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、陳藹宇（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)、《泼根》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。

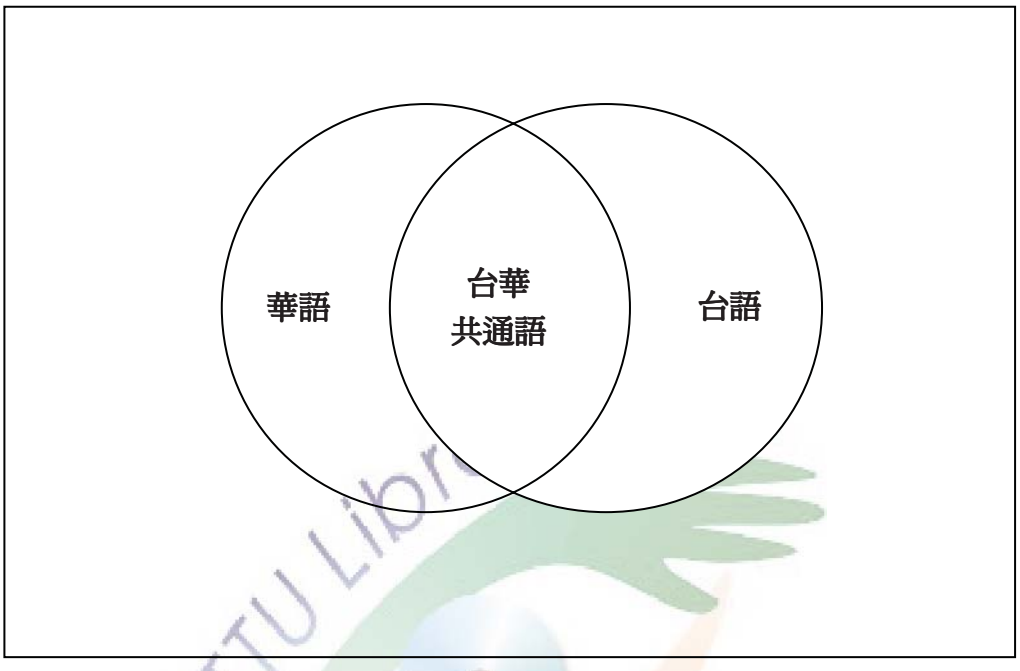

圖 1　台華共通語示意圖

（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