

台語文語料處理 *kah*

線頂資源研究

Research on Corpus Processing and
Online Resource for Written Taiwanese
Tâi-gú-bûn Gú-liâu Chhú-lí kah
Sòaⁿ-téng Chu-goân Gián-kiù

lûⁿ Ún-giân

楊允言 著

允言兄長期投入台語文運動，真早就建立台語文數位化 ê 語料庫，免費提供社會大眾 kah 研究者 tī 線頂查詢。M-koh，一般人 kan-ta 使用，可能 m̄知影背後語言科技 ê 技術 kah 原理。想欲瞭解台語文有啥物線頂資源 kah 語言科技通用 ê 讀者，這本冊真值得參考。

語言學 ê 研究者有一个定著愛處理 ê 倫理問題：咱提著語言研究 ê 語料，chia ê 語料 beh 按怎進一步為語言保存 kah 復興有所貢獻？π 這方面，允言兄有真好 ê 示範。伊 kā 語言科技 kah 語言文獻做結合，用「傳遞成果」ê 角度，將收集來 ê 文獻轉化做線頂 ê 資源，對語言學習、語言記錄 kah 保存攏有幫贊。

這本冊記錄台語文語料處理 ê 方法，koh 有提供台語文線頂資源予咱參考。語料庫 kah 線頂資源是台語文現代化 ê 基礎工程。咱除了需要 kā 台語文數位化，koh 需要有實際 ê 讀寫實踐，台語才會行出數位博物館，變做有性命 ê 語言。

國立台東大學華語文學系教授 張學謙 *Jiun Hah-khiam*

導讀 kah 說明

這本冊分做七章，第一章導論是台語文相關 ê 背景介紹，包括台灣多元 ê 語言現況，針對台語 ê 部分，koh ùi 世界 ê 角度來看咱人話（閩南語）使用 ê 地理分佈，然後討論台語 ê 名稱問題，並且說明台語 ê 處境、轉機 kah 危機。繼落來介紹台語文書寫 ê 文字系統，包括羅馬字 kah 漢字兩大系統，羅馬字會以白話字為主，白話字 kah 漢字是咱 beh 處理 ê 文字。Koh 來針對台語文語料處理，tī beh 做進前，提出對這項工課 ê 基本了解、心理準備 kah 建議。

第二章討論白話字按怎處理，因為 Unicode ê 編碼政策會使 kā 兩、三個符號鬥做一個符號，所以有 ê 白話字符號 ê Unicode 編碼會使有兩款表示方式，可比“á”，有可能是單一 ê 符號，mā 有可能是“a” kah “”兩個符號鬥起來 ê，全一個符號若用無全 ê 表示方式，會造成查詢 ê 困擾，所以阮提出用純文字 ê 白話字表示法成做內部表示法，這個表示法 ê 另外一個好處是排等（[pâi-téng]，sort）ê 時，會照咱愛 ê 聲調順序來排。繼落來討論白話字搜揣 ê 方式，用一般 ê 字串比對並 m 是好 ê 做法，因為可能會揣著真濟無需要 ê 資料，阮提出兩階段 ê 搜揣策略，先用一般 ê 字串比對，揣著 ê 資料 koh 經過篩選才呈現予使用者。另外，根據實際台語羅馬字教學 ê 經驗，有一寡學習者聲調 ê 分別 khah 有困難，有 ê 學習者對入聲（-p/-t/-k/-h）khah 無法度掌握，所以阮提出無聲調、入聲

加添使用者詞庫 ê 功能來解決這個問題。

第四章 teh 討論 beh 按怎做台語文轉做聲音 ê 線頂系統。Kah 華語比起來，台語有複雜 ê 連讀變調，問題相對困難。文字（輸入）ê 部分，因為漢字寫法 khah 無一致，而且一个漢字可能有幾 lō 个讀音，koh 愛處理斷詞，為著 kā 文轉音 ê 問題單純化，阮用白話字做文字輸入。白話字是標本調，所以需要提出變調演算法，阮先整理台語 ê 七種變調（規則變調、隨前變調、輕聲、再變調、á 前變調、三連音變調 kah 升調），訂出變調 ê 內部表示法，了後根據討論台語變調 ê 著作所整理 ê 規則來揣變調分界組，m̄-koh chia ê 規則無夠齊全[chiâu-chîng]，所以阮掠一寡文本來看，設法 ùi 內底揣出變調分界組 ê 線索，訂出 20 條變調規則，包括音節、語詞、詞性 kah 語句四个無仝層次 ê 規則。詞性 ê 部分，因為無符合阮需要 ê 帶詞性標記 ê 台語辭典，阮利用中央研究院資訊所詞庫小組針對華語 ê 46 个詞性標記做基礎，kā 簡化做對判斷變調有幫贊 ê 12 个詞性，對每一个台語語詞，透過台華辭典 kā 對應去華語，去查這個華語 ê 詞性。這個規則式 ê 台語變調演算法，正確率有 89%。阮另外請人 kā 所有台語音節錄音起來，一个音節切做一个聲音檔，使用者輸入 ê 白話字台文，透過查辭典揣著詞性，koh 經過台語變調演算法算出佢一个音節愛讀啥物聲調，了後 kā 對應 ê 聲音鬥起來做一个檔案，即時傳轉去予使用者。阮 mā 討論錯誤 ê 一寡原因，做未來改進 ê 參考。

因為台語變調處理 ê 時，詞性並無做確認 (disambiguity)，一个語詞可能對應 m-nā 一个詞性。第五章討論台語文 ê 詞性標記，目標就是 beh 確認每一个語詞 ê 詞性。因為資源有限，無法度利用人工來做大量 ê 詞性標記 (成做訓練語料)。阮提出透過 kā 每一个語詞對應去華語 (有 khah 豐富 ê 資源)，利用華語現有 ê 詞性標記家私來做台語文 ê 詞性標記。語料來源是國立台灣文學館台語文數位典藏資料庫 ê 台語文本，伊是有白話字 kah 漢羅合用兩種文字一段落一段落對齊 ê 語料，(舉例：“Tâi-ôan tē-it kôan ê Giòk-san ê hū-kūn khah kē ê só-chāi ...” 對應 “台灣第一懸 ê 玉山 ê 附近較低 ê 所在...”)，先 kā 轉做語詞對齊 (“台灣/Tâi-ôan 第一/tē-it 懸/kôan ê/ê 玉山/Giòk-san ê/ê 附近/hū-kūn 較/khah 低/kē ê/ê 所在/só-chāi ...”)，查台華辭典揣出對應 ê 華語詞 (“台灣{台灣} 第一{第一;絕頂} 懸{高} ê {的} 玉山{玉山} ê {的} 附近{附近} 較{較} 低{低} ê {的} 所在{去處;地方;角頭;所在;處所;場所;間量} ...”)，了後利用 Hidden Markov Model 統計模型，提詞庫小組千萬字 ê 華文語料庫做訓練語料，ùi chia ê 華語候選詞內底揀出系統認為上適當 ê 華語對應詞 (“台灣[台灣] 第一[第一] 懸[高] ê [的] 玉山[玉山] ê [的] 附近[附近] 較[較] 低[低] ê [的] 所在[地方] ...”)，一个華語詞可能有 m-nā 一个詞性，最後利用 Maximal Entropy Markov Model 統計模型決定詞性，對應轉來台語詞 (“台灣/Tâi-ôan(Nc) 第一/tē-it(Neu) 懸/kôan(VH) ê/ê(DE) 玉山/Giòk-san(Nc) ê/ê(DE) 附近/hū-kūn(Nc)

域，其實已經累積真濟有心 ê 人拍拚 ê 心血。希望後來加入 ê 友志，會當以這個基礎繼續向前行。

這本書 ê 寫作，得著國科會人文處 ê 專書寫作計畫補助，計畫編號是 NSC 101-2410-H-142-015。書 ê 部分內容是以筆者 ê 博士論文（英文）做基礎，用台語文重寫，並且做真大程度 ê 修改，koh 補充真濟新 ê 發展。其中，台語變調處理 ê 部分 pat 發表 tī 第 17 屆計算語言學研討會（2005 年，華文）kah *International Journal of Computational Linguistics and Chinese Language Processing* (IJCLCLP) 12 卷 4 號（2007 年，英文），台語文詞性標記 ê 部分 pat 發表 tī 第二十屆自然語言與語音處理研討會（2008 年，華文）kah IJCLCLP 14 卷 3 號（2009 年，英文），台語雙文字轉換 ê 部分 pat 發表 tī 第五屆台灣羅馬字國際研討會（2011 年，台文）。所開發 ê 相關系統，除了家己以外，koh 有劉杰岳、李盛安、戴嘉宏、雷明漢鬥跤手，特別愛 kā in 說謝，辭典資料主要是鄭良偉教授提供，台語文語料 ê 提供 kah 整理者，就算袂清 ah，無法度一个一个列名。總是會當有 chia ê 成果，除了我 ê 指導教授高成炎老師 kah 陳克健老師當初時提供我真豐富 ê 研究資源，濟濟牽成我 ê 師長、親人 kah 台文界 ê 朋友（心內小算一下，成百人），攏是我愛感謝 ê 對象。

目 錄

序 運用語言科技來補破網	I
導讀 kah 說明.....	III
第一章 導論	3
第一節 背景.....	4
一、台灣所使用 ê 語言 kah 語言人口	4
二、通世界咱人話現況	8
三、台灣 ê 「閩南語」 beh 叫啥物名？	11
四、台語 ê 處境	13
第二節 無仝形式 ê 台語書寫	16
一、漢字	17
二、羅馬字	19
三、漢羅合用	22
四、其它書寫系統	23
五、本文所 beh 處理 ê 書寫系統	23
第三節 台語文語料處理相關問題	23
本章參考書目	25
第二章 白話字 ê 編碼、搜搵 kah 顯示	31
第一節 白話字編碼.....	31
第二節 白話字 ê 兩款表示法	34
第三節 白話字文本搜搵	36
一、白話字文本搜搵相關問題	36
二、兩階段搜搵策略：先比對字串才 koh 篩選.....	37
三、體貼 ê 查詢擴展：無聲調、束喉音、入聲、韻母	39
四、實作結果舉例	44

二、未來改進方向	87
本章參考書目	89
第五章 台語詞性標記	93
第一節 台語詞性標記相關問題	93
第二節 實驗方法	95
一、語料來源	95
二、語詞對齊	95
三、揣出對應 ê 華文候選詞	95
四、揀出上適當 ê 華文對應詞	96
五、根據華文詞串揀出詞性	98
六、實例說明	99
第三節 實驗結果	100
第四節 錯誤分析	105
一、揀 m 著華文對應詞造成 ê 錯誤	106
二、台華辭典欠缺適當 ê 華文對應詞造成 ê 錯誤	107
三、未知詞造成 ê 詞性標記錯誤	107
四、傳播錯誤	108
五、其它情形	108
六、詞性標記錯誤總歸納	110
第五節 相關討論 kah 未來改進方向	110
一、敢有可能 koh 改進？	110
二、連字符問題、台語 kah 華語 ê 差異	111
三、無全文類抑是時代 ê 台語文本敢會有差別？	112
本章參考書目	113
第六章 台語雙文字轉換系統	117
第一節 現有相關系統介紹	118
第二節 互訊息 kah 相關度	118

第三節 轉換演算法.....	121
一、羅馬字轉漢羅演算法.....	122
二、漢羅轉羅馬字演算法.....	124
第四節 實驗結果.....	124
第五節 進一步 ê 應用：台語文發音系統	126
本章參考書目.....	127
第七章 台語文線頂資源評介	131
第一節 字型.....	131
一、白話字字型.....	131
二、漢字字型.....	132
第二節 字、辭典.....	133
一、台華線頂辭典.....	133
二、台語線頂字典.....	134
三、台日大辭典台語譯本.....	135
四、教育部台灣閩南語常用詞辭典.....	135
五、甘字典（廈門音新字典）.....	137
六、中研院語言所閩客語典藏計畫.....	137
第三節 語料.....	138
一、閩南語俗曲唱本「歌仔冊」全文資料庫.....	138
二、台語維基百科.....	138
三、台語文語料庫.....	139
四、台語聖經語料庫.....	139
五、台語文數位典藏資料庫.....	139
六、台灣白話字文獻館.....	140
七、台灣民間文學館.....	140
八、教育部台灣閩南語字詞頻調查工作.....	141
九、台灣民間文學歌仔冊資料庫.....	141

表目錄

表 1 - 1 台灣 ê 語言現況.....	4
表 1 - 2 台灣各語言使用人口.....	5
表 1 - 3 2011 年台灣各族群人口 kah 比例.....	7
表 1 - 4 1931 年台灣各語言人口 kah 比例.....	7
表 1 - 5 通世界咱人話人口 kah 比例.....	8
表 1 - 6 台語 ê 各種名稱.....	12
表 1 - 7 無仝世代 ê 母語態度.....	16
表 2 - 1 POJ Unicode 編碼	32
表 2 - 2 POJ 符號 ê 純文字表示.....	34
表 2 - 3 台語文三款表示法舉例.....	34
表 4 - 1 台語變調整理.....	72
表 4 - 2 觀察文本.....	76
表 4 - 3 試驗文本.....	76
表 4 - 4 中文詞知識庫小組簡化詞類集.....	77
表 4 - 5 台語變調處理系統所採用 ê 詞性集.....	79
表 4 - 6 變調標記.....	79
表 4 - 7 台語變調處理演算法.....	80
表 4 - 8 變調處理 ê 錯誤數 kah 正確率.....	85
表 4 - 9 錯誤原因 kah 可能 ê 解決方案.....	87
表 5 - 1 試驗語料列表.....	101
表 5 - 2 試驗語料詞性標記結果.....	102
表 5 - 3 台語詞性標記結果舉例.....	103

表 5 - 4 系統選 ĩ 著 ê 華文對應詞 kah 標記.....	106
表 5 - 5 欠缺適當 ê 華文對應詞造成 ê 錯誤.....	107
表 5 - 6 未知詞 ê 詞性標示錯誤.....	108
表 5 - 7 其它詞性標記錯誤舉例.....	109
表 5 - 8 詞性標記錯誤原因.....	110
表 5 - 9 無仝文類 ê 詞性標記正確率.....	112
表 5 - 10 無仝時代 ê 詞性標記正確率.....	112
表 6 - 1 “有人”各種發音頻率表	119
表 6 - 2 “有人” <i>MI kah rel</i> ê 值	121

第一節 背景

一、台灣所使用 ê 語言 kah 語言人口

根據 2013 年 ê *Ethnologue : Languages of the World* (《世界語言年鑑》) ê 記錄，台灣有 26 種語言，除去其中四種語言已經死去，koh 有 22 種 iah-koh 活 leh ê 語言。(Lewis *et.al.* 2013) 台灣 ê 語言現況，請參考表 1 - 1。

表 1 - 1 台灣 ê 語言現況

語言狀態等級 kah 說明	語言
1 國家語言 National	華語 Mandarin
3 普遍使用 Wider Communication	台語（閩南語）Min nan、日語 Japanese
4 教學用 Educational	台灣手語 Taiwan sign language、太魯閣語 Taroko
5 發展中 Developing	阿美語 Amis、泰雅語 Atayal、布農語 Bunun、客語 Hakka、排灣語 Paiwan、魯凱語 Rukai、雅美語 Yami
6a 有活力 Vigorous	鄒語 Tsou
6b 受威脅 Threatened	賽夏語 Saisiyat
7 無傳承 Shifting	卑南語 Puyuma
8b 接近滅絕 Nearly Extinct	荳蘭阿美語 Amis Nataoran、卡那卡那布語 Kanakanabu、噶瑪蘭語 Kavalan、沙阿魯阿語 Saaroa、邵語 Thao
9 賸文獻 Dormant	貓霧揀語 Babuza、西拉雅語 Siraya
10 滅絕 Extinct	巴賽 Basay、凱達格蘭 Ketangalan、龜崙巴宰 Kulon-Pazeh、拍瀑拉洪雅 Papora-Hoanya
資料來源：Lewis <i>et.al.</i> (2013)，語言狀態分級請參考 http://www.ethnologue.com/about/language-status	

為著好理解，原住民語ê名稱採用華文寫法。《世界語言年鑑》所描寫ê台灣語言現況，可能kah咱ê認知無完全一致，包括日語是m̄是有親像台語hiah-nī普遍，台灣原住民語言官方ê資料是 14 族、42 語，²是按怎kan-na太魯閣語ê語言狀態等級達到 4，koh贏過客語，抑是滅絕ê語言敢kan-na這四種等等。這可能表示in對咱ê現況無夠了解，mā表示咱需要有專家學者抑是官方單位提供khah符合現況ê調查資料thang予《世界語言年鑑》做參考。

阮以 Lewis *et.al.* (2013)為主，koh 參考客委會(2011)推估 ê 台灣各族群人數 kah 原民會(2013)提供 ê 2013 年 2 月 ê 各族人口資料，表 1-2 列出台灣各語言 ê 使用人口。

表 1-2 台灣各語言使用人口

語言	使用人口	引用資料 發表時間	族群人口數
台語	15,000,000	1997	15,642,000 [*]
華語	4,320,000	1993	1,777,000 [*]
客語	2,370,000	1993	3,147,000 [*]
阿美語 Amis	138,000	2002	195,335 ⁺
手語	104,000	2004	
泰雅語 Atayal	79,000	1989	83,837 ⁺
排灣語 Paiwan	66,100	2002	93,795 ⁺
布農語 Bunun	38,000	2002	54,660 ⁺

² 原民會網站提供「原住民族 14 族語 42 個方言別分布參考表」
<http://material.apc.gov.tw/material/file/distribution.htm>。

根據客委會(2011)，台灣各族群 ê 人口 kah 比例，請看下表：

表 1-3 2011 年台灣各族群人口 kah 比例

族群	人口數	比例
福佬人	15,642,000	67.5%
客人	3,147,000	13.6%
外省人	1,777,000	7.6%
原住民	409,000	1.8%
說明：來源是客委會(2011) p172，樣本數 65,566 人，自我單一認定，有 ê 人 ê 選擇超出這四項，包括台灣人、其他、m̄知影等等。		

咱 koh 來看日本時代 ê 資料，東方孝義(1931)有彼時台灣 ê 語言人口概況：

表 1-4 1931 年台灣各語言人口 kah 比例

語言	人口數	比例	
南部福建語漳州語	132 萬	33.08%	75.19%
南部福建語泉州語	168 萬	42.11%	
客人語	59 萬	14.79%	
其它支那語	3 萬	0.75%	
熟蕃語	5 萬	1.25%	
生蕃語	13 萬	3.26%	
日本語	19 萬	4.76%	
說明：語言名稱直接引用東方孝義(1931)ê 寫法。			

族群人口 kah 語言人口是小可無全款 ê 概念，為著方便，咱暫時無 beh 對這兩項 ê 差異做討論。

Üi 頂面 ê 表 thang 發現，「台語」ê 名稱有真濟款，總是這個語言是台灣使用人口上濟 ê 語言。本文 beh 針對「台語」做討論。

二、通世界咱人話現況

表 1-5 通世界咱人話人口 kah 比例

國家	總人口數 (千人)A	咱人話人口 B/ 比例(B/A)	語言 狀態	2009 咱人話 人口/比例	2005 咱人話 人口/比例	2000 咱人話 人口/比例
台灣	23,120	15,000,000/ 64.88%	3	15,000,000/ 68.01%	15,000,000/ 68.01%	15,000,000/ 69.74%
新加坡	4,930	471,000/ 9.55%	5	1,170,000/ 27.04%	1,170,000/ 26.87%	1,170,000/ 33.66%
馬來西亞	28,639	2,660,000/ 9.29%	3	2,660,000/ 10.37%	1,946,698/ 8.28%	1,946,698/ 9.09%
澳門	503	18,900/ 3.76%	6a	-	-	-
汶萊	406	13,100/ 3.23%	3	13,100/ 3.50%	12,147/ 3.33%	10,000/ 3.17%
中國	1,339,725	25,700,000/ 1.92%	6a	25,700,000/ 1.96%	25,725,000/ 1.98%	25,725,000/ 2.04%
泰國	67,312	1,080,000/ 1.60%	5	1,080,000/ 1.71%	1,081,920/ 1.67%	1,081,920/ 1.79%
香港	6,864	77,610/ 1.13%	5	-	-	-
菲律賓	88,707	592,000/ 0.67%	5	592,200/ 0.70%	592,200/ 0.69%	592,200/ 0.81%
印尼	237,642	700,000/ 0.29%	7	700,000/ 0.28%	700,000/ 0.29%	700,000/ 0.34%
	排名	第 25 名		第 24 名	第 24 名	第 21 名

根據 Lewis *et.al.* (2013)、Lewis(2009)、Gordon(2005) kah Grimes(2000), 筆者 kā 內底 kah 咱人話相關 ê 資料整理 tī 表 1-5, thang 看著 2000 年到 2013 年這十幾冬中間, 世界上咱人話人口 ê 變化情形。其中, 使用咱人話 ê 人口 tī 彼國 ê 比例是筆者根據彼

是 tī 語言認同上，其中有 ê 人已經無 koh 繼續使用這個語言。

黃宣範(1993 p21)針對台灣各族群人口 ê 估計，伊認為台語人比例佔 73.3%，若照伊 ê 估計，對照這陣台灣 ê 人口數，台灣 ê 台語人口就將近 1700 萬人，kah Lewis *et.al.* (2013) ê 統計，精差 200 萬人。當然，kah 新加坡 ê 問題全款，族群人口 kah 語言使用人口，可能會有落差，只是無 khah 詳細 ê 資料。

Lewis *et.al.* (2013)對台灣 ê 台語 ê 描述，有寫一句：「De facto language of national identity（實際上國家認同 ê 語言）」。若是 beh 討論台灣 ê 文化軟實力，台語絕對是真要緊 ê 因素。

Chinese, Min Nan	[nan] Tainan, Penghu Archipelago, east coast cities, western plain. 15,000,000 in Taiwan (1997 A. Chang). <i>Status:</i> 3 (Wider communication). De facto language of national identity. <i>Alternate Names:</i> Min Nan, Minnan <i>Dialects:</i> Taiwanese (Amoy, Chaenzo, Hokkien, Sanso). <i>Classification:</i> Sino-Tibetan, Chinese <i>Comments:</i> Taiwanese people are called Hoklo or Holo. Traditional religion, Buddhist, Christian.
-------------------------	--

圖 1 - 1 《世界語言年鑑》對台灣 ê 台語 ê 介紹

除了表 1 - 5 所列 ê 國家，若根據維基百科（Wikipedia）ê 描寫，包括越南、緬甸、寮國，mā 攏有咱人 ê 分佈。

目前英語是國際語言，講英語 ê 人，去到世界各所在，大概攏會使用英語來溝通。若用地圖來看咱人 ê 人口分佈，會當看出，咱人 ê 人口分佈大概是以南海做中心，環南海 ê 國家，差不多攏有講咱人話 ê 人口，南海會使算是咱人語族 ê 內海。台灣若 beh kah 東南亞連結，咱人話 mā 是一个媒介。當然，環南海 ê 咱人語族，中國福建是主要輸出地，台灣是上重要 ê 基地。



圖 1-2 咱人人口分佈，大概以南海做中心

三、台灣 ê 「閩南語」 beh 叫啥物名？

Tī 台灣，這個語言 beh 叫啥物名，是一個不止困擾 ê 問題。

李勤岸指出，這個語言至少有 17 个無仝 ê 名稱：台語、閩南語、福建話、Hō-ló 語、...(李勤岸, 洪惟仁 2007)。會有 chiah-nī 濟無仝 ê 名稱，反應出這個語言 ê 狀態 kah 處境 m̄ 是真好。

一開始，外國人 kā 這個語言號做「廈門話 (Amoy)」，因為清國時期廈門開港通商，in 對這個語言開始有 khah 大量 ê 接觸，出版 ê 辭典，攏 kā 叫廈門話。日本統治台灣 ê 時，捌叫做「福建話」，後來改稱「台灣話」。戰後，國民政府來台灣，kā 這個語言叫做「閩



圖 1-3 利用 Google trend 看台語各種名稱 ê 熱門度

另外，kan-na 看「閩」這個漢字有「虫」，會當知影這是帶歧視 ê 稱呼，beh 用「閩南語」做這個語言 ê 泛稱並 m̄ 是 hiah-nī 妥當，蔣為文、康培德等人建議用「咱人話」這個 khah 中性 ê 稱呼。(蔣為文 2012; Kang 2013)

因為本文 beh 處理 ê 文本，以台灣為主，所以阮用「台語」稱呼這個語言，用「咱人話」做這個語言 ê 泛稱。

四、台語ê處境

使用人口上濟 ê 語言，無予執政者訂做官方語言，除了受殖民統治，世界上大概揣無這款例。因為世界政治 ê 因素，台灣雖然有政治實體，聯合國 soah 無承認台灣是一個國家。生活 tī 台灣

ê 人，普遍無認為台灣是殖民地，m̄-koh 台語 tī 台灣 ê 處境，koh 有符合殖民地 ê 特點，這的確是真奇怪 ê 代誌。

日本時代，台灣開始行向現代化，統治者辦教育，m̄-koh 教學語言是國語(日語)。國民政府時代，教學語言全款是國語(華語)，入學考試抑是國家考試攏愛考國文(華文)，利用政治力提升華語文 ê 地位，kā 欠缺華語文能力 ê 人排除 tī 體制外(受完整日本教育 ê 台灣人是主要受害者)。1950 年代了後，進一步沒收教會 ê 台語聖經，限制台語 tī 學校 kah 媒體 ê 使用，教育學生講台語是無水準 ê 方言，tī 學校講台語愛罰錢抑是掛狗牌。(張學謙,施正鋒 2003)

教育體制方面，1987 年解嚴了後，本土意識漸漸浮頭，1990 年代初期，ùi 民進黨執政 ê 縣市開始利用非正式課程 ê 時間(社團活動)開鄉土課程，1996 年國小三年到六年設「鄉土教學活動」，其中包括鄉土語言教學，2001 年鄉土語言(後來改名叫做本土語言)成做國小一年到六年 ê 正式課程。(林瑞榮 2012) M̄-koh 一禮拜 kan-na 一節，而且實施十外冬以後，師資 iah-koh 是真大 ê 問題。

真濟無受本土語言專業教學訓練、無通過本土語言認證 ê 現職老師 teh 教本土語言課，教育部 tī 2013 年發公文予國小，要求 ùi 2016 年開始(成做正式課程 ê 15 冬後)現職教師若 beh 教台語，一定愛通過教育部台語語言能力認證中高級才會使教。另外有一批體制外 ê 台語支援老師，教育部稱呼 in 是教學支援人員，是領

表 1-7 無仝世代 ê 母語態度

出生年	華語	台語	客語
1931-1940	12.80%	65.80%	12.00%
1941-1950	18.50%	63.50%	10.00%
1951-1960	30.00%	46.90%	7.90%
1961-1970	37.20%	37.40%	8.50%
1971-1980	50.80%	25.70%	6.10%
1981-1990	61.10%	18.90%	2.20%

說明：「請問在家裡，您最希望您的小孩跟
您說哪一種話？」回答結果統計

第二節 無仝形式 ê 台語書寫

台語 kah 華語是有相關，m̄-koh 無仝款 ê 兩種語言，包括語音、構詞、語法各方面。(Tiunn 1998) 若 ùi 文言文 kah 現代華文 ê 觀點來看，台語有以下 ê 特色：

- (1) 保留真濟古漢語成分；
- (2) 文白異讀[ī-thók] ê 漢字特別濟；
- (3) 有音無漢字 ê 語詞特別濟；
- (4) 英、日外來語特別濟；
- (5) 吸收真濟和製漢語詞，而且用台語發音；
- (6) 合音詞濟。(鄭良偉 1990 p222-224)

以上 chia ê 台語 ê 特色，是咱 teh 文字化 ê 過程需要來斟酌 ê。台語文書寫真多元，到底有佻濟套台語書寫系統？楊允言、

張學謙(1999)列出至少 64 套，chia ê 書寫系統會當分做四大類，分別是：漢字、注音符號、假名 kah 羅馬字。

Tī chia，先 beh 介紹其中上主要 ê 書寫系統，包括漢字、白話字（羅馬字系統其中一種），koh 有漢字、白話字混合使用 ê 漢羅合用。

一、漢字

咱人話漢字書寫會當 ùi 西元 1566 年開始算，是 tī 中國閩南地區 ê 南管戲文。(吳守禮 1995) 彼陣漢字書寫並 m̄ 是書寫口語，總是南管戲文內底有咱人話 ê 詞彙出現。另外，19 世紀 tī 台灣民間社會流傳 ê 歌仔冊 mā 是用漢字書寫(Klötter 2005)，歌仔冊使用 ê 漢字，借音字 khah 濟。以上兩種文本算是 khah 特殊 ê 文類，kah 口語。

上早發現 ê 咱人話口語書寫 ê 文本，是 1593 年 tī 菲律賓（當時受西班牙統治）發行 ê 《基督要理》。(Cobo 1593) 日本時代官方 ê 台語書寫，mā 是以漢字為主，漢字邊 a 用修改過 ê 假名注台語音，日本官方所使用 ê 漢字，訓讀字比例 khah 懸。

Tiunn(1998)指出，用漢字書寫台語，chia ê 漢字會當分做四類：

- (1) 訓讀字（借漢字 ê 意思）
- (2) 本字

漢字 khah 罕用，若 kan-na 看 BIG-5 內碼 5,401 个常用字集 ê 漢字，其中有 5,337 無仝 ê 漢字分佈 tī 字典其中 13,176 條，按呢平均起來，一个漢字有 2.5 个發音。

ㄟ 台灣，因為華文教育普及，逐家看著漢字書寫 ê 台語文，普遍 khah 袂排斥，至少袂像繼落來 beh 講 ê 羅馬字按呢，掠做羅馬字是外國字，心內會 khah 驚惶。

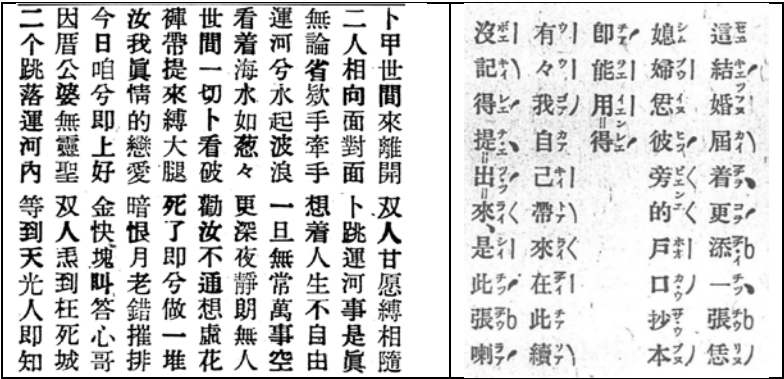


圖 1-4 漢字文獻舉例

說明：倒片是歌仔冊，正片是日本時代台灣語教科書

二、羅馬字

ㄟ 幾 lō 十套台語羅馬字書寫系統內底，1990 年以來，khah 重要 ê 有三个系統，分別是：

- (1) 白話字（教會羅馬字，下面簡稱 POJ）
- (2) 台灣語言音標方案（簡稱 TLPA）
- (3) 通用拼音

POJ 上早 ê 文獻，是麥都思 (Medhurst) tī 1832 年編 ê 《福建方言字典》，麥都思 tī 1820 年就編過一本詞彙集，所以 POJ 有超過 190 外冬 ê 歷史。POJ ê 出版品包括辭典、宗教作品、報紙、啟蒙讀物、教科書、...等。(張裕宏 2001) 其中，《Tâi-oân Hú-siân Kàu-hōe-pò》(《台灣府城教會報》，這馬改名做《台灣教會公報》) 1885 年 tī 台南開始發行，1970 年進前 ê 書寫文字就是 POJ，這份刊物是台灣出刊上久 ê 刊物。

POJ ê 文獻誠濟，1913 年，甘為霖 (William Campbell) 真驕傲 kā 咱講，in 用南台灣 ê 腔口，出版至少 700,357 頁 ê POJ 書刊。(Klōeter 2002) 另外根據呂興昌所執行 ê 「台灣白話字文學資料蒐集計畫」所蒐集著 ê 資料來看，戰後 POJ 書刊猶原繼續出刊，高峰 tī 1950、1960 年代，(楊允言，陳鄭弘堯 2004) 後來 soah 因為政治力 ê chhap 手，致到 tī 1970 年代急速衰微，1969 年《台灣教會公報》予執政當局全面禁止使用 POJ，暫時停刊了後，1970 改做華文版。1990 年代才 koh 開始有復興 ê 影跡，m̄-koh 若 kah 進前相比並，1970 年進前出版 ê 書刊，讀者群以教會為主，而且是全 POJ，1990 年以後 ê 台語文出版品，主要讀者群換做社會大眾，而且是漢羅合用成做主流。

蔣為文(2013)指出，POJ文化至少包含下面幾項有形、無形資產：(1) 口頭傳說kah表述，包括做為無形文化遺產ê台語kah客語；(2) 由台語kah客語延伸出來ê民間念謠、歌曲kah文學；(3) 白話字

ê 差異。(蔡志浩 2000) 台語部分，因為已經有白話字，kah 通用拼音 ê 方式明顯無仝，所以這部分 koh 分做甲式 kah 乙式拼音，甲式就是白話字，乙式才是通用推廣者真正想 beh 推廣 ê 通用拼音。(楊允言 2003b)

2001 年國校開始有每禮拜一節 ê 鄉土語言課程，一開始拼音符號並無規範，造成教學上一寡困擾，經過教育部開會討論，2006 年 10 月公佈「台灣閩南語羅馬字拼音方案」，簡稱台羅，會使 kā 看做是 POJ kah TLPA ê 混合體(有人講是 TLPA ê 身軀·POJ ê 頭殼)。

Tiunn (1998) 指出，漢字 kah 羅馬字有互補 ê 關係，漢字 ê 一寡缺點（欠表音功能，一个漢字可能有幾 lō 个發音）會使用羅馬字得著解決，倒頭講 mā 是按呢（仝一个羅馬字，對應幾 lō 个無仝意思 ê 漢字）。

三、漢羅合用

1970 年代，鄭良偉提出漢羅合用 ê 想法，伊同時 mā 是實踐者。簡單講，漢字使用上 khah 無爭議 ê，就寫漢字，若無就用羅馬字。Tī 1980 年代尾期以後開始 khah 興旺 ê 台語文運動，漢羅合用成做主流，大部分 ê 台語文出版品 ê 文字型式攏是漢羅合用。

漢羅合用 ê 羅馬字比例是佻濟？可能無仝作者會有一寡差異，蔡秀俐(2009)根據台語文語料庫內底 ê 漢羅文本來計算，羅馬字 ê 比例佔 21.56%。

四、其它書寫系統

除了以上提起 ê 系統，koh 有人提出無仝 ê 書寫系統，攞是偏向拼音 ê，包括假名、諺文（親像韓文 ê 形式，聲母、韻母、聲調門 tī 一个四角格仔內）kah 注音符號。（楊允言，張學謙 1999）

五、本文所 beh 處理 ê 書寫系統

咱若 beh 利用電腦來處理台語文相關 ê 議題，上基礎 ê 部分是愛收集夠額 ê 台語文語料。雖然台羅目前是官方 ê 羅馬字，總是若用語料量來看，POJ iah 是上濟 ê。所以阮揀 POJ kah 漢羅合用做咱 beh 處理 ê 台語文對象。到 taⁿ 為止，這兩種書寫形式 ê 台語文文本，阮已經收集著超過一千萬音節 ê 語料 ah。另外，因為官方 ê 羅馬字是台羅，m̄ 是 POJ，總是 POJ kah 台羅，符號並無衝突，阮 teh 討論 POJ 按怎處理 ê 時，mā 會順繼講台羅 ê 部分。

第三節 台語文語料處理相關問題

Beh 處理 POJ kah 漢羅合用兩種文字形式，咱需要先了解一寡相關 ê 問題，包括：

- (1) 台語文到 taⁿ iah 未標準化；雖然教育部針對漢字 kah 羅馬字訂出規範，m̄-koh kan-na 讀國校彼六冬，一禮拜有一節課，按呢敢 thang 寄望學習者有法度寫出遵照教育

1997), 台語文本若 beh 做標記, 建議愛標記實際發音(表層結構) kah 變調種類, 方便後續 ê 應用;

- (7) 想 beh 參與台語文處理工課 ê 技術人員, 本身上好是台語文書寫 ê 實踐者, 按呢 khah thang 知影處理 ê 鉅[mî]角, mā 上好愛 kah 台語文工作者有接接[chi-h-chiap], khah 會當了解台語文工作者實際 ê 需求。

本章參考書目

- Chen, Matthew Y. 2000. *Phonological Phrase as a Sandhi Domain. In Tone Sandhi : Patterns Across Chinese Dialects*. Cambridge Univ. Press
- Cobo, Juan. 1593. *Doctrina Christiana* (《基督要理》) <http://ip194097.ntcu.edu.tw/memory/tgb/thak.asp?id=178> (2013/12/20 檢索)
- Gordon, R. G. J. (ed.) 2005. *Ethnologue : Languages of the World*(15 ed.). Dallas, TX: SIL International
- Grimes, B. F. (ed.) 2000. *Ethnologue : Languages of the World*(14 ed.). Dallas, TX: SIL International
- ISO/IEC JTC1/SC2/WG2. 2004. Universal Multiple-Octet Coded Character Set(UCS) <http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2754.pdf> (2013/9/20 檢索)
- Kang, Peter. 2013. Field Report: On the Romanization Archives of Lán-lâng Language Collected in Leiden University. *Journal of Taiwanese Vernacular* 5(1) p98-108

- Klöter, Henning. 2002. The History of Peh-oe-ji. 《2002 台灣羅馬字教學 kap 研究國際學術研討會》
- Klöter, Henning. 2005. *Written Taiwanese*. Wiesbaden: Harrassowitz
- Lâu Kiát-gák. 2002. Sú-iōng Unicode lâi Chhú-lí POJ. 《2002 台灣羅馬字教學 kap 研究國際學術研討會》
- Lewis, M. Paul (ed.) 2009. *Ethnologue : Languages of the World*(16 ed.). Dallas, TX:SIL International
- Lewis, M. Paul, Gary F. Simons, Charles D. Fennig (eds.) 2013. *Ethnologue : Languages of the World*(17 ed.). Dallas, Texas:SIL International
- Tiunn, Hak-khiam. 1998. Writing in Two Scripts : A Case Study of Digraphia in Taiwanese. *Written Language and Literacy* 1(2) p223-231
- 丁鳳珍. 2013. 〈竟然能夠用台語聊天？〉. 《自由時報》2013.4.17 自由廣場
- 吳守禮. 1995. 《閩台方言研究集 1》. 台北:南天
- 李勤岸, 洪惟仁. 2007. 〈沒有名字的語言？－「台灣話」、「閩南話」還是 Hòh-ló 話？〉. 《台灣文學館通訊》15 期 p36-41
- 東方孝義. 1931. 《台日新辭書》. 台北:台灣警察協會
- 林瑞榮. 2012. 〈我國鄉土教育沿革與發展趨勢〉. 《教育資料與研究》105 期 p163-184
- 客委會. 2011. 〈99 年至 100 年全國客家人口基礎資料調查研究〉. 《行政院客家委員會委託研究報告》. <http://www.hakka.gov.tw/dl.asp?fileName=1521131271.pdf> (2013.6.30 檢索)
- 原民會. 2013. 〈原住民人口數統計資料〉. <http://www.apc.gov.tw/portal/docList.html?CID=940F9579765AC6A0> (2013.6.30 檢索)

- 蔡志浩. 2000. 〈通用拼音與漢語拼音的相似性：在「音節」與「詞」層次上比較〉. <http://research.chtsai.org/papers/pinyin-comparison-big5.html> (2013.12.10 檢索)
- 蔡秀俐. 2009. 《羅馬字 tī 台語文漢羅文本中 ê 使用分析—以台語文語料庫為基礎》. 國立台東大學台灣語文教師碩士班碩士論文
- 蔣為文. 2012. 〈是「台語」kap「咱人話」, m̄是「閩南話」!〉. 《台文通訊罔報》 219 期 p14-16
- 蔣為文. 2013. 〈教會內台語白話字使用人口 kap 現況調查〉. 《台語研究》 5(1) p74-97
- 鄭良偉. 1990. 《演變中的臺灣社會語文—多語社會及雙語教育》. 台北: 自立
- 鄭良偉. 1997. 《台語的語音與詞法》. 台北: 遠流

第二章 白話字 ê 編碼、搜揣 kah 顯示

電腦處理上，漢字 khah 無問題。本章 beh 針對白話字（除了標題以外，以下簡稱 POJ），介紹按怎處理 POJ，包括編碼、顯示，koh 有按怎搜揣。

第一節 白話字編碼

台語無官方地位，POJ koh m̄是唯一 ê 台語書寫系統，所以白話字符號申請加入 Unicode ê 過程並無順利。1997 年戴凱序向 Unicode 聯盟(consortium)申請將 POJ ê “o”加入 Unicode 字元集 (Te, Everson 1997)，中央 koh 有一寡過程，一直到 2004 年才成功 (ISO/IEC JTC1/SC2/WG2 2004; 台北市電腦公會 2003)。符號分散 tī 無全區，包括 *basic Latin*(0020-007F, English)、*Latin-1*(00A0-00FF, French, Germany)、*Latin Extended-A*(0100-017F: Latvian, Romanian, Polish, Lithuanian, Croatian, Esperanto, Maltese, Irish, Czech)、*Latin Extended-B*(0180-0233: Zulu, Pan-Nigerian, Zhuang, Acrican, Ewe, Pinyin)、*Latin Extended Additional*(1EF0-1EF9: Livonian, Vietnamese)、kah *Combining Diacritical Marks*(0300-306F)。(Lâu Kiát-gák 2002)

若申請 ê 字元會當用 Unicode 字元集內 ê 兩個字元抑是三个字

元門起來，in 就無 beh 批准新符號 ê 申請，根據這個政策，就 kan-na “o”正片頂面彼點（“ȯ”）有需要申請。提案 tī 2004 年 6 月通過。

根據這個編碼政策，一个 POJ 符號可能有兩款編碼方式，可比“á”，有可能是單一符號，mā 可能是“a” kah “ ”兩個符號鬥做伙，假使文本內全一个符號有兩種編碼方式，查詢 ê 時，其中一款編碼會揣無，這是 Unicode 等價（equivalence）問題。（Unicode 2013）

表 2 – 1 POJ Unicode 編碼

POJ	ASCII	Unicode 編碼	POJ	ASCII	Unicode 編碼
Á	A2	00C1、0041+0301	á	a2	00E1、0061+0301
À	A3	00C0、0041+0300	à	a3	00E0、0061+0300
Â	A5	00C2、0041+0302	â	a5	00E2、0061+0302
Ã	A7	0100、0041+0304	ã	a7	0101、0061+0304
Ä	A8	0041+030D	ä	a8	0061+030D
É	E2	00C9、0045+0301	é	e2	00E9、0065+0301
È	E3	00C8、0045+0300	è	e3	00E8、0065+0300
Ê	E5	00CA、0045+0302	ê	e5	00EA、0065+0302
Ë	E7	0102、0045+0304	ë	e7	0103、0065+0304
Ê	E8	0045+030D	è	e8	0065+030D
Í	I2	00CD、0049+0301	í	i2	00ED、0069+0301
Ì	I3	00CC、0049+0300	ì	i3	00EC、0069+0300
Î	I5	00CE、0049+0302	î	i5	00EE、0069+0302
Ī	I7	012A、0049+0304	ī	i7	012B、0069+0304
Ĭ	I8	0049+030D	î	i8	0069+030D
Ŭ	M2	1E3E、004D+0301	ŭ	m2	1E3F、006D+0301
Ų	M3	004D+0300	ŭ	m3	006D+0300
Ŵ	M5	004D+0302	ŭ	m5	006D+0302
Ŷ	M7	004D+0304	ŭ	m7	006D+0304
Ÿ	M8	004D+030D	ŭ	m8	006D+030D

純文字，無需要考慮編碼問題。

第二節 白話字 ê 兩款表示法

台語是聲調語言，爰有聲調符號標示聲調。假使韻母是複母音，基本原則是 kā 聲調符號 khng tī 上主要 ê 母音頂頭(a > o,e >u,i > m,n)，總是一寡文獻有無全 ê 寫法，像“oe”這個韻母，聲調符號有 ê 寫 tī “o”頂頭(親像 “óe”)，有 ê 寫 tī “e”頂頭(親像 “oé”)。這造成查詢 ê 困擾，咱若用“óe”來查，可能會揣無“oé”，除非經過標準化。另外是頂一節所講 ê POJ 符號編碼 m̄-nā 一款。

內部表示法若用純文字 (ASCII)，數字表示聲調，問題就解決，無需要特別處理 Unicode 等價問題。因為第一聲 kah 第四聲(入聲) 無標聲調符號，所以數字聲調有 2、3、5、7、8。另外，鼻音符號“n” kah 韻母“o”分別改做“N” (抑是“nn”) kah “ou”。

表 2-2 POJ 符號 ê 純文字表示

POJ 符號	純文字表示	舉例
ˊ	2	hái → hai2
ˋ	3	làng → lang3
ˊˊ	5	bêng → beng5
ˊˊˊ	7	siā → sia7
ˊˊˊˊ	8	jòah → joah8
ˋˋ	N / nn	taˋˋ → taN / tann
ˋˋˋ	ou	koˋˋ → kou

表 2-3 台語文三款表示法舉例

POJ	Chá-chêng tī Se-kok ū bó-lâng, ang-bó n̄ng-ê
-----	--

表 2-3 台語文三款表示法舉例

	sio-hóe-thòan ⁿ teh tō-òah; sī kèng-khiân hòk-sāi Chú, gâu hō khóan-thāi lāng. í-keng gō-cháp gōa hòe, iáu-bōe ū kiá ⁿ -jī. Sī chin sàh-hiong, ke-lāi bô pòan ⁿ -hāng mih, chí-ū chit-tè chhù thang khiā-khi, iok-liōng tát 600 gín nā-tiā ⁿ . Chiāh-chheng siōng-siōng khiām-khoeh.
數字調號 POJ	Cha2-cheng5 ti7 Se-kok u7 bou2-lang5, ang-bou2 nng7-e5 sio-hoe2-thoann3 teh tou7-oah8; si7 keng3-khian5 hok8-sai7 Chu2, gau5 ho7 khoan2-thai7 lang5. l2-keng gou7-chap8 goa7 hoe3, iau2-boe7 u7 kiann2-ji5. Si7 chin san3-hiong, ke-lai7 bo5 poann3-hang7 mih8, chi2-u7 chit8-te3 chhu3 thang khia7-khi2, iok-liong7 tat8 600 gin5 na7-tiann7. Chiah8-chheng7 siong5-siong5 khiam3-khoeh.
漢羅合用	早前 tī 西國有某人，翁某兩個燒火炭 teh 渡活；是敬虔服事主，gâu 號款待人。已經五十外歲，iáu-bōe 有子兒。是真散 hiong，家內無半項物，只有一塊厝 thang 企起，約量值 600 銀 nā -tiā ⁿ 。食穿常常欠缺。
資料來源：台語文數位典藏資料庫（第二階段） ⁶	

咱若對帶聲調 ê 符號排序，可比用 a、á、à、â、ā、á來排，結果順序會是 a（第一聲）à（第八聲）á（第二聲）à（第三聲）â（第五聲）ā（第七聲），m̄-koh 若用 a a2 a3 a5 a7 a8 來排，就會照聲調 ê 順序，這是數字調號 ê 另外一個好處。

若 beh 採用數字調號這個策略，咱需要一套 POJ kah 數字調號 POJ ê 轉換系統，內部轉做數字調號來 khng，顯示 ê 時 koh 轉做 POJ。

因為有一寡 POJ ê 符號是用兩個 Unicode 字元鬥起來 ê，一寡作業系統抑是字型無支援鬥字，結果，tī 瀏覽器無法度正常顯示

⁶ <http://ip194097.ntcu.edu.tw/nmtl/dadwt/index.htm>。

著。⁷

- (2) SQL 提供大小寫無分 (case insensitive) ê 搜揣，總是這是為著英文等 ê 文字所設計，對數字調號 POJ 來講，鼻音 ê “N” kah 鼻音韻尾 ê “n”並無全款，使用者查“tan”[單]，系統無應該 kā “taN”[擔]揣予伊。M̄-koh 若分大小寫 (case sensitive)， “M̄” kah “m̄” 無全 mā m̄著。

對頂面 ê 例咱 thang 知，若 beh kā POJ 搜揣功能做予好，袂使 kan-na 用一般 ê 字串比對 kah 大小寫無分/有分 ê 功能。

二、兩階段搜揣策略：先比對字串才 koh 篩選

咱提辭典 ê 搜揣做例，詞條 khng tī 資料庫 ê 資料表內底。若 beh 搜揣，SQL 指令通常是按呢寫：

“select * from dict_table where poj_field like ‘%input_str% ’ ”

初步 ê 搜揣結果可能有一寡是無需要 ê，若 beh 解決頂面 ê 問題(1)，咱愛 tī 音節層次 (m̄是語詞層次) koh 做一擺比對，若輸入 “a”，就愛 kā 查詢結果中 “a2”、“hah”、“khoan5”、...等等攞提掉。

頂面講 ê 問題(2)，若是鼻音用 “nn”表示，就無問題，若是用 “N”表示鼻音，就愛檢查這個音節倒頭 (ùì 正片) 算過來第一個非數

⁷ 教育部台灣閩南語常用詞辭典 ê 搜揣方式，會使用「精確查詢」kah「模糊查詢」，所舉 ê 例是用「模糊查詢」得著 ê 結果，若用「精確查詢」，kan-na 會當查著詞條是 “a” ê，無法度查著包含 “a” 這個音節 ê 所有語詞 (親像 “a-phian 鴉片”、“oo-a 烏鴉”、...等等)。阮認為 kā 字串比對講做是「模糊查詢」是無妥當 ê。

字（數字表示聲調）是 $\bar{m}$ 是“N”。另外，專有名詞第一字母大寫，查詢 $\hat{e}$ 時無需要分大小寫，這點 kah 英文全款，會使 $\bar{m}$ 免另外做處理。

所以，數字調號 POJ $\hat{e}$ 搜揣需要用兩階段來做，咱先用 SQL 指令得著初步 $\hat{e}$ 查詢結果，koh 寫程式篩掉 $\bar{m}$ 是咱 beh 愛 $\hat{e}$ ，得著最後 $\hat{e}$ 結果。下面是針對語詞 $\hat{e}$ POJ 比對演算法：

1. 執行 SQL 指令得著初步查詢結果
2. $k\bar{a}$ 查詢 $\hat{e}$ 字串 $kh\bar{h}\bar{g}$ 入 array **A**，一位 $kh\bar{h}\bar{g}$ 一音節
3. 每一位 $\hat{e}$ 第一字母， $k\bar{a}$ 改做小寫(lowercase **A**[i][0])
4. $m \leftarrow \text{array A } \hat{e} \text{ 長度 (音節數)}$
5. for each 初步搜揣結果
6. $k\bar{a}$ POJ 欄位內容 $kh\bar{h}\bar{g}$ 入 array **B**，一位 $kh\bar{h}\bar{g}$ 一音節
7. 每一位 $\hat{e}$ 第一字母， $k\bar{a}$ 改做小寫(lowercase **B**[i][0])
8. $n \leftarrow \text{array B } \hat{e} \text{ 長度 (音節數)}$
9. for $i = 0$ to $(n-m)$
10. if (**A**[0]..**A**[$m-1$]=**B**[i]..**B**[$i+m-1$]) then
11. 顯示 **B**
12. break
13. end if
14. next i
15. next

舉例來講，設使用 POJ 查詢“Sin-si”，利用 SQL 查著 $\hat{e}$ 初步結果有“siN-si2[生死]”、“siN-siuN3[生相]”、“siN-sian[生銑]”、“sin-si [新詩]”、“sin-si5[申時]”、“sin-sian[新鮮]”、“sin-siat[新設]”、“sin-siu3[新秀]”攏總八个，beh 做第二階段 $\hat{e}$ 篩選，**A**[0]=“sin”（第一字母改做

等等。搜揣 ê 功能若會當針對一般台語學習者 ê 困難點來做改進，會減少使用者搜揣過程 ê 困擾，增加 in 使用搜揣功能 ê 意願。

經過實務 ê 台語教學，有一寡是學習者感覺 khah 困難 ê 所在，包括：

- (1) 聲調問題：華語有四個聲調，m̄-koh 台語有七個；
- (2) 束喉音：束喉音(-h)是弱化 ê 入聲，甚至有部分台語人 ê 發音已經漸漸感覺缺著束喉音，有 ê 人無啥會曉分辨“a” kah “ah” ê 差異；
- (3) 入聲：華語無入聲，m̄-koh 台語有入聲，而且有四種入聲（-p|-t|-k|-h），有 ê 人無啥會曉分辨。

若是 POJ ê 搜揣，會當允准使用者免輸入聲調，免指定是 m̄ 是束喉音，抑是免指定是佗一款入聲，就會當查著伊想 beh 愛 ê 結果，按呢對使用者是加真體貼 ê 設計。

M̄免輸入聲調 ê 搜揣，會使 kah 無輸入束喉音-h ê 搜揣做伙設計（對一寡學習者來講，“ah” kah “à”無好分辨，“àh” kah “á”抑是“a” mā 無好分辨）。對一個音節來講，上尾後 ê 部分是“φ|2|3|h|5|7|h8”（φ 表示無數字，就是第 1 聲），以外 ê 部分若相全，就攞 kā 揣出來。

下面是無輸入聲調 ê POJ 搜揣演算法（設使使用者有輸入聲調，mā 會使 mài 管伊輸入 ê 聲調），假設輸入是一個音節：

1. if 上後壁是“[1|2|3|h4|h|4|5|7|8|h8]” then kā 提掉

2. $str \leftarrow$ 修改過 ê 使用者輸入
3. 執行 SQL 指令
`“select * from dict_table where poj_field like ‘str%’”`
4. for each 初步查詢結果
5. $x \leftarrow$ 查詢結果 *poj_field* 後壁 ê 部分 (*poj_field* – *str*)
6. if $x = [\phi | 2 | 3 | h | 4 | 5 | 7 | 8 | h8]$ then 顯示
7. next

舉例來講，若輸入“ka3”，先 kâ 後壁 ê “3”提掉， $str = “ka”$ ，了後用 SQL 指令查，包括“ka”、“ka2”、“ka3”、“ka5”、“ka7”、“kah”、“kah8”、“kai”、“kai2”、...、“kak”、“kak8”、“kam”、“kam2”、...、“kan”、“kan2”、...、“kang”、“kang2”、...、“kap”、“kap8”、“kat”、“kat8”、“kau”、“kau2”、...等等，經過篩選，符合條件 ê 結果是“ka”、“ka2”、“ka3”、“ka5”、“ka7”、“kah”、“kah8”。

第 1~2 逝，假使輸入 ê 最後部分是[1|2|3|h4|h|4|5|7|8|h8]，就 kâ 這部分提掉，修改後 ê 輸入定做 str ，雖然第 1 聲 kah 第 4 聲會使免寫數字，總是先假設有 ê 輸入者可能會輸入 1 抑是 4；第 3 逝，執行 SQL 指令，揣出 POJ 欄位包含 str ê 資料；第 4~7 逝針對初步查詢結果隨个隨个檢查，查詢結果 *poj_field* 扣去 str ，chhun--ê 後壁 ê 部分，若是[$\phi | 2 | 3 | h | 4 | 5 | 7 | 8 | h8$]表示是符合 ê 資料，就 kâ 顯示出來。

以上是無輸入聲調 ê POJ 音節搜揣，若是 beh 用這個方法搜揣 POJ 語詞，會小可 khah 複雜，爰先分解做音節處理，才 koh 鬥起

韻母搜揣會比進前提起 ê 演算法 khah 複雜。查詢者若是 beh 揣鼻音，可比“aN”，m̄是 kan-na 揣出包含“aN” ê 音節就好，koh 需要揣“[ma|nana|nga]”開頭 ê 音節。另外愛注意 ê 是“m”、“n” kah “ng”，in 會使是聲母，mā 會使是韻尾 ê 部分，假使“m”抑是“ng” tī 音節中單獨出現，in 是韻母（台語一个合法 ê 音節會使無聲母，m̄-koh 袂使無韻母）。另外，韻母搜揣 mā 是無需要輸入聲調。

下面是以音節做單位 ê POJ 韻母搜揣演算法：

1. 對正片到倒片，kā 輸入 ê 音節拆做聲調、韻母、聲母三部分，分別指定做 **st**、**ub**、**sb**（kan-na 需要 **ub**）
2. if right(**ub**,1)= “N” then //鼻音處理，有 ê 鼻音成份 tī 聲母
3. **ubx**=left(**ub**,length(**ub**)-1);**ub1**=“m”+**ubx**;**ub2**=“n”+**ubx**;**ub3**=“ng”+**ubx**
4. 執行 SQL “select * from dict_table where poj_field like ‘%**ub**%’ or poj_field like ‘%**ub1**%’ or poj_field like ‘%**ub2**%’ or poj_field like ‘%**ub3**%’”
5. else “select * from dict_table where poj_field like ‘%**ub**%’”
6. enf if
7. for each 初步查詢結果
8. l ← POJ 欄位 **ub** or **ub1** or **ub2** or **ub3** 倒片 ê 部分
9. r ← POJ 欄位 **ub** or **ub1** or **ub2** or **ub3** 正片 ê 部分
10. if l = “[φ|p|ph|m|b|t|th|n||k|kh|g|ng|h|ch|chh|s|j]” then
11. if r = “[φ|2|3|5|7|8]” then 顯示
12. end if
13. next

第 1 逝，kā 輸入 ê 音節用倒頭 ê 順序拆做聲調、韻母、聲母三部分，咱 kan-na 需要韻母 ê 部分，後壁會 khah 詳細來說明伊 ê

演算法；第 2~5 逝，根據是 $\tilde{m}$ 是有鼻音執行對應 ê SQL 指令，揣出 POJ 欄位包含[**ub**(|**ub1**|**ub2**|**ub3**)] ê 資料，**ub** 若帶鼻音，可比是 “iN”，就會指定 **ub1**= “mi”，**ub2**= “ni”，**ub3**= “ngi”（後壁 ê “N”提掉，頭前分別加“m”、“n”、“ng”），這四个字串攏入去查；第 7~13 逝針對初步查詢結果隨个隨个檢查，先 kâ **ub**（抑是 **ub1,ub2,ub3**）倒片（聲母）kah 正片（聲調）ê 部分分出來，假使倒片是[$\phi|p|ph|m|b|t|th|n||k|kh|g|ng|h|ch|chh|s|j$]而且正片是[$\phi|2|3|5|7|8$]，這是咱 beh 愛 ê，就 kâ 顯示出來。

比一个例，假使輸入“a”來進行韻母搜揣，經過初步查詢結果，“pa2”、“kia7”、“khang”、...等等攏會揣著，這三个 kan-na “pa2”是符合 ê（“p”是聲母“2”是聲調）， $\tilde{m}$ -koh “kia7” kah “khang” $\tilde{m}$ 是（“kia7”ê 倒片“ki” $\tilde{m}$ 是聲母，“khang”ê 正片“ng” $\tilde{m}$ 是聲調）。

四、實作結果舉例

用無聲調羅馬字查 詞條是[**hoe-chhia**] e 結果

Chhih白話字iah是台語漢羅會tàng進去台語文語詞檢索系統，查台語文語料庫內底這個詞ê前後文。

	台語羅馬字	漢羅	華文	英文
1	h^hoe-chhia / hé-chhia	火車	火車	railway train
2	h^hoe-chhiah / hé-chhiah	火鋸	火鑪	
3	h^hoe-chhia / hè-chhia	貨車	貨車	truck; freight-car; freight-train;
4	hoe-chhia	花車	花車	vehicle decorated with flowers (used for funerals)

圖 2 - 2 台華辭典用「無聲調羅馬字」查“hoe-chhia”

提台華線頂辭典 kah 台語線頂字典實際查詢 ê 例。台華線頂辭典用「無聲調羅馬字」功能查“hoe-chhia”，結果會像圖 2 - 2 。

第四節 白話字顯示

一、白話字文本顯示相關問題

根據頂面所講，咱用數字調號 POJ 做內部表示 ê 款式，m̄-koh 顯示 ê 時猶原應該愛顯示正式 ê POJ，所以咱需要 POJ kah 數字調號 POJ 互相之間 ê 轉換系統。

以 Windows 作業系統來講，Windows 7 作業系統了後，真濟字型攏有支援 Unicode 兩字鬥起來 ê 政策，所以大部分情形下，POJ 攏會當正常顯示，若是 Windows XP 作業系統抑是部分應用程式，無支援 Unicode 鬥字 ê，顯示就會有一寡問題，第二個符號，會變做四角格仔。

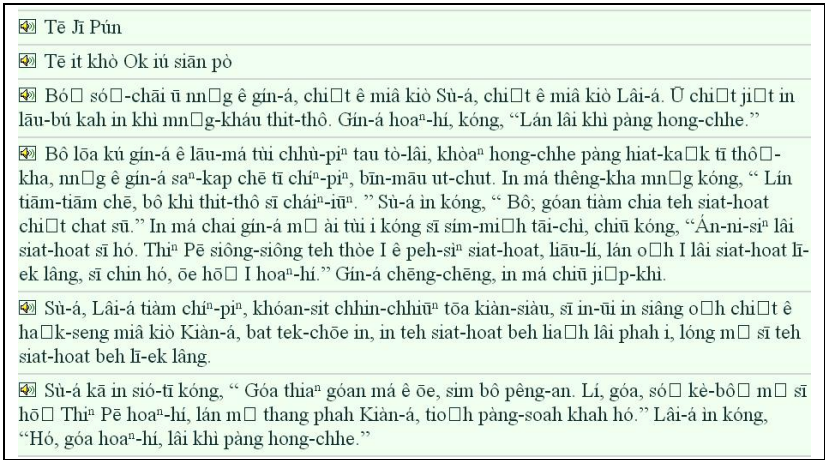


圖 2 - 5 Windows XP 作業系統頂面 ê POJ 顯示

Ti Windows XP 作業系統若拄著 POJ 顯示問題，會當安裝有支援 Unicode 編碼政策 ê 免費字型，親像 Taigi Unicode、Doulos SIL、Charis SIL、Gentium、DejaVu、...等等 ê 字型。(楊允言 2006b; Laenen et al. 2004; 劉杰岳 2005; SIL 2005a; SIL 2005b; SIL 2005c)

M-koh，m̄是每一个使用者攏對電腦相關操作真熟手，提供免費字型無表示使用者就會當順利 kā 安裝安裝；而且瀏覽 ê 網頁內底 ê 字型 mā m̄是咱會使去修改 ê；另外，使用者 mā 無一定是用家己 ê 電腦，有 ê 公用電腦，使用者並無權限來安裝字型。換一个思考方向，若是利用圖形來顯示 POJ，這個問題就會當解決。

二、白話字 kah 數字調號白話字轉換方法

若 beh kā POJ 轉做數字調號 POJ，是用音節做單位，一个音節一个音節做落去，先 kā 聲調符號改用數字表示，了後徙到這音節上尾 ê 位。可比“lâng”，咱先 kā “â”改做“a5”，變做“la5ng”，了後 kā “5”徙去上後壁，變做“lang5”。

lâng→ la5ng→ lang5

下面是 POJ 轉數字調號 POJ ê 演算法：

1. for each POJ 音節
2. kā 聲調符號轉做對應 ê 數字
3. kā 數字徙去上後壁
4. next

nng7	<table><tr><td>o</td><td>a</td><td>e</td><td>u</td><td>i</td><td>n</td><td>m</td></tr><tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td><u>2</u></td><td>0</td></tr></table>	o	a	e	u	i	n	m	0	0	0	0	0	<u>2</u>	0	nn7g	nñg
o	a	e	u	i	n	m											
0	0	0	0	0	<u>2</u>	0											
mi5	<table><tr><td>o</td><td>a</td><td>e</td><td>u</td><td>i</td><td>n</td><td>m</td></tr><tr><td>0</td><td>0</td><td>0</td><td>0</td><td><u>2</u></td><td>0</td><td>1</td></tr></table>	o	a	e	u	i	n	m	0	0	0	0	<u>2</u>	0	1	mi5	mî
o	a	e	u	i	n	m											
0	0	0	0	<u>2</u>	0	1											
koai2	<table><tr><td>o</td><td>a</td><td>e</td><td>u</td><td>i</td><td>n</td><td>m</td></tr><tr><td><u>2</u></td><td>3</td><td>0</td><td>0</td><td>4</td><td>0</td><td>0</td></tr></table>	o	a	e	u	i	n	m	<u>2</u>	3	0	0	4	0	0	ko2ai	kóai
o	a	e	u	i	n	m											
<u>2</u>	3	0	0	4	0	0											

另外，咱 mā 需要 kā “nn”（抑是“N”）kah “ou”轉做對應 ê 符號。調符 ê 位確定了後，若是這音節上後壁兩個字母是“nn”抑是上尾 ê 字母是“N”，就 kā 轉做“n”。可比講：

phainn2 → pha2inn → pháinn → pháinⁿ

愛注意 ê 是，咱袂使 kā 一个音節開頭 ê “nn”抑是“N”提來換，一个音節頭前 ê “nn”是“nng” ê 一部分，第一个“n”是聲母，第二个“n”是韻母 ê 一部分；另外，一个音節開頭 ê “N”，應該是專有名詞所以大寫，m̄是代表鼻音韻尾。

“ou” ê 部分，因為頭前 ê 做法會 kā 數字 khng tī “o”後壁，會造成轉換錯誤（“ou3”→“o3u”→“òu”，正確 ê 符號是“ò”）；咱會使先 kā “o<num>u”轉做“ou<num>”才轉調符。可比下跤 ê 例：

kounn5 → ko5unn → kou5nn → kôⁿ

若是感覺這個方法傷費氣，實務上，上簡單 ê 做法是查表來轉換，對長度上長 ê （可比 “iong2”→“íong”、 “iong3”→“ìong”、...）一直到上短 ê （可比 “a2”→“á”、 “a3”→“à”、...）隨个隨个比對然

後轉換，愛注意 ê 是，短 ê 袂使先轉換，若無，“iau3”有可能予轉做“iaù”（應該是“iàu”）。

三、白話字圖形顯示

咱若 beh tī 網站頂面顯示 POJ，koh 無想 beh 勞煩瀏覽者安裝任何字型，會使考慮用圖形顯示。

觀念上並無困難，只要 kā 每一個 POJ 符號換做對應 ê 圖形就會使。因為網頁可能有背景圖抑是背景色，POJ 符號 ê 圖形上好是透明 ê，所以建議用 GIF 抑是 PNG 格式。

楊允言(2006a)有收集一寡 PNG 格式 ê POJ 符號 ê 圖。下面是數字調號 POJ 轉圖形 ê 演算法

1. for each 數字調號 POJ 音節
2. kā 數字徙去主要元音彼个字母後壁
3. kā 所有 POJ 符號轉做對應 ê 圖形
4. next

四、白話字顯示舉例

下面兩個圖展示 POJ 文本兩款無仝 ê 顯示策略。第一張圖是文字，Windows 7 作業系統頂面顯示並無問題，m̃-koh Windows XP 作業系統就無 hiah-nī 好勢，網頁 ê 文字會使鉸落來貼去別位；(楊允言 2006b) 第二張圖是圖形，所以 kah 漢字鬥做伙 ê 時，會感覺

SIL. 2005b. Doulos SIL Font Home. <http://scripts.sil.org/DoulosSILfont>
(2013/11/20 檢索)

SIL. 2005c. Gentium - a typeface for the nations. <http://scripts.sil.org/Gentium> (2013/11/20 檢索)

Te, K.-s., Everson M. 1997. N1593: Proposal to add Latin characters required by Latinized Taiwanese languages to ISO/IEC 10646. <http://www.evertype.com/standards/la/taioan.html> (2013/12/10 檢索)

Unicode. 2013. Unicode Standard Annex #15 Unicode Normalization Forms. <http://unicode.org/reports/tr15/> (2014/1/10 檢索)

台北市電腦公會. 2003. N2628: Proposal to add COMBINING RIGHT DOT ABOVE as required by Taiwanese Holo language Romanization to ISO/IEC 10646. <http://std.dkuug.dk/JTC1/SC2/wg2/docs/n2628.pdf>
(2013/9/20 檢索)

楊允言,張學謙,呂美親編. 2008. 《台語文運動訪談暨史料彙編》. 台北: 國史館

楊允言. 2006a. 〈白話字字母圖形〉. <http://ip194097.ntcu.edu.tw/TG/POJJB/POJjb.asp> (2013/11/20 檢索)

楊允言. 2006b. 台語文數位典藏資料庫 (第二階段) (備站). <http://ip194097.ntcu.edu.tw/nmtl/DADWT/pbk.asp> (2013/12/10 檢索)

劉杰岳. 2005. Taigi Unicode Font. http://taigi.fhl.net/TP/taigi_unicode_font.htm (2013/11/20 檢索)

國立台灣文學館. 2006. 台語文數位典藏資料庫 (第二階段). <http://xdcm.nmtl.gov.tw/dadwt/pbk.asp> (2013/12/10 檢索)

第三章 台語文處理 ê 一寡家私

這章 beh 介紹台語文處理 ê 一寡家私 (utilities)，包括 POJ 聲韻調分拆，POJ 拼字檢查，POJ 文本音節數、語詞數、語句數統計，漢羅合用文本 ê 音節數統計，kah 漢羅合用文本 ê 斷詞等等。Chia ê 家私，是阮 teh 處理台語文文本 ê 時，因為實際 ê 需要所做出來 ê，mā 提供予逐家來使用。

第一節 白話字聲韻調分拆 kah 拼字檢查

對一个文字系統來講，拼字檢查是真要緊 ê 家私。阮針對 POJ，做拼字檢查系統，幫贊 POJ 文本 ê 整理 kah 校對，因為一个台語音節會當分做聲母、韻母、聲調三部分，做 POJ 拼字檢查 ê 時，是 kā 一个音節拆做這三部分，來確認這個音節是 m̄ 是合法 ê 音節。

Khah 正規 ê 做法，會使用有限狀態機 (finite state machine) ê 技術來實作，先畫出狀態轉換圖 (state transition diagram)，了後 kā 轉做程式，總是按呢做有 khah 囉嗦。(Carroll, Long 1989) 另外一个 khah 簡單 ê 做法是 kā 所有合法 ê 音節攏列做一个表，利用查表 ê 方式來判斷是 m̄ 是合法 ê 音節，假使 beh 按呢做，這個表大概有 3,000 筆資料。

kā 一个 POJ 音節分拆做聲母、韻母 kah 聲調三个部分來判斷是 m̄ 是合法 ê 音節，是相對 khah 簡單 koh 有效率 ê 做法，總是伊 m̄ 是無缺點，合法 ê 聲母、韻母、聲調鬥起來 ê 音節，未必是合法 ê，親像“man7”並 m̄ 是台語合法 ê 音節，m̄-koh “m”（聲母）、“an”（韻母）kah “7”（聲調）攏合法。若是 beh 解決這個問題，愛加一寡規則入來才處理會好勢。

合法 ê 台語音節，會使無聲母，m̄-koh 袂使無韻母，咱若照倒片到正片 ê 順序看，有一寡部分咱無法度判斷伊到底是聲母抑是韻母，可比咱若看著頭前 ê 部分“ng”，假使這個音節是“nga2”，按呢“ng”是聲母，設使這個音節是“ng2”，按呢“ng”是韻母。若是用倒頭 ê 方向來檢查，照倒頭 ê 順序分別揣出聲調、韻母 kah 聲母，按呢就無這欸問題。

對每一個數字調號 POJ ê 音節來講，若是對正片看到倒片，第一部分是聲調，可能是一個數字（[2|3|5|7|8]）抑是無數字（第 1 抑是第 4 聲）；第二部分是韻母，beh kā 韻母揣出來，愛對長度 khah 長 ê 到 khah 短 ê 順序來揣，可比“iau”（長度 3）愛排 tī “u”（長度 1）頭前，才袂先揣出“u”這個韻母 soah kā “ia”留落來；第三部分是聲母，mā 有可能無。

下面是檢查一个 POJ 音節是 m̄ 是合法 ê 演算法：

1. //對正片到倒片 kā 數字調號 POJ 音節拆做聲調、韻母、聲母
2. lowercase (第 1 个字母)
3. // 聲調

```

29.      siannbo ← 這部分
30.      else
31.          return false //無合法
32.      end if
33.      return true //合法

```

先處理聲調，第 3~12 逝 kā 分做四種情形：(a) 若上尾是數字，而且是 [1|2|3|4|5|7|8] (雖然 1 kah 4 會使免寫，為著 khah 周至，mā kā 考慮入來)，這個數字就是這音節 ê 聲調，數字 kā 提掉；(b) 上尾 ê 部分若是 [hN|p|t|k|h] (入聲)，聲調就是 4；(c) 上尾 ê 字元是 [N|a|e|g|i|m|n|o|u]，聲調就是 1；(d) 其它 (無合法)。

第 13~26 逝確認韻母，咱 kā 所有 ê 韻母照長度長到短 ê 順序一个一个 kā 檢查，若是對有，這部分就是韻母，koh kā 提掉；若對無，表示這 m̄ 是合法 ê 音節。

第 27~32 逝確認韻母，chhun 落來 ê 部分若是 [ɸ|p|ph|m|b|t|th|n||k|kh|g|ng|h|ch|chh|s|j]，這部分就是聲母，若無，就 m̄ 是合法 ê 音節。

若 m̄ 是合法 ê 音節，會 tī 第 11、25、31 逝跳離開，行到 32 逝，表示這是合法 ê 音節。

第二節 白話字文本音節/語詞/語句統計

MicroSoft Office 軟體內底 ê Word，有提供字數統計 ê 功能，m̄-koh 伊提供 ê 功能，對 POJ 文本並無全然適用，伊 kan-na 會當

算有幾個語詞，咱受漢字影響，實務上，POJ 文本投稿，若 beh 算稿費，應該是算音節數，m̄是算字元數，mā m̄是算語詞數。

POJ 基本上 kah 英文 khah 倚，語詞 ê 界線是空白(包括<Tab>、<Enter>等等)、標點、括號等等，音節 ê 界線除了頭前所講 ê，koh 加連字符(hyphen)。有一个愛注意 ê 所在是連繼 ê 數字，“1234”會當讀做“it-jī sam-sù”(兩個語詞、四個音節，咱讀數字 ê 時，通常攞是 kā 切做兩個、三個抑是四個音節做一個變調分界組)，mā 會使讀做“chit-chheng n̄ng-pah saⁿ-cháp-sì”(三個語詞，七個音節)。數字 ê 問題，有需要做 koh-khah 幼路 ê 分析，m̄-koh 為著利便，咱暫時先 kā “1234”當做是一個語詞、四個音節。

若 beh 算語句就 khah 費氣，句點(“.”)會使算是語句 kah 語句 ê 分界，m̄-koh 連繼 ê 句點(“...”) m̄是分界，“<Enter>”有時是句 ê 分界，有時 koh m̄是，“?”、“!”、“!?”有時是句 ê 分界，有時 soah koh m̄是。

為著執行計畫，beh 統計文本數量，劉杰岳設計一個 POJ 文本音節/語詞/語句/段落統計 ê 家私，網址 tī <http://ip194097.ntcu.edu.tw/nmtl/dadwt/count.asp>，音節數 kah 語詞數基本上 khah 無問題，阮用[.|?|!|!|?]來判斷是 m̄是一個語句，用“<Enter>”來判斷是 m̄是一個段落，雖然用按呢來判斷 m̄是全然正確，總是提供一個參考 ê 數字。

第四節 漢羅合用文本斷詞

斷詞系統是咱 beh 針對漢羅合用台語文做處理，真基本 ê 家私，其實全漢字文本 mā 會使包含 tī 這內底。

漢羅合用台語文ê斷詞，比全漢字ê斷詞加khah複雜，⁸一个漢字基本上mā是一个音節（至少台灣是按呢），POJ ê一个字元（以Unicode ê角度，一个漢字mā是一个字元）通常是一个音節ê一部分，tī數字調號POJ，一个音節上長有 8 个字元（親像“chhiang5”）。tī漢羅合用文本，一个POJ音節ê分界，除了空白（空白包括<Tab>、<Enter>等）、標點、連字符以外，koh有漢字。

咱若會當 kā 漢羅合用文本先分拆做一个一个 ê 音節，觀念上就會使用類似華文 ê 斷詞系統來處理，總是按呢看起來並無 khah 好做。實務上，咱會使暫時 mài 管音節，就 kā 一个語詞當做是一个字串來處理。斷詞系統需要有辭典 ê 詞條做分詞 ê 基礎，阮利用台華線頂辭典 ê 詞條，內底攏總有 62,000 外條，其中，上長 ê 詞條是 20 字元（“ngiauh8-ngiauh8-soan”，三音節）。

阮利用台華線頂辭典 ê 詞條 kah 倒頭上大比對演算法（backward maximal matching algorithm，簡稱 BMM），實作一个線頂 ê 漢羅合用文本斷詞系統，網址 tī http://poj.likulaw.info/hanlo_hunsu.php。下面是漢羅合用文本斷詞倒頭上大比對演算法：

⁸ 親像華文，當然華文文本內底有可能摻一寡 abc，in 是當做外國字，直接跳過無處理。

1. Initial : **S1** 是輸入字串, **S2** 是空字串
2. while **S1** is not empty
3. **w** = right (**S1**, *MaxLen*)
4. while Lookup(**w**)==*false* and IsNotPOJWord(**w**) and
 length(**w**)>1
5. **w** = right (**w**, length(**w**)-1)
6. end while
7. **S2** = '[' + **w** + ']' + **S2**
8. **S1** = Left (**S1**, length(**S1**)-length(**w**))
9. end while
10. output **S2**

S1 輸入 ê 字串，以標點做分隔 ê 單位，一份文本若有 10 个標點，就 kā 切做 10 部分來輸入；**S2** 是輸出（斷詞 ê 結果）。*MaxLen* 是詞條上長 ê 字元數（20）；第 3 逝，掠 **S1** 正片 *MaxLen* 長度 ê 子字串（substring）做 **w**，第 4~6 逝，就 kā 子字串提來 kah 所有詞條比對，若比著，表示子字串是一个詞，就 kā 輸出，而且對 **S1** kā 這個子字串提掉，若是無比對著，就 kā **w** 倒片一个字元提掉，koh 入去比，一直比到 **w** ê 長度到 1 為止。一直比到 **S1** 變做空字串才停落來，這時，**S2** 就是斷詞了後 ê 結果。

咱直接用實例來說明，kā “Ui3 台語聖經看台語語詞變化”這段文字做斷詞，這段字串 ê 長度是 14，比 *MaxLen* (=20) khah 細，所以一開始 **w** = “Ui3 台語聖經看台語語詞變化”，m̄-koh 這個字串 tī 詞條內底揣無，所以 kā 第 1 字元提掉，**w** 變做“i3 台語聖經看台語語詞變化”，詞條全款揣無，**w** ê 第 1 个字元隨个隨个提掉，一直

輸入 ê 文本：

過去 chit 9 工，我 tak8 工 chhoe7 攔頭延緩離開 e5 時間，一直到
今仔日天光，Koxinga 講 e5 對期到，就準 kui 暝無眠，心內 koh-khah
an2-choaN2 艱苦，透早，ma7 是 tioh8 chhoa7 最後 1 陣阿兵哥舉
旗 a2、pun5 鼓吹，齊齊整整、英威驕 kiah 行 ng3 pha 碇 ti7 台江
內海 e5 Erasmus，這 ma7 是荷蘭 ti7 大員 e5 上尾一隻軍艦。

使用者詞庫：

驕 kiah
pha 碇
台江
大員

斷詞結果1：

[過去][chit][9][工]，[我][tak8][工][chhoe7][攔頭][延緩][離開][e5][時間]，[一
直到][今仔日][天光]，{Koxinga}[講][e5][對期][到]，[就][準][kui][暝][無眠]，
[心內][koh-khah][an2-choaN2][艱苦]，[透早]，[ma7是][tioh8][chhoa7][最後]
[1][陣][阿兵哥][舉]{旗a2}、[pun5][鼓吹]，[齊齊][整整]、[英威][驕kiah][行]
[ng3][pha碇][ti7][台江][內海][e5]{Erasmus}，[這][ma7是][荷蘭][ti7][大員][e5]
[上尾][一][隻][軍艦]。

斷詞結果2（用空格隔開）：

過去 chit 9 工，我 tak8 工 chhoe7 攔頭 延緩 離開 e5 時間，一直到 今仔日
天光，Koxinga 講 e5 對期 到，就 準 kui 暝 無眠，心內 koh-khah an2-
choaN2 艱苦，透早，ma7 是 tioh8 chhoa7 最後 1 陣 阿兵哥 舉旗 a2、pun5
鼓吹，齊齊 整整、英威 驕kiah 行 ng3 pha碇 ti7 台江 內海 e5 Erasmus，這
ma7 是 荷蘭 ti7 大員 e5 上尾 一隻 軍艦。

圖 3-2 線頂漢羅台語文斷詞系統結果展示⁹

本章參考書目

俞士汶編. 2003. 《計算語言學概論》. 中國北京:商務

曾金金. 1997. 〈台語斷詞原則討論〉. 《台灣文學出版物收集、目錄、
選讀編輯計畫結案報告》p45-72. 台北:文建會 [http://ip194097.
ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm](http://ip194097.ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm) (2013/12/30 檢索)

⁹ 這是劉承賢 ê 作品。

第四章 台語變調處理

台語有複雜 ê 變調系統，漢字表音 ê 功能 khah 弱，而且一个漢字可能有幾 lō 个讀音，所以咱 beh kā 台語文轉做聲音，用 POJ 會比漢字效果加 khah 好，m̄-koh POJ 是標本調（深層結構），tī 轉做聲音（表面結構）進前，咱需要處理變調問題。本章根據台語變調 ê 規則 kah 阮對台語語料 ê 觀察，提出規則式 ê 變調處理演算法，來實作一个 POJ 文本轉做聲音 ê 系統。

台語變調處理系統是為著執行國立台灣文學館 ê 委託計畫，所做出來 ê 成果。基本 ê 想法是，大多數台灣人並無受台語讀寫 ê 訓練，雖罔有濟濟台語文本，假使無法度讀，chia ê 文本袂當發揮伊 ê 功能，咱若有法度予 chia ê 文本發聲，會當幫贊 chia ê 讀者運用 chia ê 豐富 ê 文本資源。換另外一個角度來看，助讀器 ê 開發，對語言學習是真有幫贊 ê。

第一節 台語 ê 變調問題

台語文若 beh kā 轉做聲音，變調問題會使講是上關鍵 ê 因素。每一个音節，攏有伊本來 ê 聲調，m̄-koh 若 kā chia ê 音節鬥做一句話 ê 時，有連讀變調 ê 問題，有時一句話 kan-na 一个變調分界組，除了最後一个音節以外，其它攏讀變調，m̄-koh 有 ê 一句話內底就

有幾 lō 个變調分界組。咱有必要 kā chia ê 問題先變清楚。

台語 ê 聲調，根據傳統 ê 講法，有平、上、去、入[piāⁿ, chiūⁿ/sióng, khì, jip], koh 分陰陽，m̄-koh 上聲無分陰陽，所以攏總有七個聲調，根據陰平、上、陰去、陰入、陽平、陽去、陽入 ê 順序，若用數字表示，分別是 1（懸平）、2（懸降）、3（低）、4（中短）、5（低升）、7（中平）、8（懸短），刮號內底是對這個聲調 ê 懸低、長短 ê 描寫。

一、台語變調種類

咱來討論上基本 ê 規則變調。先看下面 ê 例：

tài 台

Tài-gí 台語

Tài-gí-bûn 台語文

Tài-gí bûn-hák 台語文學

Tài-gí bûn-hák-sú 台語文學史

畫底線 ê，表示這個音節讀本調，無畫底線 ê 就愛讀變調。若照順序讀落來，攏是最後一音節讀本調，其它攏讀變調，這是規則變調。

歸納起來，台語攏總有七種變調，下面一个一个來介紹。

- (1) 規則變調：用疊[tháh]詞來做說明，音節後壁上標 ê 數字是實際讀出來 ê 聲調。

例 (i) 「Tân⁵--sian³-sin³ (sin³-seⁿ³)陳先生」

(「sian⁷-sin¹ (sin⁷-seⁿ¹)先生」原來聲調是 7,1)

(ii) 「kiāⁿ⁵--chhut⁴-lâi³ 行出來」(「chhut⁸-lâi⁵ 出來」原來聲調是 8,5)

(4) 再變調：大部分出現 tī 束喉音(-h) 4 聲，規則變調兩擺 (4→2→1)，變調了後讀第 1 聲。

例 (i) 「beh¹ thák⁴-chu¹ 欲讀書」(「beh 欲」4 聲應該變第 2 聲，實際變第 1 聲)

(ii) 「khi¹ gōa³-kháu² 去外口」(「khi 去」第 3 聲應該變第 2 聲，實際變第 1 聲)

(5) á[仔]前變調：á 前 ê 音節，變調情形真規則，m̄-koh kah 規則變調 koh 小可無全。

(i) 1 聲→7 聲：「sun⁷-á² 孫仔」

(ii) 2 聲→1 聲：「cháu¹-á² 草仔」

(iii) 3 聲→1 聲：「tàⁿ¹-á² 擔仔」(kah 規則變調無全)
(規則變調 3→2 eg: 「擔頭 tàⁿ²-thâu⁵」)

(iv) 4 聲→8 聲(-p/t/k)：「tek⁸-á² 竹仔」

4 聲→1 聲(-h)：「thih¹-á² 鐵仔」(kah 規則變調無全)
(-h 規則變調 4→2 eg: 「thih²-ki¹ 鐵枝」)

- (v) 5 聲→7 聲：「lô⁷-á² 爐仔」(kah 規則變調無啥全)
 (規則變調 5→7/3 eg: 「lô^{7/3}-chú² 爐主」)
- (vi) 7 聲→7 聲：「bāng⁷-á² 網仔」(kah 規則變調無全)
 (規則變調 7→3 eg: 「bāng³-lō⁷ 網路」)
- (vii) 8 聲→7?聲(-p/t/k)：「chhàt^{7?}-á² 賊仔」(kah 規則變調無全)
 (-p/t/k 規則變調 8→4 eg: 「chhàt⁴-thâu⁵ 賊頭」)
- 8 聲→7 聲(-h)：「chiòh⁷-á² 石仔」(kah 規則變調無全)
 (-h 規則變調 8→3 eg: 「chiòh³-thâu⁵ 石頭」)

(6) 三連音變調：三連音疊詞 ê 第 1 音節，第 2、3、4 聲 kah 規則變調全款，其它 ê 無全。

- 例 (i) 1 聲→5 聲：「chheng⁵-chheng⁷-chheng¹ 清清清」¹⁰
- (ii) 2 聲→1 聲：「ún¹-ún¹-ún² 穩穩穩」
- (iii) 3 聲→2 聲：「hèng²-hèng²-hèng³ 興興興」
- (iv) 4 聲→8 聲(-p/t/k)：「sip⁸-sip⁸-sip⁴ 濕濕濕」
 4 聲→2 聲(-h)：「bah²-bah²-bah⁴ 肉肉肉」
- (v) 5 聲→5 聲：「kôaⁿ⁵-kôa^{n7/3}-kôaⁿ⁵ 寒寒寒」¹¹

¹⁰ 有 ê 人第一音節讀做 “chheng⁷-éng²” ê 合音。

¹¹ 有 ê 人第一音節讀做 “kôaⁿ⁷-aⁿ¹” ê 合音。

二、變調組分界

頂面所整理 ê 台語變調，主要是以語詞抑是詞組做單位，若是到語句 ê 層次，看起來是 koh-khah 複雜。

變調組大概是以詞組做單位，總是按呢講並 m̄ 是足精確 ê 講法，可比講一个詞組若是包含另外一个詞組，內底彼个詞組敢是一个變調組？另外，“ê” mā 是一个真重要 ê 標記，因為“ê”頭前彼个音節（設使 m̄ 是人稱代名詞，親像：我、你、伊、阮、恁、in）是變調組 ê 界線。

假使咱有句法剖析器（parser），問題可能就解決一大半 ah，可惜咱無。咱需要 kā chia ê 變調規則條列整理出來，才 thang 繼續做落去。

參考一寡台語變調 ê 冊，m̄-koh 並無看著完整 ê 台語變調處理 ê 通則。（洪惟仁 1985; 鄭良偉 1997; 盧廣誠 1999; Chen 2000; 駱嘉鵬 2008）所以阮決定直接揀一寡文本來讀，對 chia ê 文本內底，揣出一寡變調處理 ê 規則。

實際提文本來看 ê 時陣，阮發現 beh 設計一个 kā 語句分做幾個變調組，並無真簡單，m̄-koh 詞性（part of speech）、連字符抑是某一類 ê 語詞，會當提供變調 ê 線索，若會當 kā chia ê 線索轉做變調規則，對處理變調問題是有幫贊 ê。

第二節 台語文發音系統實作

一、系統 ê 架構

輸入是數字調號 POJ，會當輸入規篇文章，系統經過變調處理，算出變調以後 ê 聲調，kā 系統 chhoân 好 ê 每一個音節 ê 聲音鬥起來做一個聲音檔，輸出予使用者。

輸入是規篇文章，chia ê 資料先利用標點符號做分界，kā 切做幾 lō 個單位，針對每一個單位，先查辭典，揣出伊可能是啥物詞性（一個語詞可能有 m̄-nā 一種 ê 詞性），了後根據訂出來 ê 變調規則，算出每一個音節變調了後 ê 聲調。所以除了變調處理 ê 演算法，koh 需要有包含語詞詞性 ê 台語辭典，kah 錄音錄好 ê 聲音檔案。

阮利用台華線頂辭典 ê 資料。辭典有 62,000 外筆詞條，每一個詞條有漢羅、POJ、華文解說，m̄-koh 並無詞性。（楊允言 2003）

因為揣無有標記詞性 ê 台語語詞電子資源，所以阮利用中央研究院詞庫小組整理 ê 中文八萬目詞¹⁴ ê 資料，透過台華辭典華文解說 ê 欄位，對去中文八萬目詞，揣著這個詞所有可能 ê 詞性。因為一個台語詞，有一個抑是以上 ê 華文解說，每一個華文解說攏會連去中文 8 萬目詞去揣伊 ê 詞性。

¹⁴ 計算語言學會負責管理，若有會員資格會使申請使用。
http://www.aclclp.org.tw/use_ced_c.php。

表 4-2 觀察文本

文本	年	作者	文類	編號
“Sin-bûn ê chap-liók” (新聞 ê 雜錄)	1913	編輯室	報導	1,2
“Cháp-hāng kóan-kiàn” (十項管見)	1925	蔡培火	議論文	3,4
“Chháu-tui téng ê bîn-bāng -- Jī-tông chong-kàu kò-sū” (草堆頂 ê 眠夢-兒童宗教故事)	1955	黃懷恩	小說	5,6
“Tang-pō thōan-tō kiàn-bûn kì” (東部傳道見聞記)	1961	陳降祥	報導	7,8

為著 beh 評估系統 ê 效果，變調 ê 正確率有侏懸，全款對台語文數位典藏資料庫揀四篇文本，每篇揀兩至三個段落，攞總揀九個段落，檢查實際輸出 ê 聲音有幾個音節聲調讀 m 著。下面是試驗文本：

表 4-3 試驗文本

文本	年	作者	文類	編號
“Pèh-ōe-jī ê lī-ek” (白話字 ê 利益)	1885	葉牧師	議論文	1,2
“Kau-chiàn ê Siau-sit” (交戰 ê 消息)	1905	編輯室	報導	3,4
“Thiàⁿ lí iáⁿ kè thong sè-kan” (疼你贏過通世間)	1955	賴仁聲	小說	5,6
“Ài lí kap ài i pīⁿ-á chōe” (愛妳 kap 愛伊平仔濟)	1997	盧誕春	散文	7,8,9

試驗文本計共有 962 個音節 (656 個語詞)，時間範圍全款包含日本時代 kah 戰後。

三、詞性集

台語到 taⁿ 為止，ná 像並無一套逐家普遍接受 ê 詞性集，所以阮暫時採用華語 ê 標準，中央研究院詞庫小組所訂出來 ê 詞性集，是為著華文 ê 資訊處理來訂 ê，khah 符合咱 ê 需要，資訊處理用 ê 詞性集是已經經過簡化 ê 詞性集，有 46 个詞性。(中文詞知識庫小組 2007)

表 4-4 中文詞知識庫小組簡化詞類集

形容詞	A 非調形容詞
連接詞	Caa 對等連接詞(如：和、跟)、Cab 連接詞(如：等等)、Cba 連接詞(如：的話)、Cbb 關聯連接詞
副詞	Da 數量副詞、Dfa 動詞前程度副詞、Dfb 動詞後程度副詞、Di 時態標記、Dk 句副詞、D 副詞
名詞	Na 普通名詞、Nb 專有名稱、Nc 地方詞、Ncd 位置詞、Nd 時間詞、Neu 數詞定詞、Nes 特指定詞、Nep 指代定詞、Neqa 數量定詞、Neqb 後置數量定詞、Nf 量詞、Ng 後置詞、Nh 代名詞、Nv 名物化動詞
感嘆詞	I 感嘆詞
介詞	P 介詞
語助詞	T 語助詞
動詞	VA 動作不及物動詞、VAC 動作使動動詞、VB 動作類及物動詞、VC 動作及物動詞、VCL 動作接地方賓語動詞、VD 雙賓動詞、VE 動作句賓動詞、VF 動作謂賓動詞、VG 分類動詞、VH 狀態不及物動詞、VHC 狀態使動動詞、VI 狀態類及物動詞、VJ 狀態及物動詞、VK 狀態句賓動詞、VL 狀態謂賓動詞、V_2 有、D 的、之、得、地、SHI 是

表 4-5 台語變調處理系統所採用 ê 詞性集

A	形容詞	I	感嘆詞	R	人稱代名詞
C	連接詞	M	特別標記	S	時間詞
D	副詞	N	名詞	T	語助詞
G	後置詞	P	介詞	V	動詞

若拄著未知詞(辭典查無 ê)，假使是疊詞(“XX”抑是“XXX”，X 指一个音節) ê 形式，阮就 kâ 標記做 **A** (形容詞)，其它 ê 標記做 **N** (名詞)。

四、變調標記

因為變調有幾 lō 種類型，系統會先根據演算法，tī 輸入 ê 文本做變調標記，繼落來才根據 chia ê 變調標記進一步轉做變調了後 ê 聲調。下面是阮所訂 ê 變調標記，這是內部表示 ê 符號。

表 4-6 變調標記

符號	說明	符號	說明
(無符號)	規則變調	\$	再變調
#	本調	&	á 前變調
@	隨前變調	~	三連音變調
%	輕聲	^	升調

第三節 規則式台語變調演算法

規則式台語變調演算法是這個即時系統上核心 ê 部分。下面是這個演算法 ê 內容：

表 4-7 台語變調處理演算法

規則	說明
1 所有音節攏無做標記（假設是規則變調）	
2 標點符號頭前彼音節攏標記#（本調）	
3 “è”頭前彼音節攏標記#（本調）	è 是特別標記
4 A-A 詞對(pair):第一個形容詞 ê 最後一音節標記#（本調）	詞性層次，有含糊性 (ambiguity)
5 N-V、N-A、N-P、N-R、N-D 詞對 5.1 N-V 詞對：名詞最後一音節標記#（本調） 5.2 N-A 詞對：名詞最後一音節標記#（本調）N-P 詞對：名詞最後一音節標記#（本調） 5.3 N-R 詞對：名詞最後一音節標記#（本調）N-D 詞對：名詞最後一音節標記#（本調）	詞性層次，有含糊性
6 C：連接詞頭前彼個語詞 ê 最後一音節標記#（本調）	詞性層次
7 G：後置詞 kah 伊頭前彼個語詞 ê 最後一音節標記#（本調）	詞性層次，無含糊性
8 S：時間詞 ê 最後一音節標記#（本調）	
9 POS R 9.1 i / in : kā 標記提掉（規則變調） 9.2 góa / lí / gún / góan / lán / lín : 若 m 是 tī 句尾，就 kā 標記提掉（規則變調）	詞性/語詞層次
10 句尾 ê “kóng”：假使“kóng” 頭前 ê 語詞是代名詞，而且標點是 “,” 抑是 “:”，就 kā 標記提掉（規則變調）	語詞層次，對觀察文本歸納 ê
11 á 頭前彼音節標記&（á 前變調）	音節層次

表 4-7 台語變調處理演算法

16.5 chiū / tō : 頭前 ê 語詞若是 A (形容詞), 就 kā A 最後一音節標記# (本調) 16.6 sî-kàu 時到: 兩個音節攏標記# (本調)	語詞層次, 對觀察文本歸納 ê 語詞層次, 對觀察文本歸納
17 T (語助詞): 語詞詞性是 T 而且 tī 句尾, kā 這個語詞頭前彼音節標記# (本調)	詞性層次
18 其它變調: 18.1 “teh”: 語詞“teh”kah“tī-teh” ê “teh”標記做^ (升調)	語詞層次, 阮 ê 觀察
19 輕聲變調: 19.1 chhut-lâi 出來、chhut-khì 出去、lòh-lâi 落來、lòh-khì 落去、kòe-lâi/kè-lâi 過來、kòe-khì/kè-khì 過去: 頭前彼個語詞詞性若是 V, 就 kā 頭前彼個語詞標記# (本調), 這個語詞每一音節攏標記% (輕聲) 19.2 sian-siⁿ/sin-seⁿ/sian-seⁿ 先生: 頭前彼音節 ê 第一字母若是大寫, 頭前彼音節就標記# (本調), 這個語詞每一音節攏標記% (輕聲) 19.3 bô 無: bô 頭前彼個語詞若 m̄ 是 á / á-sī / iah / iah-sī / ah / ah-sī, 就 kā 頭前彼音節標記# (本調), bô 標記% (輕聲) 19.4 bē/bōe 袂/未若 tī 句尾: 19.4.1 這句話有 “ê/ōe”、“ê-hiáu/ōe-hiáu”, 就 kā“bē/bōe”標記% (輕聲) 19.4.2 頭前 ê 語詞若是 á / á-sī / iah / iah-sī / ah / ah-sī, 就 kā“bē/bōe”標記# (本調) 19.4.3 其它情形 mài 處理(有可能是另外一個意思「賣」)	趨向動詞 語詞層次 語詞層次 語詞層次, 對觀察文本歸納 ê 語詞層次, 對觀察文本歸納 ê
20 R tī 句尾 góa / lí / i / gún / góan / lán / lín / in : 假使頭前彼個語詞詞性是 V, 就 kā 彼語詞最後一音節標記# (本調), 這個人稱代名詞標記@ (隨前變調)	語詞層次

阮根據頂面 ê 變調處理演算法來實作變調系統。Chia ê 變調規則包括音節、語詞、詞性 kah 句型四個層次，有 ê 語詞會受著一條以上規則 ê 影響（可比人稱代名詞會受第 1 條、第 2 條、第 3 條、第 9 條、第 20 條影響），所以規則 ê 順序有經過考慮，後壁 ê 規則會當 kā 頭前 ê 規則 khàm 過去（overwrite）。

第四節 實驗結果 kah 正確率

一、實驗結果

下面是觀察文本“Cháp-hāng kóan-kiàn”（《十項管見》，蔡培火，1925）ê 一部分（觀察文本編號 4），正片附轉寫 ê 漢羅文本。

Chhin-chhiūⁿ án-ni lâi kóng , chāi lán
Tâi-oân kîn-kîn chit-tiap-á-kú ê
kang-hu , ài soaⁿ chiū ũ soaⁿ , ài hái
chiū ũ hái , beh joah chiū ũ
joah , kôaⁿ chiū ũ kôaⁿ . Só-í thang
kóng Tâi-oân sī chit-ê sío Tang-iūⁿ .
Lán Tâi-oân ũ chit-khoán thian-jiân ê
hó-kéng , hó khi-hâu , chiong-lâi nā-sī
ēng-sim ke lāng ê kang-hu tōa-tōa lâi
chéng-tùn , tek-khak ē chiâⁿ-chò
Tang-iūⁿ ê tōa kong-hôh , hō
Tang-iūⁿ ê lāng chip-óa lâi hióng-hok
an-lók .

親像按呢來講，在咱台灣
近近一 tiap 仔久 ê 工夫，
愛山就有山、愛海就有
海；欲熱就有熱、寒就有
寒。所以通講台灣是一個
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，予東洋 ê
人集倚來享福安樂。

經過詞性標記 kah 變調規則處理了後，輸出是：

二、正確率

若針對攏總 614 音節 ê 觀察文本做處理，結果有 17 个錯誤，正確率是 97.2%；試驗文本有 962 个音節，結果有 106 个錯誤，正確率是 89.0%。實作出來 ê 系統，輸入 100 音節，大概有 11 个錯誤。

表 4－8 變調處理 ê 錯誤數 kah 正確率

觀察文本					試驗文本				
編號	語詞數	音節數	錯誤 音節	正確 率(%)	編號	語詞數	音節數	錯誤 音節	正確 率(%)
1	38	51	4	92.2	1	130	184	16	91.3
2	85	110	4	96.4	2	56	85	12	85.9
3	97	144	6	95.8	3	53	84	13	84.5
4	72	103	2	98.1	4	96	143	16	88.8
5	44	70	0	100.0	5	66	97	10	89.7
6	33	52	0	100.0	6	63	86	9	89.5
7	27	30	1	96.7	7	32	43	3	93.0
8	42	54	0	100.0	8	38	58	2	96.6
					9	122	182	25	86.3
計共	438	614	17	97.2	計共	656	962	106	89.0

第五節 相關討論 kah 未來改進方向

一、相關討論

若是針對錯誤 ê 部分做分析，咱會當看著一寡問題。因為 m̄ 是用句法剖析器做，有一寡變調問題用詞性 ê 排列順序是無才調判斷變調 ê，親像下面 ê 例：

(a) “phah²-piàⁿ²(V) chò²(V) khang⁷-khòe(khè)³(N)”拍拚做工課

(b) “kiáh³-bák⁸(V) khòaⁿ²(V) hnh⁵(N)”擲目看園

這兩個例攏有 V-V 詞對，第一個動詞 ê 最後一音節，(a) (piàⁿ) 愛讀變調，m̄-koh (b) (bák) 應該愛讀本調，雖然咱看會出來，“kiáh-bák”雖然是動詞，m̄-koh 伊 ê 構詞方式是動詞“kiáh”加名詞“bák”，名詞 ê 後壁若有動詞，這個名詞會是一個變調分界組 ê 尾溜。當然，除非咱 ê 辭典資料有標註構詞 ê 形式，若無這個問題 oh 得改進。

Koh 看下面 ê 例：

(a) “siong⁷-giáp⁴(N) kóng¹-kò³(N)”商業廣告

(b) “thâng^{7/3}-thōa⁷(N) chiáu¹-chiah⁴(N)”蟲豸鳥隻

這兩個例攏有 N-N 詞對，第一個名詞 ê 最後一音節，(a) (giáp) 愛讀變調，m̄-koh (b) (thōa) 愛讀本調，(b)是兩個名詞並列，設使

台語文學文本，koh有聲音，雖然變調結果m̄是全然正確。另外，阮有做一个「台語羅馬字發音試驗」系統¹⁷，提供使用者輸入台語文句，了後輸出聲音，m̄-koh因為需要輸入POJ，使用ê人可能無濟。後來阮進一步處理漢羅kah全羅文字互轉，就有提供輸入台語漢羅、輸出聲音ê系統，請參考第六章。

經過實作台語變調系統，予阮深深感覺，beh 要求電腦學會曉台語變調智識，並 m̄是 hiah-nī 輕鬆 ê 代誌。M̄-koh 對一个三歲囡仔，假使伊自細漢 tī 台語 ê 環境，講出正確 ê 台語聲調並無困難，這點實在不止仔奧妙。

Ti chia 提出幾點可能會當 koh 進一步改進 ê 方向：

- (a) 初胚 ê 台語變調系統已經實作出來，希望有語言學專家協助訂出 koh-khah 幼路、正確 ê 變調規則；
- (b) 改進斷詞 kah 詞組 ê 判斷，特別是對構詞、定量詞 kah 專有名詞 ê 判斷 kah 處理；
- (c) 改進詞性標記，解決含糊性 ê 問題；
- (d) 增補、改進台語變調規則；
- (e) 試用其它 ê 方法來處理台語變調問題。

¹⁷ 台語羅馬字發音試驗系統網址：<http://ip194097.ntcu.edu.tw/tts/tts.asp>。

本章參考書目

Chen, Matthew Y. 2000. *Phonological Phrase as a Sandhi Domain. In Tone Sandhi : Patterns Across Chinese Dialects*. Cambridge Univ. Press

中文詞知識庫小組. 2007. 《中研院平衡語料庫詞類標記集》 <http://godel.iis.sinica.edu.tw/CKIP/paper/poslist.pdf> (2013/12/20 檢索)

洪惟仁. 1985. 《台灣河佬語聲調研究》. 台北:自立

楊允言. 2003. 〈台文華文線上辭典建置技術及使用情形探討〉.《第三屆全球華文網路教育國際學術研討會論文集》 p132-141

楊允言. 2007. 〈台語白話文學 ê 全新表現——台語文數位典藏資料庫計畫簡介〉. 《台灣文學館通訊》15 期 p20-22

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

盧廣誠. 1999. 《臺灣閩南語詞彙研究》. 台北:南天

駱嘉鵬. 2008. 〈閩南語連讀變調域的劃分〉. 《第十屆閩方言國際研討會》

時間詞、地方詞、...等等。¹⁸一層當然相對khah簡單，m̄-koh咱大概會當知影，這馬若kan-na做一層，未來早慢愛koh重做。所以，至少兩層，是khah實際ê想法。

學界並無特別針對台語 ê 詞性訂一套標準，所以，阮暫時採用中央研究院詞庫小組所訂 ê 中文詞性集，攏總有 46 個詞性。(中文詞知識庫小組 2007) 按呢做主要 ê 考慮是，有詞性集了後，若是 beh 做資訊處理，咱需要利用所訂 ê 詞性集對文本做詞性標記，m̄-koh 暫時咱並無充足 ê 人力資源對文本做這項工課，包括對電腦做出來 ê 詞性標記結果做大量 ê 人工校對。若是用華文 ê 詞性集，有一寡所在會使直接借用華文 ê 成果，實務上，按呢會 khah 好進行，而且中央研究院詞庫小組 ê 詞性集主要是針對資訊處理所設計出來 ê。

咱並無充足 ê 人力，針對台語文文本做人工 ê 詞性標記，因為需要有 chia ê 人工處理好 ê 訓練語料，提供電腦利用統計方法來學習，進一步來做自動詞性標記。

π 這個情形下，阮提出利用現有 ê 華文資源，透過台華線頂辭典做對應，用統計方法來進行台語詞性自動標記 ê 工課。

¹⁸ 陳輝龍(1934) kā 詞性分做名詞、代名詞、數詞 kah 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 11 類；李獻璋(1950) kā 詞性分做名詞、代名詞、動詞、助動詞、形容詞、副詞、介詞、連接詞、助詞 9 類，有分到第二層，親像形容詞 koh 分做性狀形容詞、數量形容詞、指示形容詞 kah 疑問形容詞四類。

第二節 實驗方法

一、語料來源

阮利用國立台灣文學館台語文數位典藏資料庫（第二階段）ê 部分文本來試詞性標記。這份語料有 258 萬外 ê 音節，POJ kah 漢羅合用兩種文字形式，一段落一段落對齊。（楊允言 2007）

二、語詞對齊

阮開發語詞對齊程式，kā 原來是一段落一段落對齊 ê 台語雙文字文本改做一個語詞一個語詞對齊 ê 形式。因為 POJ ê 書寫形式 kah 英語類似，每一個語詞有空白分開，所以利用 POJ ê 語詞，算出 POJ 語詞 ê 音節數，對去漢羅文本掠全款音節數 ê 語詞，按呢就無需要另外處理斷詞。

另外，阮將這個語詞 ê POJ kah 漢羅提來 kah 台華線頂辭典中 ê 詞條比對，假使這個語詞無同時出現 tī 全一詞條 ê POJ、漢羅欄位，系統會標示出來做提醒，有可能是未知詞，mā 有可能是漢字用字 kah 辭典無一致，抑是其它親像拍字錯誤等等 ê 原因。

三、揣出對應 ê 華文候選詞

繼落來，阮繼續 koh 利用台華線頂辭典，kā POJ、漢羅配對 ê 語詞，揣出伊對應 ê 華文解說。這 m̄ 是一對一 ê 對應，一個台語

假設一語句有 m 個語詞，第一個語詞 w_1 是對 $w_{11}, w_{12}, \dots, w_{1n_1}$ 候選詞內底揀出來 ê 華文詞，第二個語詞 w_2 是對 $w_{21}, w_{22}, \dots, w_{2n_2}$ 候選詞中揀出來 ê 華文詞，第 m 個語詞 w_m 是對 $w_{m1}, w_{m2}, \dots, w_{mn_m}$ 候選詞中揀出來 ê 華文詞。咱 beh 對其中揀選出可能 ê 串列 $\hat{S} = w_1 w_2 \cdots w_m$ ，予 $\prod_{i=1}^m P(w_i | w_{i-1})$ 上大，因為 HMM 模型假設語詞 w_i kan-na 受頭前彼個語詞 w_{i-1} 影響，所以

$$P(\hat{S} = w_1 w_2 \cdots w_m) \cong P(w_1) \times \prod_{i=2}^m P(w_i | w_{i-1})$$

也就是予 $\log P(w_1) + \sum_{i=2}^m \log P(w_i | w_{i-1})$ 上大。

M-koh 因為這是 kā 台文（來源文）轉做華文（目標文），雙連語詞 $w_{i-1} w_i$ 有可能無 tī 訓練語料內底出現。阮用 Laplace smoothing 方法來解決 $P(w_i | w_{i-1}) = 0$ ê 問題。(Field 1988)

實務上，阮用 Viterbi 演算法來降低複雜度，設使一個語句 S 有 m 個語詞，每一個語詞有 n 個候選語詞，計算 ê 時間複雜度會是 $O(n^m)$ ，利用 Viterbi 演算法會當 kā 降到 $O(n^2 \times m)$ 。(Manning, Schütze 1999)

需要說明 ê 是，台語俚華語 ê 句法並無仝款，這個串列 $\hat{S}$ ，可能 m̄ 是合法 ê 華文語句。

下面是一個實例，選出來 ê 華文對應詞用粗體加框表示：

台 文	對/ Tùi	古早/ kó·chá	以來/ í-lâi	琴/ khîm	有/ ū	濟濟/ chê-chê	款/ khóan
華 文 對 應 詞	從 _{w₁₁} 對 _{w₁₂} 對子 _{w₁₃} 對於 _{w₁₄}	以前 _{w₂₁} 古代 _{w₂₂} 古時候 _{w₂₃} 從前 _{w₂₄}	以來 _{w₃₁}	琴 _{w₄₁}	有 _{w₅₁}	濟濟 _{w₆₁} 很多 _{w₆₂}	樣子 _{w₇₁} 樣式 _{w₇₂} 整理 _{w₇₃}
	$w_1 = w_{11}$	$w_2 = w_{21}$	$w_3 = w_{31}$	$w_4 = w_{41}$	$w_5 = w_{51}$	$w_6 = w_{62}$	$w_7 = w_{71}$

這是統計模型，若是用 khah 淺白 ê 方式來講，就是一個語句包含幾 lō 个語詞，每一個語詞愛 ùi 幾 lō 个候選詞內底 kā 揀出來，咱用統計模型來 kā 每一個語詞揀出來。

五、根據華文詞串揀出詞性

這部分 mā 是中央研究院詞庫小組協助完成 ê，阮採用 Maximal Entropy Markov Model (MEMM) 來決定語詞 ê 詞性。

MEMM 包括一組包含語詞 kah 詞性 ê 歷程集合 H ，kah 詞性集合 T ， $p(h,t) = \pi \mu \prod_{j=1}^k \alpha_j^{f_j(h,t)}$ ，其中， $h \in H, t \in T$ ， π 是常數， $\{\mu, \alpha_1, \dots, \alpha_k\}$ 是比 0 khah 大 ê 參數， $\{f_1, \dots, f_k\}$ 是特徵， $f_j(h,t) \in \{0,1\}$ ，參數 α_j 對應特徵 f_j 。

針對目標語詞 w_i ê 詞性 t_i ，揀出 10 个特徵，包括：

- (a) 語詞：有 $w_i, w_{i-1}, w_{i-2}, w_{i-1}, w_{i+1}, w_{i+1}, w_{i+2}$ 五个特徵；
- (b) 詞性：有 $t_{i-1}, t_{i-2}, t_{i-1}$ 兩個特徵；
- (c) 構詞： m_1, m_2, m_n 三个特徵

第三步，利用 HMM 揀出上適當 ê 華文對應詞：

“台灣/Tâi-ôan[台灣] 第一/tē-it[第一] 懸/kôan[高]
ê/ê[的] 玉山/Giòk-san[玉山] ê/ê[的] 附近/hū-kūn[附
近] 較/khah[較] 低/kē[低] ê/ê[的] 所在/só-chāi[地
方] ...”

第四步，利用 MEMM 決定詞性：

“台灣/Tâi-ôan(Nc) 第一/tē-it(Neu) 懸/kôan(VH) ê/ê(DE)
玉 山/ Giòk-san(Nc) ê/ê(DE) 附 近 /hū-kūn(Nc) 較
/khah(Dfa) 低/kē(VH) ê/ê(DE) 所在/só-chāi(Na)...”

第三節 實驗結果

阮利用頂面所講ê方法來做台語文詞性標記，m-koh因為無標準答案thang好檢查系統ê正確率，阮只好抽部分資料，用人工來檢查結果，人工檢查ê時，主要參考中央研究院詞庫小組ê中文斷詞系統。¹⁹

阮揀七篇文章，時間包括清國、日本 kah 戰後三个無仝ê時代，文類包括散文（七篇）、小說（五篇）kah 劇本（兩篇），每一篇文章揀第一段，若第一段文字傷少，就揀第二段。下面是試驗語料列表：

¹⁹ <http://ckipsvr.iis.sinica.edu.tw/>。

表 5 - 1 試驗語料列表

編號	年	文類	作者	篇名	音節數
1	1885	散文	葉牧師	Pèh-ōe-jī ê lī-ek (白話字 ê 利益)	162
2	1893	散文	甘為霖	Chhiⁿ-mî òh (青盲學)	66
3	1919	散文	H S K	Phín-hēng ê ûi-thôan (品行 ê 遺傳)	180
4	1935	散文	汪宗程	Chín-chai kì (震災記)	122
5	1954	散文	胡文池	Tōa-soaⁿ chhiùⁿ-koa (大山唱歌)	74
6	1990	散文	陳義仁	Lāu-lâng ê kè-tàt (老人 ê 價值)	75
7	2000	散文	陳明仁	Sûn-chêng Ông Pó-chhoan(純情王寶釧)	112
8	1890	小說	(m̄知)	An-lòk-ke (安樂家)	101
9	1924	小說	賴仁聲	Ān-niá ê Bák-sái (俺娘 ê 目屎)	133
10	1955	小說	黃懷恩	Chháu-tui téng ê bîn-bāng(草堆頂 ê 夢)	116
11	1990	小說	楊允言譯	Hái-phīⁿ Sin-niû (海鼻新娘)	94
12	2006	小說	劉承賢	Chiàh-chōe (食罪)	92
13	1924	劇本	林茂生	Hì-chhut: Lō̍-tek kái kàu (戲齣：路得改教)	77
14	1950	劇本	陳清忠譯	Venice ê Seng-lí-lâng (Venice ê 生理人)	92
說明：編號 11 原作者是宋澤萊，編號 14 原作者是 Shakespeare					

下面是試驗語料 ê 詞性標記結果，包括每一篇 ê 音節數、語詞數、有幾個語詞選 m̄著，有幾個語詞詞性標記錯誤，正確率是佻濟：

另外，有時一个台語詞會對應到兩個華文詞，親像台語 ê “piah-téng[壁頂]”對應華文 ê “牆壁”、“上”兩個詞；mā 有兩個台語詞對應到一个華文詞 ê，親像台語 ê “中國”、“字”兩個詞對應到華文 ê “中國字”。第一種情形會 kā 當做未知詞來處理；第二種情形，因為拆開 ê 兩部分 mā 攏是詞，所以就 kā 當做兩個詞來處理。這個部分，假使詞性標記是正確 ê，阮就 kā 當做是正確。

下面是實際 ê 詞性標記結果舉例，這是編號 7 ê 一部分。其中，第一欄是漢羅，第二欄是 POJ，第三欄是華文候選詞，頭前加‘@’符號 ê，表示第一欄 kah 第二欄 ê 台語文並無出現 tī 辭典全一筆詞條內底，所以直接提漢羅當做華文候選詞，第四欄是揀出來 ê 華文詞，最後一欄是系統決定 ê 詞性。阮 kā 揀出來 ê 華文詞錯誤抑是詞性標記錯誤 ê 加底線，頭前 koh 加‘**’，錯誤 ê 詞性後壁，加註正確 ê 詞性標記（粗體，用小括號表示）。

表 5-3 台語詞性標記結果舉例

漢羅	POJ	華文候選詞	揀ê華文詞	詞性
我	góa	我	我	Nh
將	chiong	將	將	D
草帽仔	chháu-bō-á	@草帽仔	草帽仔	Na
掛	kòa	帶;掛;戴	帶	VC
tī	tī	在	在	P
壁頂	piah-téng	牆壁上	牆壁上	Nc
,	,	,	comma	

表 5-3 台語詞性標記結果舉例

漢羅	POJ	華文候選詞	揀ê華文詞	詞性
行李	hêng-lí	行李	行李	Na
khêng	khêng	收拾;盤點	收拾	VC
khêng	khêng	收拾;盤點	收拾	VC
leh	leh	咧	咧	T
,	,	,	comma	
坐	chē	坐	坐	VA
tòa	tòa	住	** <u>住</u>	** <u>VCL(P)</u>
小店	sió-tiàm	@小店	小店	Na
ê	ê	的	的	DE
tha-thá-mì	tha-tha-mì	塌塌米	塌塌米	Na
頂 kôan	téng-kôan	上面	上面	Ncd
,	,	,	comma	
看	khòa ⁿ	看	看	VC
窗外	thang-gōa	@窗外	窗外	Nc
ê	ê	的	的	DE
光景	kong-kéng	風光;氣象;光景;風景;景氣	** <u>景氣</u>	Na
,	,	,	comma	
看	khòa ⁿ	看	看	VC
起起	khí-khí	@起起	起起	<u>Nb(VA)</u>
落落	lòh-lòh	@落落	落落	VA
ê	ê	的	的	DE
海湧	hái-éng	海浪;海潮	海浪	Na
,	,	,	comma	
因為	in-ūi	由於;因為	因為	Cbb
等待	tán-thāi	留待;等待	等待	VK
朋友	pêng-iú	友人;朋友	朋友	Na
,	,	,	comma	

一、揀 m 著華文對應詞造成 ê 錯誤

表 5-4 系統選 m 著 ê 華文對應詞 kah 標記

台語詞	系統揀 ê 華文 詞 kah 標記	khah 適當 ê 華 文詞 kah 標記	說明
押/ah	強制(D)	押(VC)	
bat/bat	知道(VK)	曾(D)	
無/bô	不(D)	沒有(VJ)	2 擺
chham/chham	和(P)	摻(VC)	
進前/chìn-chêng	之前(Ng)	向前(P Nes)	
這號/chit-hō	這樣(VH)	這種(Nep Nf)	2 擺
轉/chōan	上(Ncd)	轉(Vac)	2 擺
外/gōa	外(Ng)	開外(Neqa)	2 擺
夭壽/iáu-siū	非常(Dfa)	早夭(VH)	
加/ke	上(Ncd)	多(Dfa)	
價值/kè-tát	值得(VH)	價值(Na)	
腳/kha	個(DE)	下(Ncd)	
黃 hóa ⁿ /âng-hóa ⁿ	罕(D)	淺黃(A)	
倚/óa	依(P)	靠(VJ)	
活/óah	生活(Na)	活(VH)	
破相/phòa-siù ⁿ	破(VHC)	殘廢(Na)	
細漢/sè-hàn	小時候(Nd)	年幼(VH)	
相借問 /sio-chioh-m̄ng	招呼(VC)	打招呼(VB)	
搭/tah	地方(Na)	搭(VC)	
tióh/tiόh	就(P)	著(VCL)	
著/tiόh	就(P)	得(D)	

阮針對揀 m 著華文對應詞抑是詞性標記錯誤 ê 所在，斟酌去

kā 看，發現其中有 25 个語詞是因為揀 m̄ 著華文對應詞致到詞性標記錯誤。表 5 - 4 列出揀 m̄ 著 ê 華文對應詞 kah 標記錯誤 ê 詞性。

二、台華辭典欠缺適當 ê 華文對應詞造成 ê 錯誤

阮發現有 14 个語詞，因為台華線頂辭典內底無收 khah 適當 ê 華文對應詞，致使詞性標記錯誤，這表示台華辭典有需要 koh 增補一寡詞條 ê 華文解說。下面列出 chia ê 語詞：

表 5 - 5 欠缺適當 ê 華文對應詞造成 ê 錯誤

台語詞	系統揀 ê 華文詞 kah 標記	khah 適當 ê 華文詞 kah 標記	說明
chak/chak	促(VF)	擠(VC)	
chūn/chūn	絞(VC)	陣(Nf)	2 擺
kah/kah	和(Caa)	得(DE)	3 擺
leh/leh	咧(T)	在(P)	3 擺
煞/soah	結束(VHC)	卻(D)	
teh/teh	在(P)	著(Di)	
頂/téng	頂(VC)	上(Nes)	
tiā ⁿ -tiā ⁿ / tiā ⁿ -tiā ⁿ	常常(D)	而已(T)	
轉 / tŕg	調解(VC)	轉(VAC)	

三、未知詞造成 ê 詞性標記錯誤

有 10 个錯誤是因為未知詞 ê 詞性標示錯誤，下面 ê 表 kā chia ê 語詞列出來：

另外有一个台語詞“對/tùi”，tī 大部分 ê 語境，伊 ê 華文對應詞是“從”。這個語詞 tī 試驗語料內底攏總出現十擺，m-koh 系統有八擺選“對”做華文對應詞，兩擺選“從”做華文對應詞。總是詞性標記攏全款是“P”（介詞），對詞性標記正確率並無影響。

Koh 有 30 个詞性標記錯誤，阮暫時無法度歸納出錯誤 ê 原因，下面 kā 其中幾個列出來：

表 5 - 7 其它詞性標記錯誤舉例

倒片 ê 語境	語詞 kah 詞性	正確	正片 ê 語境	編號
	lūn/論(Na)	VE	thák/讀(VC) pèh-ōe-jī/白話字(Na) khah-iâⁿ/ khah 贏(VJ) ...	1
chò/做(VC) thài-lâng/刳人(VA)	hōan/犯(VC)	Na	siū/受(P) sí-hêng/死刑(Na) ê(DE) 7(Neu) lāng/人(Na)	2
lāng/人(Na)	chit-ē/一下(Nd)	D	chiáh-lāu/食老(VH)	3
tùi/對(P)	khí-thâu/起頭(VH)	Nv	chiū/就(D) chin/真(Dfa) tâng/動(VAC)	4
Má-lī/(人名)(Nb) ê(DE) lāu-pē/老爸(Na)	sí/死(Dfb)	VH	ê(DE) sí/時(Na)	10
khòaⁿ/看(VC)	khí-khí/起起(Nb)	VA	lòh-lòh/落落(VA) ê(DE) hái-éng/海湧(Na)	11
âng-hóaⁿ/黃 hóaⁿ(VH) àm-tām/暗淡(VH) ê(DE) lō-teng/路燈(Na)	chhiō(D)	VC	lóng/攏(D) bē/袂(D) hng/遠(VH)	12

六、詞性標記錯誤總歸納

下面 kā 攞總 87 个詞性標記錯誤內底，分析出詞性標記錯誤原因 ê 部分列表，看每一款錯誤 ê 比例：

表 5-8 詞性標記錯誤原因

原因	幾擺	比例	說明
揀 m̄ 著華文對應詞	25	28.7%	
辭典欠缺適當 ê 華文對應詞	14	16.1%	
未知詞	10	11.5%	
人名	4	4.6%	
傳播錯誤	5	5.7%	包括一个未知詞
計共	57	65.5%	扣調重覆算 ê

第五節 相關討論 kah 未來改進方向

一、敢有可能 koh 改進？

上蓋理想 ê 情形是，假使頂面知影原因 ê 57 个錯誤攞有法度改進，按呢，台語詞性標記會當達到 97.1%（ $= (87-57)/1038 * 100\%$ ） ê 正確率。總是，這實在有困難。

台語語詞順序 kah 華語語詞順序畢竟有差別，所以選 m̄ 著華文對應詞致到詞性標記錯誤是比例上懸 ê。無適當 ê 華文對應詞 thang 選 ê 問題，是造成標記錯誤 ê 第二個主要原因，這會使透過增補台華線頂辭典 ê 詞條來得著解決。

按算修改文本 ê 前提下，假使對應 ê 華文詞是未知詞 ê 時，會使考慮翻頭 kā 台語這個語詞 ê 連字符拆開，才繼續試看覓。按呢做，無的確會使降低因為未知詞所造成 ê 詞性標記錯誤 ê 機會。

三、無仝文類抑是時代 ê 台語文本敢會有差別？

這個實驗所揀 ê 試驗文本，有三種文類，mā hān 過台灣歷史上 ê 三個時期，無仝文類抑是無仝時期所書寫 ê 台語文，敢會影響詞性標記 ê 結果？根據表 5 - 2，咱分別用文類 kah 時代來算，得著下面 ê 列表：

表 5 - 9 無仝文類 ê 詞性標記正確率

文類	語詞數	標記錯誤數	正確率(%)
散文	549	43	92.2
小說	372	36	90.3
劇本	117	8	93.2

表 5 - 10 無仝時代 ê 詞性標記正確率

時代	語詞數	標記錯誤數	正確率(%)
清國	232	18	92.2
日本	359	27	92.5
戰後	447	42	90.6

若以文類來分，小說類文本 ê 詞性標記正確率有 khah 低；若以時代來分，戰後文本 ê 詞性標記正確率有 khah 低，總是並無真明顯 ê 差異。當然，咱 ê 樣本數並無濟。

翻頭 koh 來看表 5 - 2，若用各別 ê 文本看，詞性標記正確率上懸 ê 是 98%（編號 5），上低 ê 是 83.6%（編號 12），這兩份文本攏是戰後 ê。看起來，無的確作者本身 ê 寫作風格才是影響詞性標記正確率上大 ê 因素。

本章參考書目

- Carroll, J., Long, D. 1989. *Theory of Finite Automata with an Introduction to Formal Languages*. NJ:Prentice Hall
- Field, D. A. 1988. Laplacian smoothing and Delaunay triangulations. *Communications in applied numerical methods* 4(6) p709-712
- Manning, C., Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge:MIT Press
- McCallum, A., Freitag, D., & Pereira, F. C. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. *International Conference on Machine Learning* p591-598
- Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing* Vol.1 p133-142
- Tsai, Y. F., Chen, K. J. 2004. Reliable and Cost-Effective PoS-Tagging. *International Journal of Computational Linguistics and Chinese Language Processing* 9(1) p83-96

第一節 現有相關系統介紹

台語信望愛網站提供漢羅文本 kah 羅馬字文本 ê 轉換系統，漢羅轉羅馬字系統 tī “<http://taigi.fhl.net/lohan/>”，羅馬字轉漢羅系統 tī “<http://taigi.fhl.net/hanlo/>”。基本上，背後是利用辭典做轉換。

台語信望愛網站是一個台語文綜合性網站，提供真濟台語文文本。對早期宗教性 ê 羅馬字文本，in 想 beh kā 轉寫做漢羅台語文，m̄-koh 愛用真濟時間，而且人工轉換可能會拄著一寡問題，包括人有時會拍 m̄ 著字，têng-tâⁿ 去，轉寫 ê 漢字 mā 可能無一致。利用辭典來轉換，會使加減避免頂面講著 ê 問題發生。總是辭典無可能 kā 所有 ê 語詞攏包含入來，koh 有人名、地名等等 ê 問題，所以系統提供使用者家已增加詞庫 ê 功能。

M̄-koh 這個方法並 m̄ 是全然無問題。一個羅馬字，可能對應幾 lō 個漢字，一個漢字 mā 可能對應幾 lō 個羅馬字，若是雙音節（iah 是以上）ê 語詞，相對 khah 袂有問題，m̄-koh mā m̄ 是攏無問題。台語平均一個語詞 ê 長度是 1.4 個音節，單音節語詞佔 ê 比例真懸。為著解決這個問題，除了單純用辭典來轉換，in koh 提供詞組 ê 轉換，來增加轉換 ê 正確率。（蔡哲民等 2006a; 蔡哲民等 2006b）

第二節 互訊息 kah 相關度

過去 10 外冬，阮收集將近一千萬音節 ê 台語文語料。漢字、

羅馬字轉換系統，除了用辭典，其實 koh thang 利用台語文語料做線索來改進轉換 ê 品質。

對仝一个漢字，beh 揣對應 ê 羅馬字，辭典 ê 做法是揀上有可能 ê 羅馬字（就是頻率上懸彼个），其實語料內底，這個語詞 ê 前後文提供真好 ê 線索。

可比講，用漢字文本 ê 內容是「有人...」（設使辭典並無「有人」這個詞），其中，「有」可能讀做“ū”抑是“iú”，人可能是“jîn”抑是“lâng”，辭典無法度判斷佢一个正確。若用統計 ê 方法，「有人」ê 發音有 4(=2*2)種可能：“ū jîn”、“ū lâng”、“iú jîn” kah “iú lâng”。

咱提台語文語料庫 ê 資料來檢視，其中，羅馬字文本有 346 萬外音節，這 4 个羅馬字 kah 組合了後 ê 結果，tī 語料庫內底出現 ê 次數請參考下表：

表 6-1 “有人”各種發音頻率表

音節	次數	可能組合	次數
ū	41,009	ū jîn	55
iú	2,580	ū lâng	1,821
jîn	6,232	iú jîn	1
lâng	51,525	iú lâng	6

對頂面 ê 表，咱大概隨會當認定，“ū lâng”就是標準答案。雖然答案正確，m̄-koh 假使用这个方法來判斷，頻率懸 ê 一定會 hōng 揀出來，頻率低 ê soah 全然無翻身 ê 機會。

rel 值攏是正數，值愈大，結合程度愈強。

根據頂面 ê 公式，所算出來 ê 4 種可能 ê 組合 ê *MI kah rel* 值。兩種算法 ê 排名結果並無相仝，*m̄-koh*，“*ū lâng*”攏是排名第一 ê 組合。

表 6 - 2 “有人” *MI kah rel* 值

可能 ê 組合	<i>MI</i>	排名	<i>rel</i>	排名
<i>ū jîn</i>	-0.2352	2	5.05	3
<i>ū lâng</i>	1.7655	1	2296.03	1
<i>iú jîn</i>	-2.0260	3	2.02	4
<i>iú lâng</i>	-2.4886	4	21.42	2

羅馬字轉漢羅，所用 ê 方法 *mā* 仝款，先查字典、辭典，若是揣無抑是有超過一个可能 ê 結果，就透過對語料庫揣著 ê 資料，計算 *MI kah rel*，決定 *beh* 揀佗一个。

第三節 轉換演算法

根據頂一節 ê 說明，下面列出台語羅馬字 *kah* 漢羅文本轉換系統 ê 演算法。執行演算法 ê 時需要查表，包括台華辭典、羅馬字 ê 頻率、羅馬字 *MI kah rel* 值、漢羅 ê 頻率、漢羅 *MI kah rel* 值。

台華辭典有 62,000 外筆詞條，每一个詞條有羅馬字 *kah* 漢羅 ê 寫法；(楊允言 2003) 羅馬字 ê *MI kah rel* 值，是利用 POJ 文本語料，攏總 3,462,367 个音節，*kā* 相倚 ê 兩個音節(無包括標點符號)

掠出來，攏總有 3,036,642 組雙連音節 (syllable bigram tokens) /297,426 組無全 ê 雙連音節 (syllable bigram types)，利用頂面 ê 公式計算，得著 29 萬外筆 *MI kah rel* 值 ê 資料；漢羅 ê *MI kah rel* 值，利用台語漢羅文本語料攏總 5,568,057 个音節，相倚 ê 兩個音節計共 4,819,219 組雙連音節/570,903 組無全 ê 雙連音節來計算 ê，得著 57 萬外筆 *MI kah rel* 值 ê 資料。

一、羅馬字轉漢羅演算法

下面列出羅馬字轉漢羅演算法：

Step 1：查 台華辭典，結果有三種情形

(case 1a) 揣無

(case 1b) kan-na 一个答案 (這就是答案)

(case 1c) 超過一个答案

Step 2：(case 1a) ê 處理

(2a) 拼字檢查，若是無合法 ê 音節 (可能是外文)，直接輸出

(2b) // Spelling 正確

這個詞 ê 音節數是

(case 2b1) 1 // 單音節

ùi 漢羅候選詞表揣出對應 ê 漢羅候選詞，候選詞數目 n

(case $n=0$) 直接將這個羅馬字 ê 詞當做輸出

(case $n=1$) 當做 (case 1b)來處理

(case $n \geq 2$) 當做 (case 1c)來處理

(case 2b2) 2 // 雙音節

拆做兩個單音節，分別揣出漢羅候選音節

假使第一音節有 m 个可能，第二音節有 n 个可能，按呢就有 $m*n$ 个組合，揀選雙連音節 bigram 分數上懸 ê (若 m pat 出現過，分數就是 0)

值分別對應去 0 到 10,000，假使對應了後 ê 分數分別是 *MI' kah rel'*， $bigram(AB) = (MI'(AB) + rel'(AB)) / 2$ ，值 mā tī 0 到 10,000 ê 範圍內。另外，假使無 bigram thang 用 ê 時，阮就採用 unigram， $unigram(A) = P(A) * 10,000$ 。

二、漢羅轉羅馬字演算法

下面是漢羅轉羅馬字演算法：

Step 1：斷詞

Step 2：

Case 1：台華辭典有

Case 1-1 kan-na 一个：就是結果

Case 1-2 單音節詞，對應羅馬字 m nā 一个：bigram / unigram

Case 2：辭典查無

Case 2-1 漢字: ùi 這個漢字 ê 羅馬字候選字做 bigram/unigram

Case 2-2 羅馬字：直接輸出

Bigram kah unigram ê 方法 kah 羅馬字轉漢羅全款，差別是所查 ê *MI kah rel* 表無全。斷詞請參考第三章第三節，採用長詞優先，倒頭上大比對法（Backward Maximal Matching）。(Goh et.al. 2005)

第四節 實驗結果

阮將台語文漢羅、羅馬字互轉 ê 系統實作出來，掛 tī 網站，這個線頂轉換系統網址 tī “<http://ip194093.ntcu.edu.tw/cl2hl/choan-lai-choan-khi.php>”。

阮提 1925 年蔡培火寫 ê 《十項管見》內底 ê 一段內容當做輸入，這篇文章用 POJ 書寫，有人將 chia ê 文本轉寫做漢羅。原文是：

“Gōa-kok-lâng o-ló lán ê Tâi-ôan siù-lê ; ū-ia̍n, láu-sit bô chhò-gō̍ . Lán Tâi-ôan ê soaⁿ kék hiông-chòng , kék ū piàn-hòa , koh-chài sī kék chhong-chhùi , kék chhéⁿ-bák , sī-siông ū gîn-sek ê hûn , chāi hiah ê chheⁿ-leng-leng ê soaⁿ-chiam ê tiong-kan teh kún teh giá.”

利用羅馬字做輸入，系統轉換 ê 結果是：

“外國人 o-ló 咱 ê 台灣秀麗；有影，老實無錯誤。咱台灣 ê 山極雄壯，極有變化，閣再是極蒼翠，極醒目，時常有銀色 ê 魂(雲)，在 hiah ê 青瓏瓏 ê 山尖 ê 中間 teh 筊(滾)teh giá。”

其中，畫底線 ê 字是系統轉寫錯誤 ê，後面括號內是正確 ê 結果。62 个音節攏總出現 2 个錯誤，正確率是 96.77%。

若用台語信望愛網站提供 ê 系統來轉換，有 10 个音節錯誤，分別是「秀麗」(轉做獸 lē)、「雄壯」(轉做 hiông 葬)、「蒼翠」(轉做 chhong 嘴)、「醒目」(轉做 chhéⁿ-目/木)、「銀色」(轉做 gîn 瑟)、「giá」(轉做鵝)。

專家轉寫 ê 漢羅是：

“外國人呵咾咱 ê 台灣秀麗；有影，老實無錯誤。咱台灣 ê 山極雄壯、極有變化，閣再是極蒼翠、極醒目，時常有銀色 ê 雲，在 hiah ê 青 leng-leng ê 山尖 ê 中間 teh 滾 teh giá。”

用這做輸入，系統轉換 ê 結果是：

文漢羅、羅馬字 ê 轉換工具，若會當將這兩項成果結合，使用者就會使輸入台語漢字，予系統發音。

這個線頂台語發音系統網址 tī “<http://ip194093.ntcu.edu.tw/cl2hl/taigi2siann.php>”，會使輸入漢字，mā 會使輸入羅馬字，羅馬字用 POJ 抑是台羅，用調符抑是數字調號攏會使。

本章參考書目

- Church, K. W., Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1) p22-29
- Goh, C. L., Asahara, M., & Matsumoto, Y. 2005. Chinese word segmentation by classification of characters. *International Journal of Computational Linguistics and Chinese Language Processing* 10(3) p381-396
- 楊允言. 2003. 〈台文華文線上辭典建置技術及使用情形探討〉. 《第三屆全球華文網路教育國際學術研討會論文集》 p132-141
- 劉開瑛. 2000. 《中文文本自動分詞與標注》. 中國北京:商務
- 蔡哲民, 李仁傑, 林俊育. 2006a. 〈一套以拼音規則為基礎 ê 台語羅馬字編碼轉換與網頁展示工具〉. 《TANet 2006 研討會論文集》

蔡哲民, 林俊育, 周家弘. 2006b. 〈以片語轉換為基礎 ê 漢羅轉換、華台翻譯系統〉. 《2006 第一屆台灣語文暨文化研討會論文集》

第七章 台語文線頂資源評介

本章 beh 介紹台語文 ê 線頂資源，對部分資源 mā 會提出阮 ê 淡薄仔想法。包括字型、字詞典、語料庫、電子冊等等，除了這，koh 會介紹一寡台語文處理 ê 家私，包括輸入法、斷詞、文字系統轉換。其中有 ê 系統，tī 頭前有提起過。

第一節 字型

一、白話字字型

因為 tī 台灣，使用 Windows 作業系統 ê 人比例 khah 懸，所以這部分主要以 Windows 作業系統來介紹。POJ ê 部分，Tī 2004 年進前，包括 DOS 作業系統時代 1990 年鄭良偉開發 ê TW301、Windows 3.1 作業系統時代 1994 年左右蘇芝萌開發 ê HOTSYS，Windows 2000 作業系統時代 2001 年左右劉杰岳開發 ê Taiwanese Serif 字型，基本 ê 做法攏是揀 POJ 用袂著 ê 符號 ê 位，kā 改做 POJ ê 符號，可比講 kā “ä”改做“ā”，所以若有裝 chia ê 字型，看著 ê POJ 是正常 ê，m̄-koh 假使無裝，有一寡符號看起來是亂碼。(鄭良偉 1990; 曾正義 1997; 劉杰岳，楊允言 2002)

經入去 Unicode 字元集內底，可比“迨”，若是用預設 ê 新細明體抑是標楷體無法度顯示，必須愛用親像「細明體_HKSCS-ExtB」等後壁有“ExtB” ê 字型才會當正常顯示，m̄-koh 目前 iah 無標楷體 ê ExtB 字型，所以假使台語漢字 beh 用楷體來顯示，會拄著一寡漢字是無法度用標楷體，愛變做細明體。為著解決這個問題，烏衣行（廖偉成）有製作「台灣楷體」kah「台灣宋體」字型，會當免費掠來使用，網址 tī “http://wesingkasu.blogspot.tw/2012/01/blog-post_1223.html”。

另外，中央研究院文獻處理實驗室為著 beh 處理中國古籍內底無 tī Unicode 字元集 ê 漢字 ê 顯示，開發漢字構形資料庫。教育部台灣閩南語字詞頻調查工作（本章第三節會介紹）teh 整理文本 ê 時，拄著無 tī Unicode 字元集 ê 漢字 ê 時，mā 有向文獻處理實驗室申請 kā chia ê 漢字加入，利用漢字構形資料庫來處理顯示問題。（莊德明，鄧賢瑛 2008）

第二節 字、辭典

一、台華線頂辭典

台華線頂辭典是楊允言，tī 2000 年 11 月上線，目前詞條有 62,000 外條，詞條 ê 來源，主要是鄭良偉提供，koh 有部分是台語界熱心 ê 朋友提供。包括 POJ（有無仝腔口）、漢羅、華文解說等欄位，網址 tī “<http://ip194097.ntcu.edu.tw/q/q.asp>”。

2006 年詞條增加台語發音，就是利用第四章所講 ê 台語文發音系統，2007 年利用程式比對，kā《台英辭典》(Embree 1984) 內底大約 10,000 條英文詞加入來，所以查詢結果有部分看會著英文解說。

這個線頂辭典，是目前唯一提供免輸入聲調 ê POJ 查詢，可比輸入 “hoe-chhia”，會當查著 “花車 hoe-chhia”、“火車 hóe-chhia”、“貨車 hòe-chhia”、“火鋤 hóe-chhiah”，第三章第三節有說明無聲調輸入 ê 做法。另外，這個系統 mā 有提供網頁插件 ê 介面，thang 好 khng tī 別个網頁內底，包括 khng tī Blog 邊 a ê 欄位。網頁插件 ê 說明 tī “<http://ip194097.ntcu.edu.tw/TG/CK/chhahkiann.asp>”。

本冊提起 ê 一寡實作 ê 系統，有袂少攏有利用這份辭典 ê 詞條。(楊允言 2003a)

二、台語線頂字典

台語線頂字典是楊允言開發，2003 年上線 ê 系統，資料來源有大部分是甘字典，攏總有 22,000 外條，包括 POJ kah 漢字欄位兩個欄位。kah 台華線頂辭典全款，2006 年增加台語發音。網址 tī “<http://ip194097.ntcu.edu.tw/TG/jitian/tgjt.asp>”。

這個線頂字典會當用漢字查 POJ，mā 會當用 POJ 查漢字，POJ ê 部分，有提供包括無聲調、入聲、韻母 ê 查詢，相關技術 tī 第三章第三節有說明。

常用詞辭典攏總有 18,000 外條詞條（包括附錄），有漢字、台羅、真人台語發音、華文解說，另外，大部分詞條攏有例句。

漢字 ê 部分，因為有一寡字無 tī Unicode 字元集內底，所以採用圖形來顯示（m̄-koh 其中有一寡已經 tī Unicode 字元集內底，這個系統猶原用圖形，親像“勢”）；羅馬字查詢 ê 部分，會當用數字聲調，mā 會使用帶調符 ê 符號做輸入，可比講，輸入“bi2-le7” kah “bí-lē”攏會得著全款 ê 結果。

系統有提供模糊查詢（預設）kah 精確查詢，可比“𠵿”是造字，用圖形表示，所以使用者若 beh 知影這字按怎寫，伊會使用羅馬字“in”查詢，m̄-koh 羅馬字有包含“in” ê，親像“kin”、“āinn”等等 ê 音節攏會查著，結果攏總查著 1,359 筆，分做 68 頁，若是改用精確查詢，kan-na 有 7 筆，真緊就揣會著。（駱嘉鵬，李良一 2010）

系統所謂 ê 「模糊查詢」，其實是子字串（substring）比對，阮認為這 kah 模糊查詢 ê 概念並無全款，提頂面 ê “in”做例，kah “in” khah 倚 ê 發音可能是“im”、“ing”（POJ 是“eng”）、“inn”（POJ 是“i”），所以若用羅馬字“in” ê 模糊查詢，顛倒是“im”、“ing”、“inn” khah 需要 kā 揣出來，“āinn” khah 無需要。

另外，辭典 ê 附錄有外來語，m̄-koh 聲調是用下標 ê 音階數字來表示，音節之間無連字符，親像“bat₃ te₅₅ lih₃”（電池）、“kha₃₃ tian₅₁”（窗簾）、“lai₅₁ tah₃”（拍火機），並 m̄是寫“bát-té-lih”、“kha-tián”、“lài-tah”，這部分感覺有 khah 奇怪。

五、甘字典 (廈門音新字典)

甘為霖 (William Campbell) tī 1913 年出版《廈門音新字典》，一般 kā 號做《甘字典》，重印真濟擺，2009 年，台灣教會公報社重拍字出版，號名叫《甘為霖台語字典》。線頂 ê 甘字典 kah 台日大辭典台語譯本全款，是林俊育整理，蔡哲民負責開發系統，mā kah 中央研究院閩客語典藏計畫合作，提供 ê 查詢功能 kah 台日大辭典台語譯本全款，精差內容無全，網址 tī “<http://taigi.fhl.net/dick/>”。

甘字典是第一本台台字典，就是台語詞條、台語解說，而且有分文白，文音 ê 羅馬字第一字母寫大寫，可比“大”有兩個詞條，一个是“Tài”（文讀），另外一个“tōa”（白讀）。

六、中研院語言所閩客語典藏計畫

閩客語典藏計畫是中研院語言所語言典藏計畫下面 ê 子計畫，第二期（2007~1012）ê 時有 kā 一寡辭典數位化，除了進前提起 ê 台日大辭典台語譯本 kah 甘字典以外，koh 有 1873 年杜嘉德（Carstairs Douglas）編 ê 《廈英大辭典》、1883 年麥嘉湖（John Macgowan）編 ê 《英廈辭典》kah 1923 年巴克禮（Thomas Barclay）編 ê 《增補廈英大辭典》。這三本辭典是 khng 做伙提供整合查詢，系統開發者是雷明漢。網址 tī http://minhakka.ling.sinica.edu.tw/bkg/chong-su-tian.php?gi_gian=hoa。

三、台語文語料庫

這是楊允言開發，tī 2003 年上線 ê 語料庫，語料分 POJ 文本 kah 漢羅兩部分，主要來源是台語文刊物 ê 文章，mā 有 khah 早期 POJ 文獻等等。POJ ê 文本有 349 萬外音節，漢羅文本有 581 萬外音節，有提供語詞檢索功能，會當查詢關鍵詞 ê 前後文。網址 tī “<http://ip194097.ntcu.edu.tw/TG/concordance/form.asp>”。根據網站所提供 ê 統計資料，上懸 ê 記錄是一工有 3 萬外擺 ê 查詢。

四、台語聖經語料庫

台語信望愛網站有提供無仝版本 ê 台語聖經，包括巴克禮聖經、紅皮聖經、全民台語聖經、...等等，會當用章節來查詢，而且無仝版本會使做伙比較。除了台語聖經，mā 有連結其它語文版本 ê 聖經。網址 tī “<http://taigi.fhl.net/list.html>”。

五、台語文數位典藏資料庫

呂興昌 tī 2001~2004 年執行國立台灣文學館委託研究 ê 「台灣白話字文學資料蒐集計畫」，針對 POJ 文獻，有將 224 萬外音節 ê 語料拍字整理；2004 年高成炎執行「台語文數位典藏資料庫」計畫，主要是做 POJ 轉聲音。以這做基礎，2006 年楊允言執行「台語文數位典藏資料庫（第二階段）」計畫，koh 進一步 kā chia ê 文本上網，另外 koh 補充一寡文本，予這個資料庫 ê 文本有 2163 篇，

攏總 258 萬外音節。文本有 POJ kah 漢羅對照，mā 有人工合成 ê 台語聲音輸出（請參考第四章），會使用年代、文類、作者、篇名來檢索，mā 會使全文檢索。網址 tī “<http://ip194097.ntcu.edu.tw/nmtl/dadwt/pbk.asp>”。(楊允言 2007)

六、台灣白話字文獻館

這是這是國科會數位典藏與數位學習國家型科技計畫 ê 內容分項公開計畫，李勤岸執行 ê 「台灣教會公報白話字文獻數位典藏計畫 1885~1969」，將《台灣教會公報》1885~1969 ê 文本，kah《芥菜子》1-22 期（是北部教會 ê 刊物，23 期以後就併入去《台灣教會公報》內底），揀部分 ê 內容拍字，而且轉寫做漢羅。除了文字，koh 有掃描一寡影像，包括台大圖書館特藏 ê 楊雲萍文庫內底 ê POJ 文獻。網址 tī “<http://www.tcll.ntnu.edu.tw/pojbh/script/>”。(呂美親 2008)

七、台灣民間文學館

這是文建會支持，羅鳳珠開發 ê，2007 年上線。資料來源是各縣市採集 ê 民間文學資料，大部分是胡萬川執行 ê 計畫，有台語 kah 客語，koh 有少部分 ê 文本有濫華語、日語。台語 ê 文本有 389 萬外音節。網址 tī “<http://cls.hs.yzu.edu.tw/tfl2010/>”。

十、白話字數位典藏博物館

這是廖瑞銘執行 ê 國立台灣文學館計畫成果，內底有 21 本 POJ 書寫 ê 冊，轉寫做 POJ kah 漢羅，所以會使做查詢，會當分段看文本（有 POJ kah 漢羅），mā 會當線頂看紙本。線頂看紙本 ê 時陣，會當利用系統提供 ê hām 鏡功能，kā 游標指 ê 位局部放大。網址 tī “<http://ip194093.ntcu.edu.tw/koa-a-chheh/bang-cham/>”。

十一、其它

教育部對 2008 年開始辦母語文學獎，2009 年改做台灣閩客語文學獎，mā 改做兩年辦一擺，到 2013 年已經辦四擺。歷屆閩客語文學獎作品集 ê 網址 tī “<http://www.topwin.com.tw/nativelanguage/works.html>”。另外，教育部 2007 年出版《台灣閩南語朗讀文章選輯》，是為著全國語文競賽台語朗讀 ê 部分，逐年競賽進前會公佈朗讀篇目，文本 khng tī 全國語文競賽網站頂懸。

《台灣教會公報》台文世界專欄，登記有台語文章刊出，chia ê 文本攏有上網，網址 tī “<http://www.tcn.org.tw/news.php?ntid=43&pid=7>”。台語信望愛內底 mā 有一寡台語文專欄。

近來 mā 有真濟台語文工作者用台文寫 blog，包括杜建坊、Lahjih、潘科元、王昭華、劉承賢、呂美親、胡長松、陳金順、黃元興、...等等，chia ê 台語文本攏 tī 網路頂面。

第四節 電子冊

一、台語文記憶系統

這是楊允言建立 ê 網站，主要是收集台語文相關書刊，盡量揣 khah 無版權問題抑是作者同意授權 ê，到 2014 年 1 月，有 1,725 份書刊，攏總 96,552 頁影像上網。建立這個網站 ê 用意是，雖然台語文過去有真濟出版品，m̄-koh tī 目前華文為主 ê 市場，chia ê 文本並無容易予人揣著，所以真濟人 m̄知影台語文有 chiah-nī 輝煌 ê 文字歷史，khòg tī 網路頂會當增加 chia ê 資料 ê 能見度。網址 tī “<http://ip194097.ntcu.edu.tw/memory/TGB/mowt.asp>”。

系統會當用書刊名、作者、出版社、出版年等等來查詢。可惜無像 Google books ê 技術，會當針對全文搜揣。(楊允言 2008)

二、其它

日本時代 ê 冊，包括國立台灣圖書館、國立台中圖書館、中研院、台大 kah 其它公立圖書館，有將 chia ê 冊掃描影像起來，內底有一寡台語相關 ê 冊。總是 chia ê 冊散 tī 一堆冊內底，khah 無好揣，另外，編目者可能對台語文 khah 無了解，所以內底有部分 POJ ê 冊，冊 ê 基本資料 (metadata) 註明是「台羅拼音」(台羅是 2006 年以後才有 ê) 抑是「日文」，這是 khah 可惜 ê 所在。

另外，陳柏中 tī Keyman 輸入法 ê 架構下面開發 POJ 輸入法，使用者若輸入“`a`a^a=a|a”，會顯示“áàâãä”。

林裕凱用陳鄭弘堯 tī Firefox ê 附加元件 Transliterato POJ 輸入法這個基礎，開發 Chrome 瀏覽器 ê 附加元件 Topoj，會當 kā 理解做 Transliterato POJ 輸入法 tī Chrome 瀏覽器 ê 版本。

三、信望愛台客語輸入法

信望愛台客語輸入法是信望愛資訊文化藝術基金會負責，透過台語界朋友募款，請電腦公司開發出來 ê 台客語輸入法，包括 Windows kah Mac 平台攏有，台語 ê 詞庫大概有 10 萬條以上，所以真濟語詞攏有，詞庫是幾 lō 冬來累積 ê，m̄-koh 因為詞庫資料 tī 教育部漢字規範出來進前就 teh 建立、累積，所以輸入法 ê 漢字用字並無完全照教育部 ê 規範，另外，少數 ê 詞條有錯誤，²¹需要 táuh-táuh-á 修改。

就是因為詞庫 iah-koh teh 增補，所以新版 ê 輸入法會自動檢查網路頂是 m̄ 是有 khah 新 ê 詞庫，若有就會通知使用者，點落去就會當 kā 新詞庫掠落來並且安裝好勢。

²¹ 有部分是羅馬字 ê 錯誤，另外有部分是漢字 kah 教育部 ê 規範無一致，總是教育部 ê 規範是 khah 後壁 ê 代誌，tī 教育部 ê 規範進前，漢羅台語文就 teh 使用，所以 mā 袂使講是錯誤，總是既然教育部 ê 漢字規範已經公佈，台語漢字建議盡量以教育部規範用字為主。

羅馬字 ê 部分會當選 POJ 抑是台羅，mā 有自訂詞庫，使用者 thang 好家已增加語詞。另外，雙音節抑是以上 ê 語詞會當免拍聲調、連字符，三音節抑是以上會使 kan-na 拍聲母抑是第一个字母（無聲母 ê 音節），可比講若 beh 拍出“台灣”就輸入“taioan”，若 beh 拍出“台灣人”就輸入“taioanlang”抑是“tol”（若台羅，分別是“taiuanlang” kah “tul”）。

Koh 有一个好用 ê 功能是 ruby，html ê ruby 標籤會當提供文字頂下排列 ê 功能，m̄-koh 標籤 ê 寫法長 lò-lò，這個輸入法會使設定 beh 拍 ruby，tī 這個設定下，若拍“pahhpiann”，會輸出：“<ruby><rb>拍拚</rb><rp></rp><rt>pahh-pià</rt><rp></rp></ruby>”，貼去 html 檔案，就會有頂下排列 ê 效果。



信望愛台客語輸入法 ê 網址 tī “<http://taigi.fhl.net/TaigiIME/>”。



圖 7-1 網頁頂面 ruby 排版結果

潘科元有寫文章介紹台語文羅馬字、漢字字型 kah 台語輸入法軟體，提供伊 ê 做法，網址 tī “<http://blog.xuite.net/khoguan/blog/73486235>”。

智慧型手機頂面 mā 有台語輸入法，而且是法國朋友 Pierre Magistry kā 咱設計 ê，會使用台羅 kah 注音符號，網址 tī “<https://play.google.com/store/apps/details?id=fr.magistry.taigime>”。(自由時報 2014)

第六節 斷詞

劉杰岳 tī 2006 年開發線頂漢羅台語文斷詞系統，利用台華線頂辭典 ê 詞條做基礎，輸入漢羅（羅馬字用數字調號 POJ），系統會顯示斷詞結果，mā 會提供這段文本 ê 詞頻表。假使斷詞結果有錯誤，會當增加使用者詞庫，改進斷詞結果，可惜使用者詞庫並無記錄落來，網頁若關掉，使用者詞庫就無去。斷詞演算法採用長詞優先、倒頭上大比對演算法，請參考第三章第三節。線頂斷詞系統網址 tī “http://poj.likulaw.info/hanlo_hunsu.php”。

第七節 雙文字轉換系統

一、白話字、台羅互轉

台語羅馬字文獻所使用 ê 文字系統是 POJ，台羅是 2006 年教育部公佈 ê 台語拼音系統，兩個系統雖然小可無全，m̄-koh 符號並

無衝突（若全一个符號 tī 無全書寫系統 ê 發音無全，表示符號有衝突）。

劉杰岳針對 POJ ê 輸入，開發 Taiwanese Package，利用 MicroSoft Word ê 巨集（Macro）來做轉換，可比使用者輸入“Goa2 si7 Tai5-oan5-lang5”，Taiwanese Package 提供 ê 巨集就會 kā 轉換做“Góa sī Tâi-oân-lâng”，mā 會當 kā 調符轉做數字。Taiwanese Package 支援到 Word 2003，khah 新 ê Word 版本無繼續支援。

林裕凱用 Taiwanese Package 做基礎繼續開發 TaigiLMJ101，會當用 tī Word 2003/2007/2010，提供 POJ kah 台羅，數字聲調 kah 調符 ê 互相轉換功能。使用者會當用這個家私，kā 規篇用台羅寫 ê 文本轉做 POJ，mā 會當 kā 規篇 ê POJ 轉做台羅。

台語信望愛網站提供線頂 POJ kah 台羅互轉 ê 功能，除了 POJ kah 台羅、數字聲調 kah 調符會當互相轉換，也顧慮著 POJ 符號 iah 未完全 tī Unicode 字元集進前一寡過渡性 ê 編碼，包括 HOTSYS kah Taiwanese Package 1.5 版進前 ê 編碼。

二、漢羅、羅馬字互轉

台語信望愛網站提供漢羅文本 kah 羅馬字（POJ）文本 ê 轉換系統，漢羅轉羅馬字系統 tī “<http://taigi.fhl.net/lohan/>”，羅馬字轉漢羅系統 tī “<http://taigi.fhl.net/hanlo/>”。基本上，背後是利用辭典做轉換。系統是蔡哲民開發，詞條是林俊育提供。

本章參考書目

- Embree, Bernard L.M.A. 1984. *Dictionary of Southern Min* (《台英辭典》). 台北:台北語文學院
- 自由時報. 2014. 〈台語輸入法 APP 法籍學生研發〉. 《自由時報》2014.1.12 生活新聞 <http://www.libertytimes.com.tw/2014/new/jan/12/today-life2.htm> (2014.1.20 檢索)
- 呂美親. 2008. 〈天書 e 化：台語白話字文獻數位典藏〉. 《數位島嶼電子報》22 期 http://content.teldap.tw/main/epaper_detail.php?doc_id=1168&epaper_id=42 (2013/12/30 檢索)
- 林俊育. 2007. 〈分享翻譯《台日大辭典》〉. 《台灣語言學一百周年國際學術研討會論文集》
- 莊德明, 鄧賢瑛. 2008. 〈文字學入口網站的規畫〉. 《第四屆中國文字學國際學術研討會》
- 曾正義. 1997. 《台語電腦強強滾：HOTSYS 台文系統操作手冊》. 台中:雙語
- 楊允言. 2003a. 〈台文華文線上辭典建置技術及使用情形探討〉. 《第三屆全球華文網路教育國際學術研討會論文集》p132-141
- 楊允言. 2003b. 〈近年來台語文資料處理的成果與展望〉. 《大漢學報》18 期 p71-78
- 楊允言. 2007. 〈台語白話文學 ê 全新表現——台語文數位典藏資料庫計畫簡介〉. 《台灣文學館通訊》15 期 p20-22

- 楊允言. 2008. 〈台語文數位典藏—以台語文記憶系統為例〉. 《第四屆台灣羅馬字國際學術研討會/中山醫學大學第三屆台灣語文暨文化研討會論文集》 p119-136
- 楊允言. 2010. 〈教育部臺灣閩南語字詞頻調查工作簡介〉. 《2010 第三屆台語認證研討會論文集》
- 劉杰岳, 楊允言. 2002. 〈白話字電腦文書處理 ê 研究〉. 《第四屆台灣語言及其教學國際學術研討會論文集》 p341-349
- 鄭良偉. 1993. 《台語電腦文書處理輸入法系統: TW301 使用手冊》. 台北:前衛
- 駱嘉鵬, 李良一. 2010. 〈《臺灣閩南語常用詞辭典》暨《閩南語字彙》網路版擴增連結計畫〉. 《教育部國語會《閩南語常用詞辭典》成果檢校研討會》 <http://w3.uch.edu.tw/luo4/jiapeng/d1-20.pdf> (2013/12/30 檢索)

本章提起ê網路資源

(以下網頁 ê 檢索時間攏 tī 2004 年 1 月。)

Charis SIL 字型. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=charissil_download

Chrome Extension Topoj. <http://khai.mtwww.mt.au.edu.tw/front/bin/ptdetail.phtml?Part=013>

DejaVu 字型. <http://dejavu-fonts.org/>

Doulos SIL 字型. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=doulossil_download

台語線頂字典. <http://ip194097.ntcu.edu.tw/TG/jitian/tgjt.asp>

台語線頂聖經. <http://taigi.fhl.net/list.html>

台灣民間文學歌仔冊資料庫. <http://ip194093.ntcu.edu.tw/koa-a-chheh/bang-cham/>

台灣民間文學館. <http://cls.hs.yzu.edu.tw/tfl2010/>

台灣白話字文獻館. <http://www.tcll.ntnu.edu.tw/pojbh/script/>

台灣教會公報台文世界專欄. <http://www.tcn.org.tw/news.php?ntid=43&pid=7>

台灣楷體、台灣宋體. http://wesingkasu.blogspot.tw/2012/01/blog-post_1223.html

台灣閩南語朗讀文章選輯. <http://140.111.34.54/MANDR/minna/first.html>

甘字典. <http://taigi.fhl.net/dick/>

白話字數位典藏博物館. <http://taibun.nmtl.gov.tw/bang-cham/>

老人囡仔性 閒閒罔話仙 blog. <http://chhantionglang.blogspot.tw/>

杜建坊个台灣厝 blog. <http://blog.xuite.net/getting.rich/hopefully>

花埕照日 blog. http://blog.roodo.com/cit_lui_hoe

信望愛台客語輸入法. <http://taigi.fhl.net/TaigiIME/>

南方的溫度 blog. <http://blog.yam.com/tionsgiong>

春天 e 霧 blog. <http://mypaper.pchome.com.tw/19660329>

烏衣行 blog 閩南語相關. <http://wesingkasu.blogspot.tw/search/label/閩南語>

國立台中圖書館日文舊籍數位典藏. <http://jdlib.ntl.gov.tw/>

教育部閩南語輸入法. <http://www.edu.tw/pages/detail.aspx?node=3683&Page=15638>

教羅台羅轉換工具. <http://taigi.fhl.net/CTS/>

智慧型手機台語輸入法. <https://play.google.com/store/apps/details?id=fr.magistry.taigime>

港邊惜別 blog. <http://blog.roodo.com/bichhin2007>

黃元興 blog. <http://blog.xuite.net/twgadang/twblog>

漢字構形資料庫. <http://cdp.sinica.edu.tw/cdphanzi/>

漢羅全羅互轉系統. <http://ip194093.ntcu.edu.tw/cl2hl/choan-lai-choan-khi.php>

漢羅轉羅馬字. <http://taigi.fhl.net/lohan/>

閩南語俗曲唱本歌仔冊全文資料庫 <http://hanji.sinica.edu.tw/?tdb=kua-a-chheh>

閩客語典藏計畫辭典整合查詢 http://minhakka.ling.sinica.edu.tw/bkg/chong-su-tian.php?gi_gian=hoa

潘科元台語文理想國 blog. <http://blog.xuite.net/khoguan/blog>

歷屆台灣閩客語文學獎作品集. <http://www.topwin.com.tw/nativelanguage/works.html>

幫助:如何輸入白話字. <http://zh-min-nan.wikipedia.org/wiki/幫助:如何輸入白話字>

雙倒水別莊 blog. <http://blog.roodo.com/senghian>

羅馬字轉漢羅. <http://taigi.fhl.net/hanlo/>

一扇窗，原來資訊 ê 專業會使運用 tī 語言處理方面。當其時讀 ê 教科書是 teh 討論英文 beh 按怎處理，華文彼陣才 beh 起步 niă。

因為這個因緣，退伍了後我就去中研院資訊所詞庫小組食頭路，tè 陳老師學工夫，陳老師成做後來我碩士 kah 博士 ê 共同指導老師。這期間考著清華資訊所去 hia 讀碩士，mā 繼續 tī 詞庫小組做兼任助理。彼陣開發中文 ê 語詞檢索系統，中文語料 ê 內碼 ùi 電信碼、shift BIG-5 beh 轉過去 BIG-5，轉袂過 ê 漢字就用“★”來表示，若去看 hia ê 語料，會看著滿天星，實務 ê 經驗予我深深感覺，資料若清氣，人生是彩色 ê（至少 khah 好食睏），資料品質若 bái，寫程式以外愛 koh 做真濟 dirty work。

除了語詞檢索，koh 有開發文稿校對系統，因為古文 beh 用人工校對實在真艱苦 mā 無效率，khah 省錢 ê 方法是請兩間拍字行拍全一份文本，了後用電腦系統揣出兩個電子檔案無全 ê 所在，人工看這部分就好。碩士論文做文件自動分類，tī 無用辭典 ê 情形，ùì 文章內底揣出會當做分類標準 ê 關鍵詞，用 chia ê 關鍵詞來做文件分類，彼陣 WWW 才當起步，Google iah 未出世。

Kā 時間 sèh 翻頭，做兵退伍轉台灣 ê 時，有接觸鄭良偉教授 kah 伊開發 ê TW301 台語輸入法，對有電腦背景 ê 筆者來講，beh 了解這個台語輸入法 ê 運作並無困難，會當家己修正、增補詞庫內容。讀碩士班 ê 時，真濟間學校已經有台語文相關社團，逐家做陣 tī 1992 年成立學生台灣語文促進會，我做第一任會長，會長

有一个重要 ê 工課，就是發落《台語學生》這份機關刊物 ê 發行，
《台語學生》一個月出三期，各校社團輪流編，負責 ê 社團愛家
己寫稿、拍字，我這片才彙[lūi]整，TW301 輸入法（包括我增補 ê
詞庫）對刊物 ê 發行有真大 ê 幫贊，彼時陣 ê 台語文刊物，親像
《台文通訊》等等，mā 受著 TW301 大幫贊。

1999 年筆者去大漢技術學院教書，思考家己 ê 研究方向，才
開始考慮台語文 kah 電腦結合 ê 可能性，總是心內小可驚驚，一方面，
進前並無人開創這個領域，是 m̄是會當行出一條路，一屑仔
把握都無，另外一方面 mā 無啥物基礎資源 thang 好運用，彼陣 iah
無線頂台語辭典，無充足 ê 語料。會記得有一擺請公假去世界台
文營報告台語相關 ê 電腦工具，轉來 beh 報出差，會計室 kā 出差
單退轉來，kā 你寫講「與專業不符」。一直到 2004 年決定 beh 讀
博士班，想 beh 做台語計算語言學方面 ê 研究，有幾個老師好心
kā 我提醒，做這可能無法度畢業。

有資源當然就 kā sa 來用，無資源就家己想法度 kā 生出來，進
前 tī 詞庫小組練工夫，這馬會當發揮 ah，啥物是台語無、華語有
ê 物件，kā 做看覓就著 ah，橫直有樣看樣無想家己想。總是台語
文 ê 書寫加減無一致，這表示愛做比中文 koh-khah 濟 ê dirty work。
Ùi 2000 年利用鄭良偉教授提供 ê 詞條開發台華線頂辭典，了後 kā
家己 ê 庫存語料加上請台文刊物編輯提供，以 chia ê 語料做基礎，
2003 年開發台語文語詞檢索系統，繼落來 koh kā 書架仔頂頭古早

人生 oh 得按算 ê 代誌 m̄-nā 這幾項，我 tī 30 歲 ê 時陣有做一擺大冒險，放棄台北穩定 ê 頭路走去花蓮度日子，頭前三年換五個頭路，換去第五個頭路，大漢技術學院做老師彼年（1999），校長 kā 阮講，舊年（1998）規年台灣 ê 出生人口比大專校院 ê 招生人數 koh-khah 少，所以 18 冬後會有危機，彼時學生濟，上課上甲歪腰，一點仔 mā 感覺袂出來，總是過無幾冬，學生數就開始直直落，敢 ná teh 趨流籠。有一擺參加研討會，洪惟仁老師 kā 我講：你 tī 大漢[Tài-hàn]袂大漢[tōa-hàn]！

去花蓮生活是為著 beh 過逐工看山看海 ê 日子，m̄-koh beh 大漢，必須愛有改變。後來到台中教育大學台灣語文學系教書（tī chia 上課攏用台語，有夠好 ê 好），koh 後來，離開蹓 16 冬蹓甲四序四序 ê 花蓮，規家伙仔搬來台中，雖然心內 iah-koh 定定 teh 數想，拍開窗仔門就會當看著青 leng-leng ê 太平洋。

決定 beh 寫作了後，這段期間搬兩擺厝，加減影響著寫作 ê 進度，總是 mā kā 交出來 ah。

Tī 這本冊，kā 阮所知 ê，最近 10 外冬來台語文語料處理方面相關 ê 進展 kah 成果，做初步 ê 整理 kah 記錄，希望 chia ê 成果，會當成做以後相關研究 ê 基礎，mā 希望透過相關線頂系統 ê 介紹，逐家 thang 好好仔來運用 chia ê 工具，加速台語文 ê 推展。總是個人見識有限，難免有落勾 ê 所在。

另外，mā 特別感謝秀蓮 kah 理揚幫贊校對、編輯、...等等出版進前 li-li-khok-khok ê 工課，予這本書會當順利出版。

歐美一寡大學有計算語言學 ê 相關系所，中國 ê 北京大學有計算語言學研究所，比較起來，台灣 ê 學界，對計算語言學這門學科是相對 khah 無重視 ê，華語是按呢，台語就 koh-khah m̄免講。（台灣 ê 語言學界另外一個問題是，普遍 khah 無注重本土語言。）

可能是因為台語文處理相對 khah 冷門，mā 可能同時有電腦技術 koh 對台語文有興趣 ê 人並無濟，加上伊 tī 資源分配方面本成就足弱勢。一路行過來 ê 感覺是，這條路不止孤單、辛苦，進展 mā 算無緊。

對台語文 ê 發展，到底應該愛樂觀抑是悲觀，逐家可能有無全 ê 看法。一方面，台語 tī 家庭 ê 使用 ná 像愈來愈少，少年家對台語愈來愈無感情；另外一方面，透過國校一禮拜一節 ê 本土語言課，koh 看著台語生澁 ê 一寡成果，甚至部分已經放棄使用母語 ê 家庭，因為囡仔學習 ê 需要，koh kā 母語 khioh 轉來用，有淡薄仔恢復 ê 趨勢出來。

大環境予咱警覺著台語可能會綴時間漸漸仔衰微去，因為當阿公、阿媽彼沿勻勻仔凋零，爸母彼沿因為讀書 ê 時 hōng 禁止講家己 ê 母語，造成母語能力降低，囡仔會當接觸母語 ê 機會減少，會自然失去對母語 ê 感情。M̄-koh 換另外一個角度來看，網路這個媒介，有伊深遠 ê 影響力，koh 予咱感覺會使透過這個工具，對母

附錄

附錄一 所參與開發ê系統列表

下面每一筆資料列出系統名稱、上線時間 kah 網址

- 台華線頂辭典, 2000, <http://ip194097.ntcu.edu.tw/q>
- 台語文語詞檢索系統, 2003, <http://ip194097.ntcu.edu.tw/TG/concordance>
- 台語線頂字典, 2003, <http://ip194097.ntcu.edu.tw/TG/jitian>
- 台灣俗語鹹酸甜, 2003, <http://ip194097.ntcu.edu.tw/TG/chu/kiamsngtiN>
- 台語笑諛代, 2003, <http://ip194097.ntcu.edu.tw/TG/chhiokhe>
- 台語羅馬字教學進修網站, 2004, <http://ip194097.ntcu.edu.tw/TG/TGLMJ>
- 社區母語資源中心, 2005, <http://ip194097.ntcu.edu.tw/Life/MTSociety>
- 家庭母語資源中心, 2006, <http://ip194097.ntcu.edu.tw/TG/MTFamily/>
- 台語音節詞頻統計報告, 2005, <http://ip194097.ntcu.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp>
- 白話字 Word count 工具, 2006, <http://ip194097.ntcu.edu.tw/nmtl/dadwt/count.htm>
- 國立台灣文學館台語文數位典藏資料庫, 2006, <http://ip194097.ntcu.edu.tw/nmtl/dadwt/>
- 台語文斷詞系統, 2007, http://poj.likulaw.info/hanlo_hunsu.php
- 台語文記憶, 2007, <http://ip194097.ntcu.edu.tw/memory/TGB/>
- 台語文全羅漢羅轉換系統, 2009, <http://ip194093.ntcu.edu.tw/cl2hl/choan-lai-choan-khi.php>

- ♦ 教育部台灣閩南語字詞頻統計, 2010, <http://wordfreq-tm.moe.edu.tw/>
- ♦ 國立台灣文學館白話字數位典藏博物館, 2010, <http://taibun.nmtl.gov.tw/bang-cham/>
- ♦ 台語文發音系統, 2011, <http://ip194093.ntcu.edu.tw/cl2hl/taigi2siann.php>
- ♦ 國立台灣文學館台灣民間文學歌仔冊資料庫, 2011, <http://koaachheh.nmtl.gov.tw/bang-cham/>
- ♦ 台語文音節統計, 2011, <http://ip194093.ntcu.edu.tw/sng-im-chat.php>
- ♦ 台語民間書寫, 2013, <http://ip194097.ntcu.edu.tw/TG/binkan/>

白話字 kah 台羅主要差異

1. 聲母

POJ	ch	chh
台羅	ts	tsh

2. 韻母

POJ	o̍	ⁿ	oa	oe	eng	ek
台羅	oo	nn	ua	ue	ing	ik

3. 舉例

POJ	cháu-chhōe	lō̍-ēng	khùì ⁿ -oáh	hōe-ek
台羅	tsáu-tshuē	lōo-īng	khùìnn-uáh	huê-ik
漢字	走揣	路用	快活	回憶

Research on Corpus Processing and Online Resource for Written Taiwanese

Table of Contents

Chapter 1 Introduction

Section 1 Background

Section 2 Different Types of Written Taiwanese Scripts

Section 3 Issues Related to Written Taiwanese Processing

Chapter 2 Coding, Searching and Display of POJ

Section 1 Character Code of POJ

Section 2 Two Kinds of POJ Representation

Section 3 Search Problem with POJ Text

Section 4 POJ Text Display

Chapter 3 Some Text Processing Utilities for POJ

Section 1 POJ Phoneme Segmentation and Spelling Checker

Section 2 POJ Syllable/Word/Sentence Count

Section 3 Syllable Count for Han-Romanization Mixed Script

Section 4 Word Segmentation for Han-Romanization Mixed Script

Chapter 4 Taiwanese Tone Sandhi Processing

Section 1 Tone Sandhi Problem of the Taiwanese Language

Section 2 Implementation of the Taiwanese Pronunciation System

Section 3 Rule-based Taiwanese Tone Sandhi Algorithm

Section 4 Results and Accuracy Rate

Section 5 Related Discussion and Future Direction