全球客家研究，2021 年 11 月
第 17 期，頁 49-100

臺灣客語語料之數位化

葉秋杏 *
國立政治大學博士級研究人員
賴惠玲
國立政治大學英國語文學系特聘教授

本文旨在闡述臺灣客語語料庫之語料數位化，耙梳其流程整體脈絡並廓清文本授權與客語用字問題。語料作業流程係由「前置作業」與「數位化及檔案管理」兩大階段串聯，在「前置作業」中包含「語料盤點」、「語料徵集與授權」兩大步驟；而「數位化及檔案管理」則囊括「語料建檔與後設資料標註」、「語料數位化與資料清理」（含語料轉寫校訂）和「語料儲存與管理」三個部分。臺灣客語語料庫的重要性在於其為臺灣第一個書面語料與口語語料兼具且附口語錄音檔的帶標記語料庫，以系統化方式收錄臺灣客語六腔語料。藉由臺灣客語語料庫實際建構經驗，本文希望能發揮「鑒往知來」之效，提供其他專家學者參考，以應用到臺灣其他語言之語料庫建構，更希冀能為語言學與資訊科學之跨領域研究開創新機。

關鍵字：臺灣客語語料庫、語料數位化、授權、後設資料、語言典藏

* E-mail: csyeh.corpus@gmail.com
投稿日期：2021 年 3 月 9 日
接受刊登日期：2021 年 9 月 28 日

The Digitalization of Corpus Data in Taiwan Hakka Language

Chiou-shing Yeh**
Post-Doctoral Research Fellow, National Chengchi University
Huei-ling Lai
Distinguished Professor, Department of English, National Chengchi University

This paper lays out the digitization of corpus data in Taiwan Hakka Corpus, resolving the issues of texts authorization and Hakka character at the same time. The main task encompasses two stages: "preprocessing operation" and "digitization of corpus data and document management". Taiwan Hakka Corpus with both written and spoken varieties (audio recordings available) of Taiwan Hakka language collected in a systematic manner is the first part-of-speech-tagged corpus among Taiwanese native languages. Its construction has taken the initiative in setting a model for corpus construction of other national languages in Taiwan. This paper demonstrates a significant reference for the development of interdisciplinary research on linguistics and computer science.

Keywords: Taiwan Hakka Corpus, Digitalization of Corpus Data, Authorization, Metadata, Language Archive
** Date of Submission: March 9, 2021
Accepted Date: September 28, 2021

一、前言

臺灣是一個多元語言文化的社會，語言使用以臺灣華語最為廣泛，本土語言則包含臺灣閩南語、臺灣客語，及臺灣原住民族語。然而，社會變遷及經濟發展造成臺灣本土語言面臨了嚴重的語言流失。客家委員會（以下簡稱客委會）為瞭解各縣市客家民眾使用客語的情況，自 2004 年迄今多次進行全國客家人口暨語言基礎調查，最近一次報告為客委會 2017 年發佈之《105 年度全國客家人口暨語言基礎調查資料研究》（客家委員會 2017）。研究報告中顯示，會說客語與會聽客語的民眾數量比例呈現逐年下降的趨勢，且該期調查中「完全聽得懂客語」的客家民眾占樣本數的 46.5%，「會說流利客語」則占 46.8%，比例皆未達樣本數的五成；60 歲以上會聽、說客語的比例分別為 87.4% 及 77.8%；13 歲以下則下探至 31% 及 13%。此外，臺灣客家民眾使用腔調最高者為四縣腔，占比 58.4%；其次是比例為 44.8% 的海陸腔；而同屬少數腔之大埔腔、饒平腔及詔安腔僅分別占 4.1%、2.6%、1.7%，顯示少數腔面臨著更加嚴峻之語言流失困境。聯合國教科文組織（UNESCO）於 2010 年發布《世界瀕危語言地圖》（*Atlas of the World's Languages in Danger*）（Moseley 2010）將語言瀕危性鑑定分爲 6 個等級：無危（safe；此語言被群體內所有年齡層使用，且跨年齡層傳播順暢無阻）、脆弱（vulnerable；此語言絕大多數小孩使用，但可能侷限於某特定場域（如在家））、明確瀕危（definitely endangered；此語言不再被小孩作為母語在家學習）、嚴重瀕危（severely endangered；此語言多被祖父母或更

上一代年齡層使用；父母可能可理解，但不會對小孩或同齡層使用）、極度瀕危（critically endangered；此語言僅被祖父母一代局部或鮮少使用）、滅絕（extinct；此語言自 1950 年以來已失去所有使用者）。以此鑑定臺灣客語的現況，確實已接近明確瀕危。賴惠玲（2008）指出臺灣客語正處於瀕臨死亡的危機，當使用客語的人數越來越少，客語就處於凋零，甚至於走向死亡的階段，當時即呼籲應積極推動臺灣客語語料庫的建置，以忠實保存與記錄客語的樣貌。語料庫具有語言典藏、保存紀錄與永續傳承之重要價值，對於面臨急遽流失之臺灣客語而言，建置數位化的書面及口語語料庫，保存典藏書面文字以及口說語音資料，不僅刻不容緩，也是讓臺灣本土語言永續傳承之有效方式。有鑑於此，客委會自 2017 年底開始推動「建置臺灣客語語料庫」計畫，[1] 並委託政治大學團隊執行，目標為國際間第一個同時收錄客語書面與口語文本的語料庫，希冀藉由系統性地蒐集六腔[2] 語料並將客語進行數位化處理與保存，以利後續研究、教學、推廣及延伸應用，而語料庫之三個少數腔的語料收錄比例以占總體至少 3% 為目標，期能為客語稀缺腔調留下彌足珍貴之語言紀錄。

接近明確瀕危的臺灣客語，在語料庫建置時面臨到許多挑戰，語料蒐集之初即遭受語料盤點的困境。首先是用字問題，儘管官方與民間已有為數不少的客語書面文本或口語音檔，然客語仍有許多詞彙有音無

1　此案執行日程為 2017 年 12 月 25 日起至 2022 年 10 月 31 日止，共計五年，主要項目為語料庫架構與資訊系統建置、系統設計與各項相關功能開發擴充與維護（如搜尋功能、詞性標記功能、斷詞功能、介面管理功能等），以及語料蒐集目標達到至少 600 萬字書面語料以及 40 萬字口語語料。

2　根據 2010 年公布的《客家基本法》，客語指的是臺灣通行之四縣、海陸、大埔、饒平、詔安等客家腔調。教育部於 2012 年將「南四縣」自四縣腔中獨立，列為第六個客語腔調。

字，且大部分有音無字之用字制訂仍有爭議；而客語與華語同屬漢字系統，許多客語文法與詞彙在使用上與華語類似，加上不同作者對於客語文字掌握度與熟悉度不一，有些客語母語人士在創作時借用華語進行書寫，導致有許多客語文本內容夾雜華語漢字以及不同系統之拼音（多為漢語拼音或通用拼音）。此外，因客語特殊性，許多語料文本會選用特殊字體來表達客語罕用字（如揇[3]），然而大多數的輸入法及網頁無法支援罕用字之輸入與顯示。在仍有大量客語用字尚待官方制訂之下，上述種種因素皆造成客語用字系統紊亂，導致臺灣客語各家用字遣詞百家爭鳴之狀況。儘管教育部為了改善坊間本土語言教材之用字不一情形，已分別於民國 98 年與 100 年發布第一批與第二批「客家語書寫推薦用字」，然總數僅達 564 字，無法因應臺灣客語所有的詞彙量。雖然教育部另設有「臺灣客家語常用詞辭典」，客委會也制訂「客語認證詞彙資料庫」，然官方用字多有不一，許多特殊字也依舊只有拼音沒有相對應的文字。此外，即便是客語大腔（四縣（含南四縣）、海陸）的語料來源，仍遠遠不及華語語料的豐碩；而具獨特性之客語少數腔（大埔、饒平、詔安），語料徵集尤為艱難。書籍絕版無法獲得或是口語檔案品質欠佳影響文字轉寫正確率等，也是常遭遇到的問題。

另一個棘手挑戰即是語料授權歸屬釐清。在授權取得上，早期的著作權概念不盡成熟，過往許多官方（如政府部門或各地方政府機構）補助或策劃執行活動產製的出版品，常有原作者或著作權不清的問題，導致版權歸屬不明，因此在進行語料數位化之前，必須妥當處理著作權相關問題。江俊龍（2013）在其計畫《東勢客語故事採集整理暨「臺灣客

3 華語釋義為「幫忙」。字型資訊為 Unicode: U+22BED，屬「中日韓統一表意文字擴展區 B」。

家語語料庫」的增建》報告中也提及：「…惟因配合語料庫後設資料的更新擴充，以及版權問題尚未全部解決，因此未能全面開放給網路大眾使用。有鑑於此，深感妥善取得版權實為數位典藏工程的當務之急…」，顯見語料數位化及後續應用加值工作每每涉及著作權議題。而在聯繫過程中常遇授權主體不明之情形，或是共同作者並非每位都願意授權等情況，因此，在授權階段必須確認出版方及作者方的權利結構，方可釐清兩造間擬定之授權形式及著作權歸屬等相關規範。由於具法治結構之授權體系發展至今已日趨完備，即便順利與作者取得文本授權，為求謹慎，仍必須進一步廓清公部門與私人單位、個人之間的權利義務關係，施以不同處理原則洽談授權。因此，授權書擬定與效力釐清、授權歸屬狀態之確認格外重要。

再者，無論國內外，相較於書面語料庫，口語語料庫較少，原因之一為書面語式較為正式，多數經過潤飾或修訂，文句較為連貫且合乎語法，階層結構也較完整緊湊並具規範性，斷詞處理較為容易；而口語則多為非正式語式，文句較豐富多變，然較不連貫，甚或支離破碎，階層結構較鬆散且相對不具規範性，對於斷詞是一大挑戰。此外，口語語料庫須額外規劃音檔相關設計，相較於單一模態的書面語料庫，口語語料庫的建構更為困難。

語言本就帶有多模態特色，文字或是話語即屬於不同類型的信號，隨著科學技術不斷發展，自然語言訊息處理以及語音合成、語音辨識、人機互動等相關應用與發展皆為日益受到關注的重要議題。儘管口語語料蒐集與採錄相當不易，但真實且自發性的（spontaneous）口語語料，可提高自然口語語音合成的表現效率，亦有助於語音辨識精確度；口語

也反映母語使用者的口音、自然語流及情境化的慣用詞藻，因此客語口語語料的收錄彌足珍貴。本文將透過臺灣客語語料庫建置經驗，介紹臺灣客語語料數位化流程，並就上述語料庫所面臨的幾項挑戰（如用字規範問題、語料授權歸屬釐清、書面與口語語料轉寫校訂以及口語語料之聲音處理等）提出一些建議。兼具書面與口語語料的臺灣客語語料庫，其語料數位化的每一階段繁瑣且環環相扣。如何將為數眾多的語料予以組織分類至為關鍵，不但可讓工作人員清楚地掌握語料數量以及處理進程，進而系統性地控管，也提供未來使用者進行語料探勘觀察與比對分析。分類後的語料如何進行資料清理同樣相當關鍵，由於臺灣客語腔調種類繁多，六腔各自之特有用法及字詞產生的歧異性，也會影響電腦處理及判斷。用字規範對於詞頻統計數據亦有著重要影響力，在語料的文字轉寫與校訂過程中，如何顧全客語用字的規範性，同時留意跨腔調之間的特殊性，甚或區分客語和其他語言之間的混用情況等，皆需謹慎處理。此外，臺灣客語語料庫中的所有書面或口語語料均會標註斷詞與詞性標記，與一般傳統的多半只提供文字的一般傳統無標語料庫相比，帶標記語料庫更增加語料之附加價值，使用者可經由斷詞與詞性標記更加理解客語詞彙的語言資訊，而透過斷詞系統程式，即可在龐大的資訊量中計算獲取所需的統計資訊，如詞彙頻率與共現關係等等。藉由臺灣客語語料庫的建置，本文擬聚焦於臺灣客語書面及口語語料數位化流程與內容，分享臺灣客語語料庫前期作業的語料盤點徵集與授權處理之實務經驗，同時將語料處理作業進行脈絡性爬梳與歸納性整理，嘗試建構一套語料庫數位化流程之概貌，兼顧臺灣六腔詞彙之共性與殊性，希冀能為其他語料庫的建置提供建設性的參考。

二、文獻探討

語料庫建置無論是世界之強勢語言或瀕危語言均有成功的案例。本節將以世界具代表性的兩大語料庫：英國國家語料庫（British National Corpus（BNC））及美國當代英語語料庫（Corpus of Contemporary American English（COCA））為引，介紹國內外不同類型語料庫的建置實例及其功能設計。以下將針對英語語料庫與臺灣語言語料庫進行概述。

（一）英語語料庫

相較於國內而言，國外語料庫的建置起步較早。最早的機讀語料庫為英國布朗大學於 1960 年代所建立之布朗大學現代美式英語標準語料庫（Brown University Standard Corpus of Present-Day American English（Brown Corpus），通稱為布朗語料庫）。此語料庫收集 500 個語料文本，總計達約 100 萬字，並由專家歸類為 15 種不同的文類，包含報刊、宗教、技藝與嗜好、熱門影片、學習、小說、幽默、其他等（Kučera and Francis 1967）。自布朗語料庫之後，國外語料庫蓬勃發展，現今較著名的國家型英語語料庫，當屬英國國家語料庫（British National Corpus（BNC））及美國當代英語語料庫（Corpus of Contemporary American English（COCA））。英國國家語料庫（BNC）建置於 1980 年代至 1990 年代初期，並由產學界組成集團共同運作，產業界包括英國牛津出版社（Oxford University Press）、朗文出版（Longman，現為培生教

育出版（Pearson Education））、樂思出版（Larousse Kingfisher Chambers）；學界則包含牛津大學計算服務中心（Oxford University Computing Services）、蘭開斯特大學的計算機語言庫研究中心（University Centre for Computer Corpus Research on Language（UCREL））以及大英圖書館的研究與創新中心（British Library's Research and Innovation Centre）。該語料庫廣泛收錄 20 世紀後半的語料共約 1 億字，其中 90% 為書面語料（含報紙、學術期刊、學術專書、小說、書信手稿等）、10% 為口語語料（含非正式自然會話、正式的商業或政府會議、廣播節目等）。此語料庫檢索頁面之呈現如圖 1 所示。

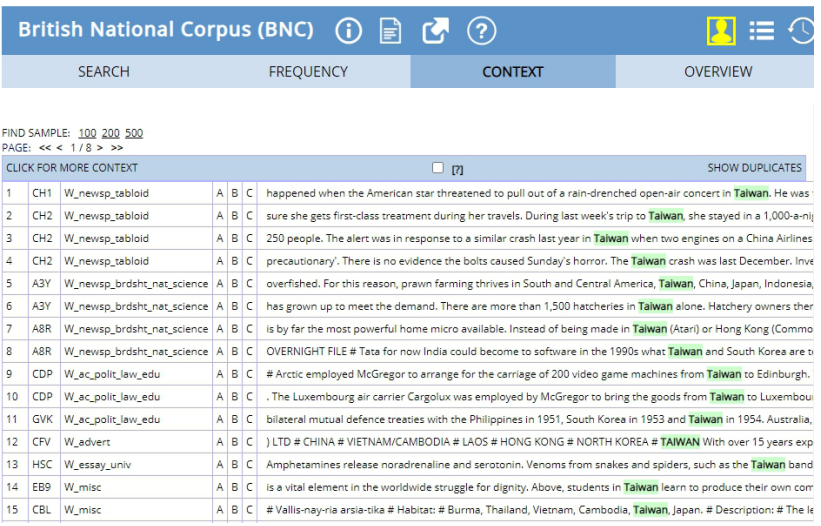


圖 1　BNC 語料檢索顯示頁面

資料來源：作者截圖自 The British National Corpus（2007）。

美國當代英語語料庫（COCA）則於 2008 年由美國楊百翰大學語

言學教授 Mark Davies 建立，收錄 1990 年至 2019 年共約 10 億字語料，可供一般使用者使用，文類包含八個類型：口說、小說、流行雜誌、報紙、學術文章，以及於 2020 年 3 月新增之電視與電影字幕、部落格以及其他網頁。該語料庫檢索之頁面如圖 2 所示。

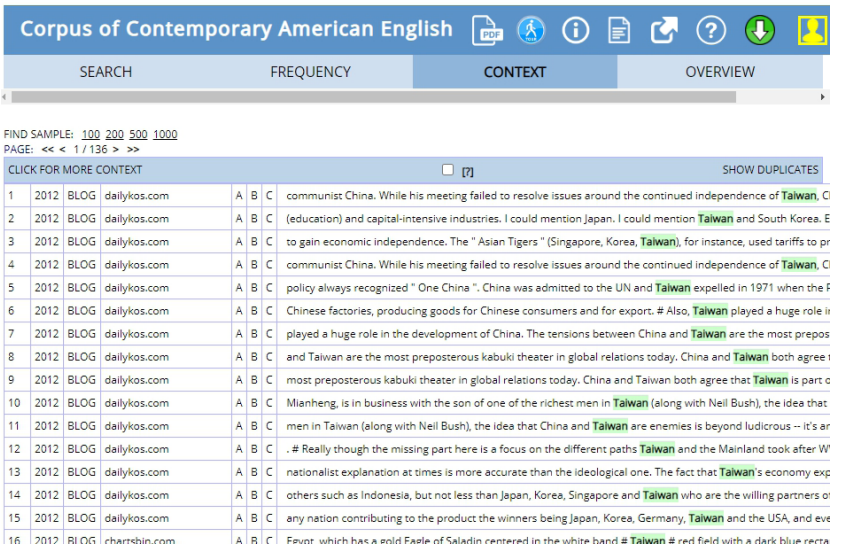
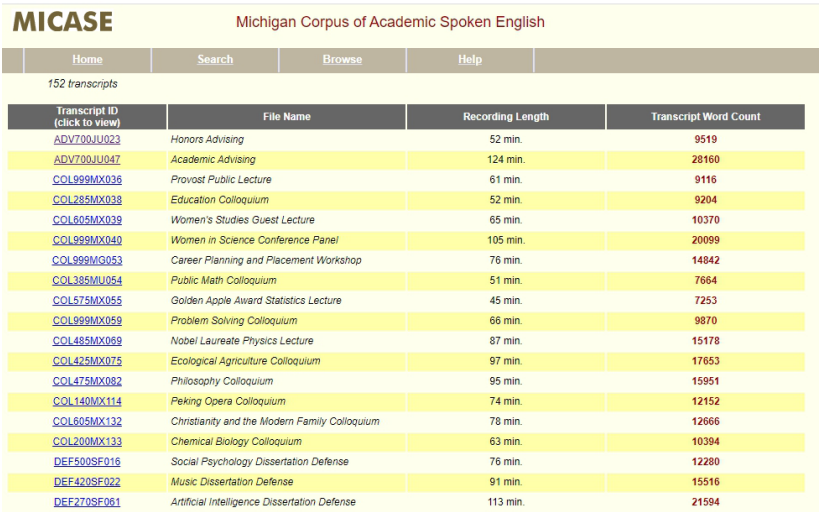


圖 2　COCA 語料檢索顯示頁面

資料來源：作者截圖自 Corpus of Contemporary American English（Davies 2008）。

儘管英國國家語料庫（BNC）及美國當代英語語料庫（COCA）皆同時收錄書面語料及口語語料，此二語料庫的口語語料呈現方式皆為轉寫後的文本材料，而非原始的口語語音素材，網站並無提供相對應之音檔（英國國家語料庫（BNC）之語音錄音檔需額外註冊申請），因此無法忠實呈現美國與英國國內各地口音之變換差異性與多元豐富性。

大多數英語語料庫以書寫語料為主，而也有少數英語語料庫專門收錄口語語料。密西根學術英語口語語料庫（Michigan Corpus of Academic Spoken English（MICASE））即是收集當代學術口語語料的語料庫，由美國密西根大學於 1997 年至 2001 年建構，語料主要為大學校園中各類型的情境與說話者及不同形式的互動（Simpson et al. 2002）。目前網站提供 152 個語料轉寫稿，共計 1,848,364 個單詞，並設計多種篩選條件提供使用者檢索，內容主要為分為兩大類，包括說話者資訊（含學術職位、母語使用者狀態、第一語言）以及文本資訊（含事件類型、學術部門、學術學科、參與者程度、互動等級）。然而，該語料庫目前只提供語料瀏覽及關鍵字搜索，聲音檔之連結已毀損無法使用。[4]



MICASE Michigan Corpus of Academic Spoken English

Home | Search | Browse | Help

152 transcripts

Transcript ID (click to view)	File Name	Recording Length	Transcript Word Count
ADV700JU023	Honors Advising	52 min.	9519
ADV700JU047	Academic Advising	124 min.	28160
COL999MX036	Provost Public Lecture	61 min.	9116
COL285MX038	Education Colloquium	52 min.	9204
COL605MX039	Women's Studies Guest Lecture	65 min.	10370
COL999MX040	Women in Science Conference Panel	105 min.	20099
COL999MG053	Career Planning and Placement Workshop	76 min.	14842
COL385MU054	Public Math Colloquium	51 min.	7664
COL575MX055	Golden Apple Award Statistics Lecture	45 min.	7253
COL999MX059	Problem Solving Colloquium	66 min.	9870
COL485MX069	Nobel Laureate Physics Lecture	87 min.	15178
COL425MX075	Ecological Agriculture Colloquium	97 min.	17653
COL475MX082	Philosophy Colloquium	95 min.	15951
COL140MX114	Peking Opera Colloquium	74 min.	12152
COL605MX132	Christianity and the Modern Family Colloquium	78 min.	12666
COL200MX133	Chemical Biology Colloquium	63 min.	10394
DEF500SF016	Social Psychology Dissertation Defense	76 min.	12280
DEF420SF022	Music Dissertation Defense	91 min.	15516
DEF270SF061	Artificial Intelligence Dissertation Defense	113 min.	21594

圖 3　MICASE 語料檢索顯示頁面

資料來源：作者截圖自 The Michigan Corpus of Academic Spoken English（Simpson et al. 2002）。

4　網際網路可搜尋到「密西根大學學術口語英文語料庫」聲音檔附文字檔之網址連結（https://micase.elicorpora.info/sound-files-online），但該網頁顯示「找不到頁面」。

至於許多正遭受瀕危的少數語言，許多國家也正積極採取搶救措施，將影音文字等語言資源彙整並記錄保存，然受限於語料稀缺且蒐集不易，瀕危語言多半是以檔案庫或資料庫的方式呈現。拉丁美洲原住民族語言資料館（The Archive of the Indigenous Languages of Latin America（AILLA））即是致力於收藏逐漸流失的拉丁美洲原住民族語，此於2000 年由美國德克薩斯大學奧斯汀分校的李拉斯本遜拉丁美洲研究和收藏館（LLILAS Benson Latin American Studies and Collections）共同設立，典藏之資源主要為拉丁美洲原住民族語言的影音檔與聲音檔及文字轉寫檔，另也收藏圖像，如相片、圖畫、地圖等。資料類型包含敘事、聖歌、演講、對話、歌曲，許多錄音檔也被轉錄為英語、西班牙語或葡萄牙語，而語言文檔資料以語法、詞典、民族誌以及田野筆記為主。截至目前為止（2021 年 9 月），拉丁美洲原住民族語言資料館已收藏多達 25 個國家，合計 420 種拉丁美洲原住民語的多媒體材料，[5] 並運用後設資料（metadata）的描述，以數位檔案方式保存語言資源（詳圖 4）。

5　資料來源為拉丁美洲原住民族語言資料館之臉書資訊：https://www.facebook.com/AILLAarchive/posts/2024854964335328。

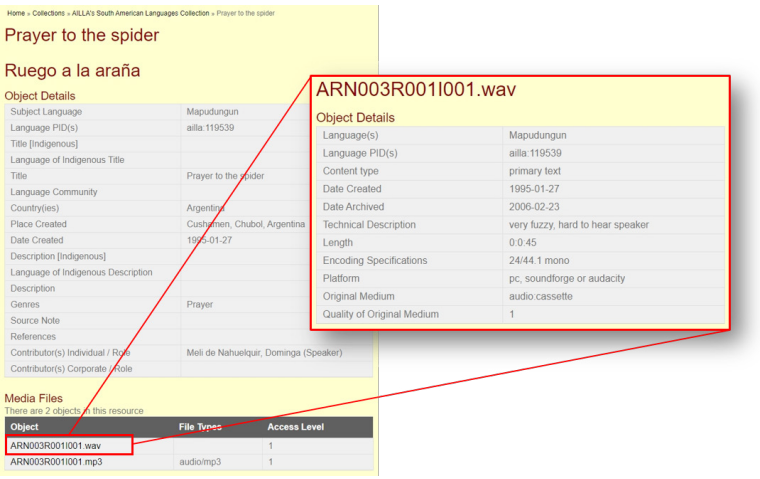


圖 4　AILLA 語言檔案顯示頁面

資料來源：作者截圖與製圖自 Archive of the Indigenous Languages of Latin America（2002, revised 2015, 2017）。

同樣面臨瀕危困境的，為澳洲的原住民族語：達利語。達利語是由澳大利亞北領地之戴利地區中 4 到 5 個澳大利亞原住民語言群體所組成，為保存此瀕危語言，墨爾本大學的副教授 Rachel Nordlinger 與阿德萊德大學的博士 Ian Green 兩人合作建立澳洲達利語檔案館（The Daly Languages (Australia)），並於 1980 年至 1996 年進行達利語實地考察與記錄。此檔案館語料包含 11 種語言共 157 小時的錄音檔，以及豐富的田野筆記、詞典材料與未出版的手稿（其中記錄語法描述、語言分析以及歷史重建資料）。網站使用「語料庫（corpus）」描述所收錄的內容，然而根據網頁語言檔案的呈現方式（如圖 5 所示），此資料館應亦較偏向以數位典藏的形式儲存語言材料。

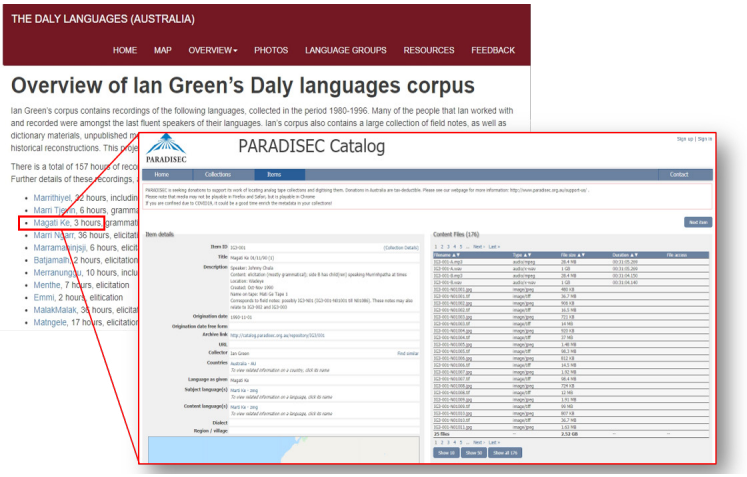


圖 5　The Daly Languages (Australia) 語言檔案顯示頁面

資料來源：作者截圖與製圖自 The Daly Languages (Australia)（Green and Nordlinger 2021）。

由於數位科技突飛猛進，基於語言運用實例進行的語言研究已然成為一種趨勢，以計算機為載體呈載真實語料的語料庫日益受到重視，國內近年來也陸續開啟了在地語料庫建構。藉由國外語料庫與相關語言資料網站的成果，得以對國外語料庫與資料庫現況獲得大致地了解，以下則將基於〈國家語言發展法〉定義之「臺灣各固有族群使用之自然語言」，對於臺灣明定為國家語言之臺灣華語、臺灣閩南語、臺灣客語、臺灣原住民族語等幾項現有較具代表性之語料庫的整理發展進行簡覽與介紹。

（二）臺灣華語語料庫

1. 中央研究院漢語平衡語料庫

中央研究院漢語平衡語料庫（中央研究院 2021）為中央研究院（以下簡稱中研院）從 1990 年起至 2013 年共歷時 24 年所建置第一個帶詞類標記的現代漢語平衡語料庫，所蒐集的語料為 1981 年到 2007 年之間的文章。詞庫小組（1998）指出，為建構漢語平衡語料庫，平衡語料之抽取以中研院詞庫組已收集之現代漢語語料為優先，再同時透過不同管道取得語料以達平衡，來源諸如：（1）經合作計畫交換取得（如中國時報、洪建全基金會、師大國語中心等）；（2）向版權所有單位取得（如天下雜誌社、國語日報社、資訊傳真雜誌社、節目製作單位、中研院內單位、多位教授提供之轉寫口語資料等）；（3）由公共區域取得的公共資料（如 BBS 電子布告欄、蕃薯藤等）。語料收集後便進行語料整理，包含語料清潔、語料分類、加詞類標記等，以足夠的人力條件進行耗時費力的漢語平衡語料庫建構，並以計算語言學常用規模之五百萬詞做為目標建置。現行為 4.0 版本，網路可供檢索的語料量為一千萬，語式（文檔的呈現方式）包含書面語及口語，收錄共計 17,554,089 個字數（character token）與 11,245,330 個詞數（word token）。

而在李佩瑛等（2010）所整理的《語料庫建置入門數位化工作流程指南》中，介紹該語料庫的數位化流程，並依照六項步驟進行，依序為：（1）詞類分析、定義及確定；（2）選擇語料文本來源；（3）程式抓取電子語料；（4）程式自動分合詞及詞類標記；（5）人工詞類檢查；（6）匯入語料庫。其中「程式抓取電子語料」係以人工操作電子語料抓取程式，同步分類文章主題以及匯入系統，並於「人工詞類檢查」時

利用中文斷詞編輯介面由助理以人工方式作詞類檢查（李佩瑛等 2010：35-42）。

2. 國教院語料庫索引典系統（含國教院華語中介語索引典系統）

此索引典系統（國家教育研究院 2021）由國家教育研究院（以下簡稱國教院）依據教育部《推動全球華語文教育八年計畫 (102-109) 》執行之「建置應用語料庫及標準體系」工作計畫，為 2014 年至 2020 年所建置「華語文語料庫及標準體系整合應用系統」之子系統之一，整合共計 10 個語料庫，包含書面語、口語、華語中介語三種類型。根據《「建置應用語料庫及標準體系」108 年工作計畫期末報告》（國家教育研究院 2019），書面語料的文字檔係以臺灣華語之正體字為限，來源為近二十年包含具 ISBN 之書籍語料及新聞資料類型；口語語料為自 2005 年起首播之電視節目，每集長度為 20 ～ 50 分鐘（不含商業廣告），以授權方式取得，語言以臺灣華語為限；至於華語中介語之語料來源則分屬各大華語文中心提供、華測會試場收集、華測會授權共三類。索引典系統累計字數為書面語料約 4 億 3,500 萬字、口語語料約 4,600 萬字（僅文字無聲音）、中介語語料約 156 萬字。

3. 政治大學中文口語語料庫（The NCCU Corpus of Spoken Taiwan Mandarin）

該語料庫[6]（The NCCU (National Chengchi University) Corpus of Spoken Taiwan Mandarin (政治大學中文口語語料庫) 2021）為國立政治大學語言學研究所徐嘉慧教授所建置，自 2006 年開始錄製與蒐集語料迄今，共收錄 49 筆會話型語料，每筆語料大約為 20 分鐘左右，

6　前身為「國立政治大學漢語口語語料庫」，分別收集華語、閩南語及客語語料。詳見 Chui et al.（2017）與 Chui and Lai（2008）。

少數幾筆長達半小時以上。此語料庫提供口語語料之文字轉寫檔，部分語料聲音檔可於多語言語料庫 TalkBank 之 CABank（MacWhinney and Wagner 2010）閱聽（TalkBank 成立於 2002 年，為卡內基梅隆大學 Brian MacWhinney 教授所開發創立）。語料皆依照口語轉寫標記（speech transcription convention）標註（標記方式主要依循 Du Bois et al.（1993）之言談標記轉寫格式），如語輪轉換（turn transition）、重疊（overlaps）及語碼轉換（code-switching）等。每筆語料皆依流水編號規則命名，以 NCCU-TM001-CN-FM 為例，TM 為臺灣華語（Taiwan Mandarin）、001 為順序編號、CN 為會話（conversation）、F 為女性（female）、M 為男性（male），用以表示該筆語料的參與者。

4. 現代漢語口語對話語料庫

此語料庫[7] 為中研院語言學研究所籌備處於 2000 年至 2002 年執行的口語語料庫，蒐集共 30 個對話的錄音資料，共計 25.6 個小時，每個對話平均為 50 分鐘。曾淑娟、劉怡芬（2002）於技術報告中介紹語料蒐集的工作項目主要包含：（1）選取發音人：隨機抽樣選出 16 ～ 25 歲、26 ～ 35 歲、36 ～ 45 歲三大年齡層之臺北市市民，依意願同意而前往參與錄音者取前 60 位（即 30 個對話中皆含兩位參與者），其中含 37 位女性與 23 位男性；（2）錄音過程說明與指示：主要過程為向發音人說明計畫目標與錄音過程，發音人瞭解同意後進行同意書簽署、基本資料及語言使用問卷填寫、過程說明閱讀，最後是正式錄音；（3）錄音設備與格式：錄音地點為普通房間，兩位發音人分別錄於左右聲道；（4）

7 「現代漢語口語對話語料庫」（The Mandarin Conversational Dialogue Corpus（MCDC））無對外開放；另有「中研院漢語對話語音語料庫」（Sinica MCDC8）提供 8 個對話之聲音檔與文字轉記檔，此須付費申請使用，申請網址為：http://www.aclclp.org.tw/use_mat_c.php#mcdc。

錄音資料處理：每個對話轉為數位聲檔，檔名以 mcdc-01 到 mcdc-30 命名，雙聲道音檔則個別存為 ptk 及 wav 的單聲道格式；（5）對話內容整理：將檔名、音檔長度、發音人的性別年齡、對話主題製表建檔。

（三）臺灣閩南語之語料庫

1. 臺灣閩南語口語語料庫

此閩南語口語語料庫（蔡素娟、麥傑 2013）係由國立中正大學語言學研究所麥傑教授與蔡素娟教授自 1999 年至 2014 年共同執行之五個國科會（現為科技部）計畫所建構，將閩南語的廣播節目轉錄成文字。語料來源為雲嘉電臺（《Q 鬆鬆太太俱樂部》、《歡喜看臺灣》）及中廣寶島網（《輕鬆麻辣鍋涼拌小黃瓜》、《四神湯》、《美麗人生》、《中午茶》、《有緣來做伙》、《寶島向前行》、《下班真輕鬆》、《幸福萬事通》），共計 10 個廣播節目，以 MP3 格式儲存，並依《臺灣閩南語辭典》、《臺灣話大辭典》、《廈門方言詞典》、《閩南語詞彙》之順序參考四本辭書之用字，將廣播轉記為閩南語漢字或拼音，同時將段落做斷句工作，每一筆語料會由兩位轉記者確認用字及音標在斷詞上無誤後，以自動檢測程式進行詞頻計算與詞條比對，最後由人工檢測完成確認工作（蔡素娟、麥傑 2013）。目前已公開完成轉記的錄音約 28 個小時、詞頻數約 315,069 詞。

2. 臺灣閩南語兒童語料庫（Taiwanese Child Language Corpus）

臺灣閩南語兒童語料庫（Taiwanese Child Language Corpus 2021）為國立中正大學語言學研究所蔡素娟教授所主持之國科會（現科技部）研究計畫《臺灣話聲調習得的發展之研究》，執行期間（1997 年至 2000 年）

採錄 14 名兒童（男童 9 名、女童 5 名）長期自然語料，共計 430 個錄音檔案，錄音總長為 330 小時。參與計畫的閩南語母語兒童主要來自嘉義縣民雄鄉，錄音年齡為一歲兩個月至五歲三個月，錄製同時還包含其他參與人員（訪問者、兒童的父母親、兒童的祖父母、兒童的兄弟姐妹等）。錄音檔以 MP3 格式儲存，錄音語料經過錄音內容的編輯，文字檔案以標準萬國碼（Unicode）方式編碼。根據蔡素娟（2011）《臺灣閩南語兒童字詞統計分析》之研究結果，「臺灣閩南語兒童語料庫」之總詞頻（包括其他參與錄音者）為 1,741,408 詞，兒童的詞頻為 499,618 詞（占總詞頻 30%）。

（四）臺灣客語之語料庫

1. 國立政治大學客語口語語料庫

政治大學於 2007 年推行「國立政治大學漢語口語語料庫」，下轄三個子語料庫，分別為國語語料庫、客語語料庫、[8] 閩南語語料庫。其中，客語口語語料庫由國立政治大學賴惠玲教授所主持建置，係臺灣首座開放性之臺灣客語口語語料庫。此語料庫記錄臺灣客語的真實使用狀況，每一份語料皆註明參與者的客語腔調、年齡、性別、語料內容主題與內容簡介、談話地點及參與者彼此的關係。語料涵蓋臺灣客語不同次方言，包含北四縣、南四縣、海陸、大埔、饒平等腔調，音檔內容依言談分析標記系統轉錄方式（主要遵照 Du Bois et al.（1993））將真實語

8 「國立政治大學客語口語語料庫」網址為：http://140.119.172.200/，然現已關閉。2017 年 7 月起與 Carnegie Mellon University 的 Brian MacWhinney 教授合作，部分客語口語語料聲音及轉寫之文字公開於 TalkBank 之 CABank（MacWhinney and Wagner 2010）提供全球使用，可於以下之網址取得（the CABank of TalkBank https://ca.talkbank.org/access/TaiwanHakka.html）。

言轉寫為客語文字。文類為不同主題的「對話」和口述故事的「獨白」，主題式對話口語語料是由兩至三位具有客語母語口說能力之發音人不限主題以聊天的方式自由進行談話，攝錄時間為 1 小時，語料庫工作人員則會從中擷取語料內較為自然之 20 分鐘片段；口述故事則以 Mayer（1980）無文字的圖片故事書 *Frog, where are you?* 為引，請受錄者以客語講述故事，長度約 10 分鐘。語料處理工作包含語料建檔、語料加工（文字轉寫及校訂、聲音檔切割、影像隱私處理等）、語料標記等，完成後便上傳至語料庫。語料庫為會員制，提供一般會員語料瀏覽功能與語料檢索功能，除了可瀏覽每一筆語料的整篇文字內容，亦可查詢欲搜尋的關鍵字詞，並可檢視字詞的前後文。語料庫也設有字／詞頻系統（僅開放予語料庫管理者），包含個別詞彙頻率查詢、詞彙排序查詢、個別字元頻率查詢、字元排序查詢。

王勻芊（2016）從建置、典藏與應用的視角，對國立政治大學客語口語語料庫語料建置與典藏的三個主要過程進行全面性檢視：（1）前置作業（如語料收錄、語料發音人選擇、錄影準備、標記規範制定、系統評估規劃等）；（2）數位化作業（語料數位化處理及系統開發建構）；（3）保存作業（語料歸檔保存、異地備援、後續維護）。根據 Yeh（2017：18）之統計資料，語料庫自 2007 年起建置起至 2016 年 10 月共收錄 77 筆口語語料（含對話 31 筆，獨白 46 筆；已公開語料筆數共為 44 筆，含對話 15 筆及獨白 29 筆）。[9] 所有經過一校與二校的公開與未公開語料，累計文字數量共 248,556 字。

9　44 筆已公開之口語語料目前存放 TalkBank 之 CABank（MacWhinney and Wagner 2010）：https://ca.talkbank.org/browser/index.php?url=TaiwanHakka/。

2. 臺灣客家語語料庫

此語料庫[10]為國立中央大學客家語文暨社會科學學系江俊龍副教授所建置，根據其 2010 年與 2013 年的國科會（現為科技部）計畫報告（江俊龍 2010，2013），語料庫建置主要分為兩個階段。第一階段為臺灣客家語語料庫的建構工作（2008 年至 2010 年），語料來源主要為兩項，一是取得現有客語文獻及客語口語語料授權，共計約 220 萬字；二是自產語料，除了由具備客語母語能力之研究生從各新聞網站中蒐集語料並進行華語和客語的轉碼工作，累計共約 21 萬字的新聞稿轉寫（含南四縣 19 萬、海陸 2 萬）之外，也進行田野調查，將錄音語料轉換為文字電子資料，完成約 19 萬字。語料總字數為 2,608,026 字，並依照六種腔調進行分類，詞條也陸續進行字型校對、詞性標記、人工斷詞等工作。第二階段則是重點進行大埔語料的擴充（2010 年至 2013 年），主要為訪查地方耆老並進行客語故事採集，田野採集工作以全程錄音和筆記方式記錄訪談內容，並標記發音人的性別、出生年、語言別等相關基本資料。口述語料收錄完成後則進行逐字稿轉錄，並會經過用字與標音規範、字碼轉換等工作。兩年計畫執行期間共訪查 13 位耆老、採集 48 篇口述故事，並集結為《新編臺中東勢客語故事》第一輯與第二輯出版，其中第一輯已匯入語料庫，其文字內容皆已完成用字與標音檢查、字碼轉換、文句分詞、詞類標記等工作，且網頁擴建錄音資料頁面以提供使用者聆聽實地語音音檔。根據江俊龍（2013），該語料庫已匯入 90 萬字，並陸續完成約 70 萬字之字型校對、人工斷詞、詞性標記等工作，語料分類為六種不同腔調，各自累計字數如下：北四縣累計 70,929 字次、

10 「臺灣客家語語料庫」網址為：http://163.16.82.253/hakkacorpus，然現已連線失效。

南四縣累計 203,877 字次、海陸累計 112,538 字次、大埔累計 88,845 字次、饒平累計 113,115 字次、詔安累計 110,953 字次。

（五）臺灣原住民族語之語料庫

1. 臺大臺灣南島語多媒體語料庫

此多媒體語料庫（國立臺灣大學語言學研究所 2021）原屬於國立臺灣大學資訊電子科技整合研究中心「多媒體整合實驗室」子計畫之一（2001 ～ 2003 年），由臺灣大學語言學研究所黃宣範教授、蘇以文教授及宋麗梅教授共同主持，自 1998 年便以團隊合作方式致力於研究臺灣南島語語言，並於 2005 年建立語料庫雛型。目前語料庫中已建置好的南島語，分別為：噶瑪蘭語 4 筆口述語料（附聲音及影像檔）、賽夏語 22 筆口述語料（附聲音檔）、阿美語 2 筆口述語料（附聲音及影像檔）、鄒語 2 筆語料（附聲音檔）（Sung et al. 2008）。該語料庫透過田野調查方式每週定期採集第一手口語語料，語料內容為發音人的生活對話、傳說故事，或為發音人觀看無對白影片 （Pear stories）或不含文字的圖書（Frog story）後再口述觀看過的影片或圖書內容。語料以數位錄音機及錄影機記錄，並予以轉寫成文字，文字採用由原住民族委員會（以下簡稱原民會）與教育部於 2005 年共同公告之〈原住民族語言書寫系統〉，並參照 Du Bois et al.（1993）的轉寫標記法，進一步將語料以語調單位（intonation unit（IU））切分成句，並記錄口語現象如停頓、重複、修正、音調等；而在語法標記方面，則遵循「萊比錫標記系統」（Leipzig Glossing Rules）之規定。語料查詢系統除了可查詢任一字詞，亦提供詞綴及語法或篇章標記之檢索功能。

2. 蘭嶼達悟語口語資料典藏網

此典藏網由靜宜大學於 2005 ～ 2006 年所執行的達悟語蒐集計畫，計畫主持人為何德華教授、共同主持人為楊孟蒨助理教授，以收錄口語紀錄為主，該網頁所釋出的語料共計 88 筆，每筆皆含聲音檔及轉寫後的文字，然每篇語料各自獨立，尚無法如一般語料庫提供檢索功能。依其網站所載之計畫內容概述，口語紀錄之操作細節係由靜宜大學研究生配合基金會進行採集與收錄，針對所採之語料加以整理編輯，使用文字來描述口語語料並整理達悟語言的字彙庫，且以電腦資料表形式儲存。目前提供關鍵字查詢[11]的文本為《達悟語多媒體語言教材》（四冊）、二十篇語料（網站無指明語料來源）、《聖經》。

以上為臺灣華語、臺灣閩南語、臺灣客語、以及臺灣原住民族語幾項較具代表性的語料庫介紹。綜觀國內外這些語料庫，大多對於各語料採集與整理等數位作業歷程較少著墨，因此，本文彙整正在建置之臺灣客語語料庫的實際操作經驗，詳述書面語料與口語語料之數位化作業流程，以期做為援引，希冀能夠提供其他語料庫建置的參考。

三、臺灣客語語料庫之語料數位化

為將客語語料以系統性且數位化方式進行收錄、處理與保存，臺灣客語語料庫團隊（以下簡稱語料庫團隊）藉由實務操作經驗的累積，建立一系列語料處理作業流程，主要可分為「前置作業」和「數位化及檔案管理」兩大部分。

11 「蘭嶼達悟語口語資料典藏網」之「關鍵字查詢」網址為：http://yamiproject.cs.pu.edu.tw/yami/yami_ch/database.htm（靜宜大學 2021）。

（一）前置作業

1. 語料盤點

前置作業係指語料在文字數位化處理前須先進行的工作，包含語料盤點以及語料徵集與授權。語料盤點著重於出版品篩選及審核，並依類型分為書面語料（如實體書籍、電子書）和口語語料（如電視節目、演講）。語料庫團隊致力於全面性地清點盤查全臺灣客語著作，並予以系統性地分類，以篩選符合臺灣客語語料庫收錄需求的語料。客語書面出版品盤點涵蓋 1990 年代迄今近三十年的政府與民間客語出版品，並同時對客語著作發展歷程進行耙梳，盤點過程中發現早期客家文學作品仍屬華語文字書寫階段，如黃恒秋（2005）介紹之吳濁流《亞細亞的孤兒》、鍾理和《笠山農場》、李喬《寒夜三部曲》等膾炙人口的著作（以及而後相繼透過鍾肇政、羅肇錦、黃恒秋、龔萬灶、朱真一等作家投入並出版的客語文學作品）；真正的客語寫作則是以客語詩歌成濫觴，[12] 如杜潘芳格率先以客語從事現代詩的寫作。客語作品中詩歌類的創作量明顯多於其他文類，一方面簡短的句型較能夠掌握，有利於客語寫作，另一方面詩歌的重複性及朗誦效果有助於形成感染力較強的效果，因此許多作品在客語詩的文學領域逐漸遍地開花，包含葉日松《一張日曆等於一張稿紙》、邱一帆《田螺》與《油桐花樹下介思念》、曾貴海《原鄉・夜合》、張芳慈《天光日》等。客語散文也開始逐漸發展，根據黃恒秋（2005），范文芳於 1998 年所著《頭前溪个故事》為臺灣出版的第一

12　客語作品雖以詩歌類為大宗，然目前臺灣客語語料庫處於建置初期階段，就系統工程的角度而言，為了有效地訓練機器斷詞，目前斷詞階段仍仰賴人工過濾分類詞條並標示斷詞標記，須使用大量包含標點符號的完整語句方可較順利地訓練機器學習，因此目前語料收錄以具完整篇章結構的文章優先，詩歌類的著作須待系統建置穩定之後，再評估其列為未來徵集標的之適切性。

部客語散文集，可將之視為客語散文集的出發，張捷明《客家少年》、龔萬灶《阿啾箭个故鄉》等以客語撰寫之著作也隨之出版。隨著客委會與原民會先後成立，母語寫作的風氣開始受到重視，其中以閩南語寫作起步最早、客語居次、原住民族語殿後。雖然母語寫作已經開始上路，但閱讀族群仍屬於小眾，出版量極為稀少（邱各容 2016）。儘管臺灣客語文本數量遠不及臺灣華語，透過書面語料盤點可揭示臺灣目前客語作品類型之概貌。語料庫現已蒐集為數不少採用客語文字書寫的語料，如經典著作、一般書籍、獲獎之單篇作品、報章雜誌刊登文章、客語認證試題等。

至於客語口語出版品的盤點，則是從現有的影音媒體檔案進行評估與揀選；另因口語語料須仰賴人工逐一聽打，檔案品質的清晰度為優先篩選的門檻。臺灣客語口語出版品最初始的萌芽形式為客語廣播節目，發展至今已有近三十年的歷史，最早是由「寶島新聲客家臺」（現為「財團法人寶島客家廣播電臺」）自 1994 年起開播，之後隨著社會逐步風氣逐步開放、母語意識逐漸抬頭，其他的民間電臺也如雨後春筍般地設立，例如 1997 年創立的「中廣客家頻道」（現為「I go 531」），以及 2002 年於桃竹苗地區開始播送節目的「大漢之音調頻廣播電臺」等。南部地區的「高屏溪廣播電臺」也於 2002 年正式成立，製播內容包含客語及原住民族語的節目。至於客委會最早的廣播節目為 2002 年委外製作的《哈客一族》，並透過國立教育廣播電臺進行全國性播送。然而，傳統廣播媒體性質屬於特定頻段、特定時間的即時播送，現今難以透過公開平臺取得早期節目清單，亦無法有效取得檔案資源，語料內容與檔案音質的審核工作難以推進；且廣播性質主要為聲音傳遞，無發

音者之臉部畫面，工作人員在進行語料轉寫時若遇到模糊不清之處，無法依據發音者之口型進行文字輔助判斷。幾經權衡之下，臺灣客語語料庫以同時具影像與音檔之口語語料納入優先收錄考量，如以客語發音的電視節目。臺灣第一部客語連續劇當屬 2002 年由臺灣公共電視臺與客委會合製的一齣客家文學大戲《寒夜》，然而，演員錄製是以華語發音演出，所有客語發音皆為後期配音添加，如此演員嘴型和客語音軌無法對應，不利於音檔轉客語文字工作。考量後續語料數位化需求，口語語料之收錄聚焦於以真人客語發音演出的客家電視臺影視節目。客家電視臺於 2003 年開播，為臺灣最具客語代表性的電視媒體，早期的節目雖多以錄影帶錄製，但隨著數位化技術逐漸普及，多數錄影帶已轉檔為數位電子檔，並可於「好客 ING －客家影音網路平臺」瀏覽，國際性影音平臺 YouTube 亦可查詢到近年製播的節目，資源蒐整管道更為便捷，有利於盤點工作進行；且電視節目多提供華語字幕做輔助，對於客語轉寫工作人員亦為一大助益。而參與客家電視臺演出之客語發音者經由電視臺選角篩選，其母語能力須達到一定流利程度，客語語言能力具有標準可信度。目前語料庫取得授權且已收錄的口語出版品，包含客家電視臺產製的戲劇節目以及人物報導節目，至於新聞等其他類型的節目也納入未來持續蒐整標的。

通過篩選的語料則進入初步審核。書面出版品的初審重點在於文本是否為客語文字書寫或發音，以及是否符合客語用法標準，多半可藉由著作的出版資訊、名稱、目錄等內容來檢視，有些著作書名即帶有明確客語特徵詞彙，部分書籍也提供線上簡介或電子目錄做為參考，然而語料是否使用客語書寫仍須進行全文內容審查方可確認，且需倚賴客語母

語工作人員進行語料判別與審核。客語的書面文本在眾多不同的文學類型上仍屬發展中階段，且根據國家圖書館（2020）《108 年臺灣圖書出版現況與趨勢報告》可以得知，申請 ISBN 圖書出版使用語文統計中，相較於以正體中文出版作品約占全部新書總數之 93.68%，以客語出版之作品屬於「其他」語文別之中的一個小次分類（「其他」類僅占整體之 2.94%，且內容主要為東南亞語文、雙語、多語對照等，故可推估以客語書寫之作品比例應更低），此外，許多作品列屬或標示為客語書籍，但經過實際檢核後，發現內容參雜大量華語或甚至不符合客語語法，於盤點過程中在在影響客語語料篩選，因此語料必須符合審核標準，方可收錄於語料庫。至於口語出版品的初步查核，則特重客語母語口說能力以及音檔語音品質。語料庫每筆口語語料均設計附帶音檔，為利轉寫工作人員進行文字繕打與文本數位化，口語發音者客語能力須達標準程度，口語音質必須清晰可辨，盡量避免環境噪音或其他聲音源如震動或反射音等干擾因素而致使語音識別率降低。因此，客語口語出版品盤點與審核優先以使用客語發音的電視節目或新聞為主要標的，因影視出版品事先已經由電視臺執行把關，品質及內容相對穩定。

語料盤點作業流程可對客語文本內容或語料發音者客語能力進行評估與篩查，初步審核語料或語料發音者是否符合收集採錄標準，審核通過後再進行下一步聯繫洽談，因此是一個相當重要且不可或缺的作業環節。若書面文本或口語言談之中出現大量非客語語流文字或話語，對於語料庫後續斷詞流程之系統自動斷詞與機器學習而言將會遭遇到許多困境；因此，從作業流程的開端即進行把關，對於語料庫的基礎工程建置具有相當大的助益。

2. 語料徵集與授權

語料盤點完成即進入語料徵集與授權階段，主要為聯繫出版品之著作財產歸屬人並洽談語料授權事宜。為使書面與口語語料得以正當且合法地收錄、使用、加工或重製（如標註詞性標記），同時保障著作人之權利，語料在數位化工作前必須遵循相關法律規範。授權狀況相當繁雜，例如有些公部門或民間團體作品之授權不清（常見於比賽得獎作品合集等）；有些典藏單位或機構對典藏作品僅享有物權而無著作權；有些典藏作品不只有著作財產權，還有著作人格權問題，甚至有部分出版品更進一步衍生出其視聽及錄音作品之相關著作權。因此，臺灣客語語料庫以〈著作權法〉為依據，結合實務使用之範圍，在諮詢專業法律顧問並制訂授權同意書內容後執行授權洽談。由於著作品版權不必然歸屬於作者，而著作權常見的法律糾紛中，又莫過於作者與出版社之間的著作權歸屬約定不明確，因此授權洽談作業中，釐清「著作財產權歸屬人」為首要工作。口語語料採錄影音之授權人即為採錄對象（發音者）本人，因此僅須向個人取得授權同意即可。出版品的著作財產權歸屬人之認定，則須藉由出版品所登錄的「出版單位」與「作者」資訊來確認。不論是出版單位或是作者，都再依「公部門」、「民間團體」、「個人」三種身份進行區分，因此出版品之著作財產權歸屬狀況可彙整為以下五種組合類型：

（1）出版單位及作者皆為「公部門」

近年來，政府出版品多半以機構或單位名義出版，例如教育部於2008 年出版的《臺灣客家語朗讀文章選輯》作者即為「教育部國語推行委員會」。由於出版品的著作財產權為政府所有，故部會之間以行文

方式無償取得授權即可。

（2）出版單位為「公部門」、作者為「個人」

許多出版品的出版單位為公部門，作者則為個人，這類屬政府出版品的著作財產權歸屬，悉依政府單位當初與作者簽訂授權書的條款內容而定，因此須先向出版方（政府單位）確認授權細節，並須確認著作財產權之歸屬，再視情況向個人作者進一步聯繫。著作財產權歸屬狀況可分為以下三種：

（i）著作財產權歸「公部門」所有

此類型多半為公部門早期辦理徵件活動、或是彙編出版品時，與作者之間簽署「著作財產權讓與同意書」，由作者「專屬授權」予政府單位，如《107 年苗栗縣文學集 — 兒童文學創作》之系列書籍。然而，由於此類作品為專屬授權，雖政府單位取得全部的著作財產權，但無法逕予授權第三方使用，因此，此類型的授權處理，除了以行文方式取得公部門無償授權之外，亦須向作者取得授權同意書。值得注意的是，依照政府單位與作者之間的「著作財產權讓與同意書」，著作財產權屬於政府單位所有，故不須另行支付授權金予個人作者。

（ii）著作財產權由「公部門」與「個人」共有

若出版品為公部門與作者共同擁有著作財產權，則需同時取得兩方的授權同意書，並依比例分配支付個人作者授權金（公部門單位為無償授權）。

（iii）著作財產權歸「個人」所有

有極少數政府出版品（多為早期作品）之著作財產權全數歸作者擁有。以「臺南縣立文化中心」（現整合為「臺南市政府文化局」）於

1996 年出版的《向日葵》為例，經向臺南市政府文化局確認後，著作財產權歸屬個人作者，因此逕行向個人作者取得授權並支付授權金即可，免向公部門取得授權。

（3）出版單位及作者皆為「民間團體」

民間團體包含出版社、學協會等，其部分出版品係以機構名義發行。以屏東縣六堆文化研究學會於 2004 年出版的《六堆人揣令子》為例，作者即為學會的編輯小組，因此向該單位申請授權並支付授權金。

（4）出版單位為「民間團體」、作者為「個人」

（i）著作財產權歸「民間團體」所有

有些作者將著作財產權讓與給出版單位或學協會。以《安徒生童話全集 國家語言（臺灣客語 - 海陸腔）》為例，作者（此案例為譯者）為謝杰雄等人，著作財產權歸屬發行單位「中華臺灣客家國際藝文交流協會」，故僅須向單位取得授權並支付授權金即可。

（ii）著作財產權由「民間團體」與「個人」共有

若「民間團體」與「個人」共享著作財產權，兩方均須簽署授權同意書，授權金則按比例（通常為均分）計算。

（iii）著作財產權歸「個人」所有

此類型多屬報刊雜誌所刊登的單篇文章，經向出版單位確認後，逕向作者取得授權同意即可（著作權歸屬於「個人」），例如國立中央大學客家學院電子報所刊載之著作。此外，由於早期授權觀念尚未普及，有些出版單位（民間團體）和個人作者並無在出版當時簽署任何契約文件或授權書。以語料庫接洽的一案為例，在此狀況下，儘管出版單位表示某作品著作權為其與個人作者雙方共同擁有，但作者當時並未明示

共享或讓與著作財產權予出版單位，亦無文書契約為證，因此作者主張擁有全部的著作財產權。後續透過語料庫團隊居中溝通並多次向雙方確認，最後出版單位及作者雙方達成共識，同意著作財產權歸屬為「個人」所有，授權金支付予作者。此實證案例可更加佐證著作財產權制訂與授權書簽署之重要性。

（5）出版單位及作者皆為「個人」

市面上有些作品是由個人作者自行出版，因此財產權歸屬人即為作者本人，例如龔萬灶於 2004 年出版的《阿啾箭个故鄉》。若遇多位作者共同出版之情形，須取得所有作者之同意，授權金則按比例支付。

以上所列舉之個案係以書面出版品為例，口語語料的影視出版品處理方式亦然，著作財產權的歸屬人釐清與認定可彙整為五種類型，簡列如下：

（1）出版單位及作者皆為「公部門」

（2）出版單位為「公部門」、作者為「個人」

（3）出版單位及作者皆為「民間團體」

（4）出版單位為「民間團體」、作者為「個人」

（5）出版單位及作者皆為「個人」

語料在取得著作權人的授權同意之後，即進行檔案分類及管理。已取得授權之著作，臺灣客語語料庫係遵循「創用 CC：姓名標示 - 非商業性 - 相同方式分享 臺灣 2.5 版」（Creative Commons liscense: Attribution-NonCommercial-Sharealike 2.5 Taiwan）之規範，在臺灣客語語料庫正式上線運營後，使用者可進行合理使用，惟須標註來源且限於非商業用途。若使用者修改語料庫素材，亦須依照同樣的方式進行分享，讓臺

灣客語語料庫的內涵與價值，得以獲得永續性的傳遞與推廣。

除了書面與口語出版品，語料庫語料的來源亦囊括來自臺灣客語六個腔調發音者的口語採錄音檔。採錄對象主要為具有客家代表性或特殊專長且具備流利客語口說能力之客語專家人士，經語料庫彙整其經歷、專長、志趣等背景資訊，並同步盤點評估語料庫缺稀主題後，列為特邀對象，正式提出邀訪並經對方同意授權後，隨即啟動口語採錄作業，執行項目包含主題與題目設定、採錄日程聯繫與安排、錄製場地商借、攝錄器材（如：動態攝影機、指向性麥克風等）與授權文件準備，以及實際拍攝等一系列語料採錄流程，語料形式多以獨白式敘事與對話式日常生活會話為主。採錄作業完畢後，核查影片畫面與影音品質，確認符合收錄標準後請採錄對象（發音者）填寫相關個人資訊，包含腔調、年齡、性別等後設資料，不僅做為建檔之依據，更為語料庫量化的基礎。

影音採錄品質的控管，可讓語料轉寫者的錄音檔聽打及文字轉寫的過程更加順利，因此於語料錄製的過程中需注意「安靜」、「明亮」、「平穩」、「清晰」的錄影四大要素。前兩項強調環境因素的控管，必須儘量避免外在音訊的干擾（如車聲、電話鈴聲、鬧鈴、狗吠聲等）和語料發音者以外不相干人等的人聲（如路人、郵差及快遞送貨的說話聲），影音錄製盡可能選在安靜的室內取景；此外也需注意場地明亮度，利用自然光或室內燈光進行影片拍攝。而後兩項則是遵循取景的基本概念，利用腳架固定攝影機以避免畫面晃動模糊，採用平視、近景角度拍攝，完整呈現語料發音者的臉部與嘴型。語料影音檔品質穩定與影音清晰，有助於語料內容辨識且利於語料標記選用；轉寫後的口語語料文字搭配

六腔聲音檔，更有益於未來在其他研究或教學用途上的應用。[13]

綜上所述，語料前置處理作業流程藉由語料盤點來審核語料是否符合收錄標準，而語料徵集與授權之繁瑣嚴謹則有利後續進行語料數位化作業。對於整合各時期客家創作與各式客語語料，可謂是盤點整理之功效。

（二）數位化及檔案管理

1. 語料建檔與後設資料標註

語料於徵集完成後，須先進行基本後設資訊建檔並予以編碼管理。後設資料標註不僅有利於資料管理與查找，更是作為語料入庫時，供系統抓取後設欄位資訊的重要依據。

書面語料之建檔，係於語料庫後臺選取書面語式後，即依序輸入書面文本相關後設資訊，包括語料名稱、單元名稱（如章節、篇名、單元等）、作者、出版者、出版年份、文本內容（即語料文字）。而後設資訊如文類、主題、載體等，則是以下拉式選單方式選取。書面語料後設資料格式，請見圖 6；[14] 口語語料的後設資訊大致上也與書面語料相同（如圖 7）。[15]

13 臺灣客語語料庫之口語語料提供轉寫文字以及聲音檔，影片檔僅作為內部語料處理之用，提供轉寫人員參照影片發音者臉部表情與發音嘴型，以做為文字轉寫或校訂之參考依據。

14 後設資料欄位中，標示「*」者表示「必填」；空白欄位處代表無資訊（例如某些書籍僅有單一版本，故無版本資訊）。此外，文本斷詞後結果係透過語料庫斷詞系統生成後，再由工作人員貼至「文本內容」下方之「文本斷詞標記」欄位。語料庫斷詞系統因正在擴建更新中，且受限於篇幅，未來將另文專論。

15 相較於書面語料多有再版、修訂版之版本差異，口語語料以聲音為主，較無一語料多版本狀況，因此口語語料後設無設置「版本」欄位。

圖 6　書面語料後設資料格式擷取畫面

資料來源：作者製圖。

圖 7　口語語料後設資料格式擷取畫面

資料來源：作者製圖。

口語語料另會再登錄採錄對象（發音者）的基本資料（呈現如圖8），包含姓名、腔調、代號、性別、年齡，以及發音者於錄影畫面上之相對位置。[16] 口語語料音檔的相關資訊，則填入「背景資訊」分頁，裡面內容包含影片錄製日期、錄製時間、語料長度、影片錄製地區、場景，以及備註（請見圖9）。

圖 8　口語語料後設資料格式擷取畫面：發音者資訊

資料來源：作者製圖。

圖 9　口語語料後設資料格式擷取畫面：背景資訊

資料來源：作者製圖。

16 發音者資訊欄的「畫面上位置」係僅供轉校人員搭配影片檔轉寫之參考依據。

2. 語料數位化與資料清理

語料建檔完成後，即進行文字數位化工作。而語料數位化的首要階段，即屬語料轉檔（或轉寫）與語料校訂。書面語料多為轉檔作業，文本取得形式可分為「Word 檔／ txt 檔」、「PDF 可編輯檔」、「PDF 不可編輯檔」及「實體書籍」四類，為使書面文本轉成可機讀的語料，檔案皆統一為 Word 格式。圖片掃描檔或限制編輯檔案則運用文字識別軟體（Optical Character Recognition（OCR））來進行轉檔輔助。轉檔過程中經常出現電子轉檔或機讀錯誤、客語特殊字判讀錯誤等各式狀況，因此高度仰賴人工判讀檢核並進行文字查漏補缺，亦屬資料清理的一環。口語語料則多為影音檔素材，因此由工作人員將口語語音進行文字化轉寫工作。臺灣客語用字相較於臺灣華語發展更晚，而坊間各家客語書寫的使用紛雜不一、用字紊亂，不僅詞彙涵蓋量不足，臺灣客語次方言各家選用文字也紛雜迥異；另一個客語文字化遭遇到的問題則是部分詞彙有音無字，因此許多作者使用客語其他詞彙予以借代，或是使用漢字取代，亦或自行創字甚或直接以拼音表記，這些用字差異等問題皆會影響詞彙詞頻、詞彙共現詞等語料庫功能計算。因此，為求客語用字盡量達到一致性，語料庫用字規範主要遵照教育部發佈之第 1 批與第 2 批「臺灣客家語書寫推薦用字」以及教育部「臺灣客家語常用詞辭典」，用字收錄不足之部分則輔以客委會「客語認證詞彙資料庫」以做參考。而對於教育部及客委會均無規範或收錄之用字，例如著書內容出

現「𫰛[17]」、「癆[18]」、「辽[19]」等罕用生僻字，語料庫暫採「忠於原著」方式保留作者原用字。部分客語罕用字還會遭遇到一個問題，即是在Windows等系統中無法正確顯示，變成方塊「□」（一般慣稱為「tofu」，意即「豆腐」）。一般字型如微軟系統內建之新細明體與標楷體等系統字型，可支援Unicode 3.0版所引進的「擴充A區漢字」（共27,496字），然而仍不敷使用，無法涵蓋所有客語用字，以致於有些客語用字仍無法正常顯示，儘管是屬於「臺灣客家語書寫推薦用字」如「偌[20]」、「偬[21]」等，也面臨相同狀況；再者，有些「字」本身具有多個以上的Unicode碼，以「𠊎」為例，此字屬擴展E區，其Unicode編碼為2B8C6，而過去於私人使用之編碼有E000、E700、F307、FA40等。這些格式的「𠊎」字形非常相近，難以用肉眼精準辨識，因此若不同使用者選用到不同編碼的「𠊎」，即可能造成不同的搜尋結果有所差異甚至是搜尋失敗。為解決客語難字輸入以及Unicode碼不一等問題，本語料庫支援客語字Unicode字集擴充，可對新舊字碼進行轉換，另也將Unicode擴展A、B、C、D、E區的字型進行合併轉檔，解決客語一字多碼（Unicode）以及異體字的問題。此外，語料庫系統使用雲端字型（Web Font）嵌入技術，語料庫頁面字型可自動從雲端獲取字型資料，使用者可免任何事前字型安裝工作，網站也依照國家發展委員會訂頒之無障礙網頁開發規範以及

17 華語義為「碎裂」。字型資訊為Unicode: U+203B7，屬「中日韓統一表意文字擴展區B」。
18 華語義為「疲累」。字型資訊為Unicode: U+24E01，屬「中日韓統一表意文字擴展區B」。
19 華語義為「繞行」。字型資訊為Unicode: U+2B7E7，屬「中日韓統一表意文字擴展區D」。
20 華語義為「我」。字型資訊為Unicode: U+2028E，屬「中日韓統一表意文字擴展區B」。
21 華語義為「我們」。字型資訊為Unicode: U+2B8C6，屬「中日韓統一表意文字擴展區E」。

符合響應式網頁設計規範，偵測使用者的螢幕大小並進而自動調整網頁圖文內容，提供使用者能以跨平臺形式在不同系統或行動裝置間皆可達到一致的顯示效果。

每一筆語料都會經過轉寫與校訂階段，遵循交叉校訂用字及格式，並遵照用字規範及轉寫原則，以確保客語用字、客語拼音以及轉寫標記都完全正確，甫能再進一步處理數位化文本斷詞及詞性標記等工作。[22] 轉寫與校訂作業由母語本腔人士進行專家校訂，而客語六個腔調（四縣、海陸、大埔、饒平、詔安、南四縣）各自具有其獨特性，最能體現六腔差異的，即是調值與詞彙。在調值方面，四縣腔與海陸腔各聲調的調值幾乎相反，以「文字」為例，四縣音為 vun11 sii55，海陸音則為 vun55 sii33。與四縣腔較相近的是南四縣腔與大埔腔（南四縣音為 vun11 sii55，大埔音為 vun113 sii53），饒平音則為 vun55 sii24，與海陸較相似；至於詔安腔則與其他腔調差異最大（拼音為 bbun53 cu55）。另依據教育部 2012 年〈客家語拼音方案使用手冊〉（教育部 2012），除了海陸具七個聲調外，其他五腔皆為六個聲調（陰平、陽平、上聲、去聲、陰入、陽入），海陸的去聲則是再區分為「陰去」、「陽去」二聲調。此外，各腔調間也存在明顯的詞彙差異，特別是少數腔的詞彙。舉例而言，大埔腔詞彙通常不帶詞綴「仔」，如華語「木板」在四縣為「枋仔」，大埔則為「枋」。饒平腔的詞彙相對於其他客家腔調也具有特殊性，以華語「起身」為例，在四縣、海陸、大埔、南四縣的客語詞彙為「趴」，饒平則為「跆」（詔安則為「跋」）。而使用比例最少的詔安腔，其語音系統受到臺灣閩南語的影響，導致許多詞彙音韻與閩南

22 本文主要聚焦於語料數位化作業流程，考量篇幅所限，關於用字規範、語料階層屬性、轉寫標記與原則、斷詞與詞性標記等細部內容，將另以他文闡述。

語相似（例如華語「風箏」在客語其他五腔的用法為「紙鷂（仔）」，詔安則是「風吹」），在這種語言接觸的情況下，致使臺灣詔安客語與臺灣閩南語的區辨更添判斷困難性（特別是口語語料），因此轉寫與校訂人員皆須經過篩選並進行教育訓練，主要為客語薪傳師、客語教材編輯委員、客語認證命題或閱卷等典試人員等，其中具有字音字形比賽得獎或客語各項比賽指導經驗者不乏少數。語料中也多有客語和其他語言之間的混用情況，例如客語語料中出現華語、閩南語、原住民族語、英語、日語等。因此，本語料庫也採用轉寫標記「<CS- 語言代碼 > 語料 </CS- 語言代碼 >」以示語碼轉換中其他語言的角色，語言代碼係依循國際標準化組織為全世界各語言所制訂的標準代碼 ISO639（International Organization for Standardization），每種語言以兩個字母（639-1）或三個字母（639-2 和 639-3）的小寫代碼來標示。以客語夾用日語的文本為例，客語語料原文「逐日朝晨，毋係駛車就係騎オ－トバイ送佢去學校。」（華譯：每天早上，不是開車就是騎オ－トバイ（摩托車）送他去學校。）即會在「オ－トバイ」前後加註語碼轉換標記，標記後文本為「逐日朝晨，毋係駛車就係騎 <CS-ja> オ－トバイ </CS-ja > 送佢去學校。」；或是客語夾用英語之語料，加上標記後呈現如下：「為了支持在 <CS-en>Facebook</CS-en> 發起个「世界無車日」活動，在厥 <CS-en>Blog</CS-en> 裡肚有一張佢自家揹等背包、騎等自行車个相片。」（華譯：為了支持在 Facebook 發起的「世界無車日」活動，在他 Blog 裡面有一張他自己揹著背包、騎著自行車的相片。）。

此外，相較於書面語料多使用正式用語，篇章層次較複雜縝密，句型結構較完整連貫，口語語料則多為非正式用語，話語層次較鬆散或缺

乏條理，句型結構則較簡單，常有破碎不完整的文句出現。因此，口語語料轉寫除了將影音檔素材內容轉譯並逐一繕打為客語文字外，對於一些言談表現如停頓、發音不清無法辨識的音節等也加註轉寫標記，以呈現真實且自然的口語語流樣貌。而每筆口語語料除了文字檔亦會附聲音檔，若該筆語料屬於多人對談，則會加以發音者代碼以及時間戳記。特屬口語語料的轉寫標記整理如表 1：

表 1　臺灣客語之專屬口語語料轉寫標記及其範例

標記意義	標記格式
發音者代碼	男、女
時間戳記	00:00:00-00:00:00 （時 - 分 - 秒）
停頓	…
無法辨識的音節	<IS>X</IS>
個人隱私	<PP>○</PP>

資料來源：作者製表。

發音者代碼依照發音人之性別標記，同性別兩人以上，則會依照說話順序在標記後方加註數字，如「女 1」、「女 2」。時間戳記標註每個語輪於媒體檔案之起始時間與結束時間，於括號中以連字號作為連結，連字號兩端之時間格式為「hh:mm:ss」。發音者言談之停頓處，標以「…」符號。遇到語料不清楚，如發音模糊以致無法辨識，則以「<IS>X</IS>」標示。而言談中若涉及個人隱私部分，如個人住家地址、電話號碼或身份證字號等，即以 <PP>○</PP> 標示。

轉檔（或轉寫）完畢的語料由語料管理員檢核後，交由第二位轉寫人員進行文字與格式校訂（視語料狀況安排二次（以上）的交叉校訂）。所有轉校工作完成後，語料即匯入語料庫後臺，進行後續的儲存與管

理。

3. 語料儲存與管理

「語料儲存與管理」涵蓋所有與語料相關的檔案管理，為不可或缺的環節。書面語料在徵集與授權時便會取得著作的原始電子檔案或實體書，而實體書亦由團隊逐一掃描成為 PDF 檔，以便同時列為數位檔案進行保存與管理。至於口語語料則是由具有影音剪輯技術的語料庫工作人員使用影音軟體（如 EDIUS 或 Premiere 等）處理，工作項目包含影像與聲音檔剪輯，以及加上流水編號與標誌（浮水印）。經過處理的影像檔僅供轉校人員標記時間碼以及作為文字轉寫之輔助，基於隱私權及肖像權保護原則，不對外公開；而聲音檔則會於語料庫頁面上提供。語料的聲音檔必須符合網路串流格式，也須考慮檔案大小不能超過語料庫後端的負荷，因此將剪輯後的聲音檔轉檔為 mp3，並另依計畫標規要求轉存為 wma 及 wav 檔備存。所有檔案（包含剪輯完畢、轉檔完畢之影音檔案以及口語語料原始影音檔）皆進行異地備援，包含網路儲存裝置（Network Attached Storage（NAS））以及隨身硬碟等多項儲存裝置，以達到一定程度之備份妥善率。

資料清理完畢並完成文字轉寫校訂的語料，即接續進行語料斷詞，[23] 語料會經由系統進行適當的斷詞處理，完成後即成為可檢索的資源。語

23　語料斷詞主要仰賴語料庫最底層的資料模型，其中包含字詞庫之建立。目前字庫提供部首、單字總筆畫數以及字頻統計，詞庫則提供斷詞標記、字元數以及詞頻統計，其中在詞頻方面，由於客語在兩個語式（書面與口語）以及六個腔調之間有許多詞彙上的差異，因此詞頻統計再依照語式以及腔調分開計算。此外，本語料庫於五年計畫期間，字詞庫的階段性任務以詞目及其斷詞標記建構為首要，系統先行開發六腔之拼音、釋義及音檔匯入功能。語料庫字詞庫係以「教育部臺灣客家語常用詞辭典」之詞彙為基礎，因此目前具拼音與華語釋義資訊之客語詞目皆屬「臺灣客家語常用詞辭典」之詞彙，其他詞彙之拼音未來將有待客語學者專家考證討論與統整制訂。未來也規劃可增加華語對譯詞等功能。

料庫檢索功能具兩種檢索方式：一般檢索和進階檢索，目前的一般檢索提供關鍵詞以及搭配詞之查詢；進階檢索功能則是基於資料模型提供之屬性分類為基礎而發展之多重檢索功能，語料庫的大量文本搜尋結果可再依使用者需求逐步縮小聚焦搜尋範圍，例如可選定特定腔調與屬性標籤（如文類、主題、載體），提高檢索精確率，此項設定皆適用於關鍵字檢索與搭配詞檢索。口語語料提供完整音訊，口語檢索介面整合語料時間戳記與媒體音訊，並支援跳轉至音訊相對應之文字段落。語料庫也提供斷詞結果，使用者可根據其關鍵詞檢索結果選取語料並點選「顯示斷詞標記」，系統即會進行自動斷詞將語料文句依詞斷開並標記詞類。[24]

以上為客語語料數位化流程的介紹。綜上所述，語料作業主要分為「前置作業」與「數位化及檔案管理」，前置作業可細分為「語料盤點」和「語料徵集與授權」兩大步驟（對應到口語採錄音檔則為「語料採錄評估」和「語料採錄與授權」）；數位化及檔案管理則包含「語料建檔與後設資料標註」、「語料數位化與資料清理」和「語料儲存與管理」三個部分。每個作業環節均採取滾動式修正，持續各項環節優化精進，臺灣客語語料作業處理流程的整體脈絡詳見圖 10。

24 臺灣客語語料庫之系統功能與檢索介面仍處於滾動式修正與擴充優化階段，完整介面與功能以 2022 年之上線版為主。

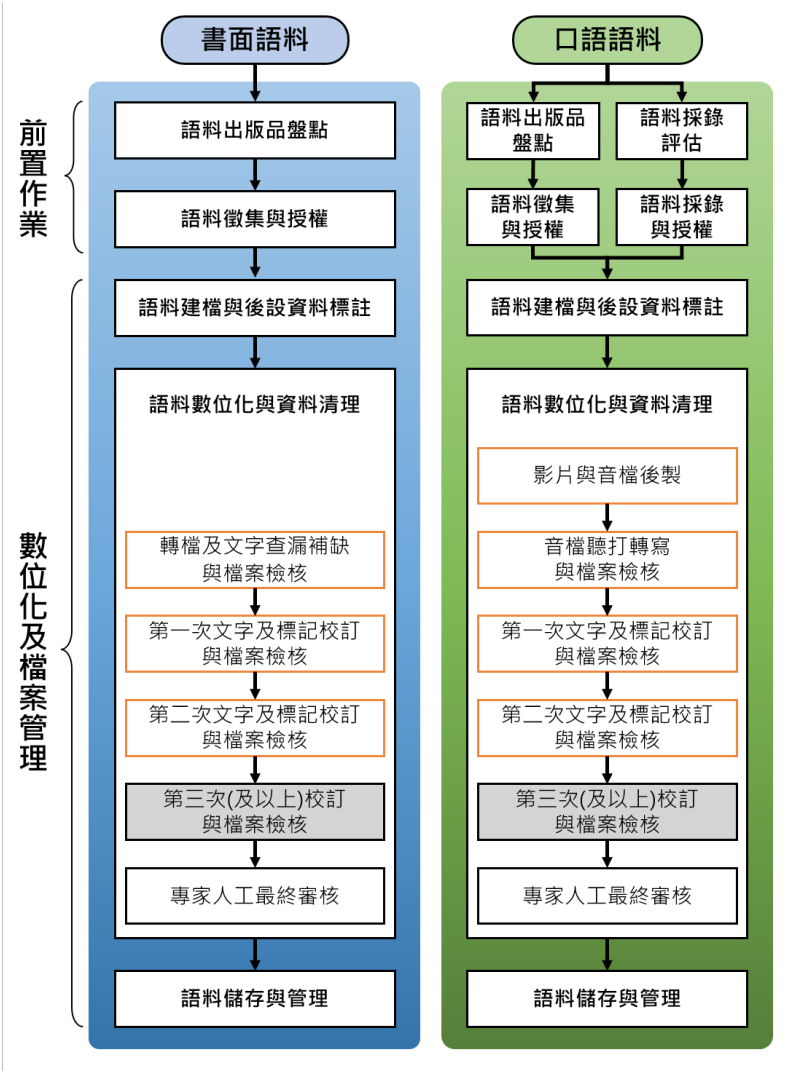


圖 10　臺灣客語語料作業處理流程

資料來源：作者製圖。

四、結論及意涵

臺灣客語語料庫預計為臺灣首座同時收錄客語書面語料及口語語料（附錄音檔）之帶標記語料庫，以記錄語言事實為目的，有規模地保存臺灣客語六腔文字與語音。藉由本語料庫的建置經驗，本文旨在介紹臺灣客語語料的數位化流程並闡述其重要意涵。語料流程的前置作業包含「語料盤點」和「語料徵集與授權」，其中最複雜的即為各種出版品之著作權歸屬。語料授權來源除了公部門之外，亦包含民間團體及個人作者，甚至有授權歸屬者同時為「公部門及個人」、「民間團體及個人」之共有情形。尤其，我國智慧財產權之發展，在過去三十年間有極大的變化，儘管近期社會對著作財產權的觀念提升，逐漸發展出法治結構完備之授權體系，然早期較無著作權概念，許多出版品存有著作權爭議。為此，臺灣客語語料庫針對全臺灣客語著作進行全面性盤點，並且建立語料徵集流程，對授權程序進行相當程度的耙梳，釐清公部門與民間團體、個人之間的權利義務關係，針對不同狀況制訂相對應的處理原則。而在「數位化及檔案管理」範疇，則是涵蓋了「語料建檔與後設資料標註」、「語料數位化與資料清理」和「語料儲存與管理」三個階段。書面及口語語料均以數位化方式處理，作業包含語料後設標註與客語文字轉寫及校訂，語料文字主要遵循教育部用字規範，文本中出現的語言混用等現象加註轉寫標記，並根據文句結構標註詞性與斷詞標記。這些文字校訂與標記可謂語料庫之精髓，除了展示臺灣客語語言之真實樣貌，用字校訂可讓語料庫字詞得以規範化，避免用字紊亂不一等狀況；詞性

標記標示出每個字詞於句中扮演的詞性及其句法功能；轉寫標記主要功能之一則是標註非客語用詞，便於客語斷詞系統偵測辨識，將轉寫標記及其包圍起來的非客語文字內容進行忽略處理，確保非客語文字不會被誤篩選入客語詞庫中。所有書面及口語語料在經過多次交叉審核校訂、詞性標記與斷詞修訂後，匯入語料庫後臺予以線上化管理，同時進行檔案所有版本之異地備援。

國內外大部分的口語語料庫多以提供文字轉寫檔為主，臺灣客語語料庫每一筆口語語料除了轉寫為文字外，亦皆提供相對應之音檔。目前口語語料收集之文類包括會話、敘事、演講以及戲劇，語料參與者（發音者）若為兩位以上，口語文字以語輪為單位進行轉寫，並標註口語參照時間點戳記，提供使用者逐句聆聽與文字聲音對照。「聽、說、讀、寫」為人類學習語言的四個階段，初始階段為「聽」，接下來為「說」，後面才進階為「讀」與「寫」，可見於語言學習歷程中，口語語料可為熟悉一種語言提供最迅速的管道，保留音檔有其必要性與迫切性，而客語語料庫的六腔分類功能，更可將不同腔調之語音流變透過音檔完整保留與記錄。藉由真實口語語料，不僅可窺見各年齡層的客語使用狀況，也可記錄耆老的口傳記憶與親身見聞，以及各行各業的經歷等具有歷史意義的口述語料，豐富客家文化發展歷程，並呈現出臺灣客語在不同時期、不同地區的多樣面貌。

臺灣客語語料庫正在逐步建構中，建置完成之後將具備資訊檢索分析功能，並客觀地呈現語言真實的使用情境，未來可提供教學應用及研究豐富的素材，並達到多領域之加值應用願景。在語言學方面，舉凡語音、構詞、語法、語意、言談等各領域，都可透過語料庫真實語料進行

鑽研與探索，無論是進行質性語言結構剖析或是量化統計應用或數據歸納等，皆可讓語言研究擴展到更深更廣的範域。對於教學而言，教師可於教學過程中，運用語料庫檢索，讓學生觀察字、詞、句等資訊來理解語言的使用；可查詢語料庫中詞語和語法頻率以及詞彙之共現詞，作為語言教學之教材；或是可透過詞彙抓取語境前後文組成的句子，成為生詞教學的例句，有助於學生對於詞彙、句型的認知與習得；也可藉由詞頻的高低，作為詞彙或語法教學難易度分級的參考。[25] 再者，所有客語文本均予以屬性分類，此有助於使用者根據其研究需求篩選語料的後設資訊，如腔調、語式、文類、主題等，從中獲取所蘊藏的客家文化資訊與詞句用法，使客語文學與客家文字得以發展與傳承。而在翻譯層面，語料庫可協助華客翻譯者快速參照用字以及詞彙共現詞之搭配，並藉由大量真實自然語料輔助文法結構與詞彙之使用。此外，語料庫未來亦可應用於 AI 領域之自然語言處理，結合各項客語智能創新運用，如輔助客語教材編纂、提供辭典收集與編輯之材料、協助客語對譯的開發、訓練語音辨識的前導工程等。自然語言處理須經由語素分析、語法分析以及語意分析，方可讓人工智慧理解人類語言文字以及話語。目前語料庫進行的斷詞處理，即是為 AI 之自然語言處理建構藍圖。自然語言處理的目標在於訓練電腦了解與運用人類的語言，人類的語言承載思緒、意識，是極其複雜的行為表現，機器同時也須具備強大的靈活性才得以歸納並建立模型。語言在 AI 領域之應用，尚須更多人力及技術資源的投入，結合語言專家之領域知識及科技專家之技術，共同合作來克服與突破。

25 關於臺灣客語之詞頻研究可參閱葉秋杏等（2020）。

自然語言處理道阻且長，理解語言所需要的知識又是不斷變化且無止境，客語在這未知的領域迎接著諸多挑戰與困難。臺灣客語語料庫所建構的基礎內涵，記錄客語語言事實，其跨學科、跨領域、跨時代之特質，為客語在廣袤無垠的 AI 領域踏出了第一步。

參考文獻

中央研究院，2021，《中央研究院漢語平衡語料庫（第 4.0 版）》。http://asbc.iis.sinica.edu.tw/，取用日期：2021 年 1 月 9 日。

王勻芊，2016，《口語語料庫之建置典藏與應用：以臺灣客語口語語料庫為例》。國立政治大學圖書資訊與檔案學研究所碩士論文。

江俊龍，2010，《臺灣客家語語料庫之建置及應用》。臺北：行政院國家科學委員會補助專題研究計畫。

_____，2013，《東勢客語故事採集整理暨「臺灣客家語語料庫」的增建》。臺北：行政院國家科學委員會補助專題研究計畫。

李佩瑛等，2010，《語料庫建置入門數位化工作流程指南》。臺北：數位典藏拓展臺灣庫位典藏計畫。

邱各容，2015，〈2015 年臺灣童書出版觀察報告〉。《全國新書資訊月刊》206: 36-40。

客家委員會，2017，〈105 年度全國客家人口暨語言基礎資料調查研究〉。《客家委員會》。https://www.hakka.gov.tw/File/Attach/37585/File_73865.pdf，取用日期：2021 年 1 月 9 日。

國立臺灣大學語言學研究所，2021，《臺大臺灣南島語多媒體語料庫》。

http://203.66.168.190/index.asp，取用日期：2021 年 1 月 9 日。

國家教育研究院，2019，《「建置應用語料庫及標準體系」108 年工作計畫期末報告》。臺北：教育部華語文教育八年計畫（102-109）。

_____，2021，《國教院語料庫索引典系統（含國教院華語中介語索引典系統）》。https://coct.naer.edu.tw/cqpweb/，取用日期：2021 年 1 月 9 日。

國家圖書館，2020，《108 年臺灣圖書出版現況與趨勢報告》。臺北：國家圖書館。

教育部，2012，《客家語拼音方案使用手冊》。臺北市：教育部。

曾淑娟、劉怡芬，2002，《現代漢語口語對話語料庫標註系統說明》。臺北：中央研究院詞庫小組技術報告（編號：02-01）。

詞庫小組，1998，《中央研究院平衡語料庫的內容與說明（修訂版）》。臺北：中央研究院詞庫小組技術報告（編號：95-02/98-04）。

黃恒秋，2005，〈打造客語創作文學：現階段臺灣客語文學發展的觀察〉。http://ip194097.ntcu.edu.tw/giankiu/GTH/2005/TGBH/lunbun/3-3.pdf，取用日期：2021 年 1 月 9 日。

葉秋杏等，2020，〈臺灣客語語料庫建置與初步資料分析〉。論文發表於「第十一屆數位典藏與數位人文國際研討會」（DADH 2020），高雄市：中央研究院數位文化中心與臺灣數位人文學會主辦，12 月 1-4 日。

蔡素娟，2011，《臺灣閩南語兒童字詞統計分析》。http://www.ccunix.ccu.edu.tw/~lnglab/TAICORP.htm，取用日期：2021 年 1 月 9 日。

蔡素娟、麥傑，2013，《臺灣閩南語口語語料庫》。http://lngproc.ccu.

edu.tw/SouthernMinCorpus/，取用日期：2021 年 1 月 9 日。

賴惠玲，2008，〈面對瀕臨死亡的語言我們能做些什麼？——以臺灣客語為例〉。頁 77-101，收錄於李旺龍編，《閱讀科學大師 2》。臺南：成大出版社。

靜宜大學，2021，《蘭嶼達悟語口語資料典藏網》。http://yamiproject.cs.pu.edu.tw/yami/，取用日期：2021 年 1 月 9 日。

Archive of the Indigenous Languages of Latin America, 2002, revised 2015, 2017, *Archive of the Indigenous Languages of Latin America.* https://ailla.utexas.org/ (Date visited: September 6, 2021).

Chui, Kawai et al., 2017, "Taiwan Spoken Chinese Corpus." Pp. 257-259 in *Encyclopedia of Chinese Language and Linguistics*, edited by Rint Sybesma. Netherland: Koninklijke Brill NV.

Chui, Kawai, and Huei-ling Lai, 2008, "The NCCU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern Min." *Taiwan Journal of Linguistics* 6(2): 119-144.

Davies, Mark, 2008, *Corpus of Contemporary American English (COCA).* https://www.english-corpora.org/coca/ (Date visited: January 9, 2021).

Du Bois, John W. et al., 1993, "Outline of discourse transcription." Pp. 45-89 in *Talking Data: Transcription and Coding in Discourse Research*, edited by Jane A. Edwards and Martin D. Lampert. Hillsdale, NJ: Lawrence Erlbaum.

Green, Ian, and Rachel Nordlinger, 2021, *The Daly Languages (Australia).*

http://dalylanguages.org (Date visited: September 6, 2021).

Kučera, H., and W. N. Francis, 1967, *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

MacWhinney, B., & Wagner, J., 2010, "Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository." *Gesprachsforschung*, 11, 154-173. http://ca.talkbank.org/access/TaiwanMandarin.html

Mayer, Mercer, 1980, *Frog, where are you?* New York: Dial Books.

Moseley, Christopher, ed., 2010, *Atlas of the World's Languages in Danger*, 3rd ed.. Paris: UNESCO Publishing.

Simpson, R. C.et al., 2002, *The Michigan Corpus of Academic Spoken English.* Ann Arbor: Regents of the University of Michigan. https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase (Date visited: January 9, 2021).

Sung, Li-May et al., 2008, "Developing an Online Corpus of Formosan Languages." *Taiwan Journal of Linguistics* 6(2): 79-118.

Taiwanese Child Language Corpus, 2021, *Taiwanese Child Language Corpus*. https://phonbank.talkbank.org/access/Chinese/Taiwanese/Tsay.html (Date visited: anuary 9, 2021).

The British National Corpus, version 3 (BNC XML Edition), 2007, *The British National Corpus, version 3 (BNC XML Edition)*. http://www.natcorp.ox.ac.uk/ (Date visited: January 9, 2021).

The NCCU (National Chengchi University) Corpus of Spoken Taiwan

Mandarin (政治大學中文口語語料庫), 2021, *The NCCU (National Chengchi University) Corpus of Spoken Taiwan Mandarin (政治大學中文口語語料庫)*. http://spokentaiwanmandarin.nccu.edu.tw/ (Date visited: January 9, 2021).

Yeh, Chiou-shing, 2017, *The Emergence of Epistemicity and Subjectification: A Cognitive-Pragmatic Approach of Modality Disjuncts in Taiwan Hakka*. Ph.D. diss., Graduate Institute of Linguistics, National Chengchi University.