第三章　研究途徑與步驟

語料庫語言學的方法是對大規模的語料進行研究分析，可以減低少數文本對研究結果的影響；使用相關的電腦分析工具，如自動斷詞、詞頻統計等，可以有效率的處理語料，並以量化的方式呈現研究結果，是目前語言學界重要的研究方法。

本研究是採用語料庫語言學的方法，並參考 Biber 將文類分成學術類與小說類比較詞彙使用差異的研究（Biber、Conrad、Reppen 1998：43-44），將文類分成學術與非學術兩類探討台語詞彙的使用差異。語料來源主要是「白話字台語文網站」以及「台語文數位典藏資料庫（第二階段）」所蒐集的電子文本，使用「漢羅台語文斷詞系統」進行斷詞、詞頻統計，輔以相關的電腦軟體進行資料整理，分析台語詞彙在學術類與非學類文本的使用差異。

本研究分成三個主要步驟進行：第一個步驟：建立研究語料；第二個步驟：語料處理；第三個步驟：研究分析。研究流程如圖 1 所示。

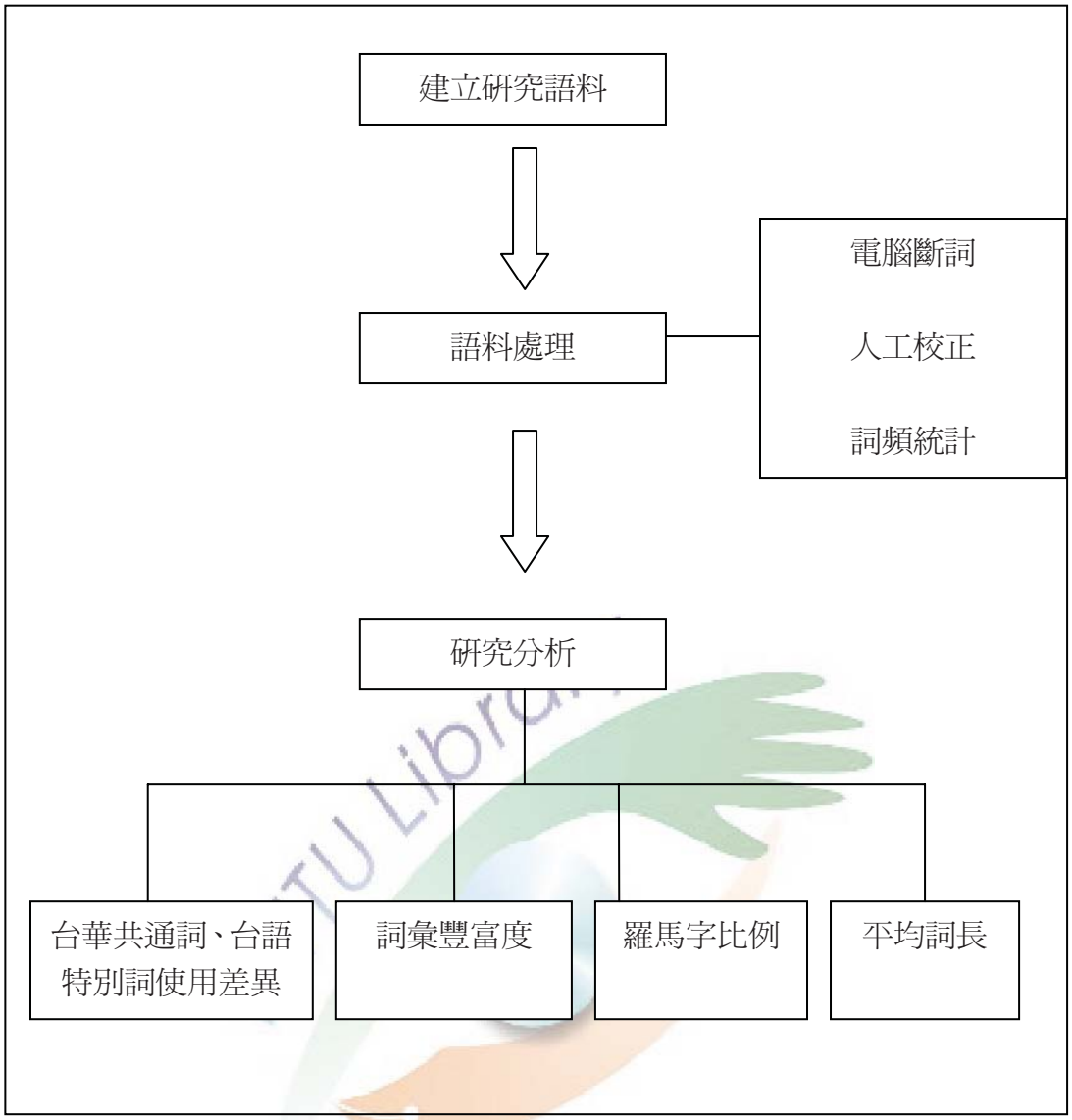

圖 2　研究流程圖

第一節　建立研究語料

本研究的語料以台語漢字書面語或漢羅書面語爲主，語料分成學術類與非學術類語料。學術類語料的來源是「白話字台語文網站」所蒐集的台語文研討會論文；非學術類的語料來源是「台語文數位典藏資料庫（第二階段）」所蒐集的文本，這些文本目前亦收錄於「台語文語料庫」。本研究的語料是經「白話字台語

文網站」與「台語文語料庫」管理者同意取得的，從語料中分別抽樣學術類與非學術類（包括小說類、散文類、劇本類）各約 10 萬音節，合計建立約 20 萬音節的研究語料。

本節分成語料簡介、語料限制、語料抽樣、抽樣結果四個部分說明研究語建立的過程。

一、語料簡介

本研究語料分成學術類與非學術類兩種，選擇小說、散文、劇本三種文本做爲非學術語料，以便與學術類語料做詞彙使用上的比較。以下分成兩部分簡介學術類語料以及非學術類語料：

（一）學術類語料

本研究的學術類語料來源是「白話字台語文網站」所蒐集的台語文學術研討會論文電子文本，經網站管理者同意，下載做爲學術研究之用，語料下載取得的時間點爲 2008/7/12，該網站網址爲：

http://iug.csie.dahan.edu.tw/giankiu/GTH/gth.asp。

學術類語料包括鄭良偉、張學謙、楊允言、李勤岸、方耀乾、呂興昌、蔣爲文、丁鳳珍等學者討論台灣羅馬字、語言人權、台語文學等議題所發表的論文，有全羅、全漢以及漢羅三種書寫形式，時間從西元 2002~2007 年，共計 6 個台語文學術研討會。

本研究以全漢與漢羅書寫形式的文本爲主，全羅書寫形式的文本不在本研究範圍。刪除全羅書寫形式以及僅收錄題目或摘要的論文，取得的全文論文有 75 篇，音節數約有 885,454 音節（音節數的計算是採用 Microsoft Word 工具選項

中的字數統計功能得到的數據）。

6 個台語文學術研討會，共計 75 篇全文論文的語料概述如下表：

表 12　學術類語料表

年份	研討會名稱	篇數	音節數
2002	台灣羅馬字教學 kap 研究國際學術研討會	11	123,320
2004	台灣羅馬字國際研討會	14	144,100
2004	語言人權與語言復振學術研討會	9	69,350
2005	台語文學學術研討會	11	150,046
2006	台灣羅馬字國際學術研討會	16	190,528
2007	台語文學學術研討會	14	208,110
		計 75	885,454

（二）非學術類語料

非學術類的語料來源是「台語文數位典藏資料庫（第二階段）」所蒐集的文本，這些文本亦收錄在「台語文語料庫」，本研究的語料是經由「台語文語料庫」管理者同意取得的，取得的時間點爲 2008/7/12。

「台語文數位典藏資料庫（第二階段）」所蒐集的文本包括小說、散文、劇本以及詩四種文本，以全羅與漢羅對照的方式書寫，年代從西元 1885~2006 年。作者從早期的巴克禮、賴仁聲、鄭溪泮、蔡培火，到近期的陳雷、陳明仁、李勤岸等數百位作者，作品內容涵蓋宗教、時政、生活雜記、笑話等題材。

本研究以漢羅書寫形式的文本爲主，全羅書寫形式的文本不在本研究範圍。詩的表達方式最不接近口語，與小說、散文、劇本三類的差異較大，因此不列入研究的語料，本研究僅選擇小說、散文、劇本三類文本做爲非學術語料，用來和學術類語料做詞彙使用差異的對照比較。

非學術語料扣掉詩以及全羅書寫形式的文本，得到小說、散文、劇本三類文本共計 1,560 篇，音節數有 2,452,075 音節。語料概述如下表：

表 13　非學術類語料表

類別	年代	篇數	音節數	比例
小說類	西元 1890~2006 年	386	1,051,375	42.88%
散文類	西元 1885~2006 年	1,125	1,264,609	51.57%
劇本類	西元 1924~2004 年	49	136,091	5.55%
		計 1,560	2,452,075	100.00%

說明：小說類、散文類、劇本類比例計算至小數點以下第二位四捨五入。

二、語料限制

從表 12、表 13 可知，本研究取得的學術類與非學術類語料分佈並不平均，歸納有以下幾點限制：

（一）語料年代：學術類語料的年代集中在西元 2000 年以後，非學術類語料的年代從西元 1885~2006 年，橫跨三個世紀。

（二）語料數量：學術類語料有 80 多萬音節，非學術類語料有 240 多萬音

節，非學術類音節的數量大約是學術類的三倍。

（三）文本種類：非學術類文本僅收錄小說、散文、劇本三類，未能涵蓋其他文類。

（四）文本數量：學術類僅有 75 篇，非學術類有 1,560 篇，非學術類的文本數量大約是學術類的 20 倍左右，相差懸殊。

因爲上述語料的限制，本研究在語料抽樣上僅以語料音節總數、單一文本音節數、文本年代、不同作者爲抽樣主要考量，無法兼顧兩種文類的隨機抽樣；非學術類文本亦僅有小說、散文、劇本等三類，未能涵蓋其他文類，無法呈現非學術類台語詞彙使用的全貌。以上語料的限制與問題，期待日後建立大型公開的台語書面語語料庫後能夠有進一步的解決途徑。

三、語料抽樣

研究者根據所獲得的語料以及語料限制，擬定本研究的語料抽樣原則與抽樣步驟，說明如下：

（一）抽樣原則

1.學術類與非學術類語料各抽樣約 10 萬音節數做爲研究語料。

2.非學術類語料抽樣以西元 2000 年後的文本爲原則，抽樣不足的部分依年代往前抽樣。

3.同一文類同一作者抽樣以一篇文本爲限。

4.同一文本超過兩位（含兩位）以上作者，以第一位作者爲抽樣目標。

5.每篇文本抽樣最多以 5,000 音節左右爲原則。

6.第 5,000 音節該段全部取樣，以保留語意之完整。

（二）抽樣步驟

學術類與非學術類實際的抽樣步驟說明如下：

1.學術類

（1）檢視文本內容：刪除外語（例如：日語）比例過重和以華語語法夾雜少量台語詞彙（例如： ê、个）撰寫的文本。

（2）刪除篇名、作者、任職單位、圖表、參考書目、附表、附記、附錄、註腳、謝詞等部分。

（3）刪除連續 50 音節（含 50 音節）以上非台語詞彙之語句。

（4）摘要、序論、結論等具代表性章節先行抽樣。

（5）若無摘要，直接從正文抽樣；若無註明序論、結論之文本，則以正文第一段做爲序論，最後一段做爲結論。

（6）若摘要、序論、結論未滿 5,000 音節，由正文第二章節（或第二段）依序往後抽樣，至第 5,000 音節該段落爲止。

（7）扣除步驟 2，不足 5,000 音節的文本整篇抽樣。

2.非學術類

（1）調整語料比例：從「台語文語料庫」獲得的語料中，小說佔 42.88%，散文佔 51.57%，劇本佔 5.55%。若依三類文本比例抽樣，劇本僅約佔 5%，代表性略嫌不足，因此調整比例爲：小說約佔 45%，散文約佔 45%，劇本約佔 10%。

（2）刪除篇名、作者、出處、日期等部分。

（3）由正文第一段依序往後取樣至第 5,000 音節該段落爲止。

（4）扣除步驟 2，不足 5,000 音節的文本整篇抽樣。

四、抽樣結果

根據抽樣原則與抽樣步驟實施操作後，得到學術類文本 21 篇，101,349 音節；非學術類文本 67 篇，102,335 音節；共計 88 篇文本，203,684 音節，語料抽樣結果如表 14、表 15：

表 14　學術類語料抽樣表

年份	研討會名稱	選取篇數	音節數
2002	台灣羅馬字教學 kap 研究國際學術研討會	5	25,374
2004	台灣羅馬字國際研討會	4	17,666
2004	語言人權與語言復振學術研討會	2	8,724
2005	台語文學學術研討會	7	34,969
2006	台灣羅馬字國際學術研討會	2	10,074
2007	台語文學學術研討會	1	4,542
		計 21	101,349

表 15　非學術類語料抽樣表

類別	年代	選取篇數	音節數
小說類	西元 1991～2006 年	16	48,008
散文類	西元 2001～2006 年	49	46,286
劇本類	西元 1966、2004 年	2	8,041
		計 67	102,335

第二節　語料處理

抽樣所得之語料尚需要經過電腦斷詞、人工校正以及詞頻統計三個處理程序，之後整理出學術類與非學術類兩個詞頻統計表，方可做爲分析台語詞彙的資料。以下分成三個部分說明語料處理過程。

一、電腦斷詞

本研究電腦斷詞採用「漢羅台語文斷詞系統」[19]，此系統是以「台文華文線上辭典」六萬多筆詞條做爲詞庫，以「逆向最大比對法」找出詞庫裡有的詞進行斷詞。但是目前台語尚缺乏完善的分詞規範，「漢羅台語文斷詞系統」亦非以分詞規範進行斷詞，而且辭典也不可能收錄所有的詞條，因此電腦所斷之詞彙並非

[19] 漢羅台語文斷詞系統：http://poj.likulaw.info/hanlo_hunsu.php。2008/8/22。

完成正確，有其限制。以下分成三點說明「逆向最大比對法」、電腦斷詞步驟以及電腦斷詞限制：

（一）逆向最大比對法的斷詞步驟

電腦針對輸入的句子，從句尾往句首比對電腦詞庫裡有的語詞，先比對最長的音節，再依序比對到最短的音節，與詞庫語詞相符的則判斷爲詞彙。茲以「這个囡仔真古錐」爲例說明逆向最大比對法的斷詞步驟。

1.由句尾往句首數 4 個字：「仔真古錐」。

2.比對電腦詞庫找不到「仔真古錐」，因此不是詞彙。

3.由句尾往句首數 3 個字：「真古錐」。

4.比對電腦詞庫找不到「真古錐」，因此不是詞彙。

5.由句尾往句首數 2 個字：「古錐」。

6.比對電腦詞庫找到「古錐」，「古錐」斷詞爲詞彙。

7.剩餘的字「這个囡仔真」，回到步驟一、步驟二...依序操作至整句話斷詞完畢。

8.將斷詞完畢的詞彙順序顛倒。例如：上述的例子斷詞完後是「古錐、真、囡仔、這个」，順序顛倒後「這个、囡仔、真、古錐」。

補充說明：僅剩下一個字時，不論詞庫有無該字，皆斷詞爲詞彙。

（二）電腦斷詞步驟

1.將語料逐篇輸入「漢羅台語文斷詞系統」，由電腦斷詞。

2.電腦檢索「台文華文線上辭典」詞庫裡六萬多筆詞條。

3.詞庫裡有的詞用［　］表示；詞庫裡沒有的詞用｛　｝表示。例如：[教學]表示詞庫有，電腦能夠斷詞；{的}表示詞庫沒有，電腦無法斷詞。

4.電腦無法斷詞的部分則由人工校正。

（三）電腦斷詞的限制

「漢羅台語文斷詞系統」是以「台文華文線上辭典」中的詞條做爲斷詞依據，因爲台語書寫形式尚未規範，而且「台文華文線上辭典」不可能收錄所有的詞條與書寫形式；再者目前亦缺乏一套完善的台語分詞準則，「漢羅台語文斷詞系統」也不是以分詞準則進行斷詞，因此斷詞結果不一定完全正確，此爲電腦斷詞的限制。

二、人工校正斷詞

因爲電腦斷詞有其限制，因此需要以人工校正的方式輔助「漢羅台語文斷詞系統」之不足。以下分成人工校正斷詞的限制、電腦斷詞錯誤的詞彙、人工校正斷詞的原則、人工校正斷詞的步驟四個部分說明：

（一）人工校正斷詞的限制

本研究語料經過電腦斷詞後共有 15,000 多個詞型，礙於時間與人力的考量，無法一一校對電腦斷詞詞彙的正確與否。因此「台文華文線上辭典」收錄的詞條經電腦斷詞爲詞彙後，即不再以人工進行校正。研究者僅能針對電腦斷詞錯誤的前後三個詞彙予以人工校正（請參考下面人工校正斷詞步驟 1），此爲人工校正斷詞的限制之一。

目前台語缺乏一套完善的分詞準則，在人工校正斷詞時，研究者主要以詞意完整爲主要考量，在人工校正時很難避免主觀的成分，亦無法兼顧語法、詞性分類等問題，此爲本研究人工校正斷詞的限制之二。

（二）電腦斷詞錯誤的詞彙

「台文華文線上辭典」沒有收錄的詞條以及書寫形式不一致，「漢羅台語文

斷詞系統」無法比對造成斷詞錯誤，這些詞彙可分成以下兩種類型：

1.「台文華文線上辭典」沒有收錄的詞條（包含書寫不一致的詞彙），例如：暗 bong-bong。

2.詞彙中插：兩個詞彙中間插入「-」符號，例如：{1- ê }， [出]-{去}。

（三）人工校正斷詞的原則

根據人工校正斷詞的限制以及電腦無法斷詞的詞彙，擬定以下幾項人工校正斷詞原則：

1.「漢羅台語文斷詞系統」已完成斷詞的詞彙，不再進行人工校正。

2.僅針對「漢羅台語文斷詞系統」斷詞錯誤的台語詞彙進行人工校正。

3.非台語詞彙部分（例如：英語或日語），詞彙之間以空白分隔，沒有斷詞的問題，故不須人工校正。

4.人工校正以保留詞意完整爲原則。

（四）人工校正斷詞步驟

依據上述的人工校正斷詞原則，擬定以下幾點人工校正斷詞步驟：

1.以電腦斷詞錯誤的詞彙爲中心，將前後三個詞彙（如少於三個詞彙，取到前後標點符號爲止）視爲未完成斷詞部分，納入人工校正範圍，保持{　}前後詞意之完整。例如：『[　][　][　]{　}[　][　][　]』，或『，[　][　]{　}[　][　][　]。』，範例『，[將][課]{程的}[目標][設定][做][20]』。

2.人工校正{　}前後三個詞彙，詞意完整者斷詞成爲一個詞彙。

[20] 擷取自「台語文語料庫」學術語料：梁淑慧。2002。〈「幼兒台語班」的教學實務 kah 成果〉。《2002 台灣羅馬字教學 kap 研究國際學術研討會》。

3.將人工斷詞後的台語詞彙加入「漢羅台語文斷詞系統」使用者詞庫，改進電腦斷詞結果。

4.再執行一次電腦斷詞。

5.將文本再檢視一次，若有遺漏沒有斷詞的部分，再執行步驟 1~步驟 4。

表 16 爲人工斷詞處理實例說明：

表 16　人工斷詞實例表

例句	電腦斷詞	人工斷詞	備註
hit 個 gín-á 豆油 moh--leh[21]	[hit][個][gín-á][豆油][moh]--[leh]	*	人工斷詞原則 1
將課程的目標設定做[22]	[將][課]{程的}[目標][設定][做]	[將][課程] [的] [目標] [設定][做]	人工斷詞原則 2、4
上尾 1-pái 喘氣[23]	[上]{尾 1--pái }[喘氣]	[上尾] [1] -[-pái]	人工斷詞原則 2、4

[21] 擷取自「台語文語料庫」非學術語料：Abon。2000。〈魚肉〉。

[22] 擷取自「台語文語料庫」學術語料：梁淑慧。2002。〈「幼兒台語班」的教學實務 kah 成果〉。《2002 台灣羅馬字教學 kap 研究國際學術研討會》。

[23] 擷取自「台語文語料庫」非學術語料：Voyu Taokara 劉。2006。〈目睭〉。

表 16　人工斷詞實例表

		[喘氣]	
仙人飛--出-去[24]	[仙人][飛] --[出]-{去}	[仙人][飛] --[出-去]	人工斷詞原則 2、4
chhōa 300 ê 兵 á[25]	[chhōa] [300] [ê]{兵 á}	[chhōa] [300] [ê] [兵 á]	人工斷詞原則 2、4
Beh ài 伊乖乖 á 做頭前[26]	[Beh] [ài] [伊]{乖乖 á}[做][頭前]	[Beh] [ài][伊] [乖乖] [á][做] [頭前]	人工斷詞原則 2、4
鄭良偉[27]	{鄭}[良]{偉}	[鄭良偉]	人工斷詞原則 2、4
E. G..Lewis [28]	[E]. {G}. {Lewis}	*	人工斷詞原則 3

下表爲使用者詞庫舉例說明：

[24] 擷取自「台語文語料庫」非學術語料：Abon。2000。〈魚肉〉。

[25] 擷取自「台語文語料庫」非學術語料：Voyu Taokara 劉。2006。〈目睭〉。

[26] 擷取自「台語文語料庫」非學術語料：Voyu Taokara 劉。2006。〈目睭〉。

[27] 擷取自「台語文語料庫」學術語料：梁淑慧。2002。〈「幼兒台語班」的教學實務 kah 成果〉。《2002 台灣羅馬字教學 kap 研究國際學術研討會》。

[28] 擷取自「台語文語料庫」學術語料：張學謙。2002。〈東是東，西是西，永遠 bē 相 tú？台灣人對台語文字 ê 態度研究〉。《2002 台灣羅馬字教學 kap 研究國際學術研討會》。

表 17　使用者詞庫表（舉例）

使用者詞庫	使用者詞庫	使用者詞庫	使用者詞庫
黑板	上尾	干乾	攏
kōaⁿ-kōaⁿ-kōaⁿ	tò-tńg--來	漢羅	流程
摃搥仔	到	態度	旗 á
韻母	今 á 日	多元	黃 gīm-gīm
咁閣	án-chóaⁿ	現主時	光 sih-sih

三、詞頻統計

經過電腦斷詞和人工校正斷詞後，由「漢羅台語文斷詞系統」輸出詞頻統計表，學術類 21 篇、非學術類 67 篇，共計 88 份分篇詞頻統計表。接著使用 Microsoft Excel 軟體和楊允言撰寫的程式，將 21 篇學術類以及 67 篇非學術類分篇詞頻統計表，分別合併整理出學術類詞頻統計表以及非學術類詞頻統表各一份。如表 18、表 19：

表 18　學術類詞頻統計表（範例）

編號	詞型	詞次	比例	比例總合
1	ê	2,817	4.8521%	4.8521%
2	的	1,235	2.1272%	6.9793%

表 18　學術類詞頻統計表（範例）

3	[NUMBER][29]	1,209	2.0824%	9.0618%
4	是	839	1.4451%	10.5069%
5	有	620	1.0679%	11.5748%
6,853	喚[30]	1	0.0017%	99.9931%
6,854	出世	1	0.0017%	99.9948%
6,855	步數	1	0.0017%	99.9966%
6,856	泪泪	1	0.0017%	99.9983%
6,857	靭性	1	0.0017%	100.0000%

表 19　非學術類詞頻統計表（範例）

編號	詞型	詞次	比例	比例總合
1	ê	4,082	5.2364%	5.2364%
2	[NUMBER]	1,251	1.6048%	6.8411%
3	是	1,218	1.5624%	8.4036%
4	我	1,129	1.4483%	9.8518%

[29] [NUMBER]是數字 1,2,3……，因爲可以無限衍生，影響統計結果，因此全部歸爲一個詞型[NUMBER]計算。

[30] 學術類詞頻編號 6,853「喚」、編號 6,854「出」、編號 6,855「步」是打字造成的錯誤，屬於雜訊的一種，在做分析時會考慮將低頻詞拿掉，避免雜訊對研究分析造成干擾。

表 19　非學術類詞頻統計表（範例）

5	講	1,019	1.3072%	11.1590%
9,079	靈堂	1	0.0013%	99.9949%
9,080	觀看	1	0.0013%	99.9962%
9,081	觀音山	1	0.0013%	99.9974%
9,082	讚嘆	1	0.0013%	99.9987%
9,083	鑼	1	0.0013%	100.0000%

第三節　研究分析

使用學術類與非學術類詞頻統計表進行詞彙分析，分析工作分成四個部分：第一部分分析台華共通詞在學術類與非學術類的使用差異；第二部分分析詞彙豐富度在學術類與非學術類的差異；第三部分分析台語羅馬字詞彙在學術類與非學術類的使用情形；第四部分分析台語平均詞長在學術類與非學術類的差異。

因爲低頻詞[31]中包含電腦雜訊，會影響研究結果，分析統計時會以不同覆蓋率計算的方式設法排除電腦雜訊的干擾，以求研究結果之精確。

[31] 低頻詞就是在語料中使用頻率比較低的詞彙。低頻詞有兩種：一是冷僻的詞彙，本身使用率就比較低，例如：「蹔箴」或專有名詞等；另一種是電腦雜訊（打字錯誤），例如：「出世」的「出」，是由兩個「山」組合而成的，並不是「出去」的「出」字，因此電腦會將「出世」與錯誤的「出世」分成兩個不同的詞彙，錯誤的「出世」詞頻自然就比較低，其他還有「唍」、「步數」等，無法一一列舉。

一、台華共通詞的使用差異分析

這一部分是分析台華共通詞在學術類與非學術類的使用情形，步驟如下：

（一）將學術類與非學術類詞頻統計表中所有詞彙（詞型）分別與詞庫小組「中文詞庫（八萬詞目）[32]」逐一比對，字形、字義皆相同者視爲台華共通詞，字形、字意其中有一個不同者視爲台語特別詞。

（二）使用 Microsoft Excel 軟體的函數 vlookup 功能，利用電腦程式與詞庫小組「中文詞庫（八萬詞目）」逐一比對，將學術類 6,857 個詞彙（詞型），非學術類 9,083 個詞彙（詞型），分別區分爲台華共通詞和台語特別詞。

（三）人工校對電腦分類後的台華共通詞和台語特別詞，校對工作有四個部分：

1.台華共通詞轉爲台語特別詞：電腦比對結果爲台華共通詞，但實際上華語已不使用，或台語已改變詞意的詞彙，轉歸類爲台語特別詞。例如：上好、攏。

2.台語特別詞轉爲台華共通詞：電腦比對結果爲台語特別詞，但實際上華語仍在使用，且詞意與台語相同的詞彙，轉歸類爲台華共通詞。例如：台語、全部。

---

[32] 由中央研究院中文詞知識庫小組執行、研究，授權中華民國計算語言學學會發行，爲一包含八萬目詞的電子辭典。詞庫收的詞包含一般用詞、常用專有名詞、成語、慣用語、常用派生詞、異體詞、合併詞以及少數特殊領域用語和古漢語詞語。每個詞項包含的訊息有: 注音、頻率、詞類、名詞語義分類等。

資料來源：中華民國計算語言學學會：http://www.aclclp.org.tw/use_ced_c.php。2008/9/6。

3.無法明確歸類的詞彙：於統計詞型、詞次時，台華共通詞與台語特別詞以一半計算（即詞型、詞次乘以 1/2）。例如：「足」，台語、華語皆有「腳」的意思，但台語亦有副詞「很」的意思。

（四）人工校對範圍：學術類覆蓋率（比例總合）達 80%且詞次達 8 次（含 8 次）以上之詞彙，計 1,250 詞；非學術類覆蓋率（比例總合）達 80%且詞次達 7 次（含 7 次）以上之詞彙，計 1,657 詞。

（五）計算學術類與非學術類覆蓋率 100%、覆蓋率 80%的台華共通詞與台語特別詞的比例[33]。

（六）結果分析。

二、詞彙豐富度分析

本研究詞彙豐富度的計算方式爲：

$$\text{詞彙豐富度} = \frac{\text{詞型}}{\text{詞次}}$$

步驟如下：

（一）分別計算學術類與非學術類的詞型與詞次。

（二）詞型÷詞次即可得到學術類與非學術類的詞彙豐富度。

（三）分別計算學術類與非學術類覆蓋率 100%、覆蓋率 95%、覆蓋率 90

[33] 因爲覆蓋率 100%包含許多打字錯誤等電腦雜訊，會影響研究結果，本研究以不同覆蓋率的計算方式，降低電腦雜訊的影響程度。

％、覆蓋率 85％、覆蓋率 80％的詞彙豐富度[34]。

（四）結果分析。

三、台語羅馬字詞彙使用分析

這一部分探討學術類與非學術類台語羅馬字詞彙的使用情形，本研究的台語羅馬字詞彙包括全羅詞彙和漢羅詞彙，步驟如下：

（一）使用 Microsoft Excel 軟體的函數 code、left、right 功能，利用電腦程式將台語羅馬字詞彙與台語全漢字詞彙做初步分類。

（二）人工校對：挑出電腦錯誤歸類爲台語羅馬字詞彙的部分。例如：英文詞 power、solution，數詞「3-6」，以及電腦無法分辨的漢字：團圓的「圓」等。

（三）分別計算學術類與非學術類台語羅馬字在詞型、詞次所佔的比例。

（四）結果分析。

四、台語平均詞長分析

這一部分探討學術類與非學術類語料的平均詞長，計算方式爲：音節數÷詞型數。步驟如下：

（一）分別計算學術類與非學術類的音節總數以及詞型總數。

（二）音節（syllable tokens）總數÷詞次總數即可得到學術類與非學術類

[34] 同註 33。

的平均詞長。

（三）分別計算學術類與非學術類覆蓋率 100%、覆蓋率 80%的台語平均詞長[35]。

（四）結果分析。

[35] 同註 33。


NTTU Library