國立台東大學華語文學系

台灣語文教師碩士班碩士論文

指導教授：楊允言 先生

台語學術類和非學術類的詞彙使用比較

研究生：曾國榕撰

中華民國九十七年九月

國立台東大學
學位論文考試委員審定書

系所別：華語文學系台灣語文教師碩士班

本班 曾國榕 君

所提之論文 台語學術類和非學術類的詞彙使用比較

業經本委員會通過合於 ☑ 碩士學位論文 ☐ 博士學位論文 條件

論文學位考試委員會：丁鳳珍

（學位考試委員會主席）

張學謙

楊允言

（指導教授）

論文學位考試日期：2008年 9月 20日

國立台東大學

附註：1. 一式二份經學位考試委員會簽後，送交系所辦公室及註冊組或進修部存查。

2. 本表為日夜學制通用，請依個人學制分送教務處或進修部辦理。


NTTU Library

博碩士論文授權書

本授權書所授權之論文為本人在 國立臺東大學 華語文學 系(所)
台灣語文教師碩士班 組 97 學年度第 暑 學期取得 碩 士學位之論文。
論文名稱：台語學術類和非學術類的詞彙使用比較。

本人具有著作財產權之論文全文資料，授權予下列單位：

同意	不同意	單　位
☑	☐	國家圖書館
☑	☐	本人畢業學校圖書館
☑	☐	與本人畢業學校圖書館簽訂合作協議之資料庫業者

得不限地域、時間與次數以微縮、光碟或其他各種數位化方式重製後散布發行或上載網站，藉由網路傳輸，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

☑同意 ☐不同意 本人畢業學校圖書館基於學術傳播之目的，在上述範圍內得再授權第三人進行資料重製。

本論文為本人向經濟部智慧財產局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：_____，請將全文資料延後半年再公開。

公開時程

立即公開	一年後公開	二年後公開	三年後公開
✓			

上述授權內容均無須訂立讓與及授權契約書。依本授權之發行權為非專屬性發行權利。依本授權所為之收錄、重製、發行及學術研發利用均為無償。上述同意與不同意之欄位若未勾選，本人同意視同授權。

指導教授姓名：楊允言　(親筆簽名)

研究生簽名：曾國發　(親筆正楷)

學　　號：3895002　(務必填寫)

日　　期：中華民國 97 年 11 月 29 日

1.本授權書 (得自 http: //www.lib.nttu.edu.tw/theses/ 下載) 請以黑筆撰寫並影印裝訂於書名頁之次頁。

2.依據 91 學年度第一學期一次教務會議決議:研究生畢業論文「至少需授權學校圖書館數位化，並至遲於三年後上載網路供各界使用及校內瀏覽。」

授權書版本:2008/05/29


NTTU Library

誌謝辭

寫作論文是一項與意志力搏鬥艱辛的過程，一路走來老師的指引、同學的鼓勵以及家人的支持，使我能夠走完全程，要感謝的人很多。

首先要感謝的是我的指導教授楊允言老師，我們的討論時間大都是他剛上完課身體正疲備的時候，他還是不厭其煩的設法解決我論文寫作上的問題。他常常和我們這組的同學講的兩句話就是「吓免煩惱啦」、「不要太累、身体吓通拍壞去」，但是我覺得他想的、做的都比我們多。從他身上，學到的不只是做學問的方法，更是做學問的態度。

再來要感謝的是台文所的創辦人，也是我的授課老師---張學謙先生，使我了解語言現象有很多是值得研究的（因爲張老師上課常常講的一句話就是「這嘛真值得研究」）。除了張老師，還要感謝許秀霞老師提供研究室，讓我有一個可以安靜寫論文的地方，以及這段日子曾教過我的教授，傅濟功老師、林雅玲老師、林淑慧老師；還有感謝審查這本論文的張學謙老師、丁鳳珍老師，提供我很多寶貴的意見。

接著要感謝的是我的岳父陳嘉興先生、岳母莊純美女士，在暑假期間幫忙照顧我五個月大的女兒，讓我可以專心撰寫論文；還要感謝我的老婆慧蓮，在我修改論文期間操持家務，照顧女兒，讓我可以無後顧之憂；還有我的女兒子芩（古錐），能夠乖乖的和外公、外婆渡過一個爸媽不在身邊的暑假。

同時也要感謝同組的淑玲總是扮演火車頭帶領大家向前；欣珉常給予我適時的幫助；還有秀俐、廷運一路走來的相互鼓勵。以及相伴三年的同學，有緣一起渡過這段人生中難忘的旅程。

最後要感謝成立「台語文數位典藏資料庫」的國家台灣文學館，負責管理「台語文語料庫」的台灣羅馬字協會，和「白話字台語文網站」的管理者，讓我免費使用其中的語料，順利完成論文，以及所有曾經協助過我完成本論文的人士。


NTTU Library

台語學術類和非學術類的詞彙使用比較

作者：曾國榕

國立台東大學台灣語文教師碩士班

摘　要

本論文的目的在於比較學術和非學術的台語書面語的風格差異。研究採用語料庫語言學的方法進行資料的收集和分析。本論文探討學術和非學術這兩種不同文體在語詞選用方面的差異，包括台華共通詞、詞彙豐富度、羅馬字詞彙、平均詞長這四方面的比較。

研究發現：1.台華共通詞的使用比例學術類高於非學術類；2.詞彙豐富度學術類略高於非學術類 3.台語羅馬字詞彙的使用比例非學術類高於學術類；4.台語詞彙平均長度學術類高於非學術類。顯示台語借用華語詞彙是一種趨勢，愈正式的文體借用的比例愈高。

關鍵詞：語料庫語言學、台華共通詞、詞彙豐富度、計算語言學、台語文


NTTU Library

Comparison with the Usage of Academic and Non-academic Taiwanese Words

Chan Kok-iong

Abstract

The purpose of this thesis was to compare academic and non-academic written Taiwanese style. Corpus linguistic method was used to collect and analyze written Taiwanese data. Specifically, this study focus of style related to word usage, such as Taiwanese-Chinese common word, word richness, the use of Roman script and the average length of Taiwanese word.

The results of this study can be summarized as follows : 1. Academic style is higher than non-acacdemic style on the rate of Taiwanese-Mandarin common words. 2. Academic style is slightly higher than non-academic style on word richness. 3. Non-academic style is higher than academic style on the use of Taiwainese Roman script. 4. Academic style is higher than non-academic style on the average length of Taiwanese word. The higher rate of borrowed words from Mandarin has been founded in more formal genre.

Keywords：Corpus linguistics, Taiwanese-Chinese common word, Word richness, Computational linguistics, Written Taiwanese.

目錄

表目錄

圖目錄


NTTU Library

第一章　緒論

第一節　研究背景與動機

一、研究背景

台語過去長期遭受到政治手段的打壓，以及書寫系統尚未完全標準化，著作流通不廣，以致於造成一般民眾觀念上的偏差，認爲台語是沒有文字的次等語言。就語言地位而言，華語是強勢語言，台語是弱勢語言；就語言的功能來看，華語屬於高階語言，主要用於正式的場合，例如：議會、法院、學校...等；台語屬於低階語言，主要用於非正式場合，例如：家庭、私人聚會、民俗文學...等，有不同的功能，但是現在華語已經侵入台語家庭，台灣漸漸走向雙語但非雙言的社會（黃宣範 2004：14-15）。借用華語詞彙已成爲台語寫作的趨勢。

台語目前至少有三種主要的書面語，第一種是全漢文字，目前已知最早的文獻是明嘉靖 45 年（1566）的《荔鏡記》戲文，主要是記錄文言文，至今有四百多年的歷史；第二種是台語羅馬文字，當初傳入台灣是教會爲了方便傳教使用，又叫做「白話字」，最早的文獻一般認爲是 1837 年 W. H. Medhurst 編纂的《福建方言字典》，至今也有一百多年的歷史（張裕宏 2001）；第三種是漢羅文字，1964 年王育德提出的理論，1980 年代後期，由鄭良偉引進台灣，並且用於實際寫作（張學謙 2003），目前已有愈來愈多人使用。

台語書面語的發展，鄭良偉（1990：5）認爲：「已經從第一期的口傳文學

筆錄，第二期的詩歌創作，進入第三期的散文試寫初期」。小說創作近年來也蓬勃發展，宋澤萊、陳明仁、陳雷等人皆有豐富的作品；而且已有不少學者用於撰寫正式的學術著作，例如；鄭良偉在 1989、1990 年以漢羅文字先後完成《走向標準化的台灣話文》以及《演變中的台灣社會語言－多語社會及雙語教育》兩本台語研究專書；張學謙、呂興昌、楊允言、蔣爲文、丁鳳珍、廖瑞銘、何信翰等學者相繼發表多篇學術論文；2000 年以後更陸續舉辦多場學術研討會，例如：2002 年「台灣羅馬字教學 kap 研究國際學術研討會」、2004 年「語言人權與語言復振學術研討會」、「台灣羅馬字國際研討會」……以及 2007 年「台語文學學術研討會」等；方耀乾、李勤岸等人更是學者兼作家，著作豐富多樣。台語書面語的使用層面愈來愈廣泛，已由非正式的民間文學初步發展至正式的學術著作，寫作的人也愈來愈多。

「語料庫語言學」是結合語言學與電腦，研究自然語言的新興學科，在語言學界已是重要的發展。語料庫研究的濫觴一般認爲是美國布朗語料庫（Brown Corpus）（黃居仁 1997）；在台灣，華語語料庫的建立與研究起步比較早，中央研究院建立的「中央研究院平衡語料庫」，簡稱「中研院平衡語料庫」（Sinica Corpus），是世界上第一個有完整詞類標記的漢語平衡語料庫[1]，並且是世界華語語料庫研究領先的中心（黃居仁 1997）。

台語語料庫的起步比較晚，建立的過程中並沒有得到太多的支援，在有志之士的努力下，還是克服資源缺乏的問題，建立數個公開或尚未公開的台語語料庫。「台語文數位典藏資料庫（第二階段）」[2]是國家台灣文學館成立，少數公開

[1] 中文詞知識庫小組：http://ckip.iis.sinica.edu.tw/CKIP/20corpus.htm。2008/9/7。

[2] 台語文數位典藏資料庫（第二階段）：http://iug.csie.dahan.edu.tw/nmtl/dadwt/pbk.asp。2008/8/23。

上網的台語文學語料庫；「台語文語料庫」[3]是楊允言、張學謙共同主持建立的書面語語料庫，目前雖然尚未完全公開，但是已提供部分功能與資料供研究者使用；「台灣兒童語料庫」和「閩南語口語語料庫」是蔡素娟主持建立的台語口語語料庫，「台灣兒童語料庫」已經收錄於國家數位典藏（蔡素娟 2007）。除了以上較具規模的語料庫外，其他多半是個人基於研究需要所建立小型語庫料。

台語雖然至少有三種書面語，並且有豐富的語料，但是因爲政治以及語言本身的因素，是一個弱勢、低階語言，受到強勢華語的影響，借用華語詞彙寫作已是一種趨勢。電腦問世後，語料庫語言學成爲語言研究另一條重要途徑，語料庫是其中重要的基礎建設，就一個弱勢語言來說，更是語言保存的重要管道。「資訊時代的殘酷事實是不在電腦資訊傳遞上使用的語言可能被加速淘汰」（黃居仁 1997：261）。國內目目前也已成立數個公開或未公開的台語語料庫，做爲語言保存以及學術研究之用，以上是本研究的背景。

二、研究動機

「子芩」是研究者的女兒，今年二月出生，她的台語小名叫「古錐」（kó͘-chui）------取諧音就叫「曾古錐」（真古錐）（chin kó͘-chui），方便古錐的阿嬤叫她。研究者當初在思索「子芩」台語小名時，發現台語詞彙的精妙是值得進一步了解的學問。

如何以「有音無字」尚未規範的台語撰寫正式學術論文，在選詞上與非正式的文本有何差異，也是一個值得深入研究的問題。

[3]「白話字台語文網站」---台語文語料庫建立蒐集計畫：http://iug.csie.dahan.edu.tw/TG/guliaukhou/。2008/8/22。

第二節　問題意識

台語和華語都是屬於漢語的一個支系，詞彙共通性高，相互借用容易；台語因爲缺乏一套規範的書面語，屬於弱勢、低階的語言，寫作上借用華語詞彙已是時勢所趨。不同文體因爲訴求的重點不同，借用華語詞彙的情形應該會有所差異。學術類著作比較正式，詞彙選用要求客觀、精確，非學術著作比較非正式，詞彙選用著重作者情感的表達與觀眾的共鳴。本研究即是嘗試探討台語詞彙在學術與非學術兩類不同文體的使用差異。

台語詞彙的特性可用許多不同的面向來探討包括：合音詞、外來詞、文白異讀、台華共通詞、羅馬字詞彙、平均詞長等，本研究擬就台華共通詞、羅馬字詞彙、平均詞長等三項進行討論，並比較詞彙豐富度在學術與非學術文本的差異。研究者擬定以下四個問題進行探討：

一、台華共通詞在學術類與非學術類書面語有何使用差異？

二、台語詞彙豐富度在學術類與非學術類書面語有何差異？

三、台語羅馬字詞彙在學術類與非學術類書面語有何使用差異？

四、台語平均詞長在學術類與非學術類書面語有何差異？

第三節　研究範圍

本研究語料主要以全漢字書面語或漢羅混用書面語爲主，語料分成學術與非學術兩類；學術類主要是台語研討會論文，非學術類則包含：小說、散文、劇本三類。學術類語料的來源是「白話字台語文網站」所蒐集之台語文學術研討會論文；非學術類語料的來源是「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」。本研究語料經「白話字台語文網

站」與「台語文語料庫」管理者同意後取得做爲研究之用，研究者從取得的語料中分別抽樣學術與非學術語料各約 10 萬音節，合計約 20 萬音節語料。學術與非學術類語料的簡介說明請參閱本研究第三章第一節建立研究語料，文本篇名、作者、年代等詳目，請參閱本研究附錄一與附錄二。

第四節　名詞解釋

一、台華共通詞

台華共通詞是指台語和華語共通的詞彙而言，本研究定義爲：和華語詞形相同且詞義相近的台語詞彙爲台華共通詞。詳細說明請參見本研究第二章第三節台華共通詞的部分。

二、台語特別詞

台語特別詞是相對於台華共通詞而言，意思相同的詞彙使用和華語不同的書寫形式就是台語特別詞。本研究定義爲：和華語詞義相同但寫法不同的台語詞彙爲台語特別詞。詳細說明請參見本研究第二章第三節台語特別詞的部分。

三、逆向最大比對法（Backword Maximal Matching Algorithm）

電腦自動斷詞的一種方式，就是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。詳細的操作步驟請參考本研究第三章第二節的相關說明。

四、詞型(word types)、詞次(word tokens)

詞型是指文章中詞彙的類型而言；詞次是指文章中某個詞彙類型的出現頻率。例如：「這个囝仔真古錐」，詞型有「這个、囝仔、真、古錐」四個，詞次爲「這个：1 次」，「囝仔：1 次」，「真：1 次」，「古錐：1 次」，詞次共有 4 次。第二個例子：「千江有水千江月」，詞型有「千、江、有、水、月」五個；詞次爲「千：2 次」，「江：2 次」，「有：1 次」，「水：1 次」，「月：1 次」，詞次共有 7 次。

五、詞彙豐富度（word richness）

詞彙豐富度系指文本中詞彙的豐富程度，亦即作者所能掌握運用於寫作的詞彙。詞彙豐富度因作者、文體、年代等因素而有程度上的差異。本研究是參考（楊允言 2003）的計算方式：詞型÷詞次，得到的值愈高表示詞彙愈豐富，反之愈低。

六、覆蓋率

將詞彙出現的頻率由高到低排序，依序累積計算詞彙比例，即爲覆蓋率（表 1 中的比例總合）。假設有一份語料，共有 108 個詞型，400 個詞次，將此語料

中每個詞彙的詞頻由高到低排列，得到如表 1 的詞頻統計表。茲以表 1 爲例說明覆蓋率：

「ê 」的頻率是 5%，如果只計算「ê」的覆蓋率也是 5%；「的」的頻率是 2.5%，「 ê」加「的」的覆蓋率是 7.50%；「是」的頻率是 1%，「 ê」加「的」加「是」的覆蓋率是 8.50%；依序累加至編號 108「古錐」的覆蓋率是 100.00%。

表 1　詞頻統計表

編號	詞型	詞次	比例	比例總合
1	ê	20	5%	5.00%
2	的	10	2.5%	7.50%
3	是	4	1%	8.50%
4	有	4	1%	9.50%
105	出世	1	0.25%	99.25%
106	冷支支	1	0.25%	99.50%
107	稀微	1	0.25%	99.75%
108	古錐	1	0.25%	100.00%


NTTU Library

第二章　文獻探討

語料庫語言學是以電腦做爲工具研究自然語言現象的一門科學，目前已經是語言學界重要的發展，語料庫是其中重要的基礎建設。從 1950 年代開始各國已相繼投入資源成立各種語料庫，進行語言研究與開發應用。台語語料庫的建置起步比較晚，得到的資源有限，但是在台語有心人士共同努力下，目前已建立幾個公開或尚未公開的語料庫，例如：「台語文數位典藏資料庫」、「台語文語料庫」、「台灣兒童語料庫」和「閩南語口語語料庫」等。

台語詞彙的成份相當複雜，可以從許多面向探討分析。本研究僅就台語詞彙層、台華共通詞、詞彙豐富度、台語羅馬字詞彙以及台語平均詞長進行討論。

第一節　語料庫語言學

本節分成四點說明語料庫的定義、語料庫語言學的定義、中、英文語料庫簡介、語料庫的應用與研究。

一、語料庫的定義

語料庫顧名思義是指存放大量自然語言材料的倉庫，可以是書面語或是口語，以前以人工方式處理，現在以電子形式保存於電腦中，可作爲語言研究的基礎，廣泛用於語言研究和語言工程，現在所說的語料庫通常是指電腦語料庫而言（黃昌寧、李涓子 2002）。

語料庫是指可由機器辨讀的書面或口語抽樣文本的集合，可以多種不同形式

的語言訊息標註加工（Anthony McEnery, Richard Xiao, Yukio Tono 2006：345）。

語料庫裡的語料必須能夠永續使用，永續使用包含兩個層面：一是語料重複使用而不會耗損；其次是語料實質內容的永續性，亦即語料量夠大且足以代表語言本體，少量特殊的語料沒有永續使用的價值。紙本書籍、錄音磁帶容易毀損，人力能夠處理的資料量有限，因此，電腦可重複性、儲存記憶量大、運算快的特點在語料的永續使用性上扮演關鍵性的角色，所以，現在所說的「語料庫」和「機讀語料庫」基本是同義詞（黃居仁 1997）。「使用電腦儲存並處理語料，已成了『語料庫』基本定義的一部分」（Atkins et al. 1992；轉引自黃居仁 1997：258）。

二、語料庫語言學的定義

關於語料庫語言學，學者有以下看法：

（一）語料庫語言學是以文體研究做爲語言描述、立論的基礎，以具體量化的方式描述語言現象（Kennedy, Graeme D 1998：7）。

（二）Biber, Conrad, and Reppen 認爲以語料庫爲基礎的研究方式有以下特徵：1.基於大規模、有系統收集的自然語料的實證分析；2.廣泛的應用電腦工具進行分析，使用自動和互動的技術；3.同時運用質性和量化的分析技術（轉引自張學謙 2005：2）。

（三）Biber et al.指出使用以語料庫爲本的分析可以對自然言談的龐大語料進行使用模式的實證分析（轉引自盧慧娟 2006：161）。

（四）Kennedy 也指出以語料庫爲基礎的研究有助於語言學的描述與分析（轉引自盧慧娟 2006：161）。

語料庫語言學是語言學界的重要發展，以建立語料庫爲研究起點，運用電腦做爲研究工具，對大規模的自然語言進行分析，以定量的方式描述語言實際使用情形的一門科學。

三、中、英文語料庫簡介

語料庫是語料庫語言學的基礎工程，而且應用廣泛，例如：語言研究、辭典編纂、語言教學、教材開發等。但是語料庫的建置涉及結構、規模、語料選擇、語料加工以及語料庫管理等工作，是一項高資源、高成本的建設。各國大型的語料庫多由政府或學術機構建立，例如：「英國國家語料庫」（BNC）由政府出資一半，參與的單位有英國國家圖書館、牛津大學、蘭開斯特大學、朗文集團、錢伯斯出版社等；「日語言語數據庫」是由日本教育科學文化省組織三百多位學者共同完成的（黃昌寧、李涓子著 2002）。

自從 1959 年倫敦大學 Randolph Quirk 建立第一個大型電腦語料庫 SEU（The Survey of Engilsh Usage）以來，語料庫發展快速，各國政府、組織陸續建立各種語料庫，以下是幾個較具代表性的中、英文語料庫簡介：

表 2　中文語料庫簡介表

語料庫名稱	年代	主持或組織	語料	規模	特色
中央研究院平	1997	中研究中文知	中文，書面	500 萬	第一個有完

表 2　中文語料庫簡介表

衡語料庫 3.0 版[4]		識庫小組	語，6 種文類		整詞類標記的漢語平衡語料庫
中國國家現代漢語語料庫[5]	1993	中華人民共和國國家語言文字應用委員會	中文，書面語	7,000 萬	目前最大的漢語平衡語料庫

表 3　英文語料庫簡介表

語料庫名稱	年代	主持或組織	語料	規模	特色
SEU 語料庫	1959	Randolph Quirk 英國倫敦大學	英語，口語 50%，書面 50%	100 萬	第一個大型電腦語料庫
布朗語料庫	60 年代	Francis、Kucera 美國布朗大學	英語（美國），書面語，15 種文類	100 萬	共時平衡語料庫
LOB 語料庫	70 年代	Geoffrey Leech Lancaster 大學和 Oslo 大學	英語，書面語，15 種文類	100 萬	TAGIT 系統提高詞性標注準確率
LLC 口語語料庫	1981	Svartvik Lund 大學	英語，口語，五種文類	50 萬	第一個口語語料庫索引系統
COBUILD 語料庫	80 年代	John Sinclair Collins 出版社，Berminhan 大學	英語（英國 70%，美國 25%，其他 5%），書面語 75%，口語 25%	3.2 億	動態語料庫，編纂 COBUILD 詞典
朗文語料庫 (Longman)	1988~1990	朗文語料庫委員會	英語（英國 50%，美國 40%，其他 10%），書	2,800 萬	歷時語料庫 1990~目前

[4] 資料來源：中文詞知識庫小組：http://ckip.iis.sinica.edu.tw/CKIP/20corpus.htm。2008/9/7。

[5] 資料來源：中國國家語委現代漢語語料庫：http://www.clr.org.cn/retrieval/index.html。2008/9/7。

表 3　英文語料庫簡介表

			面語，10 種文類		
英國國家語料庫(BNC)	1991～1995	英國政府、國家圖書館、牛津大學、蘭開斯特大學、朗文集團、錢伯斯出版社等	英語，書面語 90％，10 種文類；口語 10％，5 種文類	口語 1,000 萬	最大的英語口語語料庫
國際英語語料庫(ICE)	1988	Sidney Greenbaum	英語，書面語，口語	20 個平行子語料庫	不同國家的英語比較

資料來源：黃昌寧、李涓子著（2002），表格由研究者整理。

四、語料庫的應用與研究

（一）語料庫的應用

語料庫的應用領域十分廣泛，楊惠中（2002）歸納出幾個比較重要的領域，例如：語言頻率統計、詞典編纂、詞彙搭配（collocation）研究、語言教學等。

本研究僅就幾項重要的應用領域，說明國內語料庫的應用概況。

1.詞典編纂

常用詞頻統計常用來編輯詞典與編寫教材，John Sinclair 編輯的 COBUILD 詞典，開啓以語料庫編纂詞典的先河。在台灣，1997 年黃居仁、陳克健和賴慶雄主編的「國語日報量詞典」，是台灣首次採用語料庫方法的範例（黃昌寧、李涓子 2002：170）。教育部國語推行委員會根據常用詞頻編纂「台灣閩南語常用詞辭典」，收錄國中、小學生日常生活用語，目前共有 1 萬 3 千餘詞，已推出網

路試用版[6]。

2.教材編輯

教育部自 1994 年開始規劃年度語詞調查統計工作，逐年進行統計提出報告，做爲教材以及語文工具書編輯參考（八十六年常用語詞調查報告書 1999）。

3.語言教學

學習者可利用語詞檢索（concordance）到語料庫中查詢詞的實際用法、搭配等資料。「台語文 concordance」即是提供線上台語語詞檢索學習的網站，目前有漢羅文本大約 5,816,250 音節；白話字文本大約 3,490,476 音節[7]。

4.第二外語教學

對比語料庫常應用於第二外語教學，盧慧娟（2006）比對「成功大學西班牙語學習者語料庫」（CATE-NCKU-3）與西班牙語語料庫（CLE）（篩選自西班牙皇家學院的現代西語語料庫），分析成功大學西班牙語系三年學生的常用詞彙和詞語搭配組合的模式類型與分佈傾向，作爲教學與設計教案的參考。

5.語言翻譯

關於應用語料庫進行翻譯的情形，以下學者有詳細的介紹：鄧敏君 （2005）介紹語料庫中日、日中翻譯的應用；陳瑞清（2003）將這幾年應用語料庫翻譯中英文的進展做詳細介紹；高照明（2002）則是簡介翻譯檢索系統在中英雙語

[6] 資料來源：教育部閩南語常用詞辭典試用版：http://twblg.dict.edu.tw/tw/index.htm。2008/11/3。

[7] 資料來源：台語文 concordance 網站：http://iug.csie.dahan.edu.tw/TG/concordance/form.asp。2008/11/3。

近譯句的應用。

（二）語料庫語言學的研究

語料庫語言學的研究大致可以分成兩個部分討論：一是對自然語料進行加工、標注；二是用已經標注好的語料進行語言研究和應用開發（黃昌寧、李涓子 2002）。本小節著重於以語料庫爲本的語言研究。

Biber 曾做過多項不同文類的詞彙研究，例如：以 Longman-Lancaster 語料庫 570 萬詞次語料，比較 big、large、great 在學術類、小說類的使用差異，發現 large 在學術類文本使用的頻率最高，great 在小說類文本使用的頻率最高（Biber、Conrad、Reppen 1998：43-44）。又以 Longman-Lancaster 語料庫爲基礎，探討小說類與學術類 begin 和 start 的語法關聯；以 Longman-Lancaster 語料庫學術語料，和英國國家語料庫（BNC）對話語料，研究 little 和 small 在兩種不同語域中謂語形容詞用法的差異（黃昌寧、李涓子著 2002）。

在台灣，華語方面利用「中央研究院平衡語料庫」進行的研究有很多：廖小婷（2003）採用中研院平衡語料庫的語料分析中文的施力動詞---「拉、拖、扯」這組近義詞的詞彙語意特徵。研究方法主要是透過詞語搭配(collocation)辨別每個動詞詞彙語意的基本特徵，主要目的是要從句法中的互補分佈，定義出近義詞組「拉、拖、扯」的語意特質。黃郁純、陳薾宇（2005）以中研院平衡語料庫爲基礎，探討「擺」和「放」的詞語搭配及近義關係。研究顯示「擺」是比較靜態的對物體描述，「放」比較屬於動態的對物體處置；「擺」比「放」更具有深層意義。陳珮嘉（2000）探討漢語動詞單位詞與動詞搭配關係；余明憲（2005）探討現代漢語中的三個「框架觸發動詞」--「玩」、「弄」和「搞」在「動賓」格

式中之格式語意；謝佳玲（2006）研究漢語情態詞的語意界定。

其他以語料庫語言學方法的研究還有：王萸芳（1995）研究漢語口語與書面語中副詞子句的訊息順序，發現出現在主要子句前的副詞子句爲引述下文之功用，在主要子句後的副詞子句是爲補充解釋前面的句子，通常出現在主要子句前的副詞子句所修飾的範圍較大。劉賢軒（2005）比較台籍應用語言學研究者與相同領域的英美籍學者所寫的論文，探討三種應用語言學論文中的態度成分：評斷符號、強調符號和謹慎符號。發現台灣應用語言學研究者已經具備基本的學術論文寫作能力，但是英文能力和學術寫作的成熟度仍比不上英美籍作者。盧慧娟、林柳村、白芳怡（2007）以語料庫爲本應用語詞搭配的語言教學研究。洪嘉馡、黃居仁（2008）以語料庫爲本的兩岸對應詞彙發掘。

以語料庫語言學方法研究台語的起步比較晚，不過到目前爲止也已累積不少研究資料，相關的研究計畫也在陸續進行中，台語語料庫語言學的發展已愈來愈受到重視。

早期台語語料庫的研究有顏國仁（1995）台語口語的詞頻調查，口語語料的來源主要是電台錄製的台語談話節目以及日常生活對話，以漢羅文字的形式轉錄成 12 萬字的書面語，經過斷詞、詞頻統計後，得到字頻表、詞頻表、以及雙字組頻表三個常用詞頻統計表，整個研究只是建立台語口語語料庫的初步報告。研究過程中遇到的台語文字標準化、詞彙定義、斷詞等問題，這些也是目前台語語料庫研究主要的困難。

台語語料庫尚未建立之前，基於台語語料庫的研究，語料多爲研究者自行蒐集建立。張學謙（2000）是最早以語料庫語言學的方法比較台語口語和書面語的研究。該研究建立了 94 篇口語語料（9 類）與 91 篇書面語料（8 類），總計

144,942 個詞的研究語料，進行台語口語與書面語的多面向分析，主要在找出影響台語語體變異的深層言談面向，同時刻畫語體的篇章關係，經過分析之後得出五個深層言談的面向。李勤岸（2000）蒐集 1920 年代 112,964 個詞以及 1990 年代 92,539 個詞，建立總計 205,503 個詞的研究語料，比較兩個年代台語詞彙流失與台語借詞的情形，發現 1990 年代華語借詞大量增加、日語借詞不減反增、教會用語大量減少。楊允言（2003a）蒐集卓緞女士 25 首白話詩歌，共 2,878 個詞，以李勤岸（2000）的研究爲基礎，從語域及借詞的觀點探討台語文寫作風格。由個人建立語料庫進行研究是一件耗時耗力、繁雜的工作，而且建立的語料亦未能公開供後續的研究者使用，形成一種資源浪費，如能建立一個公開的台語語料庫，對以後的台語研究將是一大助益。

近幾年陸續建立台語書面語與口語語料庫，「台語文數位典藏資料庫(第二階段)——台語文學線上博物館」就是一個公開的書面語語料，「台語文語料庫」雖然尚未正式公開，但是已提供台語語詞檢索、各類統計表做爲查詢與研究之用，楊允言（2004）以其中收錄的 1916 年巴克禮聖經和 1974 年紅皮聖經爲研究語料，比較發現台語詞彙在六十年裡流失了 43%。蕭如卿 （2006）以「台灣兒童語料庫」探討台灣兩歲一個月到四歲之兒童閩南語量詞習得，結果發現兒童與外在世界接觸的經驗會影響量詞習得的順序。以「台灣兒童語料庫」進行的研究還有 Hung（2005）研究動詞的習得；Hung, Li & Tsay（2004）研究語尾助詞的習得…等。

其他的研究還有：謝昌運（2007）分析戲劇、小說、散文、社論、學術論文等五種台語文本，低調詞、退讓詞、擴充詞、強調詞等四種加強詞的使用差異。

顯示強調詞最常使用，低調詞使用的頻率最少；五種文類裡，散文最常使用加強詞，學術論文使用的頻率最少。李珮甄 （2007）以口語語料庫探討台灣閩南語「是講」、「著是講」的語用功能，分析發現「是講」和「著是講」從原本的繫詞作用，延伸出多種的語用功能。

第二節　台語語料庫

台語語料庫建立的起步比較晚，過程中並沒有得到太多支援，「台語文所能運用的資源，大概不及華語的千分之一」（楊允言 2003c），台語語料庫的建立大部分是倚靠個人力量和政府單位少許經費補助下進行的。1990 年鄭良偉在 DOS 作業系統平台上開發 TW301 軟體，1994 年蘇芝萌在 Windows 作業系統上開發 HOTSYS 軟體，解決台語電腦輸入法與文書處理的問題；1999 年鄭良偉與 Roderick Gammon 合作開發的 TMLAP，功能包括斷詞、詞性標示、詞頻統計…等；劉杰岳 2001 年開發 Taiwanese Package（簡稱 TP），解決台語符號在網路顯示的問題後，台語網站發展快速，擴展台語文在網際網路的流通性；目前台語文已累積不少數位化的作品與刊物，台語語料庫的發展可說已經達到成熟的階段（楊允言 2003c）。

目前台灣已建立數個公開與未公開的語料庫，有「台語文數位典藏資料庫（第二階段）」、「台語文語料庫」、「台灣兒童語料庫」和「閩南語口語語料庫」。除了以上的語料庫之外，還有許多個人基於研究需要所建立的小型語庫料。

以下分成五個部分簡介「台語文數位典藏資料庫（第二階段）」、「台灣兒童語料庫」、「閩南語口語語料庫」、「台語文語料庫」以及小結。

一、台語文數位典藏資料庫（第二階段）

「台語文數位典藏資料庫」[8]是國家台灣文學館建立的台灣文學語料庫，委託呂興昌執行「台灣白話字文學資料蒐集整理」計畫，蒐集到一千餘本白話字書刊；高成炎執行「台語文數位典藏資料庫(第一階段)——台語文全羅文字語音輸出系統」，將全羅馬字的台語文資料轉成聲音，透過網路放播放出來；楊允言執行「台語文數位典藏資料庫(第二階段)——台語文學線上博物館」，此計畫承接前述兩個計畫的成果，將已經打字建檔且取得授權的資料上網。

「台語文數位典藏資料庫」目前已完成兩個階段，建立漢羅、全羅對齊語料，各 258 萬音節，分爲清國、日治以及終戰後三個時期，文本分爲詩、散文、小說以及劇本四類，以漢羅、全羅文字對照的方式呈現，並且附有語音輸出可供學習，現在已將資料上網供使用者查詢。

二、台灣兒童語料庫（Taiwan Child Language Corpus, 簡稱 TAICORP）

台灣兒童語料庫是由蔡素娟主持建立的，語料來源是十四名嘉義縣民雄鄉一歲二個月至五歲三個月的兒童，共有 431 人次錄音檔案，約 330 小時，以世界標準的兒童語料交換系統（Child Language Data System, CHILDES）爲格式建構的語料庫，有 46 個詞類標記，是世界上第一個有詞類標記的台語電腦語料庫，

[8] 台語文數位典藏資料庫（第二階段）：http://iug.csie.dahan.edu.tw/nmtl/dadwt/pbk.asp。2008/8/23。

目前收錄於國家數位典藏，並且架設網站提供資料作爲學者研究之用，網址：http://www.ccunix.ccu.edu.tw/~lngcorp/Taicorp-Homepage1/index.htm（蔡素娟 2004）。

台灣兒童語料庫的語料統計如下表：

表 4　台灣兒童語料庫語料統計表

項目	行（lines/utterances）	詞（words）	平均句長（MLU）
兒童	161,253	434,557	2.695
成人	336,173	1,211,946	3.605
合計	497,426	1,646,503	3.150

資料來源：Tsay（2005）；轉引自蔡素娟（2007：359）。

說明：語料內容是兒童活動時，與家人（或研究員）自然言談的錄音，因此包含兒童與成人的語料。

三、閩南語口語語料庫

「閩南語口語語料庫」的建置過程基本上和「台灣兒童語料庫」類似，不同的是這個語料庫是蒐集成人的口語語料爲主，語料來源是雲林、嘉義地區的台語節目錄音，目前已經轉記成文字的錄音約有 37 個小時，40 萬詞左右（蔡素娟 2007）。

四、台語文語料庫

「台語文語料庫」[9]是由楊允言與張學謙共同主持建置的，是目前爲止收錄音節數最多的台語書面語語料庫，建立的目的是爲了台語的相關研究建立基礎，提升台語文的地位，並且促進台語文計算語言學的發展。

「台語文語料庫」的簡介如下：

（一）語料

1.語料來源

（1）台文刊物：包括《台文通訊》(1991 年創刊)、《台文罔報》(1996 年創刊)、《TGB 通訊》(1999 年創刊)、《蓮蕉花》(1999 年創刊)、《台灣字》(2000 年創刊，全羅馬字)、《滾根》母語文雜誌(2002 年創刊，現已停刊)、《台灣公論報》蕃薯園台文專刊(2003 年創刊)、...等。

（2）專書或論文：主要由作者或編者提供。

（3）研究計畫成果：主要爲國家台灣文學館的「台灣白話字文學資料蒐集整理計畫」中已經數位化的電子檔，執行的時間至 2004 年 12 月止。

2.語料規模

本研究所取得台語文語料庫規模是截至 2005/7/31 爲止的資料。

[9]「白話字台語文網站」---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：http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/kiatanpoko/kiatanpoko.asp。2008/8/22。

表 5　台語文語料庫規模表

	音節次	音節型	詞次	詞型
漢羅合用台語	5,568,057	8,527	4,051,195	47,130
台語羅馬字	3,462,367	3,525	2,436,599	73,258
合計	9,030,424		6,487,794	

資料來源：「白話字台語文網站」---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：http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/kiatanpoko/kiatanpoko.asp。2008/8/22。表格爲研究者整理。

3.語料分佈

下表是台語文漢羅、全羅語料文類的比例分佈表。

表 6　台語文語料文類分佈表

文類[Bûn-lūi]	**漢羅[Hàn-lô]**	**全羅[Chôan-lô]**
學術**[Ha̍k-su̍t]**	7.48%	2.01%
報導**[Pò-tō]**	4.23%	2.54%
訪談**[Hóng-tâm]**	1.42%	0.00%
傳記**[Tōan-kì]**	2.90%	5.03%
評論**[Phêng-lūn]**	4.87%	4.39%
其它**[Kî-tha]**	1.20%	0.34%

表 6　台語文語料文類分佈表

小說[**Siáu-soat**]	29.31%	59.08%
散文[**Sàn-bûn**]	35.78%	17.16%
新詩[**Sin-si**]	5.30%	3.42%
劇本[**Ke̍k-pún**]	3.43%	3.42%
兒童[**Gín-á**]	0.41%	0.97%
笑話[**Chhiò-khe**]	0.27%	0.24%
寓言[**Gū-giân**]	0.24%	0.12%
對話[**Tùi-ōe**]	0.38%	0.04%
書信[**Phoe-sìn**]	1.04%	0.58%
民間文學[**Bîn-kan bûn-ha̍k**]	0.72%	0.11%
演講[**Káng-ián**]	1.02%	0.54%

資料來源：「白話字台語文網站」---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：http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/kiatanpoko/kiatanpoko.asp。2008/8/22。

（二）台語文語料庫的應用

1.台語語詞檢索（concordance）：分爲漢羅和全羅兩個部分，提供學習者學習欲查詢語詞的用法。

2.各類統計表：目前將語料庫全羅、漢羅的音節、語詞相關統計所得上網，提供台語文研究使用。

表 7　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

	羅馬字(P)	漢羅(H)
音節層次 (S)	頻率統計(Frequency Count) 互訊息(Mutual Information) 相關度(Correlation)	頻率統計(Frequency Count) 互訊息(Mutual Information) 相關度(Correlation)
語詞層次 (W)	頻率統計(Frequency Count) 互訊息(Mutual Information) 相關度(Correlation)	頻率統計(Frequency Count) 互訊息(Mutual Information) 相關度(Correlation)

資料來源：「白話字台語文網站」---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：http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/ guliau-supin.asp。2008/8/22。

五、小結

台語語料庫的建立雖然起步比較晚，在有志之士的努力下，目前已經成立數個公開和尚未公開的語料庫；語料庫是一件高成本的工程，台語語料庫的建立需要政府更積極的參與。

第三節　台語詞彙

台語是從福建閩南地區傳入台灣的，因爲歷史及地理的因素，已經發展成不同於福建地區的閩南語。就詞彙來說，根據王毓仁的調查，台語和福建閩南語至

少有 10%的差異（張學謙 1998）。

台語詞彙的成份相當複雜，有源自福建閩南語底層的古漢語、古畲語、閩越語、吳語以及楚語等（周長楫 1996），傳入台灣之後先後接觸了原住民語以及荷蘭語、西班牙語、日本語、英語等外來語，堆疊覆蓋、摻雜融合，所以郭一舟形容台語詞彙像「九重粿」一樣，層層複雜（楊允言 2003a）。鄭良偉（1990）歸納台語詞彙有六點特色：保留古代漢語成份、文白異讀漢字特別多、有音無字的語詞特別多、英日外來語特別多、借用日語的漢語語詞以及合音詞等。

台灣語言自然的演變趨勢就是雙語共通，華語和台語的共通比例會愈來愈高（鄭良偉 1990）。台語過去因爲政策的打壓以及書面語尚未完全標準化等問題，在社會及語言地位上台語是屬於弱勢、低階的語言，華語屬於強勢、高階的語言，Allard & Landry 認爲：「弱勢族群常被迫放棄母語，轉向強勢語言」（張學謙 2004：61）。而且現在華語已經侵入台語家庭，台灣漸漸走向雙語但非雙言的社會（黃宣範 2004）。台語詞彙借用華語已是時勢所趨，比例也會愈來愈高。

台語詞彙有多種不同面向，本節擬探討台華共通詞與台語特別詞、詞彙豐富度、台語羅馬字詞彙以及台語平均詞長等五個面向。

一、台華共通詞與台語特別詞

「語言層」（linguistic stratum）是語言因接觸發生變化的時候，不同語言變體的作用力在語言結構上表現出來的痕跡，包括音韻、詞彙、構詞、句法各方面（何大安 1996）。本研究所要探討的是台語「語言層」中詞彙的部分。

台語詞彙層層堆積有如沈積岩，詞彙層如何畫分，學者有不同的觀點。張學謙(1998)將台語虛詞分爲三層：文言層、台華共通語層及本土層；李勤岸(2000)將台語詞彙分爲兩層：本土語層和非本土語層(又稱爲移借語層)；林香薇(2003)分爲四個層次：台華共通語層、文言層、本土層、移借層。本研究參考以上學者台語詞彙的畫分方法，將台語詞彙畫分爲兩層：台華共通層和台語特別層。

（一）台華共通詞

台華共通詞就是台語和華語共通的詞彙，可以分成兩部分：一部分是台語、華語共同繼承古漢語的語詞；一部分是透過漢字轉讀的「對音詞」(張學謙1998)。台語、華語共同繼承古漢語同語源的部分，鄭良偉（1984）認爲漢字寫法比較固定，例如：國家、社會等詞。「對音詞」的部分，是台語借用華語詞的主要方式，主要爲多音節的新語詞，借詞不借音，凡是用漢字書寫，用台語發音，就可以借用（張學謙 1998）。

綜合以上所述可知，台華共通詞的漢字書面語標準化的程度比較高，主要有兩個來源，一是台語、華語有共同語源的詞，二是台語向華語移借來的對音詞。本研究參考以上兩位學者的觀點，將台華共通詞定義爲：和華語文詞形相同而且詞義相近的台語詞彙。

（二）台語特別詞

台語特別詞係指在台語文與華語文書面語中，意義相同的詞彙有不同的書寫方式，例如：曝日頭、暗 bong-bong 等詞就是台語特別詞。姚榮松（2000）認爲特別詞與通用詞是相對的，凡是與共同語或鄰近方言詞形相同的便不是特別詞。鄭良偉（1984）認爲台語特別詞的漢字寫法比較不固定。

綜合以上所述可知，台華共通詞與台語特別詞是指相對的意義而言，只要書寫的形式不同，就不是共通詞；台語特別詞比較沒有一致的漢字書寫形式。本研究參考以上學者的觀點將台語特別詞定義爲：和華語文詞義相同但寫法不同的台

語詞彙。

從圖 1 可以比較清楚了解台語、華語和台華共通詞的關係。

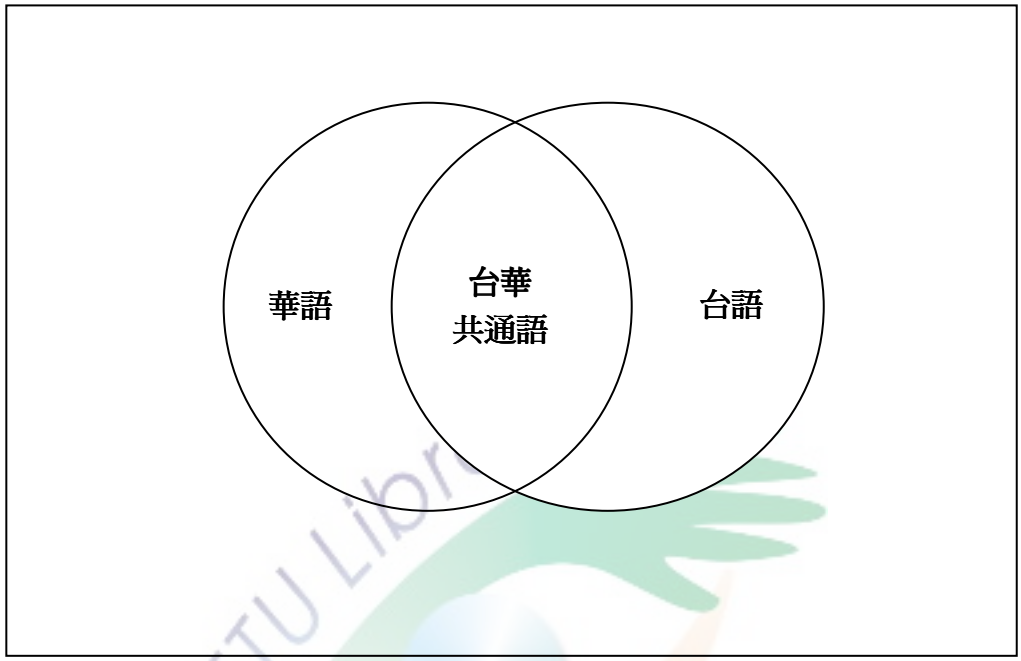

圖 1　台華共通語示意圖

（三）統計研究

關於台華共通詞和台語特別詞在書面語以及不同文體的使用情形，學者有以下的研究與看法：

鄭良偉（1990）認爲台灣各種語言自然演變的特色之一是雙語共通化，1960 年前台語及華語的共通語詞只有百分之六十，但是 1990 年共通的語詞差不多有百分之八十五。

鄭良偉統計村上嘉英編的「現代閩南語辭典」所收錄的台語詞彙，台華共通的語詞佔 65.6%，不同的詞彙佔 34.4%（張學謙 1998）。黃宣範（1998）調查同一本辭典的結果，多音節詞（雙音節及雙音節以上）有 7,860 個，台華共通詞

有 5,730 個，約佔 73%，台語特別詞有 2,130 個，約佔 27%。

綜合上述鄭、黃兩位學者的研究，可以歸納出兩點結論：第一點，兩人所做的是屬於普遍性的詞型統計；第二點，台華共通詞的比例約介於 66%~73%，台語特別詞的比例約介於 27%~34%。

鄭、黃兩位學者的調查是一般台華共通詞的使用情形，Sander 及 Hsieh 的研究則是探討不同文體的詞彙使用差異，統計結果如下表（張學謙 1998）：

表 8　不同文類共通語層及本土語層詞彙使用差異表

項目	新聞報告（散文）	可愛的仇人（小說）	情歌（詩）	諺語及俗語
共通語層	266（77%）	628（67.5%）	182（67.4%）	102（67.5%）
本土語層	79（23%）	302（32.5%）	88（32.6%）	49（32.5%）
合計	345（100%）	930（100%）	270（100%）	151（100%）

資料來源：Sander 及 Hsieh，轉引自張學謙（1998）；表格爲研究者整理。

從上表可知：共通語層小說、詩以及諺語、俗語的比例相當一致，大約佔 67%左右；散文類（新聞報告）的比例最高約 77%，比其他三類多出 10%左右。

鄭良偉認爲一般文章中台華共通詞約佔 70%，有差異的語詞約佔 30%；日常會話、詩歌、俗語台語特別詞的使用比例有時佔文章的一半左右，散文大約佔百分之二十到四十之間（張學謙 1998）。

綜合以上所述，台華共通詞的使用比例超過 60%，有愈來愈高的趨勢；而

且不同的文體有比例上的差異。

（四）台華共通詞在學術類與非學術類使用情形

台語屬於弱勢、低階語言，目前尚無一套高度標準化的書面語，部分詞彙並沒有固定寫法，學術類與非學術類著作因爲訴求的重點不同，選用台華共通詞的情形應該也會有所差異。

David Crystal（1995）認爲：高階語言用於布道、講演、演說、新聞廣播、報刊社論等正式場合；低階語言用於日常會話、討論、民間文學及其他非正式的語境。

David Crystal（1995：584）認爲：「科學的方法論連同客觀性、系統性和準確性，對語言產生了不少影響。……日常語言用于科學研究意義太含糊。」

宋澤萊在《一枝煎匙》〈序言〉裡說：「只有以母語寫詩才能充份把握詩的韻律及民族的情感，才不會有『隔』的感覺」（林香薇 2003：114）。姚榮松（1990）認爲台語詞彙在小說中扮演的角色功能有三點：第一點，爲小說人物的社會階層定位；第二點，營造小說的社會背景；第三點，強化鄉土文學的色彩。

綜合以上所述可知，學術類著作比較屬於正式的文體，著重事理現象客觀、系統、準確的論述；非學術類著作，例如詩、小說、劇本等比較屬於非正式的文體，重視的是情感的傳達，以及如何讓讀者融入著作之中。因此，對於詞彙的選用，學術類文本會比較傾向有標準化書面語的高階語言，非學術類文本會比較傾向有鄉土味的低階語言。本研究據此推論：台華共通詞的使用比例，學術類高於非學術類。

二、詞彙豐富度

詞彙豐富度即文本中詞彙的豐富程度，也就是作者所能夠掌握運用的詞彙，

這裡所說的詞彙是指詞型而言。衡量詞彙豐富度有許多計算方式，本研究是參考（楊允言 2003a）的方法：詞型÷詞次，得到的值愈高表示詞彙愈豐富，反之愈低。

目前國內研究台語詞彙豐富度的文獻並不多，楊允言（2003a）以卓緞的白話詩以及李勤岸整理的 1920 年和 1990 年小說文本，比較詩和小說不同文類的詞彙豐富度，是少數的研究之一。結果如下表：

表9　詞彙豐富度比較表

	詞型	詞次	詞彙豐富度
1920 年代文本	12,941	112,764	11.48%
卓緞作品	829	2,878	28.80%
1990 年代文本	12,969	92,539	14.01%

資料來源：楊允言（2003a）。

上表的統計資料，因爲詩和小說的語料數量差異太大，很難以此結果比較詩和小說之間的詞彙豐富差異。但確實是比較不同文體詞彙豐富度可行的方法之一。

由於國內研究台語詞彙豐富度的文獻有限，研究者嘗試以鄭錦全（1998）整理的中國歷代古書總用字數與用種字類爲基礎，比較不同文體的詞彙豐富度，以期有初步的了解。

鄭錦全（1998）文中所說的「字數」以及「字種[10]」和本研究「詞次」與「詞

[10] 鄭氏所說的「字種」即本研究的「單音節詞型」，「字數」即本研究的「單音節詞次」。漢字屬

型」的概念相似。鄭氏認爲個人所能掌握運用的詞彙數量大約八千個左右，以詞型÷詞次計算詞彙豐富度，若單純只考慮數學問題，詞次愈多詞彙豐富度愈低，反之亦然。

本研究將鄭氏整理的資料，詞次由低而高依序排列，以詞型÷詞次計算，比較不同文體詞彙豐富度[11]的差異；並觀察詞彙豐富度因詞次增加而下降的情形。統計結果如下表：

表 10　中國古籍詞彙豐富度統計表

書目	字種（詞型）	總字數（詞次）	詞彙豐富度
風俗通[12]	2,716	34,431	7.89%
孟子[13]	1,913	35,417	5.40%
毛詩[14]	2,989	37,438	7.98%
大戴禮記[15]	22,59	38,597	5.85%
北齊書	4,032	212,506	1.90%

於表意文字，每個漢字大都能代表一個詞素（morpheme），單獨出現成爲一個詞語（word）。

[11] 鄭氏的研究僅整理單音節詞的部分，並未包含多音節詞，下表詞彙豐富度也僅是以單音節詞所計算出的結果。

[12]《風俗通》東漢應劭著，內容爲記錄考釋當時的社會風俗故事。

[13]《孟子》由孟子及其弟子公孫醜、萬章等人編著。以答問的方式記述孟子思想的著作。

[14]《毛詩》即《詩經》，毛享注釋，爲西周初期至春秋中葉的詩歌集，作者多不可考。

[15]《大戴禮記》漢朝戴德編著，以散文的方式記載孔子的學生及戰國時期儒學學者的作品。

表 10　中國古籍詞彙豐富度統計表

紅樓夢後 40 回	3,217	234,980	1.37%
周書	4,161	262,659	1.58%
日知錄[16]	5,225	459,357	1.14%
紅樓夢前 80 回	4,293	496,855	0.86%
隋書	5,592	701,698	0.80%
紅樓夢 120 回	4,501	731,835	0.62%
漢書	5,833	742,298	0.79%
舊五代史	5,109	790,879	0.65%

資料來源：鄭錦全（1998），表格由研究者整理計算。

從表 10 可知：詞彙豐富度：故事小說（風俗通，7.89%）高於口語（孟子，5.40%）[17]；詩（毛詩，7.98%）高於散文（大戴禮記，5.85%）[18]；散文（日知錄，1.14%）高於小說《紅樓夢前 80 回，0.86%》。以上結果是以單音節詞比較不同文體的詞彙豐富度，並未考量作者、年代等相關因素，僅可做爲初步參考，若要進一步了解需蒐集更多的語料進行分析。

史書（北齊書、周書、漢書、隋書、舊五代史）的詞彙豐富度雖然高於小說

[16]《日知錄》爲顧炎武平日讀書心得的札記，顧炎武死後由弟子潘耒蒐集手稿校定編寫而成。

[17]《孟子》雖非一人所作，但多爲記錄孟子一人言行之作，故以之與一人之作的《風俗通》做文體之間比較。

[18]《毛詩》與《大戴禮記》皆爲多人的作品彙編而成，故以二書做文體之間詞彙豐富度的比較。

《紅樓夢》（後 40 回、120 回），但是這可能是史書多半出於眾人之手的集體創作，小說則多爲一人之作（紅樓夢爲一人或二人所作有待考查）的結果，作者人數才是影響詞彙豐富度的主要因素，並非文體。

若僅從詞次的角度觀察詞彙豐富度，可以發現詞彙豐富度隨著詞次增加而遞減，也間接說明個人所能掌握的詞彙有其極限性。

從上述的研究可以初步了解不同文體的詞彙豐富度有其差異存在。David Crystal（1995：584）認爲：「科學的方法論連同客觀性、系統性和準確性，對語言產生了不少影響。……日常語言用于科學研究意義太含糊。」在詞彙選用方面，客觀、系統、精確以及有規範固定寫法，是學術性文本的主要考量。相對而言，非學術性的文本，例如小說、散文、劇本等屬於想像虛幻的情節，內容可以是報導、說理、傳教、故事、神話等，自由發揮的空間比較大，相同意義的詞彙，可依不同劇情場景做更換，主要的訴求是作者情感的表達與讀者的反應，不在詞彙本身的客觀、系統或有無標準規範。由以上論述可知，學術與非學術性文本對詞彙訴求的重點不同，相對而言，非學術性文本可以有比較多的選擇與發揮空間。因此，本研究據以假設：台語詞彙豐富度，非學術類多於學術類。

三、台語羅馬字詞彙

目前用以表記台語的書面語至少三種，全漢字，全羅馬字以及漢羅合用的文字，本研究所指台語羅馬字系指全羅文字或漢羅文字而言。

羅馬字最早的文獻一般認爲是 1837 年 W. H. Medhurst 編纂的《福建方言字典》，1885 年《台南府城教會報》創刊，是台灣第一份報紙，1913 年 William Campbell 年出版《廈門音新字典》，流傳甚廣（張裕宏 2001）。所以，以羅馬

字表記台語已有一百多歷史，而且有不少文獻可供參考。

1931 年，林鳳岐提議用羅馬字代替有音無字的台語詞，1964 年王育德提出漢羅合用理論，不過並沒有實際用於寫作，1980 年代後期，鄭良偉以實際行動將漢羅文字引進台灣，並且用以著作發表（張學謙 2003）。以漢羅文字出版過的刊物有《台語通訊》、《台文通訊》、《台語風》、《台灣學生》、《台文罔報》以及全羅、漢羅合用的《台灣羅馬字協會通訊》等（楊允言、張學謙、呂美親 2008）。漢羅文字雖然起步較晚，但已爲愈來愈多人接受與使用。

張學謙（2003）研究台語五種文字的社會評價，全羅字的地位權勢最高，親和力最差；漢羅字在地位權勢與親和力上都排名第二。顯示羅馬字在一般民眾心中地位是比較高的，如果漢字和羅馬字一起使用，除了有權勢之外而且也比較容易爲一般民眾所接受。

台語「有音無字」的詞彙，楊秀芳（1995）歸納有擬聲詞、外來語、譬況詞、合音詞、閩南語底層非漢語詞彙、以及音字脫節等幾種類型。許極燉（1994）認爲台語的擬聲詞、擬態詞與外來語要用羅馬字書寫。這些「有音無字」的台語詞比例有多少？郭秋生表示不過是百字中五個半找不到（黃宣範 2004），王育德（2000）認爲約有四分之一找不到正確的漢字。

學術類的文章需要一定程度規範的書面語，因此，選用「有音無字」台語詞的機率應該不高，使用羅馬字的比例自然也較低。非學術類的文章，注重情感的表達，以及與讀者的距離，因此選用台語特別詞創作的機會應該比較高，羅馬字的比例相對也會比較高。所以本研究假設台語羅馬字詞彙的比例，非學術類高於學術類。

四、台語平均詞長

台語平均詞長，台灣目前少有學者實際做過相關的調查。本研究嘗試以音節數的多寡推論比較台語、華語的平均詞長，並以台華共通詞在學術類與非學術類不同的使用比例，分析比較學術類與非學術類的台語平均詞長。

語意由詞彙來表達，詞彙是由音節組成，從語言音節數的多寡約略能夠推測出詞彙的平均長度。音節數比較多的語言，意謂著一個詞彙只用單音節表示的選擇比較多，可能性也比較高；反過來說音節數比較少的語言需要用比較多的音節來表示一個詞彙。因此，相對而言，音節數多的語言平均詞長比較短；音節數少的語言平均詞長比較長，音節數與平均詞長呈反比。

一個音節包括聲母、韻母、聲調三個部分，可能的音節數是聲母數乘以韻母數乘以聲調數，但是很少語言會使用到所有可能音節數，例如：台語有「l」聲母、「oai」韻母，「loai」卻不是合法的台語音節。黃宣範（1988）以聲韻組合數乘以聲調估計台語的音節數約有 5,460 個（780 個聲韻組合數乘以 7 個聲調），華語的音節數約有 1,644 個（411 個聲韻組合數乘以 4 個聲調）。以上的估計是理想的音節數，實際使用時有許多是沒有用到的空音節，但是由此可以初步看出台語的音節數比華語多。

楊允言（2003a）調查「台語線上字典」，共有 2,728 個音節；華語的部分，以詞庫小組技術報告 98-01 詞頻詞典的索引計算，共有 1,081 個音節，台語約是華語的 2.5 倍。

黃宣範（1988）另一項調查是以「國語日報詞典」與「現代閩南語詞典」爲例，比較台語和華語多音節化的程度。研究結果「國語日報詞典」雙音節詞是

單音節詞的 8 倍，「現代閩南語詞典」雙音節詞僅是單音節詞的 2 倍，顯示台語雙音節的程度低於華語。由此推論：華語的平均詞長應該會比台語長。

綜合以上兩位學者的研究推論可知：台語的音節數應該會比華語多；也就是華語的平均詞長比台語長。如果本研究的第一項假設：「台華共通詞的比例，學術類多於非學術類」成立，亦即學術類文本使用較多的華語詞彙，本研究的第四項假設：「台語平均詞長，學術類多於非學術類」應該能夠成立。

漢語的平均詞長中國學者曾做過研究，根據《現代漢語詞語頻率辭典》的調查約爲 2.0928（轉引自湯志祥 2001）。湯志祥（2001）亦以「兩岸三地漢語語料庫」爲基礎調查漢語平均詞長。茲將《現代漢語詞語頻率辭典》與湯志祥（2001）的調查詳列如下表：

表 11　漢語平均詞長比較表

項目	現代漢語詞語頻率辭典	兩岸三地漢語語料庫			
		台灣	中國	香港	台灣、中國、香港
平均詞長	2.0928	2.1770	2.2049	2.1788	2.2706

資料來源：湯志祥（2001）；表格爲研究者整理。

從上表可知：三個地區的華語平均詞長差別不大，大約是 2 左右。

五、小結

台語詞彙有許多面向，本研究從台華共通詞、台語特別詞、詞彙豐富度、台

語羅馬字、平均詞長探討在學術類與非學術類的使用差異，從文獻分析中歸納出四項假設：

（一）台華共通詞使用比例，學術類多於非學術類。

（二）詞彙豐富度，非學術類多於學術類。

（三）台語羅馬字詞彙使用比例，非學術類多於學術類。

（四）台語平均詞長，學術類多於非學術類。

第四節　本研究特色

語料庫語言學研究方法是對大規模的自然語言進行分析，以量化的方式描述語言實際使用情形。在台灣，台語詞彙的相關論述很多，但是以台華共通詞、詞彙豐富度、台語羅馬字詞彙、平均詞長等面向分析不同文體詞彙使用差異的著作有限，以量化方式呈現的文獻更少。本研究即是對於上述研究不足之處進行探討的，特色在於：

一、第一個將文體分成學術類與非學術類，以語料庫語言學方法探討台華共通詞、詞彙豐富度、台語羅馬字詞彙、平均詞長等面向的研究。

二、第一個以量化方式描述台語羅馬字詞彙使用比例的研究。

三、第一個以量化方式描述台語平均詞長的研究。


NTTU Library

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。

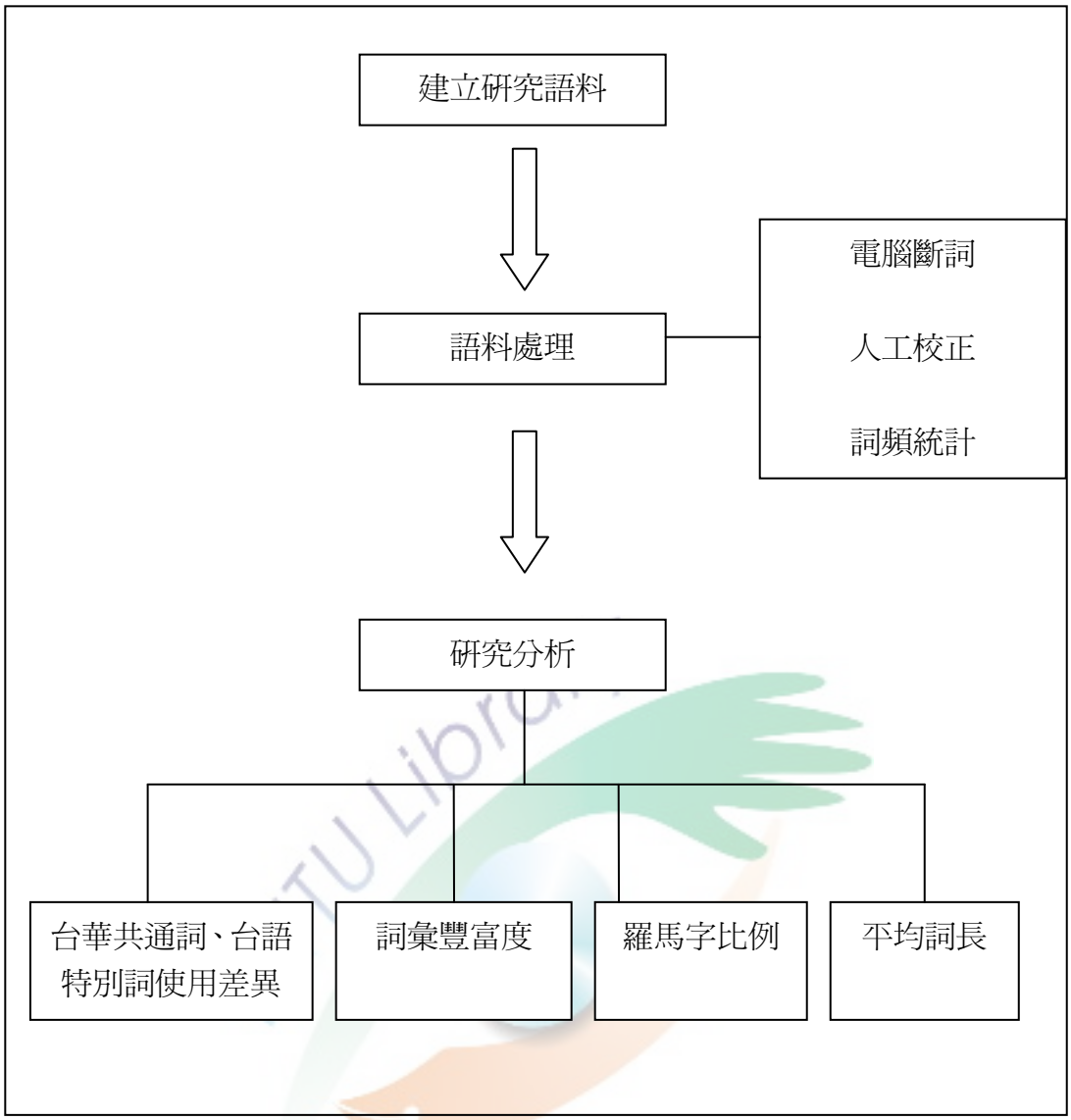

圖 2　研究流程圖

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
呣閣	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	韌性	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：「踅篦」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「喚」、「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。

第四章　結果分析

本研究採用語料庫語言學的研究方法，探討台語詞彙在學術和非學術書面語的使用差異。建立學術類語料 101,349 音節，非學術類語料 102,335 音節，合計 203,684 音節，經過電腦斷詞、人工校正、詞頻統計後，得到學術詞頻統計表與非學術詞頻統計表各一份，做爲分析台華共通詞、台語詞彙豐富度、台語羅馬字詞彙、台語平均詞長等四項台語詞彙特性的使用差異。分析的結果分成四節做說明：第一節台華共通詞的使用差異分析；第二節台語詞彙豐富度分析；第三節台語羅馬字詞彙分析；第四節台語平均詞長分析。

第一節　台華共通詞的使用差異分析

本節分成詞型、詞次、詞型與詞次比較分析、小結四個部分討論台華共通詞在學術與非學術書面語的使用差異，在詞型和詞次的部分再細分成覆蓋率 100％、覆蓋率 80％、不同覆蓋率的比較以及小結四個項目討論。

一、詞型分析

分成覆蓋率 100％、覆蓋率 80％、覆蓋率比較分析、小結四個部分討論。

（一）覆蓋率 100%

表 20　台華共通詞與台語特別詞使用差異統計表（詞型，覆蓋率 100%）

項目	學術類		非學術類	
	詞型	比例	詞型	比例
台華共通詞	4,588	66.91%	5,003.5	55.09%
台語特別詞	2,269	33.09%	4,079.5	44.91%
合計	6,857	100.00%	9,083.0	100.00%

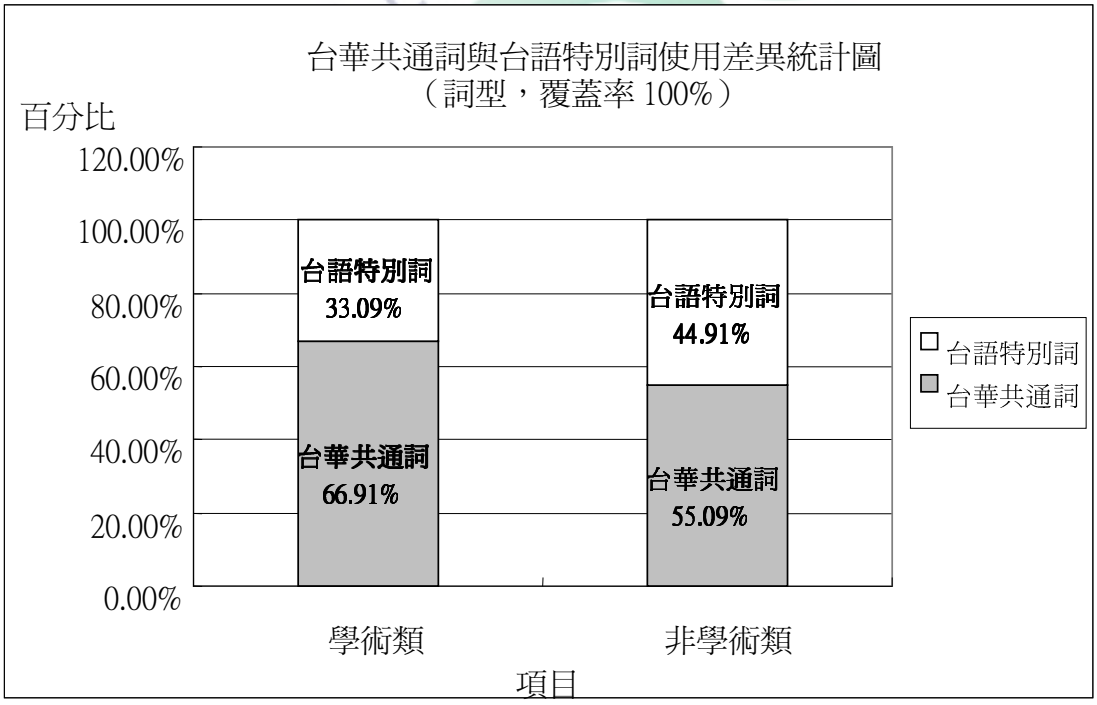

圖 3　台華共通詞與台語特別詞使用差異統計圖（詞型，覆蓋率 100%）

從上表和上圖可知：台華共通詞在學術類佔 66.91%，非學術佔 55.09%，

學術類比非學術類多出 11.82%；兩種文類所佔的比例都超過 50%，學術類達到近 70%的水準。統計的結果支持本研究第一項假設：台華共通詞的使用比例，學術類多於非學術類。

（二）覆蓋率 80%

表 21　台華共通詞與台語特別詞使用差異統計表（詞型，覆蓋率 80%）

項目	學術類		非學術類	
	詞型	比例	詞型	比例
台華共通詞	1,030	82.40%	1,088.5	65.69%
台語特別詞	220	17.60%	568.5	34.31%
合計	1,250	100.00%	1,657.0	100.00%

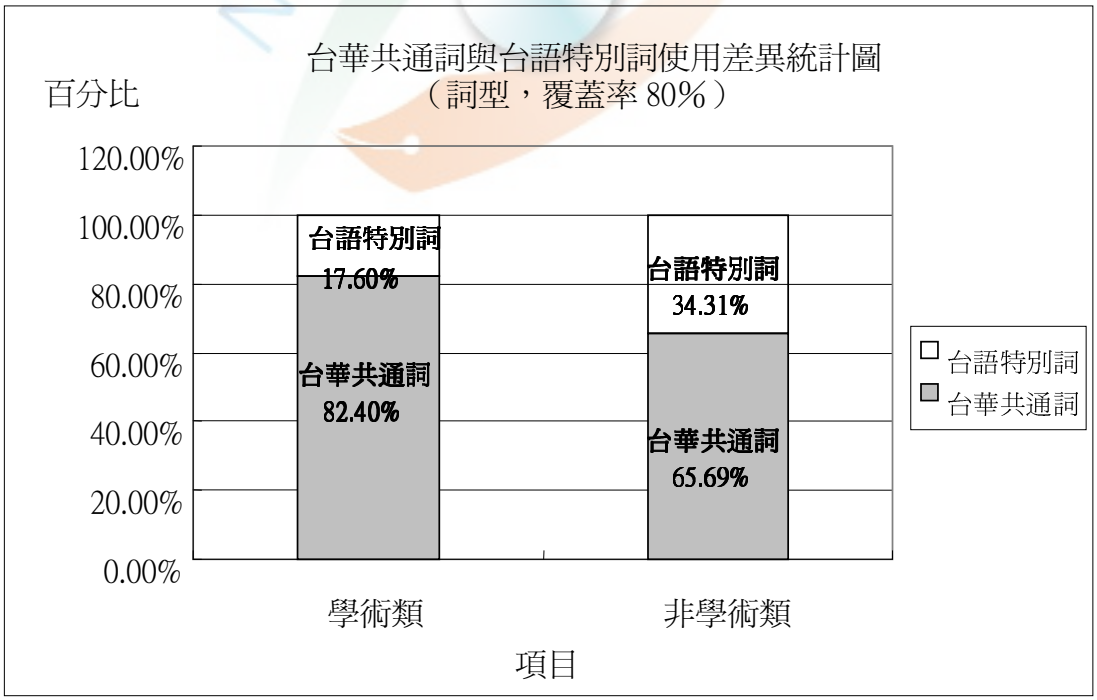

圖 4　台華共通詞與台語特別詞使用差異統計圖（詞型，覆蓋率 80%）

從上表和上圖可知：台華共通詞在學術類佔 82.40%，非學術佔 65.69%，學術類比非學術類多出 16.71%。兩種文類所佔的比例都超過 65%，學術類達到 80%以上的水準。統計的結果支持本研究第一項假設：台華共通詞的使用比例，學術類多於非學術類。

（三）覆蓋率比較分析

本小節透過比較台華共通詞不同覆蓋率的比例，分析低頻詞對統計結果的影響；並與鄭良偉（1987）、黃宣範（1998）的研究比較，探討台華共通詞實際的使用情形。

表 22　台華共通詞覆蓋率 100%VS 覆蓋率 80%比較表（詞型）

項目	覆蓋率 100%	覆蓋率 80%
學術類	66.91%	82.40%
非學術類	55.09%	65.69%

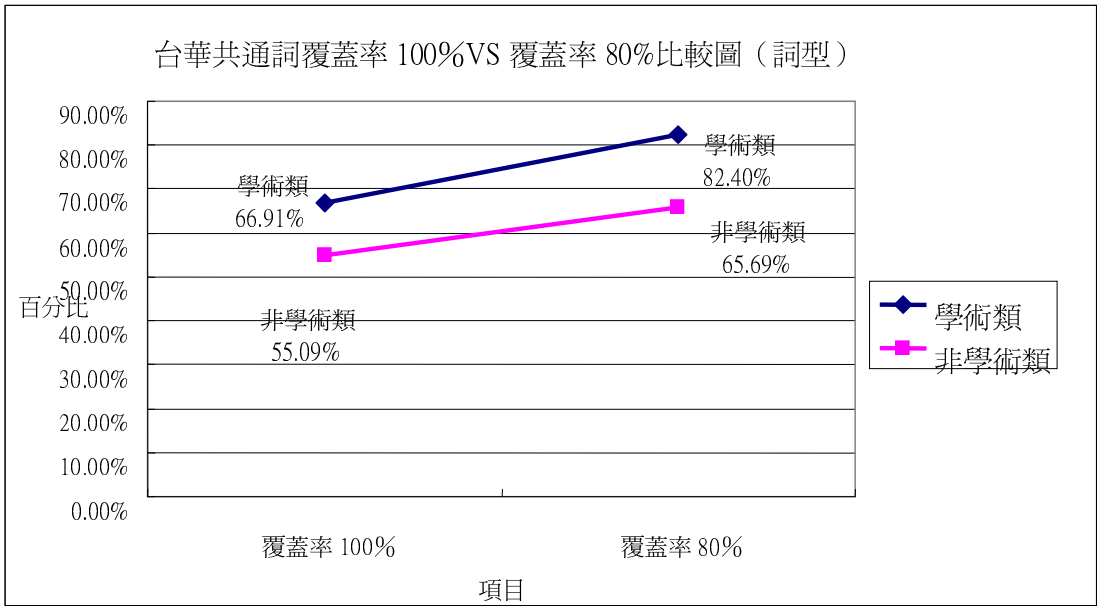

圖 5　台華共通詞覆蓋率 100%VS 覆蓋率 80%比較圖（詞型）

從上表及上圖可知：排除低頻詞後（即覆蓋率 100%調整至覆蓋率 80%），台華共通詞在學術類的比例由（覆蓋率 100%）66.91%提升至（覆蓋率 80%）82.40%，增加 15.49%；在非學術類的比例由（覆蓋率 100%）55.09%提升至（覆蓋率 80%）65.69%，增加 10.60%。

由上述可知，計算台華共通詞在學術類與非學術類的比例時，低頻詞對於統計結果會產生影響，影響的程度至少超過 10%。可能的原因是低頻詞中的電腦雜訊是以台語特別詞計算，而且以詞型計算，低頻詞和高頻詞的比重是一樣的，因此增加了低頻詞對統計結果的影響程度。

將兩個不同覆蓋率的台華共通詞使用比例與鄭良偉（1987）、黃宣範（1998）的統計做比較，探討台華共通詞的使用情形。比較結果如下表：

表 23　台華共通詞使用比較表（詞型）

項目	鄭良偉	黃宣範	覆蓋率 100%		覆蓋率 80%	
			學術	非學術	學術	非學術
台華共通詞	65.6%	73%	66.91%	55.09%	82.40%	65.69%
台語特別詞	34.4%	27%	33.09%	44.91%	17.60%	34.31%

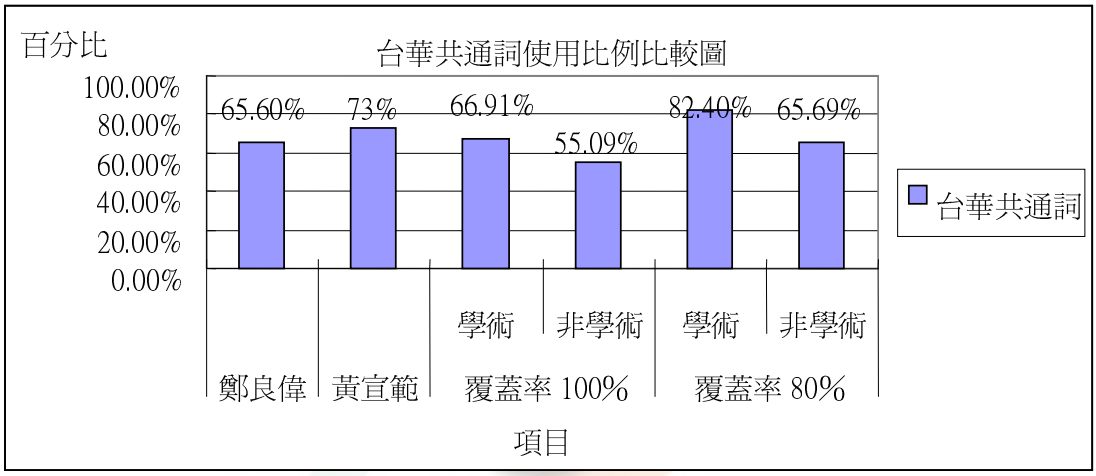

圖 6　台華共通詞使用比例比較圖

從上表及上圖可知：鄭、黃兩位學者台華共通詞的比例約介於 66%~73%；本研究台華共通詞覆蓋率 100%的比例介於 55.09%（非學術）~66.91%（學術），台華共通詞覆蓋率 80%的比例介於 65.69%（非學術）~82.40%（學術）。

鄭、黃兩位學者統計「現代閩南語辭典」台華共通詞是沒有分文類，普遍性的調查。如果以鄭、黃兩位學者的統計爲比較基準，學術類的文本對書面語標準化的程度要求比較高，使用台華共通詞的比例應該會高於鄭、黃兩位學者的調查結果；非學術性文本含蓋的文體種類較廣，使用台華共通詞的比例應該和鄭、黃

兩位學者的調查差不多。因此，我們可以歸納出一個簡單的比較方式，了解那一個台華共通詞的覆蓋率比較接近實際的使用情形，爲比較方便，以數學符號「＞」表示「大於、多於」，以「＝」表示「等於」。敘述如下：

由上述的推論可知，學術類文本＞鄭、黃≧非學術類文本是比較接近台華共通詞實際的使用情形，因此以台華共通詞覆蓋率 100%與覆蓋率 80%與之比較就可以很容易看出那一個覆蓋率是比較符合實際的使用情形。

根據上述方式比較台華共通詞覆蓋率 100%、覆蓋率 80%，結果如下：

覆蓋率 100%：鄭、黃（66%~73%）≧學術類文本（66.91%）＞非學術類文本（55.09%）。

覆蓋率 80%：學術類文本（82.40%）＞鄭、黃（66%~73%）≧非學術文本（65.69%）。

由上述的比較可知，台華共通詞覆蓋率 80%的統計結果（65.69%~82.40%）是比較符合實際的使用情形。

（四）小結

覆蓋率 100%和覆蓋率 80%的統計結果皆支持本研究的假設：台華共通詞的使用比例，學術類多於非學術類。

低頻詞對統計結果支持本研究假設方面沒有造成影響；但是在計算台華共通詞的比例時，對於學術類或非學術類大約有 10%以上的影響，排除低頻詞影響後（覆蓋率 80%）所得到的結果是比較接近實際的使用情形。

二、詞次分析

分成覆蓋率 100%、覆蓋率 80%、覆蓋率比較分析以及小結四個部分：

（一）覆蓋率 100%

表 24　台華共通詞與台語特別詞使用差異統計表（詞次，覆蓋率 100%）

項目	學術類		非學術類	
	詞次	比例	詞次	比例
台華共通詞	43,301	74.58%	44,289	56.81%
台語特別詞	14,756	25.42%	33,666	43.19%
合計	58,057	100.00%	77,955	100.00%

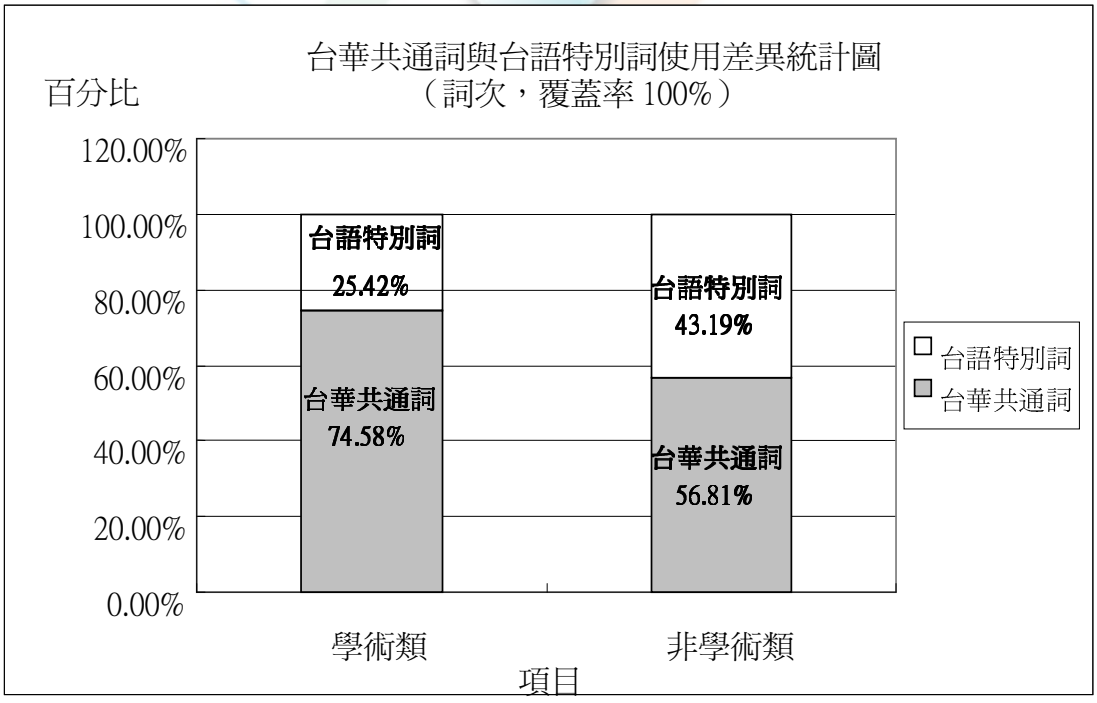

圖 7　台華共通詞與台語特別詞使用差異統計圖（詞次，覆蓋率 100%）

從上表和上圖可知：台華共通詞在學術類佔 74.58%，非學術佔 56.81%，學術類比非學術類多出 17.77%；兩種文類所佔的比例都超過 55%，學術類達到 70%以上的水準。統計的結果支持本研究第一項假設：台華共通詞的使用比例，學術類多於非學術類。

（二）覆蓋率 80%

表 25　台華共通詞與台語特別詞使用差異統計表（詞次，覆蓋率 80%）

項目	學術類		非學術類	
	詞次	比例	詞次	比例
台華共通詞	35,469	75.98%	36,197	57.01%
台語特別詞	11,215	24.02%	27,298	42.99%
合計	46,684	100.00%	63,495	100.00%

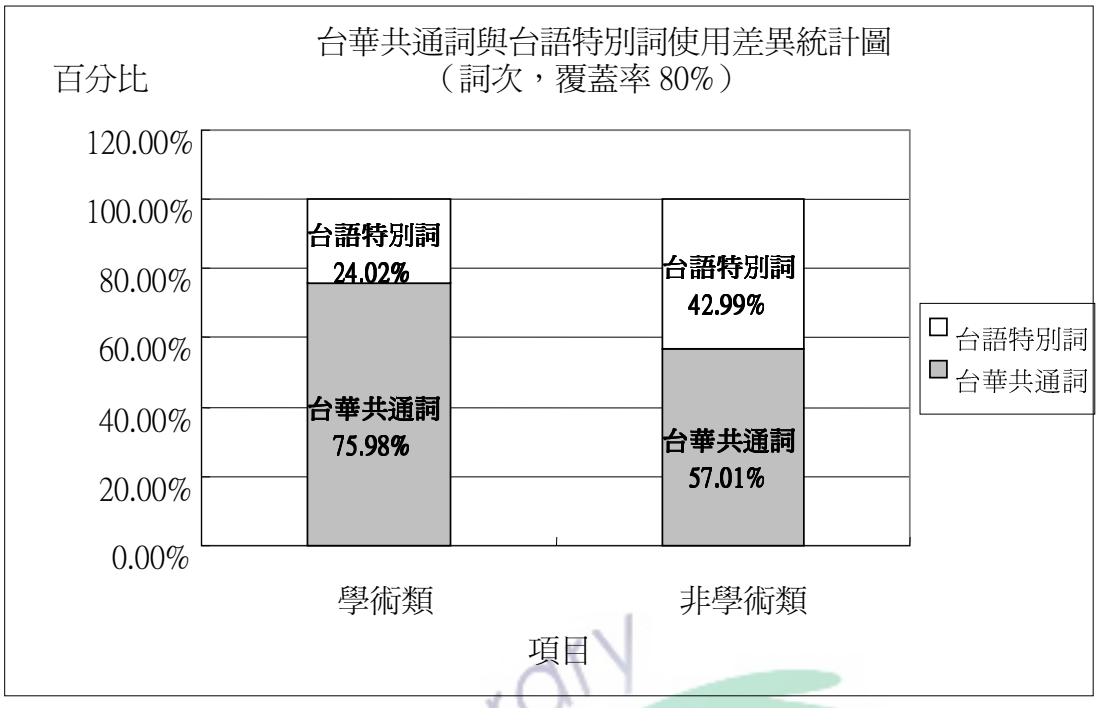

圖 8　台華共通詞與台語特別詞使用差異統計圖（詞次，覆蓋率 80%）

從上表和上圖可知：台華共通詞在學術類佔 75.98%，非學術佔 56.68%，學術類比非學術類多出 19.30%；兩種文類所佔的比例都超過 55%，學術類達到 75%以上的水準。統計的結果支持本研究第一項假設：台華共通詞的使用比例，學術類多於非學術類。

（三）覆蓋率比較分析

表 26　台華共通詞覆蓋率 100%VS 覆蓋率 80%比較表（詞次）

項目	覆蓋率 100%	覆蓋率 80%
學術類	74.58%	75.98%
非學術類	56.81%	57.01%

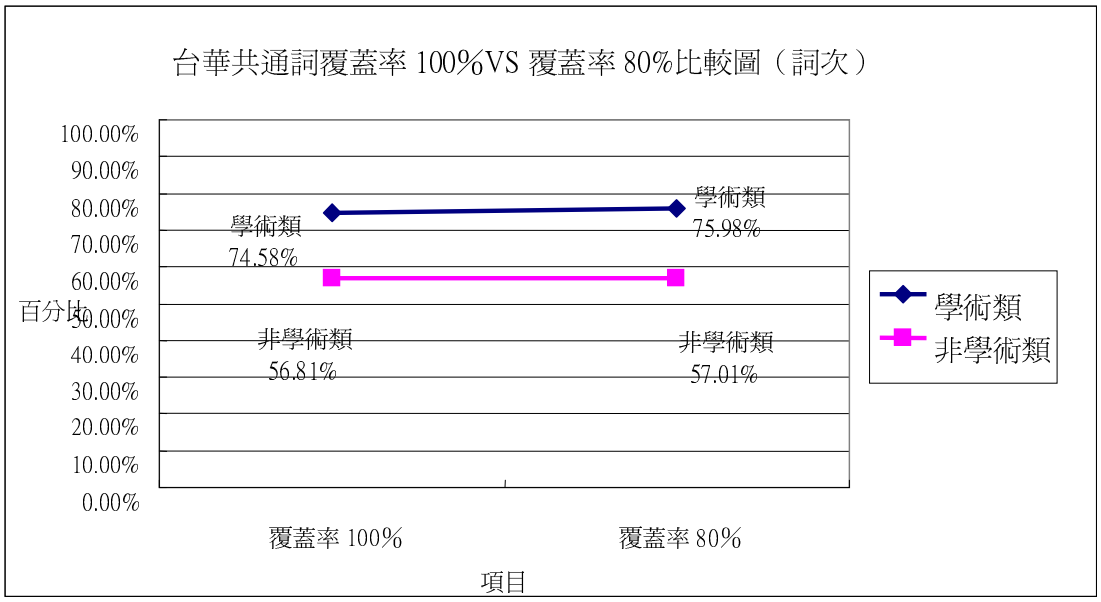

圖 9　台華共通詞覆蓋率 100%VS 覆蓋率 80%比較圖（詞次）

從上表及上圖可知：排除低頻詞後（即覆蓋率 100%調整至覆蓋率 80%），台華共通詞在學術類的比例由（覆蓋率 100%）74.58%提升至（覆蓋率 80%）75.98%，增加 1.4%；在非學術類的比例由（覆蓋率 100%）56.81%提升至（覆蓋率 80%）57.01%，增加 0.2%。

由上述可知，低頻詞在計算台華共通詞的比例時產生的影響很小。可能的原因是低頻詞中電腦雜訊雖然以台語特別詞計算，在總詞次中所佔比例不高，因此產生的影響有限。

將統計結果與 Sander 及 Hsieh （1981）的研究做比較，探討台華共通詞的使用情形，比較結果如下表：

表 27　台華共通詞使用比較表（詞次）

項目	新聞報告（散文）	可愛的仇人（小說）	情歌（詩）	諺語及俗語	學術 覆蓋率 80 %	非學術 覆蓋率 80 %
共通語層	77.0%	67.5%	67.4%	67.5%	75.98%	57.01%
本土語層	23.0%	32.5%	32.6%	32.5%	24.02%	42.99%
合計	100.0%	100.0%	100.0%	100.0%	100.00%	100.00%

資料來源：Sander 及 Hsieh （1981）；轉引自張學謙（ 1998）。

說明：1.本研究之台華共通詞和 Sander 及 Hsieh 研究之共通語層是指同一概念的詞彙；台語特別詞和本土語層亦同。

2.覆蓋率 80%是比較接近實際的使用情形，因此以之做爲比較。

從上表可知：Sander 及 Hsieh 新聞報告共通語層佔 77%、可愛的仇人佔（小說）67.5%、歌詩佔 67.4%、諺語及俗語佔 67.5%。新聞報告是屬於比較正式的文體，其他三類屬於比較非正式的文體；共通語層的使用比例，新聞報告（正式）多於小說、歌詩、諺語及俗語（非正式）。因此 Sander 及 Hsieh 的調查支持本研究的第一項假設：台華共通詞的使用比例，學術類多於非學術類。

再來比較各類文體台華共通詞的比例，新聞報告和學術類同屬於比較正式的文體，因此兩類文體台華共通詞所佔的比例差不多（77%和 75.98%）。非學術類和小說、歌詩、諺語及俗語同屬於非正式的文體，理論上台語特別詞的使用比例應該也是差不多，但是 Sander 及 Hsieh 調查的三類文體皆比本研究少了約 10%。可能是 Sander 及 Hsieh 的調查樣本數較少，單一作者個人的用詞習慣所造成的差異。

本研究非學術台語特別使用比例的統計結果（42.99%），和鄭良偉預估的使用比例差不多：日常會話、詩歌、俗語有時佔文章的一半左右，散文大約佔百分之二十到四十之間（轉引自張學謙 1998）。

（四）小結

覆蓋率 100%和覆蓋率 80%的統計結果皆支持本研究的假設：台華共通詞的使用比例，學術類高於非學術類。低頻詞沒有影響統計結果支持本研究假設，在計算台華共通詞的比例時也沒有太大的影響。

三、詞型 VS 詞次比較分析

詞型是文本中詞彙的類型，詞次是文本中詞型的實際使用次數或頻率；本研究透過比較兩者的差異探討台華共通詞詞型與詞次實際使用的情形。

以下分成覆蓋率 100%、覆蓋率 80%兩部分討論。

（一） 覆蓋率 100%

表 28　台華共通詞詞型 VS 詞次比較表（覆蓋率 100%）

項目	詞型	詞次
學術類	66.91%	74.58%
非學術類	55.09%	56.81%

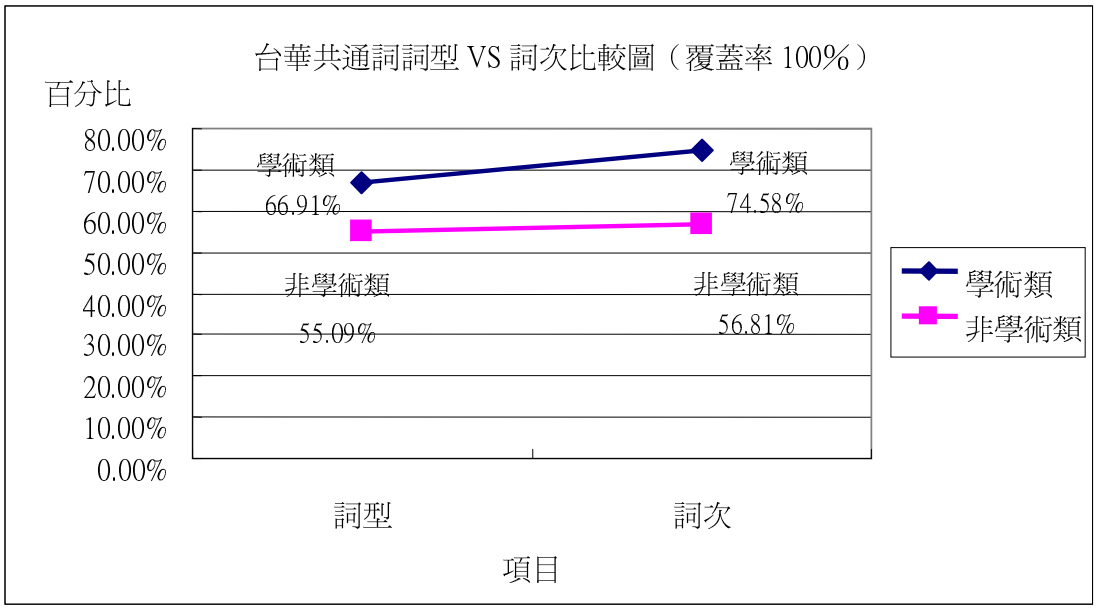

圖 10　台華共通詞詞型 VS 詞次比較圖（覆蓋率 100%）

從上表及上圖可知：學術類詞次比詞型多 7.67%；非學術類詞次比詞型多 1.72%。

學術類與非學術類詞次多於詞型的原因可能是：

1.低頻詞的影響：低頻詞包含許多電腦雜訊，統計時是以台語特別詞計算，因此會低估台華共通詞詞型與詞次的比例，但是對詞型的影響遠大於對詞次。

2.台華共通詞在文本中實際的使用頻率比較高。

因爲電腦雜訊是以台語特別詞計算，若以覆蓋率 80%（排除電腦雜訊後）來看，台華共通詞的詞型應該會大幅提升，詞次提升的幅度應該會低於詞型。

（二）覆蓋率 80%

表 29　台華共通詞詞型 VS 詞次比較表（覆蓋率 100%）

項目	詞型	詞次
學術類	82.40%	75.98%
非學術類	65.69%	57.01%

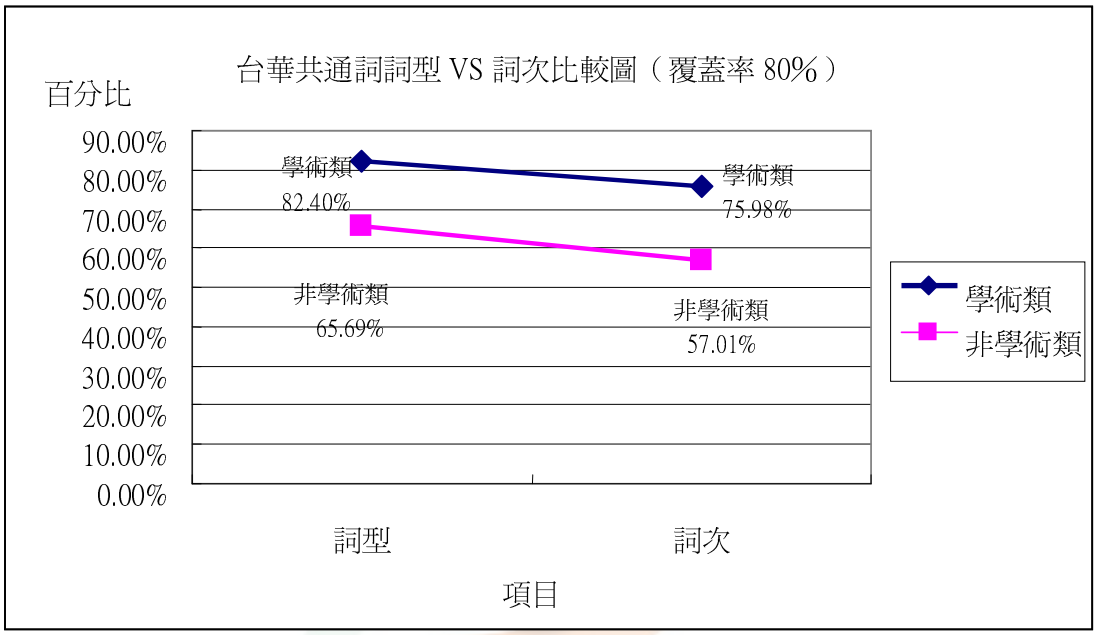

圖 11　台華共通詞詞型 VS 詞次比較圖（覆蓋率 80%）

從上表及上圖可知：學術類詞型比詞次多 6.42%；非學術類詞型比詞次多 8.68%；亦即台華共通詞在實際文本中的使用比例低於詞型。覆蓋率 80%已是排除電腦雜訊影響後的統計結果，因此，原因可能是一部分的台華共通詞是屬於使用頻率低的詞彙，非學術類又比學術類有更多的低頻詞。

四、小結

本研究分別以詞型和詞次以及不同覆蓋率探討台華共通詞的使用比例，並與

鄭良偉（1987）、黃宣範（1998）以及 Sander 及 Hsieh（1981）的調查做比較，結果皆支持本研究的第一項假設：台華共通詞的使用比例，學術類多於非學術類。低頻詞並不會影響統計結果支持本研究假設，但是在計算台華共通詞詞型的比例時會有影響的程度比較大，排除低頻詞後，以覆蓋率 80%的計算結果比較符合實際的使用情形。台華共通詞詞型的比例高於詞次，也就是說台華共通詞有一部分可能是屬於低頻詞。

第二節　台語詞彙豐度分析

本研究詞彙豐富度的計算方式爲：詞型÷詞次，分別計算學術與非學術類的詞彙豐富度後比較其中的差異。低頻詞中包含有電腦雜訊，因此本研究計算五個不同覆蓋率的詞彙豐富度，降低電腦雜訊對研究結果的影響。

研究的結果五個覆蓋率的詞彙豐富度都是學術類多於非學術類，沒有支持本研究假設；因爲差距不大，比例都低於 0.25%，因此研究者另外計算此差距在非學術類佔有多少比例（即差異比），以探討假設沒有得到支持的原因，計算方式如下：

$$\text{差異比} = \frac{\text{學術類 WR} - \text{非學術類 WR}}{\text{非學術類 WR}}$$

說明：WR 爲 word richness 的縮寫，即詞彙豐富度。

詞彙豐富度與差異比的統計結果如下表與下圖：

表 30　詞彙豐富度比較表

項目	覆蓋率 100%	覆蓋率 95%	覆蓋率 90%	覆蓋率 85%	覆蓋率 80%
學術類	11.81%	7.17%	4.77%	3.44%	2.63%
非學術類	11.65%	7.00%	4.62%	3.26%	2.40%
差異	0.16%	0.17%	0.15%	0.18%	0.23%
差異比	1.37%	2.40%	3.24%	5.56%	9.58%

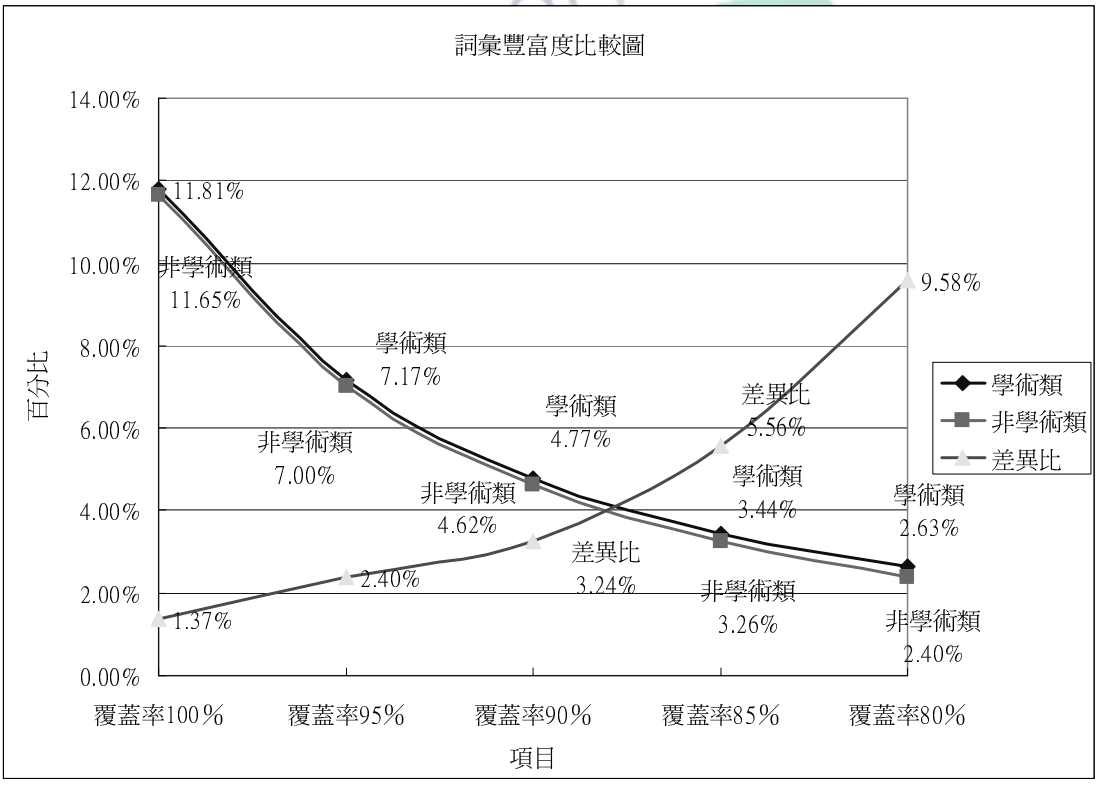

圖 12　詞彙豐富度比較圖

從上表及上圖可知：五個不同覆蓋率的詞彙豐富度都是學術類多於非學術類，相差的比例都低於 0.25%。統計結果並沒有支持本研究的第二項假設：台

語詞彙豐富度非學術類多於學術類。

本研究假設雖然並沒有得到支持，但是五個覆蓋率的差異都不大（差距都少於 0.25%，差異比介於 1.37%~9.58%），本研究試圖分析假設沒有得到支持的原因：

一、本研究已用不同覆蓋率計算的方式將電腦雜訊可能造成的影響降低，因此統計結果未能支持研究假設，初步可以排除電腦雜訊的因素。

二、樣本文體種類：本研究非學術類文體僅有小說、散文、劇本三類，未能包含詩、傳記、演講稿等其他文體，可能是研究假設沒有得到支持的因素之一。

三、樣本內部差異：《風俗通》（7.89%）詞彙豐富度多於《孟子》（5.40%）[36]，文本雖然分屬不同時代、且《孟子》出於多人手，但是差異比達到 46%，初步可知小說和口語兩種文體存在一定程度的差異。劇本在內容上，口語佔有相當的比例；因此，小說、散文和劇本的內部差異，可能是研究結果未能支持假設的因素之一。

四、樣本語料量：本研究的語料量約 20 萬音節（學術、非學術約各 10 萬），可能是少數特殊的文本影響研究結果，以致研究假設未能得到支持。

以上是研究未能得到支持的可能因素，如能排除文本內部差異，廣泛蒐集各種文體，並且有更多的語料量再行研究，對於學術類與非學術類的詞彙豐富會有更全面性的了解。

第三節　台語羅馬字詞彙分析

[36] 見本研究第二章第三節台語詞彙豐富度部分。

本節分成詞型與詞次兩個部分探討台語羅馬字詞彙在學術類與非學術類的使用差異。

一、詞型

表 31　台語羅馬字詞彙使用差異統計表（詞型）

項目	學術	非學術
詞型個數	861	1,882
詞型總數	6,857	9,083
百分比	12.56%	20.72%

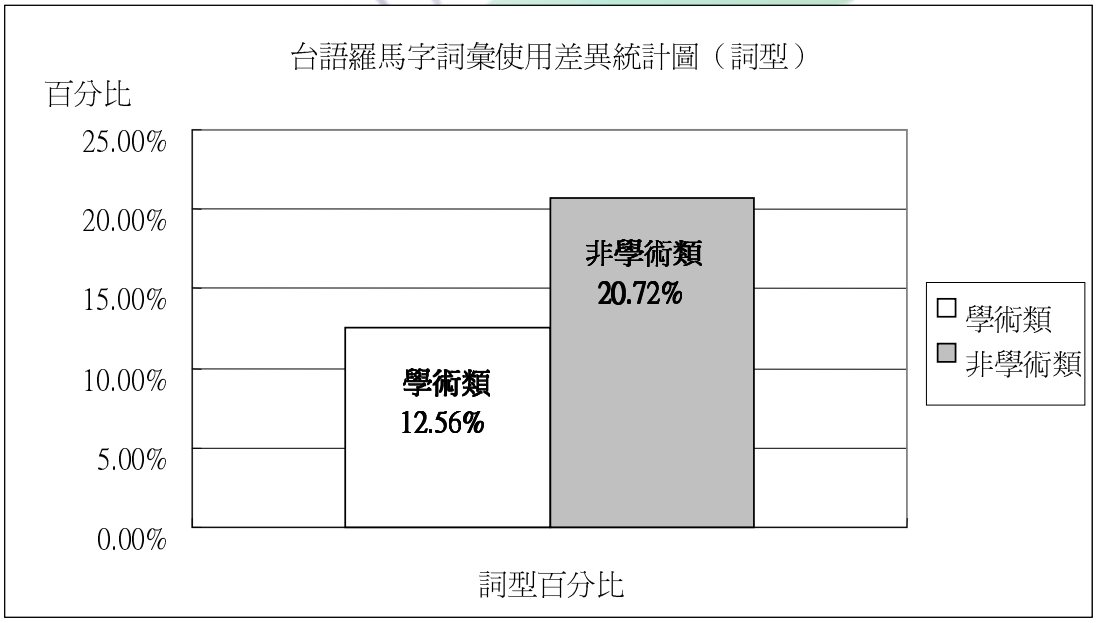

圖 13　台語羅馬字詞彙使用差異統計圖（詞型）

從上表及上圖中可知：學術類台語羅馬字詞彙的使用比例佔 12.56%，非學術類台語羅馬字詞彙的使用比例佔 20.72%，非學術類比學術類在台語羅馬字詞

彙的使用上多 8.16%。兩種文類使用台語羅馬字詞彙的比例皆超過 10%以上，非學術類達到 20%以上的水準。

上述的統計結果支持本研究的第四項假設：假設台語羅馬字詞彙使用比例非學術類高於學術類。

二、詞次

表 32　台語羅馬字詞彙使用差異統計表（詞次）

項目	學術	非學術
詞次個數	9,073	23,725
詞次總數	58,057	77,955
百分比	15.63%	30.43%

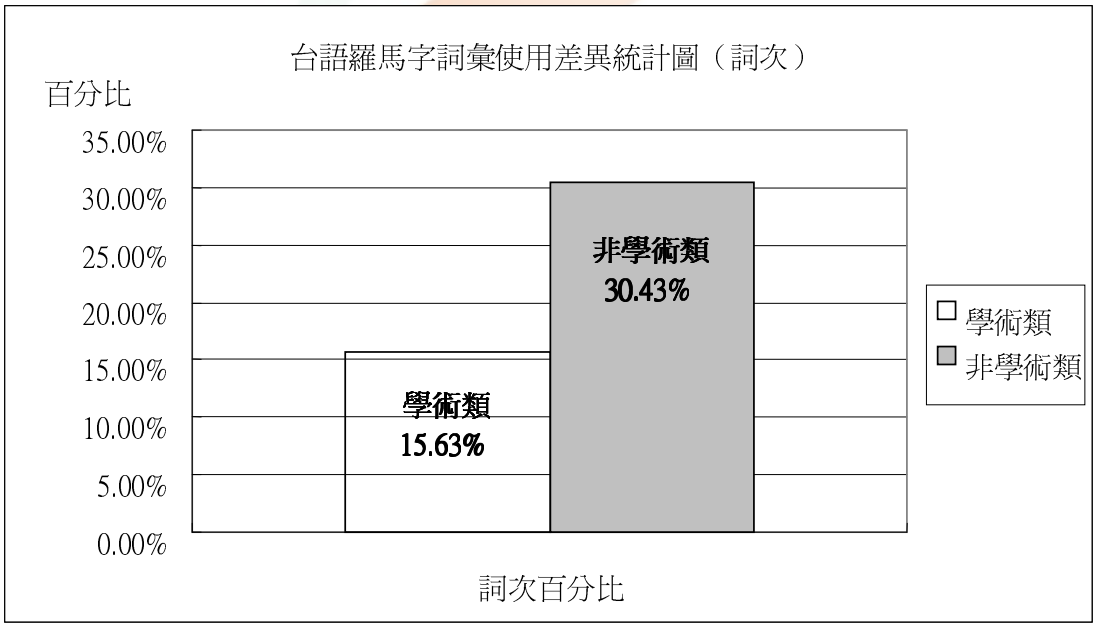

圖 14　台語羅馬字詞彙使用差異統計圖（詞次）

從上表及上圖中可知：學術類台語羅馬字詞彙的使用比例佔 15.63%，非學術類台語羅馬字詞彙的使用比例佔 30.43%，非學術類比學術類在台語羅馬字詞彙的使用上多出 14.80%。兩種文類使用台語羅馬字詞彙皆超過 15%以上，非學術類達到 30%以上的水準。

上述的統計結果支持本研究的第四項假設：台語羅馬字詞彙使用比例非學術類高於學術類。

三、詞型 VS 詞次比較分析

本研究透過比較詞型與詞次的差異探討台語羅馬字詞彙實際使用的情形。

表 33　台語羅馬字詞彙詞型 VS 詞次比較表

項目	詞型	詞次
學術類	12.56%	15.63%
非學術類	20.72%	30.43%

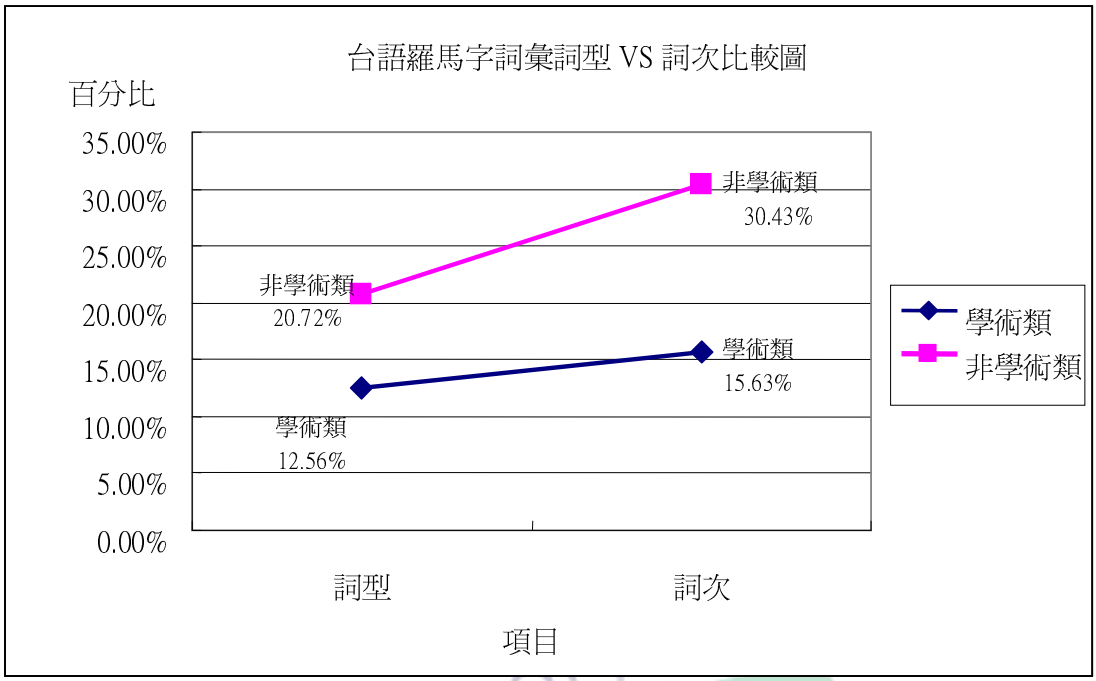

圖 15　台語羅馬字詞彙詞型 VS 詞次比較圖

從上表及上圖中可知：學術類詞次比詞型多 3.07%；非學術類詞次比詞型多 9.71%。可能的原因是：有一部分的台語羅馬字詞彙在文本中實際的使用頻率比較高；非學術類又比學術類有更多的高頻詞。

第四節　台語平均詞長分析

本研究計算平均詞長的方式爲：音節數÷詞次，分成四個部分比較學術類與非學術類的平均詞長：覆蓋率 100%、覆蓋率 80%、覆蓋率 100%VS 覆蓋率 80%、台語 VS 華語平均詞長比較分析。

一、覆蓋率 100%

表 34　台語平均詞長統計表（覆蓋率 100%）

項目	學術類	非學術類	台語詞彙
音節數	101,349	102,335	203,684
詞次	58,057	77,955	136,012
平均詞長	1.75	1.31	1.50

從上表可知：學術類平均詞長爲 1.75，非學術類平均詞長爲 1.31，學術類比非學術類多 0.44。統計結果支持本研究的第四項假設：台語平均詞長，學術類多於非學術類。

二、覆蓋率 80%

表 35　台語平均詞長統計表（覆蓋率 80%）

項目	學術類	非學術類	台語詞彙
音節數	66,617	77,863	144,480
詞次	46,452	62,368	108,820
平均詞長	1.43	1.25	1.33

從上表可知：學術類平均詞長爲 1.43，非學術類平均詞長爲 1.25，學術類比非學術類多 0.18。統計結果支持本研究的第四項假設：台語平均詞長，學術類多於非學術類。

排除低頻詞後的平均詞長比較符合實際情況。

三、覆蓋率 100%VS 覆蓋率 80%

表 36　台語平均詞長覆蓋率 100%VS 覆蓋率 80%比較表

項目	覆蓋率 100%	覆蓋率 80%
學術類	1.75	1.43
非學術類	1.31	1.25
台語詞彙	1.5	1.33

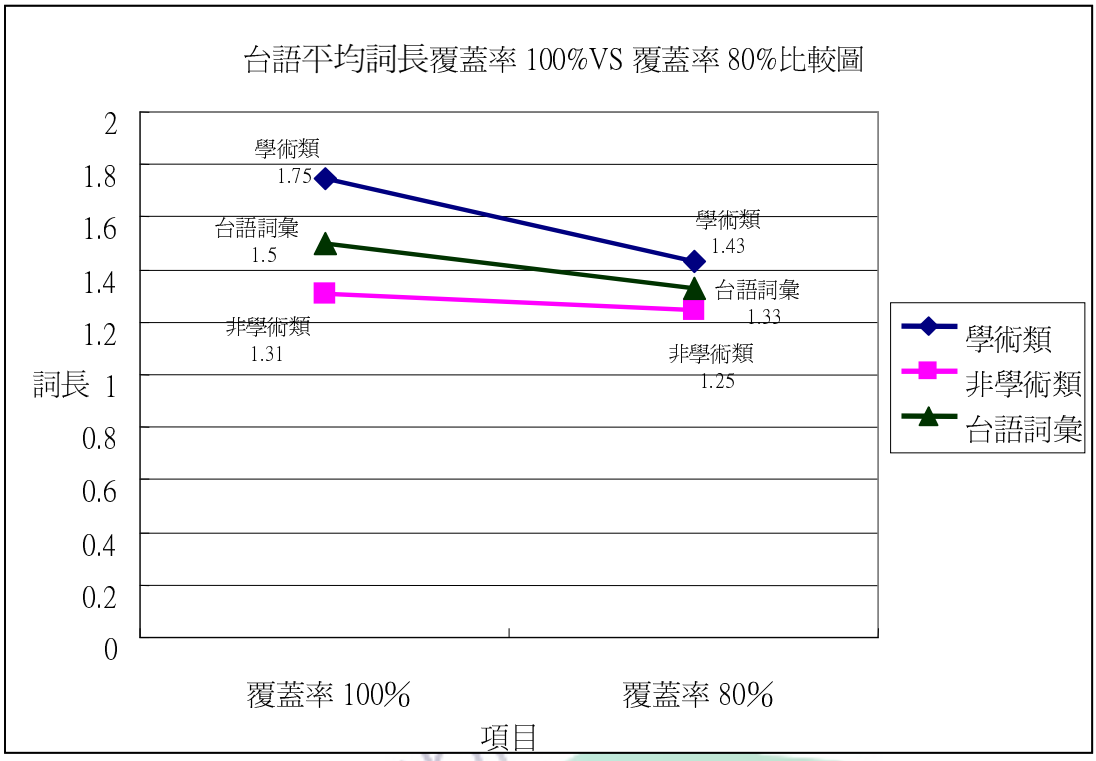

圖 16　台語平均詞長覆蓋率 100%VS 覆蓋率 80%比較圖

從上表及上圖可知：

（一）覆蓋率 100%與覆蓋率 80%的平均詞長都向下調整。

（二）平均詞長學術類多於台語詞彙多於非學術類。

四、台語 VS 華語平均詞長比較分析

上表台語平均詞長爲 1.50，此數據是以本研究的語料音節總數÷詞次總數計算得到的，雖然學術類與非學術類各文體並非以 1：1 的比例抽樣，亦未能涵蓋其他文體，但仍可做爲初步了解台語平均詞長的參考。茲將本研究數據與漢語做比較，以了解其間的差異，結果如下表：

表 37　台語 VS 華語平均詞長比較表

項目	台語詞彙 覆蓋率 80%	現代漢語詞語頻率辭典	兩岸三地漢語語料庫（台灣、中國、香港）
平均詞長	1.33	2.0928	2.2706

從上表可知：台語平均詞長：1.33；《現代漢語詞語頻率辭典》：2.0928；《兩岸三地漢語語料庫》：2.2706。平均詞長《現代漢語詞語頻率辭典》多於《兩岸三地漢語語料庫》多於台語詞彙。以上結果支持本研究以音節數多寡推論比較平均詞長的論點。

第五章　結論與建議

第一節　結論

本研究採用語料庫語言學的研究方法探討台語詞彙在學術與非學術類的使用差異，研究的語料共有約 20 萬音節左右，探討台華共通詞、台語詞彙豐富度、台語羅馬字詞彙以及台語平均詞長在學術類與非學術類書面語的情形。研究結論如以下幾點說明：

一、台華共通詞

台華共通詞的統計結果支持本研究假設：台華共通詞的使用比例，學術類多於非學術類。詞型、詞次、覆蓋率各項統計結果如下所述：

（一）詞型

1.覆蓋率 100%：學術類多於非學術類，多出 11.82%。

2.覆蓋率 80%：學術類多於學術類，多出 16.71%。

（二）詞次

1.覆蓋率 100%：學術類多於非學術類，多出 17.77%。

2.覆蓋率 80%：學術類多於非學術類，多出 19.30%。

二、詞彙豐富度

台語詞彙豐富度的統計結果：非學術類多於學術類，研究假設並沒有得到支持。可能是樣本的文體種類、內部差異、語料量等因素影響統計結果對於假設的支持。

三、台語羅馬字詞彙

台語羅馬字詞彙的統計結果支持本研究假設：台語羅馬字詞彙的使用比例，非學術類多於學術類。詞型、詞次的統計結果如下所述：

（一）詞型：非學術類多於學術類，多出 8.16%。

（二）詞次：非學術類多於學術類，多出 14.80%。

四、台語平均詞長

台語平均詞長的統計結果支持本研究假設：台語平均詞長，學術類多於非學術類。統計結果如下所述：

（一）覆蓋率 100%台語平均詞長：學術類比非學術類多 0.44 個詞。

（二）覆蓋率 80%台語平均詞長：學術類比非學術類多 0.18 個詞。

第二節　建議

台語著作目前已累積有多種不同文體與豐富語料，基於台語語料庫的研究是未來重要的發展，研究者以語料庫語言學的方法研究台語詞彙的使用差異，發現台語借用華語詞彙已經是時代的趨勢與事實，不同文體有借用比例上的差異。研

究過程中，研究者將所遇到的困難與限制，以及新發現許多值得深入探討問題，提出幾點建議做爲爾後相關研究與單位之參考：

一、台語詞彙豐富度本研究結果未能支持假設，未來若能蒐集更多樣的文體、建立充足樣本的語料量、排除樣本內部差異進行研究，相信會有完整的了解。

二、高頻詞與低頻詞的使用差異是研究過程中發現但未探討的問題，例如「去」在非學術語料中是屬於高頻詞（編號 14），但是在學術語料中卻不是（編號 80），這也是值得討論的問題。

三、一個華語詞對應兩個以上（含兩個）台語詞彙，例如：「可愛」（華語）與之對應的台語詞彙有「可愛（khó-ài）」、「古錐（ kó͘ -chui）」、「巧裝（khiáu-chng）」等同義詞，其使用情形亦是值得研究的問題。

四、研究者在做台華共通詞的比對時，發現中文詞庫(八萬目詞)重複出現一次的詞有 35 個之多，例如：邊、點、還、頭、調、數、種、當...等，這也是值得進一步了解的問題。

五、研究者在做詞彙斷詞時發現台語缺乏一套完善的電腦斷詞系統以及分詞規範，這是台語語料庫語言學研究的基本問題。

六、籲請政府積極投注資源建立一個公開的加工台語書面語語料庫，提供相關領域使用，以提昇研究品質與擴展應用範籌，並且保存台語免於流失。


NTTU Library

參考文獻

專書

Anthony McEnery, Richard Xiao, Yukio Tono .2006. Corpus-based language studies : An advanced resource book .London : Routledge.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. Corpus Linguistics ： Investigating language structure and use. London：Cambridge University Press.

Kennedy, Graeme D . 1998. An introduction to corpus linguistics . London : Longman.

David Crystal 著，方晶譯。1995。《劍橋語言百科全書》。北京：中國社會科學。

David Crystal 著，周蔚譯。2001。《語言的死亡》。台北市：貓頭鷹出版社。

王育德。2000。《台語入門》台北市：前衛出版社。

王育德。2000。《台灣話講座》。台北市：前衛出版社。

王育德。2002。《台灣語常用語彙》。台北市：前衛出版社。

何大安。1996。《聲韻學中的觀念和方法》。台北市：大安出版社。

周長楫。1996。《閩南話的形成發展及臺灣的傳播 》。台北市：台笠出版社。

林慶勳。2001。《臺灣閩南語概論》。台北：心理出版社。

俞士汶。2003。《計算語言學概論》。北京：商務印書館。

施炳華。2001。《行入台語文學的花園》。台南市:真平企業有限公司。

張裕宏。2001。《白話字基本論：台語文對應＆相關的議題淺說》。台北：文鶴出版社。

張學謙。2003。《行向多文字 ê 台語文：文字態度 kap 政策論文集》。屏東：睿煜出版社。

教育部國語推行委員會。1999。《常用語詞調查報告書》。台北市：教育部。

許極燉。1991。《台灣話流浪記》。臺北市:臺灣語文研究發展基金會。

許極燉。1994。《台語文字化的方向》。台北市：自立晚報社文化出版部。

許極燉。1998。《台灣語概論》。台北：前衛出版社。

湯志祥。2001。《當代漢語詞語的共時狀況及其嬗變》。上海：復旦大學出版社。

詞庫小組。1996。《「搜」文解字------中文詞界研究與資訊用分詞標準技術報告NO.96-01》。中央研究院資訊科學研究所中文詞知識庫小組。

黃昌寧、李涓子著。2002。《語料庫語言學》。北京：北京商務印書館。

黃宣範。2004。《語言、社會與族群意識－台灣語言社會學的研究》。台北：文鶴出版社。

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

楊秀芳。1998。《閩南語字彙（一）》。教育部。

楊秀芳。1998。《閩南語字彙（二）》。教育部。

楊秀芳。2005。《台灣閩南語語法稿》。台北市：大安出版社。

楊惠中主編。2002。《語料庫語言學導論》。上海：上海外語教育出版社。

董忠司主編。1996。《臺灣閩南語概論講授資料彙編》。台北：臺灣語文學會。

董忠司編。2001。《福爾摩沙的烙印：臺灣閩南語概要》。台北市：行政院文建會。

蔣爲文。2005。《語言、認同與去殖民》。台南：國立成功大學。

蔣爲文。2007。《語言、文學 kap 台灣國家再想像》。台南：國立成功大學。

鄭良偉。1989。《走向標準化的台灣話文》。台北：自立晚報文化出版部。

鄭良偉。1990。《演變中的台灣社會語言－多語社會及雙語教育》。台北：自立晚報文化出版部。

研討會及期刊論文

Hung,Jia-Fei, Cherry Li, and Jane Tsay. 2004. The Child's utterance final particles in Taiwnaese: a case study. Taipei:National Taiwna University. Proceedings of the IsCLL-9, 477-498.

李勤岸、洪惟仁。2007。〈沒有名字的語言？－「台灣話」、「閩南話」還是 Hoh-lo 話〉。《臺灣文學館通訊》15：36- 41。

林香薇。2003。〈論宋澤萊台語詩《一枝煎匙》的用字與用詞〉。《師大學報》48，2：95-118。

姚榮松。1990。〈當代台灣小說中的方言詞彙－兼談閩南語的書面語〉。《國文學

報》19：223-264。

姚榮松。1990。〈閩南語書面語的漢字規範〉。《教學與研究》12：77-94。

姚榮松。2000。〈臺灣閩南語歌仔冊的用字分析與詞彙解讀－以《最新落陰相褒歌》爲例〉。《國文學報》29：193-230。

姚榮松。2004。〈閩南語漢字書寫的檢討與文字化的方向〉。《國文天地》19，10：81-86。

洪嘉馡、黃居仁。2008。〈語料庫爲本的兩岸對應詞彙發掘〉。《語言暨語言學》9，2：221-238。

高照明。2002。〈中英雙語近譯句翻譯檢索系統〉。《翻譯學研究集刊》。台北：台灣翻譯學會。7：75-107。

張學謙、楊允言。2005。〈台語語體演變分析〉。行政院國家科學委員會補助專題研究計畫成果報告，計畫編號：NSC 93-2411-H-143-005。

張學謙。1998。〈Ho-Lo 台語虛詞的語層及語用〉。載於《第二屆台灣語言國際研討會論文選集》。451-463。

張學謙。2000。〈台語口語及書面語體的多面向分析〉。《語言暨語言學》1，1：89-117。

張學謙。2004。〈弱勢語言的地位規劃與語言復振：從語言歧視主義到語言公平〉。載於《語言人權與語言復振學術研討會》。台東：國立台東大學。61-76。

陳瑞清。2003。〈語料庫翻譯學：英漢翻譯的研究與應用初探〉。載於《第八屆

口筆譯教學研討會論文集》。台北：台灣翻譯學會。7-1 至 7-28。

曾金金。1997。〈台語斷詞原則〉。《台灣文學出版物收集、目錄、選讀編輯計畫結案報告說明》。台北：行政院文化建設委員會。45-72。

黃居仁。1990。〈計算語言學– – 人工智慧、語言學、認知科學的結合〉。《科學月刊》21，4：281-287。

黃居仁。1997。〈科技整合與整合科技---談計算語言學與語料庫語言學之角色與發展〉。《中央研究院計算中心通訊》13，23：257-262。

黃居仁。2006。〈大數與求真：如何以十億字語料庫進行語言分析與研究〉。載於《第四屆台灣華語文教學研討會 2006 論文集》。國立高師範大學華語文教學研究所。1-5

黃宣範。1988。〈台灣話構詞論〉。載於《現代台灣話研究論文集》。鄭良偉、黃宣範主編。台北市：文鶴出版有限公司。121-146。

黃郁純、陳薾宇。2005。〈以語料庫爲本分析詞語搭配及近義關係〉。《華語文教學研究》2，2：57-71。

楊允言、張學謙。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〉。行政院國家科學委員會補助專題研究計畫成果報告。NSC93-2213-E-122-001。

楊允言、劉杰岳、李盛安、高成炎。2004。〈語言流失 kap 變化 e 探討---以台語新約聖經做例〉。載於《語言人權與語言復振學術研討會論文集》。台東：

國立台東大學。237-249。

楊允言、劉杰岳。2006。〈台語文計算語言學基礎建設—介紹台語線頂辭典 kap 語料庫〉。載於《第一屆台灣語文暨文化研討會論文集》。台中：中山醫學大學。168-179。

楊允言。1993。〈台語文字化个過去佮現在〉。《台灣史料研究》1：57-75。

楊允言。2003a。〈從語域及借詞觀點探討台語文寫作風格〉。載於《第 15 屆計算語言學研討會論文集》。新竹：清華大學。73-86。

楊允言。2003b。〈台文華文線上辭典建置技術及使用情形探討〉。載於《第三屆全球華文網路教育國際學術研討會論文集》。台北：圓山大飯店。132-141。

楊允言。2003c。〈近年來台語文資料處理的成果與展望〉。《大漢學報》18：71-78。

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〉。《臺灣文學館通訊》15：42-44。

楊允言。2008。〈台語文數位典藏---以台語文記憶系統做例〉。載於《第三屆台灣語文暨文化研討會》台中：中山醫學大學。119-136。

楊秀芳。1995。〈閩南語書寫問題平議〉。《大陸雜誌》90，1：15-24。

劉賢軒。2005。〈應用語言學論文中的態度成份〉。《清雲學報》25，1：295-305。

蔡素娟。2004。〈台灣兒童語料庫 III〉。行政院國家科學委員會補助專題研究計畫結案報告，計畫編號：NSC91-2411-H-194-029.

蔡素娟。2007。〈閩南語語料庫的建構與相關問題〉。載於《語言政策的多元文

化思考》。鄭錦全等編。台北市：中央研究院語言研究所。357-372。

鄭良偉。1984。〈台語能以漢字書寫嗎？（上）〉。《臺灣風物》34，4：111-128。

鄭良偉。1985。〈台語能以漢字書寫嗎？（下）〉。《臺灣風物》35，1：133-154。

鄭錦全。1998。〈從計量理解語言認知〉。載於《漢語計量與計算研究》。鄒佳彥、黎邦洋、陳偉光、王士元主編。香港城市大學語言資訊科學研究中心。15-30。

鄧敏君。2005。〈語料庫輔助翻譯研究之方法論---中日、日中翻譯的應用〉。《翻譯學研究集刊》。台北：台灣翻譯學會。9：405-426。

盧慧娟、林柳村、白芳怡。2007。〈以語料庫爲本之語言教學應用研究--語詞搭配之分析〉。《外國語文研究》6：39-57。

盧慧娟。2006。〈以對比語料庫爲之「詞語搭配」研究〉。《淡江外語論叢》8：159-175。

謝佳玲。2006。〈漢語情態詞的語意界定：語料庫爲本的研究〉。《中國語文研究》21：45-63。

謝清俊。1985。〈計算的科學〉。《科學月刊》16，10：769-775。

謝清俊。1990。〈中文計算語言學的發展〉。《科學月刊》21，4：300-305。

碩博士論文

Hung, Yu-Jun. 2005. The Child's Acquisition of Verbs in Taiwanese. Chiayi：

National Chung Cheng University MA thesis.

王萸芳。1995。《漢語口語與書面語中副詞子句的訊息順序》。國立臺灣師範大學英語學系碩士班碩士論文。

余明憲。2005。《漢語之框觸動詞——「玩」、「弄」和「搞」在「動賓」格式中之研究》。國立交通大學外國文學與語言研究所碩士論文。

李珮甄。2007。《以口語語料庫爲本探究台灣閩南語「是講」、「著是講」的語用功能》。高雄師範大學台灣文化及語言研究所碩士論文。

李勤岸。2000。《Lexical Change and Variation in Taiwanese Literary Texts, 1916--1998 —A Computer-Assisted Corpus Analysis》。美國夏威夷大學語言學研究所博士論文。

陳珮嘉。2000。《漢語動詞單位詞與動詞搭配關係之初探》。國立臺灣師範大學華語文教學研究所碩士論文。

廖小婷。2003。《中文施力動詞『拉、拖、扯』之語意初探--以語料庫爲本的近義詞研究》。國立交通大學語言與文化研究所碩士論文。

蕭如卿。2006。《台灣閩南語量詞習得》。國立中正大學語言研究所碩士論文。

謝昌運。2007。《台語加強詞的研究：語料庫語言學的分析》。國立臺東大學語文教育學系碩士班碩士論文。

顏國仁。1995。《台語口語常用詞頻率調查初步報告》。國立清華大學統計研究所碩士論文。

網站

「白話字台語文網站」---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：http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp。2008/8/22。

「白話字台語文網站」---白話字台語文相關研討會：http://iug.csie.dahan.edu.tw/giankiu/GTH/gth.asp。2008/7/12。

「白話字台語文網站」---台語文語料庫建立蒐集計畫：http://iug.csie.dahan.edu.tw/TG/guliaukhou/。2008/8/22。

中文詞知識庫小組：http://ckip.iis.sinica.edu.tw/CKIP/20corpus.htm。2008/9/7。

中國國家語委現代漢語語料庫：http://www.clr.org.cn/retrieval/index.html。2008/9/7。

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

台語文 concordance 網站：http://iug.csie.dahan.edu.tw/TG/concordance/form.asp。2008/11/3。

台語文數位典藏資料庫（第二階段）：http://iug.csie.dahan.edu.tw/nmtl/dadwt/pbk.asp。2008/8/23。

教育部閩南語常用詞辭典試用版：http://twblg.dict.edu.tw/tw/index.htm。2008/11/3。

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

附錄一：學術類語料抽樣一覽表

編號	作 者	篇　名	日　期	研 討 會	音節數
1	張學謙	東是東，西是西，永遠 bē 相 tú？台灣人對台語文字 ê 態度研究	2002/7/14	2002 台灣羅馬字教學 kap 研究國際學術研討會	5,255
2	梁淑慧	「幼兒台語班」的教學實務 kah 成果	2002/7/14	2002 台灣羅馬字教學 kap 研究國際學術研討會	5,019
3	林雅雯	國小學生學習白話字 ê 課程安排及發現	2002/7/14	2002 台灣羅馬字教學 kap 研究國際學術研討會	5,028
4	蔡澄甫	白話字教學疏失的討探	2002/7/14	2002 台灣羅馬字教學 kap 研究國際學術研討會	5,096
5	楊允言	台語拼音符號 ê 競爭—以 TLPA kah 白話字做例	2002/7/14	2002 台灣羅馬字教學 kap 研究國際學術研討會	4,976
6	程俊源	台華語對應及轉換	2004/10/9	2004 台灣羅馬字國際研討會	3,669
7	方耀乾	台語白話文學 ê 起源 kap 發展：一個學界疏忽去 ê 存在	2004/10/9	2004 台灣羅馬字國際研討會	5,005

編號	作 者	篇　名	日　期	研 討 會	音節數
8	吳仁瑟	台語基督教詩歌文學初探	2004/10/9	2004 台灣羅馬字國際研討會	4,417
9	高成炎 李盛安 陳豐惠	台語文文化推廣網站建構發展計畫	2004/10/9	2004 台灣羅馬字國際研討會	4,575
10	陳錦玉	衝破殖民烏影、迎接台灣主體文化 ê 日光—論台灣多語言教育 ê 實踐方案	2004/12/18	語言人權與語言復振學術研討會	5,271
11	張復聚	台灣母語教學的基本思考	2004/12/18	語言人權與語言復振學術研討會	3,453
12	呂興昌	憑什麼台語？爲什麼文學？：台語文學的新思考	2005/10/29-30	2005 台語文學學術研討會	4,725
13	蔣爲文	語言、文學 Kap 民族國家 ê 建構：台語文學運動史初探	2005/10/29-30	2005 台語文學學術研討會	5,145
14	陳恒嘉	狂歡吧！諸神！許丙丁《小封神》的狂歡詩學	2005/10/29-30	2005 台語文學學術研討會	4,991
15	廖瑞銘	在地異鄉人 ê 詩情 ：論陳明仁《流浪記事》以後 ê 詩作	2005/10/29-30	2005 台語文學學術研討會	5,100
16	李勤岸	蔡培火白話字散文集《十項管見》ê 關鍵：論述中 ê 譬	2005/10/29-30	2005 台語文學學術研討會	4,958

編號	作 者	篇　名	日　期	研 討 會	音節數
		論			
17	陳慕真	母親、母土 kap 母語：鄭雅怡台語作品中 ê 女性書寫	2005/10/29-30	2005 台語文學學術研討會	5,078
18	呂美親	召喚苦澀原汁，延長賞味期限：論析清文小說〈虱目 á ê 滋味〉中失落語詞 ê 積極意涵	2005/10/29-30	2005 台語文學學術研討會	4,972
19	鄭良偉	母語文學 tī 母語教育中所扮演的角色	2006/9/9-10	2006 台灣羅馬字國際學術研討會	5,035
20	何信翰	“主導” ê 觀念 kap 伊 tī台語詩研究頂 koân ê運用	2006/9/9-10	2006 台灣羅馬字國際學術研討會	5,039
21	林俊育	探討台語文學書寫 ê 電腦工具	2007/10/6-7	2007 台語文學學術研討會--中山醫學大學第二屆台灣語文 kap 文化學術研討會	4,542
計					101,349


NTTU Library

附錄二：非學術類語料抽樣一覽表

編號	作　者	篇　　名	年代	類型	音節數
1	Voyu Taokara 劉	目睭	2006	小說	3,670
2	吳昭新譯	蜘蛛之絲	2006	小說	1,829
3	陳廷宣	憨子婿	2003	小說	617
4	陳德樺	駛計程仔 ê 永仔	2003	小說	1,511
5	林長昇	人間百合_05	2003	小說	1,830
6	台灣羅馬字協會	伊索寓言_04	2002	小說	2,424
7	鄭詩宗	菜 chhoah 面 ê 故事	2001	小說	1,547
8	Pasuya	我 beh tńg 來種作 ！	2001	小說	3,799
9	Abon	魚肉	2000	小說	4,243
10	張聰敏	阿瑛！啊_11	1999	小說	3,155
11	Taibunun	勇敢 ê Aukele_02	1997	小說	2,708
12	Loonng	OPERA 內 ê 魔神仔_04	1996	小說	2,399
13	許惠悰	天光前 ê 戀愛故事	1996	小說	5,334
14	莊惠平	化學品 ê 玄機_02	1996	小說	4,819
15	蔡承維	富戶人 ê 歷史	1996	小說	4,614
16	楊允言譯	海鼻新娘_02	1991	小說	3,509

編號	作　者	篇　　名	年代	類型	音節數
17	鄭碧燕	對話	2006	散文	460
18	陳威志	種菜過日子	2006	散文	1,380
19	Noya	Satsuma 精神	2006	散文	966
20	吳宜璘	Siuⁿ Chē	2006	散文	733
21	Sakoas	煌--á	2005	散文	933
22	黃淑惠	老人退休 beh án-chóaⁿ	2005	散文	852
23	王寶漣	Kavalan a-má	2004	散文	2,846
24	陳怡君	永遠 ê 櫻花	2004	散文	537
25	陳昭宏	Hông 犧牲 ê 農民	2004	散文	527
26	林俊育	電腦 ê 性別	2004	散文	326
27	楊振裕	酒友煉仙	2004	散文	577
28	葉信志	請老師 hō͘學生正確 ê 觀念	2004	散文	614
29	蔡 Sek-gî	聰明 vs 愚戇	2004	散文	249
30	大目降	語言接觸 ê 笑話	2003	散文	582
31	張學謙	時鐘時間無 kâng	2003	散文	132
32	鄭雅怡	教台語受迫害 ê 經驗	2003	散文	4,407
33	鄭良光	學北京話	2003	散文	72
34	鄭玉鳳	Siàu 念 A 公	2003	散文	1,211

編號	作　者	篇　　名	年代	類型	音節數
35	陳星旭	公民入籍考試	2003	散文	214
36	陳柏壽,黃真救	Lín pē ê 事件	2003	散文	469
37	陳俊宏	危機	2003	散文	458
38	孫健智	Hioh-kôaⁿ	2003	散文	1,176
39	王嚥	A 公 kap a-má ê 故事(1)	2003	散文	1,464
40	黃真救	學漢文	2003	散文	224
41	賴青松	Tāi-sam 出世 Hit 1 工	2003	散文	4,121
42	林清祥	『發球』kap “fxck you”	2003	散文	342
43	廖立文	想 boeh hō͘ A-pâ 知影我心目中 ê A-pâ	2003	散文	1,641
44	廖麗雪	肛溫	2003	散文	194
45	劉杰岳	號英文名	2003	散文	195
46	許隼夫	阿公，我 beh 唱歌	2003	散文	232
47	江澄樹	緊張老母條直子	2003	散文	192
48	葉永德	SARS 麻雀間	2003	散文	275
49	吳清江	阮丈人	2003	散文	172
50	蔡其達	無「夠長」無愛 kap 你結婚	2003	散文	170
51	洪隆建	褪 (thǹg) 口罩	2003	散文	308
52	丁連宗	牛乳糖仔 chhák 破椰子殼	2002	散文	980

編號	作　者	篇　　名	年代	類型	音節數
53	陳宇碩	Tī 高等教育體制中台語教育 ê 重要性	2002	散文	1,449
54	陳玲玉	數念所消失 ê 台灣農村風景	2002	散文	371
55	陳慕真	任務	2002	散文	1,004
56	Makatau Loan	論翻譯	2002	散文	2,289
57	盧永芳	Sorting	2002	散文	1,651
58	林宜慧	Góa ê 電話簿 á	2002	散文	465
59	林艾萱	故鄉 ê A-pà	2002	散文	414
60	李和惠	我心內所想	2002	散文	245
61	Hoanya kc	已經 tiàu 遠 ê 日子	2002	散文	2,244
62	吳柏鴻	來去宜蘭	2002	散文	2,665
63	顏信星	阿爸 ê 遺產	2002	散文	912
64	蔡雅祺	入秋 ê 淡水街 á	2002	散文	884
65	張復聚	保存地球 ê 重要資源：語言	2001	散文	1,462
66	郭文卿	江湖答嘴鼓	2004	劇本	3,699
67	蔡愛義	話劇林學恭牧師	1966	劇本	4,342
計					102,335