

從語料庫建構探討臺灣客語難字、缺字與 異體字議題*

葉秋杏、賴惠玲
國立政治大學

臺灣客語文本中有許多難字、缺字及異體字，在在造成語料庫建置過程之語料用字處理作業繁複且紛雜。本文首先簡述臺灣客語的用字現況，包含民間具代表性的客語辭典與官方標準，其次依據《臺灣客語語料庫》建置經驗，介紹本語料庫的用字規範，並基於文本資料清理，探析文本用字校訂類型，包含客語拼音校訂為客語漢字、客語用字統整、多字刪除、缺字補齊、顛倒字序調換、形似字勘誤等。接續則檢視客語文本中難字無法正常顯示時出現的四種情形，包括拼音、借音或借義字、空格或符號（缺字）、漢字部件拆解，並展示相對應的處理方式。本文最後以探討如何克服字碼不一以及異體字等問題作結。

關鍵詞：難字、缺字、異體字、一字多碼、臺灣客語語料庫

* 本文初稿曾宣讀於 2022 年 8 月 27 至 28 日「第十四屆臺灣語言及其教學暨臺灣學『蛻變的聲音』國際學術研討會」，感謝與會學者葉瑞娟教授與邱湘雲教授賜教，以及王勻芊、陳采蘋、王勻采、林欣穎在語料搜整與文字校訂上的協助。審查期間承蒙兩位匿名審查委員及編委會悉心更正並惠賜寶貴建議，謹申謝忱。「建置臺灣客語語料庫」計畫之共同主持人劉吉軒教授與劉慧雯教授惠予跨領域的專業指導，特此致謝。

1. 前言

語料庫 (corpus) 係大量文字的集合，語料庫語言學即是運用電腦科技管理與操縱這些可機讀的文本，結合統計與電腦演算法，以語言學方法為文字進行標記，並從中挖掘與汲取重要資訊，如詞頻或共現等 (Sinclair 1991, Stefanowitsch and Gries 2003, Hoey 2005, McEnery and Hardie 2013)。語料庫記錄的書面語或口語等語言真實使用情境，亦即大量文本中的客觀語言資訊，提供其他領域相當豐富的語言研究及分析素材，可應用於對比語言學 (Johansson 2007)、言談分析 (Aijmer and Stenström 2004, Baker 2006)、語言學習 (Chuang and Nesi 2006, Aijmer 2009)、語意學 (Ensslin and Johnson 2006)、社會語言學 (Gabrielatos et al. 2010)，以及理論語言學 (Wong 2006, Xiao and McEnery 2004)。對於進行語言描述與分析的詞典編纂者以及語法學家而言，語料庫亦是項重要的資源 (Halliday 1994, Carter and McCarthy 1997, McEnery and Hardie 2013)。各類型語料庫的建置在強勢語言 (dominant language) 開始得很早，例如美國的布朗大學現代美式英語標準語料庫 (Brown University Standard Corpus of Present-Day American English) 與現代美語語料庫 (Corpus of Contemporary American English)，英國的蘭卡斯特—奧斯陸—卑爾根語料庫 (Lancaster-Oslo-Bergen Corpus) 與英語國家語料庫 (The British National Corpus)，法國的法語文本語料庫 (Frantext)，德國的語料庫搜尋、管理與分析系統 (Corpus Search, Management and Analysis System)，西班牙的現代西班牙語語料庫 (Corpus de Referencia del Español Actual) 等。臺灣華語語料庫則有中央研究院漢語平衡語料庫 (Academia Sinica Balanced Corpus of Modern Chinese) 以及國家教育研究院的臺灣華語文語料庫 (Corpus of Contemporary Taiwanese Mandarin)。除強勢語言語料庫，少數族群為了復振其瀕危的母語，亦紛紛建置語料庫或語言資料庫，如太平洋和區域瀕危文化數字資源檔案館 (Pacific and Regional Archive for Digital Sources in Endangered Cultures)、拉丁美洲原住民族語言資料館 (The Archive of the Indigenous Languages of Latin America) 等多媒體語言資料檔案平臺。

隨國家語言政策推進，臺灣非強勢語言如臺灣閩南語、臺灣客語、臺灣原住民族語等，效仿國外語言復振經驗，近年來亦開始逐步推動語料庫建置。為保存臺灣客語書寫文字與口說語言，客家委員會於 2017 年 12 月底委託政治大學執行「建置臺灣客語語料庫」計畫，目標為建構一個含書面及口語語料之客語語料庫。《臺灣客語語料庫》於 2022 年 10 月正式上線，包括四縣、海陸、大埔、饒平、詔安、南四縣六種腔調，語料涵蓋 1990 年代迄今之臺灣客語書口文本，收錄書面語料字數達 600 餘萬字，口語達 40 餘萬字，其中少數腔（大埔、饒平、詔安）字數達到目標值（占總體至少 3%）。語料庫目前主要系統功能包含關鍵詞檢索系統、共現詞檢索系統以及斷詞系統，書面及口語關鍵詞檢索採上下文關鍵詞（**Keyword in Context**）之索引方式；口語語料文字與媒體音檔介接，支援音訊檔案播放功能，可供使用者點選時間戳記聽取與口語文字相對應之音訊內容；共現詞檢索系統提供使用者查詢緊鄰或近鄰的關鍵詞及其共現語料；斷詞系統則會自動進行語料斷詞，將書面及口語語料之語句斷開為詞彙並標示斷詞標記。¹

語料庫基礎工程的穩健性乃系統功能是否充分發揮之關鍵，將已數位化的語料進行文本資料清理係語料庫建置過程十分重要的一環，包括轉檔以及處理資料缺失與雜訊，舉凡移除各文本格式不一的設定（如縮排、換行符號、多餘空格）、剔除不必要的資訊（如文本註腳、圖表）、修正轉檔錯誤（如轉檔造成的漏字或錯字）等皆是。而與強勢語言語料庫相比，臺灣客語尚遭遇到另外幾項挑戰，在在增添了語料庫建構的難度。首先，客語在臺灣屬瀕危語言，根據行政院主計總處（2021）所發布之《109 年人口及住宅普查初步統計結果》，6 歲以上本國籍常住人口計 2,178.6 萬人，主要或次要使用語言為臺灣華語者占 96.8%，閩南語占 86.0%，客語則僅占 5.5%，客語出版品數量也與客語使用人數面臨一樣的狀況，與華語相比，客語語料相對稀缺。此

¹ 關於語料庫其他資訊，可參閱葉秋杏、賴惠玲、劉吉軒（2021）；更多詳細資料與更新，未來將另專文介紹。

外，相較於臺灣閩語主要三腔調（偏漳腔、偏泉腔以及混合腔）²僅有部分發音不同，彼此之間仍屬高度相通，臺灣客語腔調則因地而異，目前較活躍且有完整分布區的主要為六個腔調：四縣、海陸、大埔、饒平、詔安、南四縣，³其中四縣腔與海陸腔為使用人口最多的前二大腔調，除了少部分詞彙與發音差異外，兩者各聲調的調值幾乎相反，雖然彼此有聲調轉換的關係，但兩腔的單腔使用者相互對話仍有一定難度。而少數腔與其他腔調之間差異更大，這些差異也造成了用字競合現象，各腔彼此間的內部互通性與閩語三腔之間相對更低，⁴因此客語面臨著越加嚴峻的瀕危形勢。⁵再者，不同於強

² 洪惟仁（1992, 2013）指出，臺灣閩南語三大腔為偏漳腔（通行於近山地帶，又稱「內埔腔」）、偏泉腔（流行於沿海地區，故被稱為「海口腔」）以及混合腔（如臺灣南部地區的閩南語已經完成漳泉融合，因此被歸類為混合腔，又稱「漳泉濫」(Tsiang-Tsuân-lām)）。教育部（2011）《臺灣閩南語常用詞辭典》所採用之高雄腔即屬於漳泉融合相當顯著的混合腔。

³ 客家委員會（2022）《110 年全國客家人口暨語言基礎資料調查研究》資料顯示，客家民眾用以與他人溝通之客語腔調（複選），使用且能分辨出來的客語腔調中，人數最眾者為四縣腔，比率为 57.7%，亦為目前官方或公開場合主要使用的腔調；其次則為海陸腔（44.4%）；另外四個腔調使用比率則較低，依序為南四縣腔（5.8%，主要分布於南部六堆地區）以及三個少數腔：大埔腔（5.9%）、饒平腔（2.4%）與詔安腔（1.6%）等（客語民眾使用客語腔調之縣市分布完整數據，可見客家委員會（2022））。

⁴ 臺灣閩南語三腔的相通性以及臺灣客語六腔的特殊性，可由教育部（2011）《臺灣閩南語常用詞辭典》和教育部（2019）《臺灣客家語常用詞辭典》略見一二。《臺灣閩南語常用詞辭典》之詞目音讀以提供第一優勢腔（以高雄音為主）為代表（附錄區的詞彙如地名等，若具區域特殊性，則酌收在地腔或第二優勢腔（以臺北音為主）），而《臺灣客家語常用詞辭典》則採六腔平行的架構進行編修（教育部 2011, 2019）。

⁵ 據客家委員會最新一期發布之《110 年全國客家人口暨語言基礎資料調查研究》（客家委員會 2022）所示，自 105 年至 110 年度，客家民眾整體聽的能力從 64.3% 下滑至 56.4%，五年之間下降了 7.9 個百分點；整體說的能力也由 46.8% 下滑至 38.3%，跌幅為 8.5 個百分點。此份報告中也指出，臺灣客語使用率呈現隨年齡層下降而明顯降低的趨勢，60 歲以上聽懂客語以及會說客語的比例分別為 85.3% 及 75.2%，未滿 13 歲的比例則急降至 21.6% 及 9.6%；另根據 95 年度《臺灣客家民眾客語使用狀況調查》（行政院客家委員會 2006），若缺少政府與民間持續推動，客家民眾的客語口說能力每年會自然流失 1.1 個百分點，據此推估，在當時調查時間的 40 年以後（民國 135 年），臺灣客語使用人口即可能不復存在。令人更為擔憂的是，根據 110 年的報告，近五年（自 105 年起）客語聽說能力下降的幅度已皆高於推估的自然流失率 5.5 個百分點，因此需要更多的資源投入以協

勢語言歷經幾世紀的發展早已建立了標準化的書寫系統，現代臺灣客語的文字化於近 50 年內才甫興起，書寫系統標準規範現今仍刻正進行，在文字系統尚未成熟健全的情況下，客語不僅有許多特殊用字，亦有部分詞彙仍為有音無字以及找不到對應文字，許多客家作者在寫作時無可依據，導致部分客家出版品文本內容出現自行造字、以拼音代字、使用華語文字替代等，造成客語用字凌亂分歧，甚或是有些語句已不符合客語結構或句法，也因此影響文本品質，加劇語料搜整的難度。另一方面，系統字碼不一或是因字碼間轉換而衍生的亂碼或缺漏字而無法正確顯示等問題也造成查索不易。

客語語料作業處理過程中，除了面臨客語作品出版量有限之外，最大考驗即是難字顯示與用字統一問題，種種以上顯示臺灣客語文本在數位化過程中，遭遇到比強勢語言更為嚴苛的挑戰，尤其在語料用字處理方面更為繁複紛雜。在臺灣客語用字尚未全面規範化的狀況下，本文將基於語料庫的建置經驗，首將簡述臺灣客語用字現況，並介紹官方用字制訂標準與分享實際經歷的用字校訂狀況類型。而客語用字與語料庫系統建構之字型編碼與網頁顯示等功能息息相關，因此將接續探討如何克服資料清理中的難字、缺字、一字多碼與異體字等問題，最後則為全文結論。

2. 臺灣客語用字現況簡述：民間百家爭鳴與官方標準制訂

相較於臺灣華語穩定發展，呈現出完整的書面樣貌與口語豐富度，臺灣客語的文字化起步較晚，有一部分用語至今仍無文字可表示，須借用古漢語或現代華語之用字，甚或只能以拼音表示。現代客語書寫可追溯至法國傳教士 Charles Rey 於 1901 年出版之 *Dictionnaire Chinois-Français, Dialecte Hac-Ka. Hong Kong: Imprimerie de la Société des Missions Etrangères* (臺灣通

助客語之推廣與傳承(客家委員會 2022)。除此之外，該期調查中客家民眾「完全聽得懂客語」的人數占樣本數之 39%，「會說很流利的客語」則占 25.1%，兩者比例均未達樣本數的五成，在在顯示臺灣客語已處於持續萎縮的狀態(羅肇錦 1990，黃宣範 1993)。

稱為《客法大辭典》)以及蘇格蘭傳教士 Donald MacIver 於 1905 年出版之 *An English-Chinese Dictionary in the Vernacular of the Hakka People in the Canton Province* (臺灣通稱《客英大辭典》) 兩部早期客語文獻, 兩本辭書分別記錄了廣東、梅縣、興寧、平遠、蕉嶺、五華、大埔一帶當時的客語成語、諺語、教會用語等, 部分書寫用字採「記音不記字」或「借字」的方式處理(江敏華、黃彥菁、宋柏賢 2009)。臺灣客家民間文學中, 代表臺灣客家族群自覺意識發展的重要文獻為〈渡臺悲歌〉, 流傳數個版本, 最廣為人知的為黃榮洛將手抄本整理加註後於 1989 年出版的《渡臺悲歌: 臺灣的開拓與抗爭史話》,⁶內容屬客家山歌詩。根據曾學奎(2003)整理各版〈渡臺悲歌〉的詞彙比較, 發現俗體字或簡體字使用比例偏高、同一詞彙缺乏一致性寫法(例如閩南人有「學老」、「貉老」、「福佬」三種用法), 或是因無法從漢字找出對應詞彙, 選以羅馬拼音書寫。早期客家文學作品多屬華語書寫, 為了因應客語寫作, 許多客語工具書應運而出, 各家學者開始針對書寫用字進行考源, 坊間可取得的幾本主要客語辭典, 包括《客話辭典》(中原週刊社客家文化學術研究會 1992)、《客語辭典》(楊政男、龔萬灶、徐清明 2013)、《海陸客語字典》(詹益雲 2003)、《六堆詞典》(曾彩金 2019)、《客語詞庫》(何石松、劉醇鑫 2007)、《臺灣四縣腔海陸腔客家話辭典》(徐兆泉 2009)等。《客話辭典》及《客語辭典》之語系皆以苗栗四縣腔為主, 《客話辭典》係由中原週刊社聘請客家菁英, 如羅肇錦、徐清明、龔萬灶、楊政男、宋聰正等人投入彙編, 內容有不少用字尚未確定者, 依古韻書及字書之音、義, 或依六書造字原則借音、附義或另造新字, 所收錄之詞彙以客語和華語措辭不同者為主, 並依音標排序。該辭典主張「客家話有音即有字」, 以考證本字為第一原則、採用俗字為第二原則、採堪用字為第三原則、採同源字為第四原則、借音為第五原則, 若以上五種均無法處理者則從缺並待後人考訂。而《客語辭典》源自 90 年代初期母語課程開設, 客語工具書需求大增, 因此作者群將《客語字音詞典》(由《客話辭典》增修)重新修訂後出版為《客

⁶ 據黃榮洛(1989)推斷, 〈渡臺悲歌〉之寫作年代約在清嘉慶、道光年間。

語辭典》，書中聲母、韻母符號皆依循教育部 2012（民 101）年公告之《客家語拼音方案使用手冊》。

另一方面，因有感於海陸語系之客語語音工具書貧乏，詹益雲（2003）完成編纂並出版《海陸客語字典》，內容以客語單字為條目，標音以羅馬拼音為主，注音符號、同音字、切音字為輔。六堆南四縣腔的客語專家也投入詞彙編撰行列，《六堆詞典》由曾彩金總編輯並於 2019 年出版發行，此係基於《六堆客家詞彙庫編纂計畫》（曾彩金 2010）之基礎上彙編專門通行於六堆的南四縣常用詞彙，包含日常用語、片語、俗諺語，收錄單音（字）詞約 7,000 個、複音詞 12,000 餘條，並採用教育部 2012（民 101）年公告《客家語拼音方案使用手冊》之聲母符號表、韻母符號表與南四縣腔聲調符號表。

四縣腔與海陸腔為客語主要兩大腔調，因此也陸續出現同時收錄此二腔的辭典，例如《客語詞庫》（何石松、劉醇鑫 2007）及《臺灣四縣腔海陸腔客家話辭典》（徐兆泉 2009）。《客語詞庫》係根據《現代客語詞彙彙編》（何石松、劉醇鑫 2002）及《現代客語詞彙彙編續編》（何石松、劉醇鑫 2004）增補修訂而成，蒐羅現代客家話詞語 28,000 餘條，每條詞彙為四縣、海陸、華語相互對應，並以臺北市推行客語教學所使用的兩種音標符號分為兩種版本發行（客語音標版、注音符號版），對於尚未形成共識之借用字，以標楷體書之表尚待研究。由徐兆泉（2009）增修之《臺灣四縣腔海陸腔客家話辭典》，則是以《臺灣客家話辭典》（徐兆泉 2001）為基礎改寫，使用教育部通過的「通用拼音」作為標音系統，語詞條目按四縣腔拼音開頭的 26 字母排序，詞目首為該語詞的四縣腔拼音，斜線後為海陸腔拼音，語詞後附加例句。

除了四縣、海陸兩腔已有辭典出版外，徐登志於 1993 年籌辦「寮下文化工作室」，自編《臺灣大埔音客語辭典》（徐登志 2005），收錄道地大埔客語 26,008 條詞目，依照 21 個客語聲母音序排列，辭典採用臺灣客語拼音與注音符號兩套系統逐一標注發音，並附華語釋義。而饒平、詔安兩腔目前暫無出現專屬於自身腔調的客語辭典，僅於學者的個人研究及著作中提及，如徐貴榮所著之《臺灣饒平客話》（徐貴榮 2005）與《饒平客家調查與語言論

輯》(徐貴榮 2018)，書中章節介紹饒平腔之音韻結構、詞彙特色與重要詞彙等。至於詔安腔，則以學者張屏生於 2002 年自行出版《雲林縣崙背鄉詔安腔客家話語彙初集稿》為代表，依照不同主題，將詔安腔客語詞彙分為 35 類。饒平與詔安兩腔雖有學者投入詞彙彙編，但目前仍未有一完整、正式的辭典專書。

如上所述，民間各家辭典的不同編纂者對於較特殊的本字未定詞彙有著各自偏好之用字習慣與處理方式(如使用借義字或借音字、參考其他方言或古文、選用特殊罕見字，甚或是自行造字等)，而數十餘年間各家辭書相繼出版，用字體例勢必會隨時間而有所改變；此外，少數腔面臨更嚴重的瀕危問題，不僅有詞彙流失之困境，坊間出版的臺灣客語少數腔(大埔、饒平、詔安)辭典更是稀缺(目前僅搜索到徐登志(2005)所著之《臺灣大埔音客語辭典》)，加上各腔內部之間存在著明顯的讀音與用字歧異性，這些難點與問題實為推行客語書寫用字系統整合之挑戰。許多學者也針對客語用字進行討論與研究分析，探尋客語用字的起源與流變(如羅肇錦 1991，李如龍 1993，姚榮松 1998，涂春景 2004，邱湘雲 2004, 2013，鍾榮富 2014，滕暢 2017)。

臺灣客語用字紛然雜陳的現象造成書寫困難，不利於文字記錄與知識傳承，也可能會加深語言溝通交流上的隔閡，因此由公部門統一制訂標準並發布客語用字規範便顯得尤其重要。有鑑於此，教育部於 2008 年成立「臺灣客家語書寫推薦用字小組」，研訂〈臺灣客家語書寫推薦用字漢字選用原則〉，並根據此原則陸續於 2009(民 98)年及 2011(民 100)年推出第 1 批客家語書寫推薦用字(305 字)與第 2 批客家語書寫推薦用字(259 字)，內容包括四縣、海陸、大埔、饒平、詔安等五腔之資料(含推薦用字及其拼音、華語注音、構詞用例與釋義等)。推薦用字係針對客家語本字、堪用字、俗用字、借用字等之選用提出認定方式與原則，除了將字音義的學理依據納入考量外，亦兼顧幾項重要因素，如跨語言比對、教學實務以及電腦資訊等(教育部 n.d.)。隨後教育部也基於 2003(民 92)年《臺灣客語通用拼音方案》與客家委員會「客語能力認證通用拼音」之整合修訂成果，於 2012(民 101)年出版《客家語拼音方案使用手冊》，並新增南四縣腔，將原五腔架構增修

為六腔。教育部進行客語用字與拼音修訂的期間，於 2008（民 97）年 5 月公開《臺灣客家語常用詞辭典（試用版）》，並於 2019（民 108）年 11 月改版釋出《臺灣客家語常用詞辭典》正式版，共計 15,454 條詞目，收錄 7,402 筆大埔、饒平、詔安、南四縣等腔相關資料。此辭典行文所使用之漢字也以用字統一為主要原則，⁷大埔、饒平、詔安、南四縣等腔之用字，也盡量與四縣與海陸腔一致。⁸

客家委員會則自 2003 年起開辦「客語能力認證」測驗，鼓勵並推廣客語學習，程度分為初級、中級暨中高級，腔調分為四縣（含南四縣）、海陸、大埔、饒平、詔安共五腔，⁹並將認證的詞彙彙整建置為《客語認證詞彙資料庫》網站，共收錄初級與中級暨中高級共 26,925 個詞彙，2022 年更首度舉辦高級認證；另《哈客網路學院》網站也提供認證教材與試題下載，包含初級 1,200 個詞彙、中級 1,800 個詞彙、中高級 2,000 個詞彙，以及高級 3,000 個詞彙，各級認證詞彙類別共分為 18 個單元。此外，為搶救饒平、大埔、詔安三個流失情況較嚴重的少數腔，客家委員會於 2003 年啟動「臺灣饒平、大埔、詔安客語辭典編纂工作計畫」，搜整此三少數腔客語字彙，除了審定三腔字彙之形、音、義，也以適合大眾應用為目標，著手進行臺灣饒平、大埔、詔安三部客語辭典之編纂，為瀕臨消失危機的三個少數腔客語做及時的

⁷ 詳見教育部（2019）《臺灣客家語常用詞辭典》編輯說明第四點「編輯凡例」內之「統一用字」

（<https://hakkadict.moe.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=NFBwfl/description?id=MSA00000045&opt=opt2>）。

⁸ 客語與漢語之用字發音有不少明顯分歧處，尤其許多客語詞彙有音無字，在在增添定字的難度。例如在《臺灣客家語常用詞辭典》以「全文」輸入「本字未定」所得之的條目筆數為 68 筆、「從俗暫用」的條目達 69 筆（查詢日期為 2022 年 11 月），再加上六腔之間的詞彙與發音差異，顯見臺灣客語定字仍存在著一定程度上的困難。

⁹ 相較於教育部（2003）《臺灣客語通用拼音方案》以及教育部（2019）《臺灣客家語常用詞辭典》均將南四縣獨立為一腔，亦即將臺灣客語腔調架構增建為六個腔調，客家委員會之《客語認證詞彙資料庫》之詞彙目前仍採以五腔方式呈現。然為順應教育部於 101 年將「南四縣」自四縣腔調中獨立出來，客家委員會 110 年度之人口調查報告也增列「南四縣」腔為第六個客語腔調（客家委員會 2022: 87）。

編錄。此三部辭典於 2008 年網路正式公告試用（行政院客家委員會 2008c, 2008d, 2008e），收錄詞條數分別為饒平腔 21,432 條、大埔腔 20,274 條、詔安腔 20,006 條。辭典為 PDF 檔形式，內容主要為客語字詞及其拼音、華語釋義以及客語例句等資訊。

有鑑於教育部與客家委員會為臺灣教育與客家事務的最高主管機關，臺灣客語語料庫用字依據係以教育部兩批規範用字以及《臺灣客家語常用詞辭典》詞彙作為第一順位；其次，教育部若未收錄，則以客家委員會《客語認證詞彙資料庫》詞彙為輔助，為用字依據之第二順位；最後若屬教育部與客家委員會均未收錄，且無法從原始文本透過拼音或註釋查證用法之客語字，並經審核非屬須被修正的狀況（例如文字勘誤、一字多碼整併、異體字替換等），則以忠於原著的方式，維持作者原用字。待未來公部門以及各領域專家學者將這些未定字標準化後，即可透過系統程式搭配人工審核進行更新取代。關於用字校訂等相關議題，將於下一章節進行詳細介紹。

3. 臺灣客語語料庫之文字處理

臺灣客語語料庫的每筆語料，均會依據腔調不同，請本腔的語料轉校人員進行轉寫以及二次以上的交叉校訂。轉校人員為語料庫建構期間所招募且經過條件篩選與教育訓練的客語六腔人士，主要為客語薪傳師、客語能力認證命題或閱卷等典試人員、客語教材編輯委員等。本節將依序簡介臺灣客語語料庫的用字規範，以及文本校訂遇到的問題類型與處理方式。

3.1 客語用字規範

臺灣客語語料庫用字依據主要遵循教育部兩批規範用字以及《臺灣客家語常用詞辭典》，再者則是客家委員會的《客語認證詞彙資料庫》。此二公部門所制訂的客語規範用字大致相同，小部分仍存有差異。以客語時相詞（phase marker）之四縣拼音與調型「doˊ」為例（常用以表示前述動作完成，華語多為「到」），《臺灣客家語常用詞辭典》使用「著」（圖一），而《客語

認證詞彙資料庫》則使用「着」(圖二)。¹⁰

詞目	【燙著】	詞性：動
四縣音	▶ ben do`	
海陸音	▶ ben` do´	
大埔音	▶ ben` do^	
饒平音	▶ ben` do`	
詔安音 ^ㄅ	詞目	詔安音
	燙蹄	▶ pen dai^
南四縣	▶ ben do`	
釋義	用身體碰觸。例：高壓電電壓大，毋好燙著。(高壓電電壓大，不可碰到。)	
近義詞	【碰著】	
對應國語	碰到	

圖一 《臺灣客家語常用詞辭典》「著」相關詞目舉隅

¹⁰ 《臺灣客家語常用詞辭典》之拼音查詢功能支援以調值、調型或調類進行搜尋，《客語認證詞彙資料庫》之拼音查詢則僅支援調型。根據教育部(2012)《客家語拼音方案使用手冊》，客語聲調之調類包含陰平、陽平、上聲、陰去、陽去、陰入、陽入，客語六腔因古今聲調歸併情況不同，除海陸腔為七聲調外，其他則為六聲調。調類依序對應到聲調調號，而調號亦可以調值或調型標示。調值係採五度制調值法將聲調分為五度，並以阿拉伯數字代表聲調的高低，如 53 或 24；調型則為聲調符號，包括「ˊ」、「ˋ」、「ˊˋ」、「ˊˊ」、「ˊ+」等，其中表高平調(調值 55)及高促調(調值 5)之調型以「空白」表示。更多客語拼音聲調及符號等詳細資訊以及示例，請參見教育部(2012, 2019)。

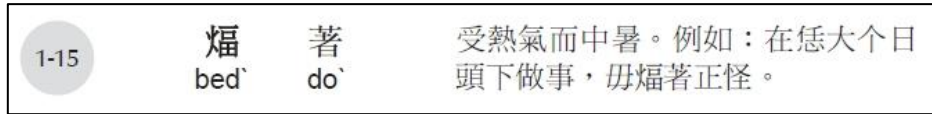


圖二 《客語認證詞彙資料庫》「着」相關詞彙舉隅

因此，在進行資料清理與用字校訂時，若文本出現「着」，語料庫工作人員須先比對教育部「著」與客家委員會「着」在客語文句中的語音、語法、語意各用法均一致，且確認教育部規範用字中並無使用「着」作為其他用詞之規範用字後，方可將「着」替換成「著」。

值得注意的是，《客語能力認證參考詞彙·高級（四縣腔上冊）》一書中之編輯說明提及：「用字依教育部公告之標準用字為原則，教育部尚未公告者，得以『暫用字』標示之。惟教育部刻正進行客語辭典重編作業，部分尚未及更正、然已有學界共識之用字，則以客家委員會之客語認證詞彙資料庫為準。高級詞彙中含有教育部客語辭典尚未收詞者眾，為便於考生學習，以暫用字呈現……」（客家委員會 2022: 12），顯見兩大公部門對於客語用字標準化持續不斷地進行討論與研訂更修，¹¹以《客語認證詞彙資料庫》中的「着」為例，此用字於 110 年的高級認證參考詞彙中已與教育部統一，修訂為「著」（如圖三所示）。

¹¹ 《臺灣客家語常用詞辭典》近年來也發布多次辭典修訂公告（2016 年、2020 年、2021 年、2022 年），語料庫團隊持續追蹤規範用字修訂訊息，並進行文本用字滾動式修正。



圖三 《客語能力認證參考詞彙·高級（四縣腔上冊）》「焮著」

3.2 文本用字校訂

3.2.1 客語拼音校訂為客語漢字

將文字資料數位化時，必須先進行文字處理(王雅萍、張如瑩、陳秀華、蕭貴徽 2012，江敏華、黃彥菁、宋柏賢 2009)。在出版年份較早期的書面文本中，由於官方尚無公布用字規範，對於作者而言，許多客語詞彙只知其音，不知其字，因此有部分文本中之用字係以拼音呈現。相關示例如下：

(1) 「dem`」修訂為「蹬」

原文：大聖一急速速念矣咒語同時用力用腳一 **dem`**，當地个山神土地便匆匆忙忙現身：「大聖有何吩咐？……」

校訂：大聖一急速速念歛咒語同時用力用腳一**蹬**，當地个山神土地便匆匆忙忙現身：「大聖有何吩咐？……」

華譯：大聖一急速速念了咒語同時用力用腳一踏，當地的山神土地便匆匆忙忙現身：「大聖有何吩咐？……」

妖怪捉人水裡，啊！究竟係何方妖怪？大聖一急速速念矣咒語同時用力用腳一 dem` ，當地个山神土地便匆匆忙忙現身：「大聖有何吩咐？……」
--

圖四 拼音校訂舉例：「dem`」

(資料來源：《中大湖个風：桃園地區新舊兩隻移墾地區个故事客語文選集》(2018: 265))

詞目	【蹬】	詞性：動
四縣音	dem`	
海陸音	dem´	
大埔音	dem^	
饒平音	dem`	
詔安音	dem	
南四縣	dem`	
釋義	用腳踩踏。例：蹬地泥（以腳踏地）。	

圖五 以拼音「dem`」查找《臺灣客家語常用詞辭典》相關詞目：「蹬」

3.2.2 客語用字統整

早期文本亦常有借音字、借義字等作者自行選用字的狀況，舉隅如下：

(2) 寮（教育部規範字，華語釋義為「休閒、聊天、玩耍」）

a. 燎

原文：逐日都係燎日喔！

校訂：逐日都係寮日喔！

華譯：每天都是放假日喔！

拜六、拜日正有鬧熱阿公記性無恁好哋，佢想恁好啊！逐日都係燎日喔！

圖六 「寮」之非規範用字舉例一：「燎」

（資料來源：《臺灣客家語朗讀文章選輯》（2008: 172））

b. 聊

原文：阿春伯，會無閒無？恁久就無下來聊哋？

校訂：阿春伯，會無閒無？恁久就無下來寮欸？

華譯：阿春伯，會忙嗎？那麼久沒下來玩了？

「阿春伯，會無閒無？恁久就無下來聊地？」

圖七 「寮」之非規範用字舉例二：「聊」

（資料來源：《阿啾箭个故鄉》（2004: 14））

c. 遷

原文：今這下催食飽了，愛轉了，看哪久正來遷哪？

校訂：今這下催食飽了，愛轉了，看哪久正來寮哪？

華譯：現在我吃飽了，要回去了，看哪時再來玩哪？

今這下催食飽了，愛轉了，看哪久正來遷哪？

圖八 「寮」之非規範用字舉例三：「遷」

（資料來源：《龍潭鄉廖德添客語專輯（一）》（2006: 109））

⑦正來遷 $cang^3\ loi^5\ liau^3 / cang^3\ loi^5\ liau^7$ ：再來坐坐，再來玩。

圖九 非規範用字「遷」之文本註釋

（資料來源：《龍潭鄉廖德添客語專輯（一）》（2006: 109））

d. 料

原文：彭祖八百二十歲鬚當長，有一擺揸張果老在路項共下料，

校訂：彭祖八百二十歲鬚當長，有一擺揸張果老在路項共下寮，

華譯：彭祖八百二十歲鬚鬚很長，有一次跟張果老在路上一一起玩，

彭祖八百二十歲鬚當長，有一擺⁽⁶⁾撓張果老在路項共下料⁽⁷⁾，

圖十 「寮」之非規範用字舉例四：「料」

(資料來源：《花蓮客家民間文學集》(2009: 43))

(07) 共下料：音kiung+ ha \ liau+，
一起玩。

圖十一 非規範用字「料」之文本註釋

(資料來源：《花蓮客家民間文學集》(2009: 43))

e. 翹

原文：吾介奈久做戲，你來吾介翹啦。

校訂：吾个哪久做戲，你來吾个寮啦。

華譯：我這裡什麼時候演酬神平安戲的話，你來我這玩啦。¹²

吾介奈久做戲，你來吾介翹⁽⁹⁾啦。

圖十二 「寮」之非規範用字舉例五：「翹」

(資料來源：《苗栗縣客語故事集》(1998: 20))

⑨ 翹：音 liau⁴，休閒、遊玩。

圖十三 非規範用字「翹」之文本註釋

(資料來源：《苗栗縣客語故事集》(1998: 20))

¹² 其中一位匿名審查人指出，「哪久做戲」為民間文學述說故事時常用的方法，如用「某名」指稱某個人，此處的「哪久做戲」則是指「某個時間」，在此謝謝審查人的更正。筆者參考審查人的建議以及原作者提供之翻譯，並根據前後語境，將「吾个哪久做戲」之翻譯修訂為「我這裡什麼時候演酬神平安戲的話……」。

從文本訊息可得知，在官方的規範用字制訂之前，客語用字使用混亂，而這些不同用字之間多半存在著字符、發音或語意的相似性。例如，從字符來看，「療」之偏旁同「寮」，兩者之音韻結構也皆由「liau」所組成（以四縣腔為例，然調值有所差異，前者為「liau11」，後者為「liau55」）；從發音方面，「料」與「寮」之四縣腔發音皆為「liau55」；而從語意角度，「寮」可表休閒、聊天、玩耍，部分語意與「聊」（表聊天）相同，且音韻結構也皆由「liau」所組成（二字調值不同，「聊」為 liau11）。至於採用「遯」表「遊玩」之作者，推測應係根據其語音相似性（「遯」之漢語拼音為「liào」）；「𪗇」之選用則可能是其語意相近（根據教育部（2021）《重編國語辭典修訂本》，其漢語拼音為「niào」，語意為「戲弄」）。更多例子如客語「舖娘」（華語「妻子」，其他不同選字包括「媠娘」、「輔娘」、「夫娘」等）、「俵仔」（華語「兒子」，各式用法諸如「賴仔」、「賴子」、「賴兒」、「孺仔」等）也是類似的狀況。這些用字不一的現象，皆會由語料庫工作人員確認詞彙的語音、語法、語意以及文本篇章上下文語境後，依據公部門規範用字予以校訂。

3.2.3 多字刪除

文本中若檢查到多字情況，在確認文本內容後，則將多字刪除。唯須留意腔調之詞彙差異，並盡量多方佐證，除了查閱《臺灣客家語常用詞辭典》之外，若文本有提供註釋，可作為校訂之參考，或是若語料為多腔文本形式（例如全國語文競賽客家語朗讀文章），語料庫工作人員即會參照對應並確認後（如下方例），刪除多字部分。

(3) 「大同『𪗇』鄉」修訂為「大同鄉」

原文：毋過蹶山个人會去到宜蘭个大同𪗇鄉蹶起，也算係宜蘭異有名个高山湖。

校訂：毋過蹶山个人會去到宜蘭个大同鄉蹶起，也算係宜蘭已有有名个高山湖。

華譯：不過爬山的人會到宜蘭的大同鄉開始爬，也算是宜蘭很有名的高山湖。

源頭。佢个行政區在新北市个烏來區，毋過蹶山个人會去到宜蘭
个大同膠鄉蹶起，也算係宜蘭具有名个高山湖。因為山路溜溜¹⁶⁸

圖十四 多字舉例：「大同『膠』鄉」

（資料來源：《105 年全國語文競賽客家語朗讀文章社會組（饒平腔）》（2016））

源頭。佢个行政區在新北市个烏來區，毋過蹶山个人會去到宜蘭
个大同鄉蹶起，也算係宜蘭已有名个高山湖。因為山路溜溜²⁰⁶難

圖十五 上圖例（饒平腔）之其他腔對照版（海陸腔）：「大同鄉」

（資料來源：《105 年全國語文競賽客家語朗讀文章社會組（海陸腔）》（2016））

3.2.4 缺字補齊

客語文本也檢查到缺字狀況，如直接漏字，此時即須依據作者提供的文本註釋將缺漏文字補齊。舉例如下：

（4）「毛末節」修訂為「『微』毛末節」

原文：板圓愛煮來好食，各個毛末節就愛當注意。

校訂：板圓愛煮來好食，各個微毛末節就愛當注意。

華譯：湯圓要煮起來好吃，各個枝微末節都要很注意。

板圓愛煮來好食，各個**毛末節**就愛當注意。好比煮板圓個湯頭，愛用大骨炆個、無就愛用雞仔燻過個，湯頭凝油，正會香、正有好味緒。食慣地阿媽煮個板圓，阿媽煮板圓放個配料，正係真工夫：油放落去，蝦蜆仔、三層肉切砵(kud)、香菇切片、紅蔥頭、辣椒仔，加兜仔鹽共下燻(biaq)熟嚟香。接好個板圓分佢燥水過，放到滾水裡肚，等到拋拋滾，一粒一粒個板圓，在滾水面頂浮浮胖胖，熟哋！就好撈到湯頭裡肚，配料落落下去，再煮分佢滾來，艾菜、芹菜放落去，蓋仔拿好、火搵成，就做得撿碗筷，準備好食哩啲！

在這天時恁冷個時節，食一碗燒燒個板圓，做得燒暖自家霜凍個手腳，做得燒暖自家個圓身。恆著從細到大，佢兜姊妹仔同阿媽共下接板圓，共下食板圓個情景，心肝肚實在當燒暖哦！

作者：邱一帆

☆詞彙學習☆

【呼呼滾】fu fu gun`：形容風聲。	【打眼】da` ngien`：搶眼、顯目。
【掃掃滾】so so gun`：形容雨聲。	【微毛末節】mi` mo` mad jied`：枝微末節。

圖十六 缺字舉例：「毛末節」

(資料來源：《教育部電子報「閱讀越懂閩客語」專欄(客家語文章 106 年)》(2017))

3.2.5 顛倒字序調換

字序顛倒的情況較少發生，例(5)的案例為機關名稱，因此確認正確名稱後直接校訂即可。

(5) 「員委」修訂為「委員」

原文：「行政院客家**員委會**客家貢獻獎」係恩客家界第高榮譽表彰，
校訂：「行政院客家**委員會**客家貢獻獎」係恩客家界第高榮譽表彰，
華譯：「行政院客家委員會客家貢獻獎」是我們客家界最高榮譽表彰，

在會發cii v 驚，在這存亡接續中，「行政院客家**員委會**客家貢獻獎」係恩客家界第高榮譽表彰，當重要個獎，佢

圖十七 字序顛倒舉例：「員委」

(資料來源：《2012 苗栗縣第 15 屆夢花文學獎得獎作品專輯(一)》(2012: 370))

3.2.6 形似字勘誤（兩者非正異體字關係）

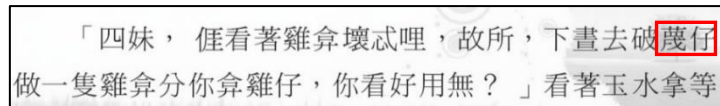
語料庫客語用字之正異體關係，主要遵照 Unihan Database（中日韓統一表意文字數據庫）所制訂之判定標準。¹³若文本用字與規範用字之間非嚴格的正異體字關係，而僅是用字誤選（多半為形似字），且誤用字與規範字之語意及用法不相同，則須予以校訂之。相關示例如下：

（6）「『蔑』仔」修訂為「『箴』仔」（表「竹子」之意）

原文：下晝去破蔑仔做一隻雞畀分你畀雞仔，你看好用無？

校訂：下晝去破箴仔做一隻雞畀分你畀雞仔，你看好用無？

華譯：下午去砍竹子做一個雞籠給你關小雞，你看好用嗎？



「四妹， 佢看著雞畀壞忒哩，故所，下晝去破蔑仔
做一隻雞畀分你畀雞仔，你看好用無？」看著玉水拿等

圖十八 形似字舉例：「『蔑』仔」

（資料來源：《2013 苗栗縣第 16 屆夢花文學獎得獎作品專輯（一）》（2013: 309））

¹³ 關於正異體字以及中日韓統一表意文字等相關介紹，將於下一節闡述。

詞目	【竹箴仔】		詞性：名
四縣音	▶ zug`med e`		
海陸音	▶ zhug med`er		
大埔音	詞目	大埔音	
	竹箴	▶ zhug^ med`	
饒平音	詞目	饒平音	
	竹箴	▶ zhug`med	
詔安音	詞目	詔安音	
	竹箴	▶ zhu' med`	
南四縣	▶ zug`med e`		
釋義	以弓鋸截取材料，刮除竹青，將竹片細劈，去竹簧，最後用整箴刀修整，用劍門定寬。劈竹箴是學習竹細工最基本的入門工作。例：用竹箴仔編籠床底个功夫，現下無幾多儕會。（用竹箴編蒸籠底的功夫，現在沒幾個人會。）		
近義詞	【箴仔】		

圖十九 以釋義「竹子」查找《臺灣客家語常用詞辭典》相關詞目：「竹箴仔」及其近義詞「箴仔」

詞目	【蔑】
四縣音	▶ med
海陸音	▶ med`
大埔音	▶ med`
饒平音	▶ med
詔安音	▶ med`
南四縣	▶ med
釋義	藐視、侮辱。例：輕蔑。

圖二十 《臺灣客家語常用詞辭典》「蔑」

例(6)為「蔑」與「箴」，二字相同的部分在於位處下方之部件，文字上側之部首則有不同（分別為「艸部」(艹)與「竹部」），原文文句中表竹子，然「蔑」於客語中為「輕蔑」的意思，與「竹子」之語意有所差異，因

此須進行文字校訂。

3.3 客語難字與缺字處理

客語有許多詞彙採用罕用字或難字，部分作者往往會以漢字部件組合的方式自行造字，例如：𪗇、埕、𪗇……等；而教育部在處理客語特殊用字時，為表達臺灣客語的特殊性，無可避免地會選用「罕用字」作為客語規範用字。在電腦應用環境中，許多罕用字對於一般使用者在操作輸入法時無法利用繕打，對於瀏覽者也常有無法透過網頁正確看到字形顯示的狀況。在文字數位化的過程中，必然會使用文書軟體處理電腦「字型 (font)」，即電腦程式記錄某個字體的完整符號集 (柯志杰、蘇煒翔 2014)。電腦使用文字時須為每一個字加以編碼，才有辦法儲存、傳遞、交換資料，而要對文字進行編碼前，確立「字元集 (Characters Set)」是至關重要的。隨著資訊科技日漸普及，世界各國制訂了各自的字元碼，但卻導致嚴重的字元碼競合現象，更進而阻礙資訊上的交流。為容納全世界各種語言的字元和符號，自 1984 年起，國際標準組織 (International Organization for Standardization, 簡稱 ISO) 制訂通用字元集 (Universal Character Set, 簡稱 UCS)，由 ISO/IEC JTC1/SC2/WG2 (表意文字標準字集討論組，簡稱 WG2) 負責擬定，發展了一套 ISO/IEC 10646 的國際編碼標準；另外一個組織 Unicode Consortium 也於 1989 年開始研發統一碼 (Universal Code, 簡稱 Unicode)。最終兩個協會的標準互相整合，同步發展 ISO/IEC 10646 及 Unicode，ISO 提供 ISO/IEC 10646 內的字元及編碼資料，Unicode Consortium 則對這些字元及編碼提出應用的方法以及語義資料作補充，並且編印 Unicode 標準。故 Unicode 又稱為「統一碼」或「萬國碼」，為電腦的字元編碼，每種語言中的每個字元 (character) 皆設定了統一並且唯一的二進制編碼，以利電腦以簡單的方式來呈現和處理文字。而只要是涉及漢字的使用，便勢必存在著字元編碼及字形顯示的根本問題，其中難字問題更是亟需克服以順應數位化。「Unicode 擴展漢字」目前已編訂至 Unicode 15.0 版 (2022 年發布)，編碼表包含中日韓統一表意文字、中日韓兼容漢字、中日韓兼容補充漢字、擴展 A 區至 H

區漢字等，其容納世界各種語言的字元和符號，字集量相較於其他字碼更為完整。《臺灣客家語常用詞辭典》於 2019 年上線的新版系統，更新項目之一即是儘量將原版使用字圖呈現的詞目文字化，這些難字即屬於擴展 A 至 E 區，其中有些字無法運用一般電腦內建之輸入法繕打出來（如「𪗇」¹⁴、「𪗈」¹⁵、「𪗉」¹⁶等）。目前語料庫蒐集到的文本中，可觀察到文字無法正確顯示時的四種問題類型，通常也多為難字。第一種類型是直接以拼音呈現，可於《臺灣客家語常用詞辭典》以「詞目音讀」搭配其釋義進行查詢：

(7) 「kioi」修訂為「瘵」¹⁷

原文：騎 kioi 吔，兩子哀坐在大樹下，

校訂：騎瘵欵，兩子哀坐在大樹下，

華譯：騎累了，母子倆坐在大樹下，

騎kioi吔，兩子哀坐在大樹下，

圖二十一 拼音舉例：「kioi」（「瘵」之拼音）

（資料來源：《2011 苗栗縣第 14 屆夢花文學獎得獎作品專輯》（2011: 343））

¹⁴ 客語「𪗇」，量詞，計算楊桃、西瓜、柑橘等果肉剖開後的稜片單位。編碼為 U+3F13，屬擴展 A 區。

¹⁵ 客語「𪗈」，華語釋義為「蒙住、蓋住」。編碼為 U+20584，屬擴展 B 區。

¹⁶ 客語「𪗉」，華語釋義為「繞行」。編碼為 U+2B7E7，屬擴展 D 區。

¹⁷ 客語「瘵」，華語釋義為「累」。編碼為 U+24E01，屬擴展 B 區。

詞目	【瘵】	詞性：形
四縣音	▶ kioi	
海陸音		
大埔音	▶ kioi`	
饒平音	詞目	饒平音
	瘵	▶ tiam`
詔安音	詞目	詔安音
	瘵	▶ tiam^
南四縣	▶ kioi	
釋義	累、疲勞之意。例：當瘵。	

圖二十二 以拼音「kioi」查找《臺灣客家語常用詞辭典》相關詞目：「瘵」

第二種為借音或借義字，在此以「僱」¹⁸為例，以及兩篇文本作者各自選用不同的用字作為難字替代：

(8) 涯

原文：涯愛你在腰仔項緝一條索仔，

校訂：僱愛你在腰仔項緝一條索仔，

華譯：我要你在腰上綁一條繩子，

涯愛你在腰仔項緝²⁰一條索仔²¹，

圖二十三 借音或借義字舉例一：「涯」（「僱」之非規範用字）

（資料來源：《安徒生童話全集〔第一輯〕（國家語言【臺灣客語——四縣腔】）》（2018: 5））

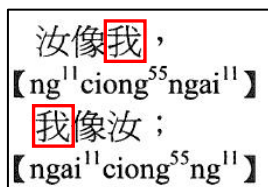
¹⁸ 客語「僱」，華語釋義為「我」。編碼為 U+2028E，屬擴展 B 區。

(9) 我¹⁹ (文本為客語發音)

原文：汝像我，我像汝；

校訂：你像𠵼，𠵼像你；

華譯：你像我，我像你；



圖二十四 借音或借義字舉例二：「我」（「𠵼」之華語用字）

（資料來源：《客家令兒 168》（2003: 151））

藉由這組例子可以觀察到坊間作者早期受限於客語許多難字無法正常顯示，採取了不同的方式呈現，除了選用字符相似的用字作為取代外（例（8）），有些作者則是直接採用華語的用字並加註客語發音（例（9）），其他例子中也常見選用同／近音字或同／近義字作為難字的替代用字。而由於「𠵼」為客語代名詞，表第一人稱單數，華語釋義為「我」，此字於客語相當常用，因此可使用《臺灣客家語常用詞辭典》查詢到正確的客語文字。

第三種是缺字，除了於語句中直接缺字以一個空格呈現外，有些變成符號（例（10）），有些則是變成空白方格（例（11）），下方兩例以「𠵼」²⁰作為示範：

(10) 「『•』爪」修訂為「『𠵼』爪」

原文：往語講：「貓拖案耙難•爪」，

¹⁹ 客語民間作品常有借用華語用字之情況，然文本中並非全然是華語借用字，有些為客語慣用語的用字，文字校訂時須額外留意。以代名詞「𠵼」為例，若文本用字為「我」，而該文本提供拼音為「ngai¹¹」（以四縣腔為例）或語意明確為第一人稱單數者，即可校訂為客語規範字「𠵼」；然客語中另有慣用語「知人我」之用法（四縣腔發音為「di²⁴ ngin¹¹ ngo²⁴」，表明白事理，知所進退分寸），即不可將「我」替換成「𠵼」。

²⁰ 客語「𠵼」，華語釋義為「從主體上落下來」。編碼為 U+3A90，屬擴展 A 區。

校訂：往語講：「貓拖糝粿難**𦉳**爪」，

華譯：俗語說：「貓抓麻糬難脫爪」（就像貓用爪子抓麻糬被黏住而難以掙脫），

往語講：「貓拖糝粿難**𦉳**爪」，

圖二十五 缺字舉例一：「『•』爪」

（資料來源：《97 年度客語能力認證中級暨中高級考試試題【口試】（大埔 D）》（2008: 7））

有一句師父話講：「貓拖糝粿難**𦉳**爪」，

圖二十六 上圖例（大埔腔）之其他腔對照版（四縣腔）：「𦉳爪」

（資料來源：《97 年度客語能力認證中級暨中高級考試試題【口試】（四縣 D）》（2008: 7））

詞目	【𦉳爪】	詞性：動
四縣音	▶ lud`zau`	
海陸音	▶ ludzau´	
大埔音	▶ lud^zau^	
饒平音	▶ lud`zau`	
詔安音	▶ lud´rhiau^	
南四縣	▶ lud`zau`	
釋義	本指爪子脫落，後亦借指人脫離困境或陷阱。例：這件事情恁麻煩，做到一半了，毋知愛仰般，實在像「貓仔拖糝粿，難𦉳爪（又讀lod`zau`南）」。（這事情這麼麻煩，做到一半了，不知如何是好，實在就像貓爪抓到麻糬一樣，難以脫身。）	

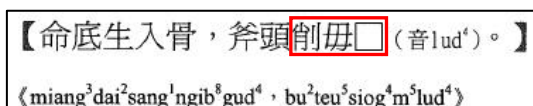
圖二十七 以「拖糝粿」查找《臺灣客家語常用詞辭典》相關詞目：「𦉳爪」

(11) 「削毋『□』」修訂為「削毋『𦉳』」

原文：命底生入骨，斧頭削毋□。

校訂：命底生入骨，斧頭**剝毋**𠵼。

華譯：命裡生入骨，斧頭削不掉（比喻命裡有時終須有，命裡無時莫強求）。



圖二十八 缺字舉例二：「剝毋『𠵼』」

（資料來源：《聽算無窮漢——有韻的客話俚諺 1500 則》（2002: 52））

詞目	【剝毋𠵼】	詞性：動				
四縣音	▶ bog4 m5 lud4					
海陸音	▶ bog4 m5 lud4					
大埔音	▶ bog4 m5 lud4					
饒平音	▶ bog4 m5 lud4					
詔安音	<table border="1"> <thead> <tr> <th>詞目</th> <th>詔安音</th> </tr> </thead> <tbody> <tr> <td>剝毋會𠵼</td> <td>▶ boo⁴ m¹ bboi⁷ lud⁴</td> </tr> </tbody> </table>	詞目	詔安音	剝毋會𠵼	▶ boo⁴ m¹ bboi⁷ lud⁴	
詞目	詔安音					
剝毋會𠵼	▶ boo⁴ m¹ bboi⁷ lud⁴					
南四縣	▶ bog4 m5 lud4					
釋義	剝不下來、剝不開。例：該螺絲仔鎖忒𠵼，剝毋𠵼（又讀bog4 m5 lod4南）。（那個螺絲鎖太緊，剝不下來。）					
對應國語	剝不下、剝不開					

圖二十九 以拼音調類「lud4」查找《臺灣客家語常用詞辭典》相關詞目：「剝毋𠵼」

在兩篇原文中，「𠵼」皆無法正常顯現，於例（10）中可能係因轉檔錯誤變成圓點符號「•」，由於此語料為多腔文本形式，因此可對照其他腔文本，查找到可能的用字；然對比文本（四縣腔）所使用之「𠵼」字非教育部規範用字，因此再使用文句中其他詞彙進行辭典查找，順利查詢到規範用字「𠵼」。而於例（11）則是以一個白色方塊「□」作為代替，這個方塊一般稱為 *tofu*（意即「豆腐」），用以表達無法正常顯示該文字。由於原文本提供

拼音資訊，因此即以拼音搭配調類的方式查詢《臺灣客家語常用詞辭典》，查找到相關詞目「剝毋𪗇」，其中「𪗇」的語音、語意皆符合原文本，因此即可替換之。

第四種則是漢字部件拆解。客語在詞彙方面與華語有許多不同之處，除了單音詞較豐富，也保留許多古漢語字或難字，而現行輸入法中多不支援這些詞彙用字，因此有許多作者採以拆解漢字部件的方式來進行繕打。然而，一旦匯入語料庫之後，拆解的部件即會被系統視為多個字元，造成系統誤判，對於詞頻統計之數據也會產生誤差。部件拆解的文本示例如下：

(12) 「虫念」修訂為「𧈧」

原文：魚塘肚有當多光雞虫念，

校訂：魚塘肚有當多江雞𧈧，

華譯：池塘裡有非常多蝌蚪，

(魚塘肚有當多光雞𧈧，泅來泅去，過一駁做下變做細蛎咧。)

圖三十 漢字部件拆解舉例一：「虫念」

(資料來源：《98 年度客語能力認證中級暨中高級考試試題【口試】(大埔 D)》(2009))

(13) 「火旁」修訂為「𤇀」

原文：在臺灣逢年過節蓋鬧熱，每年个正月半，北天燈、中火旁龍、南蜂炮、東寒單，還有燈會，來參觀个人掖麻掖米。

校訂：在臺灣逢年過節蓋鬧熱，每年个正月半，北天燈、中𤇀龍、南蜂炮、東寒單，還有燈會，來參觀个人掖麻掖米。

華譯：在臺灣逢年過節很熱鬧，每年的正月半，北天燈、中𤇀龍、南蜂炮、東寒單，還有燈會，來參觀的人萬頭攢動(人山人海)。

在臺灣逢年過節蓋鬧熱，每年个正月半，北天燈、中**火**龍、南蜂炮、東寒單，還有燈會，來參觀个人掖麻掖米¹⁰²。正月二十

圖三十一 漢字部件拆解舉例二：「火旁」

(資料來源：《105 年全國語文競賽客家語朗讀文章高中學生組（南四縣腔）》(2016))

(14) 「另(虫部)」修訂為「𧈧」

原文：上晝另(虫部)兒叫，下晝雨就到。

校訂：上晝**𧈧**仔叫，下晝雨就到。

華譯：上午青蛙叫，下午雨就到。

【上晝**另**(虫部)兒叫，下晝雨就到。】

《song³zu³guai²ie⁵gieu³/giau³ · ha¹zu³i²ciu³do³/dau³》

圖三十二 漢字部件拆解舉例三：「另(虫部)」

(資料來源：《聽算無窮漢——有韻的客話俚諺 1500 則》(2002: 14))

例子(12)與(13)之部件拆解方式較為常見，屬單純將部首與偏旁拆開為兩個字元，許多作者為了表示此為一個字，多採調整兩個字元的字元比例。以例(12)「𧈧」為例，經查原本檔案，「虫」與「念」為兩個字元(如圖三十中紅框處)，前者字元比例調整為40%('虫')，後者則為60%(念)，兩字元緊鄰即為「𧈧」，故必須進行文字校訂。另一方式則是作者先列出該難字的偏旁，後接括號說明此字的部首。如例(14)與圖三十二所示，此難字之偏旁為「另」，括號內為部首「虫」，加上作者提供之拼音及調類「guai²」，可得知此字為客語之「𧈧」(華語釋義為「青蛙」)。然因部首往往可出現在每個漢字中的不同位子(常見於上下左右四處)，此種註釋方式較不直觀，就讀者而言不利於理解，對於語料庫系統判讀也增加難度，因此，凡部件拆解之漢字，均會校訂為單個字元之正確客語字。而這三個難字中，僅有「𧈧」可於《臺灣客家語常用詞辭典》查詢到相關資料，其他二字則查無符合的詞

目，且因為皆非屬擴展 A 區漢字，無法透過輸入法繕打出來。這些輸入法沒有收錄的字，語料庫工作人員會參照 Unicode 字元對應表（目前為 15.0 版本，網址為 <http://www.unicode.org/charts/>），其中一區為「中日韓統一表意文字」（CJK Unified Ideographs (Han)），提供了擴展 A 區至 H 區的漢字 PDF 檔案，工作人員即會根據檔案內容所提供各難字及其部首，查找相對應的難字。而上述部件拆解的文字皆屬擴展區文字，如「蚬」為 U+27285，屬擴展 B 區；「焗」為 U+2AE5A，屬擴展 C 區；「蝻」為 U+2C816，屬擴展 E 區。系統字型如新細明體、標楷體、黑體—繁等，皆造滿至擴充 A 區為止的 27,496 字（柯志杰、蘇煒翔 2014: 151），而擴展 B 區至 D 區的難字，多半可藉由新細明體—ExtB 字型正常顯示（此字型專門用以支援擴充 B 區的字符），然由於臺灣部分客語難字及罕用字係收錄於擴展 E 區，為有效解決語料轉寫校訂人員使用客語難字的問題，本語料庫除了採用 Unicode 編碼並支援字集擴充外，亦搭配使用由上地宏一（2017）開發，日本 GlyphWiki 組織共同編輯管理的「花園明朝」²¹開源字型，該字型完整收錄至中日韓統一表意文字擴展 F 區的字符，與 Unicode 10.0 對應，再搭配 Google 開發涵蓋漢字、假名、諺文的開源字型 Noto Sans CJK，能消除 Windows 等系統中無法顯示的字元「□」（tofu）。工作人員於電腦將花園明朝及 Noto Sans CJK 字型檔安裝妥後，搭配文書編輯軟體如 Microsoft Office WORD，再藉由 Unicode 編碼表或《臺灣客家語常用詞辭典》將客語特殊字複製再貼上文書軟體，若出現空白、方框或亂碼，選取該字並更改其字型為 HanaMinB，便可解決客語特殊字的顯示問題。本語料庫根據教育部所公告的兩批推薦用字以及《臺灣客家語常用詞辭典》，綜整歸納出臺灣客語特殊字表，詳加載明字元所對應之 Unicode 編碼及編碼區域，提供工作人員作為參照（如下表，另因篇幅限制，僅呈現部分內容）。

²¹ 花園明朝所包含的字符數量已超過單一字型可容許的上限，因此拆為兩個字型檔。HanaMinA（花園明朝 A）支援中日韓統一表意文字區及其擴充 A 區，HanaMinB（花園明朝 B）則完整支援擴充 B 區至 F 區。更多介紹可參見 <http://fonts.jp/hanazono/>。

表一 臺灣客語特殊字表（部分節錄）

（資料來源：《臺灣客語語料庫》自行整理自《臺灣客家語常用詞辭典》與《臺灣客家語書寫推薦用字（第1批、第2批）》）

教 育 部 用 字	Unicode 編碼	編碼 區域	四縣	海陸	大埔	饒平	詔安	南四縣	釋義	用 法
礮	U+40D7	擴展 A 區	bog2	bog5	bog21	bog2	boo24	bog2	石隄，又作「壘」。	礮坎
箊	U+25BE5	擴展 B 區	cab2	cab5	cab21	cab2	cab24	cab2	一種盛物的農具。	箊箕
迂	U+2B7E7	擴展 D 區	din24	din53	din33	din11	/	din24	繞行。	迂一 圈
偲	U+2B8C6	擴展 E 區	en24	en53	en33	en11	een11	/	稱含說話者在內的「我們」	偲俚
跣	U+47D8	擴展 A 區	hong55	hong11	hong53	/	/	hong55	起身、起來。	跣起 來
孩	U+39E1	擴展 A 區	kai24	kai53	kai33	kai11	kainn11	kai24	挑。	孩水
釧	U+20803	擴展 B 區	qiam24	ciam53	ciam31	ciam11	ciam11	qiam24 qiam11	刺殺。	釧一 刀
僮	U+203B7	擴展 B 區	sab2	sab5	sab54	sab2	sab24	sab2	碎裂。	僮碎

教 育 部 用 字	Unicode	編碼	四縣	海陸	大埔	饒平	詔安	南四縣	釋義	用 法
	編碼	區域								
搵	U+22BED	擴展 B 區	ten55	ten11	ten53	ten53	ten31	ten55	幫助、幫忙。	搵手
脷	U+3B39	擴展 A 區	zang24	zang53	zang33	zang11	zang11	zang24	腳踵，即腳跟 的部位。	腳脷

3.4 一字多碼整併與具體字替換

如前所述，每個字皆有字元編碼，人工校訂用字時，僅能修訂形體明顯差異的文字，然而文本中有許多字形相同卻擁有不同編碼的文字，對於肉眼而言難以辨識。而只要是不同編碼，對於系統而言就是不相同的字元，因此必須透過系統輔助來判斷及整理這類一字多碼的狀況。少數客語特殊字因應早期電子化環境，使用 Unicode 中私人使用區（Private Use Area）碼點作為編碼準則，直到現代客語特殊字多數已被收錄於 Unicode 擴展 A 至 E 區，因此遵循 Unicode 編碼已成為世界趨勢。以「𪗇」為例，此字所屬區段為擴展 D 區，Unicode 編碼為 U+2B7E7，而過去於私人使用區曾使用之編號有 U+E0C8、U+E72C、U+F444、U+FB6B 等。語料庫系統可對新舊字碼進行轉換，將這些難以辨識其差異性的私人編碼區字符，統一替換成中日韓統一表意文字及擴展區段文字。此外，語料庫網頁採用雲端字型（web font）嵌入技術，且將 Unicode 字集擴展 A 至 E 區字型加以整併，克服以往特殊字型須安裝方可顯示之限制，解決網頁端、跨平臺以及行動裝置端客語字形顯示的問題（圖三十三、圖三十四）。

Pk	客語詞目
12422	□
12423	□
16699	□
9068	□
17659	□
9073	□
2864	□人
16631	□俚
14793	□兜
76044	□兜儕

圖三十三 客語雲端字型置入前（難字無法正確顯示）

Pk	客語詞目
12422	儂
12423	齡
16699	吭
9068	辽
17659	鯿
9073	儂
2864	儂人
16631	儂儂
14793	儂兜
76044	儂兜儕

圖三十四 客語雲端字型置入後（難字自動正確顯示）

另外在檢索系統介面，本語料庫開發客語拼音輸入的輔助工具，透過資料模型內部建立的資料，將拼音轉換客語字的功能置入檢索系統，並且提供不限腔調、不分調值的客語字詞推薦，亦即若輸入時省略調值，仍可由清單選取對應的客語字，使用者毋須另外安裝輸入法，目前已有部分用字支援拼音查詢功能。

而無論實體書籍或電子書，出版時可能會因為字型、排版、印刷等關係產生勘誤，常見誤植為異體字的狀況，此係由於漢字歷經悠久的歷史，許多文字產生了形體上的演變，再者因為周邊國家如日本、韓國、越南等地受到漢文化影響，書寫系統也使用或借用漢字，甚或與其固有文字混合使用，久而久之也造成了同一個字在不同環境下有著形體以及語意方面的差異，異體字也因此應運而生。根據教育部（2017）《異體字字典》之編輯說明，異體字係與「正字」相同音義但不同形體，多半為些微的部件差異；而正字即為標準字形，包含常用字、次常用字、罕用字三個標準字體表，異體字則涵括古文字字形、書法特殊字形及簡體字。關於異體字是否應當替換，會因為不同的語言背景面向而有著不同的處理方式。如周亞民與黃居仁（2005）引用裘錫圭（1995）之論述，異體字可區分為全同異體和部分異體，前者指的是音義完全相同而字形不同的字，後者則是只需部分用法相同即可，用法完全相同者稱為狹義異體字，廣義異體字則包含部分異體字和全同異體字。周亞民與黃居仁（2005）更進一步指出，由於計算機編碼系統採用一個字形搭配一個字碼，對於計算機而言，只要字碼不同就是不同的字，例如「說≠說」，因此會造成中文資訊處理（尤其是檢索方面）的問題。

臺灣客語語料庫主要依據 The Unicode Consortium (2018, 2022) 之「漢字統一原則」，凡屬於中日韓統一表意文字之兼容性表意文字與兼容性擴展的文字，均統一為正字。另也依循 Unicode 三維模型來確定文字之間的關係，其中 X 軸代表語意（meaning）、Y 軸代表抽象外觀形狀（abstract shape）、Z 軸則代表風格差異（stylistic variations）。首先以「說」和「貓」為例，此二字在 X 軸上的位置不同，代表這兩個字的語意不同，也就是意味著「說」和「貓」是相異的兩個字，兩者並無變體關係。而「貓」（U+8C93）和「猫」

(U+732B) 在 X 軸的位置相同，在 Y 軸則落在不同位置，表示他們的語意與發音皆相同，抽象形狀不同，因此「貓」和「猫」互為 Y 變體 (Y-variants)。至於「說」(U+8aaa) 和「説」(U+8aac) 在 X 軸和 Y 軸上所落位置皆相同，Z 軸上則相異，代表兩者不僅有著同樣的語意和發音，抽象外觀形狀也一致，僅有在風格差異上不同（例如較小的印刷差異），因此這兩個字即為彼此的 Z 變體 (Z-variants)，原則上語意、發音、外觀形狀皆相同的 Z 變體即可替換為正體字。值得注意的是，Y 變體可再區分為簡體與繁體變體 (Simplified and Traditional Chinese Variants) 以及語意變體 (Semantic Variants)，凡語意完全相同之簡體變體（如「书」(U+4E66)），也可一併予以替換為繁體變體（正體字）（如「書」(U+66F8)）；而語意變體可再區分為 k 語意變體 (kSemantic Variants，²²如「兎」(U+514E) 和「兔」(U+5154)) 以及 k 特殊語意變體 (kSpecialized Semantic Variants，如「井」(U+4E3C) 和「井」(U+4E95))，前者表兩個字擁有完全相同的語意，後者則為重疊語意 (The Unicode Consortium 2022)，因此 k 語意變體可替換為正體字，k 特殊語意變體則不可任意替換（例如，若「井」在文本中的語意與「井」之釋義不同，即不可將「井」替換為「井」）。語料庫另也搭配教育部 (2017)《異體字字典》以及數位發展部 (2022)《CNS11643 中文標準交換碼全字庫》作為參考，凡字音、字義、用法與正字相同之異體字，即可藉由系統建立規則的方式，將異體字替換為正字（或是將簡體變體替換為繁體變體）。目前語料庫收集到的異體字以 Y 變體中的簡體與繁體變體與 k 語意變體為主，簡體變體與其繁體變體如：「酿」(U+917F) 與「釀」(U+91C0)、「罵」(U+99E1) 與「罵」(U+7F75)、「来」(U+6765) 與「來」(U+4F86)，k 語意變體與其正體之範例如下：「寶」(U+5BF3) 與「寶」(U+5BF6)，兩者音義皆同，僅部分形不同；「出」(U+5C80) 與「出」(U+51FA)，兩字音義皆同，俗字筆法小異。異體字替換標準係採較嚴格之判定，二字之間的發音、語意必須均

²² 關於“k”的定義，Unicode (2022) 解釋，Unicode Han 數據庫 (Unihan) 係由許多字段所構成，這些字段名稱則是 ASCII 字母與數字的組合，基於一些歷史原因，名稱都以小寫“k”開頭 (原文：For historical reasons, they all start with a lowercase “k.”)。

相同，通常外形也多有一定程度的相似度，若不符合上述條件，亦不屬於 Z 變體、Y 變體之簡繁變體以及 k 語意變體者，原則上採保留原文樣式示之。有時 Unicode 與教育部（2017）《異體字字典》的分類方式有些微不同，如「祖」與「𠂔」，前者為「衣部」（衤），後者則為「示部」（礻），目前無論是《臺灣客家語常用詞辭典》、《客語認證詞彙資料庫》或《重編國語辭典修訂本》均無收錄「祖」字（衣部），而根據《異體字字典》，「祖」（衣部）有兩種音義，其一發音為ㄗㄨˇ，表事好；²³其二為ㄗㄨˊ，為「祖」之異體，此判定方式即與 Unicode 的 k 特殊語意變體相似。然 Unihan Database（中日韓統一表意文字數據庫）並無認列兩者有著正異體字關係，因此這個例子對語料庫而言，即歸屬於形似字之校正（如 3.2.6 之「形似字勘誤」），修訂如下：

（15）「媽『祖』」修訂為「媽『𠂔』」

原文：媽祖生个時節，路脣人家會準備茶水分進香个人啲。

校訂：媽𠂔生个時節，路脣人家會準備茶水分進香个人啲。

華譯：媽祖生日的時候，路邊人家會準備茶水給進香的人喝。

²³ 教育部（2017）提供「祖」（衣部）為正字之釋義如下：「《說文解字·衣部》：『祖，事好也。』清·段玉裁·注：『事好猶言學好。』」

詞目	【茶水】	詞性：名
四縣音	ca ^v sui [`]	
海陸音	ca shui ^ˊ	
大埔音	ca ^v shui [^]	
饒平音	ca fi [`]	
詔安音	ca [`] fi [^]	
南四縣	ca ^v sui [`]	
釋義	泛指各類飲料。例：媽祖生个時節，路厝人家會準備茶水分進香个人啲。 (媽祖生日的時候，路邊人家會準備茶水給進香的人喝。)	
對應國語	茶水	

圖三十五 形似字舉例：「媽『祖』」

(資料來源：《臺灣客家語常用詞辭典》更新前畫面截圖，取自 2021 年 12 月)²⁴

²⁴ 《臺灣客家語常用詞辭典》之詞目例句係由專家學者所擬，語料相當珍貴，因此語料庫也申請授權並收錄於語料庫。客語定字相當繁複且重要，然詞目例句中仍不免會出現格式或文字錯誤，而該辭典採創用 CC「姓名標示—禁止改作」3.0 臺灣授權條款釋出，故無法依據語料庫用字規範進行文字修訂。團隊工作人員將疑義處彙整後，陸續回饋給教育部權責單位，擬請編輯委員參酌與審訂，目前教育部已排除許多舛訛並同步於網頁上更新，在此也向教育部單位負責人員以及辭典編輯委員致上感謝。

詞目	【茶水】	詞性：名
四縣音	ca ^ˇ sui [`]	
海陸音	ca shui ^ˊ	
大埔音	ca ^ˇ shui [^]	
饒平音	ca fi [`]	
詔安音	ca [`] fi [^]	
南四縣	ca ^ˇ sui [`]	
釋義	泛指各類飲料。例：媽祖生个時節，路脣人家會準備茶水分進香个人欸。 (媽祖生日的時候，路邊人家會準備茶水給進香的人喝。)	
對應國語	茶水	

圖三十六 客語例句內文之「媽『祖』」修正為「媽『祖』」

(資料來源：《臺灣客家語常用詞辭典》更新後畫面截圖)²⁵

另一方面，部分在 Unicode 屬簡繁體變體關係的兩個字於客語中各為正字，因此亦不可替換之。如「個」(U+500B)與「个」(U+4E2A)，前者多擔任客語量詞，為計算人的單位，或表「獨自的」之意，如「個人」；後者則主要擔任客語的結構助詞，功能同華語「的」。這兩個字在 Unicode 分類中也有著多重關係，在不同的語境環境下，「个」可以是「個」的 k 語意變體、簡體變體或是 k 特殊語意變體。因此，異體字之替換必須嚴謹判定以及謹慎處理，而經確認可被更替的異體字與其對應之正體字，兩者的字形與編碼均列入語料庫的後臺資料集，在後續文本匯入時，系統會掃描符合的字符，將這些異體字字符自動轉換成正體字字符，並支援以全域文本為目標進行批次取代。

依循教育部的用字規範既已解決大部分的用字問題，其餘字詞若教育部

²⁵ 此詞目之修訂公告可參見：

<https://hakkadict.moe.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=tW17IT/fulltext?fulltext=N0000001001&dbpath=/opt/fb32/db/newsdb.db>。

及客家委員會均未規範或收錄，這一類官方未定的作者選用字，由於其著書內容幾無提供拼音或註釋可以佐證該字詞之音義，暫採忠於原著方式保留作者原用字不予校訂，未來一旦教育部或客家委員會官方發布用字更新時，確認發音與語意及用法相同者，即可依循公部門之規範予以替換。

4. 結論

臺灣客語語料庫於 2022 年 10 月完成第一階段之建置，開放大眾使用，從搜整授權取得之各式文獻與資料可以看到公部門對整合的努力以及民間不斷地嘗試創作，其中書面語料因收錄著作之成書年代跨度較長，不同作者的用字與拼音書寫多有歧異，客語用字正確性判斷之難度相對較高，往往需要較多的時間與人力進行校訂作業。臺灣客語語料庫的建置目標，除了希望可以透過現代科技將語料數位化保存客語文本外，並基於兩大公部門——教育部歷經十餘年所研擬之一套具規模之用字規範與《臺灣客家語常用詞辭典》的建置，以及客家委員會歷年來辦理之客語能力認證與基本詞彙的制訂，展示語料校訂狀況，包括客語拼音校訂為客語漢字、客語用字統整、多字刪除、缺字補齊、顛倒字序調換、形似字勘誤等。此外，客語難字無法正常顯示的四種情形，包括拼音、借音或借義字、空格或符號（缺字）、漢字部件拆解等等，也均依據官方用字標準，逐一修訂為正確的規範用字，而一字多碼以及異體字的問題，也藉由電腦技術加以整併。

本文的用字校正均有依可循，亦即可依據官方經過審音、定字等嚴謹程序後所發布的標準用字進行修訂，然而，臺灣客語仍有許多用字尚未統一，次方言間詞彙不同所造成的差異，以及客語因瀕危而衍生的詞彙丟失現象，在在讓客語本字考訂面臨重重困難。客語用字標準須六腔併陳，且本字的考證必須從文字、聲韻、訓詁等學理方面著手，以求形音義三個層面皆能契合，在客語用字研究中是一塊相當重要且正在開闢中的領域，也是個值得更深入探討的議題。本研究重點著眼於語料庫語言學之學科基礎上，透過大數據——亦即語料庫真實語料的用字表現，展示語料庫各種用字校訂狀況並提出

相關因應策略與剖析，以及如何運用電腦科技處理客語文本中的一字多碼與異體字，希冀可作為拋磚引玉的開端。藉由語料庫的建置，忠實保存臺灣客語多元的使用樣貌與豐富的自然語言表現，提供使用者獲取所需語料進行語言觀察、分析與語言學相關學理驗證；同時，語料數位化，能提供計算語言學跨領域研究與應用，結合統計與電腦演算法，透過電腦辨識或機器學習的技術以及語料庫檢索與分析工具，讓語料庫系統與客語數位化語料之應用發揮最大效益。此外，利用語料庫可作為文字保存承載利器之一大優勢，將尚未標準化、各作者選用不同用字的詞彙差異記錄下來，提供公部門以及致力於客語用字制訂的專家學者參考，期許可以為教育部以及客家委員會對於臺灣客語的用字整合與制訂略盡一己之力，延續臺灣客語書寫系統的活躍發展。

引用文獻

- Aijmer, Karin and Anna-Brita Stenström. 2004. *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: Benjamins.
- Aijmer, Karin. 2009. So er I just sort I dunno I think it's just because...: a corpus study of I don't know and dunno in learners' spoken English. In Jucker, Andreas, Daniel Schreier and Marianne Hundt (eds.), *Corpora: Pragmatics and Discourse: Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*, 151-168. Amsterdam: Rodopi.
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Carter, Ron and Michael McCarthy. 1997. *Exploring Spoken English*. Cambridge: Cambridge University Press.
- Chuang, Fei-Yu and Hilary Nesi. 2006. An analysis of formal errors in a corpus of Chinese student writing. *Corpora* 1: 251-271.
- Ensslin, Astrid and Sally Johnson. 2006. Language in the news: investigations into representations of 'Englishness' using WordSmith Tools. *Corpora* 1: 153-185.
- Gabrielatos, Costas, Eivind Torgerson, Sebastian Hoffmann and Susan Fox. 2010. A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics* 38: 1-38.
- Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar (2nd Edition)*. London: Edward Arnold.
- Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Johansson, Stig. 2007. *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.

- MacIver, Donald. 1905. *An English-Chinese Dictionary in the Vernacular of the Hakka People in the Canton Province*. Shanghai: American Presbyterian Mission Press.
- McEnery, Tony and Andrew Hardie. 2013. The history of corpus linguistics. In Allan, Keith (ed.), *The Oxford Handbook of the History of Linguistics*, 727-746. Oxford: Oxford University Press.
- Rey, Charles. 1901. *Dictionnaire Chinois-Français, Dialecte Hac-Ka*. Hong Kong: Imprimerie de la Société des Missions Etrangères.
- Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8: 209-243.
- The Unicode Consortium. 2018. Unicode standard annex 38: Unicode Han database (UNIHAN). Retrieved from <http://www.unicode.org/reports/tr38/tr38-25.html> (January 9, 2022).
- _____. 2022a. Unicode 15.0 character code charts. Retrieved from <http://www.unicode.org/charts/> (December 2, 2022).
- _____. 2022b. Unicode 15.0.0. Retrieved from <https://www.unicode.org/versions/Unicode15.0.0/> (December 2, 2022).
- _____. 2022c. Unicode standard annex 38: Unicode Han database (UNIHAN). Retrieved from <https://www.unicode.org/reports/tr38/tr38-33.html> (December 2, 2022).
- Wong, May. 2006. Corpora and intuition: a study of Mandarin Chinese adverbial clauses and subjecthood. *Corpora* 2: 187-216.
- Xiao, Zhonghua and Tony McEnery. 2004. A corpus-based two-level model of situation aspect. *Journal of Linguistics* 40: 325-363.

- 上地宏一. 2017. 「花園フォント (花園明朝) [電腦軟體]」。取自：
<http://fonts.jp/hanazono/> (查詢日期：2022.01.09)。
- 中原週刊社客家文化學術研究會. 1992. 《客話辭典》。苗栗：臺灣客家中原週刊社。
- 王雅萍、張如瑩、陳秀華、蕭貴徽. 2012. 《數位化工作流程指南：文字資料》。
臺北：數位典藏拓展臺灣數位典藏計畫。
- 安徒生著；謝杰雄等譯. 2018. 《安徒生童話全集〔第一輯〕(國家語言【臺灣客語——四縣腔】)》。臺北：龍岡數位文化。
- 江敏華、黃彥菁、宋柏賢. 2009. 〈客語文獻分析與數位典藏——以客英、客法大辭典為例〉。《教育資料與研究雙月刊》91: 131-160。
- 行政院主計總處. 2021. 《109 年人口及住宅普查初步統計結果》。取自：
<https://www.dgbas.gov.tw/public/Attachment/1831151816OM26MHO7.pdf>
(查詢日期：2022.04.09)。
- 行政院客家委員會. 2006. 《95 年度臺灣客家民眾客語使用狀況調查》。取自：
<https://www.hakka.gov.tw/file/Attach/1990/1/891015293071.pdf> (查詢日期：2022.04.09)。
- _____ . 2008a. 《97 年度客語能力認證中級暨中高級考試試題【口試】(大埔 D)》。臺北：行政院客家委員會。
- _____ . 2008b. 《97 年度客語能力認證中級暨中高級考試試題【口試】(四縣 D)》。臺北：行政院客家委員會。
- _____ . 2008c. 《臺灣饒平、大埔、詔安客語辭典——大埔分冊》。取自：
<https://cloud.hakka.gov.tw/site/hakka/public/Attachment/810161145271.pdf>
(查詢日期：2022.04.09)。
- _____ . 2008d. 《臺灣饒平、大埔、詔安客語辭典——詔安分冊》。取自：
<https://cloud.hakka.gov.tw/site/hakka/public/Attachment/8101611493671.pdf>
(查詢日期：2022.04.09)。

- _____. 2008e. 《臺灣饒平、大埔、詔安客語辭典——饒平分冊》。取自：
<https://cloud.hakka.gov.tw/site/hakka/public/Attachment/8101611405671.pdf> (查詢日期：2022.04.09)。
- _____. 2009. 《98 年度客語能力認證中級暨中高級考試試題【口試】(大埔 D)》。臺北：行政院客家委員會。
- 何石松、劉醇鑫. 2002. 《現代客語詞彙彙編》。臺北：臺北市民政局。
- _____. 2004. 《現代客語詞彙彙編續編》(初版)。臺北：臺北市政府客家事務委員會。
- _____. 2007. 《客語詞庫：客語音標版》。臺北：臺北市政府客家事務委員會。
- 李如龍. 1993. 〈從詞彙看閩南話和客家話的關係〉，曹逢甫、蔡美慧主編《第一屆臺灣語言國際研討會論文集》，5.01-5.25。臺北：文鶴出版社。
- 周亞民、黃居仁. 2005. 〈異體字語境關係的分析與建立〉。第十七屆自然語言與語音處理研討會論文。2005 年 9 月 15-16 日。臺南：成功大學。
- 邱湘雲. 2004. 〈臺灣閩客方言比較研究的意義及其語言比較〉。《問學》6: 55-84。
- _____. 2013. 〈海陸客家詞彙的趨同趨異表現〉。《臺灣語文研究》8.2: 61-98。
- 姚榮松. 1998. 〈閩客共有詞彙中的詞源問題〉。《中國學術年刊》19: 659-672。
- 客家委員會. 2022a. 《110 年全國客家人口暨語言基礎資料調查研究》。取自：
https://www.hakka.gov.tw/File/Attach/37585/File_96737.pdf (查詢日期：2022.07.07)。
- _____. 2022b. 《哈客網路學院》。取自：
<https://elearning.hakka.gov.tw/mooc/index.php> (查詢日期：2022.09.29)。
- _____. 2022c. 《客語能力認證參考詞彙·高級(四縣腔上冊)》。新北：客家委員會。
- _____. 2022d. 《客語認證詞彙資料庫》。取自：
<https://elearning.hakka.gov.tw/hakka/dictionary> (查詢日期：2022.09.29)。

- _____. 2022e. 《臺灣客語語料庫》。取自：<https://corpus.hakka.gov.tw> (查詢日期：2022.10.28)。
- 柯志杰、蘇煒翔. 2014. 《字型散步：日常生活的中文字型學》。臺北：臉譜。
- 洪惟仁. 1992. 《臺灣方言之旅》。臺北：前衛出版社。
- _____. 2013. 〈臺灣的語種分布與分區〉。《語言暨語言學》14.2: 315-369。
- 胡萬川編. 2006. 《龍潭鄉廖德添客語專輯(一)》。桃園：桃園縣政府文化局。
- 徐兆泉. 2001. 《臺灣客家話辭典》。臺北：南天。
- _____. 2009. 《臺灣四縣腔海陸腔客家話辭典》。臺北：南天。
- 徐登志. 2005. 《臺灣大埔音客語辭典》。臺中：臺中縣寮下文化學會。
- 徐貴榮. 2005. 《臺灣饒平客話》。臺北：五南。
- _____. 2018. 《饒平客家調查與語言論輯》。臺北：五南。
- 涂春景、廖月娥. 2003. 《客家令兒 168》。臺北：涂春景。
- 涂春景. 2004. 〈客語教學與漢字——從客話的異讀探尋客語本字〉。《臺灣語文研究》2: 155-170。
- 涂春景編. 2002. 《聽算無窮漢——有韻的客話俚諺 1500 則》。臺北：涂春景。
- 張屏生. 2002. 《雲林縣崙背鄉詔安腔客家話語彙初集稿》。著者自印。
- 張捷明. 2018. 《中大湖个風：桃園地區新舊兩隻移墾地區个故事客語文選集》。臺北：華夏書坊。
- 教育部. 2003. 《臺灣客語通用拼音方案》。取自：
https://acdm.tcssh.tc.edu.tw/teach/parent_%20language/parent/book/book_17.pdf (查詢日期：2022.01.09)。
- _____. 2009. 《臺灣客家語書寫推薦用字(第1批)》。取自：
https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=440&content_sn=24 (查詢日期：2022.01.09)。
- _____. 2011a. 《臺灣客家語書寫推薦用字(第2批)》。取自：
https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=440&content_sn=25 (查詢日期：2022.01.09)。

- _____. 2011b. 《臺灣閩南語常用詞辭典》。取自：
https://twblg.dict.edu.tw/holodict_new/（查詢日期：2022.04.09）。
- _____. 2012. 《客家語拼音方案使用手冊》。取自：
<https://ws.moe.edu.tw/001/Upload/6/refile/7803/38403/67447334-bef4-4c69-bc24-32090b745031.pdf>（查詢日期：2022.01.09）。
- _____. 2017. 《異體字字典》。取自：<https://dict.variants.moe.edu.tw/variants>
（查詢日期：2022.04.09）。
- _____. 2019. 《臺灣客家語常用詞辭典》。取自：<https://hakkadict.moe.edu.tw/>
（查詢日期：2022.01.09）。
- _____. 2021. 《重編國語辭典修訂本》。取自：<https://dict.revised.moe.edu.tw/>
（查詢日期：2022.01.09）。
- _____. (n.d.). 〈臺灣客家語書寫推薦用字漢字選用原則〉。取自《語文成果網》http://ws.moe.edu.tw/001/Upload/6/RelFile/6507/7822/hkwords_principle.pdf
（查詢日期：2022.07.07）。
- 教育部編. 2008. 《臺灣客家語朗讀文章選輯》。臺北：教育部。
- _____. 2012.《105 年全國語文競賽客家語朗讀文章高中學生組(南四縣腔)》。
取自：
<http://campus.ckgsh.ntpc.edu.tw/mediafile/2058002/fdownload/248/1234/2017-4-12-16-12-29-1234-nf1.pdf>（查詢日期：2022.04.09）。
- _____. 2016a. 《105 年全國語文競賽客家語朗讀文章社會組（海陸腔）》。取自：
<https://eb1.hcc.edu.tw/edu/pub/downfiles.php?recid=64101&fid=1>（查詢日期：2022.04.09）。
- _____. 2016b. 《105 年全國語文競賽客家語朗讀文章社會組（饒平腔）》。取自：
https://jweb.kl.edu.tw/userfiles/1238/document/25904_105%E5%B9%B4%E5%AE%A2%E5%AE%B6%E8%AA%9E%E6%9C%97%E8%AE%80%E6%96%87%E7%AB%A0_%E9%A5%92%E5%B9%B3%E8%85%94.doc
（查詢日期：2022.04.09）。

- _____. 2017.《教育部電子報「閱讀越懂閩客語」專欄(客家語文章106年)》。
取自：
[https://language.moe.gov.tw/files/people_files/%E2%80%BB%E9%96%B1%E8%AE%80%E8%B6%8A%E6%87%82%E9%96%A9%E5%AE%A2%E8%AA%9E%E6%96%87%E7%AB%A0%E7%AF%87%E7%9B%AE\(%E5%AE%A2106%E5%B9%B4\).pdf](https://language.moe.gov.tw/files/people_files/%E2%80%BB%E9%96%B1%E8%AE%80%E8%B6%8A%E6%87%82%E9%96%A9%E5%AE%A2%E8%AA%9E%E6%96%87%E7%AB%A0%E7%AF%87%E7%9B%AE(%E5%AE%A2106%E5%B9%B4).pdf) (查詢日期：2022.04.09)。
- 曾彩金. 2010.〈六堆客家詞彙庫編纂計畫〉。《六堆雜誌》140: 14-15。
- _____. 2019.《六堆詞典》。屏東：財團法人屏東縣六堆文化研究學會。
- 曾學奎. 2003.《臺灣客家〈渡臺悲歌〉研究》。新竹：國立新竹師範學院臺灣語言與語文教育碩士論文。
- 黃宜範. 1993.《語言、社會與族群意識——臺灣語言社會學的研究》。臺北：文鶴出版社。
- 黃榮洛. 1989.《渡臺悲歌：臺灣的開拓與抗爭史話》。臺北：臺原出版社。
- 楊政男、龔萬灶、徐清明. 2013.《客語辭典》。苗栗：楊政男、龔萬灶、徐清明。
- 葉秋杏、賴惠玲、劉吉軒. 2021.〈臺灣客語語料庫建置與客語詞彙使用初探〉。《數位典藏與數位人文》8: 75-131。
- 裘錫圭. 1995.《文字學概要》。臺北：萬卷樓。
- 詹益雲. 2003.《海陸客語字典》。新竹：詹益雲。
- 劉火欽編. 2011.《2011 苗栗縣第 14 屆夢花文學獎得獎作品專輯》。苗栗：苗栗縣政府。
- _____. 2012.《2012 苗栗縣第 15 屆夢花文學獎得獎作品專輯(一)》。苗栗：苗栗縣政府。
- _____. 2013.《2013 苗栗縣第 16 屆夢花文學獎得獎作品專輯(一)》。苗栗：苗栗縣政府。
- 劉惠萍、范姜烜欽編. 2009.《花蓮客家民間文學集》。花蓮：花蓮縣文化局。
- 數位發展部. 2022.《CNS11643 中文標準交換碼全字庫》。取自：
<http://www.cns11643.gov.tw> (查詢日期：2022.07.07)。

- 滕暢. 2017. 〈閩南語與客家語同源詞本字考——淙、淋、滿〉。《臺灣語文研究》12.2: 217-240。
- 鍾榮富. 2014. 〈臺灣客家話在地化現象之考察〉。《臺灣語文研究》9.1: 29-54。
- 羅肇錦. 1990. 《臺灣的客家話》。臺北：臺原出版社。
- _____ . 1991. 〈閩客方言與古籍訓釋〉。《聲韻學論叢》3: 405-433。
- 羅肇錦、胡萬川編. 1998. 《苗栗縣客語故事集》。苗栗：苗栗縣立文化中心。
- 龔萬灶. 2004. 《阿啾箭个故鄉》。苗栗：龔萬灶。

[2022年12月8日收稿；2023年2月20日修訂；2023年3月10日接受刊登]

葉秋杏
國立政治大學英國語文學系
csyeh.corpus@gmail.com

賴惠玲
國立政治大學英國語文學系
hllai.nccu@gmail.com

Rare Characters, Missing Characters and Character Variants in Taiwan Hakka: An Exploration from Corpus Construction

Chiou-Shing YEH, Huei-Ling LAI
National Chengchi University

The digitization of Taiwan Hakka data is immensely complicated due to the many rare characters, missing characters, or character variants found in Taiwan Hakka texts, and is further hindered by inconsistency between non-governmental Hakka dictionaries' writing practice and governmental standards for the Hakka writing system. This study describes how the Taiwan Hakka Corpus Project carried out character correction to ensure the Corpus's usefulness and robustness. First, the study demonstrates the various types of character correction that take place in our text cleaning process, including converting Hakka spellings into characters, unifying different forms of the same word, deleting redundant or repeated characters, filling in missing characters, swapping reversed characters, and correcting characters similar in shape but dissimilar in meaning. Second, we investigate situations in which rare characters cannot be shown properly, and we provide solutions to each situation. These situations include rare characters in Hakka texts being substituted with (1) Hakka spellings, (2) phonetic or semantic loan characters, (3) unintended glyphs such as squares or symbols (i.e., missing characters), and (4) character decomposition. Finally, issues related to multiple codes for the same character and character variants in Hakka texts are tackled.

Key words: rare character, missing character, character variant, multiple codes for the same character, Taiwan Hakka Corpus

