-khòe(khè) (N)打拼做空課」(2,2,2,7,3)
「kiảh-bảk(V) khòaⁿ(V) hn̂g(N)舉目看園」(3,8,2,5)

這是並列動詞的情形，第一個例子中，前面的動詞其詞尾音節需規則變調；而第二個例子則是讀本調，我們無法從詞類中確認該怎麼變調。當然這裡有一些線索，第一例中的「phah-piàⁿ打拼」若拆成個別音節，「phah 打」和「piàⁿ拼」都是動詞，第二例中的「kiảh-bảk 舉目」拆開的話，「kiảh 舉」是動詞「bảk 目」是名詞，不過這樣的分析，在實作上可能太繁瑣。

例 13　「tiān-chú lêng-kiāⁿ電子零件」
「thâng-thōa chiáu-chiah 蟲豽鳥隻」[18]

這是並列名詞的情形，第一例中，前面的名詞其詞尾音節需規則變調；第二例中則讀本調。目前我們暫時想不出解決的方法。

6.5 錯誤分析

我們檢視錯誤的部分，包括前面已經討論過的，大致有以下幾種情形（刮號中的敘述表可能的解決方式）：

- 因爲辭典沒有此語詞而導致的錯誤；（增加詞條）
- 因爲少了標點符號；[19]
- 因爲同音詞造成的詞類標記錯誤；
- 因爲詞類歧異而無法判別；
- 連字符號書寫問題；（修改語料，或是，部分情形可用程式處理，例如，看到「chit ê 這個」就改成「chit-ê 這個」）
- 變調規則不夠完整；（可以繼續修改變調規則，不過還要考慮 side effect）
- 定量詞問題；（加上定量詞處理應可解決）
- 專有名詞問題；（加上專有名詞處理應可解決）
- 句型問題；（繼續修改變調規則，不過此部分難度較高）
- ...

另外，當然還可能有不少我們尚未遇到的疑難雜症。

---

[18] 「thâng-thōa 蟲豽」就是昆蟲。

[19] 例如「tī ke-lō teh kiâⁿ ū tú-tiòh chit-ê ...[在]街路[在]行有[遇]著一個.....」，「kiâⁿ行」後面如果有逗點，結果會正確，可是語料中沒有逗點，導致錯誤。

7 未來工作

以目前的成果而言，台語變調規則系統應該還有很長一段路要走，未來的工作包括：

- 藉助語言學家的專長，希望語言學家能討論制訂出台語的詞類、分詞規範，並建立出符合分詞規範的分詞辭典；
- 語料的處理上，我們還需要處理構詞、定量詞、專有名詞等問題；
- 詞類標記的處理上，我們還需要處理詞類排歧(disambiguity)等問題；
- 詞類標記的部分，我們也許還可以考慮直接以 Embree 台英辭典的詞類來做做看，省去對應到中文這個步驟，也許讓詞類歧異降低，增加變調結果正確率。
- 變調處理部分，變調規則的改進當然是必須的；
- 目前的變調規則，算是 bottom-up 的方式，也可考慮運用語法理論，用 top-down 的方式實作看看。鄭良偉教授提出語法模版理論，認爲可以解決翻譯、台語變調等問題。實作上看起來似乎很困難，不過也值得一試。

對 Native Speaker 而言，一個三歲小孩對於台語變調幾乎沒有任何困難。台語變調系統，要如何達到三歲小孩的水準，需要持續的努力。在華文的大環境下，台語的資源很有限，不過，一步一腳印，我們希望這些點點滴滴的成果能夠累積，也期盼更多研究資源及人力的投入。

誌謝

本研究獲得國家台灣文學館籌備處委託計畫「台語文數位典藏資料庫計畫—台語文全羅文字語音輸出系統Tâi-gu-bûn Sò-uī-tián-chông Chu-liāu-khò Kè-ōe -- Tâi-gú-bûn Chôan-lô Bûn-jī Gú-im Su-chhut Hē-thóng」的經費支持，特此致謝。也感謝兩位審查者對本文所提出的建設性改進意見。

參考資料

1. 專書、論文

[Chan 1997] 曾金金，1997，〈台語斷詞原則討論〉，《台語文學出版物收集、目錄、選讀編輯計畫結案報告》p47-73，文建會委託計畫。
http://iug.csie.dahan.edu.tw/TG/CompLing/hunsu/hunsu.htm

[Iûⁿ 2003] 楊允言，2003，《台文華文線上辭典建置技術及使用情形探討》，2003 第三屆全球華文網路教育國際學術研討會論文集 p132-141，台北，圓山大飯店，2003/10/24-26。
http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/THsutian/THsoaNtengsutian.htm 。

[Iûⁿ & Tiuⁿ 1999] 楊允言、張學謙，〈台灣福佬話非漢字拼音符號的回顧與分析〉，《第一屆台灣母語文化重生與再建學術研討會論文集》p62-76，台南：台南市文化基金會。
http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/Huho/huho-0.htm

[Lîm 1997] 林川傑，1997，《國語-閩南語機器翻譯系統之研究》，台北，台灣大學資訊工程系(碩士論文)。

[Liang et. 2004] Min-siong Liang、Jui-Cheng Yang、Yuang-Chin Chiang、Ren-Yuan Lyu，2004，〈*A Taiwanese Text-to-Speech System with Applications to Language Learning*〉，《Proc. of the 4th IEEE Int. Conf. on Advanced Learning Technologies (ICALT'04) 》p91-95，Finland : Joensuu http://msp.csie.cgu.edu.tw/pmwiki.php/PublishMedia/PubPaper1

[Lô͘ 1999] 盧廣誠，1999，《臺灣閩南語詞彙研究》，台北，南天。

[Ông et. 1999] 王駿發、黃保章、林順傑，1999，〈國語文句翻台語語音系統之研究〉，《第十二屆計算語言學研討會》p37-53，新竹：交通大學。

[Sia et. 1999] 佘永吉、鍾高基、吳宗憲，1999，〈臺語多音調音節合成單元資料庫暨文字轉語音雛型系統之發展〉，《第十二屆計算語言學研討會》p15-36，新竹：交通大學。

[Tē^{n} 1997] 鄭良偉，1997，《台語、華語的結構及動向 I 台語的語音與詞法》，台北：遠流。

[Tē^{n} 2002] 鄭良偉，2002，〈語法模板上的聲調變化—認知及測驗〉，《2002 台語羅馬字教學及研究國際學術研討會論文集》，台東：台東大學。
http://iug.csie.dahan.edu.tw/iug/ungian/POJ/siausit/2002/2002POJGTH/lunbun/K1-Liong-ui.pdf

[Tiu^{n} 2001] 張裕宏，2001，《白話字基本論：台語文對應&相關的議題淺說》，台北：文鶴。(第一章導論：http://iug.csie.dahan.edu.tw/iug/Ungian/patlang/POJkpl/POJkpl01. htm)

2. 網站資料

Taiwanese Package http://www.phahng.idv.tw

中研院資訊所詞庫小組中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/

台華線上辭典 http://iug.csie.dahan.edu.tw/TG/sutian

台語文語詞檢索系統 http://iug.csie.dahan.edu.tw/TG/concordance

台語羅馬字教學進修網站 http://elearning.lib.nttu.edu.tw/tglmj/index.htm

白話字&萬國碼：字型及軟體開發 http://iug.csie.dahan.edu.tw/TG/Unicode/

白話字書目資料 http://iug.csie.dahan.edu.tw/iug/Ungian/Soannteng/subok/poj.htm

白話字文物展覽 http://www.de-han.org/pehoeji/exhibits/index.htm

長庚大學 多媒體訊號處理實驗室 http://msp.csie.cgu.edu.tw/pmwiki.php

台大資訊系 台語文研究室 http://tb.csie.ntu.edu.tw/

台大資訊系 自然語言處理實驗室 http://nlg3.csie.ntu.edu.tw

交大電信系 語音處理實驗室 http://speech.cm.nctu.edu.tw/

成大資工系 多媒體人機通訊實驗室 http://chinese.csie.ncku.edu.tw/chwu/home.htm

成大電機系 多媒體通訊 IC 系統設計實驗室 http://140.116.156.179/