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%