臺灣四縣及大埔客家語詞頻比較研究

江俊龍 著

高雄復文圖書出版社 印行

臺灣四縣及大埔客家語詞頻比較研究

江俊龍 著

高雄復文圖書出版社 印行

國家圖書館出版品預行編目資料

臺灣四縣及大埔客家語詞頻比較研究 / 江俊龍著. -- 初版. -- 高雄市 : 高雄復文, 2009.08
面 ; 公分
參考書目：面
ISBN 978-957-555-980-9（平裝）

1. 客語 2. 詞頻 3. 比較研究

802.5238 98012999

臺灣四縣及大埔客家語詞頻比較研究

著　者：江俊龍
發行人：蘇清足
出版者：高雄復文圖書出版社
地址：802 高雄市苓雅區五福一路 57 號 2 樓之 2
電話：（07）226-5267
傳真：（07）226-4697
登記證：局版台業字第 1804 號
版　次：2009 年 8 月初版一刷
ISBN：978-957-555-980-9

定　價：350 元

臺灣四縣及大埔客家語詞頻比較研究

摘要

本文蒐集臺灣四縣腔(區分北四縣及南四縣)和大埔腔客家語語料共256,199個詞，進行詞頻比較。客家語詞頻研究所牽涉的問題比國語或其他語言複雜，在文字書寫方面的問題尚未解決之前，從事書面文獻的詞頻研究時，首先將面臨語料中同音、同義、但不同字形的判別問題，這當中也可能因為次方言的差別，而導致同義但不同音、不同字形的狀況。縱使上述問題都得以解決，現有客語書面文獻的種類，不外乎山歌、童謠、俗諺、故事，以及少數詩、散文、小說、劇本的創作，這些語料恐怕無法反映客家語言的實際面貌。況且，國語或其他語言的語料收集，可將報刊、雜誌，國防、科技等專業報告都納入研究取材範圍，使成果更形周延客觀，這方面的客家語文獻則付之闕如，使得詞頻統計的解釋性有所侷限。因此，本文的語料來源，除了文獻蒐集而得之外，還從事訪談及田野調查，以補文獻之不足。

接著要解決的問題包括：客家語分詞原則的擬定、文字書寫的異用處理以及訂定現階段客家語的詞類標記種類。

本文最後所呈現的成果包括：客家語詞頻的分級、分類分腔分級詞的比較、虛詞詞頻比較分析、華語與客家語詞頻比較分析以及客家語詞頻覆蓋率的比較分析等。

關鍵詞：客語、詞頻、比較研究

臺灣四縣及大埔客家語詞頻比較研究

目錄

圖表目錄

第一章　緒論

臺灣具有許多不同的族群與文化，各自擁有不同的語言特徵，過去因爲過度強調國語的主體性，造成漠視其他語言的現象，已經有許多族群的語言流失，重建及挽救都需要花費大量的心力。臺灣客家語有四縣腔、海陸腔、大埔腔、饒平腔、詔安腔等種類，其中四縣腔又有南北之別，雖然祖籍來源背景相似，但由於使用區域不同，發音、用詞都各有其特殊性，但是差異性究竟有多大，一直都礙於有限的人力與資源，欠缺全面性的比較工作。本文以四縣、大埔腔做爲觀察對象，建立客家語語料庫，以詞頻統計的方式，進行分析比較工作。[1]

第一節　研究動機

近年來，臺灣在客家語方面的學術建置工作，取得了豐碩的研究成果，但是在教育推廣、語言傳承上的成果卻顯得有限。成效不彰的原因並不是活動辦得不夠多，也不是欠缺教材可運用，而是整個社會環境是否真正的多元包容，弱勢族群對自己的母語是否自重自尊等氛圍的問題。要想真正建立多元尊重的價值觀，並且讓弱勢族群對自身的母語產生信心與信任，提供科學化的驗證成果，讓社會各界廣泛運用，發揮深遠的影響力，是學術界所能貢獻的部分。以客家語推廣和傳承而言，建置客家語語料庫，提供科學化的詞頻研究成果，因應各種研究所需；協助教材的有效編寫，讓九年一貫的客家語教材和各種客家語書籍的編寫及創作，都有其學理上的依據和公信力，讓使用者及學習者產生信賴感，就是提振母語的利器。

第二節　研究目的

客家研究和客家語的推廣工作，近年來在臺灣顯得相當蓬勃，民間與官方各種版本客家語教材紛紛出爐，傳統歌謠、俗諺等材料亦廣爲蒐羅整理，在教育推廣及文獻保存上，已有相當的成績。爲能更有效的推廣客家語，行政院客家委員會自 2005 年起，年年舉辦「初級客語能力認證考試」，更從 2008 年起，每年辦理「中級暨中高級客語能力認證考試」，工程相當浩大。

此外，教育部也已著手規畫部編本客家語教材，編纂適合各腔調、各階

[1] 感謝行政院國家科學委員會、客家委員會及教育部國語推行委員會在研究經費上的大力支持，使得臺灣客語語料庫的建置及詞頻研究工作得以大步向前。

段的分級教材，供學習者使用。目前「教育部部編版客家語分級教材編輯小組」已經開始運作，預定將於 2011 年 12 月 31 日完成編輯工作。然而在編輯詞目、撰寫教材的過程中，雖經再三研修、去蕪存菁，卻苦無客家語詞語使用頻率之數據可供參考依循，設若有客家語詞頻的分析數據可資對照援引與應用，相信在編纂相關教材或從事語言研究時，可以提供不小的助益。

本文著眼於當前國家教育長遠之需要，彙整四縣客語及大埔客語各類文獻語料，並且進行田野調查、口語訪談來收集語料，增加語料的多樣性，使統計成果能更貼近實際語言現象，以利詞頻研究的成果應用。

第三節　文獻探討

語料庫的建置工作，國外已經進行多年，其中以英語語料庫爲最早，類型也最多。一般而言，語料的類型可以分成「口語語料」及「書面語語料」兩大類。另外，根據需求的不同，可以依據地區、年代、內容、文體、發音人性別……等等進行不同的區分。賴惠玲(2008:153)[2]曾對各國語料庫的建置，做了概略的介紹：

> 除英文外，許多其他語言也紛紛建立語料庫，包括德語、義大利語、西班牙語、瑞典語、葡萄牙語、俄語、荷蘭語、威爾斯語、波斯尼亞語、保加利亞語、以色列語、塞爾維亞語、日語、泰語和中文。同樣有以書寫語料為主的，例如義大利文語料庫（A corpus of spoken Bulgarian）；也有包括詞類和語法分析的，如德語語料庫（NEGRA Corpus）

現代漢語詞頻研究，以中國大陸北京語言學院《現代漢語頻率詞典》發其端。這本歷時約六年的詞典編纂（1979/11-1985/7），不但爲語言學開拓了新的研究領域，同時更具實用及應用價值，對各有關學科的發展產生了一定的貢獻。近年來，中國大陸已制定國家標準 GB/T 13715-92《信息處理用現代漢語分詞規範》，在中文信息處理上取得了豐碩的成果。

臺灣的詞頻研究工作，多半由官方進行分年的調查，且針對國語語料進行。其中包括「84-87 年常用語詞調查工作」、「87 年口語語料調查工作」、「87 年口語問卷調查工作」、「大陸小學教科書字詞調查工作」、「國小學童常用字

[2] 見賴惠玲(2008)〈客語語法研究議題的開發：以語料庫爲本〉，收錄於行政院客家委員會《96 年度補助大學校院暨獎助客家學術研究計畫成果發表會論文集》台北：行政院客家委員會，2008 年 6 月。

詞調查工作」、「國語辭典簡編本編輯資料字詞頻統計報告」等，有效地反映出年度口語語料及辭典或教科書的詞頻狀況。中央研究院語言學研究所也已完成「現代漢語平衡語料庫」、「近代漢語語料庫」、「上古漢語語料庫」等，其中「現代漢語平衡語料庫」，除了進行文字語料的彙整、斷詞、詞類標記之外，還提供詞頻統計數據，呈現國語詞彙的使用頻率，對於臺灣語言使用的現象分析，提供龐大的觀察資料，在教學研究和學術研究上可謂成果豐碩。

臺灣閩南語語料庫的成果，依序有鄭良偉、楊允言在 2000 年開始建置的「台文華文線上辭典」；李勤岸、林俊育於 2005 年完成的「台語摘譯台日大辭典」；楊允言於 2003 年開始建置的「台語語料庫」等。鄭良偉、楊允言「台文華文線上辭典」分別收錄漳州、泉州腔調的語料，具有中文、羅馬字的查詢功能，並且提供聲音輸出。在檢索的結果中，採取並列的方式呈現，另設有網路連結，可以連結至中央研究院「現代漢語平衡語料庫」、楊允言「羅漢台語文」等備有例句語料庫的功能，提供詞條例句。李勤岸、林俊育「台語摘譯台日大辭典」備有以台北腔爲主的台語羅馬字、其他腔調、台語漢羅、語詞的台語解釋、例詞或例句等，其中在台語羅馬字的部份，備有圖片顯示的功能，讓未安裝台語羅馬字型的電腦，可以正常顯示。楊允言「台語文語料庫」提供羅漢、全羅等台語語詞檢索功能，此資料庫會將檢索的詞彙及其前後文排列出來，讓使用者可以學習詞彙的用法。

至於客家語方面，有教育部「臺灣客家語常用辭典」、行政院客委會「客家四縣語音辭典」、台北市客委會「現代客語詞彙彙編」、臺灣大學客家社「客語小辭典」和「客語有聲字典」等電子工具書。在詞頻研究上，有 2004 年羅肇錦教授應教育部之託，統計出部分客家文獻資料的詞頻；2006 年則有鍾榮富教授指導研究生劉秀珍探討《客家語教科書常用詞彙與詞頻之初步研究－－以高市版爲例》，以及羅肇錦、呂菁菁兩位教授指導研究生謝杰雄撰寫的《語料庫的建置與臺灣客家話 VP 研究》等兩篇碩士論文。在客語語料庫的建置方面，賴惠玲教授廣收現有的客家語書面語語料，並且收錄客家電視台的口語語料，陸續建置「客語篇章資料庫」、「客語歇後語資料庫」、「客語口語資料庫」等等，成績相當可觀。另外，筆者亦在 2007 年至 2008 年之間，執行行政院國科會計畫《臺灣高屏地區客語口語詞頻研究》、行政院客家委員會計畫《臺灣客家話詞頻研究（I）－東勢地區大埔客語之文獻語料分析》和《臺灣客家話詞頻研究（II）－－四縣客語之文獻語料分析》、教育部國語推行委員會計畫《臺灣客家口語文獻《東勢客語故事集》詞頻分析》進行詞頻統計工作。

第四節　研究方法

客家語詞頻研究所牽涉的問題比國語複雜，在文字書寫的問題尚未解決之前，從事書面文獻的詞頻研究時，語料中同音、同義、但不同字形的判別，是最先面對的問題，例如：人的嘴巴，四縣客語音[tsoi55]，大埔客語音[tʃoi53]，音韻條件相當，但書寫形式無論在哪一種腔調內都有「嘴」、「喙」等不同的寫法。另一種情況是因爲次方言的差別，而導致同義但不同音、不同字形的狀況，例如：四縣客語「蘿蔔」對應大埔客語「菜頭」，其實所指相同，是客家語中的同義詞。縱使上述問題都得以解決，現有客家語的文獻種類，不外乎山歌、童謠、俗諺、故事，以及少數詩作、散文、小說、劇本的創作，不像國語或其他語言的語料收集，可將報刊、雜誌，國防、科技等專業報告納入研究取材範圍，使得客語詞頻統計的解釋性有所侷限，只能反映客家語言的部分面貌。

本文將北四縣客語、大埔客語、南四縣客語做爲取材對象，鎖定大埔客語爲範圍的原因，主要是因爲大埔爲純客區、臺灣大埔客分布地最爲集中(臺中東勢地區)、臺灣大埔客語的內部一致性最高，很適合做爲客家語初次詞頻研究的對象。四縣腔客語因爲分布區域的關係，有南北之別，在語音、詞彙上已有差異，對於南北四縣腔客語的差異比較研究，是值得關注的焦點，所以具有進行比對工作的必要。

關於各個腔調語料的字形取捨，各自有不同的考量。其中大埔客語的文獻語料分析，取材自民間文學《東勢鎮客語故事集》(共七冊)，爲臺灣大埔客語的口語語料。文字的校對工作，有請原採錄者徐登志女士進行，有效還原語料的原始樣貌，以徐女士的訂正版本做爲分詞、統計對象。在四縣語料方面，以龔萬灶先生所撰寫的《阿啾剪个故鄉》做爲北四縣的分析文本，語料性質屬於書面語。因爲龔先生亦屬教育部「臺灣客家語書寫推薦用字」小組的一員，用字較爲講究，所以北四縣《阿啾箭个故鄉》的文獻語料盡可能保持原有樣貌。南四縣方面則蒐集廖金明先生撰寫的故事（書面語）、劇本（口語），以及實地訪談語料（口語），做爲分析文本。因爲廖先生的用字參雜許多個人習慣，例如否定副詞「無」、「毋」二字，皆採用「唔」字書寫，而「唔」字非約定俗成的書寫文字，所以採用廖先生的口語語料、書面語語料時，將斟酌修改書寫字形。其修改原則與實地訪查資料的字形校對原則相同，盡可能參酌教育部公布的字形，進行修正工作。

經過謝杰雄先生授權，本文使用謝先生所研發的「SanHak 分詞系統」，進行詞條切分、詞類標記等作業。不過，在技術執行上，有電腦字碼相容性的問題產生。目前世界通行的字碼有 Big5 及 Unicode 兩種，兩種字碼並不相

容，所以現行的軟體，只能選用其中一種字碼做爲程式運作的字型，這對客語用字有很大的影響。例如「个」、「𠊎」、「咩」等字爲客家語語料常見用字，但是都爲 Unicode 字碼，如果誤入以 Big5 爲字型碼的軟體中，不但無法辨識，甚至會影響軟體的運作，造成錯誤的斷詞及標記。設計軟體以 Big5 做爲字型碼，所以語料中不能出現「个」、「𠊎」、「咩」的字型。另外，因爲電腦軟體的中文字型是因應國語而設計的，而客語用字出現了許多國語的罕用字，並未收錄於現行的電腦字型中，因此，必須灌裝新的字型檔才能辨識客語用字；縱使灌裝新字型，在某些狀況下，程式運作仍會出現錯誤訊息，造成電腦操作的困難。爲了因應上述狀況，使各式等級電腦都能順利操作，本文暫時使用 Big5 字形，替代 Unicode 字形，例如將結構助詞「个（Unicode 字碼）」替代爲「介（Big5 字碼）」；如果字形可以進行拆字，原則以拆字方式處理，例如「[上亡下口]、[左手右突]、[左女右哀]、[左手右亥]、[左口右林]、[左女右麻]」等字，若構字元素具有罕用字形，拆字亦無法解決字形碼及軟體需求，則以音標方式處理，例如「then3 手」。

這些語料經過校對用字之後，還要運用詞彙學的理論將語句拆解成一個個的詞，這樣的分析工作是另一個重點，必須有嚴謹的分析方法，才不會混淆詞語單位，使得電腦軟體可以順利作業，得到可用的統計數據，以利統計結果的詮釋。下一章裡，本文將詳論這個課題。

第二章　客家語「字」、「詞」、「詞組」的區分問題

因爲客家語的書寫牽涉到方言差異和本字選用的爭論，再加上拼音文字的觀點加入等因素，客家語書寫系統的確定，是必須面對的重要課題。此外，漢語「字」、「詞」、「詞組」的使用略有重疊之處，亦須釐清觀念的差異，才能有效進行客家語詞彙分詞原則的擬定。以下將分成「客家語的文字系統」、「客家語的「字」、「詞」和「詞組」的區分問題」、「關於分詞原則的討論」、「客家語的分詞原則」等四個小節，討論詞頻統計中書寫系統、分詞原則的問題。

第一節　客家語的文字系統

文字系統可以概分爲兩種：意音文字和拼音文字。漢字是意音文字。意音文字體系一般都包含三種不同類型的字形：表意字，借表意字充當表音字，以及兼用表意表音兩種方法的字。我國傳統的文字學把第二種稱爲假借字，第三種稱爲形聲字。[3]

意音文字系統和拼音文字系統孰優孰劣？從造字方法的歷史去考察，文字的發展似乎經歷了表意階段、表意兼表音階段、表音階段。漢字停留在意音文字階段，似乎不如拼音文字。拼音文字的確有簡明方便的優點，利用幾十個字母就可以拼寫出語言中所有的音節。但是「簡明方便」並不是我們考慮文字使用系統時的唯一標準，我們更須要衡量考慮漢語的特性。

做爲記錄語言的書寫工具，文字必須適應語言的結構特點。我們的每一個漢字基本上都紀錄了一個單音節的語素，所以，方塊漢字的設計，基本上符合漢語的結構特點。用方塊漢字所紀錄下來的文獻資料，不可勝數，雖經過時間的變遷、空間的分隔，語言產生了變體。但是古代漢語和現代漢語之間、共同語和方言之間的漢字系統並非全然走樣，毫無憑藉可資理解；這和拼音文字有很大的不同，拼音文字一旦與時代脫節，便難以閱讀。所以，漢字系統可以讓悠久的文化得以繼承，讓使用不同方言的人彼此之間思想得以交流，這正是漢字的優越性。

再者，拼音文字不符合漢語的特性，還會表現在詞和詞組的區分問題上，葉蜚聲、徐通鏘（1993：186-187）：

[3] 詳見葉蜚聲、徐通鏘《語言學綱要》頁 182。

> 一旦實現拼音化，這些原來用字形來區別的不同語素就無法識別，勢必會給語言文字的使用帶來麻煩和混亂。有些單音語素雖已不單獨使用，而與其他語素結合起來構成一個複合詞，這自然可以減少一些同音語素所帶來的麻煩，但又提出了一些新的問題：哪些是詞，應該連寫；哪些是詞組，應該分開書寫，等等。這些問題一時還解決不好。總之，漢語的實際狀況還難以實現拼音化。

大陸的普通話和臺灣的國語情況尚且如此，[4]客家語的文字系統尚未標準化，內部又存在著次方言的差異，拼音化只會帶來更多認讀和文化斷裂的問題，拼音文字「簡明方便」的優點在漢語和客家語上面難以發揮。所以，客家語的文字系統還是得朝漢字書寫的方向去努力，一步一步朝向標準化。至於拼音法，充其量只能做爲注音的工具，不能完全取代漢字的功能。

第二節　關於客家語「字」、「詞」和「詞組」的討論

「字」和「詞」雖有重疊，但基本上屬性不同。「字」是文字學的單位，是音和義的結合體，偶爾也被當純表音的記音符號，如音譯詞或聯綿詞中的方塊字。古代漢語的語詞以單音節爲主要的形式，周薦(2004：63-64)對趙誠編著的《甲骨文簡明詞典---卜辭分類讀本》(中華書局，1988 年)做了統計，發現：

> 該詞典收條總數為 2050 個，其中單字有 1589 個，約佔 77.51%，複字詞語有 461 個，約佔 22.49%。從殷商時代到現代這綿長的三千年左右的時間裡，單字在漢語詞彙中所佔的比例是以平均每百年超過 2%的速度遞減的，或者說多字詞語是以平均每百年超過 2%的速度遞增的。甲骨文不過是記錄漢語的較早的書寫形式，在甲骨文產生之前，漢語已存在了若干萬年。不難推定，遠古時期的漢語確是一種單音節語。

近代和現代漢語的語詞則以雙音節爲主要的形式，明清小說已清楚呈現此一發展趨勢，孫常敘(1956：156)對《水滸傳》、《紅樓夢》、《兒女英雄傳》

[4] 教育部 2007 年 3 月所公布的「臺灣閩南語羅馬字拼音方案」中，列有連字符使用原則：「凡辭典中列爲詞條者，應爲一個詞(word)，可以連字符『-』連結。如:『辦公』應標成 pān-kong。複合詞若可再分爲多音節詞(含雙音節)，宜加以斷詞，如:『Tsóng-thóng-hú pì-su (總統府秘書)』應斷詞爲:『Tsóng-thóng-hú (總統府)』與『pì-su(秘書)』」以辭典爲依據，事實上並沒有爲辭典學提供分詞或斷詞的標準。

三部小說做過統計，結論是：在這三部書中多字詞語和單字的比例分別爲70：30，64：36，67.5：32.5。周薦(2004：63)認爲：

> 如果我們能夠把這三部小說視作明清兩代漢語詞彙使用的代表，我們就可將上述統計數字加合起來得出一個總的百分比：在明清兩代的漢語詞彙中，多字語詞和單字的比例是67.17：32.83。

上一節提到拼音文字不適用於漢語，理由之一是：哪些是詞，應該連寫？哪些是詞組，應該分開書寫？這個問題會造成書寫表達時的困擾，必須嚴肅看待。採用漢字來做爲文字系統，就沒有連寫不連寫的問題。但是一個個的方塊漢字畢竟只是文字學上研究的單位；在語言學來說，「詞」的重要性遠大於「字」，那麼，什麼是「詞」呢？

教育部國語推行委員會「八十七年常用語詞調查報告書」對詞的定義是：「語句中具有完整概念且能獨立自由運用的基本單位爲詞。」至於「詞組」，是比「詞」更高一級的單位，中國國家標準GB/T 13715-92《信息處理用現代漢語分詞規範》對詞組的定義是：「由兩個或兩個以上的詞，按一定的語法規則組成，表示一定意義的語言單位。」

「詞組」由「詞」組成，是大於詞的單位。它和詞的不同在於：詞是造句的時候能夠獨立運用的最小單位。所謂最小，就是說不能擴展，或者通俗地說，就是中間不能插入別的成分。[5]換言之，中間能插入別的成分的，理論上應該至少屬於「詞組」這一層級。然而歷來在處理漢語詞的定義和分詞原則的問題時，始終是糾纏不清，複雜棘手的。例外或特殊的情形雖然常有，但總的原則概如上述，客家語慢慢已有了漢字書寫系統，它的語法結構不脫漢語的體系，是以客家語的詞頻研究，一方面可以藉助國語的經驗進行斷詞工作；二方面也應該以客家語爲對象，進行深入細緻的描寫分析，建立以客家語爲中心的分詞原則。

第三節　關於分詞原則的討論

中國大陸的普通話詞彙頻率研究，見諸於北京語言學院出版的《漢語頻率詞詞典》，該研究有一套分詞處理原則做爲斷詞標準。最新的看法見諸中國國家標準GB/T 13715-92《信息處理用現代漢語分詞規範》，而臺灣中央研究院的華語詞頻統計，教育部國語會的詞頻統計也都有各自嚴謹的斷詞或分

[5] 詳見葉蜚聲、徐通鏘（1993：103）