教育部臺灣閩南語字詞頻調查工作簡介

國立成功大學台灣語文測驗中心
專案計畫助理研究員 楊允言

摘要

詞頻調查是一個語言 ê 基礎統計。教育部 tī 2008 年委託學術單位進行台語字詞頻調查統計，按算 beh 蒐集 100 萬詞以上 ê 台語語料，並且提供語詞檢索系統 hō͘社會大眾查詢。

本文介紹這個詞頻調查計畫，伊 ê 重要性 kah 目標，工作團隊 ê 分工情形，mā 針對所蒐集 ê 語料分佈做簡單 ê 說明，包括各文類（15 小類）ê 比例，文字型式（漢字、漢羅合用、羅馬字）ê 比例，無仝年代 ê 文本比例來對照台語文發展歷史。Soah 落來說明計畫執行 ê 實務問題，包括工作流程，輸入 ê 格式，羅馬字轉換 kah 漢字造字 ê 處理方式。

關係語詞檢索系統 ê 功能，使用者 thang 利用詞組、語詞、詞頭、詞中、詞尾、漢字造字、重疊詞等來查詢，mā 會當設定語料範圍，顯示方式、排等 kah 取樣數量。

本文 mā 列出這個計畫 tih 執行 ê 時，所 tú 著 ê 一寡問題，包括斷詞原則、文類分類方式 kah 漢字使用，希望會當提供參考。最後，討論未來咱 koh 愛做 ê 空課。

關鍵詞：台語文、語料庫、字詞頻、語詞檢索、斷詞

1. 踏話頭：詞頻統計的重要性

台灣 ê 本土語言，包括台語、客語 kah 原住民語是台灣倚起 tī 世界 ê 重要文化資產，iáh 是文化創意產業 ê 重要基礎之一。對外，透過語言 ê 情境，hō͘台灣文學、戲劇 kah 常民文化得著豐富 ê 性命力，進一步得著國際上 ê 認同，hō͘台灣文化 tī 文化創意產業方面有充分 ê 競爭力；對內，透過本土語言教學 kah 各種 ê 拍拚，hō͘咱珍貴 ê 語言資產 thang 繼續保存。

過去，tī 錯誤 ê 語言政策影響之下，台灣本土語言受著相當大 ê 打擊。聯合國教科文組織 (UNESCO) tī 2001 年公布 ê 資料，直接點名台灣是母語瀕危地區，2003 年，當時行政院客委會主委葉菊蘭對本土語言有一段生動 m̄-koh mā 真 hō͘人心酸 ê 形容：「台語掛號中，客話急診，原住民語入加護病房。」根據語言學家 ê 估計，目前全世界大約有 6,000 外種語言，若 koh 無採取任何 ê 挽救動作，到 21 世紀尾，可能 kan-na chhun 600 種語言。

所以，設使咱無拍拚，據在台灣各語言自生自滅，100 冬後 ê 台灣，可能 kan-na chhun 華語 kah 送入加護病房 ê 台語，其它本土語言 lóng 已經滅無。語言成做一陣人 ê 思考、表達、溝通、自我認同 ê 工具，是世界人類文化 ê 資產，任何一種語言 ê 死亡，就親像地球上任何一款物種死亡共款 hō͘人感覺怨嘆。

台灣本土語言中，台灣閩南語有 70%以上 ê 使用人口（超過一千五百萬人）。另外，閩南語 tī 世界上有將近四千外萬人 ê 使用人口，tī 全世界所有語言人口使用排名第 21，chia ê 事實 lóng 顯示出，無論是 beh 強調台灣 ê 重要性，iah 是 ùi 投資報酬 ê 角度來看，這個語言非常值得咱 ê 重視。

2. 教育部台語詞頻調查簡介

爲著 beh 提供教育部九年一貫課程綱要內底，台語每一個階段所學習 ê 漢字、詞彙 kah 辭典編輯、教材編輯 ê 參考，所以規劃委託辦理台語字詞頻調查工

作。Tī 教育部 ê 規劃內底，第一階段字詞頻統計資料，主要是 beh 觀察目前台語字詞 ê 使用情形。

初步 ê 規劃，希望這個委託案會當達到以下幾個目標：

(1) 建構至少有 100 萬台語語詞 ê 基礎語料庫；

(2) 根據這個基礎語料庫，統計台語 ê 詞頻；

(3) 根據台語詞頻統計 ê 結果，製作台語字詞調查報告書；

(4) 利用這個基礎語料庫，做出一個語詞檢索系統（concordancer system）提供一般民眾使用。

基礎語料庫 ê 資料蒐集範圍，包括台語雜誌期刊、教材、流行歌、囡仔歌、囡仔詩、俚俗諺語、演講比賽文章、朗讀比賽文章 kah 文學作品（小說、散文 kah 戲劇等等）；戲劇部分希望包括李天祿布袋戲劇本 kah 文建會補助 ê「雲林拱樂社」歌仔戲劇本。另外，無 beh 收錄字、辭典 ê 資料。文類愛儘量平均、多樣，kāng 一個作者 ê 作品 mài 超過總資料量 ê 千分之五。每一筆語料，lóng 愛註明出處，而且至少包含「作者」、「出版單位」、「出版年代」kah「書寫系統」。

另外，因爲台語文語料 ê 文字型式包括全羅馬字、漢羅合用 kah 全漢字，beh 做詞頻統計，需要將所蒐集 ê 語料做斷詞，原則上，以教育部《臺灣閩南語常用詞辭典》做原則，若有需要，會使參考教育部「臺灣閩南語羅馬字拼音方案連字符使用基本規則」kah 中央研究院所制訂 ê「中文分詞原則」（CNS14366）。

這個計畫其中一項重要 ê 目標，是 beh 做教材編輯 ê 參考，所以蒐集著 ê 語料，特別 kā 國校教材 ê 部份獨立出來，統計伊 ê 詞頻，thang 來做詞類分級 ê 參考資料。

這個委託案 tī 2008 年 5 月公告，6 月開審查會議，後來通過請大漢技術學院負責執行這個調查案，計畫主持人是楊允言，協同計畫主持人是張學謙，執行期間 ùi 2008 年 7 月教 2009 年 7 月。

3. 詞頻調查工作 ê 進行

3-1 工作團隊

一個計畫 ê 執行，需要真 choē 人 ê 互相配合。這個計畫，除了基本 ê 行政事務，khah 要緊 ê 是台語拍字 ê 人才，而且 mā 需要有對台語處理技術有了解 ê 程式開發人員，計畫 chiah 做會好勢。

另外，tī 資訊時代，工作團隊未必需要坐 tiàm kāng 一間辦公廳，透過網路傳資料、互相連絡 lóng 不止仔利便。下面所列 ê 人，分散 tī 台灣各所在，甚至 mā 有 tī 美國。

本計畫 ê 工作團隊，包括：石博元、莊慧娟、蔡嘉如、蔡瑋芬、李欣民、廖淑鳳、林俊育、賴淑玲、雷明漢、王寶漣、謝月華、陳德樺、鄭佳旻、…等等，協助整理語料、拍字。

另外，洪惟仁、曾國榕、蔣爲文、吳仁瑟、姚榮松、林佳怡、林麗黛、梁淑慧、施炳華、蕭藤村、張玉萍、塗素珠、…等人，協助提供資料，有 ê 並且提供電子檔案 hō͘本計畫。

李盛安、袁崇祐等人協助安裝電腦、架設 Server，知行科技公司協助開發規個系統。詹金來、李懿軒、李承泰、沈惠珠等人協助整理計畫相關資料。林素朱對斷詞提供專業 ê 諮詢。

蔣爲文、高成炎、林俊育、陳克健等人擔任諮詢委員，對本計畫指點方向 kah 提供意見。

教育部方面，有六個審查委員，mā tī 計畫進行 ê 期間提供專業 ê 建議；甘佩玉負責計畫 ê 連絡。

3-2 語料蒐集建檔

語料 ê 蒐集，愛考慮 ê 點真 choē，除了教育部原來要求 ê 文類愛平均、每一個作者 mài 超過語料量 ê 千分之五以外，本計畫 mā 考慮兩個因素：

(1) 台語文語料 ê 文字型式，有全羅馬字、漢羅合用 kah 全漢字，這三款書寫方式盡量攏愛收入來；

(2) 因爲過去成百冬來，台語語詞 ê 變化真大，所以每一個時代 ê 台語文 mā 愛盡量 lóng 收入來，並且盡量平均。

另外，因爲國校教材是其中 ê 一個重點，愛盡量，總是實際上去蒐集國校台語教材 ê 時，chiah 發現困難重重。民間出版社台語教材 ê 部份，有幾份教材，出幾冬了後，出版社無繼續出， in 家己倉庫 ê 庫存留無 chiâu；甚至 koh 有出版社已經倒 a，電話是空號，chhē lóng chhē 無。有 ê 出版社繼續 teh 出，m̄-koh 逐年攏 koh teh 修改內容。

各縣市出 ê 國校教材部份，有 ê 出無 chiâu-chn̂g，出幾冊了後無 koh 出。Beh 蒐集 chia-ê 教材 mā 無簡單。有 ê 縣市電話轉來轉去轉 kah 無去，chhē 無負責人；有 ê 請阮寫公文去，公文送出去了後 tō 全然無消息。爲著國校教材 ê 部份，開足 choē 時間。[1]

若 beh 討論詞彙分級，一個版本 12 冊 lóng chiâu-chn̂g chiah 有意義，所以 12 冊 lóng chiâu-chn̂g ê 版本 chiah 收錄 tī 教材類，無 chiâu-chn̂g ê tō 分 tī「其它教材類」。

經過 kah 計畫審查委員 ê 討論，語料粗分做「創作文學」kah「口傳文學」兩大類，大類下面 koh 分小類，請參考表 1。

[1]這個過程 hō͘筆者真大 ê 感慨，九年一貫課程 chiah 實施猶無十冬，教科書市場 ê 變化 tō chiah-nī 大。Chia ê 教材有可能藏 tī 某一間國校 ê 圖書館，隨時有可能 hông 當做糞埽清掉。無人想著來成立國校台語教科書博物館，kā 這段過程記錄起來。筆者 mā pat 直接共出版社講，既然恁已經無 koh 出 a，是 m̄是同意阮共 chia ê 寶貴 ê 資料掃瞄上網，總是 in 以著作權 ê 理由無同意這層代誌。

表 1 語料文類一覽表

大類	小類	語料比例
創作文學 (44.97%)	報導文學	8.33%
	演講訪談	5.63%
	散文	8.91%
	小說	10.26%
	論文	3.16%
	流行歌	8.68%
口傳文學 (44.16%)	褒歌	3.42%
	囡仔歌	1.35%
	謎猜	0.58%
	歌仔冊	12.13%
	俗諺	3.82%
	民間故事	19.14%
	戲劇	3.72%
教材 (10.86%)	教材	2.88%
	其它教材	7.99%

特別 kā 教材分開，主要 ê 原因是，教材 ê 內容，有創作 ê，mā 有口傳文學 ê。[2]

文本各種文字型式 ê 比例，請參考表 2：

表 2 語料文字型式一覽表

文字型式	語料比例
漢字	37.76%
漢羅	40.89%
羅馬字	21.35%

[2] 按呢 ê 講法，並 m̄是完全無問題。劇本屬戲劇類，戲劇 tī 口傳文學下面，m̄-koh 劇本是創作 ê。總是，大類是 khah 粗 ê 分類，目的是 beh hō͘人有簡單 ê 初步 ê 印象。

語料量 tī 各年代所佔 ê 比例，請參考表 3：

表 3 語料年代一覽表

年代	語料比例
1880 年代	0.84%
1890 年代	2.77%
1900 年代	0.70%
1910 年代	6.01%
1920 年代	4.74%
1930 年代	5.17%
1940 年代	0.66%
1950 年代	3.50%
1960 年代	2.98%
1970 年代	0.75%
1980 年代	0.83%
1990 年代	18.34%
2000 年代	35.44%
Unknown	17.28%

用 10 年做一個單位來算，純粹是爲著方便。其中，愈早期 ê 文本愈 oh chhē，這可能會當解說是按怎 1880 kah 1900 年代語料量少 ê 原因；另外，1940 年代應該是戰爭 ê 關係。戰後國民黨政權實施戒嚴，台語文 mā koh 有 tih 繼續。1969 年進一步禁止教會公報使用白話字發行，台語文發展進入烏暗期，1990 年代台語文開始復興，2000 年代 koh 加上台語變做國校 ê 正式課程，台語相關系所 ê 成立，創作量 koh-khah 增加。Ùi 本計畫蒐集著 ê 語料 ê 時間分佈，加減 kah 台語文 ê 發展狀況有符合。

Unknown 表示所蒐集著 ê 文本無法度確認年代。有 ê 文本雖然知影出版年，總是創作（抑是記錄）到出版 mā 可能 koh 經過一段時間。

3-3 語料整理方式

本計畫所蒐集 ê 語料，羅馬字 ê 部份，不管伊原來 ê 文本是用 tó 一款羅馬字系統，lóng 會先轉寫做教育部公佈 ê 台羅，這是爲著後來查詢 ê 利便；漢字 ê 部份完全尊重作者 ê 寫法無修改，按呢 mā 會當反應台語漢字書寫 ê 多元現實。

爲著 beh hō͘使用者 thang 對照看，每一份語料用兩種文字型式整理。原來 ê 文本假使是全漢字抑是漢羅合用，會 koh 轉寫做全羅馬字；全羅馬字文本，mā 會 koh 轉寫一份漢羅合用 ê 版本。

轉寫做羅馬字，主要會產生 ê 問題是腔口，原則上盡量以優勢腔爲主。轉寫做漢羅 mā 可能對詞頻統計造成一寡影響，因爲民間 ê 漢字書寫真無一致，可比講「chhit-thô」這個詞，本計畫蒐集 ê 語料，漢羅有 11 種無仝 ê 寫法，[3]可能有人有趣味 beh 知影 tó 一款寫法 khah choē，m̄-koh 若原來文本是全羅馬字，轉寫者寫做漢羅，轉寫 ê 漢羅寫法若算在來，tō 對原來漢字（漢羅）書寫 ê 統計造成一寡偏差。爲著解決這個問題，詞頻統計 ê 部份，有所有 ê 語料 ê 詞頻統計，mā 有 kā 全羅馬字文本提掉了後 ê 詞頻統計。

羅馬字 ê 輸入是另外一個問題。工作團隊內底負責拍字 ê 人，有 ê 拍白話字 khah 慣勢，有 ê 對台羅 khah 熟手，目前雖然已經有系統會當來轉換，m̄-koh mā 有淡薄仔費氣。好佳哉白話字 kah 台羅符號無衝突。後來這部份，tō 交 hō͘系統來轉換，不管負責拍字 ê 人用白話字抑是台羅，用調符抑是數字表示聲調，系統 tī 顯示 ê 時，lóng 會 kā 轉做台羅。

下面是輸入 ê 範例：

[3] 這 11 種寫法分別是：tshit-thô、蹉跎、佚陶、tshit 迢、彳亍、七桃、[辶＋日]迢、逸陶、佚佗、擦桃、遊戲，其中，「[辶＋日]」表示這個漢字是「辶」kah「日」二字門起來 ê，因爲 Unicode 內底無這個漢字，所以 tī 輸入 ê 時，輸入「[辶＋日]」，系統到時用圖形 ê 方式來顯示。

文類	民間故事	
書寫系統	漢羅	
作者	李瑞枝	
出版年	1997	
篇名	十二生肖的由來	
段	今年是丁丑年，講，哦，今年是牛年哦，肖牛的，逐个攏愛安太歲。	kin-ni5 si7 teng-thiu2 ni5, kong2, ouh, kin-ni5 si7 gu5 ni5--ouh, siunn3 gu5 e5, tak8 e5 long2 ai3 an thai3-soe3.
段	肖牛仔，是沖太歲，呼，著愛安太歲，所以注意著這个十二生肖。咱台灣人，每日都愛用，無論創啥，好事、歹事，無論創啥，撞著就是講：「你肖啥？」「今年幾歲？」	siunn3 gu5-a2, si7 chhiong thai3-soe3, houh, tioh8-ai3 an thai3-soe3, sou2-i2 chu3-i3 tioh8 chit e5 chap8-ji7 senn-siunn3. lan2 Tai5-oan5-lang5, mui2-jit8 to ai3 iong7, bo5-lun7 chhong3 siann2, ho2-su7, phainn2-su7, bo5-lun7 chhong3 siann2, tng7 tioh8 chiu7-si7 kong2: “li2 siunn3 siann2?” “kin-ni5 kui2 hoe3?”
	…	…

輸入分做三欄，其中第二欄是漢羅，第三欄是羅馬字。若是無需要分漢羅 kah 羅馬字 ê，主要是文本 ê 基本資料（metadata），tō kā 第二欄 kah 第三欄合併。

第一欄表示資料 ê 性質，事先規定好 ê，包括：

- 文類：填寫 tī 表 1 所列 ê 小類，lóng 總有「報導文學」、「演講訪談」、「散文」、「小說」、「論文」、「流行歌」、「褒歌」、「囡仔歌」、「謎猜」、「歌仔冊」、「俗諺」、「民間故事」、「戲劇」、「教材」、「其它教材」等 15 個類別
- 書寫系統：填寫 tī 表 2 所列 ê 文字型式，lóng 總有「漢字」、「漢羅」、「羅馬字」等三種書寫系統
- 作者
- 出版者
- 出版年
- 書名
- 篇名
- 年級別：這是教材類 chiah 需要填寫 ê 資料，有「[一|二|三|四|五|六][上|下]」攏總 12 個年級別
- 段：tō 是語料 ê 內容，一個段落做一欄

內容 ê 部份，分做漢羅（全漢字）kah 羅馬字兩欄，親像頭前講 ê，羅馬字

用白話字、台羅 lóng 會使，調符、數字聲調 mā lóng 會使。

根據語料 ê 整理方式，系統會 kā 漢羅 kah 羅馬字一個語詞一個語詞對起來。假使漢羅 kah 羅馬字文本對 boē 起來，表示輸入 ê 時有 têng-tâⁿ，系統 tī 顯示 ê 時會用紅色標示，按呢會當倒頭幫贊輸入者做校對。另外，因爲羅馬字有連字符，所以這份語料等於有做人工斷詞。

3-4 造字處理方式

中央研究院 tih 處理中國古籍 ê 時，tú 著真 choē 造字，爲著解決相觀問題，in 發展一套欠字處理 ê 系統， in 制訂「構字式」做造字 ê 編碼，利用圖形來顯示。本計劃採用中研院 ê 欠字處理系統來解決語料中 ê 漢字欠字問題，用圖形顯示，使用者無需要另外安裝造字檔案，跨網路平台 mā boē 發生問題。

構字 ê 原則，是 kā 一個漢字（造字）看做是兩個漢字組合起來 ê，組合 ê 款式有倒 pêng 正 pêng、頂面下面 kah 外面內面三種。爲著輸入 ê 利便，本計劃 tih 輸入造字 ê 時陣，採用下面 ê 符號：

- 倒 pêng 正 pêng：以 「[倒＋正]」表示，可比「[會＋勿]」
- 頂面下面：以 「[頂^下]」表示，可比「[伊^心]」
- 外面內面：以 「[外@內]」表示，可比「[門@丨]」

系統讀著 chia ê 符號 ê 時，會去中央研究院 ê 欠字系統討這個漢字 ê 圖形來顯示。M̄-koh 並 m̄是所有 ê 台語漢字造字 lóng 有收錄 tī 中研院 ê 欠字系統內底，chia ê 漢字，咱需要整理、chhoân 咱做好 ê 這字 ê 圖形，kā 中研院申請。

3-5 語詞檢索功能

系統提供 ê 功能，主要包括語詞檢索 kah 詞頻統計兩大功能，同時，mā 對使用者使用本系統 ê 情形做使用記錄，成做未來 beh 改進本系統 ê 參考。

語詞檢索系統提供使用者查詢某一個字串(string)，這個字串可能是一個音節（字）、詞 ê 一部份、一個詞抑是詞組，系統入去語料庫 chhiau-chhē，kā 出現使用者輸入 ê 字串 chhē 著，並且順 soà kā 這個字串 ê 前後文做陣 liȧh 出來。

下面是查詢「成大」ê 部份結果：

啊 去 到 半路 的 所在，有 一 塊 石壁，	成 大	塊 安呢 啦，啊 若 風颱雨 若 到，啊
啊 尾溜 才 自 安呢，名 就 叫 大甲，呼，	成 大	甲 安呢。
這 个 員外 煞 君 一 坩 羊肉，呼！啊	成 大	坩 安呢。啊 捧 來 到 這 个 風水仙
正經 成實 去 自殺，我 麼 是 會 擱 問題	成 大	。啊 當 好 啦！好心，您 做 好心 共
煞 若 干若 豬母 咧 啦！煞 一 箍 腹肚	成 大	箍，啊 得 卜 生產 這 八 个 實在
成 濟 啦。啊 兵馬 直直 偎 來，啊 聲勢	成 大	啦。啊 嘉義囉 就 順續 攻 去 府城 台南
激氣，實在 毋 是 款，毋 忍 毋 耐，小事	成 大	，人 講 人情 留 一 線，日後 好 相

這是選擇用漢羅顯示 ê 結果，使用者 mā 會使選擇用 (a) 羅馬字 (b)頂面羅馬字下面漢羅 (c)頂面漢羅下面羅馬字 來顯示，親像下面 ê 例：

(a) thinn-kong bô kong-pênn, pù--ê pù tsiūnn-thinn, sàn--ê sàn tshùn thi̍h, ai-iò-ai-iò, bô thâu-lōo ê hiann-tī.

(b) 蕃薯 遐 的 幼的 就 共 爌，爌 一 堆 成 大 堆
han-tsî hia ê iù--ê tō kā sa̍h sa̍h tsit tui tsiânn tuā tui

(c) m̄-kú 不拘 guán 阮 Tâi-uân 臺灣 lâng 人 tshuā 娶 bóo 妻 sī 是 bô 無 tshin 親 tshiūnn 像 lín 您 luē-tē 內地 lâng 人

查詢所提供 ê 功能 koh 包括：

(a) 使用者輸入 ê 字串，會使指定伊是詞組、一個詞、詞頭(prefix)、詞中(infix)、詞尾(suffix)，方便使用者 chhē 著伊真正 beh 愛 ê 語料；

(b) 若是用羅馬字查，用白話字、台羅 lóng 會使，用數字抑是調符 mā lóng 會使，hō͘使用者減輕轉換 ê 負擔；

(c) 因爲漢字造字 oh 輸入，所以系統提供一個造字表，使用者會使直接用點 ê；

(d) 提供重疊詞 ê 查詢，有「AA」、「AAA」、「AAB」、「ABB」、「ABAB」、

「AABB」、「ABAC」、「ACBC」、「ABCAB」等等；

(e) 查詢 ê 結果，會當用查詢詞 ê 前一詞抑是後一詞來排等(sort)；

(f) 有 ê 語詞頻率較 koân，若查詢結果傷 choē 筆，會造成使用者觀看 ê 負擔，所以系統提供取樣(sampling) ê 選項；

(g) 若是 beh 觀察某一個語詞 tī 無仝文類 ê 表現，使用者會當選擇 kan-na 揀選某一個抑是某幾個文類 ê 語料來查詢。

3-6 詞頻統計功能

本系統 lóng 總 chhoân 三款詞頻統計 ê 資料，包括

(a) 所有語料 ê 詞頻統計；

(b) 書寫文字是漢羅 kah 全漢字 ê 文本 ê 詞頻統計：這是爲著 beh 觀察民間漢字書寫 ê 情形，避免受著拍字者 kā 全羅馬字文本轉寫做漢羅文本 ê 時，mā kā 轉寫者 ê 漢字書寫算入去，造成偏差；

(c) 教材小類 ê 詞頻統計；主要 beh 做國校台語教材詞類分級 ê 參考。

Ùi 技術面來看，這三款詞頻統計 lóng 仝款，kan-na 精差 tī 語料集合是無仝 ê。

4. 相關問題

下面列出這個計劃 tih 執行 ê 時，所 tú 著 ê 一寡問題：

(a) 斷詞是一個大工程：

漢字書寫，詞 kah 詞之間無界線，無法度直接 kā 一個一個 ê 詞分出來。技術上，咱會使利用辭典 ê 詞條做輔助，利用電腦來斷詞，m̄-koh 無 thang 百分之百正確；

早期 ê 全羅馬字文本，詞 kah 詞之間有 làng 格，等於是經過人工斷詞過，

總是 mā 會有一寡無一致 ê 所在，可比有 ê 寫「chit-ê」，有 ê 寫「chit ê」；咱若看華文 ê 部份，其實問題 mā 是仝款，台灣 kah 中國針對華文斷詞 ê 標準 tō 小可仔有精差。台語文面對書寫系統無一致 ê 問題，標準化 ê 路途一定 boē 真平順。

曾金金(1997) pat 根據中研院 ê 華語分詞（斷詞）標準，來討論台語斷詞原則，原則 hoān 原則，假使咱無根據這個分詞標準所做出來 ê 分詞辭典，斷詞原則 ê 實踐 tō 有伊 ê 困難；

另外是斷詞 ê 判斷需要訓練，咱 ê 語文教育對斷詞 ê 訓練是無夠 ê；這對相關計畫 ê 執行 lóng 是必須愛面對 ê 大問題；

教育部針對台語 ê 羅馬字書寫，有制訂連字符 ê 規範，這對台語斷詞 ê 空課有一寡幫贊；總是咱發現，tī 人名 ê 部份，台語 kā 姓 kah 名拆做兩個詞，華語是共姓 kah 名合做夥，互相並無一致；

另外，表示輕聲抑是隨前變調 ê “--”，連字符連接 ê 兩 pêng，有時仔是一個詞 (A-bêng--a)，有時仔是兩個詞 (chiảh --chit chhùi)。實務上，這對技術處理 ê 人來講，除非 tī 語料內底藏一寡標記，若無，無法度決定到底是一個詞抑是兩個詞，m̄-koh 對整理資料 ê 人來講，beh 做 chia ê 標記 mā 是真困難，尤其語料量大、資源有限 ê 時陣，這 koh-khah 是大困難。

(b) 文類 ê 決定，語料 ê 比例：

最後 ê 文類 kah 計畫一開始所定 ê 文類並無仝款。開始蒐集著一寡語料，提出來討論了後，tō 發現是按怎某一寡文本無 tī 內底，爲著 beh khah 全面，tō 開始調整文類。調整文類 ê 時，有 ê 文本 mā 愛 ùi 某一個文類徙去另外一個文類。

語料庫蒐集 ê 空課，lóng 一定會 tú 著文類、體裁 ê 問題，以英文做例，

無仝語料庫，分類 tō 無一定仝款。

啥物是適合台語 ê 文類分類？目前可能抑無明確 ê 答案，必須愛透過進一步 ê 考察抑是計算，chiah 有新 ê 想法。

文類一旦定出來了後，爲著 beh 符合台語書寫現實抑是台語使用現實，每一個文類 beh 佔語料外 choē 比例，koh 是值得討論 ê 大問題。這部份，專家 ê 意見 mā 真無一致。

因爲文類 tī 計畫執行 ê 時修改幾遍，定案了後 koh 討論語料佔各文類 ê 比例，致使爲著 beh 符合這個比例，愛放棄一寡已經整理好 ê 文本，抑是繼續補充其它 ê 文本。這 mā 是這個計畫並無 tī 規定 ê 時間內（一冬）順利完成上主要 ê 原因。

(c) 漢字使用 ê 問題：

民間漢字使用 ê 情形，若斟酌入去深究，真正是大問題。一方面逐家寫法真無一致，有 ê 是仝一份文本，寫法 tō 無一致；另外一方面，造字 mā 是真費氣 ê 代誌，爲著 beh 表示這個漢字書寫現實，加開足 choē 冤枉時間。

咱受華語 ê 影響 mā 真大，所以「chhit-thô」有人 kā 寫做「遊戲」、「bô」有人會寫做「沒」。咱可能會認爲，會寫出「遊戲」，表示伊 ê 台語受著華語影響，講 boē 出端 tiah ê 台語；實際 ê 情形應該是，台語猶原真端 tiah，m̄-koh 台文受華文 ê 影響，寫 boē 出端 tiah ê 台文。畢竟現時咱受華文教育 ê 影響太大，換一個角度看，mā 是因爲堅持 beh 用漢字所產生 ê 問題：華文、台文膏膏纏花 boē 清。

甚至有人 kā 「in」寫做「他們」，因爲羅馬字文本 kah 漢羅文本必須愛一音節一音節對起來，chiah 會當正常運作，所以這款情形，阮只好 kā 漢字 ê 部份改做「in」。

5. 結論 kah 未來方向

Tī 台灣所有 ê 本土語言內底，雖然得著官方 ê 資源是上少 ê，m̄-koh 台語是文字化發展了上有活力 ê，一方面，這表示台語文字化 beh 成功 ê 機會真大；另外一方面，愈有活力，mā 表示書寫愈多元，假使無政策 tī 背後推 sak，標準化 ê 過程有可能拖真長。M̄-koh，標準化制訂 ê 過程，無應該是少數人門關起來討論 tō 準 tú-soah，伊應該關照著過去 ê 書寫現實，制訂 ê 結果 mā 應該接受社會大眾 ê 檢驗 kah 試用。

論真講起來，台語書寫 ê 規範教育部是先訂標準，chiah 倒頭來做調查，另外，所訂 ê 標準 mā hō͘人感覺是以漢字做中心。總是，這擺 ê 字詞頻調查是一個好 ê 起點。

咱會當做 ê koh 真 choē，包括：

- 擴大語料 ê 規模，做 koh-khah 全面 ê 調查；
- Ùi 調查結果倒頭來討論已經公佈 ê 七百字詞，甚至是台灣閩南語常用詞辭典萬外條詞條 ê 漢字用法是 m̄是妥當；
- Ùi 實際書寫 ê 漢羅合用現象，認真思考常用字詞敢一定愛用怪漢字；
- 思考調查 ê 結果 beh 按怎運用 tī 台語教學頂面；
- 針對台語分詞規範做 khah 詳細 ê 討論，kah 華語分詞標準無仝 ê 所在，需要做 khah 深入 ê 討論，mā 愛建立符合台語分詞 ê 辭典；
- 利用調查 ê 成果，進一步鼓勵台灣人 giảh 筆書寫家己 ê 語言。

參考資料

Biber, Douglas, Susan Conrad and Randi Reppen, 1998, *Corpus Linguistics : Investigating Language Structure and Us*e, New York : Cambridge University

中央研究院，現代漢語標記語料庫, http://dbo.sinica.edu.tw/SinicaCorpus/

中央研究院詞庫小組, 1998, Accumulated Word Frequency in CKIP Corpus，台北

教育部, 2007，臺灣閩南語羅馬字拼音方案連字符使用規則 http://www.edu.tw/files/site_content/M0001/lanrule.pdf

教育部, 2008，教育部台灣閩南語常用詞辭典, http://twblg.dict.edu.tw/tw/index.htm

莊德明，謝清俊, 2005，漢字構形資料庫的建置與應用，漢字與全球化國際學術研討會，台北

楊允言，2005，台語文語料庫蒐集及語料庫爲本台語書面語音節詞頻統計，國科會結案報告